FinRegLab

The Use of Machine Learning for Credit Underwriting

Market & Data Science Context

SEPTEMBER 2021

About FinRegLab

FinRegLab is a nonprofit, nonpartisan innovation center that tests new technologies and data to inform public policy and drive the financial sector toward a responsible and inclusive financial marketplace. With our research insights, we facilitate discourse across the financial ecosystem to inform public policy and market practices.

Acknowledgments

This report is part of a broader research project on the explainability and fairness of machine learning in credit underwriting. This initiative includes empirical research being conducted with Laura Blattner and Jann Spiess at the Stanford Graduate School of Business. This report lays a foundation for that empirical research, as well as a publication which will investigate how law, regulation, and market practices may need to evolve to promote fair and responsible use of these new technologies. Many of the insights in this report were derived from interviews with stakeholders in the financial services, technology, and civil society sectors.

FinRegLab would also like to recognize members of our project Advisory Board who contributed to discussions of the issues covered in this report. The Advisory Board consists of subject matter experts from computer science, economics, finance, and regulatory backgrounds. The group represents more than 40 major entities from the worlds of bank and nonbank financial services, technology, policy, advocacy, and academia. State and federal regulators participate as observers.

We would also like to thank the following individuals who provided feedback on this report:

Michael Akinwumi, National Fair Housing Alliance; Brad Blower, National Community Reinvestment Coalition; Rohit Chauhan & Nitendra Rajput, Mastercard; Marsha Courchane & Peter Zorn, Charles River Associates; Steve Dickerson & Raghu Kulkarni, Discover Financial Services; Talia Gillis, Columbia Law School; Patrick Hall, bnh.ai; Stephen Hayes, Relman Colfax PLLC; Deborah Hellman, University of Virginia Law School; Scott Lundberg, Microsoft Research; Kevin Moss, Kevin Moss Consulting LLC; Beju Rao, Amruta Inc.; David Silberman; and Scott Zoldi.

We would also like to acknowledge the FinRegLab team who worked on the development of this report:

Kelly Thompson Cochran, Colin Foos, Hilary Griggs, Tess Johnson, Mashrur Khan, Duncan McElfresh, Kerrigan Molland, P-R Stark, YaYa Sun, and Sormeh Yazdi.

CONTENTS

1. INTRODUCTION

Every week, tens of thousands of consumers and small business owners have their applications for credit assessed and effectively decided by machine learning underwriting models.[1] The models' greater accuracy and capacity to analyze large, varied forms of data create the potential to increase access to credit for millions of people—including disproportionately high numbers of Black, Hispanic, and low-income consumers—who are difficult to assess using traditional models and information.[2]

The growing relevance of machine learning underwriting models and their potentially significant effect on consumer and small business credit markets have focused the attention of all stakeholders—firms, technologists, advocates, academics, and policymakers alike—on establishing whether and in what circumstances these models can be trusted for use in such a sensitive application. Some of these questions intensify long-standing concerns about the fairness of automated underwriting systems. Other questions focus on whether machine learning models may be more prone to performance deterioration in the face of changing data conditions. In both cases, these issues are heightened by some of the very qualities that fuel the greater accuracy of machine learning models—their ability to detect more complex relationships than prior generations of models.

In assessing both the reliability and fairness of machine learning underwriting models, model transparency emerges as an urgent threshold question for internal and external stakeholders. Without sufficient transparency, neither firms nor their regulators can evaluate whether particular models are making credit decisions based on strong, intuitive, and fair relationships between an applicant's behavior and creditworthiness. Yet the same complexity that fuels the accuracy of machine learning underwriting models can make it more difficult to understand how a model was developed and how it assessed a particular applicant's creditworthiness. Absent such understanding, lenders may not be able to mitigate aspects of a model that affect its reliability and fairness or establish compliance with a range of regulatory requirements that apply irrespective of the kind of underwriting model a lender chooses to use.

[1] Machine learning refers to the subset of artificial intelligence that gives "computers the ability to learn without being explicitly programmed." Artificial intelligence (AI) is a term coined in 1956 to describe computers that perform processes or tasks that "traditionally have required human intelligence." *See, e.g.*, Financial Stability Board, Artificial Intelligence and Machine Learning in Financial Services (2017); Ting Huang et al., The History of Artificial Intelligence, University of Washington (Dec. 2006); Arthur L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, 3 IBM J. of Research & Development 211-229 (1959); Tom Mitchell, Machine Learning (1997) (defining machine learning as the "field of study that gives computers the ability to learn without being explicitly programmed"); Michael Jordan & Tom Mitchell, Machine Learning: Trends, Perspectives, and Prospects, 349 Science 255-260 (2015) (defining machine learning as "the question of how to build computers that improve automatically through experience").

[2] For instance, more than 50 million U.S. adults lack sufficient traditional credit history to generate credit scores under the most widely used models, and prior to the COVID-19 pandemic more than 80 million adults may have struggled to access credit because they were considered "non-prime." Information and modelling limitations also make it more difficult for millions of small business owners to obtain credit. FinRegLab, The Use of Cash-Flow Data in Underwriting Credit: Market Context & Policy Analysis 12-14 (2020) (hereinafter FinRegLab, Cash-Flow Market Context & Policy Analysis).

For this reason, new approaches to enabling sufficient transparency to ensure fair and responsible use of complex models have taken on great prominence in debates about the trustworthiness of AI and machine learning systems. Assessing and measuring the trustworthiness of machine learning underwriting models is not a purely mathematical or technological problem. Nor is it a challenge unique to the financial services sector. But in financial services, emerging data science techniques are critical to addressing both the transparency questions about complex models and understanding whether such models can satisfy well-established regulatory expectations regarding reliability and fairness.

However, stakeholders without modelling expertise may not understand well the choices model developers make about how to use these techniques when designing and using machine learning underwriting models. This report sets forth the range of decisions while designing and implementing underwriting models that affect their reliability, fairness, and inclusiveness and emphasizes areas where the transition to machine learning has implications for various stakeholders. It lays the foundation for a ground-breaking evaluation of emerging market practices to foster fair and responsible use of machine learning underwriting models in consumer credit. FinRegLab's empirical research with Professors Laura Blattner and Jann Spiess of the Stanford Graduate School of Business will be the first public research shaped by input from key financial services stakeholders—including executives from banks and fintechs, technologists, consumer advocates, and regulators—to address questions about explainability and fairness that are likely to shape the nature and pace of adoption in the future.

FinRegLab's empirical research will evaluate how well a set of open-source and proprietary model diagnostic tools help lenders using machine learning models:

- Demonstrate the conceptual soundness, performance, and reliability of the models to satisfy prudential model risk management expectations;
- Identify, measure, and enable mitigation of fair lending risks, particularly whether models have a disparate impact on protected classes; and
- Provide applicants with individualized adverse action notices explaining why they were denied credit or offered less favorable terms.

These research questions involve a set of diverse requirements that apply to consumer lending regardless of the type of model being used to make credit decisions. Each one focuses attention on important aspects of model transparency and implicates foundational questions about the ability to explain, understand, and manage machine learning underwriting models.

The purpose of this research is to inform decision-making by policymakers, firms, industry groups, advocates, and researchers as the financial services sector develops norms and rules to govern the responsible, fair, and inclusive use of machine learning for credit underwriting. Examining the capabilities and performances of emerging model diagnostic tools in the context of comparatively stringent financial services requirements can also inform both the use and governance of machine learning in other sectors and the development of more effective data science techniques for explaining and understanding these models.

The policy component of this project will identify ways in which existing law, guidance, compliance assessment techniques, and market practices may need to evolve in light of the features, benefits, and limitations of currently available approaches to managing the explainability and fairness of machine learning underwriting models.

Report Overview: This report is designed to:

- Document the state of adoption of machine learning underwriting models for consumer credit and emerging approaches to enabling necessary transparency and managing compliance with a range of requirements focused on reliability and fairness;
- Define the context for FinRegLab's empirical research on the capabilities and performance of model diagnostic tools designed to help lenders responsibly use machine learning underwriting models; and
- Provide a resource to stakeholders, especially non-technical ones, that explains the techniques and tools available to lenders to design, operate, and manage machine learning underwriting models.

The report highlights specific choices that model developers make when designing and using machine learning underwriting models for two reasons. First, these choices have important implications for the transparency and fairness of machine learning models and for the effectiveness of oversight processes. Each decision that a model developer makes is reviewable by modelling peers, risk and compliance personnel, and regulators. But those opportunities may come at different points in the model lifecycle than for conventional models and may require more work and different tools to answer questions about the model's reliability and fairness for various stakeholders. Second, the emergence of an evolving set of techniques and tools designed to help lenders explain, understand, and manage complex models has made the use of machine learning for credit risk assessment more realistic for firms than it was just a few years ago. As a result, this report seeks to share ongoing debates and emerging practices as firms decide how to responsibly implement machine learning underwriting models.

Report Organization: This report has four main sections:

- Market Context: This section provides an overview of why firms are interested in using machine learning underwriting models and factors likely to affect adoption and use of these technologies, including key risks and regulatory considerations. It concludes with a survey of the state of adoption of machine learning underwriting models.
- Model Transparency: This section explores the importance of model transparency as a threshold question for the trustworthiness of machine learning models, the challenges of achieving transparency when using machine learning underwriting models, and the debate about whether machine learning underwriting models should be understandable without reliance on supplemental models, analyses, or techniques. It then considers options for model developers in designing models that can meet the transparency needs of credit underwriting.
- Modelling Considerations: This section provides a more detailed description of the decisions that individual lenders will make when designing, implementing, and operating machine learning underwriting models, including formative considerations like algorithm selection and data selection and preparation. These considerations are not necessarily unique to machine learning underwriting models, but warrant attention in this context due to their potential effect on the performance, fairness, and inclusiveness of the resulting models.
- Fairness and Bias: This section addresses a range of issues related to potential bias and discrimination in the context of algorithmic lending. It begins by setting out the ways in which models and data can each be the source of bias and considers options for reducing and measuring bias. It addresses methods for addressing bias in such models and emerging approaches to measuring fairness in the model development process.

The Market Context section presents the landscape related to the use of machine learning models in largely non-technical terms. Readers interested in more technical treatment of issues and debates introduced in that section will find that content in the subsequent sections. These more technical sections present a range of modelling considerations, including specific options available to model developers to enable model transparency, in roughly the same sequence that developers will address them and with an eye toward highlighting specific decisions made about the design and use of the model.

The conclusion looks ahead to open questions that policy, industry standards, or market practices might address to promote responsible, fair, and inclusive use of machine learning for credit underwriting.

The report also contains a glossary of terms and appendices providing greater detail on legal and regulatory requirements and data science fairness metrics to provide readers with a reference resource and additional background as they read particular sections of the main text.

Methodology: To prepare this report, FinRegLab conducted a series of interviews with industry leaders, regulators, consumer advocates, and others to explore issues related to use of machine learning, including types of models and explainability techniques currently in use, experiences with various forms of machine learning underwriting models, key challenges in designing and implementing machine learning underwriting models, and emerging risk management practices related to these models. In addition to interviews, FinRegLab gained insights from discussions with the Advisory Board that it has convened for purposes of the broader research project. This Advisory Board is composed of subject-matter experts from computer science, economics, financial services, and law, and represents more than 40 major institutions from all relevant sectors: bank and nonbank financial services, technology, policy, advocacy, and academia. State and federal regulators participate on the Advisory Board as observers. FinRegLab reviewed academic and industry literature on a range of issues related to topics such as machine learning interpretability and explainability, algorithmic fairness, trustworthy and ethical AI, financial inclusion, and regulatory compliance.

Key Findings: FinRegLab's survey of market practices suggests bank and nonbank lenders are currently using machine learning underwriting models and that many more firms across the market are looking closely at adopting them. In particular, this report finds:

- Motivations to Use Machine Learning: Lenders are attracted to machine learning models' potential to improve the accuracy of credit risk assessment and reduce losses, to speed up the process of updating and refitting models, and to keep pace with market competitors. Many also cite the ability of machine learning models to leverage large, diverse datasets as a motivation. Nonbank usage is more established due to a number of factors, including reliance on digital business models, newer lending platforms, and differences in the nature and maturity of risk management and oversight processes.
- Usage by Market: Credit cards and unsecured personal loans are the markets in which use of machine learning models to make credit decisions is most advanced. This reflects the historical position of credit cards as being at the analytical forefront of consumer finance and the dominance of digital lending in unsecured personal loans. Auto lending and small business lending are also areas where machine learning underwriting models are in use.[3]
- Importance of Transparency: Irrespective of the form of machine learning used, stakeholders of all kinds agree that a high degree of transparency is needed when machine learning is used to make credit decisions in order to enable appropriate management and

[3] *See* Megan Jarrell, Artificial Intelligence at Square—Two Use-Cases, Emerj (Sept. 6, 2021).

oversight. Accordingly, as set forth in Section 3, concerns about the ability to explain and understand complex models shape lenders' decisions at every stage of the process of developing, implementing, and managing machine learning underwriting models.

- » Enabling Transparency: In light of these explainability concerns, some firms impose constraints to reduce the complexity of the resulting model and improve its transparency. Other lenders opt to use *post hoc* explainability techniques—supplemental models, analyses, or methods—to make complex or black box models more transparent. Whether constraints unduly inhibit the performance of machine learning underwriting models is contested in academia and industry. So are the capabilities, performance, and trustworthiness of common *post hoc* explainability techniques.
- » In-House Development: Decisions about whether to develop machine learning underwriting models or supplemental model diagnostic tools in-house or to rely on vendors depend on the overall size of the lender and the size and technical sophistication of the business unit considering adoption of machine learning. Many firms are likely to lack the resources—foremost among them personnel with appropriate data science and credit expertise—to develop and operate such models on their own.[4]
- » Role of Third-Parties: To meet this need, a number of third-party providers have entered this market, including score providers, technology firms, and consulting firms. The business models and offerings of these providers vary. Some offer model diagnostic tools as a stand-alone product, while others provide those tools in the context of model development services.
- » Fairness Implications: Firms and regulators are also focusing on whether and in what circumstances the use of machine learning can improve fair lending oversight, including improving available tradeoffs between performance and fairness when mitigating sources of adverse impacts in credit decisions.

Yet while interest in machine learning underwriting models is accelerating, the scope and pace of adoption going forward will depend on the extent to which a broad range of stakeholders can answer fundamental questions about the trustworthiness of machine learning models, including how to enable necessary oversight. Concerns about the trustworthiness of machine learning models are being raised in a broad range of sectors with regard to general transparency, reliability, fairness, privacy, and security. But they are particularly pressing in credit underwriting because existing legal and regulatory frameworks force consideration of risk management questions more holistically and at an earlier stage than occurs elsewhere. The balance of the report focuses on outlining the choices that lenders face in developing, implementing, and monitoring machine learning models and emerging developments on explainability and fairness from the broader data science community that may help to shape market and regulatory practices concerning machine learning underwriting models.

Building on these market and data science developments, the conclusion briefly looks ahead to open questions that policymakers or firms might address to promote fair and responsible use of machine learning for credit underwriting. In addition to the forthcoming empirical evaluation of several model diagnostic tools, future reports will explore the potential evolution of law, policy, regulation, and market practice in greater detail.

[4] According to a 2020 survey of 175 Lendit subscribers, across both large and small institutions, approximately 20% of institutions had no in-house staff for credit modelling and relied on third parties to conduct such activities. Even large institutions with credit modelling teams did not devote significant resources, as just 16% of large institutions had four or more full time modelers. Cornerstone Advisors, Credit Monitoring and the Need for Speed: The Case for Advanced Technologies 4, Figure 4 (Q2 2020).

2. MARKET CONTEXT

Lenders have relied on automated systems to assess applications for consumer and small business credit for decades. Reliance on such underwriting systems has brought significant marketwide benefits: reduced defaults, underwriting costs, and loan pricing; expanded access to credit; improved consistency of treatment of similarly situated borrowers; and increased competition for borrowers.[5] But these benefits are not distributed evenly. For instance, a particular limitation of automated underwriting systems is that they require standardized information about credit applicants that can be obtained relatively quickly and inexpensively. But the credit information system that has evolved to support automated underwriting lacks sufficient information on about 20% of U.S. adults to generate scores under the most widely used models.[6] Communities of color and low-income populations are substantially more likely to be affected by these information barriers than other applicants. For example, nearly 30% of Black and Hispanic people cannot be scored using the most widely adopted credit scoring models, compared to about 16% of whites and Asians.[7]

As advances in computing power and creation of digital information have accelerated, interest in using machine learning to develop credit underwriting models is also increasing. Traditional automated underwriting systems rely on linear and logistic regression which identifies a set of variables with the strongest correlation to a particular outcome (such as loan performance) and assigns a weight to each variable in the model. Machine learning models are able to incorporate large volumes of diverse types of data into their analysis and discern relationships that may not be detectable through incumbent regression models, thereby generating more accurate predictions.

[5] Board of Governors of the Federal Reserve System, Report to Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit S-2 to S-4, O-2 to O-4, 32-49 (2007); Allen N. Berger & W. Scott Frame, Small Business Credit Scoring and Credit Availability, 47 J. of Small Bus. Mgmt. 5-22 (2007); Susan Wharton Gates *et al.*, Automated Underwriting in Mortgage Lending: Good News for the Underserved?, 13 Housing Policy Debate 369-391 (2002); FinRegLab, Cash-Flow Market Context & Policy Analysis at 11 n.16.

[6] Consumer Financial Protection Bureau, Data Point, Credit Invisibles at 4-6, 17 (2015); FinRegLab, Cash-Flow Market Context & Policy Analysis § 2.2. Small business owners and applicants with marred credit histories are also groups that face particular information barriers and access challenges. FinRegLab, Cash-Flow Market Context & Policy Analysis § 2.2. The risk that conventional credit scoring models err in predicting default risk also appears higher for consumers with relatively thin credit files. Laura Blattner & Scott Nelson, How Costly Is Noise? Data and Disparities in Consumer Credit, arXiv:2105.07554v1 (2021); *see also* Oportun, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning 2 (Jul. 1, 2021) (estimating based on internal analyses that 55 million consumers with limited credit histories may be misscored).

[7] Racial disparities regarding access to credit are far greater than for more basic transaction accounts, for instance. For example, a 2017 Federal Deposit Insurance Corporation survey found that about 10% of Black and Hispanic households lacked bank and/or prepaid accounts, while more than 30% of both groups reported not having mainstream credit accounts of the type that are likely to be reported to credit bureaus. FinRegLab, Market Context & Policy Analysis § 2.2; Federal Deposit Insurance Corporation, 2017 National Survey of Unbanked and Underbanked Households (2018). In a subsequent survey, the percentage of Black and Hispanic households who lack bank or prepaid accounts stayed about the same, but the question about credit was not repeated. Federal Deposit Insurance Corporation, 2019 National Survey of Unbanked and Underbanked Households (2020).

Such models and data have been used by lenders for decades to detect fraud in credit card transactions and more recently to identify new variables for use in traditional underwriting models through a process known as feature engineering.[8] However, adoption of machine learning models to make credit decisions has been slower than in other sectors due to a combination of business and regulatory factors that intensify concerns about the trustworthiness of predictive models.

This overview of the market context for use of machine learning underwriting models addresses: motivations for using these models; key issues related to adoption, including risks and regulatory compliance considerations; and the state of usage by lender type, model type, and product market. Subsequent sections of the report examine in greater detail a range of model design and implementation choices available to individual firms, such as forms of machine learning relevant to underwriting models and options for enabling transparency in those models. The Glossary provides simple definitions for key terms that are referenced in Section 2 and discussed in greater depth in later sections.

2.1 Motivations for Using Machine Learning Underwriting Models

Lenders and credit score developers are exploring the use of machine learning techniques to improve their ability to predict credit risk, with or without new data sources. These techniques have the potential to achieve benefits which, if realized, would serve goals broadly shared by borrowers, firms, policymakers, and investors alike:[9]

- Expanding access to more borrowers who are creditworthy and reducing the number of people who are offered credit on terms that they are unlikely to be able to repay by improving the accuracy of predictions of default risk
- Reducing default rates and losses
- Reducing mispricing based on inaccurate estimation of the likelihood of default and improving terms at which credit is offered to some applicants
- Improving identification and mitigation of certain forms of discriminatory lending
- Facilitating less costly and faster model generation and updating

Many of these benefits may be enhanced where lenders also incorporate new forms of credit information—particularly alternative financial information such as cash-flow information from bank or prepaid accounts (see Box 2.1).[10] This is especially likely to be true as to potential inclusion benefits of using machine learning to assess applicants whose creditworthiness is not accurately scored or cannot be assessed using incumbent models and data sources.[11]

[8] Throughout this report, the terms variable, feature, and attribute are used as synonyms."

[9] FinRegLab, Cash-Flow Market Context & Policy Analysis at 8-12; CFPB, Credit Invisibles at 4-6; Peter Carroll & Saba Rehmani, Alternative Data and the Unbanked, Oliver Wyman (2017).

[10] *See, e.g.,* Peter Rudegeair & AnnaMaria Andriotis, JPMorgan, Others Plan to Issue Credit Cards to People With No Credit Scores, Wall St. J. (May 13, 2021) (announcing an effort by a group of large U.S. banks to utilize alternative financial data such as deposit account data to extend credit to applicants with no or thin traditional credit history); Brendan Pedersen, OCC Announces Initiative to Expand Credit Access in Los Angeles, Am. Banker (Oct. 30, 2020). Because machine learning models can use more information to determine creditworthiness, it is easier for lenders to incorporate alternative financial information into the modelling process. BLDS, LLC *et al.*, Machine Learning: Considerations for Fairly and Transparently Expanding Access to Credit 6 (2020).

[11] FinRegLab, Cash-Flow Market Context & Policy Analysis § 2.2; Blattner & Nelson.

BOX 2.1 SOURCES OF CREDIT INFORMATION

The shift to machine learning underwriting models may be tied both to diversifying the kinds of data used for credit risk assessment and improving the inclusiveness of lending decisions. As discussed in more detail in Section 4.2.1.1, several types of information are attracting interest for use in credit underwriting:

- **Credit Information:** Traditional credit reporting agency records typically contain applicants' personal information; public records such as bankruptcies, tradeline data which reflect that person's repayment record mainly for secured and unsecured loans; inquiries made on the applicant's credit files; and loan balance information.[a]
- **Alternative Financial Data:** Alternative financial data refers to categories of non-lending financial activity that traditional credit bureau information does not contain, such as inflows and outflows from bank or prepaid accounts.
- **Behavioral Insights in Alternative Financial Data:** This refers to information about consumers' behavior that is derived from transaction-level financial information and includes information such as where and when they shop and in some circumstances what they buy. This may also include segregating transactions into discretionary purchases and tracking metrics designed to show how an individual manages those against fluctuations in income.
- **Non-Financial Alternative Data:** This group refers broadly to data about a person's activities that are not financial in nature or derived from financial data. Social media data, search histories, and social connectedness metrics are common examples.[b]

While researchers and lenders in some developing countries are focusing on non-financial alternative data sources such as cell phone use, lenders in the U.S. are generally concentrating on alternative financial information.

[a] Carroll & Rehmani.
[b] *See* Agarwal *et al.*; Berg *et al.*

Interest in data diversification is fueling some of the interest in machine learning given its superior ability to analyze large volumes of data and data of different kinds.[12] Even for lenders who are not primarily motivated by data considerations, the operational overhaul required for widespread use of machine learning models likely creates a rare opportunity to reset lending platforms to manage use of alternative data types and feeds from a variety of sources with relatively little additional cost.

The adoption of machine learning underwriting models is also intensifying interest in other initiatives to improve the efficiency, fairness, and inclusiveness of lending. For example, the technology sector's focus on how to enable appropriate explanations of machine learning models' predictions has focused attention on whether lenders can and should improve the information given to unsuccessful credit applicants by providing more actionable information. Generating more actionable information in this context does not necessarily require machine learning, but the advent of machine learning underwriting models may nevertheless drive the industry toward this practice.

The remainder of this subsection considers in greater detail lenders' motivations for using machine learning underwriting models in the following areas: performance; fairness and inclusion; consumer protection and empowerment; and operational efficiency. Whether and in what form individual firms realize any of these benefits will depend on how they view the business rationale for using machine learning in a specific context and specific choices they make in developing, implementing, and using particular models.

2.1.1 Performance

Machine learning models' potential to provide more accurate predictions of applicants' likelihood of default is a primary motivation for firms that have replaced or are considering replacing

[12] BLDS, LLC *et al.* at 6.

BOX 2.1.1 RISK-BASED PRICING

Risk-based pricing refers to a common approach for determining the cost of credit. In this approach, lenders base the price of their offer of credit on their estimate of an individual applicant's likelihood of default. This generally means that applicants with a good credit score will be offered lower interest rates, whereas those who have previously fallen behind on loan payments or declared bankruptcy will receive more expensive offers for loans of the same kind and amount.

Stakeholders fiercely debate the fairness and inclusion effects of risk-based pricing. Proponents argue that it has increased access to credit by giving lenders confidence that they can cover somewhat higher losses when extending credit to somewhat riskier borrowers that they would otherwise reject. Critics counter that risk-based pricing systems increase the likelihood of default for higher-risk borrowers because loan payments are less affordable and sometimes impose higher charges than necessary to cover lenders' losses.

Adoption of machine learning may heighten the strengths and weaknesses of risk-based pricing. The ability of machine learning underwriting models to assess more data and data of different kinds and to identify more predictive relationships can result in more accurate credit risk assessments, especially for those who are hard to score using traditional methods and data. This improved accuracy may ultimately lower the cost of credit for some consumers and improve access to credit for others, just as it did with the onset of automated underwriting in the 1970s.

But the improved accuracy of machine learning models could lead to some applicants being assessed as higher risk than under traditional models, causing their cost of credit to increase or their applications to be denied. Some studies suggest that pricing disparities for particular groups could increase at the same time that approval disparities decline if machine learning models predict that previously excluded applicants are at somewhat higher risk of default.[a] More research is needed to understand how these effects could play out in practice, especially with respect to shifts within and across communities that are most deeply affected by flaws in the current system. If deeper insights into customer risk intensify the effects of risk-based pricing on vulnerable populations, stakeholders will need to focus on how to respond to the credit needs of those populations.

[a] *See* Fuster *et al.*

incumbent underwriting models. At its best, improved accuracy will result in fewer borrowers being offered loans they are unlikely to be able to repay and more qualified borrowers being approved for credit. Given that the decision to provide access to credit and the terms on which credit will be provided depend on predictions of default risk (see Box 2.1.1), even small improvements in predictive accuracy can produce wide-ranging benefits for firms, borrowers, and some applicants for credit.

Some public research suggests that machine learning underwriting models can offer significant improvements in performance. For example, several studies have found substantial gains in terms of predictiveness and cost savings for lenders from using machine learning models relative to conventional models. A study assessing commercial loans in Greece showed that two types of machine learning models outperformed logistic regression models in assessing credit, and a U.S.-based mortgage study found a machine learning model to outperform both linear and non-linear logistic regression models.[13] Similarly, some lenders that use machine learning underwriting models have conducted their own analyses, with one such study finding that a machine learning underwriting model would have resulted in 75% fewer defaults than the models used by three large U.S. banks.[14] Further, a study looking at machine learning models that use both credit bureau and transaction account data between 2005-2009 for the United States found that use of machine learning models

[13] Anastasios Petropoulos *et al.*, A Robust Machine Learning Approach for Credit Risk Analysis of Large Loan Level Data Sets Using Deep Learning and Extreme Gradient Boosting, Bank for International Settlements (2018) (showing that a common performance metric—the area under the ROC curve—for the study's neural network model was 9% higher than its logit model, and 18% higher for its extreme gradient boosting model); *See also* Andreas Fuster *et al.*, Predictably Unequal? The Effects of Machine Learning on Credit Markets, J. of Finance (forthcoming) (Jun. 21, 2021) (finding that a random forest model outperformed linear and non-linear logistic regression models by 1.4% and 0.8% respectively in terms of AUC when using mortgage data for the United States).

[14] Upstart, Results to Date § 2, upstart.com (visited Jul. 29, 2021).

can result in cost savings from accurately predicting the risk profiles of credit card borrowers, and under a conservative set of assumptions, estimated the savings to be 6 to 25% of total losses.[15]

For consumers, lending that uses more accurate risk assessment methods can improve access for certain individuals or groups and reduce the cost of certain products for others.[16] For firms, improved performance in making default predictions can translate to reduced costs and opportunities to expand lending within existing customer segments and to new customer segments, especially those not well served by existing risk assessment methodologies. Both sets of considerations may also be attractive to investors, whether they provide funding directly or through securitization markets.

2.1.2 Fairness and Inclusion

The potential for machine learning models to help improve the fairness and inclusiveness of lending decisions appeals to a broad range of stakeholders—lenders, risk information service providers, advocates, academics, regulators and policymakers. Stakeholders are particularly focused on the potential to improve credit risk assessment and lending decisions with respect to applicants with little to no prior credit history and those whose credit history is marred, both of which struggle to access affordable credit under current models and data sources. For these groups, the higher precision or accuracy that machine learning models can achieve may be important to the development of business models that support lending across a broader swath of the risk spectrum. For example, VantageScore Solutions reports that its use of machine learning to assess consumers who are not scorable under some third-party models because their credit histories have not had an update in the prior six months resulted in an accuracy improvement of 16.6% for bank card originations and 12.5% improvement for auto loan originations.[17]

Stakeholders are particularly focused on the potential of machine learning methods to help improve the fairness and inclusiveness of lending decisions in two additional ways: facilitating use of non-traditional credit data and improving identification and mitigation of disparities, especially with respect to assessing less discriminatory alternatives. As discussed above, the desire to assess larger datasets and incorporate less standardized information is an important general motivation for adopting machine learning underwriting models.[18] The transition to machine learning underwriting models does not in the first instance necessitate a change in the information being assessed when reviewing applications for credit. But, as discussed further in Section 4.2, the move to more sophisticated and flexible analytical methods could facilitate data diversification over time, which may significantly improve firms' ability to responsibly serve populations at the margins of current lending practices.[19]

[15] Amir E. Khandani *et al.*, Consumer Credit-Risk Models via Machine-Learning Algorithms, 34 J. of Banking & Finance 3 (2010); *see also* Florentin Butaru *et al.*, Risk and Risk Management in the Credit Card Industry, 72 J. of Banking & Finance 218-239 (2016) (finding decision tree and random forest models that considered tradeline data, credit bureau information, and macroeconomic indicators each outperformed a more traditional logistic regression model when forecasting credit card delinquencies).

[16] Upstart, for example, has argued that improved accuracy from its use of AI in underwriting decisions has allowed it to extend credit to those who would otherwise be left out by traditional models and provide those consumers with more favorable pricing. Upstart, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (Jul. 1, 2021). An internal study conducted by Upstart found that it was able to offer 50% more loans to consumers with an income less than $50,000 than its benchmark group. Upstart, Blog, Upstart By the Numbers (undated). Oportun, which estimates that it assisted more than 900,000 consumers who lacked FICO scores begin to build credit history since its inception, reports that it has developed machine learning models based on alternative data, credit bureau records, and proprietary historical data that can score 100% of applicants. Oportun, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning at 2-3.

[17] VantageScore, Our Models, vantagescore.com (visited Jul. 29, 2021).

[18] BLDS, LLC *et al.* at 6.

[19] Blattner & Nelson (suggesting that the combination of machine learning underwriting models and alternative data, such as cash-flow data, is required for greater financial inclusion).

Adoption of machine learning underwriting models may also have the potential to improve identification of discrimination risks and to offer superior mitigation options when those risks are detected.[20] This may result in detecting risks not registering fully in current processes and enabling the use of models that retain the predictive power of variables and relationships causing disparities instead of having to eliminate those features entirely. Unlike incumbent underwriting models, the development of machine learning models enables consideration of many iterations of a model, including many changes to a model's specifications, which can enhance predictive power and enable more explicit consideration of certain tradeoffs.[21] Lenders can assess those iterations to find "less discriminatory models that maintain their predictive ability."[22] The transition to machine learning is also inspiring consideration of how to incorporate growing sophistication in approaches to measuring algorithmic fairness in model development and oversight processes.[23]

2.1.3 Consumer Empowerment

The use of machine learning underwriting models can serve the broader purpose of empowering consumers. In the most basic sense, this can occur by expanding access to core foundational products and services and improving the terms on which they are provided. But debates about responsible use of machine learning underwriting models have also focused attention on potential improvements to disclosures that lenders must send customers who have been denied credit or offered it on materially less favorable terms than other applicants. What would constitute more actionable information is open to debate but a more expansive understanding of this concept points to including information on these disclosures that could help consumers understand how changes in their financial behavior and positioning could lead to more favorable credit decisions in the future.

2.1.4 Operational Efficiency

Use of machine learning underwriting models can improve the speed and efficiency of model development in a variety of respects. As noted above, use of machine learning enables generation and evaluation of a larger set of options than is feasible with conventional models, which can give developers a broader set of options to build more accurate and fairer models. It can also improve the agility of firms' ability to adapt to changing conditions. Machine learning models may reduce the need for additional underwriting guidelines—often called overlays—because their ability to assess a greater number of features and more complex relationships among features can reflect changes in lending conditions—through for example trended analysis showing interplay of the model and macroeconomic variations—that are difficult to incorporate directly into incumbent models. Finally, although implementing updates to credit underwriting models in response to economic shifts still requires accumulation of significant amounts of data, the process of creating the code automatically for models and model updates in most cases accelerates the effort to get new models ready for use. This agility can be particularly important in response to sudden shocks and useful in managing credit line changes and capital analyses.

[20] Florian Ostmann & Cosmina Dorobantu, AI in Financial Services, The Alan Turing Institute 37 (2021).

[21] BLDS, LLC *et al.* at 6.

[22] *Id.* at 22.

[23] *See* Section 5.2.

2.2 Risks and Trustworthiness Concerns

While machine learning's potential benefits in credit underwriting are appealing, these technologies have also intensified long-standing debates about automated decision-making. Data scientists, academics, industry practitioners, advocates, and policymakers are all turning their attention to how to identify and measure the trustworthiness of AI and machine learning systems. This inquiry is broader than credit or financial services alone (see Box 2.2), but examining core concerns about the potential risks of complex underwriting models offers a compelling case study in that larger conversation, given that extensive legal and regulatory frameworks force consideration of questions about machine learning's trustworthiness more holistically and at an earlier stage than occurs in other sectors.

This section considers the elements of trustworthiness that are of most concern with regard to the adoption of machine learning underwriting models, including a discussion of transparency both as an element that is important in its own right and as instrumental to diagnosing and managing other risks. Section 2.3 provides an overview of the existing regulatory frameworks for managing regulatory risks that make model transparency particularly important in managing machine learning underwriting models.

2.2.1 Performance

Performance is a fundamental element of trustworthiness since there is little reason to adopt machine learning models if they do not improve on the accuracy of incumbent systems in predicting default risk or other key outcomes. Beyond evaluating whether a particular model's predictions meet the accuracy needs for its use case, a second key aspect of performance relates to the robustness of the model's performance in unexpected conditions.

On this second aspect of reliability, machine learning models' ability to identify a wider range of relationships in training data than incumbent models may increase their susceptibility to performance problems due to two issues: (1) overfitting, or the risk that the machine learning algorithm fits the predictive model too narrowly to the specific characteristics of training data; and (2) data drift, which can occur when conditions in deployment start to differ from the data on which a model was trained, for instance due to shifts in consumer behavior, populations, or economic conditions.

2.2.2 Fairness and Inclusion

Concerns about whether and how machine learning underwriting models could negatively impact populations who have historically been subject to discrimination, exclusion, or other disadvantage are a second component of trustworthiness. This concern is broader than establishing compliance with anti-discrimination laws and includes more fundamental questions about data gaps, modelling decisions, and other issues that can affect the performance of models for particular groups.

Machine learning models' ability to identify a wider range of relationships in training data also heightens concerns about the risk of replicating or even amplifying historical disparities in the credit context. For instance, some models rely on "latent features" that are identified by the learning algorithms from relationships in the input data rather than intentionally programmed into the models by developers. This raises concerns that the models could reverse engineer applicants' race or gender from correlations in input data or create complex variables that have disproportionately negative effects on particular groups, but that developers would have difficulty diagnosing or mitigating such problems due to the complexity of the models.

BOX 2.2 INTERNATIONAL EFFORTS TO ASSESS THE TRUSTWORTHINESS OF AI AND MACHINE LEARNING SYSTEMS

Data scientists, academics, industry practitioners, advocates, and policymakers across multiple economic sectors and countries are turning their attention to how to identify and measure the trustworthiness of AI and machine learning systems. These inquiries are seeking to harmonize foundational principles with two objectives. First, this dialogue can support development of a broad consensus about the responsible use of AI and machine learning across jurisdictions, markets, and use cases. Second, that consensus can in turn serve as the basis for efforts to articulate and adapt sector-specific technological standards and regulatory requirements to encourage responsible adoption and use.

The elements of trustworthiness that are discussed in the main text—performance, fairness and inclusion, privacy and other consumer protections, security, and transparency—are among the most common cited across these various initiatives. The European Union's recent proposed regulations to "promote trustworthy AI that is consistent with Union values and interests" build on a 2019 European Commission formulation of seven key requirements for trustworthy AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability.[a] Other general frameworks for ethical/trustworthy AI include Ostmann & Dorobantu (five principles of AI ethics, which include fairness, sustainability, safety, accountability, and transparency); Brian Stanton & Theodore Jensen, Trust and Artificial Intelligence, National Institute of Standards and Technology (Dec. 2020) (nine characteristics of trustworthy AI: accuracy, reliability, resiliency, objectivity, security, explainability, safety, accountability and privacy); Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence (2019) (six key principles: inclusive growth, sustainable development and well-being; human-centered values and fairness; transparency and explainability; robustness, security and safety; and accountability).

In time, policy processes are needed to promulgate broadly applicable approaches to what transparency or fairness, for example, should mean in the context of responsibly using AI and machine learning systems. In considering these questions, it is likely that the standards that emerge will apply even to traditional models and thus the debates renewed and intensified by the adoption of this machine learning can drive the broader financial system and other sectors to enhanced efficiency, fairness, and inclusiveness.

[a] *See* European Commission, Proposal for a Regulation Laying Down Rules on Artificial Intelligence (2021); European Commission, Building Trust in Human Centric Artificial Intelligence (2019).

Recent research also questions whether using machine learning underwriting models without diversified data will produce substantial inclusion effects. One academic study of machine learning underwriting models using conventional data to assess applicants for mortgages concluded that modest improvements in application approvals among Black and Hispanic applicants would likely be more than offset by increased pricing differentials for those groups where risk-based pricing is used (see Box 2.1.1).[24] Another academic study shows that credit scores for minority groups generally reflect significantly more signal noise—that is, they are subject to more random, unpredictable errors that make it hard to isolate their effect on default risk—than other applicants. This in turn may limit the inclusion effects of using machine learning models without data diversification.[25]

The transition to machine learning may also impede the effectiveness of fair lending oversight. The task of assessing and mitigating the effects of variables or relationships used by machine learning models is potentially more complex than in incumbent models, which may make efforts to identify disparate impact risks less effective than with conventional models.[26]

[24] Fuster *et al.*

[25] Blattner & Nelson.

[26] See Section 2.3.2 of this report for a detailed description of disparate impact.

2.2.3 Privacy and Other Consumer Protections

The ability of machine learning underwriting models to analyze large, diverse datasets and create deeper, more personalized profiles of consumers is closely tied to their potential benefits for accuracy and inclusion but can also raise significant questions about privacy, fairness, and data protections. This is especially true where the models use elements that feel personally intrusive or lack an intuitive link to creditworthiness.

In credit underwriting, for example, where models rely on behavior with unintuitive connections to creditworthiness, consumers may lack meaningful opportunities to anticipate the relationship between their behavior and future assessments of their creditworthiness.[27] To the extent that inclusion or other benefits of machine learning underwriting models depend on an applicant's consent to make certain kinds of data available for credit risk assessment, this may raise fairness questions as to those unable or unwilling to do so.[28] Other concerns include what constitutes informed consumer consent to data access, rules regarding data retention and use, and provision of explanations of decisions made by automated systems in both identifying errors and enabling consumers to manage how their data are being used.[29]

While these issues are not unique to machine learning underwriting models specifically, they receive heightened attention in the machine learning context due to strong interest in pairing advanced analytical techniques with non-traditional data sources. Challenges with regard to the complexity and explainability of machine learning models also increase concern about what data are being used for what purposes.

2.2.4 Security

The potential for machine learning models to rely on more granular and sensitive data also can heighten concerns about information security.[30] In addition to increasing the potential consequences of security breaches, for instance, stakeholders have identified novel risks in some other machine learning contexts. For example, research suggests that AI systems can be manipulated without direct access to their code,[31] for example by maliciously embedding signals in social network feeds or news

[27] Ostmann & Dorobantu at 40.

[28] This issue has been raised as a point of concern with respect to data access and non-machine learning models. FinRegLab, Cash-Flow Market Context & Policy Analysis § 6.3 (considering circumstances when underserved borrowers are asked to provide this information in order to access credit while prime borrowers can rely on traditional credit data); *see also* Ostmann & Dorobantu at 40.

[29] For example, experts continue to debate the extent to which the European Union's General Data Protection Regulation (GDPR) implies a "right to explanation" related to the use of AI or machine learning models. *See* Riccardo Guidotti *et al.*, A Survey of Methods for Explaining Black Box Models, 51 ACM Computing Surveys art. 93 at 2 (2018) (assessing whether Articles 13-15 and Article 22 of the GDPR at least implicitly require a "right to explanation"); *see also* Sandra Wachter *et al.*, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law (2017) (posits that the GDPR lacks precise language and explicit and well-defined rights and safeguards concerning automated decision-making needed to bolster a right to explanation); Andrew Burt, Is There a 'Right to Explanation' for Machine Learning in the GDPR?, International Association of Privacy Professionals (2017) (provides textual analysis of the GDPR with respect to a right to explanation for machine learning and recommendations to ensure compliance in the deployment of machine learning systems).

[30] *See generally* Andrew Burt & Patrick Hall, What to Do When AI Fails, O'Reilly Radar (2020) (outlines a broad framework for responding to "*any* behavior by the model with the potential to cause harm, expected or not," including both potential privacy and security violations and incorrect predictions); Sophie Stalla-Bourdillon *et al.,* Warning Signs: The Future of Privacy and Security in an Age of Machine Learning, Future of Privacy Forum (2019) (outlines a risk-based framework for privacy and security standards in machine learning systems and suggests potential mitigation strategies).

[31] Nicolas Papernot *et al.*, Practical Black-Box Attacks against Machine Learning, Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security 506-519 (2017) (demonstrating how machine learning models, in this case deep neural networks, are vulnerable to black-box adversarial attacks).

feeds that are not detectable by humans.[32] Further, because machine learning models encode aspects of training data into the mechanisms by which they operate, they have the potential to expose private or sensitive information from the training data to users.[33] In consumer credit there may be additional concerns that certain required explanations about a model's predictions could expose information about the individual applicants or the underwriting model used to assess their creditworthiness.[34]

2.2.5 Transparency

Model transparency—the ability of stakeholders in a particular model to access various kinds of information about its design, use, and performance—is a final critical element in trustworthiness. Transparency is an important element of trustworthiness in its own right, because it increases stakeholders' confidence in procedural fairness, consistency, and accountability. In consumer credit, laws requiring lenders to provide applicants with a list of the principal reasons for adverse decisions serve this aim by enabling error correction in underlying credit information and educating consumers about what factors may affect their ability to access credit over time.

Transparency is also often critical for helping to diagnose and manage other aspects of trustworthiness, such as reliability and fairness. For example, understanding which variables are driving model outcomes can be important to assessing a model's potential sensitivity to changing conditions and to diagnosing and mitigating the sources of demographic disparities in model outcomes. As discussed further in Section 2.3, several regulatory regimes for consumer credit implicitly rely on multiple concepts of transparency to manage various risk concerns.

However, depending on their structure, data sources, and other factors, machine learning models can raise additional transparency challenges relative to incumbent models because of their size, complexity, and reliance on unintuitive data relationships. While data science techniques have produced a range of supplemental tools to increase the transparency of complex models, use of those techniques raises additional trustworthiness questions in their own right. Thus, as discussed in Section 3, questions about the ability of lenders to deliver accurate and useful adverse action notices to consumers and to generate explanations that equip a range of other stakeholders to assess and mitigate various types of reliability and fairness risks have proven to be the central question regarding the fair and responsible use of machine learning underwriting models.

2.3 Existing Regulatory Frameworks for Managing Reliability, Fairness, and Transparency

Credit underwriting is a particularly compelling use case for evaluating the trustworthiness of machine learning models because pre-existing regulatory and policy frameworks apply to all aspects of designing and using underwriting models and serve important policy goals in promoting prudent and fair lending regardless of the type of model used to make credit decisions. These frameworks are designed to assure the reliability of credit decisions, promote responsible risk-taking, prohibit

[32] Valeriia Cherepanova *et al.*, LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition, published as a conference paper at the 2021 International Conference on Learning Representations, arXiv:2101.07922 (2021) (showing how black-box adversarial attacks can be deployed to degrade the accuracy of third-party facial recognition models on social media platforms).

[33] *See* Patrick Hall, Proposals for Model Security: Fair and Private Models, Whitehat and Forensic Model Debugging, and Common Sense (2019) (highlights how surrogate models can be used to extract unauthorized information from a model through inversion or a membership inference attack).

[34] Smitha Milli *et al.*, Model Reconstruction from Model Explanations, arXiv:1807.05185v1 (2018) (certain model explanation methods, particularly gradient-based explanations, can reveal enough information about an underlying model so that it can be reconstructed or so that one could obtain sensitive training data from it).

discrimination, and provide consumers information about financial decisions. As discussed in greater depth below, they rely heavily on various forms of model transparency because neither model developers nor other stakeholders—including risk and compliance personnel, regulators, and investors—can manage risks that they cannot identify, understand, or measure.

Given the impact that poor underwriting can have on consumers, lenders, investors, markets, and communities,[35] the prudential and consumer protection requirements applicable to lending are particularly stringent even compared to other financial services use cases such as fraud, where use of machine learning is well established. Yet while these frameworks provide important concepts and processes for identifying and managing risks associated with the trustworthiness of machine learning models, they may also require adjustment given the ways in which machine learning models differ from and enhance certain risks relative to incumbent systems. Uncertainty about these questions is a significant factor affecting individual firms' decisions about whether and how to adopt machine learning in the underwriting context.

The efficacy of data science techniques for managing machine learning underwriting models is an important threshold question for lenders which are considering using such models and for their regulators. Insights from lending may also produce spillover effects in other sectors by spurring improvement in the underlying data science and informing reconsideration of how law and policy should evolve to promote responsible use.

This section briefly sets forth expectations that apply to consumer lending regardless of the type of underwriting model being used in the following areas:

- Prudential expectations regarding model governance throughout the model lifecycle;
- Fair lending requirements, particularly with regard to facially neutral practices that have an impermissible disparate impact on certain groups; and
- Disclosure requirements to provide applicants with individualized adverse action notices explaining why they were denied credit or offered less favorable terms.

The consumer protection requirements apply to both banks and nonbanks, although the degree of federal oversight by regulators is lower for nonbanks that are not subject to examination. Model risk management expectations apply only to banks, although they may inform practices of nonbank lenders. Where a supervised lender relies on a third party to design, develop, or operate tools or processes that are part of their lending operations, the lender is generally responsible for overseeing the vendor's compliance with applicable regulations.[36]

FinRegLab will take up a range of policy, legal, and regulatory questions related to the responsible, fair, and inclusive use of machine learning underwriting models in subsequent publications. This section provides an overview of key expectations and regulatory compliance issues. A more detailed discussion is provided in Appendix B.

[35] Emerging approaches to regulating the use of AI in other jurisdictions, such as the EU, have recognized the sensitivity of credit scoring and underwriting among AI applications and called for them to be treated as "high-risk" for risk management purposes. *See* European Commission, Proposal for a Regulation Laying Down Rules on Artificial Intelligence (Apr. 21, 2021); Penny Crosman, EU Proposes Restrictions on AI in Credit Scoring, Authentication, Am. Banker (Apr. 21, 2021).

[36] Board of Governors of the Federal Reserve System, Supervisory & Regulation Letter 13-19 (Dec. 5, 2013); Office of the Comptroller of the Currency, Bulletin 2013-29 (Oct. 30, 2013); Office of the Comptroller of the Currency, Bulletin 2020-10 (Mar. 5, 2020); Federal Deposit Insurance Corporation, Financial Institution Letter 44-2008 (June 6, 2008); Federal Deposit Insurance Corporation, Financial Institution Letter 19-2019 (Apr. 2, 2019); Consumer Financial Protection Bureau, Compliance Bulletin and Policy Guidance 2016-02, 81 Fed. Reg. 74410 (Oct. 26, 2016).

2.3.1 Model Risk Management

Federal prudential regulators have issued extensive guidance outlining their expectations for steps that banks should take in developing, monitoring, and using models of all types throughout all aspects of their operations. This guidance applies broadly to the range of model use cases that might create unexpected losses, compliance problems, or other negative outcomes for the firm and calls for enterprise-wide risk management processes including governance, policies, and controls.[37] Expectations are calibrated to the degree of risk posed by the particular use case, and credit underwriting is often considered to be among the highest risk activities depending upon the composition of the particular firm's business. Thus, for financial institutions subject to prudential oversight, these expectations typically require extensive pre-deployment review of credit models and monitoring during use, especially for firms that emphasize retail or consumer banking. For other financial institutions, bank regulatory expectations may broadly inform aspects of their model oversight practices, in part because funding and securitization counterparties may require some of these processes and practices.

The prudential model risk management expectations emphasize various aspects of model transparency, some of which can prove to be challenging in the context of machine learning underwriting models. At a broad level, the guidance requires documentation of the processes by which a model is developed, validated, and monitored during deployment. This includes documenting how the learning algorithm produced the final model. More specifically, the guidance creates an expectation that developers will evaluate whether models are relying on relationships in the data that are intuitive and defensible with regard to the outcome that they are attempting to predict, that firms will conduct appropriate sensitivity analyses to establish the soundness of the model for use, and that lenders will establish appropriate processes for identifying and mitigating risks relevant to the model's use, including compliance with applicable consumer protection laws.[38]

One of the ways in which the transition to machine learning poses a particular transparency-related issue concerns lenders' efforts to detect in timely ways conditions that may reduce the accuracy of machine learning models. Machine learning models are prone to brittleness—they may in effect reflect the training data too closely and not generalize to conditions that differ from that data. The emergence of tools to help lenders to improve their ability to recognize and respond to conditions in which the performance of machine learning underwriting models might rapidly deteriorate points to several additional potential inquiries: what approach has the model developer taken to enabling transparency, does that approach confer appropriate levels of transparency in practice, and how can the reliability and trustworthiness of information produced to explain the model be evaluated.

[37] Although each agency has its own issuance, the Federal Reserve Board's Supervisory & Regulation Letter 11-7 is often used as a shorthand to refer to all three agencies' guidance. *See* Board of Governors of the Federal Reserve System, Supervisory & Regulation Letter 11-7: Supervisory Guidance on Model Risk Management (Apr. 4, 2011) (hereinafter "FRB, SR 11-7"); Office of the Comptroller of the Currency, Bulletin 2011-12: Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management (Apr. 4, 2011); Federal Deposit Insurance Corporation, Financial Institution Letter 22-2017: Adoption of Supervisory Guidance on Model Risk Management (Jun. 7, 2017).

[38] *See, e.g.*, FRB, SR 11-7 (evaluating conceptual soundness involves assessing "documentation and empirical evidence supporting the methods used and variables selected for the model [to] ensure that judgment exercised in model design and construction is well informed, carefully considered, and consistent with published research and with sound industry practice."); *id.* attachment at 6 ("Developers should be able to demonstrate that such data and information are suitable for the model and that they are consistent with the theory behind the approach and with the chosen methodology."); *id.* attachment at 11 ("Key assumptions and the choice of variables should be assessed, with analysis of their impact on model outputs and particular focus on any potential limitations. The relevance of the data used to build the model should be evaluated ….").

Compliance with model risk management expectations raises the following questions for users of machine learning underwriting models:

- What considerations are relevant to identifying responsible, fair, and inclusive use of AI and machine learning systems?
- What kinds of transparency are relevant to establishing the conceptual soundness of AI and machine learning models: transparency as to the model's construction and general operations or transparency as to the bases for individual predictions made by the model?
- How should firms evaluate and measure the transparency of AI and machine learning models in the context of establishing their conceptual soundness and fitness-for-use?
- Where *ex post* explainability methods are used, how should firms evaluate the trustworthiness and utility of information produced by these supplemental analyses?

2.3.2 Fair Lending

Lenders are subject to broad anti-discrimination requirements regardless of the type of model a lender uses to predict an applicant's likelihood of default.[39]

These requirements give rise to two fair lending doctrines: disparate treatment and disparate impact.[40] Disparate treatment focuses on whether lenders have treated applicants differently based on protected characteristics. It generally prohibits consideration of race, gender, or other protected characteristics in underwriting and scoring models. Disparate impact addresses lenders' use of facially neutral practices that have a disproportionately negative effect on protected classes, unless those practices meet a legitimate business need that cannot reasonably be achieved through alternative means with a smaller adverse impact.[41] The legal analysis for disparate impact has three parts:[42]

- **Adverse Impact:** A plaintiff (such as a consumer or a regulatory agency) must make an initial showing that a particular act or practice causes a disproportionate adverse effect on a prohibited basis. This is typically analyzed by looking at whether use of particular variables or other lending practices cause approval rates or pricing patterns to differ by race, gender, or other protected characteristics.
- **Business Justification:** In response, the creditor must then show that the practice furthers a legitimate business need, such as whether the variable helps to predict the risk of default.
- **Less Discriminatory Alternative:** In response, to prevail on a claim, the plaintiff must demonstrate that the legitimate business need cited by the creditor can reasonably be achieved by using an alternative practice that would have less adverse impact.

[39] The Equal Credit Opportunity Act (ECOA) prohibits discrimination in "any aspect of a credit transaction" for both consumer and commercial credit on the basis of race, color, national origin, religion, sex, marital status, age, or certain other protected characteristics, and the Fair Housing Act (FHA) prohibits discrimination on many of the same bases in connection with residential mortgage lending. *See* 15 U.S.C. § 1691(a) (also prohibiting discrimination based on the receipt of public assistance and the good faith exercise of certain rights under federal consumer financial law); 42 U.S.C. § 3605 (prohibiting discrimination on the basis of race, color, national origin, religion, sex, familial status or disability).

[40] The Supreme Court has confirmed that both doctrines are available under the Fair Housing Act, but has not yet ruled on whether disparate impact analysis applies under ECOA. Texas Dep't of Housing & Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015). Federal regulations, agency guidance, and lower court decisions have recognized the doctrine under ECOA for decades, in part based on legislative history. *See, e.g.*, 12 C.F.R. § 1002.6(a); *id.* Supp. I, cmt. 1002.6(a)-2.

[41] For a general overview of the two doctrines and the ways that they overlap, *see* Carol A. Evans, Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks, Consumer Compliance Outlook 4-9 (Second Issue 2017).

[42] In litigation, the burden shifts back and forth between the parties to make particular showings at each stage. However in other contexts, such as where a lender's compliance team is applying this analysis to monitor its fair lending risk, one party will perform each of the steps.

Both doctrines rely on statistical tests and analyses of data inputs that can be more challenging to implement in the context of complex machine learning models. For example, the identification and management of variables that may proxy for protected class status under both disparate treatment and disparate impact theories of discrimination requires a high degree of transparency into how the models are built and how they make predictions. Machine learning models may also effectively reverse-engineer protected class status from correlations in data, even though consideration of such status is prohibited. Thus, particularly where machine learning models rely on data from more varied sources or on more complex and unintuitive features, lenders and regulators may need new tools and face new limitations in efforts to diagnose bias.[43] Certain uses of protected class features could actually increase the accuracy and fairness of machine learning underwriting models,[44] but there is substantial uncertainty as to whether various options are permitted under current law.

In the context of fair lending compliance, stakeholders are increasingly focused on these questions:

- Is managing input variables to identify and control fair lending risks effective in the context of algorithmic lending?
- How should firms choose among alternative model specifications that affect protected classes differently?
- Are firms permitted to use methodologies at any point in developing an underwriting model that involve direct consideration of protected characteristics in order to improve model fairness?

2.3.3 Adverse Action Notices

The Equal Credit Opportunity Act and the Fair Credit Reporting Act require lenders to disclose to consumers their principal reasons for denying credit applications or taking other adverse actions, including offering less favorable terms based on information in applicants' credit reports.[45] The requirements were adopted as part of broader efforts to prohibit discrimination and promote the correction of errors in credit reports, and give lenders substantial latitude as to how they determine which factors to highlight and whether to explain how the factors affected the lenders' decision. Thus, these adverse action notices must describe the facts that were "relevant to a decision, but [need not provide] a description of the decision-making rules themselves."[46]

Even though the regulations provide substantial flexibility to firms, lenders report that uncertainty about complying with adverse action requirements does shape and sometimes chill adoption of nontraditional data sources and machine learning methodology.[47] Explaining particular variables

[43] Historically, regulators have looked at whether particular variables have an "understandable relationship to an individual applicant's creditworthiness" as well as a statistical relationship to loan performance in determining whether they meet a legitimate business need. Office of the Comptroller of the Currency, Bulletin 1997-24, app. at 11 (May 20, 1997). *See also* Section 5 for further discussion.

[44] See Jon Kleinberg *et al.*, Algorithmic Fairness, 108 AEA Papers and Proceedings 22-27 (2018).

[45] The laws define "adverse action" to include denials of credit applications on substantially the same terms and in substantially the same amount as requested, unless the lender makes a counter-offer. Adverse actions also include unfavorable decisions on existing credit arrangements, such as negative changes in terms, denials of line increases, and reductions or cancellations of credit lines. 15 U.S.C. §§ 1681a(k)(1), 1691(d)(6). In 2011, a FCRA amendment took effect to require similar risk-based pricing notices where credit terms are "materially less favorable" than the terms granted to a "substantial proportion" of other consumers. 15 U.S.C. § 1681m(h); 12 C.F.R. §§ 222.70-75. ECOA's disclosure requirements apply to both consumer and commercial credit, although some details are different for business applicants. Federal agencies have excluded business credit from FCRA's disclosure requirements. 15 U.S.C. § 1681a(c); 12 C.F.R. §§ 222.70(a)(2), 1002.9(a).

[46] Andrew D. Selbst & Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085, 1100 (2018).

[47] Leslie Parrish, Alternative Data and Advanced Analytics: Table Stakes for Unsecured Personal Loans, Aite Group 16, figure 12 (2019) (reporting that surveyed industry executives view explaining model results, and specifically adverse actions, as the most significant challenge for using AI and machine learning in decisioning applications for credit); Eric Knight, Note, AI and Machine Learning-Based Credit Underwriting and Adverse Action Under the ECOA, 3 Bus. & Fin. L. Rev. 236-258 (2020).

that are influential in machine learning models can be difficult where the models develop and rely on relationships that are inherently complex, non-intuitive, difficult to assess, large in number, or dependent on other input variables or relationships. However, other stakeholders argue that the growth of open-source and other tools for more transparent and interpretable machine learning models have given lenders new options to satisfy adverse action reporting requirements.[48]

Users of machine learning underwriting models consider the following questions critical to complying with adverse action reporting requirements:

- Can models that rely on complex interactions between variables or models generate accurate adverse action notices?
- How can firms evaluate whether particular explanations of a model's prediction provide adverse action notices that meet current regulatory requirements?
- Can or should law and regulation require provision of information on adverse action notices that give recipients actionable information about how to improve the prospects of a better outcome on their next application for credit?

2.4 State of Adoption of Machine Learning Underwriting Models

Given the magnitude of potential benefits, risks, and compliance questions regarding the adoption of machine learning underwriting models, it is perhaps not surprising that use has lagged behind some other sectors and applications. Each firm faces a complex set of decisions regarding whether and how to reconfigure their lending platforms to deploy such models depending on their estimates of the potential magnitude of benefits, their ability to satisfy themselves and key stakeholders with regard to the trustworthiness of the models and their compliance with existing law, and the potential implementation costs.[49] For firms that find the investments in internal infrastructure and technical expertise prohibitive, reliance on vendor-provided underwriting models may be appealing, although this approach complicates the task of establishing appropriate oversight of the vendor's model and operations.

Though surveys of industry executives do not always distinguish between credit underwriting and other use cases,[50] they do suggest that interest and adoption are increasing generally across financial services and that events of the past year have further accelerated the trend. For example, a 2019 survey of risk management executives in financial services and insurance found that the respondents viewed AI and machine learning as a "major differentiator" in their businesses, though about half of the participating institutions lacked AI and machine learning capabilities in some or all of their platforms.[51] In a 2020 lender survey, 88% of respondents reported that they plan on

[48] BLDS, LLC *et al.* at 9, 15-18.

[49] *See* Emma Strubell *et al.*, Energy and Policy Considerations for Modern Deep Learning Research, 34 Proceedings of the AAAI Conference on Artificial Intelligence (2020) (calculating the cloud computing cost of a research and development cycle for a typical natural language processing pipeline as between $100,000 and $350,000 with almost $10,000 in additional electric costs); Lasse F. Wolff Anthony *et al.*, Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models, ICML Workshop on "Challenges in Deploying and Monitoring Machine Learning Systems," arXiv:2007.03051v1 (2020) (discussing the potential for ML models to generate vast amounts of carbon emissions, proposing a tool for tracking said emissions, and suggesting that developers can reduce a model's carbon footprint by training the model in low carbon intensity regions, during low carbon intensity hours, using efficient algorithms, and using efficient hardware and settings).

[50] A recent survey of consumer lending executives reported that 45% of respondents reported that their institution used AI for originations. Notably, the category labeled originations used in this survey included machine learning underwriting models as well as streamlining data fields and real-time decisioning based on probability inputs. Leslie Parrish, Impact Report, Consumer Lenders' Plans for Navigating the Next Normal, Aite Group (2021).

[51] Leslie Parrish, Risky Business: The State of Play for Risk Executives in the Analytics Ecosystem, Aite Group 14, figure 9 (2019).

BOX 2.4 UNDERWRITING-RELATED USES OF MACHINE LEARNING IN LENDING

Machine learning and AI models can be used in several aspects of credit decisions as discussed below and in other aspects of lenders' operations that can also have important implications for applicants' access to and use of credit, as discussed further in Box 4.1.

The phrase underwriting model may in practice refer to a suite or ensemble of models that operate in concert to determine an applicant's probability of default, assign a line of credit, and set pricing. Depending on the firm and business line, line assignment and pricing analyses may be made by different models than those that support approve/deny decisions. Pricing models may look at some but not all of the same information as approval models.[a]

Underwriting Model Development and Feature Engineering: An underwriting model using traditional modelling techniques may nevertheless apply rules or use interactions identified through analysis of large datasets using machine learning. This can give lenders the benefit of machine learning's insight and ability to analyze large volumes of diverse data without incurring the full costs of changing their lending platform or incurring certain financial, regulatory, or operational risks associated with using a machine learning model directly to make underwriting decisions. Although using machine learning to discover features—variables or relationships considered by the model—is common among technologically-enabled lenders across sectors and asset classes, individual firms' approaches may vary significantly in terms of the types of machine learning and data being used, as well as whether the models to develop these insights are treated as challenger models for model risk management purposes or used only in the earliest stages of model development to identify and validate particular features incorporated in traditional models.

Underwriting: Lenders use underwriting models to evaluate the likelihood that individual applicants for credit will repay the loan or not. In general, underwriting models are used to evaluate applicants' creditworthiness and assign them to risk tiers based on the relative probability of default that the model estimates. Lenders then decide which tiers they are willing to approve based on their willingness to take on credit risk in light of market conditions and other factors. For applicants who will be offered credit, many lenders assign interest rates based on the risk tier to which the applicant is assigned, the ability to pay as measured by capacity or income and other factors such as the likelihood to revolve balances for lines of credit. As discussed in Box 2.1, under risk-based pricing systems, the higher the probability of default associated with an approved applicant, the higher the cost of the loan, including the interest rate, will be. Finally, lenders may determine the loan amount or credit limit based on the underwriting model's assessment of an applicant's creditworthiness or based on a separate model tailored to that specific purpose.

Monitoring and Adjustment of Credit Lines: For open-end credit products like credit cards, lenders typically monitor lines on outstanding loans to assess whether they should be increased or decreased and to determine when and by what increments the amount of authorized credit should be adjusted. Here, some forms of unsupervised machine learning may help lenders assess borrower behavior, either on their own or in combination with supervised learning techniques. In the context of upward adjustments, concerns about model transparency may be somewhat reduced because such adjustments are often initiated voluntarily by lenders and are not subject to adverse action reporting.

[a] *See generally* Carroll & Rehmani.

increasing investment in AI in coming years specifically for use in credit risk.[52] General surveys in 2020 and 2021 indicate that recent events have accelerated adoption of AI across financial services and in other industries. For instance, in the face of rapidly changing economic conditions, there is a clear increase in interest in using machine learning for portfolio analysis and management as lenders consider how to adjust their credit criteria in changing and uncertain conditions. Yet the surveys also suggest that this growth is sparking increased unease: the number of financial services executives who report AI is being adopted "too fast for comfort" jumped to 37% in 2021, an increase of 20 percentage points from 2020.[53]

FinRegLab conducted interviews with a variety of bank and nonbank lenders, score and analytics providers, technology vendors providing model diagnostic tools and services, consumer advocates,

[52] Survey respondents included 175 LendIt subscribers out of a pool of more than 1,000. Brighterion, Survey Report: Using AI to Manage Credit Risk: Lenders Report on Current AI Use and Future Investments 9 (2020).

[53] KPMG, Thriving in an AI World 8, chart 3 (2021).

data scientists, and other financial services stakeholders to assess how machine learning models are being used and what choices firms relying on those models have made. This section and Section 3.5 summarize findings from those interviews.

2.4.1 Usage of Machine Learning Underwriting Models

Both banks and nonbanks are using machine learning models to analyze large datasets and identify relevant variables or relationships for use in logistic regression underwriting models. While such feature engineering is common across a broad range of credit products, the use of machine learning models to make underwriting decisions is at an early stage.[54] Among those using machine learning underwriting models, firms are using a variety of methodologies to develop, explain, and manage these models. How firms design and use these models is likely to evolve as firms and other stakeholders gain experience in operating and managing these models. Decisions of individual firms about whether to use machine learning underwriting models, what forms of machine learning to use, and how to enable appropriate transparency vary based on firm culture and strategy as well as competitive dynamics in specific sectors and asset classes.

2.4.1.1 Usage by Sector

Within the banking sector, the resources that the largest firms command—especially with respect to personnel, customer data, and computing infrastructure—make machine learning a more realistic choice than for even large regional or global banks with smaller domestic retail banking businesses.[55] Similarly, the cost-benefit analysis related to transitioning to machine learning underwriting platforms may only make sense right now for lenders who operate at a sufficiently large scale that marginal performance gains translate to substantial enough returns to warrant the necessary investments. Bank concerns about technology companies challenging their traditional franchise may also be driving the largest firms to develop and implement machine learning models across their operations.[56] Most of these lenders are developing machine learning underwriting models with minimal external support.

Mid-size and smaller banks are generally still in the process of evaluating whether machine learning underwriting models can sufficiently improve the performance of specific aspects of their businesses to warrant implementation costs and process changes.[57] Over time, the emergence of a varied community of vendors that can help develop, implement, and manage machine learning models may make adoption of machine learning more realistic for smaller firms, though reliance on third-party vendors can also increase challenges related to model governance.[58]

Nonbank lenders' use of machine learning underwriting is generally more widely established across product markets than in the banking sector, especially in segments populated by newer

[54] For purposes of this report, the term machine learning underwriting model refers to a model in use to estimate the risk of default related to applications for credit and excludes activities like using machine learning for feature engineering or as challenger models. *See* Box 4.1.2.

[55] For example, across a 2020 survey of 175 large and small institutions that subscribe to Lendit, approximately 20% of respondents had no in-house staff for credit modelling and relied on third parties to conduct such activities. Even large institutions with credit modelling teams did not always devote significant resources to the activity, as just 16% of large institution respondents had four or more full time modelers. Cornerstone Advisors at 4, figure 4.

[56] In his 2020 annual letter to shareholders, JPMorgan Chase & Co. CEO Jamie Dimon writes that banks such as JPMorgan "are facing extensive competition from Silicon Valley, both in the form of fintechs and Big Tech companies (Amazon, Apple, Facebook, Google and now Walmart), that is here to stay. As the importance of cloud, AI and digital platforms grows, this competition will become even more formidable." JPMorgan Chase & Co., Chairman & CEO Letter to Shareholders, sect. III (Apr. 7, 2021).

[57] Ostmann & Dorobantu.

[58] As discussed further in Section 3, a number of vendors are providing assistance with model development and monitoring.

firms. Certain nonbank lenders emphasize the importance of machine learning models' superior capacity to analyze large volumes of data and different kinds of data. Notable factors encouraging nonbank adoption of machine underwriting models and use of more complex machine learning models include: reliance on digital business models; use of newer lending platforms; absence of bank-like model risk management requirements; less consistent examination and oversight than in the banking sector; and funding incentives from private equity investors attracted to technology transformation. But nonbank usage is not wholly unconstrained. For example, equity and capital markets investors may impose certain kinds of limitations in practice to protect potential returns.

A range of other nonbank companies are offering products, services, and data to support lenders' potential adoption of machine learning models (as discussed further in Section 3). For instance, credit bureaus and companies that develop third-party credit scores are also using machine learning in a variety of ways, ranging from supporting development of their own models to developing custom scoring models for clients and providing tools to support their use.[59] Consulting firms are also offering services to lenders at various stages of planning, designing, implementing, and using machine learning underwriting models, including conducting algorithmic audits prior to deployment and periodically during use.

2.4.1.2 Usage by Market

Credit cards and unsecured personal loans (including point-of-sale loans) are the consumer finance asset classes in which the use of machine learning models to make credit decisions is most advanced. This reflects the historical position of credit cards as being at the analytical forefront of consumer finance and the dominance of digital lending in unsecured personal loans. Between 2015 and 2019, fintech lenders doubled their share in the latter market even as its overall size expanded, and now account for 49% of originated loans.[60]

Auto lending[61] and small business lending[62] are also areas where machine learning underwriting models are in use. In small business lending in particular, less standardized data and development of significant market share by nonbank lenders may promote the use of machine learning underwriting models.[63]

2.4.1.3 Usage by Model Type

The types of machine learning models that are most relevant to credit underwriting predictions and the options for managing those models to provide transparency for business and regulatory purposes are discussed later in Section 4 and Section 3, respectively. Among early adopters of machine learning underwriting models, firms are using a variety of model types including tree-

[59] Third-party credit scores such as those developed by FICO and VantageScore for use by a broad range of lenders are estimated to be used in more than 90% of mortgages, credit cards, and auto loans, though individual lenders may use them in different ways. For instance, some lenders set a minimum score below which they will not lend, while others use the scores or the underlying attributes as inputs to proprietary underwriting models. The scores are also frequently used as benchmarks for portfolio monitoring and securitization. Some scoring model developers and consumer reporting agencies also offer services to help lenders develop custom proprietary models. FinRegLab, Cash-Flow Market Context & Policy Analysis at 9 & n.13.

[60] Experian, Fintech vs. Traditional FIs: Trends in Unsecured Personal Installment Loans 3 (2019); *see also* DBRS, U.S. Unsecured Personal Loans—Marketplace Lenders Continue to Expand Market Share 3-4 (2019) (analysis of the growth of fintechs in the unsecured personal lending space from 2013 to 2018, measuring market share as the proportion of outstanding loan balances).

[61] Auto lenders such as Prestige Financial Services and Upstart have adopted AI underwriting models. *See, e.g.,* Becky Yerak, AI Helps Auto-Loan Company Handle Industry's Trickiest Turn, Wall St. J. (Jan. 3, 2019); Upstart, Auto Loans (undated).

[62] Trevor Dryer, How Machine Learning Is Quietly Transforming Small Business Lending, Forbes (Nov. 1, 2018).

[63] Wei Wang *et al.*, Using Small Business Banking Data for Explainable Credit Risk Scoring, 34 Proceedings of the AAAI Conference on Artificial Intelligence (2020).

based models, neural networks, and ensembles that combine several different models.[64] Further, some firms are using model architectures that are more transparent by design. Others are using more complex or black box models that rely on *post hoc* explainability techniques—secondary models, analyses, or methods—to meet necessary levels of transparency. Banks may be particularly hesitant to engage in more complex modelling as increasing complexity can make it more challenging to explain to customers, examiners, and other stakeholders how and why certain decisions are made in a clear, concise way.

Some lenders believe that tree-based models offer a workable balance of improved performance, operational efficiency, and transparency. In particular, stakeholders report that XGBoost underwriting models using monotonicity constraints and coarse classing[65] can produce performance gains of up to 15-20% over traditional regression when evaluating applicants based on traditional underwriting data, while meeting transparency needs.[66]

However, other lenders and analytics providers suggest that constrained neural networks are a preferred approach, because they provide improved predictive power and do not require relying solely on supplemental models or analyses to satisfy transparency needs.[67] For example, as discussed in greater detail in Section 4, stakeholders report that single-layer neural networks or those generated using a piecewise linear activation function and consisting of a series of locally-linear models can meet relevant transparency requirements.[68]

Further, some lenders have found ways to develop and use machine learning models that involve many more features and greater complexity than traditional underwriting models have typically included. For instance, one auto lender reports using a model with more than 2,000 features to assess the creditworthiness of loan applicants—a significant expansion compared to current automated underwriting systems which tend to use dozens rather than hundreds or thousands of features.[69] Similarly, some nonbank firms originating consumer loans use AI underwriting models with more than 1,000 features.[70]

***

The remainder of this report examines in depth the choices that lenders using machine learning underwriting models make when considering how to develop models that use strong, defensible, and fair relationships to make credit decisions. This report's analysis of model development choices begins in Section 3 with a consideration of why model transparency is so important and the options for developing sufficiently transparent underwriting models. Section 4 describes forms of machine learning relevant to credit underwriting as well as foundational data selection and preparation stages

[64] *See* Office of the Comptroller of the Currency, Comptroller's Handbook, Credit Card Lending 17 (Version 2.0, Apr. 2021) ("Some banks are applying artificial intelligence (AI) and machine learning (ML) methods [*e.g.*, gradient boosting or neural network] to credit scoring").

[65] For a description of XGBoost, see Figure 4.1.2.1.1 of this report. For an explanation of monotonicity constraints, see Section 3.4.1.2 of this report. For an explanation of coarse classing, see Section 4.2.2.2 of this report.

[66] *See also* Petropoulos *et al.* (finding that a machine learning model using XGBoost outperformed a neural network—and that both models outperform traditional logistic regression models—in assessing the credit risk of corporate loans in Greece using a combination of loan-level data and macroeconomic indicators).

[67] Agus Sudjianto *et al.*, Unwrapping the Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification, arXiv:2011.04041v1 (2020); Scott Zoldi, How to Make "Black Box" Neural Networks Explainable, FICO Blog (Jan. 14, 2019).

[68] Sudjianto *et al.*

[69] Prior to deploying a machine learning platform, this lender had used only 23 features in its underwriting model. *See* Rhagav Bharadwaj, Top 5 AI Startups in Banking by Funding – A Brief Overview, Emerj (Nov. 19, 2019); Yerak.

[70] Upstart Holdings, Inc., Form 10-K 11 (2021) ("Variables in our AI models have increased from 23 in 2014 to more than 1,000 as of December 31, 2020. These include factors related to credit experience, employment, educational history, bank account transactions, cost of living and loan application interactions."); Oportun, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning at 3 ("Oportun's lending platform leverages machine learning and processes large amounts of alternative data along with traditional credit bureau data to assess creditworthiness across more than 1,000 end nodes.").

of developing an underwriting model. Section 5 examines approaches to identifying and mitigating sources of bias in model development. That section sets out the ways that data and models can each be the source of bias, describes emerging approaches to measuring bias, and outlines methods for addressing models in machine learning models. Section 6 looks ahead to FinRegLab's empirical research on the capabilities and performance of model diagnostic tools available to lenders.

3. MODEL TRANSPARENCY

The ability to realize the potential of machine learning to significantly enhance the quality, fairness, and inclusiveness of credit decisions depends on resolving uncertainty about when to trust a specific model and when not to. Among the components of model trustworthiness, model transparency serves as a critical threshold question.[71] Machine learning models do not inherently need to be transparent to make predictions, and existing law and regulation do not generally require users of AI or machine learning model users to meet defined thresholds for model transparency. But without appropriate transparency, internal and external model stakeholders—model developers, risk managers, regulators, and investors—cannot be confident that a model can be or is being used responsibly and fairly. Transparency is also important to applicants who receive adverse action notices.

The importance of transparency in answering concerns about trustworthiness is heightened in highly regulated sectors like financial services and sensitive use cases like credit underwriting because several existing oversight frameworks rely on or require transparency in one form or another. In the context of credit decisions, this creates a need to understand how a model used in credit underwriting was developed, how it makes decisions, and what factors explain particular outcomes, such as denying certain credit applicants or charging those applicants more for their loan. Specifically, various stakeholders need to be able to assess the reliability and robustness of the model's performance, to provide an explanation for why a particular credit decision was made and a particular price was offered, and to understand if the model's predictions are fair from the perspective of individual applicants and the treatment of groups given special protection due to historical discrimination.

Concerns about model transparency shape lenders' decisions at every stage of the process of developing, implementing, and managing machine learning underwriting models. Model developers may in effect work backwards from the transparency requirements of their use case—by designing and planning their modelling approach based on the level and type of transparency required. In practice, the developer of an underwriting model needs to be able to establish that each relationship in the model has an intuitive, defensible relationship to an applicant's likelihood of default. Further, given adverse action disclosure requirements, firms need the capacity to pinpoint the primary bases of individual credit decisions. Model developers can use a variety of tools and techniques to build a model with the necessary transparency in whatever type of machine learning they choose for their underwriting model.[72] They might develop an inherently interpretable model—one that can be

[71] Ostmann & Dorobantu at 45-63.

[72] Forms of machine learning relevant to underwriting are discussed in Section 4.1.

explained and understood on its face without additional analysis. Alternatively, they might build an explainable model—one that uses more complex or black box models alongside supplemental models, analyses, and techniques designed to improve the transparency of such models.[73]

This section considers both the importance and challenge of enabling appropriate transparency for machine learning underwriting models; presents the debate about whether to achieve such transparency by constraining model architecture and/or using supplemental explainability techniques; analyzes the tools and techniques available to build both kinds of models; and surveys model diagnostic tools that have emerged to support lenders using machine learning underwriting models. As discussed further in Section 1, FinRegLab is partnering with researchers from the Stanford Graduate School of Business to conduct empirical research relating to many of these issues.

3.1 The Importance of Model Transparency

While financial services laws and regulations do not define specific thresholds for model transparency, these frameworks and in some cases firm risk management policies focus attention on two different types of transparency: transparency about how a machine learning model works and transparency about the process by which it is designed, implemented, and managed.[74] Given this, resolving questions about model transparency are a key hurdle for widespread adoption of machine learning in high stakes uses like credit underwriting.[75]

In this context, model transparency serves several critical purposes in establishing the trustworthiness of a machine learning underwriting model:

- **Promoting Sound Model Development:** Sufficient transparency lets model developers generate information about and evaluate specific tradeoffs involved with various design or implementation decisions they must make in the course of designing and developing a model. This scrutiny of the logic used to make a prediction enables activities like model debugging and bias mitigation that can directly improve the performance and fairness of the model.
- **Facilitating Pre-Deployment Model Reviews:** Given the risk of financial losses and reputational damage connected with poor underwriting, lenders typically engage in significant pre-deployment reviews of underwriting models; indeed, bank lenders must do so as a matter of regulatory compliance. Being able to explain and document the conceptual soundness of a proposed model and the process by which it was designed are critical to determining whether a model can be used responsibly and fairly. So is the ability to explain its operations and performance before deployment.
- **Enabling Model Monitoring:** Once an underwriting model is in use, model transparency allows stakeholders to track its performance across key performance and risk indicators, to assess whether changes in its performance or operations are exposing the firm to new or different financial, regulatory, or reputational risks, and to make adjustments

[73] The terms interpretable AI and explainable AI, much like the underlying terms interpretability and explainability, have no fixed meaning, and are used differently among various stakeholder communities. The terms set forth above reflect usage throughout this report.

[74] Ostmann & Dorobantu at Ch. 5. Where this section discusses the transparency or complexity of machine learning underwriting models, it primarily focuses on concerns related to the underwriting model produced by various machine learning methods rather than the learning algorithm that produced the final model. More research is needed to understand how varying approaches to managing both forms of transparency affect the explainability and fairness of machine learning underwriting models.

[75] Leslie Parrish, Alternative Data and Advanced Analytics at 16, figure 12 (reporting that surveyed industry executives view explaining model results, and specifically adverse actions, as the most significant challenge for using AI and machine learning in decisioning applications for credit).

as appropriate. It also permits internal review of the quality and consistency of decisions based on model predictions and enables governance processes to determine when particular decisions need to be challenged or when a model needs to be reviewed or updated. Periodic fair lending testing is a common example.

- **Providing Recourse and Empowering Consumers:** Model transparency facilitates review, challenge, and error correction for those adversely affected by a model's predictions. In the context of underwriting, this rationale is most often considered in the context of adverse action notices. Here, model transparency enables an applicant who was not offered credit to be provided information on an adverse action notice that enables the applicant to review the basis for the decision. Where the required disclosure reports a prior bankruptcy when no such action had occurred, for example, the applicant can seek reconsideration of the decision and pursue corrections in their credit history. Model transparency also has the potential to provide information that can help empower consumers—for example, in improving consumers' understanding of their credit scores or ways to improve their financial position and creditworthiness in the future.
- **Establishing Regulatory Compliance:** Model transparency is also critical for firms to document and conduct internal assessments of their compliance with the range of regulatory requirements applicable to providing credit. Bank and nonbank lenders will generally need sufficient insight into the operations and performance of their underwriting models to assess and document their compliance with fair lending and adverse action reporting requirements, and banks will need to do the same as to prudential model risk management expectations.

3.2 The Challenge of Model Transparency

While machine learning models can be significantly more complex than traditional underwriting algorithms, it is too facile to assume that underwriting models using logistic regression are more explainable than any machine learning model. An underwriting model assessing dozens of variables using logistic regression may be prone to many of the same explainability challenges of a neural network that is more transparent by design. In this regard, the sustained attention on how to enable and evaluate the explainability of machine learning models may raise standards applied to all kinds of models.

Nevertheless, enabling necessary levels of transparency in machine learning underwriting models poses particular challenges with both human and technological components. This section considers broadly applicable challenges to explaining credit underwriting models before describing factors affecting the complexity of machine learning models.

3.2.1 Users of Model Explanations

A variety of stakeholders have a general need to understand how a credit underwriting model works and, in some cases, a particular need to understand individual predictions made by a model:

- A firm's model developers and managers, who are responsible for designing, implementing, and operating models that meet business objectives and expectations set forth in law, regulation, and firm policies
- A firm's business executives, who need to establish the model's fitness-for-use in order to commit capital based on the model's predictions

- A firm's legal and risk management teams, which review a model's compliance with laws, regulations, and firm policies relevant to a model's specific use case
- A firm's regulators, who review decisions about model development and use from the perspective of compliance with individual consumer financial protection requirements and monitoring prudential risks where applicable
- A firm's customers and potential customers, who need to make decisions about whether to provide access to their data, are legally entitled to understand the basis for the firm's decisions on applications, and are best positioned to detect the use of errors reported on adverse action disclosures
- A firm's investors, who supply capital based on confidence in management's business judgment and performance of individual loans or asset-backed securities

The needs of each of these stakeholders to understand how an underwriting model works do not fundamentally change when machine learning is used, but firms are still working to develop and test forms of machine learning that consistently meet these needs. The diversity of relevant stakeholders is important to developing context-appropriate and usable explanations of a model's behavior, as explainability is a distinct psychological process that depends on the user of the explanation.[76] Users of explanations of model behavior rely on this information to serve different purposes and answer different questions. Further, each one will also bring different expertise and experience—about predictive models, credit risk assessment, and credit decisions—to the task of interpreting the meaning and implications of a model's explanation.

In this context, generating an accurate, detailed, and often technical explanation of how a model's outcome came to be may not always be sufficient to facilitate understanding and communicate meaning.[77] Articulating a model's operations—how it makes predictions or what the basis of a particular prediction is—can be challenging in its own right when a model lacks transparency. Indeed, the technological task of explaining machine learning models has spawned a broad, sustained inquiry in academia and industry, and that inquiry has produced a range of options for generating information about model behavior. But the information produced to explain models may not be sufficient if it doesn't also enable actions contemplated by public policy, such as the identification and mitigation of fair lending risks in the context of consumer lending. To achieve this, the information provided to explain models must be usable to support and inform strategic, governance, and risk management decisions involving the stakeholders groups noted above—that is, diverse users of model explanations need to be able to understand and act on that information. Given the early state of adoption and the absence of industry or regulatory standards, approaches to meeting these needs vary across firms and even within them.

3.2.2 Understanding Complex Models

The importance of understanding machine learning models may be no greater than incumbent models, but it may require different tools and occur at different parts of the process. When designing a traditional underwriting model, a model developer assesses relevant data, conducts analyses, and opts to include particular variables or relationships into a model precisely because he or she has

[76] David A. Broniatowski, Psychological Foundations of Explainability and Interpretability in Artificial Intelligence, National Institute of Standards and Technology 2 (2021).

[77] *Id.* at 5.

demonstrated the predictiveness of those attributes.[78] Even where this process is aided by automation, the developer will make clear, *ex ante* decisions about what relationships and analyses are incorporated in the model.

However, in underwriting models developed by machine learning algorithms, the task of understanding what relationships the model is identifying and why requires additional work to identify and assess individual features, relationships, and analyses that were identified by the learning algorithm rather than a human. The algorithm deconstructs and extracts information from training data to generate a model. A model developer does not program certain relationships to be in the model, but rather oversees, monitors, and shapes how the algorithm analyzes and uses training data to construct a model. This process is not without opportunities for oversight—each decision that a model developer makes is reviewable by modelling peers, risk and compliance personnel, and regulators. But those opportunities may come at different points in the model lifecycle and require more work and different tools to answer questions about the model's reliability and fairness for each of these stakeholder groups.

3.2.3 Factors Affecting Model Complexity

In general, the more complex a model is, the more challenging it will be to explain and understand.[79] Several characteristics of machine learning models further increase the challenge of enabling necessary model transparency and have made this area a critical focal point for academic and industry data science research. Those include:

3.2.3.1 Nature of Data

Because machine learning models are shaped by training data to a greater degree than traditional models, the size, nature, and quality of data can have a direct effect on model complexity. The greater the number of input variables used (or dimensionality of the data), the more complex the model is likely to be due to increased complexity of interactions between features in the model.[80] Further, increases in the size, composition, or number of datasets used as inputs to underwriting decisions can make data quality issues difficult to identify and increase their importance, both of which can affect complexity.[81]

Decisions about what information lenders include in credit risk assessment may also affect the complexity of the resulting underwriting model. For example, where lenders choose to use only traditional credit information, that decision itself may impose some limit on model complexity. Starting with a relatively small number of input variables and limiting usable features to those with strong, defensible relationships to creditworthiness may naturally limit the complexity of the resulting underwriting models, perhaps to the point that they produce marginal increases in complexity and performance. For lenders with sufficient scale, the marginal gains may make this an inherently attractive option for early uses of machine learning underwriting models. However, for smaller lenders these gains may not justify taking on the additional operational and compliance risks that more complex models entail.

[78] In this context, an attribute refers to a variable included in a model's dataset. This could include input variables, such as an individual's income, as well as a target or output variable (such as whether an individual is likely to default on a loan).

[79] In this report, the term model complexity is used synonymously with the term model opacity. Both speak to the challenge of explaining and understanding various determinants of a model's behavior from various perspectives.

[80] *See* Selbst & Barocas at 1096 ("The more variables that the model includes, the more difficult it will be to keep all the interactions between variables in mind and thus predict how the model would behave given any particular input").

[81] Ostmann & Dorobantu at 19.

Using additional, unconventional data can introduce other difficulties when training complex models.[82] If a machine learning model overfits to training data that has a large number of missing observations or is very noisy, for instance, it may identify signals that are not actually relevant to its prediction task. Overfitting can be managed using model constraints such as regularization (see Section 3.4.1).

3.2.3.2 Model Size

The ability of machine learning algorithms to identify and assess a greater number of features enhances model complexity and the technological challenges associated with explaining machine learning underwriting models. The difference in understanding and managing a model that considers dozens of variables or features to one that uses thousands or millions is the simplest statement of this problem. Size in turn affects the number of feature interactions, and potentially their complexity. For ensemble models, which reflect use of several individual models in a sequence, the number of sub-models, the overall size of the model, and the number of features considered likely increase complexity.

The size of a model is often reflected by the number of parameters it uses.[83] For example, a logistic regression may use up to 100 parameters—roughly one for each input feature. On the other hand, an artificial neural network with hundreds of hidden nodes may have tens of thousands of parameters—reflecting the weights of each connection between nodes in the network. Larger models with more parameters can represent more complex relationships between variables, meaning they can capture more relevant patterns in the data, and also that they can be far less transparent than smaller models.

In addition to parameters, developers' decisions about how to set hyperparameters such as the depth of a tree model or number of layers in a neural network can also make a model bigger and more complex.[84] For instance, a higher number of layers may generate more accurate predictions but can also increase model complexity. Similarly, increasing the depth of a tree may increase predictive performance, but may make the model less interpretable.

3.2.3.3 Nature of Individual Features

The nature of the features used in machine learning models can also be challenging, regardless of the number of features being used. Individual features may increase model complexity where they rely on math that is challenging to unpack on its own or in the context of models with more intricate architectures (such as neural networks with high numbers of layers). Debt-to-income ratio is an example of a simple feature, in part because it is derived from two input variables and involves a simple, obvious transformation. Machine learning models have the ability to generate

[82] *Id.*

[83] Model parameters are variables in the model that are configured using the training data and are fitted to the model. When the training is initialized, the parameters are usually set to a random value (or zero). As training progresses, these random values are updated using an optimization algorithm, which performs a search through possible parameter values to learn and update the values. The final parameters that are determined at the end constitute the trained model. Examples of parameters are coefficients (or weights) of linear and logistic regression models and weights and the biases for neural networks.

[84] Instead of being learned from the training data, hyperparameters are set manually by model developers before training and help generate a more efficient and accurate optimization process to estimate and optimize the model parameters. A developer uses search algorithms, like grid and random search, to help tune model hyperparameters and improve model accuracy. Examples of hyperparameters include depth of a tree and number of layers in a neural network.

and consider features that capture both non-monotonic[85] and non-linear[86] relationships, which enable these models to gather more detailed and granular information about the data used in the models but also increase model complexity.

3.2.3.4 Relationships Between Features

Feature interactions pose a critical challenge to enabling appropriate model transparency and oversight. This is especially true where feature interactions involve latent features, which are variables that inform a model's prediction, but are not part of the training or input data or the prediction itself. Latent features are generated by a machine learning algorithm from variables in the dataset and serve as internal or interim analyses that help determine the model's prediction. In general, the greater the number of the latent features and the more difficult those relationships are to describe on their own, the more complex the model will be.

Much of the improvement that machine learning offers may derive from feature interactions, especially those related to latent features, but many emerging explainability techniques assume those interactions do not affect the model's predictions. Complexity resulting from feature interactions can increase along with the number of input variables and types of datasets being used and the number of observations in these datasets.

3.3 The Debate About Interpretable and Explainable Machine Learning

A model's use will determine the level of necessary transparency and shape fundamental choices that developers make about how to build and manage a particular model. Some models have a higher degree of transparency by virtue of their structure and design. These models are said to be inherently interpretable or self-explanatory and can generally meet transparency needs on their own.[87] Others lack architecture that is transparent by design and are therefore less interpretable. These models require the use of additional models or analyses designed to explain the model—that is, *post hoc* explainability methods designed to improve stakeholders' ability to access and understand information about the model's behavior and the bases of its predictions. These interventions add an "observable component" to complex models to enhance stakeholders' ability to understand the models' behavior and to accept or challenge their decisions.[88]

This choice between inherently interpretable models and models that require *post hoc* explainability methods has shaped early adoption of machine learning underwriting models. Firms and researchers alike are working to understand better whether lenders should use inherently interpretable models or pair less interpretable models with supplemental explainability methods to satisfy transparency needs. Proponents of using only inherently interpretable models argue that well-designed models of this kind perform as well as more complex models and deliver the necessary transparency. Importantly, they do so without relying on secondary techniques and analyses that introduce further uncertainty

[85] Adding salt to a savory dish presents an intuitive example of a non-monotonic relationship. A small amount of salt will generally make the dish taste better. However, after a certain point, adding salt will not improve the taste of the dish and, in fact, will make the dish taste worse. This is an example of a non-monotonic relationship, as the relationship is positive in some cases and negative in others, which means the relationship is not one-directional.

[86] A non-linear relationship is one in which increases or decreases in an input variable do not always produce proportionally consistent changes in the target or output variable, where each input variable impacts the model independently (no feature interactions). For example, parents of multiple children will know that going from one child to two in a household has a larger effect on the overall amount of parental attention required than the change from zero children to one.

[87] Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Boxes Explainable (2019).

[88] Jonathan Johnson, Interpretability vs. Explainability: The Black Box of Machine Learning, BMC (2020); Leilani Gilpin *et al.*, Explaining Explanations: An Overview of Interpretability of Machine Learning, arXiv:1806.00069v3 (2019).

and a second layer of trustworthiness questions and can bypass feature engineering.[89] Proponents of inherently interpretable models also commonly question whether adding a second layer of analytical complexity compounds, rather than resolves, the challenge of establishing the trustworthiness of AI and machine learning systems and can meet specific transparency requirements.[90] Finally, stakeholders suggest that interpretable models may be preferable—all else being equal—because they are less costly than running more complex models.[91]

Proponents of complex models that rely on *post hoc* explainability techniques argue that this approach has the potential to deliver superior predictive accuracy—for lenders and applicants alike—while satisfying model transparency needs.[92] Industry proponents argue that even inherently interpretable models run the risk of being too complicated for a human to interpret completely. They point to examples of 100-layered trees, which are complex enough that *post hoc* explainability methods may still be necessary to meet transparency needs.[93]

Interpretability and explainability both address important aspects of model transparency. Although many stakeholders use the terms interchangeably, this report distinguishes between the terms to discuss critical choices model developers make about how to enable necessary transparency when designing and operating specific models. The characteristics of interpretable and explainable machine learning models will be considered in turn.

3.3.1 Interpretable Machine Learning Models

In its broadest sense, model interpretability refers to the ability to understand a model's operations based largely on its formal notation. To be interpretable, a person should be able to infer the following: (1) the types of information or input variables that a model uses, (2) the relationship between the input variables and the model's predictions or outputs; and (3) the data conditions for which the model will return a specific result (for example, to receive a credit score of 600, weekly income has to be at least $600).[94] Interpretable models are ones where stakeholders can relatively easily identify correlations or relationships used by the model to predict an outcome because of the model's design or structure.[95] Interpretable models include models with comparatively simple

[89] Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Machine Intelligence 206-215 (May 13, 2019) (reporting "no performance difference between interpretable models and explainable models" for credit scoring); Scott Zoldi, Not All Explainable AI is Created Equal, Retail Banker International (Oct. 9, 2019); David J. Hand, Classifier Technology and the Illusion of Progress, 21 Statistical Science 1-15 (2006) ("the extra performance to be achieved by more sophisticated classification rules, beyond that attained by simple methods, is small"). *See also* Agus Sudjianto *et al.*

[90] *See* Molnar; Alejandro Barredo Arrieta *et al.*, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, arXiv:1910.10045v2 (2019); Rudin; Cynthia Rudin & Joanna Radin, Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition, Harvard Data Science Rev. (Issue 1.2, Fall 2019).

[91] D. Sculley *et al.*, Hidden Technical Debt in Machine Learning Systems, 2 NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems at 2503-2511 (2015).

[92] Complex machine learning models used in various fields have outperformed inherently interpretable models. *See, e.g.,* Weiwei Jiang & Jiayun Luo, An Evaluation of Machine Learning and Deep Learning Models for Drought Prediction Using Weather Data, preprint submitted to J. of LATEX Templates, arXiv:2107.02517v1 (2021); Rishi Desai *et al.,* Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims with Electronic Medical Records to Predict Heart Failure Outcomes, 3 JAMA Network Open (2020); Andrés Alonso & José Manuel Carbó, Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation, Banco de España Working Paper 2105 (2021); Jay Budzik, Why ZAML Makes Your ML Platform Better, Zest AI (Mar. 6, 2019).

[93] Zachary C. Lipton, The Mythos of Model Interpretability, arXiv:1606.03490v3 (2017); Yan-yan Song & Ying Lu, Decision Tree Methods: Applications for Classification and Prediction, 27 Shanghai Archives of Psychiatry 130-135 (2015); Patrick Hall *et al.* Proposed Guidelines for the Responsible Use of Explainable Machine Learning, arXiv:1906.03533v3 (2019).

[94] Ostmann & Dorobantu at 49-51. *See also* Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning, arXiv:1702.08608v2 (2017); The Royal Society, Explainable AI: The Basics (2019).

[95] Johnson.

structures,[96] such as small decision tree models that can be "inspected"[97] or may have a limited number of features or parameters, which make them more likely to be intuitive and easier to parse without the use of additional models or tools. As discussed further below, developers can also apply constraints in building the model to increase interpretability. Examples include regularization to encourage model sparsity, monotonicity constraints, and constrained deep neural networks to produce an interpretable latent space (see Section 3.4.1 and Section 4.1.2.1.3).

Generally, the less complex a model is, the more interpretable it is. But there is no fixed way to measure model transparency or interpretability, and various characteristics contribute to a model's interpretability, including the number of model parameters, the number of features used, the level and extent of feature engineering, and (for ensemble methods) the number and complexity of sub-models. Accordingly, individual implementation choices will affect whether a particular logistic regression model with dozens of variables is in practice more interpretable than a neural network with limited layers. The following framework suggests a simplified schematic for assessing how interpretable a model is and where *post hoc* explainability techniques may be necessary to supplement the information that is available from the structure of the model itself.[98]

» **Inherently Interpretable Models:** Models that are typically easier to understand and explain are often described as inherently interpretable or self-explanatory. A user can easily infer how a particular model input—a data point about an applicant for example—results in a model output. This category includes linear and logistic regression, small decision trees, and credit scorecards, among others. These models can often be written down using a simple diagram or equation. Often, these models are monotonic and linear—meaning that changes in an input variable produce changes in the target variable in one direction and of a consistent magnitude.

» **Moderately Interpretable Models:** Some models are too complex to be completely understood by a user, yet they have some properties that make it easier to understand their behavior. For example, large neural networks or tree ensembles are too complex to be understood completely, however by adding certain constraints during training, these models can be required to be monotonic to improve their transparency. This means that changes in a single input variable will always push the output in the same direction.[99] Moderately interpretable models can also be used with *post hoc* explainability techniques.[100]

» **Uninterpretable Models:** Some models are so large or complex that it is impossible for a user to infer why a particular input leads to a model output. These models include large neural networks and ensemble methods that combine several steps or sub-models (such as large trees or tree ensembles including XGBoost). These models are usually non-linear and non-monotonic with respect to their inputs, meaning it is difficult to predict how changing a particular input variable will impact model output. Models in this category typically require *post hoc* explainability techniques to meet levels of model transparency necessary for the model's use case.

[96] *See* Patrick Hall & Navdeep Gill, An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI, O'Reilly 13-15, 20-25 (2nd ed. Aug. 2019); Section 4.1.2.1.1 of this report.

[97] Arun Rai, Explainable AI: From Black Box to Glass Box, 48 J. of the Academy of Marketing Science 137-141 (2019).

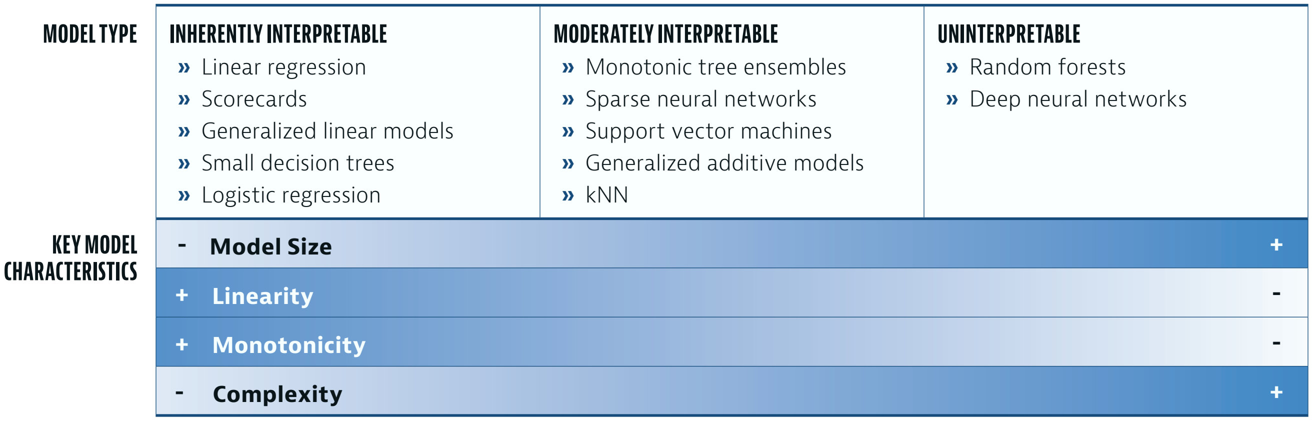
[98] Hall & Gill at 11-13.

[99] Another example is feature engineering: complex preprocessing methods can produce latent features—for example, trended analysis of a series of other features derived by the model. Using latent features can increase the predictive power for even a simple model; for example, a moderately interpretable model might use 10 latent features as the inputs to a logistic regression or decision tree.

[100] Feature importance metrics can be generated by models such as random forest or XGBoost.

This spectrum is illustrated below. Notably, different types of machine learning models—trees or neural networks—can be implemented in ways that exhibit different degrees of interpretability, depending on a model developers' choices.[101]

FIGURE 3.3.1 A FRAMEWORK FOR MODEL INTERPRETABILITY



MODEL TYPE	INHERENTLY INTERPRETABLE	MODERATELY INTERPRETABLE	UNINTERPRETABLE
	» Linear regression » Scorecards » Generalized linear models » Small decision trees » Logistic regression	» Monotonic tree ensembles » Sparse neural networks » Support vector machines » Generalized additive models » kNN	» Random forests » Deep neural networks

KEY MODEL CHARACTERISTICS

- - Model Size +
- + Linearity -
- + Monotonicity -
- - Complexity +

As shown above and applying the factors related to model complexity discussed earlier in Section 3, inherently interpretable models tend to be smaller and less complex, in part because they use linear and monotonic relationships.

Accordingly, the way in which inherently interpretable models permit review and oversight has made them an attractive option for some lenders that are using machine learning underwriting models. However, as discussed above, a structure that facilitates interpretability may limit the predictive power of the model by limiting the kinds of relationships that the model can identify and use.[102] Proponents of interpretable models suggest any such tradeoffs in this regard reflect the demands of applied use in the context of lending—that underwriting models need to reflect intuitive, defensible relationships between an applicant's financial capabilities and behavior and the model's prediction of default risk. However, for other practitioners, the prospect of better predictive power in complex models has motivated the search for alternative approaches to making complex or uninterpretable models more transparent.

3.3.2 *Post Hoc* Explanations for Machine Learning Models

Model explainability refers to the ability of model stakeholders to understand model behavior—that is, how a particular prediction was made or result was reached.[103] Two general types of explanations can serve these needs. Global explanations describe the high-level decision-making processes used by a model and are relevant to evaluating a model's overall behavior and fitness-for-use. Local explanations identify the basis for specific decisions directed by the model. Both types of explanations are important to enable appropriate human understanding and oversight of AI and machine learning models in financial services contexts.[104]

The explainability of any predictive model can be evaluated, but this issue is particularly important for models that may not be sufficiently transparent without supplemental models and

[101] Model types identified in the diagram are discussed in depth in Section 4.

[102] *See generally* Hall & Gill.

[103] *See generally* The Royal Society.

[104] *Id.*

tools. *Post hoc* techniques have been designed to satisfy transparency needs without affecting the structure or operations of the underlying model, but may impose other costs on the model user, such as requiring large volumes of data, computational power, and additional expertise.

Reliance on *post hoc* explainability techniques raises independent trustworthiness questions, as the methods that produce information about the underlying model's behavior are themselves new and complex.[105] For some stakeholders, these questions are significant enough to warrant restricting firm practice to interpretable models for credit underwriting and score development. More generally, there is neither an established methodology for evaluating the utility and quality of information produced by *post hoc* explainability techniques, nor a consensus about how to assess whether and in what circumstances that information is suitable for important oversight and governance needs.[106]

Certain technical considerations provide a starting point for evaluating individual explainability techniques based on use case, priorities, and resources. For example, the accuracy of an explainability technique's outputs can be measured in terms of consistency and stability—that is, whether the explanation produced is similar across similar applicants assessed by the same model or between different models producing similar predictions trained on the same data. How well the explainability technique approximates the underlying model—or its fidelity as measured in metrics like R^2 scores for a surrogate model—can help establish how well the technique works. The overall complexity and specific data and computational demands of explainability techniques are also relevant—as these factors will increase computation time and cost and may heighten concerns about the trustworthiness of the information being produced.

However, these considerations may not fully speak to whether the information produced by current explainability techniques is sufficiently responsive to applicable legal, regulatory, and firm policy requirements. The fact that most explainability techniques necessarily simplify or compress information about the model they describe naturally raises questions about what information about the underlying model's operation the explainability technique preserves and why.[107]

These questions speak to a critical need to make complex models sufficiently transparent to be comprehensible by a variety of stakeholders, each with their own level and type of expertise and their own need for information. A data scientist or credit risk expert will need different kinds of information about an underwriting model's operation than a financial services executive, a compliance manager, an examiner, an advocate, or a credit applicant.

3.4 Options for Enabling Transparency

This section reviews the choices model developers make and the tools available to them to improve the transparency of models—both the common constraints used to produce models with structures that make them interpretable on their own and *ex post* explainability methods available to help stakeholders analyze, explain, and understand uninterpretable or more complex models. Choices about the type of model used and about data sources can also affect transparency, as discussed further in Section 4.

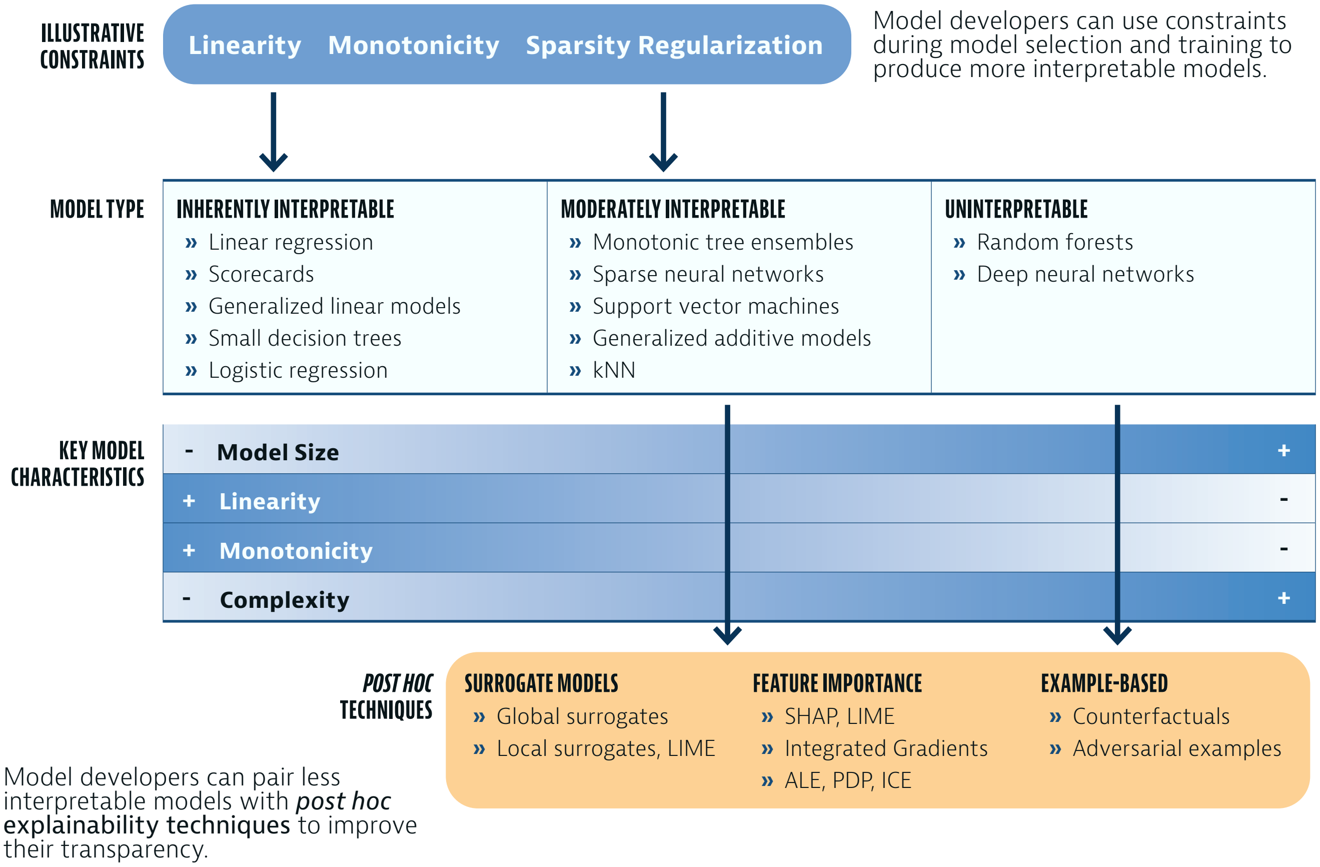
The diagram below summarizes the relationship between various options to improve model transparency.

[105] In the case of AI in the medical sector, *see, e.g.*, Boris Babic & Sara Gerke, Explaining Medical AI Is Easier Said Than Done, Stat (Jul. 21, 2021).

[106] Agus Sudjianto, What We Need Is Interpretable and Not Explainable Machine Learning, presentation at Cogilytica Machine Learning Lifecycle Conference, slides 5-6 (Jan. 2021).

[107] *See* Laura Blattner *et al.*, Unpacking the Black Box: Regulating Algorithmic Decisions (Jul. 21, 2021).

FIGURE 3.4 OPTIONS FOR ENABLING MODEL TRANSPARENCY



In order to produce a machine learning underwriting model with greater transparency, a developer can either apply constraints to the learning algorithm before training to limit the resulting model to certain characteristics or use supplemental models or analyses to explain how the underwriting model works.[108]

3.4.1 Common Constraints

When building a machine learning model, a developer can choose among a variety of approaches for limiting the complexity of the model that the algorithm produces in addition to limiting the scope of input data. In order to balance model complexity and transparency, developers may also choose to limit the model to certain kinds of relationships or limit the number of relationships considered by the model. This decision occurs when the model developer selects an algorithm and readies it for training (as discussed more fully in Section 4).

[108] For illustrative purposes, this report characterizes this as an either/or choice. But some stakeholders argue that model developers should pair inherently interpretable models with *post hoc* explainability techniques in certain contexts, to improve the accuracy of explanations produced in contexts like generating adverse action notices. *See* Patrick Hall *et al.*, Proposed Guidelines for the Responsible Use of Explainable Machine Learning, arXiv:1906.03533v3 (2019)*; see also* Scott Lundberg & Su-In Lee, A Unified Approach to Interpreting Model Predictions, 31st Conference on Neural Information Processing Systems, arXiv:1705.07874v2 (2017); Scott Lundberg *et al.,* Consistent Individualized Feature Attribution for Tree Ensembles, arXiv:1802.03888v3 (2019).

Constraints relevant to the development of underwriting models include:

3.4.1.1 Linearity Constraints

Linearity constraints are imposed to ensure a one-to-one relationship between input variables and the target variable. In other words, the terms of the constraint are of the first-order, which means that for every unit change in a given independent variable, the target variable changes by a fixed amount. In machine learning models, imposing linearity constraints improves transparency as the effect of a change in the feature has a constant change in the target variable. In the context of credit underwriting, a linear relationship may suggest that for every $100 increase in bank card balances, the probability of being approved for loan decreases by 2 percentage points, which makes the association between bank card balances and the probability of being approved for loan very transparent. This also enables a lender to provide a clear explanation for why a person was not approved or how a person can increase their chances of being approved in the future. However, the models with linearity constraints often are not able to capture important and more complex relationships to default risk, which may reduce their accuracy.

3.4.1.2 Monotonicity Constraints

Monotonicity constraints limit the algorithm to developing a model that only uses one-directional relationships between the input data and the predictions of the target variable. When such constraints are imposed, any increase in the feature value that changes the model output always leads either to an increase or decrease in the model output (such as the risk of applicant default).

In the context of underwriting, consider an example where consumers who have low credit card balances and those who have high credit card balances tend to be at higher risks of default than consumers in the middle. With monotonicity constraints, an increase in credit card balances will always lead to either an increase or decrease in predicted default risk, depending on the direction of the monotonicity constraint. This constraint can improve the intuitiveness of this model and its transparency: consumers know that an increase in their card balances will always lower their credit score. However, this monotonicity constraint means that the model cannot reflect the actual relationship between credit card balances and default: consumers in the “middle” of the credit card balances curve have lower default risk than those with higher or lower credit card balances. A monotonic model cannot reflect this distribution through consideration of credit card balances. If the model nevertheless predicts that customers in the middle of the range of values for credit card balances pose lower default risk, this effect will come from assessment of other features.

3.4.1.3 Regularization

Regularization and associated techniques create sparse models by limiting the number of features used as inputs, or by limiting the number of weights in a neural network.[109] Feature selection and engineering are ways to limit the model to the features that are the most relevant to the target variable, which can improve predictiveness and stability and lead to more transparent models. In some cases, sparsity is achieved by dropping variables that are highly correlated. For example, some types of regularization have the effect of keeping the number of parameters small—for example, a model trained with heavy L1 regularization, which is a technique designed to limit the number of parameters—will have a small number of parameters, which has the effect of creating sparsity as well as mitigating potential overfitting problems.

[109] Robert Tibshirani, Regression Shrinkage and Selection via the Lasso, 58 J. of the Royal Statistical Society 267-288 (1996).

3.4.2 *Post Hoc* Explainability Techniques

In the last decade, data scientists have made considerable strides in developing supplemental methods to analyze complex machine learning models to better explain and understand their predictions. These *post hoc* explainability techniques and taxonomies for categorizing them are rapidly evolving, especially as more evidence is gathered about their capabilities and performance in the context of specific applications.

This section first provides an overview and analysis of several individual *post hoc e*xplainability techniques that are potentially relevant to credit underwriting in the following categories: surrogate models, feature importance explainability methods, and example-based explainability methods. It then steps back to discuss potential sources of explanation errors across different types of *post hoc* techniques that cause some stakeholders to advocate for limiting machine learning in credit underwriting to inherently interpretable models.

3.4.2.1 Surrogate Models

Surrogate models are sometimes used to explain uninterpretable or black box models, such as large tree ensembles (including XGBoost) or deep neural networks. Surrogate models are typically small and interpretable models, such as shallow decision trees, rule sets, or regression models. Credit scorecards can also serve this purpose.[110] Surrogate models are designed to closely mimic the original or underlying model, and they are trained on predictions from that model.

Surrogate models come in two general types: global and local surrogate models. Global surrogate models are designed to mimic the overall behavior of the underlying model for every input value. However, global surrogates can often be too general to produce useful insight into the underlying model. For example, the features that impact a default model output for credit card applicants may be very different for applicants who already have a credit card, compared to those who don't. Even a small global surrogate model may not capture this nuance. Instead, practitioners can use local surrogate models, which mimic the behavior of the original model for feature values close to that of a particular applicant and are used to explain that particular applicant's prediction. Local Interpretable Model-Agnostic Explanations—or LIME—is one such approach that is widely used to explain models and has influenced the development of other *post hoc* explainability techniques.

Local Interpretable Model-Agnostic Explanations (LIME)

Description: Local Interpretable Model-Agnostic Explanations (LIME) is an explainability technique for complex models that uses local linear surrogate models around a particular data point to approximate the complex model's output.[111] The resulting local surrogate models are used to both explain the model's behavior around individual data points and to quantify feature importance for the overall model.

In general terms, LIME develops surrogate models by sampling several data points and obtaining the associated predicted outcomes from the complex model. LIME then assigns weights based on how far away the sample points are from the particular point being explained, giving a larger weight to the sampled points closest to the point of interest. Finally, LIME trains an interpretable model—typically a linear model—on the weighted points to produce the surrogate model. This surrogate

[110] For a detailed description of credit scorecards, see Section 4.3.2 of this report.

[111] Damien Garreau & Ulrike von Luxburg, Explaining the Explainer: A First Theoretical Analysis of LIME, Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, arXiv:2001.03447v2 (2020).

model will not altogether explain how the model arrived at the result, but instead how slight changes may affect the ultimate prediction. In the context of an underwriting model that might be sampling nearby data points to train a surrogate model to explain the prediction of a particular applicant's default risk. LIME includes a fidelity measure, giving the user insight into how well the explanation from the surrogate model approximates the underlying or original model.

FIGURE 3.4.2.1 ILLUSTRATIVE LIME ANALYSIS[112]

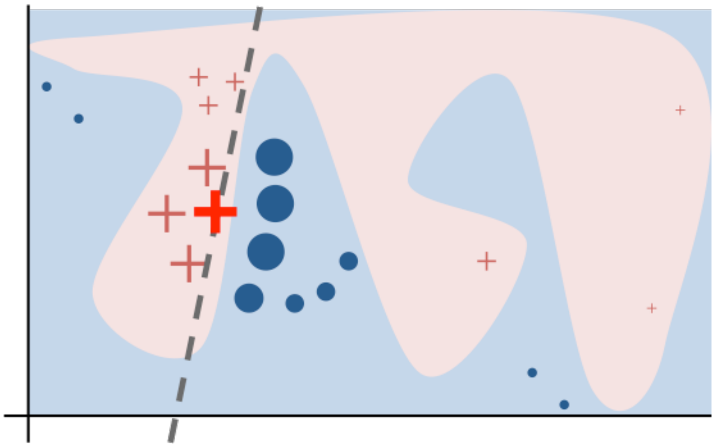


Figure 3.4.2.1 shows a complex model's output values represented by the blue and red shaded areas. The red crosses and blue dots are sampled points, while the bold red cross is the individual point LIME is explaining. To explain that point, LIME generates a dashed line that is close to the nearby border between the blue and red shaded areas.

The characteristics of LIME's surrogate models may diverge from the models they are used to explain in several significant ways. For example, the surrogate is often a linear model. It may also have substantially fewer features than the underlying model. As a result, the explanation produced by the surrogate may not perform well in capturing and explaining feature interactions. For example, credit card applicants may be at high risk of default if they have both (1) more than two credit cards, and (2) high credit utilization. On the other hand, suppose that applicants who have either (1) *or* (2) are not at high risk of default. A linear model cannot represent this effect in the underlying model.

LIME is generally used today as a baseline to compare the outputs and performance of other explainability tools against or to generate insight into feature importance as discussed further below.[113]

Analysis: LIME is versatile and adaptable since it can be used to explain a variety of types of models.[114] It also works across a variety of data types, including text, tabular data, and images. The primary challenge for LIME is derived from the inherent difficulty of using a simplified model to explain a much more complicated model. This challenge is more acute when the surrogate is a linear model, since the surrogate in this instance may not do well in mimicking the effect of non-linear relationships and feature interactions in the underlying model. To resolve this, LIME uses a local surrogate model instead of trying to mimic the underlying or original model at all points and builds

[112] Marco Tulio Ribeiro *et al.*, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, arXiv:1602.04938v3 (2016).

[113] Sérgio Jesus *et al.*, How Can I Choose an Explainer?: An Application-Grounded Evaluation of Post-Hoc Explanations, FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 805-815 (2021).

[114] Jürgen Dieber & Sabrina Kirrane, Why Model Why? Assessing the Strengths and Limitations of LIME, arXiv:2012.00093v1 (2020).

a separate surrogate model for each explanation it produces. This significantly reduces the computational speed of LIME.

When LIME is used to understand the importance of specific features in a model, the feature importance values that LIME generates have very clear meaning: they are derived from linear model weights from the local surrogate model. But these weights can be sensitive to changes in LIME parameters—such as the number of samples used (see Figure 3.4.2.1). Changes to these parameters can substantially change the local surrogate model, and feature importance values, returned by LIME. When LIME explanations are aggregated across an entire dataset, they are sometimes interpreted as global measures of feature importance—that is, how important a feature is to the model's overall behavior. This presentation may be deceptive in circumstances where the model assesses features differently for different consumers, for example. When this occurs, LIME explanations cannot convey this nuance when aggregated across an entire dataset.

Computationally, LIME requires a way of telling how "similar" two given points are, and this must be supplied by the model developer. LIME's explanations are relatively sensitive to this weighting function, which is based on the distances between a sample point and the particular point of interest. In practice, choosing a distance function that produces useful explanations can be challenging.

3.4.2.2 Feature Importance Explainability Methods

Feature importance techniques evaluate how much individual variables contribute to a model's prediction. In these methods, data are usually perturbed or permuted—meaning they are purposefully distorted or altered in a variety of ways—to determine how those changes affect the model's predictions. The aggregated effect of those changes speak to how much a variable affects the model's predictions.[115] Feature importance or variable importance scores can be presented in charts with associated predictions, or they can be aggregated together to describe the importance of a variable on the model's predictions overall, and graphed for comparison.

Feature importance explainability methods include Shapley Additive Explanation (SHAP), integrated gradients, partial dependence plots, individual conditional expectation plots, and accumulated local effects plots.[116]

Shapley Additive Explanations (SHAP)

Description: The Shapley value has been the method chiefly used for the purposes of explaining complex model outputs. SHAP uses mathematical methods derived from a significant body of cooperative game theory research[117] to analyze and explain the contributions of particular features to a model's prediction. The concept of the Shapley value method is as follows:[118] in a cooperative game with N players and a function that values how much total output is generated if all the players contribute together, the Shapley value is a method that attempts to measure the individual contribution of each player to the output generated by the cooperation of all players. If the features are the players in a given complex model, from an economic standpoint, it can be interpreted as a

[115] *See* Molnar.

[116] Although LIME uses surrogate models, some taxonomies of explainability techniques may categorize LIME with feature importance methods since it is widely used to evaluate the contributions that individual features make to a model's predictions. Further, although packages in Python and R like XGBoost include feature importance explainability methods, this section focuses primarily on model agnostic methods. *See also* Molnar.

[117] L.S. Shapley, Notes on the n-Person Game, II: The Value of an n-Person Game, U.S. Air Force Project RAND Research Memorandum (1951); Robert J. Aumann & Lloyd S. Shapley, Values of Non-Atomic Games, Princeton Legacy Library (2016).

[118] Lundberg & Lee.

weighted average of a feature's marginal contribution to every possible subset of grouped features.[119] SHAP methods currently available as machine learning explainability tools are best used on additive models.[120] SHAP is being used to explain complex models in consumer credit and other sectors such as medicine.[121]

Similar to LIME, SHAP explains how a model behaves locally. In the context of credit underwriting, local predictions can be helpful for generating adverse action notices for individuals who are denied credit.[122] However, unlike LIME, SHAP measures feature importance by conditionally averaging over features from a data point. This method measures feature importance by removing features from a data point and quantifying how much the removed features affect the model's output.[123]

Although this approach can be used with underlying models of various kinds, specialized variants of SHAP—such as tree SHAP and linear SHAP—have emerged to be used with particular model types and typically operate faster and produce more reliable outputs than generic implementations.

Analysis: SHAP is attractive to many practitioners because it is available as an open-source tool. SHAP values are also easy to interpret. SHAP values are often presented in plots that show the features and the degree to which they contribute to the target variable and facilitate interpretation. Furthermore, several model-specific versions of SHAP can be faster to calculate than other explainability techniques.

However, there are several criticisms of SHAP. First, many machine learning models cannot naturally handle "missing" features, as required by SHAP. This means that "missingness" is typically achieved by replacing a feature with a nominal value (such as the average over the entire dataset). It is unclear whether the theoretical benefits of SHAP hold up under these approximations. Second, like many other explanation methods including LIME, SHAP makes the unrealistic assumption that features are uncorrelated. This assumption glosses over real-world nuance present in real datasets including those used in financial services. Finally, calculating exact SHAP values can require significant time and computational resources even where model-specific versions of SHAP are used, so approximation or sampling methods are often used instead with some corresponding tradeoff in the quality of explanations.[124] If too few samples are used, moreover, the resulting SHAP values can be noisy, and not reflective of actual model behavior.

Integrated Gradients

Description: Integrated gradients[125] were developed to explain outputs from a differentiable model—that is, a model where the change (or derivative) in model output can be easily calculated.[126]

[119] I. Elizabeth Kumar *et al.*, Problems with Shapley-Value-Based Explanations as Feature Importance Measures, Proceedings of the 37th International Conference on Machine Learning, 119 Proceedings of Machine Learning Research, arXiv:2002.11097v2 (2020).

[120] Additive models are linear models which contain special functions that can learn non-linear relationships in the data.

[121] See Upstart, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning at 3-4.

[122] *Id.* at 4.

[123] Mathematically, this process works as follows: if the point to be explained has three associated features, x1, x2, and x3, binary features are assigned to each one representing whether the feature is known or unknown (so z1 = 0 if x1 is unknown or missing, and z2 = 1 if x2 is known). Next, SHAP values (feature importance values) are generated (a1, a2, a3) which represent a score for each of the features. The higher the score, the more important the feature.

[124] Kumar *et al.*

[125] Gradient, in simple terms, refers to the rate of change of a variable. In a machine learning model, gradient refers to a change in the target variable due to a change in the value of a feature.

[126] Mukund Sundararajan *et al.*, Axiomatic Attribution for Deep Networks, Proceedings of the 34th International Conference on Machine Learning, 70 Proceedings of Machine Learning Research 3319-3328 (2017).

Many popular machine learning models are differentiable, including neural networks. This method works by summing the gradients of the model output with respect to each feature, along some path. Features with greater summed gradients are seen as more important to the model output. Gradients are summed over a *path* of input space, between a data point to be explained (X1) and some reference point (X0). The reference point X0 is meant to be a "neutral" point. In computer vision applications, X0 is usually a blank image; in financial services, X0 might be an applicant with average features.

Analysis: Integrated gradients are attractive for a variety of reasons: they are intuitive, easy to implement, and available in several open-source formats. Although this method does not assume feature independence like LIME or SHAP, it may not do better in explaining feature interactions. Integrated gradients are defined only for continuous models,[127] though some extensions have been proposed for discontinuous models, such as tree ensembles and other piecewise continuous functions.[128] Further, defining a general reference point X0 has a significant impact on the resulting explanation, although choosing this point is difficult in practice. Nevertheless, this method remains popular in applications such as computer vision.[129]

Partial Dependence Plots

Description: Partial dependence plots (PD plots or PDPs) are common visualization methods that depict how an individual feature interacts with the model's predictions.[130] For each value of a given feature, the PD plot shows the average predicted outcome.

Consider as an example an underwriting model that analyzes the following features: number of prior loans taken, number of past defaults, and number of outstanding loans. If the model developer is interested in how the number of past defaults affects the model's prediction of the likelihood of default, a PD plot feeds the underlying model every possible value for the number of past defaults for each possible combination of features in order to understand how the model works. For a single value of the number of past defaults, it will average all those possible combinations and plot the average, then will do the same for all values of the number of past defaults. This means that for every data point, the PD plot will replace the number of defaults with zero, feed a new set of features into the underlying or original model, take the average of the resulting predicted scores, and plot them as a single point on the PD plot. This process is repeated for every value observed in the dataset for the number of past defaults. This ultimately creates a plot of averaged predicted estimates of default probability against all possible numbers for the number of defaults. This analysis can be done for any feature. The user can then see whether or not this relationship is linear, or if there is any value that is particularly surprising that might indicate inaccuracies in the dataset, in the model, or a novel relationship worth analyzing.

Analysis: PD plots are easy to understand and make identifying any remarkable behavior in the relationship between a single feature and the model's prediction easy to detect and intuitive.[131] They are designed to represent the relationship between features and the outcome at a global level,

[127] Continuous models are models which use data that can take any value, such as decimal points. In contrast, discrete models use data at fixed or discrete intervals, such as 0 and 1.

[128] John Merrill *et al.,* Generalized Integrated Gradients: A Practical Method for Explaining Diverse Ensembles, submitted to the J. of Machine Learning Research, arXiv:1909.01869v2 (2019).

[129] Examples include medicine and drug discovery. *See* Rory Sayres *et al.,* Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy, 126 Ophthalmology 552-564 (2019); Kristina Preuer *et al.*, Interpretable Deep Learning in Drug Discovery, in Explainable AI: Interpreting, Explaining, and Visualizing Deep Learning 331-345 (2019).

[130] Daniel W. Apley & Jingyu Zhu, Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, arXiv:1612.08468 (2019).

[131] *Id.*

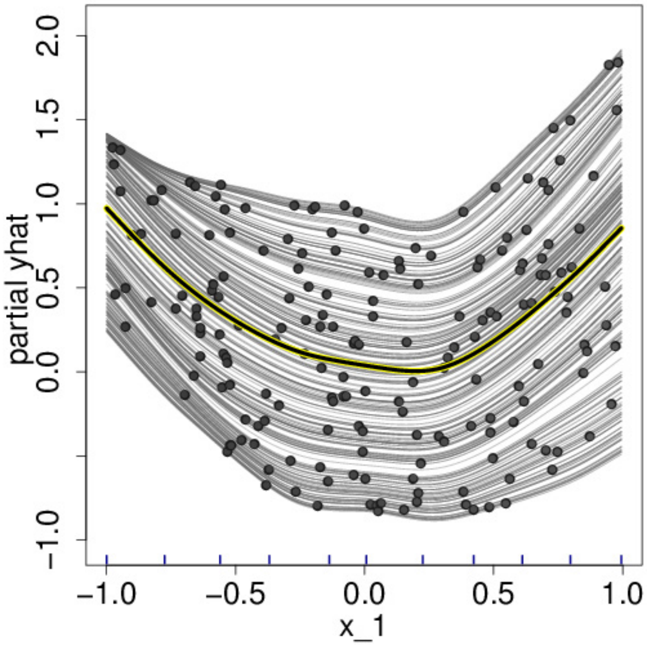
which makes PD plots suitable for model development, but not for generating individual applicant explanations. Additionally, they can be used with any type of machine learning model. However, PD plots rely on individual data points that do not exist in the original dataset. The method replaces real instances found in the dataset with synthetic feature pairings in order to make its averaged predictions over a larger set. Where the synthetic data points do not represent well the actual dataset, this can introduce bias and lead to inaccurate estimates of the effect of the feature on the results. Further, PD plots assume that each feature is independent of each other, which may often not be the case. This assumption may limit the utility of this approach in helping understand how feature interactions and correlated features affect the predictions produced by complex models. Figure 3.4.2.2.2 shows PD plots compared to the following methods.

Individual Conditional Expectation Plots

Description: Individual Conditional Expectation (ICE) plots extend PD plots by displaying the relationship between each individual input and its predicted outcome. This is in contrast to PDPs, which create one line overall for the average. ICE plots supplement PD plots by improving insight into feature interactions. PD plots are poor visualization tools for understanding a dataset that has features that interact with each other in part because averaging across all instances of a feature can often obscure relationships between two features on the output predicted by the model. ICE plots, in contrast, do not involve averaging.

For example, in a sample ICE plot such as Figure 3.4.2.2.1 below, if "x_1" axis represents the number of past defaults and "partial yhat" represents the predicted probability of default, then the curves would show the change in the predicted probability of default as the number of past defaults varies. These plots also provide insight into how the number of past defaults interacts with, for example, the number of months since the last credit card was opened. If the number of past defaults and number of months since the last credit card opened were to show some interaction on the predicted probability of default, then the curve of the lines for instances where the individual has opened a credit card one month ago for those who opened a card 24 months ago would have different shapes/slopes. In this hypothetical example depicted in Figure 3.4.2.2.1, lines representing different amounts of elapsed time since the applicant's last card was opened do not interact in the output of predicting default since the curves all have the same parabolic shape and simply show a shift along the y-axis. This means that based on this figure, number of months since last credit card opened has no non-linear relationship with number of past defaults and so does not change its influence on predicted defaults.

FIGURE 3.4.2.2.1 ILLUSTRATIVE ICE PLOT[132]



(a) ICE

Analysis: ICE plots show each instance or person in the dataset as a single line, where the value of the feature of interest varies. This makes the plot more interpretable to the user who can even focus on a given line and see how changing a feature like the number of past defaults might affect the given individual.[133] Similar to PDPs, ICE plots are not able to generate reliable estimates with correlated features, which may create congested plots and means that they cannot fully explain relationships between features.

Accumulated Local Effects Plots

Description: Accumulated Local Effects (ALE) plots go beyond PD plots by focusing only on changing the feature of interest rather than every feature involved in the model.[134] Instead of exhaustively trying to predict the relationship between a feature and the outcome of interest, ALE does not include every feature involved in the model and instead focuses only on changing the feature that will be plotted against and takes the average prediction over a small interval of the data. This means that the plotting procedure for ALE is similar to PDP, but in the example described above, the supporting features (number of prior loans, number of outstanding loans) remain constant and only the number of defaults changes for an ALE calculation. While PD plots explain the model by providing every possible combination of feature values and use synthetic data points to do so, ALE only averages over values that exist in the actual training data. Since divergences between actual and synthetic data in PD plots can introduce bias, ALE is generally more accurate than PD plots in producing explanations, but requires more data than other methods.

[132] Alex Goldstein *et al.*, Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation, arXiv:1309.6392v2 (2014).

[133] *Id.* at 10.

[134] Apley & Zhu.

Analysis: ALE is computationally more efficient than PD plots and, unlike PD plots, ALE explanations can show feature correlations as well as feature interactions.[135] However, ALE presents the user with a large range of effects a feature can have on the output, and this range can support a wider range of interpretations of the information presented. ALE plots are somewhat less intuitive than PD and ICE plots, and at the same time they convey more nuanced information about model behavior.

FIGURE 3.4.2.2.2 ILLUSTRATIVE ALE AND PDP PLOTS[136]

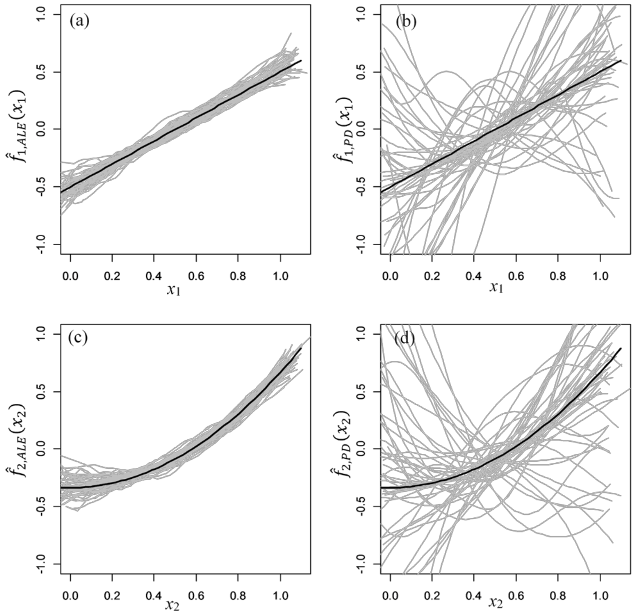


Figure 3.4.2.2.2 above shows ALE plots on the left, and PD plots on the right, with the black line representing true effects in all plots. This shows that with every iteration of the model, ALE plots are more consistent with the true effect of model features on the output than the PD plots.

3.4.2.3 Example-Based Explainability Techniques

Example-based explanations generate explanations by perturbing selected data instances from the model to see how the changes affect the associated prediction, rather than the feature-importance explanations above which create "summaries" of the effect of particular variables.[137] Conceptually, an example-based explanation is similar to a human relying on historic examples from his or her own experience to predict the outcome of a new experience.[138]

There are two primary forms of example-based explainability techniques relevant to current practice: counterfactual explanations and adversarial perturbation.[139]

[135] *Id.*

[136] *Id.* at 13.

[137] Molnar.

[138] Susanne Dandl & Christoph Molnar, Counterfactual Explanations (2019).

[139] Additional types of example-based explanations include prototype and criticism examples. Prototype examples explain a particular class or model output. For example, to describe the model output "accepted credit card application," a developer might generate 10 prototypical applicants, which are typical "accepted" applications, as well as 10 typical "rejected" applications. Criticism examples are designed to illustrate atypical examples. For instance, a criticism example for the credit card model might be an applicant who is similar to many accepted applicants, but was actually rejected.

Counterfactual Explanations

Description: Counterfactuals describe—either in plain text or in charts with data—how much a data point has to change in order to change the prediction. In other words, if someone is denied credit, a counterfactual explanation will search for the smallest possible change someone could make to the factors assessed in an underwriting analysis to change the model's prediction of his or her likelihood of default. When displayed in plain text, a counterfactual might say "If X had not occurred, Y would not have occurred", or, "If a person had not made a late mortgage payment in 2019, they would not have been denied credit in 2020." These changes might be in one factor (for example, "increase your income by $2,000") or in multiple factors (for example, "close two of your credit cards and pay off all debt"). Counterfactuals can also generate charts of numbers that represent the smallest changes which can be made to change the relevant outcome. These numbers can be graphed to represent a linear relationship between inputs and predictions. So the chart might show, if a person made $2,000 more per year, or had three fewer credit cards, they would have been accepted for credit.[140]

Counterfactuals might be especially useful for helping to generate adverse action notices used in underwriting as they provide a clear statement of how a change in a particular input would affect the model's prediction of default and the lender's decision about whether to extend credit.[141]

Analysis: Building counterfactual explanations does not require access to the underlying data used to train the model—unlike other explainability methods—which may be useful in cases where data cannot be shared. Rather, counterfactual explanations can be generated using a sample input data point (*e.g.*, a credit card applicant) and the model (an underwriting model). Counterfactual examples are identified by making small modifications to the input data point until the output changes by a sufficient amount.

Several counterfactual examples can be generated to show a range of changes to relevant characteristics might affect the model's prediction. For example, a counterfactual analysis may suggest a dozen ways that an unsuccessful applicant for a loan could improve the chances of a future approval. While this provides useful information that is not generated by various other explainability techniques, these explanations may create a false sense of precision or be confusing to applicants. For instance, some counterfactual changes—like eliminating a bankruptcy—may not be practical, while others may simply be confusing to understand and prioritize, and ultimately have limited use for consumers.

Adversarial Examples

Description: Adversarial examples are used to illustrate cases where a model makes mistakes or errors. The intent is to identify possible weaknesses or failure points for the model.

Adversarial example explainability methods work by changing features to produce a false prediction in the model. An adversarial model "perturbs" data instances to try to "deceive" the model into making false predictions. This approach is common with image recognition, for instance by changing pixels to have a model incorrectly label a picture of a dog as a car. While adversarial examples may popularly be used to find errors in image recognition models, the method can be applied across a range of use cases: adversarial examples might be identified by perturbing inputs in an underwriting model to try to change a prediction so that someone who should get credit access is denied or so

[140] Dandl & Molnar.

[141] Solon Barocas *et al.*, The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons, ACM Conference on Fairness, Accountability, and Transparency (FAT*) 5 (2020).

that someone who is not creditworthy is approved. These examples can help identify weaknesses or unexpected behavior in machine learning models. In financial services, there is growing interest in the use of adversarial models for debiasing (see Section 5.3.1.2).

Analysis: Adversarial examples are especially useful for debugging complex machine learning models: they can highlight certain inputs that cause the model to make mistakes, or behave in unexpected ways. However, it can be computationally expensive and time-consuming to identify adversarial examples, especially for complex models with many features. In addition, these require information about actual outcomes, which is not always available. The resulting adversarial examples are also sometimes difficult to interpret and impractical to use. In computer vision, adversarial examples often have added noise or artifacts that would not be found in real images.[142] Similarly, an adversarial example in a lending context that involves an applicant with 20 credit cards, no credit history, and an income of $2,000,000 may not reveal meaningful weaknesses in a model because such a case is highly unlikely in real life.

3.4.2.4 Sources of Explanation Errors

As discussed above in connection with individual techniques, several potential sources of errors in *post hoc* techniques may produce inaccurate, unstable, or unusable information. The risk of such errors is a central point in the debate over whether to limit use of machine learning in credit underwriting to inherently interpretable models.

Oversimplification: Intuitively, many *post hoc* explainability methods operate by using a simpler surrogate model to understand the behavior of a more complex model. The surrogate is fitted to the predictions of the underlying machine model and then derives a simpler, more interpretable model that approximates those outputs. However, this approach does not necessarily mean that the surrogate will reliably identify the features or relationships that caused the underlying model's predictions.

More generally, the more the surrogate and the underlying models vary in terms of complexity, the more the explainability technique may simplify the underlying model's operations and the bases for its predictions. At some point, this simplification may result in the explanation not capturing the full or true causes of the underlying model's predictions. In extreme cases, the explanation may even provide misleading information about the bases for the model's prediction. The effect of oversimplification can be described as information loss or compression, although there is no standard approach to identifying or measuring how oversimplification affects individual implementations of explainability techniques.

This may occur in a variety of situations, including when a linear surrogate model is used to explain an underlying model with non-linear relationships. In this case, the resulting explanation may not capture the effects of the underlying model's non-linear relationships and feature interactions in the underlying model. Oversimplification can also occur when building local surrogate (simple) models, as in LIME. The local surrogate may be a decent representation of the complex model for a single data point and its neighborhood, but this surrogate model may be very inaccurate for other data points or when used to provide a global explanation.

Assumption of Feature Independence: Many *post hoc* explainability techniques, including PD plots and SHAP, assume that all input variables are independent, even though this assumption is false in almost every case and feature interactions contribute to the overall challenge of understanding

[142] For a survey of adversarial methods in machine learning, *see* Naveed Akhtar & Ajmal Mian, Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey, 6 IEEE Access 14410-14430 (2018). One of the original papers on this topic is Ian J. Goodfellow *et al.*, Explaining and Harnessing Adversarial Examples, published as a conference paper at the 2015 International Conference on Learning Representations, arXiv:1412.6572 (2015).

and explaining machine learning models. This mismatch alone calls into serious question the utility of many *post hoc* techniques in applied settings. Many areas of machine learning and applied statistics also rely on the same assumption, but it can lead to particularly inconsistent and confusing results in connection with *post hoc* explainability techniques.

When multiple input features are strongly correlated or there are causal relationships, explanations may be inconsistent or inaccurate in unpredictable ways. For example, if two strongly-correlated features are closely related to a model's output, then some explanation methods may arbitrarily over- or under-weight the importance of one of these variables.

Convergence: Some *post hoc* explainability methods, including LIME and SHAP, often use sampling to arrive at a result. For example, some versions of SHAP generate several samples by perturbing the input data, and calculating how these perturbations change the output. The resulting explanation can change as more samples are used, and most methods eventually "converge" when a sufficiently large number of samples are used. If the number of samples is too small, however, the resulting explanation can be inaccurate in unpredictable ways.

The number of samples required for convergence increases with both (a) the number and range of input variables, and (b) the model complexity, but model developers will not generally be able to know with certainty *ex ante* how many samples will be needed.

User Error and Misuse of Explanations: Information produced by various *post hoc* explainability techniques can be prone to misinterpretation or misuse based on the expertise and experience of the user of that information. Explanations that are designed to answer a specific question—what caused a particular drop in model accuracy?—may not satisfy a different inquiry—how do economic downturns affect model accuracy? Further, the technical nature of information provided by explainability techniques increases not just the importance of having personnel with the relevant expertise reviewing this information, but incorporating interdisciplinary expertise into this process. For example, common explainability techniques cannot distinguish between causal effects and confounding variables that improperly suggest a relationship where none exists. In this scenario, a *post hoc* explainability method may identify a particular feature as very important to the model's prediction even though it serves as a proxy or confounding variable. For example, it may be common knowledge in financial services that the oldest tradeline is associated with a greater number of tradelines, and that applicants with fewer tradelines have a greater default risk. A machine learning model and *post hoc* explainability method may highlight the confounding variable, the oldest tradeline, as important, when in fact the number of tradelines is the variable that leads to greater default risk. This mistake would be easier to pick up by a financial services expert than a data scientist who may lack the appropriate context to ensure proper review and challenge of model explanations.[143]

In a broader sense, errors resulting from misuse of explanations can also occur when there is a mismatch between the purpose for which an explanation is needed and the technique used to derive it. When users rely on local explanations such as LIME or SHAP to describe global model behavior, differences in model output and quality across the input distribution are concealed and may undercut the accuracy of the explanation. For example, the average feature importance for a complex underwriting model may be very intuitive—such as lower credit utilization is on average associated with lower default risk. However, it is entirely possible for a machine learning model to use the opposite rule—low credit utilization is associated with high default risk—for certain applicants. An explainability technique that uses averages over the entire population would not return an explanation that reflects this counterintuitive behavior in circumstances

[143] The converse may also present problems and points to the need for diversified governance and oversight: a financial services expert may be prone to confirmation bias whereas a data scientist or other stakeholder may adhere more closely to specific findings in the data.

where it applies. Another area for mismatch may be between a lender and a regulatory authority, as they use explainability techniques for different purposes. In this case, it may be important to use a technique that is designed to mitigate the misalignment in incentives between the lender and regulatory authority.[144]

3.5 Emerging Model Diagnostic Tools

Given significant advances in the data science of explainability in recent years, a group of companies offering products that incorporate various *post hoc* explainability techniques has emerged to help firms manage AI and machine learning models. Some of these model diagnostic tools are designed to support models deployed in a variety of sectors and use cases. Others have been specifically designed to help lenders design, implement, and manage machine learning underwriting models. Some vendors have businesses focused on supporting model development teams, others primarily sell to risk and compliance units.

Although individual lenders have different appetites for relying on vendor-provided tools to manage various aspects of machine learning models, the emergence of proprietary and open-source model diagnostic tools presents two important opportunities. First, these products can help standardize model management practices across the market. They may also make adoption of machine learning underwriting models more feasible for lenders that would otherwise struggle to build and maintain the relevant internal capabilities and infrastructure.

These tools are designed to enable oversight of machine learning models across several dimensions, including the general operation of models and managing regulatory compliance issues relevant to consumer credit. For example, many of the vendors offer support for managing model fairness and bias concerns. The tools can give model developers deeper insight into individual features' predictiveness and are designed to let model developers manage accuracy-fairness tradeoffs in more nuanced ways than is possible with incumbent models.[145] If validated and realized at scale, these developments have the potential to improve outcomes for both lenders and borrowers.

Although the characteristics of individual companies' model diagnostic tools and the data science techniques embedded in them vary, emerging offerings in this market niche include the following capabilities:

- **Auto ML:** Auto ML software guides users through the development of machine learning models. Such software typically offers a more automated model development process. These products are designed to help streamline model building processes that can be iterative and time consuming—such as selecting between machine learning types, hyperparameter tuning, feature engineering, and model assessment—based on input from the model developer responsible for creating the new model. Auto ML products are typically designed to help the model developer understand and document the tradeoffs between various design and implementation options by, for example, showing metrics for the performance, stability, and fairness of various iterations of models. The diagnostics showing tradeoffs among different types of models and model specifications enable these platforms to efficiently fit many different types of models in parallel on the same dataset and allow the developer to choose a single model for use or several models to be used in an ensemble. Auto ML can also be used by model developers to customize models and

[144] Blattner *et al.* (Jul. 2021).

[145] Nicholas Schmidt & Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination, 73 Consumer Finance Law Quarterly Report 130-144 (2019).

tune models with proprietary data. Auto ML providers may not have direct insight into the nature or performance of the models developed using their platforms.

- **Model Diagnostics and Monitoring:** Model diagnostic and monitoring platforms are designed to support machine learning models designed by the user. These tools typically provide and document insights of various kinds into the operation and performance of the models and enable model users to adjust aspects of the model's operations and performance based on these insights. The kinds of information produced can vary widely based on the product's intended clientele, as well as the state of maturity of the product and provider. Some products in this category may focus on enabling oversight with respect to a single risk area—like fair lending—while others may aim to support a more broadly based set of requirements. These products typically enable users to set customized alerts to notify them when certain performance or risk thresholds have been exceeded. Firms using this approach often work closely with clients to support implementation of their analytics, train client personnel on how to use the software and interpret the information that it generates, and provide subject-matter expertise on critical interpretative issues and decisions once the platform is in use.
- **Model Development:** Some firms offer similar model diagnostics and monitoring capabilities as described above, but do so as part of a package that includes developing their clients' underwriting models. In this context, helping clients meet model validation and other risk management obligations can be essential to making them comfortable that they can rely on vendor-provided models and meet strategic and regulatory requirements.

Firms offering these products interact with their clients' use in various ways, ranging from minimal support focused on implementation and use of the software to full consulting-style advisory support on how to interpret information produced by the tools and manage individual models.

Products currently available to lenders come in both open-source and proprietary forms. Many of the open-source tools, such as SHAP and built-in features of the XGBoost package, originate with established technology firms and are designed for use across economic sectors and use cases. Proprietary tools can be provided by startup or *de novo* entrants, established technology companies, or analytics providers (such as credit bureaus and score providers). The business rationale and strategy of each provider of proprietary tools will shape key product characteristics, including their scope, methodology, specificity as to particular use case needs (such as regulatory compliance), and user interface design, among others. In some cases, proprietary model diagnostic tools leverage open-source tools and use them with refinements to adapt them to users' needs in specific use cases or to improve their operational efficiency.

Model diagnostic tools can be used in various ways. A lender may opt to use vendor-provided model diagnostic tools directly as part of developing, monitoring, and operating machine learning underwriting models. Credit risk and data science teams may also use proprietary and open-source model diagnostic tools indirectly as an additional check on the transparency of internally developed models and on the performance and capabilities of explainability techniques their teams have developed as part of building their own models. Stakeholders focused on oversight—internally in firm risk and compliance functions and externally in regulatory examination teams—may also rely in time on these kinds of tools to understand and explain machine learning underwriting models. One likely axis for further refinement of model diagnostic tools is in their ability to support use by interdisciplinary stakeholders—including those in oversight functions who may have less credit risk and data science expertise than model developers and those who need the ability to generate and document independent evaluations of various aspects of an underwriting model's operation and performance.

4. MODELLING CONSIDERATIONS

This section addresses a range of considerations that individual model development teams consider during the process of designing, implementing, and operating machine learning underwriting models, which is summarized in Box 4.1.2. These issues are not necessarily unique to machine learning models, but may take on new dimensions for lenders intent on using advanced analytical techniques to develop and operate underwriting models. The model development decisions considered here—individually and collectively—inform the accuracy, fairness, and inclusiveness of resulting models. Many of them will also affect the transparency of the resulting underwriting models, in addition to the decisions discussed in Section 3. Section 5 provides a more extended consideration of bias-related issues that typically receive sustained attention across all stages of model development and use.

4.1 Algorithm Selection

Credit underwriting presents a classic classification problem: how can a lender determine whether an applicant is likely to repay the loan for which he or she has applied.[146] Current automated underwriting systems typically use logistic regression or logit models to estimate the probability of default because of their relative interpretability. However, lenders are increasingly interested in leveraging large, complex datasets and enhanced computational power to make credit decisions using machine learning underwriting models. This section considers the various kinds of machine learning models that were designed for classification problems like credit underwriting.

4.1.1 Types of Machine Learning Relevant to Underwriting and Other Credit Activities

Machine learning refers to the subset of artificial intelligence that gives "computers the ability to learn without being explicitly programmed."[147] Figure 4.1.1 provides a simple overview of the relationship of various types of artificial intelligence.

[146] A classification problem generally refers to a situation in which a target variable is categorical or binary, in contrast to a regression problem, where the output is a continuous variable. In the context of credit underwriting, the process of sorting credit applicants into high- and low-risk buckets would be considered a classification problem.

[147] *See* Financial Stability Board; *see also* Samuel at 211-229; Mitchell (defining machine learning as the "field of study that gives computers the ability to learn without being explicitly programmed"); Jordan & Mitchell (defining machine learning as "the question of how to build computers that improve automatically through experience").

BOX 4.1.1 MACHINE LEARNING MODEL LIFECYCLE

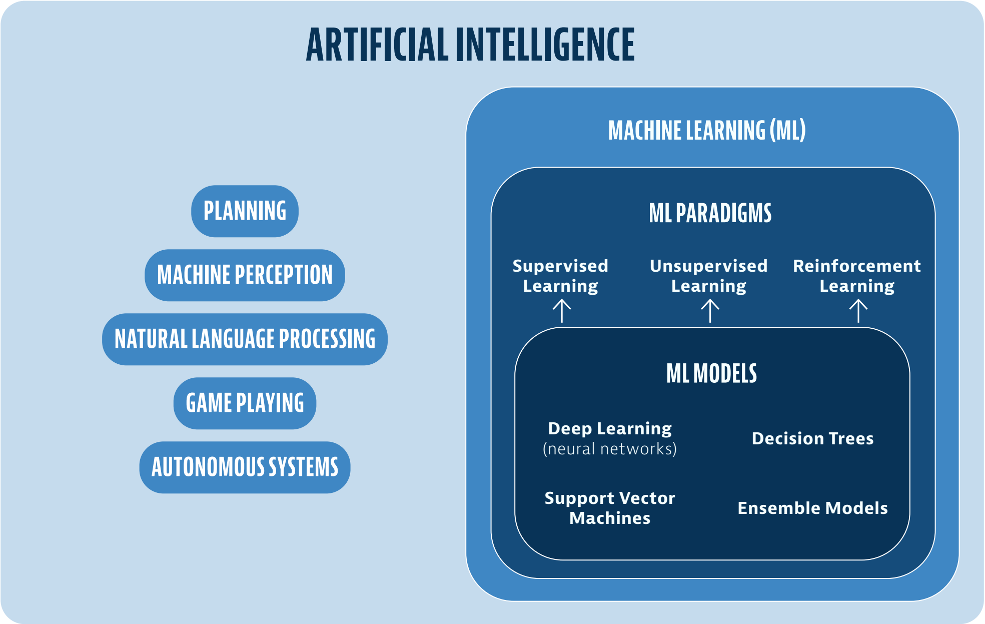
Although the importance and intensity of individual steps can vary, machine learning models are generally developed in the following steps. These steps typically occur within model development teams and business units prior to any formal validation and oversight processes that an institution might require, although how each occurs and is documented will be governed by the requirements of those processes:

- Algorithm Selection: The model builder picks a learning algorithm or algorithms based on the application for which the model is being developed and plans how to use constraints and *post hoc* methods to ensure appropriate transparency. Tradeoffs between algorithm types include the amount of data required, the types of relationships they can look for, and the ease of providing appropriate explanations of how the model works and for specific results.
- Data Selection and Preparation: Preparing data is a critical and time-consuming stage of developing a machine learning model. Choices that developers make at this stage can have broad effects on the performance, fairness, and inclusiveness of models. Decisions taken to clean data may also affect the reliability of information expressed by *post hoc* explainability techniques.
- Training: Model training is the period in which an algorithm analyzes a dataset to identify thresholds and relationships relevant to prediction of the target or output variable. Compared to traditional statistical modelling techniques, the machine learning algorithm, rather than a human coder, determines the structure of the resulting model.
- Validation and testing: After training, predictive models evaluate hold-out data—datasets other than the one on which it was trained, often including out-of-time samples—to evaluate its reliability and robustness. Test data are typically data that neither the data scientist nor model have seen. This step is particularly important in building machine learning models given the risk of overfitting—the risk that the machine learning algorithm fits the predictive model too narrowly to the specific characteristics of limited training data, which may increase the fragility of the model's performance.
- Tuning: Machine learning models are then "tuned" in order to maximize performance based on validation and testing results. Tuning, validation, and testing may occur in several iterations during model development. Tuning is a critical step to reduce overfitting problems. As discussed above, regularization is one technique used to tune a model—here, an additional term is added to constrain the model so that specific coefficients cannot take extreme values. Hyperparameters can also be used to adjust models and are set before training begins either by a data scientist or auto ML software. Their values can be changed during tuning.
- Shadow deployment: Firms typically run the developmental model parallel to models that are already in production. This permits direct comparison to incumbent models on performance, stability, and other metrics relevant to the use case and refinement of model design, implementation, and risk management plans.

In managing the development and use of machine learning models, users will monitor model performance, data conditions, and other factors to decide when to re-train and update the model. Although machine learning confers certain efficiencies on the updating process, this process typically occurs offline and is subject to oversight processes similar to those for development of a new underwriting model.[a]

[a] Bank Policy Institute, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning 17 (Jun. 25, 2021).

FIGURE 4.1.1 FORMS OF AI AND MACHINE LEARNING



Among the forms of machine learning, two types are relevant to the development of underwriting models: supervised learning and unsupervised learning.[148] Additional types of artificial intelligence and machine learning may be used in other credit-related activities, such as the use of natural language processing in processing borrower inquiries and complaints. (See Box 4.1.2)

4.1.1.1 Supervised Learning

Supervised learning refers to models that are trained using both input variables and a target (outcome) variable. The purpose of supervised machine learning models is to learn to predict the target variable from input variables. Supervised learning models are almost always used when lenders are estimating the probability of default by a potential borrower, and are commonly used for credit scoring and stress testing.[149] For example in credit underwriting, a lender might use a borrower's attributes, such as the percentage of available credit used by the borrower and employment length (the input variables), to predict whether or not they will repay a loan (the target variable). Information on past borrowers, including their repayment behavior, can be used to "train" a supervised machine learning model that predicts the repayment behavior of new borrowers. Once a new application is made for a loan, supervised learning models can then predict the likelihood of default by analyzing the applicant's information in the context of past borrowers' repayment behavior and outcomes. In practice, this means underwriting models depend almost entirely on historical lending data, which as discussed further in Section 5 can enhance challenges related to fairness and bias in assessing accurately the credit risk of individuals in groups that have had limited prior access to credit.

[148] A third type—reinforcement learning—is excluded here because it has little application to underwriting models. In reinforcement learning, a machine learning model or agent interacts with a dynamic environment by taking actions and receiving rewards. Reinforcement learning is commonly used in robotics and game-playing. This includes learning to make a series of decisions correctly—such as playing and winning games.

[149] Jie Chen, Deep Insights into Explainability and Interpretability of Machine Learning Algorithms and Applications to Risk Management, Presentation at the 2019 Joint Statistical Meetings, slide 2 (Jul. 29, 2019).

BOX 4.1.2 OTHER USES OF MACHINE LEARNING IN LENDING

While this report focuses primarily on the use of machine learning underwriting models, lenders are using various forms of AI and machine learning in other aspects of their operations that can have important implications for consumers in accessing and using credit. For example, while the use of AI and machine learning for marketing analyses is distinct from underwriting, this use case may influence underwriting by affirmatively shaping who becomes an applicant for credit or providing insight into relevant segments to be considered in credit decisions.

In addition to underwriting models, AI and machine learning can be used in the following ways throughout the lifecycle of credit products:

Marketing: AI and machine learning can help lenders identify potential customers for their products and services and support the creation of prescreened offers of credit[a] that reflect differing levels of fit with underwriting criteria. In this context, unsupervised learning is more common than it is in credit risk assessment, as it relies on diverse forms of unstructured, digital data.

Fraud: Fraud screening has become both more important and more complex in the era of digital transactions and application channels. Fraud screening is a well-established use case for both complex AI models like neural networks and varied types of digital data. Here, complex models are a natural fit for high-speed, data-intensive, iterative processes used to identify individual risks of illicit activity based on rapidly changing patterns within massive volumes of streaming data flows.[b] These models can be used both to determine which credit applications are evaluated in full underwriting processes and to evaluate individual transactions involving open-end credit, such as screening individual credit card transactions. Concerns about model transparency may not be as acute for fraud screening as in underwriting decisions because applications or transactions flagged as high risk are often subject to further review before exposing firms to financial or other forms of liability, such as declined transactions.

Loan Servicing: Lenders can use machine learning to help identify borrowers who are most likely to falter in repayment, to focus particular collection activities on those most likely and able to repay, and to determine appropriate loan terms when a modification or workout is needed. Since many of these activities carry significant financial and regulatory exposure, lenders may be more conservative about deploying machine learning for them than in some other lending-related contexts. However, in collections and recovery, activities such as using combinations of supervised learning, natural language processing, and text mining for sentiment analysis may raise significantly different risks and model transparency needs than applications that involve making decisions about loan originations or workout terms.

Portfolio Management: Lenders use machine learning to assess the performance and positioning of current credit portfolios as market and operating conditions change. The ability of machine learning models to detect non-linear relationships as well as their ability to be retrained relatively swiftly once additional data are available, may enhance their utility in this role when economic conditions shift dramatically, such as in the COVID-19 pandemic. These efforts can inform lenders' decisions to adjust their credit policies going forward, especially with respect to score cutoffs and line assignments, when changes in strategy or macroeconomic conditions affect their credit risk and capital positioning.

Customer Relationship Management: Firms are expanding their use of advanced analytics to handle and respond to customer communications in a variety of contexts. Analytical techniques like natural language processing may be useful for scanning large volumes of calls and enabling chatbot digital interfaces. When assessed in large volumes, these customer contacts may provide useful information about patterns and practices relevant to refinement of customer acquisition strategies and adjustments to credit practices.

Regulatory Compliance: Firms may also seek to enhance risk management and compliance processes using AI and machine learning. For example, firms have begun to use advanced analytical techniques to identify patterns within consumer complaints and conduct root cause analysis. Regulators are using AI and machine learning technology for similar purposes.[c]

[a] In general, a prescreened offer of credit refers to a solicitation from a lender that invites the recipient to apply for a loan based on a preliminary review of his or her credit bureau information. *See* Consumer Financial Protection Bureau, What Is a Prescreened Credit Card Offer? (2017); Federal Trade Commission, Prescreened Credit and Insurance Offers (2021).

[b] *See, e.g.,* Sushmito Ghosh & Douglas L. Reilly, Credit Card Fraud Detection with a Neural-Network, The Twenty-Seventh Hawaii International Conference on System Sciences (1994).

[c] *See* Consumer Financial Protection Bureau, 2020 Consumer Response Annual Report 5-6 (2021); Jo Ann Barefoot, A Regtech Manifesto: Redesigning Financial Regulation for the Digital Age, Alliance for Innovative Regulation (2020).

4.1.1.2 Unsupervised Learning

Unsupervised learning refers to models that detect patterns or clusters in a dataset without using a target variable. Data without a target variable are referred to as unlabeled. To use the same example, the dataset for an unsupervised learning model will consist of features, which may include credit utilization and employment length, but will not include information on loan repayment behavior. Although unsupervised learning models are not used directly to predict the probability of default, they can be used to find similarities or associations between the features and characteristics of individuals included in a dataset and to label cases of interest for a supervised model. Such clustering of borrowers is widely used for customer segmentation in marketing and in models used for image recognition. In lending, it is sometimes used to optimize credit lines.

As described above, deep learning refers to a class of machine learning models that emulate biological neural networks to identify complex patterns in data and extract higher-level information between the input data, latent features generated by the model from the input variables, and target variables.[150] Deep learning can be used in both supervised and unsupervised models.[151] The generation of these insights often depends on the transformation of input data through various layers or levels involving the generation of latent features. These models can analyze various large datasets and identify complicated and detailed patterns, which, in turn, can capture deeper insights and connections between the features and outcomes in the datasets. Deep learning has gained prominence in recent years, as these models can be highly predictive and improved computing power has continued to reduce processing times for these models. Deep learning has been used in natural language processing, computer vision, and semantic learning, among other applications, but as discussed further below use of unconstrained neural networks in credit underwriting is relatively limited due to the complexity of explaining the operation of these models and tradeoffs in performance that result from constraints to improve model transparency.[152]

4.1.2 Types of Machine Learning Models Used in Credit Underwriting

This section describes the types of machine learning models that are most relevant to credit underwriting. At the outset, all of the models discussed below can exhibit varying degrees of model complexity and transparency, depending on specific choices that lenders make when developing underwriting models as discussed in Section 3 with regard to model constraints and in Section 4.2 with regard to data inputs. Factors that affect these modelling choices include data availability; systems infrastructure; firm practice, policy requirements, and operational constraints; and regulatory considerations.

4.1.2.1 Forms of Supervised Learning Used in Credit Underwriting

4.1.2.1.1 *Tree-Based Models*

Among machine learning models, tree-based models are often used in credit underwriting because they offer lenders an attractive balance of predictive power and operational efficiency.[153]

[150] Radoslaw M. Cichy & Daniel Kaiser, Deep Neural Networks as Scientific Models, 23 Trends in Cognitive Sciences 305-317 (2019).

[151] Deep learning models can also be used with reinforcement learning techniques.

[152] Majid Bazarbash, FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk, International Monetary Fund Working Paper (2019).

[153] A 2016 study of data from six large U.S. banks found that decision tree and random forest models that considered tradeline data, credit bureau information, and macroeconomic indicators each outperformed a more traditional logistic regression model when forecasting credit card delinquencies. *See* Butaru *et al.*

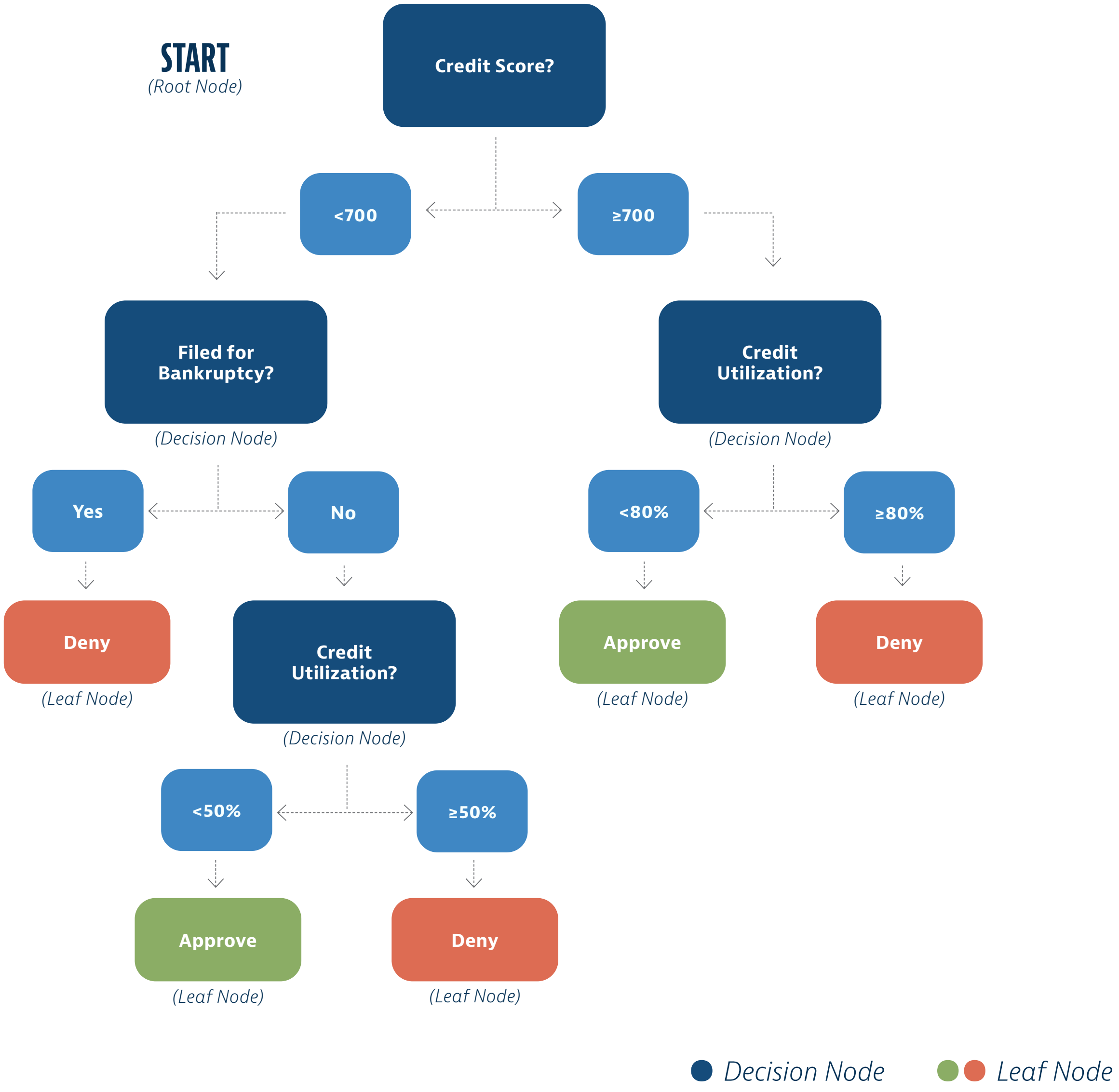
Tree-based models come in various forms and degrees of complexity—some tree-based ensemble methods may be as complex as multiple-layer neural networks—but all are built using traditional "if-then" logic to break the estimation of the target variable into a series of discrete, binary analyses. Three types of tree-based models have particular relevance for credit underwriting: decision trees, random forests, and gradient-boosted decision trees (including XGBoost).

Decision Trees

A decision tree is an algorithm that uses a hierarchical structure to estimate a target variable—such as whether an applicant should be accepted or rejected for credit because he or she has a high or low risk of default—with a series of discrete, binary rules. These smaller decisions are represented in a chain: each step of the chain is called a node, which corresponds to a simple "if-then" decision. The result of each decision at a node leads to a new node, and so on. Eventually, one of these steps leads to a leaf node, which gives an estimate of the target variable. Collectively, this set of component decisions in all of the branches of a tree will include all possible outcomes in a hierarchical structure.

A simplified decision tree for credit underwriting illustrates this approach:

FIGURE 4.1.2.1.1 DECISION TREE FOR A CREDIT UNDERWRITING MODEL



In Figure 4.1.2.1.1, a decision tree model is used to predict which applicant is likely to pay back a loan to inform a firm's decision to approve or decline an application for credit. The decision tree starts with a credit score—any applicant with a credit score of 700 or higher is assigned to the branch on the right of the diagram, which then looks at the applicant's credit utilization rate. A threshold of 80% is determined by the model to be the cutoff, where any applicant with a credit utilization rate of under 80% will be approved for credit. However, the analysis for an applicant with a credit score of below 700 will be assigned to the branch on the left that starts by looking at whether the applicant has had a bankruptcy. If so, the applicant will be declined for credit. If the applicant has not filed for bankruptcy, the decision tree then looks at the credit utilization rate of the applicant, which again has a cutoff of 80%. Anyone with utilization below 80% is approved for credit, and those with utilization at or above 80% are declined.

In practice, underwriting models will use a tree with many more branches. This allows more nuanced consideration of data relevant to prediction of default risk, but there can be tradeoffs as complexity and depth increase. The depth of a tree refers to the number of splits a tree can make before predicting the outcome/target variable. A tree with more branches and greater depth has leaves that are more pure, which means that all the data points on the leaves are from the same class (for instance, a "pure" leaf would include all data points from either high risk or low risk of default). In general, the greater the number of branches in the model the greater the risk of overfitting, which can undercut the model's accuracy. Ideally, an increase in the number of branches should be supported by an increase in the training data so that the risk of overfitting can be reduced.

Decision trees can be prone to various problems. First, they may be highly dependent on specific attributes analyzed in training and therefore pose a risk of overfitting. If a model overfits and the training and deployment data differ significantly, the decision tree is unlikely to make accurate predictions. Second, decision trees may be sensitive and inconsistent because optimal decisions are made locally at each node. But a decision that is optimal from the perspective of a single node may not produce a tree that delivers coherent or stable predictions. Third, decision trees can be biased if some groups dominate in the sample. This is particularly true in credit underwriting, as underrepresented populations may be inaccurately assessed due to lack of credit history.

Random Forest

A random forest is an ensemble machine learning method in which multiple decision trees are combined into one predictive model to decrease variance and bias and/or to improve accuracy and predictions. There are two techniques that are used in random forests to make them less biased and more accurate than individual decision trees:

- **Bagging or Bootstrap Aggregation:** Bagging or bootstrap aggregation refers to the process of generating randomly drawn subsamples from a training dataset with replacement,[154] training individual decision tree models on each of these subsamples, and then calculating the average predictions from each model to yield the final prediction.
- **Decorrelation:** Decorrelation ensures that only a subset of features is chosen at random when the decision to split is made at each node. The advantage of this randomization is that each feature is used for modelling the outcome and a single dominant feature does not drive the results, which makes the trees "decorrelated," resulting in lower variance.[155] In

[154] Replacement is a statistical technique that is designed to improve the independence of randomly drawn samples. It means that when a sample A is drawn randomly from a population, A is put back/replaced in the population before generating another sample, B, from the population. Therefore, these two sample values, A and B, are independent and their covariance is zero.

[155] Lower variance means that a small change in the training data does not result in large changes to the model predictions.

terms of the number of features selected for decorrelation, generally a square root of the number of trees is used.[156] The number of features in a sample will determine how long this process takes.

Developers will also tune certain hyperparameters, such as the number of nodes allowed in the model,[157] to increase the accuracy of the resulting model. Although the tuning may generate high accuracy in a model, the tuning process requires long processing times and high computing power.

Because individual and particularly deep decision trees tend to overfit to training data, random forests can generally yield more accurate models. A random forest deals with these issues by training several small decision trees, and training each tree on a random subsample of the training data. Often these subsamples use only a small number of features, and the resulting trees are small; this prevents overfitting. Combining all of these simple trees into a decision forest allows the model to capture more nuanced effects in the data, while also preventing overfitting. Decorrelation of features helps to ensure that certain features do not dominate the results. This is critical in decision trees generally, but has special relevance in the context of credit underwriting since indicators of creditworthiness taken from historical data may disproportionately drive the probability of default if decorrelation is not used. In this context, a random forest model can assess the various features and subset features that may be more specific to certain classes/groups and ensure that the predictions of default probabilities are driven by all features in a dataset instead of the dominant feature(s). This way, a random forest may improve the inclusiveness of lending decisions, although more research is needed to explore this potential.

A random forest can also be more complicated and harder to explain than individual decision trees due to the higher number of features included when multiple trees are generated and the need to average over many trees. However, feature importance explainability techniques can be used to make random forest models more interpretable.

Gradient-Boosted Decision Trees

Gradient-boosted decision trees are another ensemble machine learning model that uses multiple decision trees. Gradient-boosted decision tree models estimate a first tree and then estimate a second tree based on the prediction error of the first tree as the target variable (unlike random forest models where prediction errors are not utilized to generate subsequent trees). Subsequent trees in gradient-boosted models are also built based on the prediction errors of the prior models.

Unlike a random forest, the final prediction of a gradient-boosted decision tree model is the weighted sum of predictions of all trees rather than the average. The weighting ensures that gradient-boosted decision trees lead to lower prediction error rates and better predictive power compared to a random forest or individual decision trees. Similar to a random forest, developers can calculate feature importance for gradient-boosted decision tree models, which makes it easier to understand the predictions and results generated using the models.

Extreme Gradient Boosting (XGBoost) is an open-source package available in Python and R that is becoming popular for the development of credit underwriting models and other financial services applications. XGBoost uses the gradient-boosting framework and optimizes tree-based

[156] *See* Bazarbash.

[157] For each type of machine learning model, there is a set of hyperparameters that are optimized in order to help find the minimum loss or maximum accuracy for the model. The hyperparameters can be optimized using various approaches, such as random search or grid search. It is important for underwriting model developers to tune and optimize model hyperparameters as they can help to determine more accurate loan default predictions.

methods using various techniques such as L1 and L2 regularization, which lead to better predictive performance and speed (see Section 5.3.1.2). Several enhancements to the algorithms and systems have improved the accuracy and operational efficiency of XGBoost models and made them popular for underwriting. XGBoost includes tree pruning, a process controlled through a hyperparameter to remove relatively irrelevant or unimportant information from the trees and manage risks related to overfitting. Further, XGBoost can recognize areas where data sparsity may affect the model's accuracy and handle missing data better by imputing values. In addition, XGBoost includes parallelization, which sorts the data in a way that uses CPU power more efficiently, which speeds up the training process.

Further extensions of gradient-boosted decision trees (GBDTs) are also used in credit underwriting. Stochastic gradient boosting improves upon gradient-boosted decision trees by using a bagging concept similar to what is used in random forests by drawing random subsamples to generate each new tree based on the prediction errors of the previous tree.[158] This makes stochastic gradient boosting similar to random forest models, unlike gradient-boosted decision trees where the entire training data is used for classification. Open-source packages, such as CatBoost, support gradient-boosted decision trees and focus on categorical features in the data.[159] In practice, categorical features are manually converted to numbers in gradient boosting. However, this algorithm deals with categorical features during training, which is more efficient. Another advantage of the algorithm is that it uses a new method for calculating leaf values when selecting the tree structure, which reduces overfitting. Both these methods result in the algorithm to outperform other gradient-boosted trees such as GBDTs and XGBoost.[160]

4.1.2.1.2 *Support Vector Machines*

Although tree-based models are more common in credit underwriting, support vector machines (SVMs) can also be used to predict probability of default. A support vector machine is a machine learning tool that generates a separating line between observations in a dataset that belong to different classes. Support vector machines are used for classification problems such as facial and handwriting recognition and underwriting.[161] In the case of credit underwriting, a support vector machine can be used to separate applicants based on predicted "default" or "not default" outcomes.

Figure 4.1.2.1.2 below shows a support vector machine for a credit underwriting model. The dark black line in the middle shows the best separating line to split the feature space. The best separating line or boundary is calculated by maximizing the distance between the line and closest points in each class, which is the *margin*. The blue lines are the support vectors and the distance between them is the margin. As the figure shows, the support vector machine has generated a line and separated the applicants, where red stars refer to applicants who yield "default" outcomes and blue circles are applicants who yield "not default" outcomes. It is also worth noting that the support vector machine produces some overlaps, and observations are on the other side of the separating line. Thus, while this particular support vector machine has separated the most homogenous observations on either side of the separating line, it has actually misclassified some individual observations.

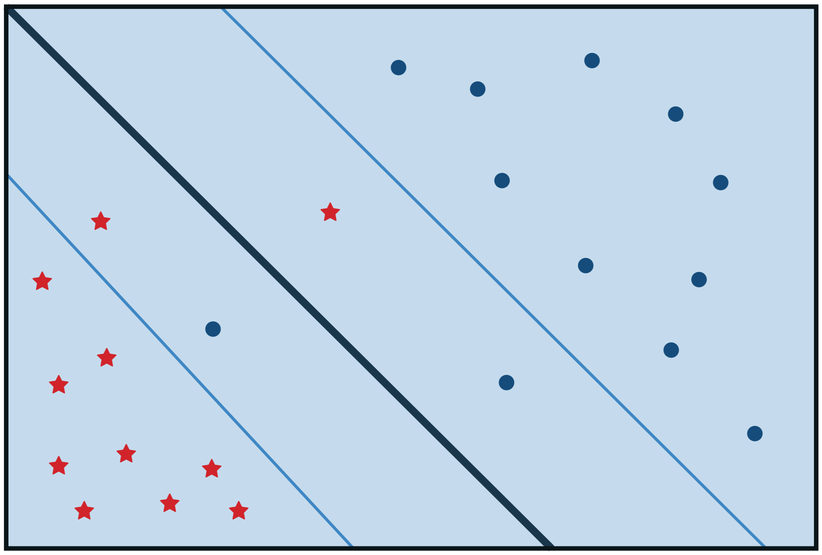
[158] Jerome H. Friedman, Stochastic Gradient Boosting, 38 Computational Statistics & Data Analysis 367-378 (2002).

[159] In the context of credit underwriting, an example of a categorical feature may be repayment status on a particular tradeline, which can be different categories, such as "paid on time", "delayed by one month", or "delayed by three months." These are not numerical features such as the number of past defaults or bank card balances, which mean that these categories need to be converted to numbers before they are used in a model.

[160] Anna Veronika Dorogush *et al.*, CatBoost: Gradient Boosting with Categorical Features Support, arXiv:1810.11363 (2018).

[161] Sheng-Tun Li *et al.,* The Evaluation of Consumer Loans Using Support Vector Machines, 30 Expert Systems with Applications 772-782 (2006).

FIGURE 4.1.2.1.2 SVM FOR A CREDIT UNDERWRITING MODEL



In terms of performance, research has shown that SVMs perform well on smaller datasets and are unlikely to overfit, as they use a subset of the training dataset to evaluate the separating line and support vectors.[162] This also means that they require less computational power. More recently, however, SVMs have gotten less attention in the context of credit underwriting due to the emergence of other methods that deliver a better balance of performance gains, interpretability, and operational efficiency.[163]

4.1.2.1.3 *Neural Networks*

Artificial neural networks can produce powerful predictions, as they learn non-linear relationships between features and the target variable through several inner layers.[164] The first layer consists of the features of the input data, which are used to generate latent features that make up the nodes in the second layer. The evaluation is conducted on a weighted sum of inputs and is based on an activation function, which combines several features into a single number (usually between 0 and 1). This process repeats until the final layer, where predictions for the target variable are generated. The layers between the first and final layers are often referred to as hidden layers. They are composed of latent features generated by the model which are used to predict the target variable. This structure can be particularly helpful to identify non-linear relationships between input features and target variable, which boost the predictiveness of the models compared to other machine learning techniques.

Neural networks have been deployed in various fields such as computer vision, speech recognition, and natural language processing, among others. In banks and fintechs, neural networks are already used extensively in fraud analytics, where the accuracy and higher predictiveness of these models allow financial institutions to better understand and detect fraud patterns in extremely large volumes of transaction data. Similarly, neural networks have potential to be very effective in credit underwriting, especially where a lender aims to use large-scale, diverse datasets for which neural

[162] Hafiz A. Alaka *et al.*, Systematic Review of Bankruptcy Prediction Models: Towards a Framework for Tool Selection, 94 Expert Systems with Applications 164-184 (2018); Li *et al.*

[163] R. Y. Goh & L. S. Lee, Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches, 2019 Advances in Operations Research art. 1974794 (2019).

[164] Neural networks can be used for supervised, unsupervised, and reinforcement learning. In this report, neural networks are primarily considered in applications using supervised learning given its prominence in underwriting.

networks' capacity to recognize complicated patterns is particularly valuable. The non-linear nature of these models may be particularly valuable after an economic shock like the onset of the pandemic because they are better able to identify and assess dynamic relationships between features and the target variable.

However, this increased predictiveness can come at some cost. In general, neural networks can be computationally taxing, although the operational constraints of running these models have eased. They can also be difficult to optimize and explain. For example, even a simple two-level neural network involving nine input features with five nodes in the first layer and five latent features in the second layer will require a total of 100 parameters to be estimated and then synthesized to explain the functioning of the model. In contrast, most regression equations using the same nine input features would require only 10 coefficients to explain the model.

Another potential disadvantage is that neural networks with more than a handful of layers can be difficult to understand, and this complexity is a particular challenge in the context of credit underwriting. Furthermore, these deep neural networks have a huge number of parameters and tend to overfit to the training data. Limiting the number of layers in the network can improve the model's transparency and prevent overfitting, but may come with equally significant performance tradeoffs. For example, a logistic regression can be represented as a neural network with a single layer of nodes—in other words, certain simple neural networks can be equivalent to logistic regression. Neural networks can also be made more interpretable by using a piecewise linear activation function,[165] such as Rectified Linear Unit (ReLU), that creates a neural network consisting of many locally-linear models that are each interpretable in the sense that the individual models use a linear combination of attributes to calculate an output. *Post hoc* explainability techniques may also make neural networks sufficiently transparent for use in contexts like credit underwriting. As discussed above in Section 3, while neural networks involve complicated algorithms, techniques such as integrated gradients and SHAP for deep learning models are designed to make these models more transparent and help users understand the relative importance of individual features to models' predictions.

4.1.2.2 Forms of Unsupervised Learning Used in Credit Underwriting

Although supervised learning models discussed above are dominant in credit underwriting, certain unsupervised learning techniques, such as cluster analysis, may also be used for other credit-related activities.

4.1.2.2.1 *Cluster Analysis (Segmentation)*

While this technique is not as commonly used as supervised learning models, unsupervised learning models can be used to generate clusters of borrowers, which may be useful to separate borrowers of different types. These may not be helpful to classify customers directly on their probability of default as part of an underwriting model, but can be used to segment potential applicants and existing borrowers for marketing purposes and for evaluating credit line amounts. For instance, the models can be structured to predict an individual's likeness to existing borrowers who did not default, were loyal, or met other criteria of value to the lender. In addition, cluster analysis is particularly useful when ground truth data about actual lending outcomes is not available, as unsupervised learning can be used on unlabeled data.

[165] Activation functions are functions that introduce non-linearity and help a neural network learn complex patterns in the data. The activation function transforms a neuron, which contains a set of inputs and associated weights. The output of the neuron is then sent as input to the neurons of another layer, which repeats the same process (weighted sum of the input and transformation with activation function) until the final layer which predicts the target variable.

K-Means Clustering

The most commonly used cluster analysis method is K-means clustering, which is very commonly used to segment customers in various markets, to determine customer loyalty, and to create targeted marketing and offers. The objective of this method is to derive *k* number of clusters from *n* observations. Each observation in the dataset is allocated to a cluster with the nearest mean or *centroid*, where the centroid refers to the mean value of all the data points in the cluster. Initially, a random number of *k* clusters is chosen, and their centroids are calculated. The data points are then assigned to the closest centroid, where the *closest* is calculated based on the distance between the data point and the cluster. Once the data points are assigned to a cluster, the cluster means are recalculated and every observation is checked to determine if it is closer to another cluster. This process continues until there is no change to the centroids and cluster assignments are no longer updating.[166] Once the k clusters have been fixed, new data points can be classified by assigning them to the nearest cluster.

For example, K-means clustering can be very effective in understanding credit card spending patterns of customers, which financial institutions can then use to define customer segments and devise marketing strategies. For instance, a bank can use K-means clustering to look at customers' credit card spending, such as amounts spent per month, individual purchase amounts, and where customers use their cards. The bank might use this analysis to offer higher credit lines to customers who regularly spend high amounts and also pay their balances on time or to offer cards providing rewards for spending at grocery stores to customers who would find them a good fit for their spending habits.

4.2 Data Selection and Preparation

The ability to assess large, diverse datasets is one significant motivation for lenders' interest in using machine learning underwriting models. In other sectors such as retail and media, the ability of AI systems to track and use broader forms of information, including data on individuals' online behavior, to substantially improve predictive power has transformed markets, business models, and consumer behavior.[167] In financial services, these changes have also been noticeable in marketing and customer engagement strategies and efforts to detect fraud and other illicit behavior.[168] However, lenders have generally been more reticent to expand underwriting data beyond credit bureau data and data about existing customers' past dealings with the lender due to regulatory concerns especially with respect to fair lending compliance, operational constraints in accessing alternative data, and securitization requirements that favor use of standardized data. Even if individual lenders decide to adopt machine learning underwriting models simply to better assess traditional data sources, the transition to machine learning underwriting models presents an opportunity to consider what data can and should be used for underwriting. Further, the choice of what data will be used to develop an underwriting model can have significant implications for the model's complexity as well as the accuracy, fairness, and inclusiveness of lending decisions.

[166] To verify that the clusters are correctly determined, the sum of the squared error (SSE) is determined after the centroids converge. The objective of K-means clustering is to minimize SSE, where SSE is defined as the sum of the square of the Euclidean distances of each point to its closest centroid.

[167] A 2020 survey of nearly 7,000 marketers globally reported a 186% increase in AI adoption since 2018 for marketing purposes such as customer personalization and data collection. Salesforce, State of Marketing Report 6, 18 (2020). *See also* Parrish, Alternative Data and Advanced Analytics.

[168] *See, e.g.,* Parrish, Impact Report (reporting that 55% of respondents employ AI in both the marketing and fraud detection stages of the loan life cycle, and 25% plan to do so in the future for both categories).

4.2.1 Data Selection

This section provides an overview of the different types of data available for use in credit underwriting and the form in which such data are delivered.

4.2.1.1 Type

The type of data used for credit underwriting plays a significant role in determining who is approved for credit.[169] This subsection provides an overview of the various types of data that can be used in credit underwriting and their implications for both predictiveness and inclusion (see Box 2.1).

» **Credit Information:** Traditionally, credit applications are approved largely based on credit histories and scores from major credit bureaus, such as Experian, Equifax, and TransUnion, and from smaller providers that focus on specialty finance records. The bureaus typically provide applicants' personal information; public records such as bankruptcies; tradeline data which reflect that person's repayment record mainly for secured and unsecured loans; inquiries made on the applicant's credit files; and balance information (including available balance for credit cards).[170] This type of data is usually standardized and relatively clean,[171] which reduces the risk of noise in predictions where the data are available. Credit bureau data are used in a variety of ways during originating and securitizing consumer loans, and are available in sufficient amounts to generate credit scores from the most widely used models for approximately 80% of adults in the United States.[172]

Some research suggests that machine learning models that evaluate only credit bureau information can improve credit risk prediction for applicants, including those who cannot be scored under some existing models.[173] However, other research has raised concerns about the potential predictiveness, inclusion, and fairness effects of relying solely on traditional data for credit risk assessment. For instance, the remaining 20% of consumers who cannot be scored using the most widely used models include disproportionate numbers of minorities, recent immigrants, and students.[174] One recent academic study found significantly more signal noise—that is, random, unpredictable errors that make it hard to reliably predict credit risk—in scores calculated based on traditional data for minority groups and consumers with marred credit records.[175] Another study found that machine learning models' ability to map more closely to traditional mortgage data sources could lead to marginal improvements

[169] *See* FinRegLab, The Use of Cash-Flow Data in Underwriting Credit: Empirical Research Findings (2019); Leonardo Gambacorta *et al.*, How Do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm, BIS Working Paper No. 834 (2019).

[170] *See* Carroll & Rehmani.

[171] A 2012 Federal Trade Commission study found that 21% of participants had errors in their credit reports, 13% had errors that affected their credit scores, and 5% were able to obtain corrections that were so large that they changed credit risk tiers. However, no comprehensive update has been performed since the Consumer Financial Protection Bureau began examining credit reporting agencies for compliance with relevant federal laws and other market developments occurred that may have affected particular sources of errors. *See* Cheryl R. Cooper & Darryl E. Getter, Consumer Credit Reporting, Credit Bureaus, Credit Scoring, and Related Policy Issues, Congressional Research Service (updated Oct. 15, 2020); Federal Trade Commission, Report to Congress under Section 319 of the Fair and Accurate Credit Transactions Act of 2003 i to vi, 57-64 (2012).

[172] Federal Deposit Insurance Corporation, 2017 National Survey of Unbanked and Underbanked Households at 10.

[173] VantageScore reports that its use of machine learning to develop scorecard models for consumers who are not scorable under some third-party models because their credit histories have not had an update in the prior six months resulted in a performance improvement of 16.6% for bank card originations and 12.5% improvement for auto loan originations. *See* VantageScore.

[174] *See* Blattner & Nelson.

[175] *Id.*; Wei Li *et al.,* The Lasting Impact of Foreclosures and Negative Public Records, Urban Institute Housing Policy Finance Center 9 (2016).

in approval rates but increase pricing disparities to the extent that certain populations are predicted to be somewhat higher risk than under conventional models.[176]

- **Alternative Financial Data:** Alternative financial data describes a variety of non-lending financial activities and can be extracted relatively easily from sources such as bank or prepaid accounts. Depending on the source and scope of data, this information may actually contain more granular and timely information about applicants' financial position than credit bureau information and can provide a more complete picture of an applicant's ability and willingness to repay a loan.[177] There is growing evidence that such data can be used to overcome the shortcomings of traditional credit information in providing credit to thin- or no-file applicants. Recent research shows that using alternative financial data such as cash-flow information from bank account records and other data sources can increase models' predictive power as well as improve access to credit for historically underserved groups.[178] In addition, information from rental, utility, and telecom records may also be useful for assessing creditworthiness, even among some populations that lack bank accounts or prepaid cards.[179]

- **Behavioral Insights in Alternative Financial Data:** Some sources of alternative financial data provide detailed information about consumer behavior, such as where and when they shop and in some circumstances what they buy. Some of this information may be relevant to credit risk assessment. For example, segregating transactions into discretionary purchases and tracking how an individual manages those against fluctuations in income may indicate how well an applicant manages financial decisions. Similarly, the time of day at which an applicant sought a loan or the temporal relationship between the application and obligations coming due may be predictive of credit risk. These kinds of behavioral insights embedded in transaction histories also raise fairness, fair lending, and privacy concerns, particularly if consumers are not aware that such information will affect underwriting decisions. For example, a lender could decide that an individual spends too much on lattes or bicycle jerseys given their income and assets and the potential obligation of loan payments. Similarly, lenders often decide that the channel by which an application is received—whether for example the application was received via the lender's app, in a branch, or through an aggregator's website—cannot be fairly used in underwriting even though it can be indicative of credit risk. These decisions reflect an understanding that emphasizing originations channels in credit risk assessment can be unfair since the lender has chosen to accept applications across various channels and may also introduce fair lending risk where the source of applications correlates with protected class characteristics.

- **Non-Financial Alternative Data:** Non-financial alternative data refers broadly to data about a person's activities that are not financial in nature or derived from financial data. Such data are mostly unstructured, as discussed further below. Social media data are one common form of non-financial data, but search histories, educational attainment, and mobile phone recharging habits are other examples of non-financial alternative data that

[176] *See* Fuster *et al.* (machine learning models using conventional data in the mortgage context concluded that such models would likely lead to modest improvements in application approvals among Black and Hispanic applicants, but would increase pricing differentials between different demographic groups due to many minority applicants being evaluated as higher risk than under conventional approaches).

[177] A 2010 study found that a machine learning model constructed using both credit bureau and transaction data from a large consumer bank improved the predictiveness of credit card delinquencies and defaults and would have resulted in the firm reducing losses by between 6% and 25% through adjusting credit lines based on the new model's predictions. *See* Khandani *et al.*

[178] FinRegLab, The Use of Cash-Flow Data in Underwriting Credit: Empirical Research Findings.

[179] Pew Research Center, Demographics of Mobile Device Ownership and Adoption in the United States (Apr. 7, 2021).

have been studied or considered for underwriting purposes in the U.S. or other countries.[180] Much of this data is behavioral and raise heightened concerns about reliability and fairness when incorporated in underwriting analyses to the extent that the data correlate but have no clear causal or intuitive links to creditworthiness. The introduction of this data also gives rise to data quality concerns if the data are available for only certain groups or if institutions lack experience working with new types of data.[181] Non-financial alternative credit data also often raise similar concerns about privacy, correlation with protected class, and other issues as articulated above in discussing behavioral data derived from alternative financial information.[182]

This type of data is not commonly used in the U.S. for underwriting, although some nonbank financial institutions consider educational factors and digital footprint information in addition to more traditional measures of creditworthiness. Outside the U.S., non-financial data are more commonly used for customers in rural areas or low-income populations who are unlikely to have previously taken loans. In such cases, digital footprint data such as browser used, calls made, and consideration of an applicant's social connectedness (such as number of connections an individual has on social media) are used in machine learning models to identify features that correlate strongly with lower probability of default. Researchers have found that this approach allows lenders to extend first-time credit to consumers who lack sufficient history to be evaluated using traditional credit information.[183]

4.2.1.2 Form

There are various forms of available data which differ in terms of the type, storage, and flexibility of access and use. The form of data has a significant impact on how often they are used and the purposes they are used for.

» **Structured Data:** Structured data refers to tabular data, which are stored in a database in columns and rows. These are typically the easiest to access and are most readily available to use. They are stored in relational databases and have relational keys, which make it easier to link the tables and combine or merge them as required. In credit underwriting, these can refer to credit report data, which usually are stored in a database within a company and then used for various purposes, and other forms of financial data, such as transaction account information in certain circumstances.[184] This type of data is most commonly available and used for approving credit applications. Systematically stored data are less likely to contain missing values, which enhances the reliability of structured data and improves the likely predictive accuracy of models trained on this data.

[180] Sumit Agarwal *et al.*, Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech, Indian School of Business (2021) (evaluating whether non-financial alternative data can improve financial inclusion, focusing in India); Asli Demirgüç-Kunt *et al.*, The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution, World Bank Group (2018) (report on the financial environment around the world, including the use of traditional and non-traditional data).

[181] *See generally* Ostmann & Dorobantu.

[182] The use of larger datasets that include alternative behavioral and non-financial data in credit underwriting may exacerbate issues related to data accuracy, representativeness, and bias more generally, in addition to the concerns outlined above. *See* Federal Trade Commission, Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues (2016).

[183] *See* Agarwal *et al.*; *see also* Tobias Berg *et al.*, On the Rise of the FinTechs: Credit Scoring Using Digital Footprints, 33 Rev. of Fin. Studies 2845–2897 (2020).

[184] Transaction account or cash-flow data may vary in format based on how it is obtained. A firm using transaction account data for accounts it holds will in all likelihood be able to access this data in structured formats. However, transaction account data acquired via screen scraping and APIs may require processing and structuring before delivery to the acquirer. FinRegLab, Cash-Flow Market Context & Policy Analysis at 46-49.

- **Semi-Structured Data:** Semi-structured data refer to data which are not stored in relational databases but have some structure which make them easier to process and use for analysis. This includes information obtained via screen scraping or otherwise extracted from webpages using Extensible Markup Language (XML)/Resource Description Format (RDF). Key-value stores in these documents give the data sufficient structure for use in various kinds of models, albeit with additional cleaning work prior to processing. Using semi-structured data to complement the information provided by structured data is a common approach. For example, lenders may combine cash-flow data scraped from banking platforms with credit bureau records in an underwriting analysis to obtain a more holistic view of the applicant's financial position.
- **Unstructured Data:** Unstructured data include information stored in text formats, audio files, video files, and images, which includes most social media data. As a result, these data are neither organized in any consistent way nor stored in a database until they are extracted from their native formats using text queries and natural language processing. As described above, non-financial data sources are often unstructured and are rarely used for credit underwriting in the United States, given serious reliability, privacy, and fairness concerns associated with these data. However, some sources of digital footprint data—such as the number of social media connections an applicant has, patterns within those networks of connections, number of phone calls made and received and the duration of the phone calls—have been used in other countries and have shown to be predictive for applicants who have thin credit histories.[185] Sentiment analysis, which is widely used in marketing and is of interest in growing digital debt collection efforts, relies on unstructured data such as social media activity.

4.2.2 Data Preparation

Preparing or cleaning data is a critical and time-consuming stage of developing a machine learning model. Choices that developers make in this stage can have broad effects on the performance, fairness, and inclusiveness of models. Decisions taken to clean data may also affect the reliability of information expressed by *post hoc* explainability techniques.

4.2.2.1 Missing Values

Deciding how to handle missing or unavailable features or explanatory variables for specific individuals in a dataset is among the most common tasks in data cleaning. One method of handling missing data is to impute the missing input with the mean or median value of that input across the entire dataset. However, in cases in which the actual value is substantially different than the mean or median, both the accuracy of the model's prediction and the accuracy of explanations about the model's performance will be affected. To the extent that the explanation provided in the context of an adverse action is inaccurate, for example, that can have a direct bearing on regulatory compliance matters. For example, suppose a loan applicant had a missing count of repaid loans in their credit bureau record, but the modelling method imputed the missing value with an average from the larger sample. If the model imputes one repaid loan, but an applicant actually had five repaid loans and the model learned that having one corresponded to higher credit risk, then the model may predict that this applicant poses higher than actual default risk based on the imputed missing value. This may also lead to generation of an inaccurate reason code on an adverse action notice. More generally, missing values of features can be indicative of certain characteristics pertaining

[185] *See* Agarwal *et al.*

to an individual. For instance, missing values for a variable that determines whether an individual owns or does not own a credit card can indicate a higher or lower risk profile. In this case, missing values can be directly used to assess the probability of default.

Machine learning models may be better able than logistic regression to assess datasets with missing values. Frequently, developers of logistic regression models have to choose between imputation methods that introduce both inaccuracy and regulatory risk and dropping a feature that is missing too often in a dataset, even if that makes the resulting model less predictive. However, XGBoost and other machine learning approaches can learn the risk pattern associated with missing values explicitly, so that they can make a risk assessment based on all the variables, including whether a particular value is missing or not. This may be a significant factor that makes machine learning models more accurate than incumbent models and be particularly beneficial for those deemed unscorable under common credit scoring approaches.

4.2.2.2 Coarse Classing

Model development teams frequently use sophisticated statistical methods, like coarse classing, to prepare data for use in developing underwriting models. Coarse classing is a method of data transformation where subcategories of a particular kind of variable are combined where they have similar probabilities of default. For example, the categorical variable “residential status” may include the following responses: “own,” “rent,” and “with family.” A method called weight of evidence or WoE can help lenders identify whether the groups associated with each possible response exhibit similar probabilities of default or number of defaults. Any subcategories showing similar probabilities of default can then be combined in the model, for instance by collapsing the categories into “own” and “not own” if the default risks between “rent” and “with family” prove to have similar default risk levels. Streamlining similar categories in this way reduces spurious correlations in the data, which can reduce noise in the data and improve the accuracy of predictions.

WoE can also convert continuous variables such as age into binned groupings. For example, credit scores can be separated into a set number of bins or groups (500 < credit score <= 600, 600 < credit score <= 700, and so on), and then the WoE for each of these groups can be calculated. Similar to the example of residential status, creating credit score bands that exhibit similar default characteristics can improve the model’s predictive accuracy.

4.3 Modelling Considerations

In addition to the data considerations discussed above, developers will typically consider the following issues when designing an underwriting model:

4.3.1 Reject Inference

A lender typically develops underwriting models using historical datasets that include information about individuals who have applied for credit and their performance in particular loans. But lenders are not always able to obtain performance data concerning applicants who they rejected or who declined their offers of credit, especially if those applicants obtained credit elsewhere. Lack of information about whether applicants predicted to default actually defaulted on loans can bias the model and make validation more difficult.

To address this issue, lenders may opt to model reject inference by credit bureau proxy or statistically imputing data on loan performance for rejected applicants had they been approved. The technique involves using data for approved applicants to impute predicted values on individuals who were denied credit or, in other words, determine if rejected applicants would have been likely to repay their loans based on data on individuals for whom the lender has default labels. These predicted values for rejected applicants are then added to historical information for approved applicants to train an underwriting model.

There are several reject inference methods available that lenders can use to address this source of bias:

- **Simple Augmentation:** Simple augmentation involves assigning rejected applicants with a cutoff, such that any individual below a threshold is in the "default" class and any individual above a threshold is in the "non-default" class. After the class is assigned to the rejected applicants, both the accepted and rejected samples are included in the final training data, which is then fitted to develop a model.
- **Fuzzy Parceling:** Fuzzy parceling is another method used for addressing bias from reject inference. Here, a logistic regression model is fitted using information on applicants who were approved. That model estimates the default probability for all applicants who were denied credit previously. Fuzzy parceling assumes that each rejected applicant has both labels $y = 1$ and $y = 0$ (or equivalently "non-default" and "default" classes), with weights given by the fitted model using only data on applicants who were accepted. Finally, a new weighted logistic regression model is developed using data on both accepted and rejected applicants,[186] which is the credit underwriting model used on future credit applicants.

4.3.2 Credit Scorecards

Credit or underwriting scorecards are widely used across asset classes and types of lenders to develop models for credit underwriting. In a credit scorecard, the model converts various characteristics of borrowers—such as debt-to-income ratios, utilization patterns, or default history—into points. These points are combined into a total score that rank orders an applicant's likelihood of default. In current practice, lenders often use a system of multiple scorecards—also known as segmented scorecards—within their underwriting processes, with each scorecard targeted to a distinct and unique subsegment of the population, such as thin- or no-file borrowers or those with marred credit histories. Segmented scorecards enable each scorecard to be tailored to the unique characteristics and risk patterns of specific subsegments of the population.

One way to build a scorecard is to bin features into different groups to distinguish between those who are likely to repay the loan ("goods") and those who are not ("bads") using the same weight of evidence described above in the discussion of coarse classing. WoE is calculated by taking the natural log of the distribution of "goods" divided by the distribution of "bads." For example, for a feature that is continuous, the data can be binned into 10 groups and the number of goods and bads are calculated within each group to calculate the WoE values for each bin. This process is repeated for all features and the WoE values are then fitted to a logistic regression model instead of the original features in the training data. Finally, the regression coefficients and the WoE values are multiplied to derive the points, which make up the total score for an applicant. A more advanced way to express the score is linear programming based optimization, which is utilized to assign weights to bins that optimize a specified objective function. Once this model is trained, it can be used on test and deployment sets to determine the total score and a consumer's likelihood of default risk and to decide whether to approve or reject individual applications for credit.

[186] Ha-Thu Nguyen, Reject Inference in Application Scorecards: Evidence from France, Economix Working Paper 2016-10 (2016).

The form of the scorecard formula is a generalized additive model which allows non-linear relationships and interactions to be accurately captured and results in highly interpretable models. It is also possible to put constraints on the score weights to restrict the score formula, for example, to follow a monotonic relationship between certain characteristics and the resulting score. These constraints allow domain knowledge to be imputed within the score weights assignment to ensure robust and palatable models are obtained when the score formula is produced.

One advantage of credit scorecards, especially when using machine learning to generate the interim analyses to which points are awarded, is that they are an approach to data transformation that reduces noise in the data and that offers built-in transparency about the basis for credit decisions. If an applicant for credit is unsuccessful, the scorecard approach makes relatively straightforward the process of identifying the categories for which his or her score was low relative to the maximum available and ranking each such category by its contribution to the aggregate score or by its distance to the mean.

5. FAIRNESS AND BIAS

News stories routinely recount that without thoughtful design and oversight, problematic biases can be built into machine learning systems and amplify the effects of discrimination in a range of everyday decisions and activities.[187] In the credit context, the transition to machine learning underwriting models has intensified attention on a range of questions related to fair lending risks and standards, similar to questions that have been raised about earlier generations of automated models. It has also intensified concern about machine learning's ability to replicate or even amplify historical biases in lending.[188] For example, if models are developed based on data that are biased because they are inaccurate, incomplete, or unrepresentative for certain groups, the models are likely to replicate or even amplify those biases particularly for machine learning models because of their sensitivity to training data. Such flaws can result in a model that makes predictions that inappropriately over- or underestimate default risk as to underrepresented groups or cannot make a prediction for certain populations altogether for lack of information.

Further, the identification and management of variables that may proxy for protected class status under both disparate treatment and disparate impact theories of discrimination can be significantly more complicated when lenders use machine learning underwriting models (see Section 2.3.2). Being able to identify and mitigate causes of discriminatory lending patterns requires a high degree of transparency into how the models work and make predictions. Particularly where models are more complex (see Section 3.2.2), lenders and regulators may need new tools and face new limitations in efforts to diagnose bias. Machine learning models may also effectively reverse-engineer protected class status from correlations in data, even though consideration of such status is prohibited.[189]

[187] *See, e.g.,* Steve Lohr, Facial Recognition Is Accurate, If You're a White Guy, N.Y. Times (Feb. 9, 2018) (citing a study of AI facial recognition systems that discovered error rates up to 35% higher for Black women compared to White men, largely due to the dominance of White males in the training dataset); Ed Yong, A Popular Algorithm Is No Better at Predicting Crimes than Random People, The Atlantic (Jan. 17, 2018) (reporting that the widely-used COMPAS tool, an algorithm used to predict violent crime, reproduced the bias against Black offenders); Starre Vartan, Racial Bias Found in a Major Health Care Risk Algorithm, Scientific American (Oct. 24, 2019) (describing a popular algorithm for determining medical need that relied on a faulty metric of healthcare spending, which differs significantly for Black populations, and therefore underestimated the needs of Black patients).

[188] Automated underwriting models are generally recognized as reducing the risk of disparate treatment because they decrease the role of personal interactions and decision-making in credit risk assessment and apply a consistent analysis across applicants based on relatively standardized information. *See, e.g.*, Board of Governors of the Federal Reserve System, Report to Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit at S-2, S-3 to S-4, O-4 to O-6, 32-49. However, over time stakeholders have become increasingly concerned about the ways that data gaps and other weaknesses continue to create systemic barriers for applicants of color and other disadvantaged populations. *See, e.g.*, Lisa Rice & Deidre Swesnik, Discriminatory Effects of Credit Scoring on Communities of Color, 46 Suffolk L. Rev. 935 (2013); National Consumer Law Center, Past Imperfect: How Credit Scores and Other Analytics "Bake In" and Perpetuate Past Discrimination (2016).

[189] For overviews of some of the issues raised by both data and machine learning models, *see* Evans; Federal Trade Commission, Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues 27-32; Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 Cal. L. Rev. 671-732 (2016); Talia Gillis, The Input Fallacy, Minn. L. Rev. (2021), forthcoming 2022; Talia Gillis & Jann Spiess, Big Data and Discrimination, 86 U. Chicago L. Rev. 459-487 (2019); Deborah Hellman, Measuring Algorithmic Fairness, 108 Va. L. Rev. 811 (2020).

The term bias can be used in a variety of ways. Statisticians and data scientists often use it to describe systematic differences between a model's predictions and actual outcomes when trying to understand and correct the causes of those deviations. Others use bias to refer to discrepancies in treatment of different demographic groups, especially for those groups which have been subject to discrimination or injustice of other forms. This kind of bias is the subject of fair lending requirements and other forms of anti-discrimination law and regulation. This section is primarily focused on causes of bias with respect to demographic groups, especially those which have been historically underserved by the financial system. However, as part of considering the causes and potential corrections for biases with respect to protected classes, this section does describe forms of statistical bias—such as representation or historical bias—that may result in fair lending or other discrimination problems.

This section addresses a range of issues related to bias and discrimination in the context of algorithmic lending, with a particular focus on how the data science community thinks about the sources of bias and techniques for measuring and mitigating bias issues in various machine learning contexts. Along with the options for enhancing model transparency and other critical model construction decisions as discussed above, these approaches to managing bias concerns may be relevant to financial services stakeholders as they consider how traditional approaches to managing compliance with anti-discrimination laws and broader notions of fairness may need to be adapted for machine learning underwriting models. The section begins by setting out the ways in which data and models can each be the source of bias and then considers options for measuring and reducing bias.

5.1 Sources of Bias

Various forms of bias—including but not limited to legally defined forms of discrimination such as disparate treatment and disparate impact—may be introduced into an underwriting model in a variety of ways throughout the process of designing, implementing, and using a machine learning model. Understanding the sources of bias can help lenders and policymakers identify and mitigate bias and discrimination in individual models and define appropriate safeguards throughout the model lifecycle.[190] However, there is no standardized taxonomy of statistical biases[191] and weaknesses in data, model design, and governance/personnel can create feedback loops and magnification effects that make it difficult and in some cases maybe impossible to pinpoint a single cause of bias or discrimination.[192] In practice, biases can be introduced at various points in developing and using machine learning models:

[190] Andrew Burt *et al.,* Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models, Immuta and Future of Privacy Forum (2018); Nicol Turner Lee *et al.,* Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms, Brookings Institute (2019).

[191] Barocas & Selbst at 677-693.

[192] *See, e.g.,* Betsy Anne Williams *et al.,* How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications, 8 J. of Information Policy 78-115 (2018); Aylin Caliskan *et al.,* Semantics Derived Automatically from Language Corpora Contain Human-Like Biases, 356 Science 183-186 (2017); *see also* Sara Hooker, Opinion, Moving Beyond "Algorithmic Bias Is a Data Problem", Patterns (Apr. 9, 2021) (arguing that the prevalent belief that a model only reflects existing bias in the dataset is misguided, and that model design choices can independently contribute to algorithmic bias).

FIGURE 5.1 BIAS IN THE DATA, ALGORITHM, AND USER INTERACTION FEEDBACK LOOP[193]

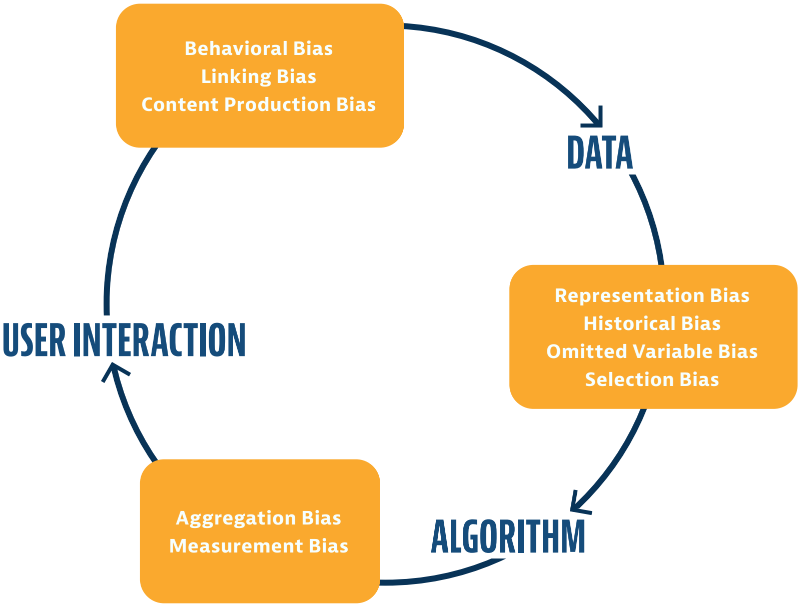


Figure 5.1 demonstrates how the cyclical nature of the model development processes can affect various types of statistical bias in machine learning models. For example, in the data selection and preparation phase of the cycle, the use of existing data may perpetuate practices or biases that existed historically (historical bias) or may produce models that do not generalize or predict well across all groups because they lacked sufficient information about particular subpopulations (representation bias). In the algorithm phase of the cycle, the training algorithm may learn biases during model development. These biases reflect the choice of the algorithms utilized in the model. Measurement bias is another type of bias that arises in the algorithm phase of the cycle when a mismeasured feature is utilized during model development, resulting in bias as to certain groups for whom the mismeasured feature is material to the model's prediction. In addition, there may also be user interaction effects where user behavior may differ in certain contexts or in use of various datasets.[194] Further, the feedback loop emphasizes that predictions produced by a particular generation of machine learning model can affect future data that are subsequently used for training subsequent generations of models.

Section 5.1.1 and Section 5.1.2 explore more deeply how bias can be introduced when models are developed and in data selection and preparation. Both sources of bias can be important for underwriting model development, although many discussions focus disproportionately on data-related concerns. Across both these sources of bias, the role of personnel and governance processes are important. Lack of representativeness among personnel who design, operate, and govern models can increase the likelihood of problems related to bias and discrimination in machine learning models and weaken organizations' ability to recognize and respond to problems in all phases of a model's development and use.[195]

[193] This figure was adapted from Ninareh Mehrabi *et al.*, A Survey on Bias and Fairness in Machine Learning, arXiv:1908.09635v2 (2019); Jongbin Jung *et al.*, Omitted and Included Variable Bias in Tests for Disparate Impact, arXiv:1809.05651v3 (2019).

[194] For more details, *see* Mehrabi *et al.* (highlighting user interactions with certain forms of statistical biases which exist in the model development cycle).

[195] Kenneth Holstein *et al.*, Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?, 2019 ACM CHI Conference on Human Factors in Computing Systems, arXiv:1812.05239v2 (2019); Turner Lee *et al.*; Judith Spitz, Why Tech Executives Must Embrace Diversity as Their First Line of Defense Against the Business Impacts of Algorithmic Bias, Forbes (Jul. 1, 2021) (suggesting three steps to combating algorithmic bias: establish diverse, cross-functional data inspection teams; establish diverse, cross-functional ethics and fairness review boards; and keep up with organizations at the forefront of AI ethics).

5.1.1 Data as a Source of Bias

Models use historical data of one kind or another to make predictions about a future event or behavior. If that data are unrepresentative, inaccurate, or contains mistakes,[196] the model's predictions will be less reliable. Biases originating in data affect regression models but their effects may be magnified in the context of machine learning underwriting models. As discussed in Section 4.2.1.1, in underwriting and credit scoring, data used to estimate which applicants are more likely to default are primarily derived from prior lending activity. As a result, data used to evaluate current applicants may not be able to assess with sufficient accuracy the credit risk posed by people who have not been able to obtain credit or have had to rely on products whose structure and terms increased their likelihood of default.[197]

The remainder of this section considers in greater depth the types of biases that can affect data used to develop underwriting models.[198]

Representation Bias: Representation bias occurs when defining and sampling a population to support development of a model. It reflects divergence in characteristics, behaviors, and outcomes for individuals in the dataset used to develop the model and the data that the model will encounter when in use. Under-representation in the training data can mean that the model's predictions do not generalize well once the model is in use. Its causes include sampling methods that only reach a narrow population (including past patterns described in the discussion of historical bias below) and changes between the population of interest and the overall sample that are not captured in data used for model development.

Historical Bias: Historical bias describes the effect that occurs when the data available from current or past practice is accurate and correctly sampled, but skewed in ways that means the model may produce outcomes that are not desirable from broader perspectives. For example, an algorithm designed to select which applicants for an engineering job merit interviews may successfully replicate the historical results from prior periods during which humans reviewed applications, but be nonetheless undesirable for institutions that want to include more women and minorities than they have historically.[199] In the context of lending decisions, historical bias addresses the potential for underwriting models to replicate past discrimination as to underrepresented groups.

Omitted Variable Bias: Omitted variable bias occurs when a model's target variable is affected by an explanatory variable that is not included in the model.[200] For example, a machine learning model can be built to determine whether a house will sell or not. In this model, several relevant variables can be included such as house price, number of bedrooms and bathrooms, and size measured in square feet. However, if the model fails to include a relevant variable such as whether other new houses are being built in the area, then the model may not be able to predict the likelihood of a sale accurately for all the houses. In this case, for example, the model may be able to predict if a house will sell in areas where no new houses are being built with high accuracy but have low accuracy for areas where new houses are being built.

Selection Bias: In credit underwriting, it is common to encounter selection bias, where outcomes for certain individuals are not available. For instance, when an applicant is rejected for a loan, it is not possible to know the actual outcomes for whether they would default, which means that the

[196] Barocas & Selbst at 684-687; Gillis at 17-22.

[197] Mark MacCarthy, Fairness in Algorithmic Decision-Making, Brookings Institute (2019); Schmidt & Stephens at 131, 143.

[198] Harini Suresh & John V. Guttag, A Framework for Understanding Sources of Harm Throughout the Machine Learning Lifecycle 6-7, arXiv:1901.10002v4 (2021).

[199] James Manyika *et al.*, What Do We Do About the Biases in AI?, Harvard Bus. Rev. (Oct. 25, 2019).

[200] Mehrabi *et al.* at 7.

data that the models are trained on comprise only applicants who are accepted. This is referred to as selection bias, as the models are biased and trained on individuals who are accepted for loans but are uninformed on rejected applicants. As discussed in Section 4.3.1, reject inference is a common technique to impute labels for the reject applicants, particularly in credit underwriting, which augment the data used for training, and can help to mitigate the selection bias issue.

5.1.2 Models as a Source of Bias

Even where data are accurate and complete, automated systems can reproduce past patterns of discrimination or introduce new forms of discrimination due to the way a model is designed, implemented, and used.[201] For example, models may be designed in ways that reflect assumptions about economic structures or business models with embedded inequalities. For example, an algorithm used to determine care for hospital patients systematically allocated Black patients less care than similarly situated White patients due to the assumption that annual accrued cost of care would be a good indicator of health needs.[202] Similar issues may result where algorithms are designed to optimize results across larger groups rather than distinct or differentiated subpopulations.[203] Decisions about optimization goals—such as overall predictiveness or privacy—can also have differential effects on certain groups. Further as discussed below in Section 5.2, decisions made about how to measure algorithmic fairness can also complicate these model design considerations.

The following list highlights key recent findings about the effect of certain model design and development decisions on the fairness of the resulting model, though some of them are focused on large neural networks deployed in non-financial settings that are designed to run quickly and to minimize energy consumption. More research is needed to determine whether the findings extend to the kinds of models most relevant to lending:

- **Training Rate:** Research suggests that observations in a training set that are difficult to learn—because they are underrepresented or present analytical complexity—are learned later in the model training process.[204] As a result, decisions to train models faster may affect how well the resulting model performs as to observations that are harder to learn. In the context of lending, this may mean that individuals who are hard to score or are generally less represented in historical lending data may be disproportionately affected by decisions about how much time is allotted for model training and the speed at which the algorithm analyzes training data.
- **Model Pruning:** In work focused on neural networks used in contexts like digital recommender systems, online advertising, and computer vision, efforts to design the models so that they run quickly and use as little energy as possible may disproportionately affect the accuracy of predictions with regard to groups that are underrepresented in the dataset. For example, pruning connections between nodes in a neural network by removing connections

[201] Schmidt & Stephens; Turner Lee *et al.*; BLDS, LLC *et al.*

[202] Heidi Ledford, Millions of Black People Affected by Racial Bias in Health-Care Algorithms, Nature (Oct. 24, 2019) (finding that risk scoring in software widely used to allocate care in U.S. hospitals consistently underestimated medical needs of Black patients when compared to equally sick white patients); *see also* Manyika *et al.* (considering algorithm used to select candidates for medical school interviews that was designed to replicate past practice with 90% accuracy).

[203] For instance, a number of advocates, policymakers, and other stakeholders have raised concerns that the use of educational information such as the school attended or the major or program of study in credit underwriting may exacerbate existing inequality in both educational and credit access for students and borrowers of color who attend or have attended minority serving institutions. *See* NAACP Legal Defense and Education Fund & Student Borrower Protection Center, LDF and Student Borrower Protection Center Announce Fair Lending Testing Agreement with Upstart Network (Dec. 2020); Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model: Initial Report of the Independent Monitor (2021).

[204] *See* Jiang *et al.*; Chirag Agarwal & Sara Hooker, Estimating Example Difficulty Using Variance of Gradients, arXiv:2008.11600v2 (2020).

with weights below a certain threshold may not affect overall model accuracy.[205] But recent research has shown that such pruning decisions can disproportionately affect underrepresented groups in the dataset.[206] More research is needed to extend this finding to the kinds of data and neural networks being used for lending, but it is possible that pruning neural networks may affect prediction accuracy more for minority borrowers than white ones or for thin or no-file applicants.

» **Privacy-Enhancing Technologies:** A model designed to deliver differential privacy[207] by preventing sensitive or "private" information from entering the training process can add statistical noise in ways that affect model accuracy.[208] In at least one study considering an algorithm designed for commercial lending, these effects were shown to disproportionately affect minorities.[209]

The following are types of statistical biases that affect how underwriting models learn and use training data:[210]

Aggregation Bias: Aggregation bias reflects the use of a generalized model for subpopulations that exhibit different distributions as to characteristics relevant to the model's prediction—such as debt-to-income ratio or credit line utilization in an underwriting model. This may result in a model that is more accurate in making predictions for a dominant population in the sample than it is for particular subgroups. For example, in the development of medications, testing results for women of child-bearing age may be unduly affected by other populations, since clinical trials tend to include fewer participants in that subpopulation.

Measurement Bias: Measurement bias arises when a variable in the model is mismeasured. For instance, if the measurement of aggregate income includes income from only one job and assumes that it is a full-time job, then the income of individuals who have more than one job will be mismeasured in ways that can bias the accuracy of the model's predictions for that group.[211] This may mean that the model leaves out important factors or that the selection or creation of features or labels introduces group- or input-dependent noise that affects model performance. It can be caused by measurement processes that vary among groups or by an oversimplified approach to defining the model's task.[212]

***

These biases can all result in faulty predictions and give rise to fair lending, discrimination, and inclusion issues in various circumstances. The shift to machine learning from incumbent underwriting

[205] A recent review paper finds that compressing deep neural networks via pruning often has very little impact on overall accuracy. *See* Davis Blalock *et al.,* What Is the State of Neural Network Pruning?, Proceedings of Machine Learning and Systems, arXiv:2003.03033 (2020).

[206] Two recent papers find that pruning can disproportionately impact underrepresented groups in the training data and can amplify existing biases. Sara Hooker *et al.,* What Do Compressed Deep Neural Networks Forget? arXiv:1911.05248v2 (2020); Sara Hooker *et al.*, Characterising Bias in Compressed Models, arXiv:2010.03058v2 (2020).

[207] Differential privacy is achieved where adding a single data point (or individual) to a dataset will not significantly change the output or reveal sensitive information about the individual. Techniques to increase differential privacy often add noise to a dataset.

[208] Ziheng Jiang *et al.*, Characterizing Structural Regularities of Labeled Data in Overparameterized Models, Proceedings of the 38th International Conference on Machine Learning, 139 Proceedings of Machine Learning Research, arXiv:2002.03206v3 (2021).

[209] Matthew Jagielski *et al.*, Differentially Private Fair Lending, Proceedings of the 36th Annual Conference on Machine Learning, 97 Proceedings of Machine Learning Research (2019).

[210] Suresh & Guttag at 5-6.

[211] Gillis at 20-22.

[212] For example, Street Bump, a smartphone app for Boston residents to report road issues, uses a data collection process that may reflect the uneven distribution of smartphone ownership across certain populations of the city rather than identifying the true geographical areas in need of road repairs, which could further disadvantage poorer, more marginalized communities. *See* Barocas & Selbst at 684-685.

models may amplify the importance of some of these risks, but it also presents an opportunity for practitioners and policymakers to rethink how underwriting models are developed and how new technologies and data can be used to help overcome, rather than further entrench, past patterns of bias and discrimination. The balance of the section discusses options for measuring and mitigating various forms of bias.

5.2 Measuring Fairness

How fairness is defined and measured is a threshold question that shapes efforts to identify and mitigate risks related to bias throughout model development and use. Regulatory oversight in financial services applies well-established definitions to assess fairness in the form of disparate treatment and disparate impact requirements (see Section 2.3.2). However, the broader community of machine learning researchers and practitioners have also devoted substantial resources to this question in recent years, producing more than 20 mathematical approaches to measuring the fairness of algorithmic models.

This work has attracted varying degrees of attention among financial services stakeholders due to several factors, including the sheer number of competing metrics, the technical nature of the source material, and the fact that there is little publicly-available research on how different measures might affect the fairness of underwriting models.[213] Nevertheless, industry and other lending stakeholders are drawing on this broader debate about defining and measuring the fairness of AI and machine learning models to consider two important questions:

- Can any of these metrics improve evaluations of compliance with traditional fair lending requirements?
- Do any of these alternative fairness metrics facilitate assessments of additional aspects of fairness that are not captured by current fair lending requirements?

Highlighting conceptual differences between the proposed definitions of fairness is a useful starting point to begin answering these questions. Toward that end, this section assesses a subset of metrics from the set of more than 20 possible approaches to measuring fairness that have generally garnered more attention in the academic literature and are more practical for use in the context of consumer credit given data available to lenders.[214] For each measure of fairness considered in depth here, this section describes the metric and its mathematical notation and assesses illustrative examples, data requirements, and tradeoffs related to using each measure.

Across the individual options, several broader considerations are likely to shape stakeholders' views about whether and how particular metrics could be useful. The first is that data availability may affect lenders' and regulators' ability to apply particular fairness tests at particular stages—for instance, during early stages of model development, while a model is operating in real time, or as part of periodic assessments of model performance. These metrics require some combination of the following primary forms of data:

[213] Most research on measuring the fairness of algorithmic models focuses on articulating definitions for these metrics. *See* Sahil Verma & Julia Rubin, Fairness Definitions Explained, FairWare'18: Proceedings of the IEEE/ACM International Workshop on Software Fairness 1-7 (2018). However, some more recent scholarship in data science and other fields has addressed the implications of proposed metrics in various contexts. *See* Hellman; Dana Pessach & Erez Shmueli, Algorithmic Fairness, arXiv:2001.09784v1 (2020).

[214] The chart provided in Appendix C provides an overview of more than 20 approaches identified in academic literature and other sources.

- **Model inputs:** The category model inputs refers to information about each applicant used by the model to make a prediction. In lending, this often includes loan-to-value ratios, credit scores, and data reflecting past credit utilization and repayment.
- **Model outputs:** The category model outputs refers to the predictions of a target variable returned by a particular model. The predictions are class labels for classification models (such as will default within six months) and estimates for regression models (such as the time before repayment is 65 days).
- **Protected class features:** Under current anti-discrimination requirements in lending, firms generally cannot consider an applicant's protected class characteristics when making a credit decision and are prohibited from collecting data about protected class status outside of mortgage lending, even though such information is critical to fairness evaluations.[215] This affects the quality and accuracy of traditional fair lending analysis as firms have to account for uncertainty associated with the protected class estimates and will have similar effects on efforts to use other proposed fairness metrics. For this reason, firms and regulators use statistical methods like Bayesian Improved Surname Geocoding (BISG) to impute an applicant's protected class characteristics in areas outside mortgage lending based on name, address, and other factors.[216]
- **Actual outcomes:** The category actual outcomes or ground truth refers to the observed values of the target variable (such as whether or not the borrower in fact defaulted within six months or the actual time before loan repayment). Lenders will typically have this data for every borrower who accepts their offer of credit and may be able to purchase such data about other consumers.

Inability to acquire or use all of these types of data naturally limits the utility of some of these alternative metrics for measuring fairness, especially in the context of credit underwriting. For example, lenders will have actual outcomes data only for a subset of their applicant pool—those to whom they offered credit and who accepted the offer—but not for those whom they denied credit or who declined an offer of credit. This may limit their ability to assess whether their models are making mistakes with respect to a set of denied applicants whose credit characteristics should have qualified them for a loan.

Further, even where data are available, some of the proposed measures may create tension with existing oversight structures. For instance, to the extent that achieving fairness under some of these definitions requires ensuring representation of various groups without regard to likely loan outcomes, those measures may prove impractical in light of safety and soundness requirements.

Moreover, even where data availability is not an obstacle, it is not generally possible to optimize model performance across all—or even several—of these definitions at once. Even if the metrics are used only in marginal ways to adjust models in development, improving fairness according to one measure may cause deterioration in others. Given this, commentators on a subset of the most widely discussed metrics—such as demographic or statistical parity, equal opportunity or predictive

[215] ECOA generally prohibits collection and use of information about an applicant's protected class characteristics outsides of residential mortgage lending (see Section 2.3.2 and Appendix B).

[216] For purposes of this section, the term protected class features contemplates use of either actual or imputed data. For the original study on BISG and applications of the method, *see* Marc N. Elliott *et al.*, A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity, 43 Health Services Research 1722-1736 (2008), Robert Letzler *et al.*, Knowing When to Quit: Default Choices, Demographics and Fraud, 127 Econ. J. 2617–2640 (2017), and Blattner & Nelson. For CFPB's exam manual and guidance, *see* Consumer Financial Protection Bureau, Supervision and Examination Manual (2020). *See also* Consumer Financial Protection Bureau, Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity (2014) (a CFPB white paper discussing the BISG method as a proxy methodology to impute information of race).

parity, and equalized odds which are discussed further below[217]—have set forth an "impossibility theorem of fairness." This suggests that although these metrics present their own set of advantages, they are mutually exclusive in normal circumstances—that is, it is not possible to satisfy all the conditions of the fairness metrics simultaneously and that building a model to be fair according to a particular definition of fairness will impose tradeoffs as to other conceptions of what it means to be fair.[218]

For these reasons, many of the proposed metrics may only be used currently in relatively narrow ways at selected stages of the development process, such as evaluating potential features in early model development stages, evaluating the effect of scores cutoffs or other similar rules, or assessing marketwide dynamics where it may be possible to assemble broadly based datasets that include consumer activities across financial institutions and over time. In practice, these metrics primarily function today as analytical tools that help firms gain insight into various aspects of a model's operations and effects in the iterative process of developing and reviewing models.[219] This means that efforts to optimize models using these metrics or simply understand what one metric might say about a model's fairness may be limited to early stages of development and occur separate and apart from traditional analyses to assess fair lending compliance prior to or after deployment.

In this context, these metrics can help lenders understand where disparities may occur and help identify ways to address them. But critically this implies that traditional governance processes remain important due to the deeply contextual nature of understanding what fairness means—that numbers generated by one metric or another do not simply automate decisions about compliance with anti-discrimination requirements.[220]

Further, these approaches to measuring fairness have not yet been adapted to measure compound or multiple protected class features.[221] For instance, they can be used to measure whether applicants who are Native American are treated fairly as compared to other race and ethnicity groups, or whether women are treated fairly as compared to men. But they cannot gauge whether Native American women are treated fairly as compared to other compound categories.

Finally, and most importantly, none of these metrics provide an entirely satisfying answer for how to deal with the existence of underlying factors that may drive differences in actual loan defaults (or other predicted outcomes) where those factors may themselves be the result of prior discrimination and historical bias. Some of the metrics focus on whether a model's predicted outcomes are consistent across different groups or individuals without regard to actual performance. But in lending, if optimizing models to be "fair" under such metrics leads to loans being made to applicants who will in fact not be able to repay them, both the applicants and the lenders will suffer as a result. In contrast, other metrics focus on whether a model is consistently accurate in predicting actual outcomes across

[217] Each of these metrics will be explored more fully in the following section (Section 5.2.1).

[218] For instance, where a model satisfies equal opportunity, it will generally not be possible to also satisfy the criteria for demographic parity and equalized odds. In the context of college applicants, for example, if the number of women and men who applied and are qualified differ, it is not possible to have equal opportunity (same percentages of qualified students admitted by gender), as well as demographic parity (same percentage of students admitted from both genders). Similarly, if equal opportunity is satisfied, it is not possible to satisfy the criteria for equalized odds that the percentage of unqualified students rejected by the college is the same across both genders. *See* Jon Kleinberg *et al.*, Inherent Trade-Offs in the Fair Determination of Risk Scores, arXiv:1609.05807v2 (2016); Alexandra Chouldechova, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, arXiv:1703.00056v1 (2017); Nengfeng Zhou *et al.*, Bias, Fairness, and Accountability with AI and ML Algorithms, arXiv:2105.06558v1 (2021).

[219] Although there are ongoing debates regarding these fairness measures among academics and practitioners, some lenders have recognized using these metrics during model development to understand better the fairness effects of different model specifications. *See* Upstart, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning at 17-19.

[220] Sandra Wachter *et al.*, Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI, Computer L. & Security Rev. (May 17, 2021).

[221] *Id.*

different groups.[222] But consistent accuracy is not sufficient to satisfy broader questions of fairness and equity if the factors driving disparities in actual outcomes are the results of prior discrimination and historical disparities in society and in lending. Such considerations further emphasize that notions of fairness are complex and often context-specific, and that no one mathematical formula can capture all relevant dimensions.[223]

5.2.1 Fairness Metrics

This section provides more in-depth consideration of a set of statistical and similarity-based measures of fairness that has been the focus of the most academic attention.[224] Some measures—like the first four measures presented below—can be computed after a prediction is made to assess disparities in the predictions of default for different groups, either in general, after taking into account certain critical factors, or focused more specifically on disparities in the accuracy of the predictions. Other measures, such as counterfactual fairness and fairness through awareness, make certain mathematical or data adjustments to achieve particular notions of fairness. Section 5.3 discusses additional methods of debasing data and models that can be applied using several of the fairness metrics that are described in this section.

This section makes two simplifying assumptions. First, the text uses a binary category—male and female—to illustrate groups that can be the subject of a fairness evaluation. Second, each metric is illustrated in the context of sorting good avocados from bad ones to help readers grasp how each metric works and how these measures compare to each other in an intuitive context, in addition to discussing potential applications in lending.

Simple examples are used throughout this section to illustrate each of the metrics. The examples start by imagining a worker in a factory processes avocados for resale to grocery stores. The avocados come from two countries—Alpha and Beta—and are processed by a worker who separates avocados that can be resold from those that cannot. If the worker determines an avocado is likely to be good when the grocery store sells them to a customer, then it will be selected for resale. If the worker determines an avocado is likely to be bad by that point in time, then the avocado will be thrown out.

The following table summarizes details used in several examples of various approaches to evaluating algorithmic fairness:

ALPHA	GOOD AVOCADOS	BAD AVOCADOS	TOTAL
Accepted	30	6	36
Rejected	15	9	24
Total	45	15	60

[222] In statistics and data science, accuracy has a technical definition: (TP [true positive] + TN [true negative]) / (TP + TN + FP [false positive] + FN [false negative]), which refers to the number of true cases out of all cases examined. In this section, "accurate" is referring more broadly to correct predictions of outcomes.

[223] *See* Kristine Gloria, Power and Progress in Algorithmic Bias, Aspen Institute 5 (2021) (discusses how situations in which a certain fairness metric is achieved still may not result in "fair" outcomes and suggests the objective of fairness in lending may need to be reframed).

[224] This section provides definitions for eight metrics of fairness selected for their relevance to credit underwriting. In the literature, there are additional metrics for measuring fairness which are not covered in this report. The additional metrics include false positive error rate balance, conditional use accuracy equality, overall accuracy equality, treatment equality, test-fairness, balance for positive class, balance for negative class, no unresolved discrimination, no proxy discrimination, fair inference, marginal effects, preferred treatment, preferred impact and equal accuracy. *See* Pessach & Shmueli; Verma & Rubin at 1-7. Appendix C provides an overview of metrics not covered in depth in this section.

BETA	GOOD AVOCADOS	BAD AVOCADOS	TOTAL
Accepted	20	4	24
Rejected	8	8	16
Total	28	12	40

In the context of machine learning terminology, the good and bad avocados are the ground-truth labels and the accepted/rejected categories/classifications are the predictions made by the worker based on his or her assessment that a particular avocado will be good at the time a grocery store sells it to a customer.

5.2.1.1 Statistical Measures

Statistical measures of fairness divide individuals into protected classes, such as race or gender, and compare some statistical measure, such as predicted ripeness or default levels, across those groups. For this reason, these definitions are said to measure group fairness, focusing on whether different groups are treated equally, instead of trying to assess fairness at the individual or application level. Depending on how models and/or decision-making processes are optimized for fairness under these definitions, it is worth emphasizing that the models may be made fairer on average but still result in individual decisions that are inaccurate.

Demographic or Statistical Parity

Metric Description: Demographic or statistical parity evaluates fairness based on whether the predicted outcomes are the same across particular subpopulations of interest. In underwriting, for example, demographic parity metrics would assess whether a particular model predicts the same risk of default across different protected class groups, for example men and women.

Illustration: To satisfy demographic parity, the total percentage of avocados considered acceptable to be sold at grocery stores from Alpha has to be equal to the total percentage of avocados considered acceptable to be sold at grocery stores from Beta. For example, if the worker determines the same percentage of avocados from both Alpha and Beta—whether that percentage is high or low—are acceptable for resale, then the worker's predictions satisfy demographic parity. The chart above satisfies this metric because 60% of each country's avocados are accepted for resale (regardless of whether they are actually good or bad). The metric does not look at how closely the predictions are tied to actual outcomes, but rather simply whether the model produces the same ratio of good to bad predictions for each group of avocados.

Analysis: This approach is used in traditional fair lending analysis and is supported by legal and regulatory precedent. Proponents argue that optimizing for this measure leads to higher inclusion effects and an increase in the number of people from protected groups. However, there are several criticisms for using this method as the sole or primary means of measuring fairness. Depending on the measures taken to maximize statistical parity, a model may not treat all individuals consistently. In the context of credit underwriting, using a model that ensures full demographic or statistical parity could lead lenders to reject individual applicants with low risks of default or to accept individual applicants with high risks of default to ensure that strict demographic or statistical parity with other groups is achieved. Such a result would lead to inconsistent treatment at the individual level and could pose risk of consumer harm through disparate treatments of the better-performing group and raise serious regulatory questions in any practical setting. Further, if a lender were to approve credit for individuals who are deemed high risk to satisfy demographic parity, there is a potential for adverse outcomes both for applicants who are given loans under that rationale and for lenders who suffer losses if such applicants default.

Data Requirements: To compute demographic or statistical parity, protected class features and model outputs are required.

Mathematical Notation:

$$P(d = 1 \mid G = m) = P(d = 1 \mid G = f),$$

where P refers to the probability, d refers to the predicted decision (for approval of credit), and G refers to gender, which can be either m (male) or f (female).

Conditional Statistical Parity

Metric Description: Conditional statistical parity measures whether the likelihood of a predicted positive outcome is the same across subgroups, once a set of control variables has been accounted for. The variables are typically chosen because they have a close link to the outcome being predicted by the model, and thus increase the accuracy of its predictions. Such variables can also often enhance the perception of fairness for different groups, because they facilitate more nuanced comparisons and consistent treatment once these key factors have been accounted for.

Illustration: If the worker sorting avocados systematically controls for factors such as size in determining whether an avocado is likely to be good or bad at the point of resale by a grocery store because small avocados ripen faster, then the ratio of good and bad avocados from Alpha and Beta can be considered an example of conditional statistical parity.

Analysis: The addition of well-chosen legitimate control factors can help to reduce the criticisms of general statistical parity because accounting for intuitive control inputs to the prediction focuses the analysis on whether groups that are consistent as to those inputs are treated consistently.[225] In credit underwriting, for example, such an analysis would consider whether there are disparities in treatment among demographic groups *after* factoring in whether applicants have sufficient income after expenses to cover the loan payments. However, the metric does not include parameters that consider how the key factors or the model as a whole performs *in practice*, and thus cannot detect whether the model may be doing a substantially better job predicting particular outcomes with regard to one group as compared to others.

Data Requirements: Computing conditional statistical parity requires similar data to statistical parity, along with additional data for the designated control factors and protected class features.

Mathematical Notation:

$$P(d = 1 \mid X = x, G = m) = P(d = 1 \mid X = x, G = f),$$

where the notations are similar to those for statistical parity. Specific to conditional statistical parity, X refers to a set of control variables, and x refers to a specific control variable or set of such variables.

[225] The use of conditional statistical parity for assessing fairness has also been studied in other parts of the world. One such study under applicable European Union requirements concluded that conditional statistical parity may close an accountability gap in anti-discrimnation oversight, but that determinations regarding fairness cannot simply be automated. *See* Wachter *et al* (2021).

Predictive Parity

Metric Description: Predictive parity focuses on the group of positive predictions generated by a particular model to determine whether there are disparities in the level of prediction accuracy across groups. Specifically, it evaluates whether various groups within a population have equal positive predictive value (PPV), where PPV refers to the number of true positives out of the total number of positive cases: PPV = True Positives/(True Positives + False Positives). In the context of lending, a "positive" label occurs when a borrower does not default on the loan. In this case, a "true positive" occurs when the model predicts that the borrower is unlikely to default, and the borrower in fact meets the obligations of their loan. On the other hand, a "false positive" occurs if the model predicts low risk of default, but the borrower actually defaults on the loan.

Illustration: In the avocado example, predictive parity is achieved if the percentage of actually good avocados predicted to be good avocados from Alpha is equal to the percentage of actually good avocados predicted to be good avocados from Beta. The worker predicts that 36 avocados from Alpha are good, of which 30 actually turn out to be good when the grocery stores sell to consumers. The worker predicts that 24 avocados from Beta are good, of which 20 are actually good. In this case, PPV is 5/6 for both Alpha and Beta, so this example satisfies predictive parity.

Analysis: Predictive parity assesses whether a model is consistently accurate in predicting true positive outcomes across different groups rather than whether it simply yields the same predictions across groups, which may address some of the criticisms of both demographic or statistical parity and conditional statistical parity. Using this measure, fairness is achieved when the probability of an applicant with a low predicted risk of default actually having a low risk of default is equal between female and male loan applicants. Thus, fairness achieved through predictive parity can be perceived to be more consistently accurate compared to demographic parity because achieving this metric ensures that if an individual is predicted to be creditworthy, then there is an equal chance of the individual to actually be creditworthy, no matter the gender of the individual.

Data Requirements: Evaluating predictive parity requires use of protected class features, model outputs, and outcome data.

Mathematical Notation:

$$P(Y = 1 \mid d = 1, G = m) = P(Y = 1 \mid d = 1, G = f),$$

where the notations are similar to those in previous definitions of fairness and Y refers to the actual classification result of an applicant.

Equalized Odds

Metric Description: Equalized odds evaluates whether groups within a population have both equal true positive rates and equal false positive rates. Fairness is achieved according to equalized odds where the probability of an applicant with an actual low risk of default to be correctly assigned a low predicted risk of default and the probability of an applicant with an actual high risk of default to be incorrectly assigned a low predicted risk of default are the same for both female and male applicants.

Illustration: In the example of avocado sorting, achieving this metric would require that both the probability that a good avocado is correctly accepted from both Alpha and Beta would have to be equal and that the probability that a bad avocado is incorrectly accepted for resale from both Alpha and Beta is equal. Using the numerical example in previous sections, the number of bad avocados incorrectly accepted as a percentage of the total number of bad avocados from Alpha and Beta respectively is not the same: 6/15 = 40% and 4/12 = 33%. In addition, the number of good avocados correctly accepted as a percentage of the total number of good avocados is not the same: 30/45 = 67% and 20/28 = 71%. This shows that in the avocado example, equalized odds is not satisfied.

It is worth noting that, in this example, both demographic parity and predictive parity are satisfied. However, equalized odds is not satisfied. In fact, except in special circumstances, there are actually no circumstances or combinations of numbers where all three statistical measures can be satisfied at the same time, which is known as the "impossibility theorem of fairness."[226]

Analysis: Proponents advance equalized odds as an appropriate measure of fairness because it ensures equal levels of accuracy across different groups. Similar to predictive parity, however, the conditions for equalized odds may be satisfied without necessarily addressing broader, societal causes of inequality in key lending variables like income and assets. In addition, similar to demographic parity and equal opportunity, fairness through equalized odds is unable to account for instances when several protected classes need to be integrated to measure fairness. However, equalized odds can be further extended to error rate parity (ERP), which states that the ratio of false negative rates (equivalently true positive rates) and false positive rates should be equal between groups. It is argued that a lack of ERP is indicative of unfairness and lack of parity in the presence of historically disadvantaged groups. In such cases, ERP can be utilized to understand if historical injustice is being perpetuated.[227]

Data Requirements: To compute equalized odds, protected class features, model outputs, and outcome data are required.

Mathematical Notation:

$$P(d = 1 \mid Y = i, G = m) = P(d = 1 \mid Y = i, G = f), i \in 0, 1,$$

where the notations are similar to previous definitions. Specific to equalized odds, **i** refers to the probability being the same for female and male applicants.

Counterfactual Fairness

Metric Description: Counterfactual fairness evaluates whether a predicted decision is the same for an individual in the actual world as well as a counterfactual world where the individual belongs to a different protected class, and applies an adjustment to generate the same prediction as the counterfactual. Unlike previous measures, counterfactual fairness does not entail a *post hoc* analysis of the applicant outcomes by group but instead involves adjusting the data to make them fairer across subgroups. For example, assume a predicted decision **d** is dependent on credit history and salary. The salary for an individual is directly correlated with a certain protected class feature, such as gender, which makes it counterfactually unfair. To be counterfactually fair, a woman's salary needs to be adjusted upwards with a counterfactual value, which will neutralize or mask the effect of gender on the model's decision and make it relatively more likely for the woman to be approved for credit. This is designed to ensure that decisions taken for individuals of a protected class, whether in credit approvals or college applications, are fair in that they do not result from differences derived from or associated with protected class characteristics.

Illustration: Using the avocado sorting example, assume that Alpha has been growing avocados for a long time and that Beta has only recently started to grow avocados. The farmers in Beta have less knowledge than those in Alpha of the type of seeds, pesticides, or methods that are needed to

[226] Zhou *et al.*

[227] *See* Hellman at 811.

produce a crop with a high percentage of good avocados. To be counterfactually fair to the farmers in Beta, the data on avocados is run through a model at the factory, which adjusts the data on the mix of agricultural inputs to be more counterfactually fair to Beta.[228]

Analysis: Adjusting the data to achieve counterfactual fairness may make the predictions more equal across particular groups, but it does not solve the underlying disparities that may have produced different outcomes in the first place.[229] For instance, if there are real disparities in income that make it less likely for women to actually be able to repay loans, simply adjusting the data does not fix the risk of a bad outcome for the consumer as well as the lender. This approach is also subject to implementation challenges due to the difficulty of quantifying the effects of certain characteristics on model predictions accurately and consistently. Finally, applying a counterfactual value to one variable may not account for interactions between variables or how other variables correlate with protected class features.

Data Requirements: To compute counterfactual fairness, protected class features, model outputs and the causal model are required.

Mathematical Notation:

A prediction $\mathbf{d}$ is counterfactually fair, given $\mathbf{X = x}$ and $\mathbf{G = g}$, for all $\mathbf{i}$ and $\mathbf{g = g'}$,
$\mathbf{iff}\ \mathbf{P\{d_{G=g} = i \mid X = x, G = g\} = P\{d_{G=g'} = i \mid X = x, G = g\}}$,

> where $\mathbf{d_{G=g}}$ is interpreted as the outcome of predictor $\mathbf{d}$ if $\mathbf{G}$ had taken value $\mathbf{g}$ and $\mathbf{d_{G=g'}}$ is interpreted as the outcome of predictor $\mathbf{d}$ if $\mathbf{G}$ had taken value $\mathbf{g'}$, which is a counterfactual value and different from $\mathbf{g}$.

Calibration

Metric Description: Fairness through calibration looks not just at different groups as a whole, but rather breaks them into segments to assess whether there are differences in predicted outcomes within each segment. Specifically, it measures the extent to which, for a predicted probability score, $\mathbf{Z}$, observations in various groups have equal probability, $\mathbf{S}$, of belonging to the positive class. In the context of credit underwriting, positive class refers to being approved for credit, whereas in the avocado sorting example, positive class refers to being accepted and passed on to be sold at grocery stores. Therefore, in the context of credit underwriting, calibration aims to achieve fairness by assigning for any predicted probability score, equal probability for both groups, for example, female and male applicants, to be approved for credit.

Illustration: To explain the fairness through calibration metrics, assume that the factory is receiving a large number of avocados and the sorting is automated. The factory workers now put the avocados on a conveyor belt and a scanner with a machine learning algorithm takes an image of an avocado and predicts the probability that an avocado will be good or bad when sold by grocery stores. To achieve fairness as calibration, if the predicted probability score is $\mathbf{z}$, then the probability of both avocados grown in Alpha and Beta and brought to the factory to belong to the positive class is $\mathbf{z}$. This can be shown in a table as:

[228] Another method would be to adjust the actual mix of agricultural inputs, for instance if Alpha shares the inputs with Beta so that any inherent differences between the avocados grown into the two countries will be eliminated.

[229] *See generally* Lily Hu & Issa Kohler-Hausmann, What's Sex Got to Do with Fair Machine Learning?, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 513 (2020).

TABLE 1 CALIBRATION SCORES

Z	0.2	0.4	0.6	0.8	1.0
$P(Y = 1 \vert Z = z, C = A)$	0.2	0.4	0.6	0.8	1.0
$P(Y = 1 \vert Z = z, C = B)$	0.2	0.4	0.6	0.8	1.0

where the predicted score $\mathbf{z}$ is calculated for all avocados on the conveyor belt and the results are binned in 5 bins, from 0.0 to 1.0. For instance, if Table 1 was an example of calibration scores for avocado sorting, it shows that avocados from both Alpha and Beta with a high prediction of being bad have an equivalent probability to actually be bad. Conversely, avocados from both Alpha and Beta with a high prediction of being good have an equivalent probability to actually be good. With the calibration scores presented in Table 1, the calibration classifier for measuring fairness is satisfied.

Analysis: For the stylized scenario in Table 1, as the predictions for the two groups are equal to actual outcomes, it may lead to higher inclusion as neither group experiences any bias. For example, when credit scoring female and male applicants,[230] it is more likely that a male applicant with a bad credit score (low values of $\mathbf{z}$) will actually have a low risk of default than a female applicant with a bad score, but that applicants of both genders with a good predicted credit score (high values of $\mathbf{z}$) have an equivalent chance to have a low risk of default. For Table 1, this means that the values in columns 1 and 2 in row 1 are higher compared to those in row 2, but they equate at higher values of $\mathbf{s}$ such as 0.8 and above.[231]

Data Requirements: To compute fairness through calibration, protected class features, model outputs, and outcome data are required.

Mathematical Notation:

Fairness through calibration can be denoted as: $P(Y = 1|Z = z, C = A) = P(Y = 1|Z = z, C = B)$.

5.2.1.2 Similarity-Based Measures

Similarity-based fairness measures evaluate whether similar individuals are treated similarly and define parity as between similar individuals based on their observed features instead of assessing fairness across groups.[232] Thus, while some of the metrics are conceptually similar to the group-based metrics described above, these definitions assess whether a model generates similar outcomes for individuals with similar characteristics.

Fairness through Unawareness

Metric Description: Fairness through unawareness is a baseline assumption used as a foil for evaluating other fairness methods. It assumes that fairness is attained as long as attributes as to which fairness requires neutrality—such as race or gender in the context of lending—are not included in the training and deployment datasets for classification.

Illustration: In the case of automated avocado sorting, fairness through unawareness is achieved if the machine learning algorithm does not incorporate the country of origin of the avocado into the prediction. As this sorting machine takes an image of an avocado to make a prediction, it is not

[230] In Table 1, male applicants can be assumed to be C = A, whereas female applicants can be assumed to be C = B.

[231] *See* Verma & Rubin.

[232] Cynthia Dwork *et al.*, Fairness Through Awareness, arXiv:1104.3913v2 (2011).

taking into account the country of origin, which means that for this case, fairness through unawareness is achieved.

Analysis: In practice, barring protected class characteristics from being included in the data used to train or operate a model may not be sufficient to lower bias or increase fairness in credit approvals due to the presence of model inputs that correlate with the information held out of consideration to achieve fairness through unawareness. This suggests that fairness through unawareness is unsuitable when protected attributes can be inferred from non-protected attributes.[233] In practice, this means that fairness through unawareness may provide a degree of formal confidence in delivering fair results without addressing complexities related to implicit sources of bias, including proxies for protected class characteristics that are found in most datasets. In lending, prohibitions on the use of protected class characteristics in making credit decisions has shifted attention to being able to detect forms of discrimination that result from data with strong enough relationships to prohibited bases to be considered a direct proxy for that information and from data or relationships that result in a model's predictions disfavoring groups on the basis of a prohibited characteristic.

Data Requirements: Fairness through unawareness does not require any additional data.

Mathematical Notation:

When assessing fairness through unawareness, the risk of default is predicted by $P(d = i \mid X = x)$,

where X is the set of features *excluding* G (gender), which is the protected class attribute.

Fairness through Awareness

Metric Description: Fairness through awareness states that fairness is attained if individuals who are similar with respect to various characteristics receive similar classifications with respect to the classification task at hand, irrespective of their protected class features. A similarity or distance measure is used to identify similarities between individuals along various characteristics and to assess individual-level fairness in classification tasks.

Illustration: In the example of automated avocado sorting, fairness through awareness is achieved if the machine learning algorithm defines a metric which calculates the similarities between the avocados and identifies the avocados that will be resold based on that metric. In this approach, two similar avocados will receive similar classifications by the machine learning algorithm in terms of whether they are accepted for resale or rejected, irrespective of what country they are from. For example, two avocados of the same size will be similarly classified.

Analysis: In theory, this approach can address concerns with "fairness through unawareness" measure and is potentially relevant to fair lending considerations if incorporating protected characteristics into the distance metric improves the context in which subpopulations are assessed. However, while similarity has been considered in the context of achieving "fair affirmative action" in hiring, there is no clear and objective way to calculate "similarity" through the distance metric that can be readily applied to other fields, such as credit underwriting. Moreover, similarity in this context cannot be related to protected class data. Defining a similarity metric that does not reflect correlations with protected class characteristics may limit severely the practical application of this approach.

Data Requirements: To measure fairness through awareness, model outputs and a similarity measure are required.

[233] Pratik Gajane & Mykola Pechenizkiy, On Formalizing Fairness in Prediction with Machine Learning, arXiv:1710.03184v3 (2018).

Mathematical Notation:

The similarity of individuals is defined by a distance metric, where the distance between the distributions of outputs for individuals should be at most the distance between the individuals. Formally, fairness can be achieved if $K(d(m), d(n)) <= k(m,n)$, where k refers to a distance metric between individuals, and is mapped for a set of applicants v to probability distributions over outcomes $d : v > \delta Z$, and a distance metric K between distribution of outputs.

5.3 Model Debiasing

In response to concerns about bias and discrimination across a variety of sectors and use cases, data scientists have produced a variety of methods for debiasing machine learning models that can be used at several different stages of the model development process. How to deploy these debiasing techniques—which techniques to use and how to pair them—depends primarily on the model's use case, as well as the type of data and model being used, the complexity of the model, and whether the relevant objective for debiasing is a particular measure of fairness, accuracy, or parity as discussed in the previous section.

In the context of developing a machine learning underwriting model, established business and risk management practices typically require sustained attention and oversight to monitoring and mitigating biases against protected classes throughout the process of developing, implementing, and operating the model. However, in practice, lenders face uncertainty when considering whether and how to use methods described in this section. Some methods require use of protected class information in ways that create tension with existing anti-discrimination laws, which has been significant enough to chill substantial adoption of these methods absent clarification from regulators. Other methods may undercut established risk management expectations. For example, banks' fair lending compliance teams are generally expected to conduct independent evaluations of lending decisions using real or imputed protected class information that is not available to model development teams. There is uncertainty about whether making this information available to model developers for the purpose of model debiasing is a practice that could potentially subject firms to regulatory criticism for compliance risk management weaknesses in addition to creating exposure to disparate treatment claims.

Further, some key approaches for debiasing real-world datasets, such as resampling or reweighting the distribution in training data, may be infeasible on a practical level given their cost and the need to know *ex ante* which sensitive features are driving the disparate outcome and to have comprehensive labels for both protected class features and proxy variables.[234]

This section provides an overview of debiasing options throughout the model development lifecycle.

5.3.1 Debiasing Approaches

Recent data science advances in debiasing fall roughly into three categories based on when they occur in the model development lifecycle. These debiasing techniques are applied to the input data, the model, or the model outputs. Debiasing activities can generally be categorized as follows:

» **Pre-processing:** Pre-processing debiasing techniques are used to transform the input data in ways to reduce bias. In most of these methods, this transformation occurs during preparation for training and primarily affects training data. Pre-processing methods shape the

[234] *See* Hooker.

data that are used to generate the model training algorithms but do not involve direct modifications to the algorithms themselves.

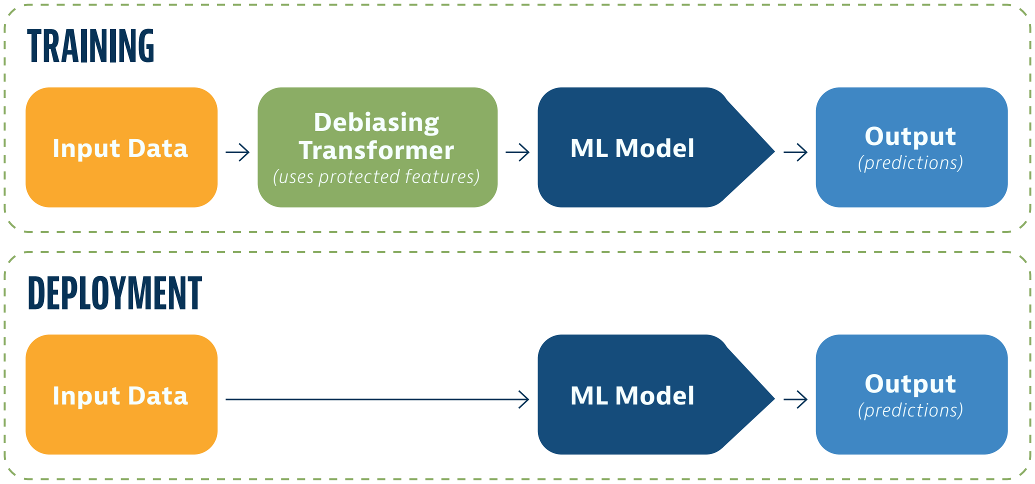
- **In-processing:** In-processing debiasing techniques use specialized machine learning training methods to reduce bias as the algorithm develops the underwriting model. These debiasing methods are not used once the model is deployed—they only modify the model training algorithm.
- **Post-processing:** Post-processing debiasing techniques transform the output of a machine learning model while the model is in use to reduce bias in the model's predictions. Most post-processing methods modify the outputs of a machine learning model using actual protected features and are therefore not relevant for financial services applications like credit underwriting due to disparate treatment requirements.[235]

Pre-processing and in-processing paradigms will be described and illustrated in turn below. Although adoption of individual techniques may be limited in the context of credit underwriting for reasons as varied as concerns about efficacy and regulatory uncertainty, interest in using these methods is growing.

5.3.1.1 Debiasing Through Pre-Processing

In this paradigm, a pre-processing method or transformer modifies data before it is used as inputs to the machine learning technique, where inputs in the training stage refer to features and outcomes data, and in the deployment stage, inputs refer only to features data. Most pre-processing debiasing methods *only* apply this transformation during training (as in Figure 5.3.1.1). Pre-processing methods use protected class characteristics—either actual data or imputed information—though the resulting machine learning model does not.

FIGURE 5.3.1.1 ILLUSTRATIVE PRE-PROCESSING DEBIASING PROCESS



[235] Emerging work on post-processing techniques may enable model debiasing in which protected class features are not required in actual deployment. Using a particular distance measure called the Wasserstein metric, protected class features of individuals are determined and then used for post-processing debiasing. This is useful in situations where actual protected class features cannot be used or are not available, which is often the case in credit underwriting. For details, *see* Alexey Miroshnikov *et al.,* Wasserstein-Based Fairness Interpretability Framework for Machine Learning Models, arXiv:2011.03156v2 (2021).

In general, debiasing techniques used during pre-processing have the advantage of using standard machine learning methods without modification. This means that pre-processing methods are compatible with "out of the box" machine learning software—since the machine learning methods do not require modification. Pre-processing methods are designed to reduce bias in any machine learning models trained on the pre-processed dataset, but such techniques will only be effective if future input data have a similar distribution to the training data. In practice, however, changing the dataset used to build a model may introduce other concerns, including risks related to disparate treatment liability where the transformations involve use of protected class status. In addition, even small divergences between deployment and training data will undercut the efficacy of pre-processing debiasing methods, just as data drift generally affects the accuracy of any machine learning model. In some cases, alterations to training data may enhance the likelihood of divergences between training and deployment conditions. Further, more research is needed to understand the circumstances in which synthetic data can be used with minimal risk of introducing accuracy, fairness, or other problems.

There are three main approaches to pre-processing debiasing methods: augmenting, transforming, or generating synthetic training data.

Augmenting Training Data

Description: In many cases, the training data reflect statistical biases—for example, it might contain a disproportionate number of White applicants with low risk profiles relative to non-White applicants when compared to the general population. Augmenting the training data is one way to correct this type of bias by, for example, adding more data points for non-White applicants assigned lower probabilities of default and for White applicants assigned higher probabilities of default. Developers may implement these methods to ensure that the training data meets the conditions of any of a number of approaches to measuring fairness as discussed above in Section 5.2.[236]

This method adds data points to the training dataset, prior to training the machine learning model. Oversampling is a common example, for instance by adding actual data from historically excluded groups to create a more balanced distribution and more accurate predictions for applicants from those groups. Where such data are not readily available, however, other forms of data augmentation draw from the existing training data and may involve the generation of synthetic data in some cases. For example, if a model developer was interested in trying to predict a very rare disease, he or she might construct a dataset using oversampling by duplicating data for people with that disease. After oversampling, the ratio of people with the disease in the dataset might be 1:50,000 rather than 1:1,000,000. Oversampling is used when there is a significant imbalance in the classes that the model is trying to predict. The focus here is on removing underlying statistical bias by generating a training dataset with a more representative distribution.

Analysis: Certain implementations of this approach improve fairness under these metrics with only a moderate effect on the model's accuracy.[237] Data augmentation has most commonly been used in the fields of computer vision and natural language processing to improve the robustness of models.[238] While data augmentation for financial time series data has yet to be systematically researched,

[236] *See e.g.,* Shubham Sharma *et al.*, Data Augmentation for Discrimination Prevention and Bias Disambiguation, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 358-364 (2020) (using data augmentation to minimize bias, achieving improved accuracy in model prediction).

[237] *Id.*

[238] *See e.g.,* Sebastian Cygert & Andrzej Czyżewski, Toward Robust Pedestrian Detection with Data Augmentation, 8 IEEE Access 136674-136683 (2020); Connor Shorten *et al.,* Text Data Augmentation for Deep Learning, 8 J. of Big Data art. 101 (2021).

one study suggests that some augmentation methods help improve model performance.[239] However, it is important for more research to be conducted using augmented data for financial time series data in order to have a better idea of the reliability of this method in this area.

Transforming Training Data

Description: Another approach to debiasing during pre-processing is to transform, rather than augment, the training data. The transformations are done with the aim of making it as difficult as possible to predict the protected feature(s) from the data. If the developer of an underwriting model were to use this approach, for example, suppose a dataset contains five attributes ("X") for each individual: debt-to-income ratio, credit score, utilization patterns, credit product mix, and account recency. The model developer also has real or imputed information about each individual's gender, the protected feature ("P") that is the focus of the debiasing effort, and whether the individual defaulted within six months, which is the "outcome" attribute ("Y").

To debias the model by training data transformation, the model developer tries to predict the protected feature P based on the non-protected features in the dataset (X). If P can successfully be predicted, then the developer modifies the non-protected features X until they are nearly impossible to predict P. One way to achieve this effect is by modifying individual features such that their distribution is equal for both groups. For instance, if the training data are transformed such that the distribution of utilization patterns between female and male credit applicants are similar, it is not possible to distinguish between the protected groups which leads to a debiased model.[240]

Analysis: This approach is theoretically appealing, and has influenced the development of a variety of other debiasing techniques, including adversarial debiasing. However as noted with respect to augmenting training data, the potential impacts of adding fabricated or modified data on accuracy are difficult to predict, making it challenging to deploy these methods in practice.

Generating Synthetic Training Data

Description: Model developers can use generative adversarial networks (GANs)[241] to produce a synthetic—meaning new and artificial—dataset designed to have a fairer representation than the original dataset with respect to particular groups or attributes. The synthetic data, rather than the original data, is then used for model training. For example, a fairness GAN might generate debiased data such that the synthetic dataset has equal representation of applicants across protected class features (which may be imputed using BISG), or is fair according to some other specified metric.[242] The model can then be trained and tested using the synthetic dataset to generate fairer predictions compared to a model using the original data.[243] Synthetic data refers to artificial data which are generated using GANs, whereas augmentation of data (see Section 5.3.1.1), can refer to either synthetic or actual data that are added to a training dataset to create more balance across different groups by probability of default (or other target variable).

[239] Elizabeth Fons *et al.*, Evaluating Data Augmentation for Financial Time Series Classification, arXiv:2010.15111v1 (2020).

[240] Michael Feldman *et al.,* Certifying and Removing Disparate Impact, KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 259-268 (2015).

[241] GANs involve two neural networks: a generator and a discriminator. The original use for GANs was to produce realistic data. In this case the generator produces new data points, while the discriminator tries to guess whether the points are real or artificial (produced by the generator). Both networks are trained simultaneously, such that the generator produces better examples while the discriminator becomes better at identifying artificial examples.

[242] Each of these approaches to measuring algorithmic fairness is explained in Section 5.2.

[243] Prasanna Sattigeri *et al.*, Fairness GAN, arXiv:1805.09910v1 (2018).

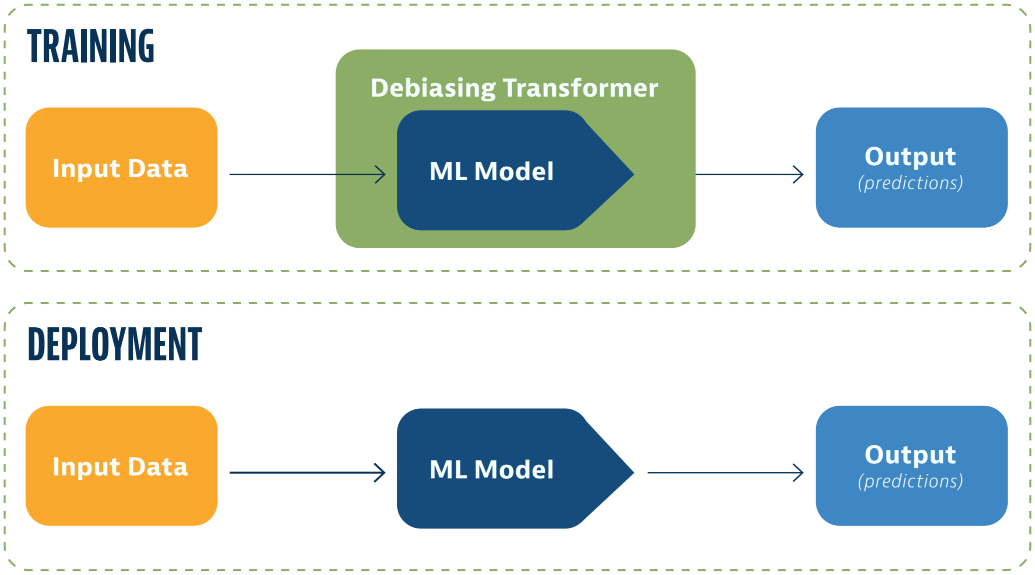
Analysis: GANs have garnered attention because they help users mitigate data security and privacy risks, since they do not generate information about real people and are flexible in the amount of data they can generate. They can also be restricted to produce data with specified properties—such as equal representation of applicants with specified characteristics. However, changing the dataset used to build a model introduces significant concerns. For example, synthetic data may diverge significantly enough from the reality in which the model is intended to operate that it raises concerns about the model's robustness. For instance, synthetic data generated for minority groups who are approved for loans may not result in reliable data points if they are produced from relatively small amounts of similar data that may not accurately reflect the full range of experience for those groups. The accuracy of the resulting model will only be as good as the design and implementation of the GAN.

5.3.1.2 Debiasing Through In-Processing

In this debiasing approach, debiasing is built into the machine learning algorithm used to train a model. Most in-processing methods use a debiasing function during training only. This debiasing function has access to real or imputed protected features, though the resulting machine learning model itself does not. In deployment, there is no debiasing function present.

In-processing debiasing approaches have the advantage of not altering training data, which can introduce accuracy and fairness risks as discussed above. Both common forms of in-processing debiasing also can be adapted easily to use a wide variety of mathematical definitions of fairness as discussed in Section 5.2. However, as with pre-processing debiasing techniques, the efficacy of in-processing debiasing depends on the deployment data conforming to a significant degree to the model's training data.

FIGURE 5.3.1.2 ILLUSTRATIVE IN-PROCESSING DEBIASING PROCESS



There are two popular approaches to in-processing debiasing: regularization and adversarial debiasing.

Regularization

Description: In addition to the use of regularization to increase transparency in models as discussed in Section 3.4.1.3, regularization can also be used to debias models. Adding regularization terms to the objective function leads the learning algorithm to select only relevant features, which decreases bias in the model and reduces the probability of overfitting.[244] Similarly, for the purpose of debiasing, regularization can add a penalty feature that is intended to increase fairness in a machine learning model. For example, a model developer training an underwriting model which can draw on 1,000 features as inputs wants the final model to use a small number of features for predictions because that will simplify the task of producing explanations.[245]

Analysis: Regularization approaches are appealing for several reasons: various types of regularization are well-understood and commonly used in machine learning, and these approaches *only* require access to protected features during training.[246] On the other hand, regularization usually requires hand-tuning—for example, to adjust the strength of a penalty—to achieve the desired outcome, which can be operationally cumbersome. However, there are practical challenges for firms considering this approach. In general, firms are expected to conduct independent fair lending assessments of their credit decisions and typically limit access to real or imputed protected class data to compliance teams to prevent improper use of that information and to reinforce the independence of compliance risk management functions. As a result, model development teams cannot practically build, tune, or test a model using information about protected class characteristics without deviating from standard practice.

Adversarial Debiasing

Description: Inspired by success in the context of image generation, adversarial debiasing methods have garnered increased attention.[247] Adversarial debiasing uses two separate machine learning models that interact during the training process. For example, if deployed in the development of an underwriting model:

- » The predictor model is an underwriting model that predicts the likelihood of default based on input data that do not include protected class characteristics.
- » The adversary model tries to predict protected features such as the applicant's race or gender based on the default probability that the predictor model produced for each individual in the training dataset. In other words, the adversarial model uses the predictor model's output as its input.

[244] Toshihiro Kamishima *et al.*, Fairness-Aware Learning Through Regularization Approach, 2011 IEEE 11th International Conference on Data Mining Workshops 643-650.

[245] Consider a simplified scenario. Without regularization, the most accurate model (M1) model produced by the learning algorithm uses 400 features and has an objective function value of 70,500. The second-most accurate model (M2) uses 100 features, and has an objective function value of 66,000. A training procedure designed to return the largest objective function value will return model M1. However, if the model developer adds regularization, a penalty of -100 will be deducted from the objective function for every feature used in the resulting model. Now, the "regularized" objective function values for two possible models are: 30,500 (or the original objective function value, 70,500, minus 100 for each of the 400 features used) for M1 and 56,000 (or the original objective function value of 66,000 minus 100 for each of the 100 features used) for M2. In this case, M2 has a larger, regularized objective since it uses fewer features than M1, and a training procedure using this regularized objective will return model M2.

[246] However, regularization and other methods to increase simplicity in models may, conversely, lead to inequity or increase bias. *See* Jon Kleinberg & Sendhil Mullainathan, Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability, EC '19: Proceedings of the 2019 ACM Conference on Economics and Computation 807-808 (2019).

[247] Brian Hu Zhang *et al.*, Mitigating Unwanted Biases with Adversarial Learning, Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society 335-340 (2018).

This approach posits that the more accurate the adversary is in identifying the applicant's protected class status, the more biased the underlying underwriting model is. During training, the predictor model is optimized according to two objectives: maximizing accuracy of the prediction of applicants' default risk and decreasing the accuracy of the adversary's predictions of the protected class characteristics of individual applicants. Simultaneously, the adversary is optimized to accurately predict the protected feature. During deployment, only the predictor model is used, without guidance from protected-feature data.

Analysis: Given the success of adversarial methods in areas of machine learning such as generative models and computer vision tasks, it is natural that firms in financial services and other sectors are interested in learning more about these techniques. This approach has the potential of making more robust and explicit firms' search for less discriminatory alternatives in the third prong of the disparate impact analysis, although concerns about potential fair lending risks associated with using protected class information in this application have chilled research. GANs support generation of a fairness-accuracy curve that can help model development teams assess tradeoffs between a multiplicity of models. However, efforts to develop adversarial debiasing techniques further and deploy them for underwriting have been substantially chilled by uncertainty related to the use of protected class information. Even if regulators clarify that such uses are permitted, additional factual questions will need to be resolved to determine their utility in the context of credit underwriting. First, adversarial methods are mainly intended for neural networks, and it is not clear if this approach will be useful for other model classes, such as linear models or decision trees. Second, the foundational research on adversarial debiasing is recent, so more research is needed to understand whether and in what circumstances this approach will work in a deployed settings, including credit underwriting.

***

Although lenders and regulators alike are interested in opportunities to make lending fairer and more inclusive, the path to realizing these aims will likely involve holistic consideration of the choices made to identify and mitigate sources of bias and discrimination problems throughout the process of developing and monitoring models. For example, choosing between particular fairness metrics often involves tradeoffs that can be difficult for the general public to understand.[248] Given this, market practice and regulatory expectations across the entire model lifecycle are relevant to defining fair and responsible use of machine learning underwriting models.

[248] Debjani Saha *et al.*, Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics, arXiv:2001.00089v3 (2020).

6. CONCLUSION

Market practices are rapidly evolving as to the use of machine learning for credit underwriting and the technology available to lenders to manage transparency and fairness concerns related to those models. So are the issues and debates presented in this report. Stakeholders in academia, industry, government, and civil society organizations are examining the capabilities and performance of current technologies and working to develop better solutions to support responsible use of machine learning underwriting models.

As highlighted throughout this report, individual lenders make a series of choices in developing and using machine learning models that are critical to managing the reliability and fairness of those models and enabling effective regulatory oversight. Although machine learning can seem entirely novel, these decisions may not be radically different from how lenders currently develop automated underwriting systems. The relevant differences may relate less to the goals or objectives developers need to meet than with the timing of specific decisions and the tools being used.

In the context of machine learning underwriting models, the decisions that model developers make about how to enable sufficient transparency into how the models work and the basis for individual predictions are particularly important. Without transparency, for example, a lender might be able to detect that its model's decisions show disparities across demographic groups, but would not be able to identify the causes of those disparities or reduce their problematic effects.

Early adopters of machine learning underwriting models are developing different approaches to building transparent models. Some opt for inherently interpretable models, while others pair more complex models with *post hoc* explainability methods. Firms and their regulators need more information about how well these approaches support responsible and fair use of new underwriting technologies to determine best practices and to assess how law, policy, and regulation may need to evolve to govern a marketplace where use of machine learning underwriting models is more common than it is today.

Those questions address not just how to comply with existing law and regulation but point to a broader reconsideration of the following topics:

- What considerations are relevant to identifying fair, responsible, and inclusive use of machine learning underwriting models?
- What kinds of transparency are relevant to establishing that a particular machine learning underwriting model is fit for use? How can or should transparency be measured?
- Where *post hoc* explainability methods are used, how should firms and regulators evaluate the trustworthiness and utility of information produced by these supplemental analyses?

- What decisions can lenders make about data to improve the reliability, fairness, and inclusiveness of machine learning underwriting models?

In the context of debate about these questions and rapidly changing technologies, FinRegLab and a team of researchers from the Stanford Graduate School of Business are conducting a study that will provide evidence relevant to many of these questions. This study will assess the capabilities and performance of various model diagnostic tools designed to support responsible use of machine learning underwriting models across a variety of dimensions:

- **Type of machine learning model:** Benchmark underwriting models will range from logistic regression and boosted trees to neural networks and ensemble models to identify whether the type of underwriting model being explained affects the accuracy and utility of information produced by the model diagnostic tools;
- **Model complexity:** Each form of machine learning being evaluated will have simple and complex forms to help us identify the tradeoffs, if any, between performance and transparency and between performance and fairness;
- **Changes in economic conditions:** Test datasets will simulate different economic environments, such as data from the 2009-2010 downturn, to help assess whether the model diagnostic tools can help lenders identify changes in data conditions and model performance once in operation; and
- **Shifts in applicant distribution:** Test datasets will encompass different kinds of borrowers with respect to geographic location and socioeconomic status to help us evaluate how well these tools detect fair lending and other risks.

The set of benchmark models have generally been designed to approximate machine learning models that lenders might use to estimate the risk of default associated with an application for credit. This evaluation will assess how a set of alternative definitions of algorithmic fairness that have emerged in academic literature work in the context of the underwriting models and model diagnostic tools used in this research.

In addition to empirical findings, this research will propose a framework that will help all stakeholders—model developers, risk and compliance personnel, and regulators—assess the accuracy and utility of accessible information about a machine learning underwriting model's decision-making. This framework will provide a substantive contribution to the current oversight approaches about model transparency by helping to define the questions to ask about the information that currently available model diagnostic tools produce. Those questions will help assess whether those tools produce information that is necessary for assessing compliance with legal and regulatory requirements and policy goals. This framework is intended to stimulate debate about and further contributions from various stakeholders regarding the development of an effective approach to promoting responsible, fair, and inclusive use of machine learning underwriting models.[249]

FinRegLab expects to report results from the empirical research being conducted with economists from the Stanford Graduate School of Business later this year and to conduct in-depth analysis of the implications of that research for law, policy, regulation, and market practice in 2022.

[249] For more on this research, visit FinRegLab's website, https://finreglab.org/ai-machine-learning/explainability-and-fairness-of-machine-learning-in-credit-underwriting/.

BIBLIOGRAPHY

Chirag Agarwal & Sara Hooker, Estimating Example Difficulty Using Variance of Gradients, arXiv:2008.11600v2 (Oct. 24, 2020), available at https://arxiv.org/pdf/2008.11600.pdf

Sumit Agarwal, Shashwat Alok, Pulak Ghosh, & Sudip Gupta, Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech, Indian School of Business (Apr. 2021), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3507827

Hafiz A. Alaka, Lukumon O. Oyedele, Hakeem A. Owolabi, Vikas Kumar, Saheed O. Ajayi, Olugbenga O. Akinade, & Muhammad Bilal., Systematic Review of Bankruptcy Prediction Models: Towards a Framework for Tool Selection, 94 Expert Systems with Applications 164-184 (Mar. 15, 2018), available at https://www.sciencedirect.com/science/article/pii/S0957417417307224

Andrés Alonso & José Manuel Carbó, Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation, Banco de España Working Paper No. 2105 (Jan. 27, 2021), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3774075

Naveed Akhtar & Ajmal Mian, Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey, 6 IEEE Access 14410-14430 (Feb. 19, 2018), available at https://ieeexplore.ieee.org/abstract/document/8294186

Daniel W. Apley & Jingyu Zhu, Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, arXiv:1612.08468 (Aug. 19, 2019), available at https://arxiv.org/pdf/1612.08468.pdf%5D

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, & Francisco Herrera, Explainable Artificial Intelligence (XAI) Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, arXiv:1910.10045v2 (Dec. 26, 2019), available at https://arxiv.org/abs/1910.10045

Robert J. Aumann & Lloyd S. Shapley, Values of Non-Atomic Games, Princeton Legacy Library (2016)

Boris Babic & Sara Gerke, Explaining Medical AI Is Easier Said Than Done, Stat (Jul. 21, 2021), available at https://www.statnews.com/2021/07/21/explainable-medical-ai-easier-said-than-done/

Bank Policy Institute, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (Jun. 25, 2021), available at https://www.regulations.gov/comment/OCC-2020-0049-0020

Jo Ann Barefoot, A Regtech Manifesto: Redesigning Financial Regulation for the Digital Age, Alliance for Innovative Regulation (Jul. 2020), available at https://regulationinnovation.org/regtech-manifesto/

Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 Cal. L. Rev. 671-732 (2016), available at http://dx.doi.org/10.15779/Z38BG31

Solon Barocas, Andrew D. Selbst, & Manish Raghavan, The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons, ACM Conference on Fairness, Accountability, and Transparency (FAT*) 5 (Dec. 12, 2020), available at https://ssrn.com/abstract=3503019

John M. Barron & Michael Staten, The Value of Comprehensive Credit Reports: Lessons from the U.S. Experience, in Margaret J. Miller, ed., Credit Reporting Systems and the International Economy (2003)

Majid Bazarbash, FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk, International Monetary Fund Working Paper (May 17, 2019), available at https://www.imf.org/en/Publications/WP/Issues/2019/05/17/FinTech-in-Financial-Inclusion-Machine-Learning-Applications-in-Assessing-Credit-Risk-46883

Jason R. Bent, Is Algorithmic Affirmative Action Legal? 108 Georgetown L. J. (Apr. 2020), available at https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2020/04/Is-Algorithmic-Affirmative-Action-Legal.pdf

Tobias Berg, Valentin Burg, Ana Gombović, & Maju Puri, On the Rise of the FinTechs: Credit Scoring Using Digital Footprints, 33 Rev. of Fin. Studies 2845–2897 (Jul. 2020), available at https://doi.org/10.1093/rfs/hhz099

Allen N. Berger & W. Scott Frame, Small Business Credit Scoring and Credit Availability, 47 J. of Small Bus. Mgmt. 5-22 (Jan. 2007), available at https://doi.org/10.1111/j.1540-627X.2007.00195.x

Raghav Bharadwaj, Top 5 AI Startups in Banking by Funding – A Brief Overview, Emerj (Nov. 19, 2019), https://emerj.com/ai-sector-overviews/top-5-ai-startups-in-banking-by-funding/

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, & John Guttag, What Is the State of Neural Network Pruning?, Proceedings of Machine Learning and Systems 2020, arXiv:2003.03033 (2020), available at https://arxiv.org/abs/2003.03033

Laura Blattner & Scott Nelson, How Costly Is Noise? Data and Disparities in Consumer Credit, arXiv:2105.07554v1 (May 17, 2021), available at https://arxiv.org/abs/2105.07554v1

Laura Blattner, Scott Nelson, & Jann Spiess, Unpacking the Black Box: Regulating Algorithmic Decisions (Jul. 21, 2021), available at https://gsb-faculty.stanford.edu/jann-spiess/files/2021/07/blackbox.pdf

BLDS, LLC, Discover Financial Services, & H2O.ai, Machine Learning: Considerations for Fairly and Transparently Expanding Access to Credit (2020), available at http://info.h2o.ai/rs/644-PKX-778/images/Machine%20Learning%20-%20Considerations%20for%20Fairly%20and%20Transparently%20Expanding%20Access%20to%20Credit.pdf

Board of Governors of the Federal Reserve System, Report to Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit (Aug. 2007), available at https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf

Board of Governors of the Federal Reserve System, Supervisory & Regulation Letter 13-19 (Dec. 5, 2013), available at https://www.federalreserve.gov/supervisionreg/srletters/sr1319.htm

Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, Supervisory & Regulation Letter 11-7: Supervisory Guidance on Model Risk Management (Apr. 4, 2011), available at https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf

Brighterion, Survey Report: Using AI to Manage Credit Risk: Lenders Report on Current AI Use and Future Investments (2020), available at https://brighterion.com/ai-for-credit-risk-lendit-survey-report/

David A. Broniatowski, Psychological Foundations of Explainability and Interpretability in Artificial Intelligence, National Institute of Standards and Technology (Apr. 2021), available at https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf

Andrew Burt, Is There a 'Right to Explanation' for Machine Learning in the GDPR?, International Association of Privacy Professionals (Jun. 1, 2017), available at https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/

Andrew Burt & Patrick Hall, What to Do When AI Fails, O'Reilly Radar (May 18, 2020), available at https://www.oreilly.com/radar/what-to-do-when-ai-fails/

Andrew Burt, Brenda Leong, Stuart Shirrell, & Xiangnong Wang, Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models, Immuta and Future of Privacy Forum (Jun. 2018), available at https://fpf.org/blog/beyond-explainability-a-practical-guide-to-managing-risk-in-machine-learning-models/

Florentin Butaru, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo, & Akhtar Siddique, Risk and Risk Management in the Credit Card Industry, 72 J. of Banking & Finance 218-239 (Nov. 2016), available at https://www.sciencedirect.com/science/article/pii/S0378426616301340

Aylin Caliskan, Joanna J. Bryson, & Arvind Narayanan, Semantics Derived Automatically from Language Corpora Contain Human-Like Biases, 356 Science 183-186 (2017), available at https://science.sciencemag.org/content/356/6334/183/tab-pdf

Peter Carroll & Saba Rehmani, Alternative Data and the Unbanked, Oliver Wyman (2017), available at https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2017/may/Alternative_Data_And_The_%20Unbanked.pdf

Jie Chen, Deep Insights into Explainability and Interpretability of Machine Learning Algorithms and Applications to Risk Management, presentation at the 2019 Joint Statistical Meetings (Jul. 29, 2019), available at https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=303053

Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, & Tom Golstein, LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition, published as a conference paper at the 2021 International Conference on Learning Representations, arXiv:2101.07922 (Jan. 2021), available at https://arxiv.org/abs/2101.07922

Alexandra Chouldechova, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, arXiv:1703.00056v1 (Feb. 28, 2017), available at https://arxiv.org/pdf/1703.00056.pdf

Radoslaw M. Cichy & Daniel Kaiser, Deep Neural Networks as Scientific Models, 23 Trends in Cognitive Sciences 305-317 (Apr. 2019), available at https://pubmed.ncbi.nlm.nih.gov/30795896/

Consumer Financial Protection Bureau, 2020 Consumer Response Annual Report (Mar. 24, 2021), available at https://www.consumerfinance.gov/data-research/research-reports/2020-consumer-response-annual-report/

Consumer Financial Protection Bureau, Compliance Bulletin and Policy Guidance 2016-02, 81 Fed. Reg. 74410 (Oct. 26, 2016), available at https://www.federalregister.gov/documents/2016/10/26/2016-25856/compliance-bulletin-and-policy-guidance-2016-02-service-providers

Consumer Financial Protection Bureau, Data Point, Credit Invisibles (2015), available at https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf

Consumer Financial Protection Bureau, Supervision and Examination Manual (Sept. 2020), available at https://files.consumerfinance.gov/f/documents/cfpb_supervision-and-examination-manual.pdf

Consumer Financial Protection Bureau, Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity (2014), available at https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

Consumer Financial Protection Bureau, What is a Prescreened Credit Card Offer? (Jul. 11, 2017), available at https://www.consumerfinance.gov/ask-cfpb/what-is-a-prescreened-credit-card-offer-en-1/ https://www.consumer.ftc.gov/articles/prescreened-credit-and-insurance-offers

Cheryl R. Cooper & Darryl E. Getter, Consumer Credit Reporting, Credit Bureaus, Credit Scoring, and Related Policy Issues, Congressional Research Service (Oct. 15, 2020), available at https://fas.org/sgp/crs/misc/R44125.pdf

Cornerstone Advisors, Credit Monitoring and the Need for Speed: The Case for Advanced Technologies (Q2 2020)

Penny Crosman, EU Proposes Restrictions on AI in Credit Scoring, Authentication, Am. Banker (Apr. 21, 2021), available at https://www.americanbanker.com/news/eu-plan-would-restrict-use-of-ai-in-credit-scoring-authentication

Sebastian Cygert & Andrzej Czyżewski, Toward Robust Pedestrian Detection with Data Augmentation, 8 IEEE Access 136674 - 136683 (Jul. 22, 2020), available at https://ieeexplore.ieee.org/document/9146161

Susanne Dandl & Christoph Molnar, Counterfactual Explanations (2019), available at https://christophm.github.io/interpretable-ml-book/counterfactual.html

DBRS, U.S. Unsecured Personal Loans — Marketplace Lenders Continue to Expand Market Share (Sept. 2019), available at https://www.dbrsmorningstar.com/research/350589/us-unsecured-personal-loans-marketplace-lenders-continue-to-expand-market-share

Asli Demirgüç-Kunt, Leora Klapper, Dorothe Singer, Saniya Ansar, & Jake Hess, The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution, World Bank Group (2018), available at https://globalfindex.worldbank.org/sites/globalfindex/files/2018-04/2017%20Findex%20full%20report_0.pdf

Rishi Desai, Shirley Wang, Muthiah Vaduganathan, Thomas Evers, & Sebastian Schneeweiss, Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes, 3 JAMA Network Open (Jan. 3, 2020), available at https://pubmed.ncbi.nlm.nih.gov/31922560/

Jürgen Dieber & Sabrina Kirrane, Why Model Why? Assessing the Strengths and Limitations of LIME, arXiv:2012.00093v1 (Nov. 30, 2020), available at https://arxiv.org/pdf/2012.00093.pdf

Anna Veronika Dorogush, Vasily Ershov, & Andrey Gulin, CatBoost: Gradient Boosting with Categorical Features Support, arXiv:1810.11363 (Oct. 24, 2018), available at https://arxiv.org/abs/1810.11363

Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning, arXiv:1702.08608v2 (Mar. 2, 2017), available at https://arxiv.org/pdf/1702.08608.pdf

Trevor Dryer, How Machine Learning Is Quietly Transforming Small Business Lending, Forbes (Nov. 1, 2018), available at https://www.forbes.com/sites/forbesfinancecouncil/2018/11/01/how-machine-learning-is-quietly-transforming-small-business-lending/?sh=101081306acc

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, & Rich Zemel, Fairness Through Awareness, arXiv:1104.3913v2 (Nov. 30, 2011), available at https://arxiv.org/abs/1104.3913

Marc N. Elliott, Allen Fremont, Peter A. Morrison, Philip Pantoja, & Nicole Lurie, A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity, 43 Health Services Research 1722-1736 (Sept. 2008), available at https://doi.org/10.1111/j.1475-6773.2008.00854.x

Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor, & Department of Treasury, Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11996 (Mar. 2, 1979), available at https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines

European Commission, Building Trust in Human Centric Artificial Intelligence (2019), available at https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence

European Commission, Proposal for a Regulation Laying Down Rules on Artificial Intelligence (Apr. 21, 2021), available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Carol A. Evans, Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks, Consumer Compliance Outlook 1-9 (Second Issue 2017), available at https://www.consumercomplianceoutlook.org/2017/second-issue/keeping-fintech-fair-thinking-about-fair-lending-and-udap-risks/

Experian, Fintech vs. Traditional FIs: Trends in Unsecured Personal Installment Loans (Sept. 17, 2019), available at http://go.experian.com/IM-20-EM-AA-FintechTrendseBook?cmpid=Insightsblog-091719-unsecured-personal-loans-infographic

Federal Deposit Insurance Corporation, 2017 National Survey of Unbanked and Underbanked Households (2018), available at https://www.fdic.gov/householdsurvey/2017/2017report.pdf

Federal Deposit Insurance Corporation, How America Banks: Household Use of Banking and Financial Services (2020), available at https://www.fdic.gov/analysis/household-survey/2019report.pdf

Federal Deposit Insurance Corporation, Financial Institution Letter 44-2008 (June 6, 2008), available at https://www.fdic.gov/news/financial-institution-letters/2008/fil08044.html

Federal Deposit Insurance Corporation, Financial Institution Letter 22-2017: Adoption of Supervisory Guidance on Model Risk Management (Jun. 7, 2017), available at https://www.fdic.gov/news/financial-institution-letters/2017/fil17022.html

Federal Deposit Insurance Corporation, Financial Institution Letter 19-2019 (Apr. 2, 2019), available at https://www.fdic.gov/news/financial-institution-letters/2019/fil19019.html

Federal Trade Commission, Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues (2016), available at https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report

Federal Trade Commission, Prescreened Credit and Insurance Offers (May 2021), available at https://www.consumer.ftc.gov/articles/prescreened-credit-and-insurance-offers

Federal Trade Commission, Report to Congress under Section 319 of the Fair and Accurate Credit Transactions Act of 2003 (2012), available at https://www.ftc.gov/sites/default/files/documents/reports/section-319-fair-and-accurate-credit-transactions-act-2003-fifth-interim-federal-trade-commission/130211factareport.pdf

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, & Suresh Venkatasubramanian, Certifying and Removing Disparate Impact, KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 259-268 (Aug. 10, 2015), available at https://doi.org/10.1145/2783258.2783311

Financial Stability Board, Artificial Intelligence and Machine Learning in Financial Services (Nov. 1, 2017), available at https://www.fsb.org/wp-content/uploads/P011117.pdf

FinRegLab, The Use of Cash-Flow Data in Underwriting Credit: Empirical Research Findings (Jul. 2019), available at https://finreglab.org/wp-content/uploads/2019/07/FRL_Research-Report_Final.pdf

FinRegLab, The Use of Cash-Flow Data in Underwriting Credit: Market Context & Policy Analysis (Feb. 2020), available at https://finreglab.org/wp-content/uploads/2020/03/FinRegLab_Cash-Flow-Data-in-Underwriting-Credit_Market-Context-Policy-Analysis.pdf

Elizabeth Fons, Paula Dawson, Xiao-jun Zeng, John Keane, & Alexandros Iosifidis, Evaluating Data Augmentation for Financial Time Series Classification, arXiv:2010.15111v1 (Oct. 28, 2020), available at https://arxiv.org/abs/2010.15111

Jerome H. Friedman, Stochastic Gradient Boosting, 38 Computational Statistics & Data Analysis 367-378 (Feb. 28, 2002), available at https://doi.org/10.1016/S0167-9473(01)00065-2

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, & Ansgar Walther, Predictably Unequal? The Effects of Machine Learning on Credit Markets, J. of Finance (forthcoming) (Jun. 21, 2021), available at http://dx.doi.org/10.2139/ssrn.3072038

Pratik Gajane & Mykola Pechenizkiy, On Formalizing Fairness in Prediction with Machine Learning, arXiv:1710.03184v3 (May 28, 2018), available at https://arxiv.org/pdf/1710.03184.pdf

Leonardo Gambacorta, Yiping Huang, Han Qiu, & Jingyi Wang, How Do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm, BIS Working Paper No. 834 (Dec. 19, 2019), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3506945

Damien Garreau & Ulrike von Luxburg, Explaining the Explainer: A First Theoretical Analysis of LIME, Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, arXiv:2001.03447v2 (Jan. 13, 2020), available at https://arxiv.org/pdf/2001.03447.pdf

Susan Wharton Gates, Vanessa Gail Perry, & Peter M. Zorn, Automated Underwriting in Mortgage Lending: Good News for the Underserved?, 13 Housing Policy Debate 369-391 (2002), available at https://doi.org/10.1080/10511482.2002.9521447

Sushmito Ghosh & Douglas L. Reilly, Credit Card Fraud Detection with a Neural-Network, 1994 Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences (Jan. 1994), available at https://doi.org/10.1109/HICSS.1994.323314

Talia Gillis, The Input Fallacy, Minn. L. Rev. (forthcoming 2022) (Feb. 16, 2021), available at https://dx.doi.org/10.2139/ssrn.3571266

Talia Gillis & Jann Spiess, Big Data and Discrimination, 86 U. Chicago L. Rev. 459-487(2019), available at https://chicagounbound.uchicago.edu/uclrev/vol86/iss2/4

Leilani Gilpin, David Bau, Ben Yuan, Ayesha Bajwa, Michael Specter, & Lalana Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, arXiv:1806.00069v3 (Feb. 3, 2019), available at https://arxiv.org/pdf/1806.00069.pdf

Kristine Gloria, Power and Progress in Algorithmic Bias, Aspen Institute (Jul. 15, 2021), available at https://www.aspeninstitute.org/wp-content/uploads/2021/07/Power-Progress-in-Algorithmic-Bias-July-2021.pdf

R. Y. Goh & L. S. Lee, Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches, 2019 Advances in Operations Research art. 1974794 (Mar. 13, 2019), available at https://doi.org/10.1155/2019/1974794

Alex Goldstein, Adam Kapelner, Justin Bleich, & Emil Pitkin, Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation, arXiv:1309.6392v2 (Mar. 21, 2014), available at https://arxiv.org/pdf/1309.6392.pdf

Ian J. Goodfellow, Jonathon Shlens, & Christian Szegedy, Explaining and Harnessing Adversarial Examples, published as a conference paper at the 2015 International Conference on Learning Representations, arXiv:1412.6572 (Mar. 20, 2015), available at https://arxiv.org/abs/1412.6572

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, & Dino Pedreschi, A Survey of Methods for Explaining Black Box Models, 51 ACM Computing Surveys art. 93 at 1-42 (Aug. 22, 2018), available at https://doi.org/10.1145/3236009

Patrick Hall, Proposals for Model Security: Fair and Private Models, Whitehat and Forensic Model Debugging, and Common Sense (Jun. 19, 2019), available at https://github.com/jphall663/secure_ML_ideas/blob/master/secure_ml_ideas.pdf

Patrick Hall & Navdeep Gill, An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI, O'Reilly (2nd ed. Aug. 2019), available at https://www.h2o.ai/wp-content/uploads/2019/08/An-Introduction-to-Machine-Learning-Interpretability-Second-Edition.pdf

Patrick Hall, Navdeep Gill, & Nicholas Schmidt, Proposed Guidelines for the Responsible Use of Explainable Machine Learning, arXiv:1906.03533v3 (Nov. 29, 2019), available at https://arxiv.org/abs/1906.03533

David J. Hand, Classifier Technology and the Illusion of Progress, 21 Statistical Science 1-15 (2006), available at https://arxiv.org/pdf/math/0606441.pdf

Deborah Hellman, Measuring Algorithmic Fairness, 108 Va. L. Rev. 811-866 (Jun. 10, 2020), available at https://ssrn.com/abstract=3418528

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, Hanna Wallach, Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?, 2019 ACM CHI Conference on Human Factors in Computing Systems, arXiv:1812.05239v2 (Jan. 7, 2019), available at https://arxiv.org/abs/1812.05239

Sara Hooker, Opinion, Moving Beyond "Algorithmic Bias Is a Data Problem", Patterns (Apr. 9, 2021), available at https://doi.org/10.1016/j.patter.2021.100241

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, & Andrea Frome, What Do Compressed Deep Neural Networks Forget? arXiv:1911.05248v2 (Jul. 13, 2020), available at https://arxiv.org/abs/1911.05248

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, & Emily Denton, Characterising Bias in Compressed Models, arXiv:2010.03058v2 (Dec. 18, 2020), available at https://arxiv.org/abs/2010.03058

David C. Hsia, Credit Scoring and the Equal Credit Opportunity Act, 30 Hastings L. J. 371 (1978), available at https://repository.uchastings.edu/cgi/viewcontent.cgi?article=2586&context=hastings_law_journal

Lily Hu & Issa Kohler-Hausmann, What's Sex Got to Do with Fair Machine Learning?, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 513 (Jan. 27, 2020), available at https://doi.org/10.1145/3351095.3375674

Ting Huang, Brian McGuire, Chris Smith, & Gary Yang, The History of Artificial Intelligence, University of Washington (Dec. 2006), available at https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, & Jonathan Ullman, Differentially Private Fair Lending, Proceedings of the 36th Annual Conference on Machine Learning, 97 Proceedings of Machine Learning Research (2019), available at http://proceedings.mlr.press/v97/jagielski19a/jagielski19a.pdf

Megan Jarrell, Artificial Intelligence at Square—Two Use-Cases, Emerj (Sept. 6, 2021), https://emerj.com/ai-sector-overviews/artificial-intelligence-at-square/

Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, & João Gama, How Can I Choose an Explainer?: An Application-Grounded Evaluation of Post-Hoc Explanations, FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 805-815 (Mar. 2021), available at https://doi.org/10.1145/3442188.3445941

Weiwei Jiang & Jiayun Luo, An Evaluation of Machine Learning and Deep Learning Models for Drought Prediction Using Weather Data, preprint submitted to J. of LATEX Templates, arXiv:2107.02517v1 (Jul. 7, 2021), available at https://arxiv.org/pdf/2107.02517.pdf

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, & Michael C. Mozer, Characterizing Structural Regularities of Labeled Data in Overparameterized Models, Proceedings of the 38th International Conference on Machine Learning, 139 Proceedings of Machine Learning Research, arXiv:2002.03206v3 (Jun. 15, 2021), available at https://arxiv.org/pdf/2002.03206.pdf

Gabbrielle M. Johnson, Proxies Aren't Intentional, They're Intentional (2021) (unpublished manuscript)

Jonathan Johnson, Interpretability vs. Explainability: The Black Box of Machine Learning, BMC (Jul. 16, 2020), available at https://www.bmc.com/blogs/machine-learning-interpretability-vs-explainability/

Michael Jordan & Tom Mitchell, Machine Learning: Trends, Perspectives, and Prospects, 349 Science 255-260 (Jul. 17, 2015), available at https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf

JPMorgan Chase & Co., Chairman & CEO Letter to Shareholders (Apr. 7, 2021), available at https://reports.jpmorganchase.com/investor-relations/2020/ar-ceo-letters.htm

Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, & Sharad Goel, Omitted and Included Variable Bias in Tests for Disparate Impact, arXiv:1809.05651v3 (Aug. 30, 2019), available at https://arxiv.org/abs/1809.05651

Toshihiro Kamishima, Shotaro Akaho, & Jun Sakuma, Fairness-Aware Learning Through Regularization Approach, 2011 IEEE 11th International Conference on Data Mining Workshops 643-650, available at https://ieeexplore.ieee.org/document/6137441

Amir E. Khandani, Adlar J. Kim, & Andrew W. Lo, Consumer Credit-Risk Models via Machine-Learning Algorithms, 34 J. of Banking & Finance 2767-2787 (May 9, 2010), available at http://dx.doi.org/10.1016/j.jbankfin.2010.06.001

Eric Knight, Note, AI and Machine Learning-Based Credit Underwriting and Adverse Action Under the ECOA, 3 Bus. & Fin. L. Rev. 236-258 (2020), available at https://gwbflr.org/wp-content/uploads/2020/04/Volume-3-Issue-2-Note-1_Knight.pdf

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, & Ashesh Rambachan, Algorithmic Fairness, 108 AEA Papers and Proceedings 22-27 (May 2018), available at https://doi.org/10.1257/pandp.20181018

Jon Kleinberg & Sendhil Mullainathan, Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability, EC '19: Proceedings of the 2019 ACM Conference on Economics and Computation 807-808 (Jun. 17, 2019), available at https://doi.org/10.1145/3328526.3329621

Jon Kleinberg, Sendhil Mullainathan, & Manish Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, arXiv:1609.05807v2 (Nov. 17, 2016), available at https://arxiv.org/abs/1609.05807

KPMG, Thriving in an AI World (Apr. 2021), available at https://info.kpmg.us/content/dam/info/en/news-perspectives/pdf/2021/Updated%204.15.21%20-%20Thriving%20in%20an%20AI%20world.pdf

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, & Sorelle A. Friedler, Problems with Shapley-Value-Based Explanations as Feature Importance Measures, Proceedings of the 37th International Conference on Machine Learning, 119 Proceedings of Machine Learning Research, arXiv:2002.11097v2 (Jun. 30, 2020), available at https://arxiv.org/pdf/2002.11097.pdf

Heidi Ledford, Millions of Black People Affected by Racial Bias in Health-Care Algorithms, Nature (Oct. 24, 2019), available at https://www.nature.com/articles/d41586-019-03228-6

Robert Letzler, Ryan Sandler, Ania Jaroszewicz, Isaac Knowles, & Luke M. Olson, Knowing When to Quit: Default Choices, Demographics and Fraud, 127 Econ. J. 2617–2640 (Dec. 2017), available at https://doi.org/10.1111/ecoj.12377

Sheng-Tun Li, Weissor Shiue, & Meng-Huah Huang, The Evaluation of Consumer Loans Using Support Vector Machines, 30 Expert Systems with Applications 772-782 (May 2006), available at https://doi.org/10.1016/j.eswa.2005.07.041

Wei Li, Laurie Goodman, & Denise Bonsu, The Lasting Impact of Foreclosures and Negative Public Records, Urban Institute Housing Policy Finance Center (Nov. 2016), available at https://www.urban.org/sites/default/files/publication/85356/the-lasting-impact-of-foreclosures-and-negative-public-records_0.pdf

Zachary C. Lipton, The Mythos of Model Interpretability, arXiv:1606.03490v3 (Mar. 6, 2017), available at https://arxiv.org/pdf/1606.03490.pdf

Steve Lohr, Facial Recognition Is Accurate, If You're a White Guy, N.Y. Times (Feb. 9, 2018), available at https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Scott Lundberg, Gabriel Erion, & Su-In Lee, Consistent Individualized Feature Attribution for Tree Ensembles, arXiv:1802.03888v3 (Mar. 7, 2019), available at https://arxiv.org/abs/1802.03888

Scott Lundberg & Su-In Lee, A Unified Approach to Interpreting Model Predictions, 31st Conference on Neural Information Processing Systems 2017, arXiv:1705.07874v2 (Nov. 25, 2017), available at https://arxiv.org/abs/1705.07874

Mark MacCarthy, Fairness in Algorithmic Decision-Making, Brookings Institute (Dec. 6, 2019), available at https://www.brookings.edu/research/fairness-in-algorithmic-decision-making/

James Manyika, Jake Silberg, & Brittany Presten, What Do We Do About the Biases in AI?, Harvard Bus. Rev. (Oct. 25, 2019), available at https//hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, & Aram Galstyan, A Survey on Bias and Fairness in Machine Learning, arXiv:1908.09635v2 (Sep. 17, 2019), available at https://arxiv.org/abs/1908.09635

John Merrill, Geoff Ward, Sean Kamkar, Jay Budzik, & Douglas Merrill, Generalized Integrated Gradients: A Practical Method for Explaining Diverse Ensembles, submitted to the J. of Machine Learning Research, arXiv:1909.01869v2 (Sept. 6, 2019), available at https://arxiv.org/abs/1909.01869

Smitha Milli, Ludwig Schmidt, Anca D. Dragan, & Moritz Hardt, Model Reconstruction from Model Explanations, arXiv:1807.05185v1 (Jul. 13, 2018), available at https://arxiv.org/abs/1807.05185

Alexey Miroshnikov, Konstandinos Kotsiopoulos, Ryan Franks, & Arjun Ravi Kannan, Wasserstein-Based Fairness Interpretability Framework for Machine Learning Models, arXiv:2011.03156v2 (Apr. 19, 2021), available at https://arxiv.org/abs/2011.03156

Tom Mitchell, Machine Learning, McGraw Hill (1997)

Model Risk Managers' International Association, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, (May 25, 2021), available at https://www.regulations.gov/comment/OCC-2020-0049-0008

Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Boxes Explainable (2019), available at https://christophm.github.io/interpretable-ml-book/

NAACP Legal Defense and Education Fund & Student Borrower Protection Center, LDF and Student Borrower Protection Center Announce Fair Lending Testing Agreement with Upstart Network (Dec. 2020), available at https://www.naacpldf.org/wp-content/uploads/FINAL-SBPC-LDF-Release-.pdf

National Fair Housing Alliance, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (Jul. 1, 2021), available at https://www.regulations.gov/comment/OCC-2020-0049-0055

National Consumer Law Center, Past Imperfect: How Credit Scores and Other Analytics "Bake In" and Perpetuate Past Discrimination (2016)

Ha-Thu Nguyen, Reject Inference in Application Scorecards: Evidence from France, Economix Working Paper 2016-10 (Feb. 2016), available at https://economix.fr/pdf/dt/2016/WP_EcoX_2016-10.pdf

Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence (2019), available at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Office of the Comptroller of the Currency, Bulletin 1997-24: Credit Scoring Models: Examination Guidance, app. at 11 (May 20, 1997), available at https://www.occ.treas.gov/news-issuances/bulletins/1997/bulletin-1997-24.html

Office of the Comptroller of the Currency, Bulletin 2011-12: Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management (Apr. 4, 2011), available at https://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12.html

Office of the Comptroller of the Currency, Bulletin 2013-29 (Oct. 30, 2013), available at https://www.occ.gov/news-issuances/bulletins/2013/bulletin-2013-29.html

Office of the Comptroller of the Currency, Bulletin 2020-10 (Mar. 5, 2020), available at https://www.occ.gov/news-issuances/bulletins/2020/bulletin-2020-10.html

Office of the Comptroller of the Currency, Comptroller's Handbook, Credit Card Lending (Version 2.0, Apr. 2021), available at https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/credit-card-lending/pub-ch-credit-card.pdf

Oportun, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (Jul. 1, 2021), available at https://www.regulations.gov/comment/OCC-2020-0049-0065

Florian Ostmann & Cosmina Dorobantu, AI in Financial Services, The Alan Turing Institute (Jun. 11, 2021), available at https://doi.org/10.5281/zenodo.4916041

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, & Anathram Swami, Practical Black-Box Attacks against Machine Learning, ASIA CCS '17: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security 506-519 (Apr. 2017), available at https://doi.org/10.1145/3052973.3053009

Leslie Parrish, Alternative Data and Advanced Analytics: Table Stakes for Unsecured Personal Loans, Aite Group (Nov. 19, 2019), available at https://aitegroup.com/report/alternative-data-and-advanced-analytics-table-stakes-unsecured-personal-loans

Leslie Parrish, Impact Report, Consumer Lenders' Plans for Navigating the Next Normal, Aite Group (May 2021), available at https://aitegroup.com/report/consumer-lenders%E2%80%99-plans-navigating-next-normal

Leslie Parrish, Risky Business: The State of Play for Risk Executives in the Analytics Ecosystem, Aite Group (Nov. 14, 2019), available at https://aitegroup.com/report/risky-business-state-play-risk-executives-analytics-ecosystem

Brendan Pedersen, OCC Announces Initiative to Expand Credit Access in Los Angeles, Am. Banker (Oct. 30, 2020), available at https://www.americanbanker.com/news/occ-announces-initiative-to-expand-credit-access-in-los-angeles

Dana Pessach & Erez Shmueli, Algorithmic Fairness, arXiv:2001.09784v1 (Jan. 21, 2020), available at https://arxiv.org/pdf/2001.09784.pdf

Anastasios Petropoulos, Vasilis Siakoulis, Evaggelos Stavroulakis, & Aristotelis Klamargis, A Robust Machine Learning Approach for Credit Risk Analysis of Large Loan Level Data Sets Using Deep Learning and Extreme Gradient Boosting, Bank for International Settlements (2018), available at https://www.bis.org/ifc/publ/ifcb49_49.pdf

Pew Research Center, Demographics of Mobile Device Ownership and Adoption in the United States (Apr. 7, 2021), available at https://www.pewresearch.org/internet/fact-sheet/mobile/

Kristina Preuer, Günter Klambauer, Friedrich Rippmann, Sepp Hochreiter, & Thomas Unterthiner, Interpretable Deep Learning in Drug Discovery, in Explainable AI: Interpreting, Explaining, and Visualizing Deep Learning 331-345 (2019), available at https://link.springer.com/chapter/10.1007/978-3-030-28954-6_18

Arun Rai, Explainable AI: From Black Box to Glass Box, 48 J. of the Academy of Marketing Science 137-141 (Dec. 17, 2019), available at https://doi.org/10.1007/s11747-019-00710-5

Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model: Initial Report of the Independent Monitor (Apr. 14, 2021), available at https://www.relmanlaw.com/media/cases/1088_Upstart%20Initial%20Report%20-%20Final.pdf

Marco Tulio Ribeiro, Sameer Singh, & Carlos Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, arXiv:1602.04938v3 (Aug. 9, 2016), available at https://arxiv.org/pdf/1602.04938.pdf

Lisa Rice & Deidre Swesnik, Discriminatory Effects of Credit Scoring on Communities of Color, 46 Suffolk L. Rev. 935 (2013)

Ralph J. Rohner, Equal Credit Opportunity Act, 34 Bus. Law. 1423 (1979), available at https://scholarship.law.edu/scholar/679/

The Royal Society, Explainable AI: The Basics (Nov. 2019), available at https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

Peter Rudegeair & AnnaMaria Andriotis, JPMorgan, Others Plan to Issue Credit Cards to People With No Credit Scores, Wall St. J. (May 13, 2021), available at https://www.wsj.com/articles/jpmorgan-others-plan-to-issue-credit-cards-to-people-with-no-credit-scores-11620898206?mod=searchresults_pos3&page=1

Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Machine Intelligence 206-215 (May 13, 2019), available at https://doi.org/10.1038/s42256-019-0048-x

Cynthia Rudin & Joanna Radin, Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition, Harvard Data Science Rev. (Issue 1.2, Fall 2019), available at https://doi.org/10.1162/99608f92.5a8a3a3d

Debjani Saha, Candice Schumann, Duncan C. McElfresh, John P. Dickerson, Michelle L. Mazurek, & Michael Carl Tschantz, Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics, arXiv:2001.00089v3 (Jul. 2, 2020), available at https://arxiv.org/abs/2001.00089

Salesforce, Sixth Edition State of Marketing Report (2020)

Arthur L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, IBM J. of Research & Development 211-259 (1959)

Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, & Kush R. Varshney, Fairness GAN, arXiv:1805.09910v1 (May 24, 2018), available at https://arxiv.org/abs/1805.09910

Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B. Sood, Greg S. Corrado, Lily Peng, & Dale R. Webster, Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy, 126 Ophthalmology 552-564 (Apr. 2019), available at https://doi.org/10.1016/j.ophtha.2018.11.016

Nicholas Schmidt & Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination, 73 Consumer Finance Law Quarterly Report 130-144 (Nov. 8, 2019), available at https://arxiv.org/pdf/1911.05755.pdf

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, & Dan Dennison, Hidden Technical Debt in Machine Learning Systems, 2 NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems 2503-2511 (Dec. 7, 2015), available at https://dl.acm.org/doi/10.5555/2969442.2969519

Andrew D. Selbst & Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085-1139 (2018), available at https://ir.lawnet.fordham.edu/flr/vol87/iss3/11/

L.S. Shapley, Notes on the n-Person Game, II: The Value of an n-Person Game, U.S. Air Force Project RAND Research Memorandum (Aug. 21, 1951), available at https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM670.pdf

Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, & Kush R. Varshney, Data Augmentation for Discrimination Prevention and Bias Disambiguation, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 358-364 (Feb. 2020), available at https://dl.acm.org/doi/10.1145/3375627.3375865

Connor Shorten, Taghi M. Khoshgoftaar, & Borko Furht, Text Data Augmentation for Deep Learning, 8 J. of Big Data art. 101 (Jul. 19, 2021), available at https://doi.org/10.1186/s40537-021-00492-0

Yan-yan Song & Ying Lu, Decision Tree Methods: Applications for Classification and Prediction, 27 Shanghai Archives of Psychiatry 2 (2015) at 130-135, available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

Judith Spitz, Why Tech Executives Must Embrace Diversity as Their First Line of Defense Against the Business Impacts of Algorithmic Bias, Forbes (Jul. 1, 2021), available at https://www.forbes.com/sites/judithspitz/2021/07/01/why-tech-executives-must-embrace-diversity-as-their-first-line-of-defense-against-the-business-impacts-of-algorithmic-bias/?sh=52c3906fc6d8

Sophie Stalla-Bourdillon, Brenda Leong, Patrick Hall, & Andrew Burt, Warning Signs: The Future of Privacy and Security in an Age of Machine Learning, Future of Privacy Forum (Sept. 2019), available at https://fpf.org/wp-content/uploads/2019/09/FPF_WarningSigns_Report.pdf

Brian Stanton & Theodore Jensen, Trust and Artificial Intelligence, National Institute of Standards and Technology (Dec. 2020), available at https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087

Emma Strubell, Ananya Ganesh, & Andrew McCallum, Energy and Policy Considerations for Modern Deep Learning Research, 34 Proceedings of the AAAI Conference on Artificial Intelligence (Apr. 2020), available at https://doi.org/10.1609/aaai.v34i09.7123

Agus Sudjianto, What We Need Is Interpretable and Not Explainable Machine Learning, presentation at Cogilytica Machine Learning Lifecycle Conference (Jan. 2021), available at https://events.cognilytica.com/wp-content/uploads/2020/12/What-we-need-is-interpretable-and-not-explainable-machine-learning-Agus-Sudjianto-ML-Lifecycle-Slides-.pdf.

Agus Sudjianto, William Knauth, Rahul Singh, Zebin Yang, & Aijun Zhang, Unwrapping the Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification, arXiv:2011.04041v1 (Nov. 8, 2020), available at https://arxiv.org/pdf/2011.04041.pdf

Mukund Sundararajan, Ankur Taly, & Qiqi Yan, Axiomatic Attribution for Deep Networks, Proceedings of the 34th International Conference on Machine Learning, 70 Proceedings of Machine Learning Research 3319-3328 (2017), available at http://proceedings.mlr.press/v70/sundararajan17a.html

Harini Suresh & John V. Guttag, A Framework for Understanding Sources of Harm Throughout the Machine Learning Lifecycle, arXiv:1901.10002v4 (Jun. 15, 2021), available at https://arxiv.org/abs/1901.10002

Winnie F. Taylor, Meeting the Equal Credit Opportunity Act's Specificity Requirement: Judgmental and Statistical Scoring Systems, 29 Buff. L. Rev. 73, 82 (1980), available at https://digitalcommons.law.buffalo.edu/buffalolawreview/vol29/iss1/4/

Robert Tibshirani, Regression Shrinkage and Selection via the Lasso, 58 J. of the Royal Statistical Society 267-288 (1996), available at https://www.jstor.org/stable/2346178

Nicol Turner Lee, Paul Resnick, & Genie Barton, Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms, Brookings Institute (May 22, 2019), available at https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Upstart, Auto Loans (undated), available at https://www.upstart.com/for-banks/auto-loans/

Upstart, Blog, By the Numbers (undated), available at https://www.upstart.com/blog/upstart-by-the-numbers

Upstart, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (Jul. 1, 2021), https://downloads.regulations.gov/OCC-2020-0049-0056/attachment_1.pdf

Upstart, Results to Date (visited Jul. 29, 2021), available at https://www.upstart.com/about#results-to-date

Upstart Holdings, Inc., Form 10-K (Mar. 18, 2021), available at https://sec.report/Document/0001647639-21-000004/

VantageScore, Our Models (visited Jul. 29, 2021), available at https://vantagescore.com/lenders/our-models#vantage-score-4

Starre Vartan, Racial Bias Found in a Major Health Care Risk Algorithm, Scientific American (Oct. 24, 2019), available at https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

Sahil Verma & Julia Rubin, Fairness Definitions Explained, FairWare'18: Proceedings of the IEEE/ACM International Workshop on Software Fairness (May 29, 2018), available at https://doi.org/10.1145/3194770.3194776

Sandra Wachter, Brent Mittelstadt, & Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation (Jan. 24, 2017), International Data Privacy Law, available at https://dx.doi.org/10.2139/ssrn.2903469

Sandra Wachter, Brent Mittelstadt, & Chris Russell, Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI, Computer L. & Security Rev. (forthcoming) (May 17, 2021), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547922

Wei Wang, Christopher Lesner, Alexander Ran, Marko Rukonic, Jason Xue, & Eric Shiu, Using Small Business Banking Data for Explainable Credit Risk Scoring, 34 Proceedings of the AAAI Conference on Artificial Intelligence 13396-13401 (Apr. 2020), available at https://ojs.aaai.org//index.php/AAAI/article/view/7055

Betsy Anne Williams, Catherine F. Brooks, & Yotam Shmargad, How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications, 8 J. of Information Policy 78-115 (2018), available at https://doi.org/10.5325/jinfopoli.8.2018.0078

Lasse F. Wolff Anthony, Benjamin Kanding, & Raghavendra Selvan, Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models, ICML Workshop on "Challenges in Deploying and Monitoring Machine Learning Systems", arXiv:2007.03051v1 (Jul. 6, 2020), available at https://arxiv.org/pdf/2007.03051.pdf

Becky Yerak, AI Helps Auto-Loan Company Handle Industry's Trickiest Turn, Wall St. J. (Jan. 3, 2019), available at https://www.wsj.com/articles/ai-helps-auto-loan-company-handle-industrys-trickiest-turn-11546516801

Ed Yong, A Popular Algorithm Is No Better at Predicting Crimes than Random People, The Atlantic (Jan. 17, 2018), available at https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/

Zest AI, Here's How ML Underwriting Fits Within Federal Model Risk Management Guidelines (May 30, 2019), available at https://www.zest.ai/insights/heres-how-ml-underwriting-fits-within-federal-model-risk-management-guidelines

Zest AI, Why ZAML Makes Your ML Platform Better, Zest AI (Mar. 6, 2019), available at https://www.zest.ai/insights/why-zaml-makes-your-ml-platform-better

Brian Hu Zhang, Blake Lemoine, & Margaret Mitchell, Mitigating Unwanted Biases with Adversarial Learning, Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society 335-340 (2018), available at https://dl.acm.org/doi/10.1145/3278721.3278779

Nengfeng Zhou, Zach Zhang, Vijayan N. Nair, Harsh Singhal, Jie Chen, & Agus Sudjianto, Bias, Fairness, and Accountability with AI and ML Algorithms, arXiv:2105.06558v1 (May 6, 2021), available at https://arxiv.org/abs/2105.06558

Scott Zoldi, How to Make "Black Box" Neural Networks Explainable, FICO Blog (Jan. 14, 2019), retrieved on Jun. 18, 2021, available at https://www.fico.com/blogs/deep-dive-how-make-black-box-neural-networks-explainable

Scott Zoldi, Not All Explainable AI is Created Equal, Retail Banker International (Oct. 9, 2019), available at https://www.retailbankerinternational.com/special-reports/not-all-explainable-ai-is-created-equal/

APPENDIX A

Common Terms and Acronyms

Accumulated Local Effects (ALE) Plots: Accumulated Local Effects plots are common visualization methods that depict how an individual feature interacts with the model's predictions and are used as a feature importance explainability technique. ALE plots depict the relationship between the variable of interest and the outcome without accounting for other factors.

Adversarial debiasing: Adversarial models are models that can be used during training to debias machine learning models. In this context, adversarial models attempt to predict the protected class status of an individual based on the output of the underlying model, with the underlying model continuing to adjust until the ability of the adversary to correctly predict protected class characteristics diminishes to an appropriate point.

Adversarial examples: Adversarial examples refer to an example-based explainability method by which weaknesses or failures of a model are identified through changes in the underlying data that cause the model to make an incorrect prediction.

Adverse action: An adverse action is a credit decision in which a lender declines to provide credit in the amount or terms requested or makes a negative change to an existing account. Federal law requires lenders to provide disclosures to consumers and small businesses after taking an adverse action to explain the principal reason(s) for the decision.

Alternative financial data: Alternative financial data are a type of credit information that describe a variety of non-lending financial activities and can be extracted relatively easily from sources such as bank or prepaid card accounts. Depending on the source and scope of data, this information may contain more granular and timely information about applicants' financial position than credit bureau information and can provide a more complete picture of an applicant's ability and willingness to repay a loan.

Attribute: An attribute generally refers to a variable or feature included in a dataset. This could include input variables, such as an individual's income, as well as a target or output variable (such as whether an individual is likely to default on a loan).

Bagging: Bagging or bootstrap aggregation is a technique used to make random forest models or gradient-boosted decision trees less biased and more accurate than individual decision trees. This approach averages the predictions of various individual decision trees that are each trained on a different subsample of observations in the training data.

Behavioral data: Behavioral data are a type of credit information firms may use in the context of credit underwriting or for other purposes such as marketing. These data include a range of possible information (such as the date, time, or place of a transaction), digital activities such as search histories, or social media data.

Bias: Bias is commonly defined by statisticians and data scientists as the variance between a model's predictions and actual outcomes. Other stakeholders use bias to refer to discrepancies across different demographic groups, especially for those groups which have been subject to discrimination or injustice of other forms.

Cash-flow data: Cash-flow data are a type of alternative financial data that shows income, expenses, and other reserves. Cash-flow data can be derived from bank and prepaid accounts, small business accounting software, and other sources.

Classification problem: A classification problem generally refers to a situation in which a target variable is categorical or binary (such as sorting credit applicants into high-risk or low-risk buckets), in contrast to a regression problem where the output is a continuous variable such as a score.

Conceptual soundness: Conceptual soundness involves an assessment of the quality of a model's design and construction as required by regulatory guidance on model risk governance. Evaluations of conceptual soundness ensure that all processes utilized to develop the model are documented thoroughly, that such documentation supports how the model operates, and that the choices made for the model are themselves supported by analysis and testing. The theoretical construction, key assumptions, data, mathematical calculations, and the usage and purpose of the data and model must all be documented.

Counterfactual explanation: A counterfactual explanation is an example-based explainability technique. Counterfactual explanations describe how much a particular data point would have to change in order to change the predicted outcome. In other words, if someone is denied credit, a counterfactual explanation will search for the smallest possible change to the factors assessed in the underwriting analysis that would cause the model to predict the applicant would not default.

Cluster analysis: Cluster analysis is an unsupervised learning technique in which observations—applicants in the case of credit underwriting—with similar attributes are grouped together in segments, without including a target variable such as loan defaults. Cluster analysis can be used for such purposes as segmenting customers for marketing purposes or creating groups of existing customers based on their spending behaviors and is useful when knowledge about actual lending outcomes is unknown.

Credit information: Credit reporting agencies provide credit applicants' personal information; public records such as bankruptcies; tradeline data which reflect an applicant's repayment record mainly for secured and unsecured loans; inquiries made on the applicant's credit files; and balance information (including available balance for credit cards) for use in lending and securitization of consumer loans.

Credit scorecard: A credit or underwriting scorecard refers to a method of modelling credit risk that converts various characteristics of an applicant's credit history (such as default history or debt-to-income ratio) to a point value and then sums these values into a total credit score that signifies an applicant's likelihood of default.

Data drift: Data drift can occur when the underlying conditions of a model's training data differ from the data it is using to model future predictions. In underwriting, this can occur when economic conditions at the time that a model is deployed differ significantly from those reflected in the training data.

Decision tree: A decision tree is a model that uses a hierarchical structure to estimate a target variable with a series of discrete, binary decisions. Beginning with a decision that separates the data into two or more subsets, each smaller decision is represented in a chain where each step of the chain corresponds to a simple "if-then" decision. This series of analyses eventually leads to an estimation of the target variable.

Decorrelation: Decorrelation is a technique used to make random forest models less biased and more accurate than individual decision trees. This approach randomly selects a subset of features from which to select at each decision point in the tree.

Deep learning: Deep learning is a form of machine learning that emulates the workings of the human brain by transforming input data through multiple layers of neural networks to identify complicated patterns and connections between input data and the target variable. Neural networks are a form of deep learning used for underwriting models.

Deployment: Deployment refers to the stage in the model lifecycle when a machine learning underwriting model is put into use to evaluate applications from consumers and make credit decisions.

Disparate impact: Disparate impact is one of two theories for establishing legal liability for discrimination against groups protected under the Equal Credit Opportunity Act (ECOA) or Fair Housing Act (FHA). It prohibits the use of facially neutral practices that have a disproportionately adverse effect on protected classes, unless those practices meet a legitimate business need that cannot reasonably be achieved through less discriminatory means.

Disparate treatment: Disparate treatment is one of two theories for establishing legal liability for discrimination against classes of persons protected under the Equal Credit Opportunity (ECOA) Act or Fair Housing Act (FHA). It prohibits treating individuals differently based on a protected characteristic. Establishing disparate treatment does not require any showing that the treatment was motivated by prejudice or a conscious intention to discriminate.

Equal Credit Opportunity Act (ECOA): The Equal Credit Opportunity Act of 1974 is a federal statute (codified at 15 U.S.C. § 1691 *et seq*.) that makes it unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction, on the basis of race, color, religion, national origin, sex, marital status, or age (provided the applicant has the capacity to contract); to the fact that all or part of the applicant's income derives from a public assistance program; or to the fact that the applicant has in good faith exercised any right under the Consumer Credit Protection Act. ECOA is implemented by the Consumer Financial Protection Bureau through Regulation B (codified at 12 C.F.R. Part 1002).

Explainability: In this report, model explainability refers to the ability of various stakeholders to understand how or why a particular decision was made or result was reached.

Explainability techniques: Explainability techniques are supplemental models, methods, and analyses used to improve the transparency of complex models. Since these tools are used after the model has been trained, they are often referred to as *post hoc* or indirect techniques. These methods do not generally affect the design or operation of the underlying model and can be used with a variety of machine learning model types.

Fair Credit Reporting Act (FCRA): The Fair Credit Reporting Act is a federal statute (codified at 15 U.S.C. § 1681 *et seq*.) enacted to protect consumers from the willful and/or negligent inclusion of inaccurate information in their credit reports and to promote the accuracy, fairness, and privacy of consumer information contained in the files of consumer reporting agencies. FCRA regulates the collection, dissemination, and use of consumer information for credit purposes as well as for activities such as employment, insurance, and housing. It is implemented by the Consumer Financial Protection Bureau through Regulation V (codified at 12 C.F.R. Part 1022).

Fair Housing Act (FHA): The Fair Housing Act refers to Titles VIII and IX of the Civil Rights Act of 1968 (codified at 42 U.S.C. § 3601 *et seq*.), which prohibit discrimination concerning the sale, rental, and financing of housing based on race, religion, and national origin. These prohibitions were subsequently extended to include discrimination based on sex, disability status, and family status. The Department of Housing and Urban Development implemented a portion of the FHA through a rule prohibiting practices with disparate impact.

Feature: Feature refers to the variables in a dataset used to predict a target variable. This term is often used synonymously with input variable or independent variable and represented in mathematical notations as X.

Feature engineering: Feature engineering refers to various methods of preparing data for training in order to maximize the accuracy of the model, such as binning numerical variables into various ranges, creating ratios or other latent features that reflect the relationship between multiple inputs, or imputing values for missing data.

Feature importance: Feature importance refers to how much impact an input variable has on the target variable in a model. Various *post hoc* explainability techniques are designed to identify and quantify feature importance within more complex models.

Feature selection: Feature selection refers to the process of determining which attributes in the dataset should be used in the machine learning model.

Fitness-for-use: Fitness-for-use refers to the effectiveness of a model in serving its purpose, which can include model accuracy, fairness, and other factors, and the quality of the plan to appropriately manage risks related to operation of a particular model. Model risk management expectations require firms to determine that a model is fit for use prior to deployment.

Global explainability: Global explainability refers to the ability to identify a model's high-level decision-making processes and is relevant to evaluating a model's overall behavior and fitness-for-use.

Gradient-boosted decision trees (GBDTs): Gradient-boosted decision trees are a form of machine learning that combines multiple decision trees, each of whose target variable is the prediction error rate of the tree that came before. The weighted sum of each tree's predictions gives the model's final prediction.

Hyperparameter: Hyperparameters refer to aspects of a machine learning model that are not learned from the data, but rather are determined by model developers, such as the number of nodes in a decision tree. Hyperparameters can affect the predictiveness and explainability of the model and are often adjusted during model tuning.

Individual Conditional Expectations (ICE) Plots: Individual Conditional Expectation plots are common visualization methods used in model development and are used as a feature importance explainability technique. These plots provide insight into feature interactions by displaying the relationship between each individual input and its predicted outcome. ICE plots show each instance or person in the dataset as a single line, where the value of the feature of interest varies.

Inherently interpretable models: An inherently interpretable model specifies the contribution that each input variable makes toward the output and enables stakeholders to understand its predictions without the use of secondary models, analyses, or methods. These models are also sometimes referred to as self-explanatory.

Input variable: Input variables refer to the variables in a dataset used to predict a target variable. This term is often used synonymously with feature or independent variable and represented in mathematical notations as X.

Integrated gradients: Integrated gradients are a feature importance explainability technique used to explain outputs from models such as a neural network or logistic regression where the change (or derivative) in output can be easily calculated. The gradients are summed to identify which feature has the most significant effect on the model's predicted output, such that features with greater summed gradients have more importance to the model output.

Interpretability: Model interpretability refers to the ability to understand a model's operations based largely on its formal notation and without reliance on secondary models, analyses, or methods. To be interpretable, a person should be able to infer the following: (1) the types of information or input variables that a model uses, (2) the relationship between the input variables and the model's predictions or outputs; and (3) the data conditions for which the model will return a specific result (for example, to receive a credit score of 600, weekly income has to be at least $600).

Latent feature: Latent features are generated by a machine learning algorithm from variables in the dataset and serve as internal or interim analyses that help determine the model's prediction. These can be derived through simple combinations of different attributes or more complex mathematical processes. In general, the greater the number of the latent features and the more difficult those relationships are to describe on their own, the more complex the model will be.

Linearity: In linear models, changes in a particular input produce a consistent rate of change in the output.

Linear regression: Linear regression refers to a statistical technique where a modeler or algorithm locates the best-fit linear relationship between input variables and a target variable.

Local explainability: Local explainability refers to the ability to identify the basis for specific decisions made by a model.

Local Interpretable Model-Agnostic Explanations (LIME): LIME is a feature importance explainability technique that uses local linear surrogate models around a particular data point to approximate a complex model's output. The resulting local surrogate models are used to both explain the model's behavior around individual data points and quantify feature importance for the overall model. LIME is generally used today as a baseline to compare the outputs and performance of other explainability tools against or to generate insight into feature importance.

Logistic regression: Logistic regression refers to a statistical technique where a modeler or algorithm locates the best-fit curve between input variables and a target variable.

Model debiasing: Model debiasing refers to a range of methods to reduce bias in a model's predictions, either by transforming the input data, building a debiasing function into model training, or transforming a model's output. Debiasing techniques vary based on the model's use case, the data being used, model complexity, and other factors.

Monotonicity: Monotonicity refers to a relationship that is one-directional (*e.g.*, increasing the value of an input variable will always cause the output to increase or will always cause the output to decrease). Imposing monotonicity constraints can help model developers limit the complexity and improve the explainability of machine learning models.

Neural network: Neural networks are a form of deep learning that consist of several hidden layers through which a model learns nonlinear patterns between features and the target variable. The model uses these patterns to create new features from the input variables in each layer, ultimately arriving at the final layer, where a prediction is made.

Non-financial alternative data: Non-financial alternative data refers broadly to data about a person's activities that are not financial in nature or derived from financial data. Examples of such data include social media data, search histories, educational attainment, and mobile phone recharging habits.

Overfitting: Overfitting occurs when a model is fitted too narrowly to the training data, which can hinder its accuracy when deployed if test or deployment data reflect conditions different than those observed in the training data.

Parameter: Model parameters are settings in the model that are determined using the training data and which are fitted to the model. When the training is initialized, the parameters are usually set to a random value (or zero). As training progresses, these random values are updated using an optimization algorithm, which performs a search through possible parameter values to learn and update the values. The final parameters that are determined at the end constitute the trained model. Examples of parameters are coefficients (or weights) of linear and logistic regression models, and in the case of neural networks, the parameters are the weights and the biases.

Partial Dependence Plots (PDP): Partial dependence plots (PD plots or PDPs) are common visualization methods used in model development and are used as a feature importance explainability technique. These plots depict a feature's effect on a model's predicted results. PD plots provide a global interpretation of more complex models.

Protected class: Like anti-discrimination statutes applicable in other areas, ECOA and FHA prohibit discrimination against people based on a common characteristic. Such characteristics include race, color, religion, national origin, sex, marital status, disability status, family status, or age (provided the applicant has the capacity to contract); reliance on a public assistance program; or the good faith exercise of any right under certain federal consumer financial laws.

Random forests: Random forest models combine multiple decision trees into a single predictive model to decrease variance and bias and/or to improve accuracy and predictions.

Regression problem: A regression problem generally refers to a situation in which the target variable is continuous, such as a model that assigns a numeric score. These contrast with classification problems where the output variable is categorical or binary.

Regularization: Regularization is a regression method that limits the coefficient estimates of irrelevant features in a model to zero or near zero. It is commonly used to reduce model complexity, manage overfitting risks, and to debias models. Regularization adds new terms—often called penalty terms—to the objective function in order to guide the optimization process toward a specified solution. For example, a model developer training an underwriting model which can draw on 1,000 features as inputs wants the final model to use a small number of features for predictions because that will simplify the task of producing explanations. To do this, he or she will add a regularization term to the objective function, which adds a penalty for each additional feature used by the model.

Reinforcement learning: Reinforcement learning trains a model on unlabeled data, identifies an action for each variable, and receives input from a human or other model that helps the model learn. It is generally not applicable to credit underwriting.

Reject inference: Reject inference is an approach used by model developers to address biases that result from the absence of loan performance data for past applicants who were rejected or declined offers of credit. It uses data for approved applicants to statistically impute predicted values on those who were denied credit, which are then added to historical information for approved applicants to train an underwriting model.

Robustness: Robustness refers to a model's ability to make accurate predictions in conditions that differ from the conditions existent in the model's training data.

Semi-structured data: Semi-structured data refers to data that are not stored in a relational database, but still retains some structure. An example is bank account transaction records from banks' online platforms that are obtained by screen scraping. Semi-structured data generally requires more cleaning than structured data prior to use.

Shapley Additive Explanations (SHAP): Shapley Additive Explanation is a feature importance explainability method that is used to explain complex models. SHAP does this by indicating the contributions of particular features in changing a model's outcome. It is similar to LIME in that it explains a model locally. This method measures feature importance by removing features from a data point and quantifying how much that affects the model's output.

Sparsity: Sparsity refers to the limiting of features or input variables that a model relies on to make predictions, such as through removing an input variable when it is highly correlated with another variable. Sparsity is one way to improve model transparency.

Structured data: Structured or tabular data refers to data that are stored in a database in columns and rows.

Surrogate models: Surrogate models refer to interpretable models that mimic and explain the behavior of more complex models.

Supervised learning: Supervised learning refers to a machine learning model that is trained using data that contains both inputs (*e.g.,* prior bankruptcies) as well as a target measure (*e.g.,* whether an individual ultimately defaulted on his or her loan). This is the most common learning approach used in credit underwriting.

Support vector machines (SVMs): Support vector machines generate a best-fit separating line between observations in a dataset that belong to different classes.

Target variable: A target variable is the dependent variable or output variable that a machine learning model predicts.

Training: Training refers to the stage in the model lifecycle when a learning algorithm analyzes data to identify relationships and rules relevant to predicting a specific target variable.

Training data: Training data refers to the data that is fed into and analyzed by a learning algorithm to produce a predictive model.

Transparency: Model transparency refers to the ability of various stakeholders in a model, including its developers, risk managers, and regulators, to access the information they need related to the model's design, use, and performance. Model transparency is generally thought of as being necessary to establish the trustworthiness of models and is important in certain use cases to evaluate and document regulatory compliance. Transparency can potentially be achieved through constraints that make a model more interpretable, *post hoc* explainability techniques, or a combination of both.

Tuning: Tuning refers to the stage in the model lifecycle that involves adjusting hyperparameters of a model to maximize performance. It can occur in conjunction with validation and testing.

Unlabeled data: Unlabeled data refers to data that excludes the target variable and can be used in models that use unsupervised or reinforcement learning.

Unstructured data: Unstructured data are neither organized in a particular way nor stored in a database, including information stored in text or image formats. Examples include information stored in text formats, audio files, video files, and images, which includes most social media data.

Unsupervised learning: Unsupervised learning detects patterns in data without the inclusion of a target variable and can be used to find similarities between observations in a dataset. Such learning is commonly used for purposes of customer segmentation in marketing as well as for image recognition models.

XGBoost: Extreme Gradient Boosting (XGBoost) is a type of tree-based machine learning model that is generated using an open-source package in both R and Python that relies on gradient boosting and is popular for use in developing underwriting models. The package has been enhanced to expedite the model training process by addressing overfitting risk, removing irrelevant information from the model, imputing missing values, and applying explainability techniques.

APPENDIX B

Legal and Regulatory Considerations

As discussed in Section 2, questions about how existing prudential and consumer protection laws and regulations apply to machine learning underwriting models are a source of uncertainty for lenders and other stakeholders. This appendix provides an overview of the core purposes and requirements in three such areas: model risk management, fair lending, and adverse action reporting.[250] It also summarizes key debates related to those expectations in the context of machine learning underwriting models.

B.1 Model Risk Management

Federal prudential regulators have issued extensive guidance outlining their expectations for steps that banks should take in developing, monitoring, and using models throughout all aspects of their operations. This guidance applies broadly to the range of model use cases that might create unexpected losses, compliance problems, or other negative outcomes for the firm and calls for enterprise-wide risk management processes including governance, policies, and controls.[251] Expectations are calibrated to the degree of risk posed by the particular use case, and credit underwriting is often considered to be among the highest risk activities depending upon the composition of the particular firm's business. Thus, for financial institutions subject to prudential oversight, these expectations typically require extensive pre-deployment review of credit models and monitoring during use, especially for firms that emphasize retail or consumer banking. For other financial institutions, bank regulatory expectations may broadly inform aspects of their model oversight practices in part because funding and securitization counterparties may require some of these processes and practices.

The guidance defines a model to include any "quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates." Similarly, model risk includes "the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports," which can lead to financial loss, poor business and strategic decision-making, or damage to a bank's reputation. Banks are expected to identify potential sources of such risk, assess their magnitude, and mitigate appropriately, both at the individual model level and in the aggregate across business lines and legal entities. In particular, higher degrees of risk management are expected where particular models pose greater risk—whether that risk derives from the model's potential impact on customers or the firm, methodology, complexity, data usage, operational structure, or other factors. Where there is higher

[250] The appendix excludes consideration of how prohibitions on unfair, deceptive, and abusive acts or practices might be applied to the use of machine learning underwriting models, including to scenarios in which models lack sufficient transparency. *See generally* 15 U.S.C. § 45(a)(1); 12 U.S.C. § 5531.

[251] FRB, SR 11-7; OCC, Bulletin 2011-12; FDIC, FIL 22-2017.

uncertainty about a particular model—about its inputs, assumptions, or methodology—model risk management programs will generally require heightened scrutiny before approving that model for use and more vigorous oversight of its operations.[252]

Because a wide variety of factors can cause financial loss, expose firms to regulatory sanction, or damage to a bank's reputation, regulators expect model developers to consider and plan for the full range of risks related to the model's use case. In the context of a credit underwriting model, this will typically include consideration of risk management issues related to both prudential and consumer protection requirements. For example, the risk that a model's predictions will start to deteriorate due to changes in economic conditions is one component of model risk management, but the framework also incorporates risks from cybersecurity threats, data quality and scope issues, changes in customer behaviors and applicant pools or behaviors, compliance with more specific legal and regulatory requirements (including for consumer credit fair lending and adverse action notice disclosures as discussed further below), and more general reputational issues.

The prudential model risk management expectations emphasize various aspects of model transparency. At a broad level, the guidance requires documentation of the processes by which a model is developed, validated, and monitored during deployment.[253] More specifically, the guidance creates an expectation that developers will evaluate whether models are relying on relationships in the data that are intuitive and defensible with regard to the outcome that they are attempting to predict,[254] that firms will conduct appropriate sensitivity analyses to establish the fitness of the model for use,[255] and that lenders will establish appropriate processes for identifying and mitigating risks relevant to the model's use, including compliance with applicable consumer protection laws.[256]

In practice, the sophistication, scope, and resources of model risk management programs varies significantly across the banking sector. Many of the largest banks typically have specialized teams with the expertise and infrastructure not only to conduct comprehensive reviews of the documentation submitted for validation, but also to develop and test their own models from the training data where warranted. In these firms, model developers can expect to defend every significant decision in model design and development prior to putting the model into use. By contrast, at smaller firms, model validation may be conducted by vendors or consist of a relatively streamlined peer review.

B.1.1 Key Issues

While the model risk management guidance generally provides a principles-based framework that can be adapted to a wide variety of firms, models, and particular circumstances, stakeholders are currently debating whether elements need to be modified or expanded upon to provide more guidance to address issues raised by the use of machine learning in various applications, including

[252] FRB, SR 11-7 attachment at 3-4.

[253] *Id.* attachment at 21 ("Without adequate documentation, model risk assessment and management will be ineffective. Documentation of model development and validation should be sufficiently detailed so that parties unfamiliar with a model can understand how the model operates, its limitations, and its key assumptions.").

[254] *Id.* SR 11-7 (evaluating conceptual soundness involves assessing "documentation and empirical evidence supporting the methods used and variables selected for the model [to] ensure that judgment exercised in model design and construction is well informed, carefully considered, and consistent with published research and with sound industry practice."); *id.* attachment at 6 ("Developers should be able to demonstrate that such data and information are suitable for the model and that they are consistent with the theory behind the approach and with the chosen methodology."); *id.* attachment at 11 ("Key assumptions and the choice of variables should be assessed, with analysis of their impact on model outputs and particular focus on any potential limitations. The relevance of the data used to build the model should be evaluated ….").

[255] *Id.* attachment at 11-13 (stating that sensitivity analyses should be performed during both development and deployment).

[256] *Id.* attachment at 17-18.

credit underwriting.[257] For example, some stakeholders have suggested that some of the guidance's language concerning sensitivity testing may not be sufficiently calibrated to how such processes are performed in machine learning contexts.[258] Others have suggested there are potential discrepancies in the standards applied to traditional linear models and more complex AI and machine learning models.[259]

As discussed in Section 2.3.1, one of the ways in which the transition to machine learning poses a particular transparency-related issue concerns lenders' efforts to detect in timely ways conditions that may reduce the accuracy of machine learning models given their tendency to overfit to training data. The emergence of various *post hoc* explainability techniques may enable lenders to improve their ability to recognize and respond to conditions in which the performance of machine learning underwriting models might rapidly deteriorate. In validation processes, this points to several additional potential inquiries: what approach has the model developer taken to enabling transparency, does that approach confer appropriate levels of transparency in practice, and how can the reliability and trustworthiness of information produced to explain the model be evaluated.

B.2 Fair Lending

Lenders are subject to broad anti-discrimination requirements regardless of the type of model they use to predict an applicant's likelihood of default. The Equal Credit Opportunity Act (ECOA) prohibits discrimination in "any aspect of a credit transaction" for both consumer and commercial credit on the basis of race, color, national origin, religion, sex, marital status, age, or certain other protected characteristics.[260] The Fair Housing Act (FHA) prohibits discrimination on many of the same bases in connection with residential mortgage lending.[261]

Fair lending enforcement actions and lawsuits are generally brought on two grounds.[262] Disparate treatment focuses on whether creditors have treated applicants differently based on protected characteristics, and generally prohibits consideration of race, gender, or other protected characteristics in underwriting and scoring models.[263] While intentional discrimination is pursued under this theory, disparate impact addresses lenders' use of facially neutral practices that have a disproportionately negative effect on protected classes, unless those practices meet a legitimate business need that cannot reasonably be achieved through less impactful alternatives.[264]

[257] Further, the kinds of fair lending issues considered below will be important substantive model risk management considerations for underwriting models.

[258] Zest AI, Here's How ML Underwriting Fits Within Federal Model Risk Management Guidelines (May 30, 2019).

[259] Model Risk Managers' International Association, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, 19 (May 25, 2021).

[260] 15 U.S.C. § 1691(a) (also prohibiting discrimination based on the receipt of public assistance and the good faith exercise of certain rights under federal consumer financial law).

[261] 42 U.S.C. § 3605 (prohibiting discrimination on the basis of race, color, national origin, religion, sex, familial status or disability).

[262] The Supreme Court has confirmed that both doctrines are available under the Fair Housing Act, but has not yet ruled on whether disparate impact analysis applies under ECOA. Texas Dep't of Housing & Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015). Federal regulations, agency guidance, and lower court decisions have recognized the doctrine under ECOA for decades, in part based on legislative history. See, *e.g.*, 12 C.F.R. § 1002.6(a); *id.* Supp. I, cmt. 1002.6(a)-2.

[263] Federal law does allow lenders to consider factors such as whether an applicant is of sufficient age to form binding contracts under state law and whether state laws regarding marital property affect their ability to repossess collateral. 15 U.S.C. § 1691(b). Models can also use applicants' age as a predictive variable under narrowly restricted circumstances involving "an empirically derived, demonstrably and statistically sound, credit scoring system" if the model does not assign a negative value to the age of older applicants. *Id.* § 1691(b)(3); 12 C.F.R. § 1002.6(b)(2).

[264] For a general overview of the two theories and the ways that they overlap, *see* Evans.

Disparate impact analysis typically follows a three-part test that was developed in the context of employment law:[265]

- **Adverse Impact:** A plaintiff (such as a consumer or a regulatory agency) must make an initial showing that a particular act or practice causes a disproportionate adverse effect on a prohibited basis. In the credit context, this is typically analyzed by looking at whether use of particular variables or other lending practices cause approval rates or pricing patterns to differ by race, gender, or other protected characteristics.
- **Business Justification:** In response, the creditor must then show that the practice furthers a legitimate business need, such as whether the variable helps to predict the risk of default.
- **Less Discriminatory Alternative:** In response, to prevail on a claim, the plaintiff must demonstrate that the legitimate business need cited by the creditor can reasonably be achieved by using an alternative practice that would have less adverse impact.

Factors that are closely correlated with protected class characteristics (which are often called proxies) can be relevant to both disparate treatment and disparate impact analyses. For instance, since collecting data about protected class status and factoring such information into a credit underwriting model are generally prohibited, a party that intends to discriminate against a particular group might incorporate a factor that is closely related to a protected characteristic into its model instead. A factor that is correlated with protected class status may also have a disproportionately negative effect on approvals or pricing among different protected classes, leading to further inquiry under disparate impact theories as to what need is being served by use of the factor and whether another variable could be substituted in its place. Accordingly, both doctrines as applied to automated underwriting models focus in substantial part on analyzing data inputs.

Statistical tests can also be important under both theories, and more generally when lenders set out to evaluate their degree of fair lending compliance risk in adopting or changing their underwriting models. However, case law and regulatory guidance do not provide precise mathematical thresholds for determining the level of problematic disparities. For instance, while federal agencies in the employment context have sometimes used a rule of thumb that hiring rates for women and applicants of color should be at least 80% of the rates for men and Whites, respectively,[266] that benchmark has not been formally recognized in financial services. Financial regulatory guidance concerning what constitutes a legitimate business need focuses on whether there is a "demonstrable relationship" between the variable or requirement and credit risk but does not specify particular quantitative evaluation methodologies or thresholds.[267] As a result, firms make decisions about tradeoffs between reducing disparities and negatively affecting model performance based on their own business judgment and risk tolerance.

[265] In litigation, the burden shifts back and forth between the parties to make particular showings at each stage. However in other contexts, such as where a lender's compliance team is applying this analysis to monitor its fair lending risk, one party will perform each of the steps.

[266] *See*, *e.g.*, Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor, & Department of Treasury, Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11996 (Mar. 2, 1979).

[267] 12 C.F.R. pt. 1002, Supp. I, cmt. 1002.6(a)-2; *see also* Office of the Comptroller of the Currency, Bulletin 1997-24, app. at 11 (May 20, 1997) (focusing on whether credit scoring variables are statistically related to loan performance and have an understandable relationship to creditworthiness).

B.2.1 Key Issues

While existing fair lending law provides a conceptual framework for considering the fairness of machine learning underwriting models that is lacking in some other sectors, stakeholders are debating whether additional guidance is needed to address issues raised by machine learning techniques and more fundamentally whether certain existing approaches can work in the context of machine learning models.

For instance, the identification and management of variables that may proxy for protected class status under both theories of discrimination can be significantly more complicated when lenders use machine learning underwriting models, particularly where those models may use data from more varied sources or in more complex or unintuitive features. Machine learning models may also effectively reverse-engineer protected class status from correlations in data, even though consideration of such status is prohibited.[268] Thus, lenders and regulators may need new tools and face new limitations in efforts to diagnose bias.[269]

In this context, some stakeholders are questioning whether reconsidering the prohibition on the use of protected class characteristics could improve the fairness and accuracy of machine learning underwriting models.[270] In its broadest form, this could involve use of protected class information to develop race-, gender- or age-specific underwriting models,[271] although more research is needed to understand the fairness and other effects of such approaches generally and in the context of diversified credit information.[272] A less sweeping reconsideration could enable use of protected class characteristics to debias machine learning underwriting models prior to their deployment.[273] In one approach, adversarial models—models designed to estimate an applicant's protected class characteristics based on the underlying model's prediction—have proven generally effective in decreasing

[268] *See* Gabbrielle M. Johnson, Proxies Aren't Intentional, They're Intentional, 2-4 (2021) (Unpublished manuscript) (arguing that machine learning algorithms have the capacity to "learn," be "aware" of, and make decisions on the basis of protected class characteristics by picking up on redundant encoding in the data and using proxies to meaningfully reason about or explicitly represent protected class characteristics, even when those characteristics are not available or provided as model inputs); *see also* Gillis.

[269] For overviews of some of the issues raised by both data and machine learning models, *see* Evans; Federal Trade Commission, Big Data at 27-32; Barocas & Selbst; Gillis; Gillis & Spiess; Hellman.

[270] Hellman, at 865 (recognizing that "[i]f algorithms use protected traits in a limited way to determine which other traits to consider within the algorithm, overall accuracy can be improved"); Model Risk Managers' International Association ("Tests that rely in the correlation of last names to ethnicity are weak and will only become weaker, Further, they say little about a range of discrimination risks. The only possible solution is for the government to change the laws around collecting protected class status. This is the single greatest obstacle to using machine learning on alternate datasets."); *see also* Barocas & Selbst; Jason R. Bent, Is Algorithmic Affirmative Action Legal? 108 Georgetown L. J. 803-853 (2020) (considering statutory and constitutional arguments for "race-aware affirmative action in the design of fair algorithms").

[271] Hellman at 846-864, 865-866 (recognizing that Constitutional law does not rule out using protected class characteristics in a limited way that may help to determine how courts should evaluate the use of race in algorithms when racial classifications are used to improve overall accuracy).

[272] A recent study in the context of mortgage lending modelled the predictiveness gains from improving the robustness of credit files and compared those gains to the performance of group-specific underwriting analyses. This evaluation suggests that inaccurate predictions of creditworthiness of low-income applicants and other historically underserved groups result more from statistical noise in available credit information—errors that might cause a credit score to over- or understate the risk of default in different cases—rather than errors that are easier to fix with modelling adjustments because they skew predictions in a single direction. The authors find that applying more specialized modelling analyses to traditional data sources may not overcome the challenges of providing credit to underserved groups, and instead emphasize the potential benefits of improving and expanding data sources. Blattner & Nelson.

[273] A coalition of consumer advocacy groups and civil rights organizations have requested that the federal financial services regulators provide more detailed regulatory guidance on various techniques that have the potential to improve model fairness and their compliance with fair lending laws, including the use of protected class information in model training as a way to decrease discrimination in lending and other financial services. National Fair Housing Alliance, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (July 1, 2021).

bias in certain circumstances, although more research is needed in financial services applications such as credit underwriting.[274] These techniques are discussed further in Section 5.3.

B.3 Adverse Action Notices

ECOA and the Fair Credit Reporting Act both require lenders to disclose their reasons for denying credit applications or taking other "adverse actions," and FCRA requires similar notices when lenders offer less favorable terms to consumers based on information in their credit reports.[275] These disclosures, which are commonly referred to as adverse action notices, are outcome-based explanations that provide "a description of the facts that proved relevant to a decision, but not a description of the decision-making rules themselves."[276]

The requirements were adopted in the 1970s as part of broader efforts to promote the correction of errors in credit reports and to prohibit discrimination, but were controversial due to concerns about burdens on creditors.[277] The laws and implementing regulations give lenders substantial latitude as to how they determine which factors to highlight, and do not require them to explain how the factors affected the lenders' decisions. For instance, while the "principal reasons" provided under ECOA must "relate to and accurately describe the factors actually considered or scored by a creditor," regulatory guidance states that the disclosure need not describe how or why a factor adversely affected an applicant or how a factor considered in a credit scoring system relates to predictions of creditworthiness. The guidance also emphasizes that the law does not require any particular method for determining which factors relating to a credit scoring model should be listed.[278]

In practice, most lenders rely heavily on a list of sample reasons that is provided in an appendix to the ECOA regulations to satisfy the requirements of both statutes.[279] Where lenders rely on a third-party credit scoring model to make underwriting decisions, the "reason codes" are often generated automatically by consumer reporting agencies and model developers to simplify production of the notices. For lenders' proprietary models, methodologies for determining which reasons to list

[274] With the aim of mitigating gender bias, a recent study created an adversarial model using census data to predict whether an individual fell into one of two income brackets and found that this method only slightly decreased the overall accuracy of the predictive model while nearly achieving equality of odds between males and females. *See* Zhang *et al.*

[275] The laws define "adverse action" to include denials of credit applications on substantially the same terms and in substantially the same amount as requested, unless the lender makes a counter-offer. Adverse actions also include unfavorable decisions on existing credit arrangements, such as negative changes in terms, denials of line increases, and reductions or cancellations of credit lines. 15 U.S.C. §§ 1681a(k)(1), 1691(d)(6). In 2011, a FCRA amendment took effect to require similar risk-based pricing notices where credit terms are "materially less favorable" than the terms granted to a "substantial proportion" of other consumers. 15 U.S.C. § 1681m(h); 12 C.F.R. §§ 222.70-.75. ECOA's disclosure requirements apply to both consumer and commercial credit, although some details are different for business applicants. Federal agencies have excluded business credit from FCRA's disclosure requirements. 15 U.S.C. § 1681a(c); 12 C.F.R. §§ 222.70(a)(2), 1002.9(a).

[276] Selbst & Barocas at 1100.

[277] Adverse action notices were a part of FCRA's initial framework in 1970 to empower consumers to correct errors in their credit reports. In 1976, ECOA was amended to require lenders to provide both a statement that discrimination is prohibited by law and a specific description of the "principal reason(s)" for taking the adverse action. 15 U.S.C. § 1691(d). For historical background, *see, e.g.*, David C. Hsia, Credit Scoring and the Equal Credit Opportunity Act, 30 Hastings L. J. 371 (1978); Ralph J. Rohner, Equal Credit Opportunity Act, 34 Bus. Law. 1423 (1979); Winnie F. Taylor, Meeting the Equal Credit Opportunity Act's Specificity Requirement: Judgmental and Statistical Scoring Systems, 29 Buff. L. Rev. 73 (1980).

[278] 12 C.F.R. § 1002.9(a)(2), (b); *id.* pt. 1002, supp. I, cmt. 9(b)(2)-2, -3, -4, -5. Similarly, where a lender has taken adverse action based on information in a consumer's credit report, the FCRA requires disclosure of "key factors," which are defined as "relevant elements or reasons adversely affecting the credit score for the particular individual, listed in the order of their importance based on their effect on the credit score." However, the law does not define a methodology for determining relative importance or effects. 15 U.S.C. § 1681g(f)(1), (2)(B), (9). FCRA generally limits the number of key factors to be disclosed at four; regulatory guidance under ECOA indicates that more than four reasons are rarely helpful. *Id.* §§ 1681g(f)(1), (9); 12 C.F.R. pt. 1002, supp. I, cmt. 9(b)(2)-1.

[279] 12 C.F.R. pt. 1002, App. C. The list is based on historically common underwriting factors and actually does provide some explanation for many of the items listed, such as "Income insufficient for amount of credit requested," "Insufficient number of credit references provided," and "Unacceptable type of credit references provided." Others such as "Length of employment" and "Length of residence" are more general. *Id.*

on an adverse action notice may vary depending on the circumstances. For example, the regulatory guidance specifically permits lenders to benchmark against either applicants whose total score was at or slightly above the minimum passing score or against the average for all applicants in determining the factors on which the individual applicant performed least well, and specifically notes that other methodologies may be acceptable.[280]

B.3.1 Key Issues

Although the applicable laws provide substantial flexibility to firms, lenders report that uncertainty about complying with adverse action requirements does shape and sometimes chill adoption of nontraditional data sources and machine learning methodology.[281] Explaining particular variables that are influential in machine learning models can be difficult where the models develop and rely on relationships that are inherently complex, non-intuitive, difficult to assess, large in number, or dependent on other input variables or relationships. And while the existing list of sample reasons provides some high-level wording with regard to traditional underwriting factors such as income and past credit defaults, it has not been updated to address less traditional data sources such as analyzing balance patterns in consumers' checking or other transaction accounts.

However, as data scientists and other stakeholders have worked to facilitate the generation of adverse action notices for machine learning models and new data types, they have helped to fuel policy conversations about how to make the notices more useful to consumers. The discussion has focused in particular on providing more actionable information to highlight ways that applicants can change their financial behavior to increase the likelihood of more favorable credit decisions in the future. Some stakeholders argue that the growth of open-source and other tools for more transparent and interpretable machine learning models have given lenders new options to satisfy adverse action reporting requirements in a more effective way.[282]

[280] 12 C.F.R. pt. 1002, supp. I, cmt. 9(b)(2)-5.

[281] Parrish, Alternative Data and Advanced Analytics (reporting that surveyed industry executives view adverse action disclosures as the most significant challenge for using AI and machine learning in underwriting); Knight.

[282] BLDS, LLC *et al.* at 9, 15-18.

APPENDIX C

Additional Fairness Metrics

As highlighted in Section 5.2, academic and industry research on machine learning has produced a robust list of mathematical approaches to measuring the fairness of algorithmic models. In the context of lending, these metrics can potentially function as part of model developers' toolkits at various ages of the model development process or be used in the context of fair lending analyses. Tables C.1-C.3 supplement the discussion of a subset of these metrics in the report to provide a more complete sense of this body of research. Many of the same limitations explored in Section 5.2 also apply to these options.

This appendix provides an overview of 21 fairness metrics (Tables C.1-C.3),[283] including narrative descriptions and mathematical notations. Metrics denoted in bold font and asterisks are described in depth in Section 5.2.1.

TABLE C.1 GLOBAL MATHEMATICAL NOTATION GUIDE

P refers to probability
d refers to the predicted decision (for approval of credit)
G refers to gender (which can be either *m* [male] or *f* [female])
X refers to a set of control variables
Y refers to the actual classification result of an applicant
$\hat{Y}$ refers to the predicted classification result of an applicant
S refers to the predicted probability score
E refers to the expected value of predicted probability assigned by the classifier
k refers to a distance metric between individuals
K refers to a distance metric between a distribution of outputs
Z refers to equal average probability score
A refers to a set of attributes

[283] These metrics are adapted from Pessach and Shmueli (2020) and Verma and Rubin (2018). *See* Dana Pessach & Erez Shmueli, Algorithmic Fairness, arXiv preprint arXiv:2001.09784 (Jan. 21, 2020); Sahil Verma & Julia Rubin, Fairness Definitions Explained, FairWare'18: Proceedings of the IEEE/ACM International Workshop on Software Fairness at 1-7 (May 29, 2018).

The following statistical metrics are needed to compute statistical measures of fairness:

TABLE C.2 STATISTICAL METRICS FOR STATISTICAL MEASURES OF FAIRNESS[284]

True positive (TP): A case when the predicted and actual outcomes are both in the positive class

False positive (FP): A case predicted to be in the positive class when the actual outcome belongs to the negative class

False negative (FN): A case predicted to be in the negative class when the actual outcome belongs to the positive class

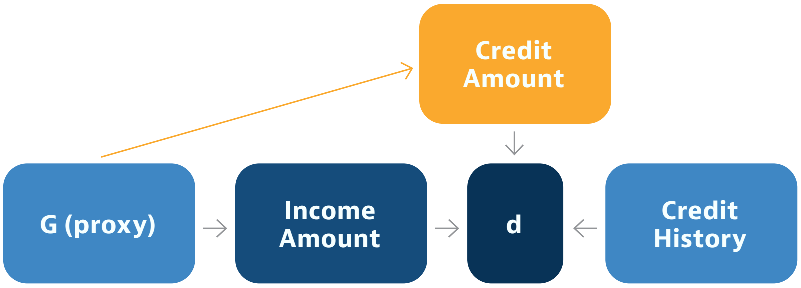
True negative (TN): A case when the predicted and actual outcomes are both in the negative class

Positive predictive value (PPV): The fraction of positive cases correctly predicted to be in the positive class out of all predicted positive cases, TP/(TP+FP)

Negative predictive value (NPV): The fraction of negative cases correctly predicted to be in the negative class out of all predicted negative cases, FN/(TN+FN)

For the fairness metrics derived from causal reasoning (*i.e.,* causal discrimination, counterfactual fairness, no unresolved discrimination, no proxy discrimination, and fair inference), a causal graph—a directed graph that contains attributes and associated relationships between the attributes and is used for building classifiers and other machine learning algorithms—has been provided below in Figure C. This causal graph maps the relationship between credit amount, credit history, income amount, the protected attribute *G*, and the predicted outcome *d*.

FIGURE C RELATIONSHIPS AMONG ATTRIBUTES



[284] Verma & Rubin at 2-3.

TABLE C.3 STATISTICAL METRICS FOR STATISTICAL MEASURES OF FAIRNESS[285]

C.1 Statistical Measures

MEASURE TYPE	MEASURE	DESCRIPTION	MATHEMATICAL NOTATION
GROUP	**DEMOGRAPHIC / STATISTICAL PARITY***	Demographic or statistical parity is achieved when the probability of a predicted positive outcome is the same across subpopulations within a dataset, such as protected class groups.	$P(d = 1 \mid G = m) = P(d = 1 \mid G = f)$
	CONDITIONAL STATISTICAL PARITY*	Conditional statistical parity is achieved when the probability of a predicted positive outcome is the same across protected class groups, once a set of designated control variables has been accounted for.	$P(d = 1 \mid X = x, G = m) = P(d = 1 \mid X = x, G = f)$, where x refers to a given control variable
	EQUAL OPPORTUNITY	Equal opportunity is achieved when true positive rates are the same across protected class groups.	$P[\hat{Y} = 1 \mid G = f, Y = 1] - P[\hat{Y} = 1 \mid G = m, Y = 1] \leq \varepsilon$
	PREDICTIVE PARITY*	Predictive parity is achieved when the PPV is the same across protected class groups.	$P(Y = 1 \mid d = 1, G = m) = P(Y = 1 \mid d = 1, G = f)$
	FALSE POSITIVE ERROR RATE BALANCE / PREDICTIVE EQUALITY	False positive error rate balance or predictive equality is achieved when the same false positive rates are observed across protected class groups.	$P(d = 1 \mid Y = 0, G = m) = P(d = 1 \mid Y = 0, G = f)$
	FALSE NEGATIVE ERROR RATE BALANCE	False negative error rate balance is achieved when the same false negative rates are observed across protected class groups.	$P(d = 0 \mid Y = 1, G = m) = P(d = 0 \mid Y = 1, G = f)$
	EQUALIZED ODDS*	Equalized odds are achieved when true positive rates and false positive rates are the same across protected class groups.	$P(d = 1 \mid Y = i, G = m) = P(d = 1 \mid Y = i, G = f)$, $i \in 0, 1$
	CONDITIONAL USE ACCURACY EQUALITY	Conditional use accuracy equality is achieved when the same PPV and NPV are observed across protected class groups.	$(P(Y = 1 \mid d = 1, G = m) = P(Y = 1 \mid d = 1, G = f)) \wedge$ $(P(Y = 0 \mid d = 0, G = m) = P(Y = 0 \mid d = 0, G = f))$
	OVERALL ACCURACY EQUALITY	Overall accuracy equality is achieved when prediction accuracy is the same across protected class groups.	$P(d = Y, G = m) = P(d = Y, G = f)$
	TREATMENT EQUALITY / ERROR RATE BALANCE	Treatment equality or error rate balance is achieved when the ratio of false negatives and false positives is the same across protected class groups.	$(FN/FP, G = m) = (FN/FP, G = f)$

[285] Verma & Rubin at 2-3.

MEASURE TYPE	MEASURE	DESCRIPTION	MATHEMATICAL NOTATION
GROUP	**CALIBRATION / TEST-FAIRNESS***	Calibration is achieved when for any predicted probability score *S*, subjects across protected class groups have equal probability to truly belong to the positive class.	$P(Y = 1 \mid S = s, G = m) = P(Y = 1 \mid S = s, G = f)$, for all s in (0, 1)
	WELL-CALIBRATION	Well-calibration is achieved if, for any predicted probability score *S*, subjects across protected class groups should not only have an equal probability to truly belong to the positive class, but this probability should be equal to *S*. That is, if the predicted probability score is *s*, the probability of both male and female applicants to truly belong to the positive class should be *s*.	$P(Y = 1 \mid S = s, G = m) = P(Y = 1 \mid S = s, G = f) = s$, for all s in (0, 1)
	BALANCE FOR THE POSITIVE CLASS	Balance for the positive class is achieved if subjects constituting the positive class across protected class groups have equal average predicted probability score *Z*.	$E(Z \mid Y = 1, G = m) = E(Z \mid Y = 1, G = f)$
	BALANCE FOR THE NEGATIVE CLASS	Balance for the negative class is achieved if subjects constituting the negative class across protected class groups have equal average predicted probability score *Z*.	$E(Z \mid Y = 0, G = m) = E(Z \mid Y = 0, G = f)$

C.2 Similarity-Based Measures

MEASURE TYPE	MEASURE	DESCRIPTION	MATHEMATICAL NOTATION
INDIVIDUAL	**FAIRNESS THROUGH UNAWARENESS***	Fairness through unawareness is achieved as long as protected class attributes are not included in the training and deployment datasets for classification.	$Ai = Aj \rightarrow di = dj$ where i and j are two individuals with the same set of attributes
	FAIRNESS THROUGH AWARENESS*	Fairness through awareness is achieved when individuals who are similar along various characteristics as defined by a distance metric—where the distance between the distributions of outputs for individuals should be at most the distance between the individuals—receive similar classifications, irrespective of their protected class features.	$K(d(m), d(n)) <= k(m, n)$

C.3 Causal Reasoning Measures

MEASURE TYPE	MEASURE	DESCRIPTION	MATHEMATICAL NOTATION
INDIVIDUAL	CAUSAL DISCRIMINATION	Causal discrimination is achieved if the same classification is produced for two subjects with the same exact attributes X.	$Xf = Xm \wedge Gf\ ! = Gm -> df = dm$
	COUNTERFACTUAL FAIRNESS*	A counterfactually fair causal graph is achieved if the predicted outcome *d* in the graph does not depend on a descendant of the protected attribute *G*. In Figure C, *d* is dependent on credit history, credit amount, and income amount. As the income amount is a direct descendant of *G*, the causal model is not counterfactually fair.	Please refer to Figure C
	NO UNRESOLVED DISCRIMINATION	A causal graph with no unresolved discrimination is achieved if there exists no path from the protected attribute *G* to the predicted outcome *d*, except via a variable that is influenced by the protected attribute in a non-discriminatory manner.	Please refer to Figure C
	NO PROXY DISCRIMINATION	A causal graph free of proxy discrimination is achieved if there exists no path from the protected attribute *G* to the predicted outcome d that is blocked by a proxy variable.	Please refer to Figure C
	FAIR INFERENCE	A causal graph satisfying n fair inference is achieved if there are no illegitimate paths from the protected attribute *G* to the predicted outcome *d*.	Please refer to Figure C

This report was made possible through operating support from Flourish Ventures. Additional support for this publication and other aspects of FinRegLab's Machine Learning in Credit Underwriting project was provided by the Mastercard Impact Fund, with support from the Mastercard Center for Inclusive Growth and JPMorgan Chase.

With support from:

Flourish, a venture of the Omidyar Group, has provided operating support to FinRegLab since its inception. Flourish is an evergreen fund investing in entrepreneurs whose innovations help people achieve financial health and prosperity. Established in 2019, Flourish is funded by Pam and Pierre Omidyar. Pierre is the founder of eBay. Managed by a global team, Flourish makes impact-oriented investments in challenger banks, personal finance, insurtech, regtech, and other technologies that empower people and foster a fairer, more inclusive economy.

The **Mastercard Center for Inclusive Growth** advances equitable and sustainable economic growth and financial inclusion around the world. The Center leverages the company's core assets and competencies, including data insights, expertise, and technology, while administering the philanthropic Mastercard Impact Fund, to produce independent research, scale global programs and empower a community of thinkers, leaders, and doers on the front lines of inclusive growth.

JPMorgan Chase is helping to advance racial equity and drive an inclusive recovery through business expertise and resources, policy advocacy, data and philanthropic investments. For more information and impact stories, visit www.jpmorganchase.com/impact.

Digital version available at finreglab.org

Published by FinRegLab, Inc.

900 19th Street NW, Suite 600
Washington, DC 20006
United States