FinRegLab

Explainability & Fairness in Machine Learning for Credit Underwriting

Policy Analysis

DECEMBER 2023

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 Policy Analysis is part of a broader research project on the explainability and fairness of machine learning in credit underwriting. Other reports in this series, including an overview of our empirical and policy findings, are available at https://finreglab.org/ai-machinelearning/explainability-and-fairness-of-machine-learning-in-credit-underwriting. Support for this publication and other aspects of FinRegLab's Machine Learning for Credit Underwriting project was provided by the Mastercard Center for Inclusive Growth, JPMorgan Chase, and Flourish Ventures. Detailed information on our funders and additional acknowledgments can be found on the inside back cover.

The empirical research described herein was conducted in collaboration with Professors Laura Blattner and Jann Spiess at the Stanford Graduate School of Business. We would also like to thank the companies that participated in the research—Arthur AI, H20.ai, Fiddler AI, Relational AI, Solas AI, Stratyfy, and Zest AI—for their time and commitment throughout this project.

We would like to extend a special note of appreciation for their leadership and facilitation of FinRegLab's policy working groups to Adam Gailey, Charles River Associates; John Morgan, Capital One; Yogesh Mudgal, Citi; Eric Sublett and Ken Scott, Relman Colfax PLLC; and Stephen Van Meter. Thanks also to the U.S. Commerce Department, National Institute of Standards and Technology, and the Stanford Institute for Human-Centered Artificial Intelligence for co-hosting a conference, "Artificial Intelligence and the Economy: Charting a Path for Responsible and Inclusive AI."

FinRegLab would also like to recognize the individuals and organizations who participated in the policy working groups and/or our project Advisory Board as listed in Appendix A, the participants of the Artificial Intelligence and the Economy conference, and stakeholders who participated in interviews.

When viewed with an Adobe Acrobat reader, elements listed in the Table of Contents or in blue text are links to the referenced section or feature. Functionality may be limited in non-Adobe readers. Adobe's reader can be downloaded for free at get.adobe.com/reader.

CONTENTS

1. Introduction 3

2. The Opportunities—and Challenges—of Machine Learning Underwriting Models 6

2.1 The Shift to Machine Learning Underwriting Models 7

2.2 Implications for Inclusion and Fairness 9

2.2.1 Potential Benefits to Inclusion and Fairness 10

2.2.2 Potential Risks to Inclusion and Fairness 12

2.2.3 The Role of Data Diversification 12

2.3 Transparency Challenges and Tools 14

2.3.1 The "Black Box" Challenge 14

2.3.2 Key Explainability Techniques 15

2.4 Broader Questions about Responsible Use 18

3. Overview of FinRegLab's Research 20

3.1 Empirical Summary 21

4. Model Risk Management 23

4.1 Regulatory and Operational Context 23

4.2 Selected Policy Topics 25

4.2.1 Updates to General MRM Frameworks 26

4.2.2 Standards for Evaluating Explainability Techniques and Conducting Conceptual Soundness Reviews 26

4.2.3 Governance Standards for Nonbanks 32

4.2.4 Validating Vendor-Provided Models and Model Management Tools 33

5. Adverse Action Notices 36
5.1 Regulatory and Operational Context 36
5.1.1 Policy Motivations 37
5.1.2 Existing Law and Guidance 37
5.1.3 Compliance Practices 39
5.1.4 Key Risks and Compliance Issues for Adverse Action Notices 40
5.2 Key Policy Issues for Adverse Action Notices and Machine Learning Models 40
5.2.1 Producing Reliable Descriptions of the Behavior of Machine Learning Models 41
5.2.2 Providing More Information about How and Why Particular Features Affected the Credit Decision 42
5.2.3 Identifying Plausible Paths for Applicants to Increase Their Chances of Approval 44
5.2.4 Incorporating Non-Traditional Data 45
6. Fair Lending 47
6.1 Regulatory and Operational Context 47
6.1.1 Policy Motivations 48
6.1.2 Existing Law and Guidance 49
6.1.3 Operational Context 50
6.1.4 Key Issues and Risks 57
6.2 Key Policy Issues 58
6.2.1 Disparate Treatment 58
6.2.2 Disparate Impact 59
6.2.2.1 Measuring Fairness 59
6.2.2.2 Use of Protected Class Information and Debiasing Techniques 60
6.2.2.3 Identifying Less Discriminatory Alternatives 62
7. Conclusion 66
APPENDIX A: FinRegLab Policy Working Groups and Advisory Board 69
APPENDIX B: Key Terms 70
APPENDIX C: Recent Research on Related Topics 75
Endnotes 81
Bibliography 95

1. INTRODUCTION

Every week, machine learning underwriting models are determining the outcomes of credit applications submitted by hundreds of thousands of consumers and small business owners.[1] While the underlying technologies and data are less complex than ChatGPT and other high-profile forms of artificial intelligence (AI), the use of machine learning (ML) for such important decisions still raises fundamental questions about whether we have adequate toolkits for building, understanding, and managing models that are reliable and fair.[2]

The stakes in the credit context are high. ML underwriting models' greater accuracy and capacity to analyze large datasets (particularly new, more inclusive sources of information) have the potential to increase access to credit for millions of people who are difficult to assess using traditional models and data. This underserved population includes disproportionately high numbers of Black, Hispanic, and lower income consumers. For instance, about 20% of U.S. adults lack sufficient traditional credit history to generate scores under the most widely used models, with almost 30% of Black and Hispanic consumers lacking such scores compared to about 16% of Whites and Asian-Americans.[3]

Yet the very quality that fuels ML models' greater predictive power—their ability to detect more complex data patterns than prior generations of underwriting models—makes them more difficult to understand and increases concerns that they could exacerbate inequalities and perform poorly in changing data conditions.[4] The complexity of many ML models has caused transparency to emerge as an urgent threshold question for both lenders and regulators in evaluating whether individual models are safe, fair, and reliable.[5] For example, many stakeholders are concerned that if users cannot assess whether an underwriting model relies on strong, intuitive, and fair relationships between an applicant's behavior and creditworthiness, it may be more difficult to diagnose and mitigate performance and fairness issues or to determine compliance with regulatory requirements.

New data science techniques—often themselves involving machine learning or other complex computational methodologies—have emerged both to explain ML models' operation and to manage concerns about their fairness and reliability. These include both *post hoc* explainability techniques that analyze key aspects of model behavior and debiasing techniques that can be used to reduce racial or other disparities in model predictions. Many vendors that are providing ML platforms and services to lenders have incorporated these techniques into their proprietary tools for diagnosing, managing, and monitoring ML models. But the techniques and tools also raise questions about whether and how to use them appropriately both to manage models and to perform regulatory compliance tasks in the credit context.[6]

To study these questions, FinRegLab has conducted extensive market context interviews, performed empirical analyses with Professors Laura Blattner and Jann Spiess of the Stanford Graduate School of Business, and engaged with a broad range of stakeholders. This Policy Analysis is the fourth in our series of research reports, and elaborates on a Policy & Empirical Findings Overview that we

released in July 2023.[7] The analyses presented below build on our prior workstreams, including the deliberations of three working groups that FinRegLab convened to solicit insights from more than 75 subject matter experts including representatives of lenders, data and technology companies, advocacy organizations, and research institutions. Stakeholder discussions shaped our understanding of various issues, but our reports reflect FinRegLab's independent analysis in all respects.

As discussed further below, we find that:

» **Some *post hoc* explainability techniques can provide reliable information about key aspects of model behavior, but stakeholders are still debating their appropriate use and sufficiency.** Our empirical research evaluated use of various techniques and tools to perform tasks relating to generating individualized consumer disclosures, managing fair lending risks, and conducting general governance activities. We found that some techniques provided reliable information about key aspects of model behavior, though there was no "one size fits all" technique or tool that performed the best across all regulatory tasks. Our results emphasize the importance of choosing the right explainability tool for the particular ML model and task, deploying it in a thoughtful way, and interpreting the outputs with an understanding of the underlying data.

  As stakeholders work to develop standards for the appropriate deployment of explainability techniques, the analytical framework we developed for our research project may be useful for evaluating the performance of tools in different settings. Additional research could also be useful to inform continuing stakeholder debates about the tradeoffs in performance and simplicity between using *post hoc* techniques and imposing upfront constraints on model architecture to create greater transparency. Beyond methodological and process issues, stakeholders are also debating whether the information produced by *post hoc* techniques is sufficiently equivalent to what can be generated about more traditional models to be relied upon for various business and compliance functions.

» **The transition to machine learning has the potential to improve fairness and inclusion, in part by giving lenders a more robust toolkit for mitigating disparities.** Despite the focus on transparency as a threshold issue for ML models as discussed above, in our empirical research the most powerful approaches to managing fairness did not necessarily hinge upon explaining the inner workings of the model as an initial step. Instead, we found that automated approaches that generated a range of alternative models produced options that had greater predictive accuracy and smaller demographic disparities than traditional strategies that assessed which input features made the biggest contribution to disparities and then omitted or made narrow adjustments to those individual features.

  Further research could help inform implementation choices for different techniques and evaluation of the different model alternatives that they produce for robustness and other considerations. Public policy questions regarding fair lending compliance have also taken on additional urgency in light of the adoption of ML models. Absent greater regulatory certainty, lenders have been hesitant to deploy certain debiasing techniques in particular ways because the techniques use data about race, gender, and other protected characteristics differently than traditional mitigation approaches. The availability of new debiasing approaches has also highlighted outstanding questions about the nature and extent of lenders' obligations to search for fairer models during the development process.

» **Defining basic concepts and expectations could be a useful first step toward updating regulatory frameworks for the machine learning era.** While ML technologies and our understanding of them are evolving rapidly, regulators can take steps now

to encourage responsible use. For instance, defining the key qualities of explainability tools and clarifying expectations about how and when lenders should search for fairer alternative models would increase consistency of practice and shape how lenders use their expanded toolkits in the ML context. And while existing model risk management guidance provides a flexible framework for governance that many other sectors lack, it does not apply to the full spectrum of lenders. Stakeholders see potential value in addressing governance concerns for particular subgroups of lenders and articulating basic elements that should be considered in determining whether particular ML models are "fit for use" in the underwriting context.

The transition to machine learning models is likely to prove the largest evolution of automated underwriting systems in at least a generation. As the fairness research results illustrate, adjusting market practices and regulatory expectations to account for ML models could provide opportunities to address longstanding concerns about prior generations of predictive credit models and the compliance frameworks that govern them. The transition may also provide opportunities to adjust existing underwriting systems to incorporate more inclusive data sources at relatively modest additional cost.

Additional public research and stakeholder dialogue will be critical to inform and shape these changes, not only within the credit ecosystem but also with other sectors that are grappling with the responsible use of AI and ML models in sensitive use cases. At the same time, lessons from deploying machine learning and secondary tools in the credit context have the potential to inform governance activities in other sectors and the development of more effective techniques for understanding and managing AI/ML applications. FinRegLab has structured this research project with an eye toward facilitating cross-sector dialogues and will continue to do so as it researches AI/ML and data topics going forward.

* * *

This report is structured as follows:

- Section 2 provides an overview of the shift toward machine learning underwriting models, including the opportunities and risks for inclusion and fairness from adoption of ML techniques and non-traditional data sources, transparency challenges and tools for managing more complex models, and broader questions about the responsible use of ML models.
- Section 3 provides a brief summary of FinRegLab's research project, including our empirical findings on explainability techniques and debiasing approaches.
- Section 4, Section 5 and Section 6 provide more detailed policy analyses of concerns and debates relating to the adoption of machine learning underwriting models in the context of three critical regulatory compliance areas for credit underwriting:
  - General risk management and model governance, particularly for banks.
  - Production of "adverse action" disclosures to certain credit applicants
  - Compliance with federal fair lending laws.
- Section 7 summarizes potential next steps for policymakers and other stakeholders as ML techniques and our understanding of them continue to evolve rapidly.

For reference, Appendix A lists the organizations whose employees participated in FinRegLab's advisory board and/or policy working group discussions; Appendix B defines common terms and acronyms; and Appendix C summarizes recent published research on related topics.

2. THE OPPORTUNITIES—AND CHALLENGES—OF MACHINE LEARNING UNDERWRITING MODELS

Lenders began using predictive models decades ago to forecast the likelihood that applicants would default on new loans, based largely on data from three nationwide credit bureaus and statistical techniques such as logistic regression.[8] Borrowers and lenders have benefitted from use of automated underwriting systems in a variety of ways: reduced defaults, processing times, underwriting costs, and loan pricing; expanded access to credit; improved consistency of treatment of similarly situated borrowers; and increased competition for borrowers.[9] However, these benefits are not distributed evenly. For instance, the fact that about 20% of U.S. adults and nearly 30% of Black and Hispanic consumers lack sufficient traditional credit history to generate scores under the most widely used models means that they may be denied credit or charged higher prices not because they are high default risks but rather because they are difficult to assess.[10]

As computational power and techniques have evolved and new sources of digital information have become more widely available, consumer and small business lending markets are facing the biggest evolution of automated underwriting systems in at least a generation. The use of machine learning algorithms to develop underwriting models is a core innovation that is being driven by the potential for substantial increases in predictive accuracy. While machine learning techniques can be applied to traditional credit history sources, they may be adopted at the same time that lenders adjust their systems to account for new types of data, such as the use of digital feeds of bank account information or other sources of cash-flow information. This combination of new data and techniques holds particularly promise for improving access to credit among historically excluded populations.

Yet the very qualities of ML models that create the potential for improvement—their ability to detect more complex relationships in historical data and to process large amounts of information from diverse sources—also create concerns about our ability to understand, manage, and rely upon the resulting models. Stakeholders are particularly concerned about risks that the models could exacerbate historical disparities and may prove "brittle" in changing conditions, causing rapid declines in predictive performance.[11] The models' increased complexity further heightens concerns about model management and regulatory compliance, making transparency and explainability critical issues with regard to the pace, breadth, and nature of future adoption.

This section summarizes (1) the shift to ML underwriting models; (2) its opportunities and risks for inclusion and fairness; (3) transparency challenges and tools for managing more complex models; and (4) broader debates about responsible use of ML models for credit underwriting. While the use of non-traditional data sources is not the primary focus of this report, Section 2.2.3 briefly discusses the ways in which incorporating new data sources can increase the potential benefits and challenges of ML adoption for greater financial inclusion.[12]

2.1 The Shift to Machine Learning Underwriting Models

Traditional credit scoring and underwriting models rely on a relatively limited number of inputs that are selected by human developers, who work to select combinations that both maximize overall predictive power and minimize correlations among inputs to simplify model operations. (See Box 2.1.1) While some companies have implemented more complex structures over time in pursuit of greater predictive power,[13] processes for documenting and analyzing traditional models have benefitted from both the relative transparency of logistic regression and decades of usage. Regression models' notation identifies both their input features and their weights,[14] and commonly used statistical analyses have become widely accepted to measure the importance of individual features for various business and regulatory purposes.[15] These include:

- **General Risk Management and Model Governance:** To protect the safety and soundness of banks and the broader financial system, banks are expected to implement risk-based governance mechanisms for the development, deployment, and monitoring of models. These processes include analyzing whether models are relying on relationships in the data that are "conceptually sound" and assessing models' stability in changing data conditions. Both activities involve identifying features that are playing important roles in the model's operation.
- **Adverse Action Disclosures:** Federal laws require lenders to provide individualized disclosures to credit applicants of the "principal reasons" for rejecting an application and the "key factors" that are negatively affecting consumers' credit scores if the lender charges higher prices based on credit report information.
- **Fair Lending Compliance:** Federal fair lending laws generally prohibit both the use of race, gender, or other protected characteristics in underwriting models ("disparate treatment") and the use of facially neutral criteria that have a disproportionately adverse impact on protected groups unless the criteria further a legitimate business need that cannot reasonably be achieved through less impactful means ("disparate impact"). Traditional disparate impact compliance approaches often focus on testing whether omitting or modifying individual features that have been identified as driving disparities can improve fairness without substantial reductions in predictive accuracy.

With advances in computational power, some lenders have begun deploying machine learning techniques to develop underwriting models. (See Box 2.1.2) Here, the algorithms themselves identify complex predictive relationships among large numbers of inputs, while developers make critical decisions about such issues as what data the learning algorithms are trained on, how the algorithms generate underwriting models, and what techniques, tools, and strategies are used in development and validation processes.[16] Two of the most commonly used machine learning approaches in the credit context are boosted tree models (particularly a variation called XGBoost)[17] and neural networks.[18]

Depending on the developer's decisions and training data, some ML underwriting models may not be significantly harder to understand than traditional underwriting models, while others are substantially more complex. The most complicated models are sometimes referred to as "black box" models. They frequently rely on hundreds or thousands of features, including in some cases "latent features" that are generated by the ML algorithms from the initial inputs, and often involve complex architectures such as multiple layers or ensembles of individual models. These factors can help machine learning models detect more complex relationships within the data, including relationships that are non-monotonic (meaning that increasing the value of an input feature may reduce the likelihood of default in some circumstances and increase it in others) and non-linear (meaning that increasing the value of an input feature by a given amount may not change the likelihood of default by the same amount in all circumstances).[19]

BOX 2.1.1 WHY DO STRONG CORRELATIONS BETWEEN FEATURES COMPLICATE ANALYSIS OF PREDICTIVE MODELS?

Highly correlated features are variables that have strong negative or positive relationships with one another, so that changes in one are generally associated with changes in the other. Credit report data and other financial information that is often used in credit underwriting tends to be highly correlated. However, the presence of correlated features within a predictive model makes it difficult to distinguish which specific feature combinations are the most important drivers of default predictions for individual applicants, demographic groups, or for operation of models as a whole.

For traditional logistic regression models, developers generally strive to reduce feature correlations in order to enhance model simplicity and interpretability. Given the linear form of these models and the relatively small number of variables typically involved, it is often feasible for developers to use an iterative process to select features that are less correlated with those already included in the model. However, this process does not entirely eliminate correlation issues, and developers may have to accept some tradeoffs between interpretability and predictive power, especially if highly predictive but correlated features are removed.[20]

In contrast, machine learning underwriting models typically use hundreds and sometimes thousands of features, many of which may be highly correlated. For instance, a machine learning model may incorporate multiple features associated with past mortgage delinquencies, each capturing somewhat different aspects of the borrower's history. While having many granular features can add predictive power, the associated correlations among features make it difficult to attribute variations in the model's predictions to individual features.

Data science techniques that have been developed to explain complex ML models vary in their assumptions and treatment of correlated features. See Section 2.3.2 for further discussion.

BOX 2.1.2 WHAT IS MACHINE LEARNING? HOW DO ML UNDERWRITING MODELS COMPARE TO OTHER FORMS OF AI/ML?

Artificial intelligence is a term coined in 1956 to describe computers that perform processes or tasks that "traditionally have required human intelligence," while machine learning is often used to refer to the subset of artificial intelligence that gives "computers the ability to learn without being explicitly programmed."[21]

While the release of ChatGPT in November 2022 sparked broad public interest in "generative AI" models that create new content (including text and images) that is similar to learned patterns in training data,[22] the types of ML techniques used in building credit underwriting models are sometimes called "predictive AI" or "discriminative AI." They use training data to develop models that will predict a particular outcome (such as the likelihood of default) when applied to additional data sets.

The scale and nature of the data used to train ML underwriting models differ substantially from those used to train generative AI models, which are often built using data scraped from large portions of the internet. This raises a broad range of questions about accuracy, bias, intellectual property rights, and other issues.[23]

By comparison, ML underwriting models are trained on much smaller, curated data sets, even in cases where they expand beyond the use of traditional credit history information (see Section 2.2.3). They are also limited in the extent to which they are allowed to update dynamically, so that changes are subject to oversight in various forms. However, even with these distinctions, the use of ML models for credit underwriting raises important questions about our ability to understand, manage, and rely upon the models for such an important use case.

Concerns about managing generative AI models are also increasing calls to regulate AI/ML applications more generally. In November 2023, a new Executive Order and supporting memo totaling more than 100 pages of directions and requests to federal agencies to update regulations and guidance concerning a broad range of AI technologies and use cases, including to federal agencies concerning credit underwriting and related activities.[24]

This ability to detect more nuanced data relationships is core to both the opportunities and challenges posed by ML adoption. On one hand, it can potentially increase predictive accuracy in general and specifically for applicants who are difficult to assess using traditional methods and data.[25] On the other, it creates concerns that ML models are more prone to "overfitting" to the data than regression models, and thus may have steeper performance deteriorations when conditions start to change, and that they could exacerbate demographic disparities relative to incumbent

models. The lack of transparency about how the models are generating their predictions further increases concerns about model management and regulatory compliance, particularly since traditional approaches often rely upon identifying and managing individual features.

At the same time, data science and machine learning techniques are providing a range of alternative options for evaluating and managing models. These include a variety of *post hoc* explainability methods that have been developed to analyze key aspects of model behavior, as described in Section 2.3.2, and debiasing techniques as described in Section 6.1.3.4. Those techniques are the primary focus of this research project as discussed in subsequent sections. The remainder of Section 2 provides additional background on inclusion and fairness considerations, transparency challenges, and broader concerns about responsible use as the adoption of machine learning models continues to expand in the credit underwriting context.

2.2 Implications for Inclusion and Fairness

While lenders are adopting ML underwriting models for a broad range of business and competitive reasons,[26] the potential implications of this transition for inclusion and fairness are important to a broad range of stakeholders (see Box 2.2.1 and Box 2.2.2). Particularly if combined with more inclusive data as discussed in Section 2.2.3, ML underwriting models have the potential to achieve a number of benefits, including:

- Expanding access to more borrowers who are creditworthy and reducing the number of people who are offered loans they are unlikely to be able to repay;
- 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; and
- Improving identification and mitigation of certain forms of discriminatory lending.

BOX 2.2.1 DEFINING INCLUSION AND FAIRNESS CONCEPTS

This section focuses primarily on the potential for ML underwriting models to improve access to affordable, responsible credit among historically underserved populations, in part because of credit's role in broader economic participation (see Box 2.2.2). As discussed in Section 6, other concepts of fairness are enshrined into fair lending laws and raised in broader policy debates about automated underwriting. These concepts include:

- **Equal treatment:** This requires that individuals be subject to the same criteria and that similarly situated applicants receive similar treatment. In anti-discrimination law, this principle is invoked to prohibit different treatment because of applicants' race/ethnicity, gender, or other protected characteristics.
- **Equity:** This focuses on whether there are equal outcomes among different groups of people, even if they may not be similarly situated in some respects. For example, the first stage of analyzing whether facially neutral practices produce a disparate impact focuses on whether the practice produces demographic disparities in approvals or pricing, without accounting for differences in financial situations.
- **Consistency of predictive accuracy:** Adopting more accurate models can potentially reduce the incidence of certain negative outcomes among particular subgroups of borrowers, for instance by reducing the number of applicants who are rejected or charged higher prices due to overestimates of default risk and/or who are granted loans that they cannot repay due to underestimates of default risk.

Other notions of fairness include whether credit criteria are arbitrary and whether applicants should have notice about those criteria. Model risk management and adverse action notice requirements can help to address such concerns, as discussed in Section 4 and Section 5 below.

BOX 2.2.2 THE IMPORTANCE OF FINANCIAL INCLUSION

Access to affordable, responsible financial services can help to increase broader economic participation, opportunity, and equity. For instance, credit can help to bridge short-term gaps in income and expenses, particularly among households and businesses that experience income or expense volatility, and to make long-term investments in work-related transportation, homeownership, and small business growth. These long-term investments in turn can help build savings to increase long-term financial health and build wealth, including providing for generational transfers.

The financial system can create significant economic multiplier effects in both positive and negative directions. For example, given that previous inequities created by historical discrimination in such fields as employment, education, housing, and lending have contributed to substantial disparities in income and assets,[27] it is not surprising that consumers of color may tend to find it more difficult to repay loans, creating a cycle of declining credit scores and increasing prices for credit. Lack of access to mainstream financial services and targeting by lenders that offer credit products with higher prices and riskier structures can also contribute to disparities in credit access and outcomes.[28]

At the same time, financial products and services can also provide a helpful bridge into the financial mainstream, for instance, by using government benefits delivery as an opportunity to provide consumers with low-cost, secure transaction accounts to manage and save their money[29] and to become a source of data for credit underwriting.[30]

Yet stakeholders are also highlighting risks that ML underwriting models might exacerbate historical disparities or prove more difficult to manage for fairness concerns. These issues extend beyond the dictates of federal fair lending law as discussed in Section 6 to include the broader net effects of underwriting and pricing practices based on credit default predictions.

2.2.1 Potential Benefits to Inclusion and Fairness

Stakeholders often focus on the potential effects of ML underwriting models on applicants who lack substantial traditional credit history and credit scores, but those consumers and business owners are just one segment of borrowers who could potentially be affected by the use of new techniques, with or without new data sources. Inclusion could potentially be advanced by reducing prediction errors in traditional credit underwriting models, implementing more effective debiasing techniques, and giving lenders confidence to extend credit to a broader range of applicants.

2.2.1.1. Improving Assessment of Applicants with Little or Marred Credit History

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 whom have long struggled to access affordable credit. 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 broader populations and groups. For example, VantageScore Solutions reports that the adoption of machine learning techniques in its most recent models resulted in an accuracy improvement of 16.6% for bank card originations and a 12.5% improvement for auto loan originations for consumers whose credit histories had not been updated in the prior six months, even though some credit scoring systems will not generate scores for such applicants.[31] Combining ML models with new data sources has the potential to produce even greater impacts on populations with sparse traditional credit histories as discussed further in Section 2.2.3.

2.2.1.2 Improving General Predictiveness

Improved accuracy can help to reduce the number of borrowers being offered loans they are unlikely to be able to repay and increase the number of qualified borrowers being approved for credit. Given that lenders rely on predictions of default risk to inform their decisions about approvals and credit terms, even small improvements in predictive accuracy can produce wide-ranging fairness and inclusion benefits for firms, borrowers, and some applicants for credit.

In practice, the net effects of moving to machine learning models—e.g., whether the increase in people approved using a machine learning model is larger than the number of people who are rejected using a machine learning model but would have been approved using a traditional underwriting model—will vary depending on a wide variety of factors such as market and product segment, economic conditions, and the sophistication of the prior model. However, many stakeholders report at least modest net inclusion gains when moving from logistic regression to machine learning underwriting models using traditional data sources, and all emphasize the fairness benefits of reducing approvals for those unlikely to be able to repay the requested loan.

At the very least, lending that uses more accurate risk assessment methods can improve access for certain individual consumers or groups and reduce the cost of certain products for others. For lenders, 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. Some fintech lenders report that machine learning models have expanded their ability to offer credit to substantial numbers of applicants who would not be approved under widely used industry benchmarks and scores.[32]

2.2.1.3 Improving Fair Lending Mitigation Strategies

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.[33] As discussed in further detail in Section 6, this may enable the use of models that retain the predictive power of variables and relationships causing disparities instead of having to eliminate those features entirely. The development of machine learning models enables consideration of many more iterations of a model than in incumbent models, including many changes to a model's specifications, which can enhance predictive power and enable more explicit consideration of certain tradeoffs.[34] The transition to machine learning is also inspiring consideration of how to incorporate additional definitions and methods of measuring algorithmic fairness into model development and oversight processes.[35]

2.2.1.4 Helping Particular Industry Segments Broaden Their Credit Boxes

One additional question is the extent to which the transition to ML occurs among smaller lenders such as community banks. Such lenders can play an important role in financial inclusion more generally by focusing on relatively underserved market segments, yet they can also be particularly dependent on traditional credit bureau data and third-party scoring and underwriting models in the consumer lending context because of analytical, technological, and human resource limitations. The transition to more accurate and inclusive machine learning models thus could have particular benefits in this market segment, yet as discussed further in Section 4.2.3, ML adoption also presents particular challenges because those same resource constraints make it difficult for smaller institutions to supervise vendors with regard to proprietary models and technologies. To date, adoption of ML underwriting models has been concentrated largely among large banks and nonbank fintech lenders although some pockets of adoption are occurring among other lender segments.[36]

2.2.2 Potential Risks to Inclusion and Fairness

For all of the potential benefits, ML underwriting models also raise concerns about potential exclusion, unfairness, and bias. These concerns are broader than establishing compliance with anti-discrimination laws and include more fundamental questions about data gaps, modeling decisions, and other issues that can affect the performance of models for particular groups.

For example, concerns that ML underwriting models may focus on financial disparities stemming from historical discrimination with even greater accuracy than traditional models have heightened many advocates' concerns about risk-based pricing, where lenders charge higher rates to higher risk applicants to offset the increased risk of loss. While proponents argue that this practice causes lenders to accept applications from borrowers who they would otherwise decline, critics argue that the higher prices increase the risk of default among the most vulnerable borrowers. Studies of both ML models and previous generations of automated underwriting suggest that pricing disparities for particular groups can increase at the same time that approval disparities shrink if models predict that previously excluded applicants are at somewhat higher risk of default (but not so high as to warrant disapproval of their applications).[37] However, such studies have not tracked whether and how patterns changed over time as newly approved applicants build credit history.

The general increases in complexity and the fact that some machine learning models rely on "latent features" that are identified by the learning algorithms from relationships in the input data also increase concerns under fair lending laws, particularly those that generally prohibit the use of demographics or other protected characteristics in credit underwriting. As discussed further in Section 6.2.1, the use of latent features raises particular 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.

Finally, while the transition to machine learning presents new options for debiasing models as noted above, stakeholders are still evaluating the utility of those options. The transition also raises questions about the application of existing regulatory frameworks and the effectiveness of traditional approaches that have evolved in the context of models that use a relatively small number of features and less complex architectures. These issues are also discussed in greater detail in Section 6.

2.2.3 The Role of Data Diversification

While this report focuses primarily on the adoption of machine learning techniques rather than non-traditional data sources, it is important to note that the two may often proceed in tandem and that the incorporation of new data sources can increase both the opportunities and challenges posed by the adoption of ML models alone.

In the U.S., automated underwriting systems have historically relied heavily on data from three nationwide consumer reporting agencies to predict credit risk. Because the credit bureau files are made up primarily of payment history on past loans, this makes it more challenging for first-time borrowers to get approved. Accordingly, lenders and other stakeholders have been trying to lever age more diverse and granular data generated about consumers' daily activities and behaviors to improve consumer and small business credit underwriting. Some sources of data, such as bank account information, are available for a much broader and more diverse range of consumers than traditional credit history.[38] Diversifying data sources also potentially gives more holistic insight into applicants' financial circumstances and behavioral patterns, such as income flows and how borrowers pay their full range of recurring obligations including bills that are not typically reflected in credit

BOX 2.2.3.1 TAXONOMY OF NON-TRADITIONAL DATA SOURCES

Potential data sources for credit underwriting can be broadly grouped into four categories:[39]

Customer Data: Lenders may capture a range of data based on their interactions with consumers during application processes and lending relationships that can be used in credit decisioning processes. This may include the channel by which the consumer submitted the application to payment and communications patterns during a prior loan from the same financial institution. Privacy, fairness, and fair lending implications of using such data may vary given the range of information included in this category.

Alternative Financial Data: Alternative financial data includes a variety of non-lending financial activities, including the payment of other types of recurring expenses, and can often be extracted relatively easily from deposit or prepaid accounts. Depending on the source, this information may contain more granular and timely information about applicants' financial position than credit bureau records and help to provide a more complete picture of an applicant's financial capabilities and behavior. While such information can be particularly important in underwriting consumers with sparse credit reports, research suggests it can improve default prediction more broadly.[40]

Behavioral Insights in Alternative Financial Data: Some sources of alternative financial data provide detailed information about consumer behavior, such as where and when people shop or what they buy. Some of this information may be considered relevant to default risk assessment, but this data can also raise fairness, fair lending, and privacy concerns, particularly if the behavior is highly correlated with protected class or consumers are not aware that such information will affect underwriting decisions or is being collected and retained.

Non-Financial Alternative Data: Examples of non-financial data include search histories and personal networks.[41] Such information tends to raise heightened concerns about reliability and fairness—even if they are statistically correlated to default risk, they may have no clear causal or intuitive links to creditworthiness—as well as potential concerns about privacy, correlation with race or other protected characteristics, and notice to consumers.[42] This type of data is not commonly used in the U.S. for underwriting, although some nonbank lenders consider educational factors and digital footprint information in addition to more traditional inputs.[43]

reports.[44] 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.[45] (See Box 2.2.3.1)

New data sources can be incorporated into traditional regression models without the use of ML techniques, but the ML transition may facilitate data adoption in at least two ways. First, ML models' capacity to detect patterns and relationships among a vast number of features can be useful in working with large and diverse data sources, particularly if they are less structured and regularized than traditional credit bureau data.[46] Second, even for lenders who are not primarily motivated by the potential inclusion benefits of incorporating new data sources, the operational overhaul required for widespread use of ML models likely creates a chance to reset lending platforms to manage use of alternative data with relatively little additional cost.

When evaluating whether to include potential variables, lenders consider not only the potential value for predictiveness, but also inclusion and fairness considerations, logistical issues regarding data access and processing, other legal and regulatory requirements, and broader reputational and customer relationship implications. For example, new data sources may make it easier to underwrite consumers with little traditional credit history, but such information is typically also evaluated to understand the extent to which it is correlated with demographics and thus may raise concerns about disparate treatment or disparate impact as discussed further in Section 6. Incorporating new data sources also requires adjusting systems to generate adverse action disclosures based on the new features as discussed in Section 5 and (for banks) validating the use of the data under model risk management guidance as discussed in Section 4. Where developers use more opaque machine learning models in conjunction with the new types of data, this further underscores the importance

of managing transparency concerns to facilitate assessing the roles that non-traditional data play in individual models.

2.3 Transparency Challenges and Tools

Along with fairness and inclusion, questions about whether we can sufficiently understand the operation of machine learning underwriting models will play a fundamental role in determining the scope and pace of adoption going forward. Advancements in data science techniques are creating both opportunities and challenges in this regard for lenders, regulators, and other stakeholders. As noted in Section 2.1, some ML models can be substantially more complicated than incumbent models due to their number of features, complex architectures, and nuanced data relationships. At the same time, advancements in *post hoc* explainability techniques could potentially be useful in probing key aspects of model behavior and have caused stakeholders more broadly to revisit debates about the importance of transparency for particular policy purposes and the extent to which existing market practices and policy standards provide it.

Given the importance of being able to understand model behavior for both business and regulatory reasons, transparency has thus emerged as an urgent threshold question that is shaping lenders' broader decisions about how and when to adopt ML underwriting models. This report defines transparency broadly as the ability of various stakeholders to access information they need related to a model's design, use, and performance. Some stakeholders use terms such as interpretability or explainability to express similar concepts, but as described below those two terms are often associated with particular approaches to obtaining greater insight into models' operations. Indeed, one of the challenges with debates about transparency and associated terms is that there are no uniform definitions or benchmarks for determining what level of information is sufficient for a particular purpose or audience.[47]

2.3.1 The "Black Box" Challenge

While lenders, regulators, and other stakeholders have grown accustomed to managing traditional underwriting models over time, such models are the culmination of a complex series of human choices during model development and validation. For example, because traditional credit information sources tend to be substantially correlated with each other, developers exercise substantial expertise and judgment in deciding which particular combination of individual features will maximize overall predictive power, simplify model operations, reduce the risk of fair lending concerns, and address other business and policy concerns.[48] While coefficients and weights help in analyzing the role that the final features play in logistic regression models' operation, other aspects of model development and operations can be complicated to understand depending on the degree of documentation created by developers and other stakeholders' technical backgrounds.

In contrast, machine learning models are developed by tasking the learning algorithm to map predictive relationships in a larger number of (often highly correlated) features. The effect is not unlike transitioning from a simple box of eight crayons, made up largely of primary colors and simple combinations, to a larger box with 128 colors that provide much more precise and subtle combinations. Rather than just picking one or two metrics concerning past delinquencies, such as the number of accounts that have been 60 or more days past due in the past two years, ML models may divide delinquency data into much more granular time periods as well as considering whether the combination of delinquencies with other features such as account balances helps to further predict future default risk. However, as complexity increases the more difficult it is to tell which precise

combinations are most important for individual applicants, demographic groups, or the operation of the model as a whole.

Firms are using a variety of strategies to manage the additional transparency concerns that arise due to the use of learning algorithms to discover predictive relationships in larger datasets. These include constraining the structure of the machine learning model to make it less complex and easier to understand, using *post hoc* explainability techniques to analyze key aspects of the model's behavior, and a combination of the two approaches. The result is a spectrum of models, ranging from what are often called "inherently interpretable" models that can generally be summarized in a single, generalizable notation which conveys their features and weights assigned to those features,[49] to highly complex ML models that are often referred to as "explainable" because they depend on the use of *post hoc* explainability methods to analyze the model's behavior and the bases of its predictions.[50]

Firms and researchers alike are working to understand better the tradeoffs between using inherently interpretable models and pairing less interpretable models with *post hoc* explainability methods to satisfy transparency needs. Proponents of using only inherently interpretable models argue that well-designed models of this type perform as well as more complex models and deliver the necessary transparency.[51] Given the limitations and potential implementation challenges in using *post hoc* techniques as described in Section 2.3.2 and subsequent sections, they also question whether relying on such techniques actually compounds the challenge of establishing the responsible use of AI and machine learning systems and meeting specific transparency requirements.[52]

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 still satisfying relevant model transparency needs.[53] Industry proponents also argue that even model types that are offered as more interpretable can run the risk of being too complicated for a human to understand and manage. For instance, while a decision tree may sound intuitively simpler than a neural network, they point to examples of trees with branches of 50 or 100 layers and ensembles of tree models, which are complex enough that explainability methods may still be necessary to meet transparency needs.[54] Others report that in the context of adverse action reporting, use of *post hoc* explainability techniques is helpful even for traditional and interpretable models to generate information about the key drivers of the underwriting model's estimation of default risk for individual applicants.

2.3.2 Key Explainability Techniques

While *post hoc* explainability techniques are not the only tools that lenders rely upon to manage machine learning models, they play a particularly important role in satisfying lenders' obligation to report the primary bases of adverse credit decisions, among other model monitoring activities. These techniques are supplemental models, analyses, or methods designed to describe the behavior of machine learning models after they have already been trained. They do not generally affect the design or operation of the underlying model and can be used with a variety of machine learning model types, although some implementations may work better than others with specific model structures.

This section will focus on feature importance explainability techniques (specifically Shapley Additive Explanations or SHAP), and surrogate models (specifically Local Interpretable Model-Agnostic Explanations or LIME).[55] While SHAP techniques have grown in popularity over time, stakeholders use a variety of approaches for different tasks and research is continuing to refine methodologies and explore additional ways of analyzing ML model operations.[56]

2.3.2.1 Feature Importance Explainability Techniques

Feature importance explainability techniques describe model behavior by determining how important each input feature is to the model's prediction. Shapley Additive Explanations—commonly referred to as SHAP—is a feature importance technique that has been particularly important in enabling lenders to use machine learning underwriting models. SHAP relies on techniques developed in game theory research to assess the impact of changing individual input features as a way to measure their significance to the model's prediction. This approach considers the various possible groups of features that can produce a prediction and assigns a score—or Shapley value—to capture each input feature's aggregate importance in the model.

In a simple underwriting model with three features—A, B, and C—this will involve calculating the possible combinations in which one of the three variables is omitted and determining how the default prediction changes in each iteration of the exercise. If the default prediction is 5 % when only A is present, 9 % when A and B are considered, and 11 % when A, B, and C are considered, this suggests that the relative contribution measured in the assigned Shapley value for this combination of input features is 5 for A , 4 for B, and 2 for C. To generate the overall Shapley value for A, B, and C, this exercise will be repeated for all combinations of A, B, C, AB, AC, BC, and ABC. In the end, values where A is present will be averaged and subtracted from the average value when A is not present to determine its Shapley value. The critical feature of Shapley values is that they can be aggregated across a large number of similar simulations, making them a useful tool for summarizing the contribution of individual input features to the overall prediction of a machine learning model.

However, various technical and implementation issues can affect the reliability of information produced by feature importance techniques. While SHAP can in theory precisely account for feature correlations, the number and complexity of the required calculations prompt users to adopt sampling techniques and other methods of reducing computational demands. Although versions of SHAP that are tailored to particular model architectures can be more efficient (which are sometimes called SHAP "implementations"), approximation or sampling methods are often used with some tradeoffs in the quality of explanations. If too few samples are used, the resulting SHAP values could be noisy and not reflective of actual model behavior.[57] Data scientists are also working to evolve techniques based on SHAP and other approaches to use more realistic assumptions about distributions and correlations in underlying data.[58]

More broadly, the fact that feature importance methods explain complex models by reference to their input features also raises questions about whether they are sufficient to detect and convey key aspects of model operations, particularly in situations where different input features are components of or contribute to other input features and situations in which relationships derived *inside* the "black box" may be the key drivers of the model's prediction. Fully expressing the role that features derived within the model play in determining the model's prediction can be both technically and practically challenging.[59] For example, a machine learning model might determine that late payments on mortgage loans are highly predictive of default risk when an applicant has an outstanding balance above $200,000. Feature importance methods might identify the number of late payments and outstanding mortgage balance as the key input features. However, reporting either or both of those as individual reasons on an adverse action notice may not fully convey the model's operations, since the combined effect of the features together caused some individual consumers to be predicted as high risk. As described further in Sections 4-6 stakeholders are grappling with the importance of being able to pinpoint feature interactions for multiple compliance requirements.

BOX 2.3.2.1 TYPES OF EXPLANATIONS

Model explainability is sometimes defined as the understanding of how a model makes its predictions.[60]

Different types of explanations may be important for different purposes. **Global explanations** shed light on a model's overall decision-making processes, relevant for assessing a model's suitability for a specific task. **Local explanations**, on the other hand, clarify the reasons behind specific decisions made by the model, such as predicting default risks for individual applicants.[61]

Explanations can either be **"true to the data,"** aiming to unveil the causal relationships between a feature and default risk, or **"true to the model,"** which analyzes the (mechanical) prediction process even if the model doesn't capture the underlying causal dynamics.[62]

The necessity for both global and local explanations, along with the examination of a model's meaningful relationships and its prediction process, underscores the importance of choosing the right tool for a particular task.

The type of explanation also holds policy relevance. For instance, as discussed in Section 5, when considering adverse action disclosure requirements, true to the data explanations might be more useful in helping recipients lower their default risks, whereas true to the model explanations could help identify errors in the information used by lenders for decision making.

2.3.2.2 Surrogate Models

Surrogate models are also sometimes used to explain uninterpretable or black box models, such as large tree ensembles (including XGBoost) or deep neural networks. Surrogate models are designed to mimic the original or underlying model, and they are trained on predictions from that model. However, surrogate models generally have characteristics that make them easier to understand—for example, they may be more parsimonious and explainable than the model they are being used to explain.[63] In practice, surrogate models are often shallow decision trees, rule sets, or regression models.

The popularity of local surrogate models to explain why ML underwriting models make individual default predictions has declined over time as many market actors have concluded that Shapley values have conceptual and performance advantages, but the models are still used for some other purposes in credit and in a variety of non-credit contexts.[64] Local Interpretable Model-Agnostic Explanations—or LIME—is one of the most widely used surrogate model methodologies and has influenced the development of other *post hoc* explainability techniques. LIME uses local linear surrogate models around a particular data point to approximate the complex model's output.[65] The resulting local surrogate models are used to both explain the model's behavior as applied to individual applicants and to quantify feature importance for inputs to the overall model.

However, 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 only (1) *or* (2) alone are not at high risk of default. A linear model cannot represent this effect in the underlying model.

LIME is versatile and adaptable since it can be used to explain a variety of types of models. 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 relying on 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 address this limitation, LIME uses a local surrogate model instead of trying to mimic the underlying or original model at all points and builds

a separate surrogate model for each explanation it produces. Different implementation choices can also be made with regard to application of LIME techniques.[66]

2.4 Broader Questions about Responsible Use

Fairness and transparency are two critical components of broader debates about whether and how machine learning models can be used responsibly in such a sensitive use case as credit underwriting. In other sectors and countries, stakeholders often include them in longer lists of qualities or principles to define what constitutes "trustworthy" artificial intelligence or machine learning along with such qualities as reliability/accuracy/robustness and oversight/governance/accountability. (See Box 2.4.1) Although the trustworthiness label has not been widely adopted by financial services stakeholders, their basic conceptual concerns about responsible use of ML models in credit underwriting and financial services more generally are quite similar.

These considerations are woven through debates about how to comply with specific regulatory regimes in the financial services context—as discussed further in Sections 4 to 6—but also operate at a more fundamental level to shape firms' and regulators' attitudes about the merits of adopting machine learning models in the first instance. In effect, they can present a chicken-or-egg conundrum: Stakeholders may be less motivated to adjust market practices and regulatory frameworks to account for ML adoption if they are not convinced that ML models can be managed responsibly to create substantial improvements over the status quo, and yet the nature of market practice and policy adjustments will help to define responsible use and determine the nature of realized benefits. As ML adoption increases in some market segments, pressure is increasing on the full range of credit ecosystem stakeholders—including skeptics—to grapple with these broader questions.

As reflected in the more detailed sections that follow, one recurring conceptual issue concerns the importance of human agency in *understanding* underwriting models. As discussed in Section 2.3, it is difficult for current explainability techniques to detect and fully convey feature interactions and relationships inside the most complex models. Data science techniques can often provide alternative means of analyzing model operations, and various measures can be used to monitor model outcomes. But stakeholders are debating whether it is critical to be able to understand ML underwriting models at the same level and in the same ways that can be achieved with logistic regression models for their results to be considered reliable and fair in a broad sense. At the highest level, these questions ask whether it is appropriate to rely upon models that we cannot fully understand.

A second and closely related issue concerns whether we can *manage* models that we cannot fully understand to address specific business and policy goals. Traditional strategies may not be as effective when applied to models that involve hundreds or thousands of input features—and potentially even more interactions or latent features within the model. Simply removing or transforming a single data point may not have the same effect as in traditional models, since one feature out of thousands is likely to make a smaller marginal contribution to the model's prediction even when correlations are accounted for. And in more complex models, the feature interactions inside the "black box" may be more critical to the models' operation relative to individual input features in isolation. This may require different approaches to optimize for particular business and policy goals at the beginning of the development process rather than making small adjustments at the end.

Whether we can achieve a reasonable level of understanding and confidence in ML models for underwriting and other financial services activities is thus essential to both the business case for adoption and debates over whether and how to adapt specific regulatory frameworks. Yet while the complexity of ML models presents substantial conceptual and practical challenges, it is also important to recognize that this moment presents opportunities for improving the standards, processes,

BOX 2.4.1 TRUSTWORTHINESS FRAMEWORKS

Dialogues about the qualities of trustworthy AI have started across multiple jurisdictions, markets, and use cases with an eye toward facilitating a broad consensus about responsible use of AI/ML technologies as well as creating a starting point for adapting sector-specific technological standards and regulatory requirements.

Reliability, fairness and inclusion, and transparency are among the most common qualities and principles cited across these various initiatives. For example, the European Union's recently adopted framework for regulating AI/ML builds 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.[67]

Other general frameworks for ethical/trustworthy AI have been published by the Alan Turing Institute under a commission by the United Kingdom's Financial Conduct Authority (five principles of AI ethics, which include fairness, sustainability, safety, accountability, and transparency); the U.S. National Institute of Standards and Technology (seven qualities: valid/reliable, safe, secure/resilient, explainable/interpretable, privacy-enhanced, fair/harmful bias managed, accountable/transparent); and the Organisation for Economic Co-operation and Development (six key principles: inclusive growth, sustainable development and well-being; human-centered values and fairness; transparency and explainability; robustness, security and safety; and accountability).[68]

and tools that were developed to manage prior generations of predictive models in financial services. In addition to new data science techniques for explaining and debiasing models, ML adoption is helping to highlight and fuel debates about unresolved questions and tensions in existing business and regulatory practices.

Stakeholders who support the adoption of ML models emphasize the importance of considering the disadvantages and tradeoffs that are embedded in traditional underwriting systems and compliance regimes when assessing both the potential risks and benefits of adopting new techniques and data. For example, judgmental underwriting by individual loan officers is both opaque and subject to risks of bias and inconsistency, as well as being difficult to scale efficiently. Prior generations of automated predictive models have increased consistency and scale, but are subject to data limitations and concerns about the effectiveness of current practice and regulatory frameworks with regard to fairness and inclusion, empowering and educating borrowers, and other topics. Moreover, different stakeholders' understanding of the models varies substantially in practice due to information asymmetries both within and among firms, their regulators, and their customers.

Thus, the debates about ML adoption may present opportunities to consider expectations for *all* types of underwriting models to determine how best to effectuate business and policy goals. This report speaks both to broader questions about the responsible use of machine learning underwriting models and to questions about adapting specific regulatory frameworks (model risk management, adverse action notices, and fair lending) to account for the growing use of machine learning and associated explainability and debiasing techniques. In so doing, it helps to surface opportunities to enhance frameworks, governance, and oversight for purposes of promoting fundamental goals such as responsible risk-taking, transparency, financial inclusion, and anti-discrimination.

3. OVERVIEW OF FINREGLAB'S RESEARCH

The issues discussed in Section 2 informed FinRegLab's decision to interrogate available techniques and tools for managing explainability and fairness concerns with machine learning underwriting models. The purpose of the project has been 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.

We conducted both quantitative and qualitative research to support the policy analyses provided in this report:

- We partnered with Stanford Business School professors Laura Blattner and Jann Spiess to produce an empirical white paper analyzing the ability of *post hoc* explainability techniques and other model diagnostic tools to help lenders understand and manage machine learning credit underwriting models. Our empirical work evaluated proprietary tools offered by seven technology companies–Arthur AI, H20.ai, Fiddler AI, Relational AI, Solas AI, Stratyfy, and Zest AI–as well as open-source techniques as applied to various tasks relating to model risk management, adverse action disclosures, and fair lending compliance.
- We conducted extensive interviews and engagement with a broad range of stakeholders to explore the implications of ML underwriting models for market practice and regulatory frameworks. In addition to convening a project advisory board to inform the initial organization of the project, FinRegLab in 2022 co-sponsored a symposium with the U.S. Department of Commerce, National Institute of Standards and Technology, and the Stanford Institute for Human-Centered Artificial Intelligence and organized three policy working groups to discuss key aspects of ML adoption. The policy working groups convened more than 75 representatives of lenders, data and technology companies, advocacy organizations, researchers, and other stakeholders to engage in extended conversations about both the challenges and opportunities associated with adoption of machine learning in credit underwriting. Representatives of federal banking regulators and the Consumer Financial Protection Bureau attended the sessions in an observer capacity.[69]

Section 3.1 provides an overview of our empirical methodology and global findings, but additional detail on the findings for each regulatory compliance area can be found in Sections 4 to 6

and in the underlying Empirical White Paper.[70] Additional discussion and market and data science context can be found in our other reports from this project.[71]

3.1 Empirical Summary

Our empirical research applied various explainability techniques and model diagnostic tools to a range of credit card underwriting models to perform various diagnostic and management tasks relating to the three regulatory compliance topics. While several other studies have demonstrated that explainability tools can produce information about machine learning credit models, they did not systematically evaluate these explanations in the context of regulatory compliance.[72]

The underwriting models were built using a representative sample of data from a nationwide credit bureau from 2009 to 2017, and included four models of varying complexity that were built by the research team as well as several machine learning models built by the participating companies using the same data.[73]

For each regulatory compliance topic, the research team then analyzed the explainability tools' outputs for three primary qualities:

- **Fidelity:** The ability to reliably identify features that are relevant to a model's prediction for the particular regulatory purpose (adverse action disclosures, fair lending compliance, or model risk management).
- **Consistency:** The degree to which different tools identify the same features to be important when they were applied to the same model.
- **Usability:** The ability to identify information that helps the user (whether a consumer or a lender depending on the circumstances) perform certain tasks, such as improving their future chances of credit approval or managing the model to address a specific regulatory concern.

We viewed fidelity and consistency as threshold technical questions about the tools' reliability, with fidelity playing the most important role. Across each of the different regulatory topics, one of the ways that we tested fidelity was to change the values in the underlying data for the features identified by a particular tool as important for the regulatory purpose and measure the effect on model predictions as compared to changing (or "perturbing") the values of features that were chosen at random or that were closely correlated to the features that had been identified as important.[74] In evaluating consistency, we compared the extent to which different tools identified the same features as important for purposes of the particular regulatory task.[75]

The usability tasks and analyses were the most varied and complicated because they tested not just what information was returned by explainability tools but how the information could be used to achieve particular goals. In the adverse action context, we focused on whether the explainability techniques could be used to identify a set of plausible changes in credit report metrics over twelve months that would cause consumers to meet thresholds for their applications to be approved. For fair lending, we evaluated recommendations by the participating companies for how to reduce racial disparities in model predictions, some of which were predicated upon information generated by the explanations and some of which were not. For model risk management, we assessed the tools' ability to determine why model performance changed when applied to data from different time periods.

As discussed in greater detail in the following sections and the empirical white paper, our empirical analyses found that some but not all of the explainability tools we tested could reliably identify features that were important to models' behavior for particular tasks. The explainability tools with

the highest fidelity generally tended to perform well when applied to different model types and to both simple and complex models. Notably, however, the gap in performance between higher fidelity and lower fidelity tools tended to be most pronounced when applied to complex models, which suggests that the choices that lenders make about which diagnostic tools to use and how to apply them becomes even more important when the models involve large numbers of features and complex techniques and architectures.

We also found that the explainability tools with the highest fidelity tended to identify more of the same features as important to the model than tools that performed poorly on fidelity tests, although there were still some variations among the higher fidelity tools particularly when they were applied to more complex models. This pattern appears to be driven in part by the fact that more complex models incorporate a large number of features that are closely correlated to each other. The level of consistency in identifying "important" features in more complex models improved substantially once we accounted for broader feature families and correlations, for example by grouping, or aggregating, features focusing on 30-, 60-, and 90-day delinquencies into a broader "delinquency" category.

While the results were encouraging, it is also important to note that no one tool performed the best across all regulatory tasks and topic areas (i.e., adverse action, fair lending, and model risk management). This underscores the importance of lenders selecting the right diagnostic tool for specific tasks and making thoughtful decisions about deployment. For example, while many tools relying on SHAP feature importance measurements performed well, some did not. The research suggests that the combination of different SHAP implementations and different sampling methods could lead to variations in the response. Further research would be helpful as academics and private sector stakeholders continue to develop new approaches and iterate on existing options.

These and other empirical results underscore the importance of interpreting the outputs of diagnostic tools in light of the broader relationships within the data. Because features that a particular tool identifies as "important" serve as approximations for patterns in model behavior that are linked to both the identified features and other features that are correlated with them, other features may also be making important contributions to model outcomes. Thus, assuming a single feature within a correlated cluster is the sole driver of model behavior is likely incomplete. This speaks to the importance of lenders having a strong understanding of the data that are being used to build, train, and deploy ML models for credit underwriting decisions.

In the fair lending context, our results also explored the potential benefits of new debiasing techniques for reducing disparate impacts. Where we tested approaches that relied on traditional mitigation strategies by identifying and modifying a narrow subset of features, we found that model performance declined with little to no improvement in fairness. But more automated approaches were able to produce a menu of options that provided larger fairness benefits and smaller accuracy tradeoffs, as discussed in Section 6. While we did not test the full spectrum of approaches, our findings illustrate the more powerful toolkit that new data science techniques can provide in searching efficiently for fairer models.

The next three sections provide more detailed discussions of explainability and fairness concerns regarding machine learning underwriting models for each of the three regulatory areas that were a focus of our empirical study. Each section provides regulatory and operational context before discussing our research findings and policy implications and debates more generally.

4. MODEL RISK MANAGEMENT

The model risk management (MRM) framework has been imposed by federal prudential regulators to promote responsible risk-taking in the banking sector by requiring both a comprehensive risk assessment prior to adopting new models and the implementation of monitoring plans and controls after deployment.[76] It not only provides a regulatory scaffolding for considering the robustness of models through stress testing and other analyses, but in effect serves as a broader governance framework for determining the trustworthiness of models of all types—not just ones used for underwriting or built with machine learning techniques. While the framework does not formally apply to nonbanks, elements may be adopted as a matter of best practice or required by investors and other contractual partners. Thus, it is little surprise that the MRM framework has become a focal point in discussions about the responsible use of AI/ML across the financial services sector.

Notions of transparency are deeply interwoven in the MRM framework, ranging from extensive documentation of model development, validation, and monitoring to processes for analyzing the conceptual soundness and performance of features that play a key role in model operations. This section focuses primarily on core model risk and governance issues that are implicated by broader concerns about the transparency of machine learning models and does not provide a full treatment of all MRM compliance considerations for ML underwriting models or for the use of AI/ML in financial services more broadly. It begins by summarizing regulatory and operational context in Section 4.1 before addressing selected policy issues in greater detail in Section 4.2. These topics include:

- The potential crossover between traditional model risk management guidance and broader trustworthiness frameworks;
- Standards for evaluating explainability techniques and conducting conceptual soundness reviews;
- Governance standards for nonbanks; and
- Validating vendor-provided models and explainability tools.

4.1 Regulatory and Operational Context

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 in order to promote responsible risk-taking.[77] This guidance applies broadly to the range of model use cases that might create unexpected losses, compliance problems, or other negative outcomes for the depository institution and calls for enterprise-wide risk management processes

including governance, policies, and controls. This framework makes firms responsible for documenting their development and validation processes, model limitations, and monitoring and mitigation strategies. It also requires firms to provide independent review of such documentation to help ensure that the institution is not exposing itself to unnecessary risk because of an erroneous or substandard model, model development process, monitoring plan, governance, or controls.

In practice, financial institutions calibrate efforts to evaluate and monitor risks related to models based on the degree of risk posed by the particular use case. Credit underwriting is often considered to be among the highest risk activities. Thus, for depository institutions, model risk management expectations typically require extensive pre-deployment review of credit underwriting models and monitoring during use, especially for firms that emphasize retail or consumer banking. For nonbank financial institutions, bank regulatory expectations may broadly inform aspects of their model oversight practices, in part because funding and securitization partners may require some of these processes and practices in their contracts. The same may be true in varying degrees for nonbank financial institutions that are public companies.

Prior to adopting a new model, MRM validation processes focus on two complementary components, each of which try to surface potential weaknesses in the model from different perspectives:

- **A review of the conceptual soundness of the proposed model evaluates the proposed model's detailed structure to assess robustness and stability.** This involves assessing the appropriateness of data sources, the suitability of the model structure and estimation methodology in light of how the model is expected to be used, the theoretical grounding of input features and estimated relationships in the model, features' statistical significance, and the intuitiveness of their directional impacts, linear/non-linear relationships, and relative magnitudes, as well as the model's consistency with business objectives and policies. Conceptual soundness can also involve quantitative testing of models for issues not addressed by outcomes assessments.
- **An outcomes assessment evaluates the model's performance under various scenarios to assess how its prediction might change.** This analysis tries to isolate potential anomalies in the model's performance based on the variability and potential trends in error rates, particularly over time and in different conditions. The assessment is typically performed for the model overall and for different sub-groups (such as different risk factor ranges or by time periods). In addition to backtesting against data from other historical periods, outcomes analysis also includes stress-testing and simulations whereby the model's predictions are evaluated over input combinations that may not be part of the training, validation, or test data. In assessing performance, stability, and robustness during validation of a proposed model, the validator is looking for evidence of prediction anomalies—model outcomes that appear to be illogical, unreasonable in direction or magnitude, or inconsistent with other benchmarks.

Different aspects of model transparency play a particularly important role in conceptual soundness processes. At a broad level, the guidance requires documentation of the processes by which a model is developed, validated, and monitored during deployment.[78] For traditional underwriting models, conceptual soundness reviews have traditionally been done in part by assessing whether the model relies on relationships that are empirically sound and draw on appropriate institutional experience, industry practice, and relevant economic theories.[79] Empirical analyses often focus on how well developers have balanced using features that offer predictive power while reducing redundancy in correlated features and managing other types of risks, such as consumer protection risk due to adverse actions notice requirements and fair lending enforcement risk due to disparities for protected classes.

BOX 4.1.1 MODEL PERFORMANCE, STABILITY, AND ROBUSTNESS

Outcomes assessments typically focus on three qualities:

- **Performance:** A model's performance refers to its effectiveness in making accurate predictions using appropriate metrics depending on the context.[80] Performance is often measured by back-testing additional data, whether drawn from the sample, from outside the sample during the same time period, or from a different time period. Performance metrics are often reviewed in comparison to a benchmark, such as a traditional credit score or the performance of an underwriting model currently in use, and often include metrics designed to capture economic performance, such as profitability metrics like net interest income and net charge-offs.
- **Stability:** Model stability refers to the model's ability to deliver consistent performance in the presence of changing conditions. In many cases, lenders use an out-of-time dataset from periods of economic stress that are of particular interest. These datasets are split up into monthly segments and performance metrics are calculated for each of the time-based segments. Significant variance in performance metrics over these segments indicates less model stability. In some cases, stability may also be assessed by examining whether there are significant variations in the coefficients for individual features when the model is applied to different samples and time periods.
- **Robustness:** Model robustness evaluates the extent to which assumptions made during the model development process impact model stability. Alternative models are often evaluated to determine their relative sensitivity to changing inputs and economic conditions.

For machine learning models, conceptual soundness processes include documenting how the learning algorithm produced the final model. Firms often produce this information by using new techniques and analyses used to retroactively unpack information about how the model works, such as the feature importance methods described in Section 2. In this context, the Office of the Comptroller of the Currency has recognized that:

> An evaluation of conceptual soundness may be difficult for some complex models (e.g., those that use AI approaches) because the underlying theory and logic may not be transparent. Transparency and explainability are key considerations that are typically evaluated as part of effective risk management regarding the use of complex models. The appropriate level of explainability of a model outcome depends on the specific use and level of risk associated with that use.[81]

4.2 Selected Policy Topics

Because traditional model risk management guidance provides a flexible approach to managing risks across a wide variety of bank models, it can provide a useful framework for managing the transition to ML techniques and new diagnostic and debiasing tools. Indeed, it is often cited as a potential tool for other sectors to consider in adopting AI/ML applications for other use cases.[82] At the same time, the existing prudential guidance was drafted before ML adoption for underwriting and other contexts accelerated, and some financial services stakeholders suggest it could benefit from refreshing and elaboration. Federal financial regulators sought feedback in 2021 on whether updates would be useful to MRM or consumer protection guidance but have not yet released a specific proposal.[83]

As this process plays out, one question is whether there is potential value in supplementing the MRM guidance at a broad conceptual level, similar to the frameworks for responsible AI that are emerging in other sectors and jurisdictions as discussed in Box 2.4.1. Stakeholders have also identified more specific areas of potential focus with regard to enhancing model risk management and general governance for ML adoption, including: (1) standards for evaluating explainability

techniques and conducting conceptual soundness reviews of ML models; (2) governance standards for nonbanks; and (3) challenges in validating vendor-provided models and tools. Each of these topics is addressed below.

4.2.1 Updates to General MRM Frameworks

The broader frameworks for trustworthy or responsible AI that are emerging in other jurisdictions and sectors as discussed in Box 2.4.1 are not always as detailed in articulating processes for risk mitigation across specific stages such as development, validation, and deployment as existing MRM guidance, yet they articulate a broad range of considerations in developing and deploying ML models such as reliability, transparency, fairness, privacy and security, and accountability. The existing MRM guidance emphasizes that "model risk" includes potential business/strategic and reputational damage as well as financial losses, yet it is most concrete in focusing on accuracy, robustness, and other primary performance considerations and on governance processes.[84] While other bodies of federal regulatory guidance address topics such as fair lending, privacy, and information security, that guidance was not developed to address ML adoption specifically and varies as to its focus and scope of jurisdiction.[85]

As a result, there is no single high level conceptual list of qualities or risks that should be evaluated in determining whether to adopt a particular ML model for use in credit underwriting or other financial services use cases.[86] Some stakeholders suggest that adopting a common sector-wide framework of core considerations for ML models could increase the general consistency of practice, provide useful guideposts for rapidly evolving areas and topics not covered by more detailed guidance, and facilitate the identification of situations in which existing frameworks and practices need to be tailored for particular use cases or technologies. For example, a common framework could encourage lenders to develop protocols for evaluating machine learning models as to various qualities and to engage in an ongoing dialogue with regulators about their analytical methodologies and substantive results. Other stakeholders are skeptical, suggesting that the relevant concerns are so use-case specific that a general list would not be helpful.

Broad debates about responsible AI are also prompting some stakeholders to call for a renewed focus by lenders on standardizing their procedures and documentation for model development, pointing to survey results suggesting that even large financial institutions have not necessarily articulated precisely which algorithms and approaches are permissible to use in processing data, building models, and managing concerns about transparency and other risk topics.[87] At least one company has begun using blockchain as a means of articulating development standards and documenting model development decisions.[88]

4.2.2 Standards for Evaluating Explainability Techniques and Conducting Conceptual Soundness Reviews

The adoption of machine learning techniques is also prompting more specific debates about MRM standards for use of particular explainability techniques and conducting conceptual soundness reviews more generally. Reviewing the conceptual soundness of traditional models focuses on measuring individual features and relationships against conventional measures of statistical validity and evaluating their alignment with economic theory and industry and institutional experience. Model developers typically provide documentation that includes descriptions of each feature in the model and the basis for its inclusion, such as empirical testing that establishes the ability of a feature to predict an outcome and an intuitive justification for the relationship of that feature to the outcome.[89] The notation of logistic regression models themselves identify both the input features and their weights, and further statistical analyses are widely accepted ways for establishing conceptual soundness. The combination of selecting features based on economic, behavioral, and other theories and the empirical information provided by

the weights and familiar tests has prompted some commentators to suggest that "conceptual soundness is a fundamental property of econometric models by their very nature."[90]

Financial institutions using machine learning underwriting models have therefore had to adapt their policies and procedures for establishing conceptual soundness to account for differences in the development process, the number of features in the models, and model transparency, as discussed in Section 2.3. Although practice is varied, establishing the conceptual soundness of machine learning models can require more effort than for logistic regression models for a variety of reasons:

- Use of more features and more complex models may require more review than incumbent modeling methods as well as involve new analyses and new or different personnel, all of which can elongate review periods;
- Explainability tools such as SHAP are often required to generate information to facilitate a more thorough review of how the model operates, which requires lenders to decide which tools to use and to train analytics staff and second-line review teams on how to use and interpret output from these tools;
- Use of machine learning models often increases scrutiny of the model because of concerns that adoption of new technologies can increase inherent risk across several risk areas considered within the MRM framework.

Among industry and other stakeholders, questions about the application of specific explainability techniques are just one component of broader debates about the ability to conduct conceptual soundness reviews of ML models and of the utility of more specific model risk guidance. This section outlines how conceptual soundness reviews of ML models are occurring (including the roles that explainability techniques can play), criteria for evaluation and selection of such techniques, and broader debates about conceptual soundness of ML models.

4.2.2.1 Evolution of Conceptual Soundness Review Processes

Conceptual soundness reviews for machine learning models in part cover similar terrain as they do for traditional models—such as selection and treatment of the data and articulation of the outcome the model is designed to predict—yet differences in how machine learning models are developed drive the use of different analyses and techniques to assess and document a model's conceptual soundness.

A conceptual soundness review for a machine learning model begins with an assessment of the suitability of the data for modeling. This step addresses questions such as whether the information is representative of the population the model is expected to evaluate, whether sufficient controls are in place to protect the integrity of the data, and if the information is of sufficient quality and reliability in light of the use case and architecture of the proposed model.

The next steps scrutinize both the data point or outcome that the model is designed to predict (often called the modeling target) and the input features in the training data set that will be used to predict the modeling target. Underwriting models are typically probability of default models, which predict the likelihood of a binary outcome such as whether a loan will go 90 days delinquent in the first 24 months of the loan. Justification must be provided as to the choice of target variable, the performance window, and any reject inference methods used during the development process.[91]

With regard to the input features, validation in both traditional and machine learning model development involves assessing relationships in the available data in light of business objectives,

risk tolerance, and prior experience. The process of selecting these features, justification for the number used, the definition of individual input features, methods of calculating individual input features, and the accuracy of those calculations are all subject to review. The process also includes consideration of the intuitive connections between the input features and the target. For example, does it make sense that the number of delinquencies on past obligations could predict the likelihood of default?

Next, model validation teams consider developers' choice of modeling method. Model developers must justify the choice of model architecture, modeling method, and learning algorithm selected. This inquiry relates to how the model development team elected between options such as a logistic regression, a gradient-boosted tree, or an artificial neural network. For example, using a neural network trained on a data set that is small and narrow may present inappropriate risk because of the risk that "overfitting" would cause performance benefits to deteriorate rapidly in the face of changes to economic conditions, customer populations, or other data. Often, when a machine learning model is proposed, developers create both a set of potential ML models and a new model using traditional techniques to help assess whether there is a need to take on a more complex model and, if so, which one best fits the lender's overall objectives. The more complex ML model will generally be proposed for use only where it yields greater predictive performance than the less complex model, although individual firms may make different judgments about the size of improvement necessary to justify use of the more complex model.

At this point, the steps in the conceptual soundness review begin to diverge from those used with traditional models. When a machine learning model is used, model hyperparameter choices are also a major point of focus.[92] For an XGBoost model, this might include justification for the number of trees and the maximum tree depth, for example. These choices are also typically justified by building alternative models using different hyperparameter choices and empirically demonstrating that the selected value yields the best performance.

The relationships the model learned between the input features and the target are also subject to different forms of scrutiny to understand the importance of individual features to the operation of the ML model. The number of features and more complex architectures used in many ML models complicate producing the same kind of feature-by-feature analyses that are typically generated for traditional models with a relatively small number of inputs. Here, practice varies, but in most cases firms use a range of analyses for this component of conceptual soundness. Three main approaches are considered here: use of Shapley values, partial dependence plots ("PDP Plots"), and individual conditional expectation plots ("ICE Plots").[93]

- **Shapley Values.** Shapley values can be used to derive a holistic picture of model behavior and to illuminate the contribution each input feature makes to the model's predictions overall and for a given segment or population. This review helps lenders review and assess whether the contributions of each variable to model predictions are intuitive and whether the model is overly reliant on any one feature or category of features which may be correlated. For example, this analysis can determine whether the model relies on positive repayment behavior more than negative repayment behavior, or whether the model relies more on self-reported application data, internal bank data, or data from credit bureaus. Analysis of the contribution of each variable or category of variables to a model's predictions can provide insight into how robust the model will be in the face of missing data, errors in the data, or temporary changes in conditions. This analysis may be repeated for various segments of the population to assess how changes in the contributions of input features to model scores affect different groups.

- **PDP Plots.** To understand the relationship between the input features and the model score, partial dependence plots associated with each of the input features are reviewed. The plots group different values of the feature into bins along the X axis and display the average model prediction associated with each bin. Partial dependence plots allow model developers and internal and external reviewers to examine the trend in model predictions associated with values of a particular variable. These plots are inspected to determine whether the model has learned a predictive relationship that meshes with the intuitions and collective experience of model developers, second-line reviewers, and other stakeholders. For example, examining the partial dependence plot for an input feature such as the count of past delinquencies would allow a model developer or reviewer to determine whether, on average, a model predicts a higher likelihood of default for applicants with more past delinquencies.
- **ICE Plots.** Another tool used to examine the relationship between an input feature and a model's prediction is the individual conditional expectation or "ICE" plot. ICE plots examine a model's predictions for each observation in the dataset by substituting a range of values for a given input feature, while keeping all other features constant. These further help model developers and reviewers assess whether a model is behaving in ways that are intuitive and predictable based on their individual and institutional experience. For example, an ICE plot would allow a reviewer to verify that when a selected applicant was scored with 0 bankruptcies the applicant would get a lower likelihood of default than if they had 1 bankruptcy, and that the model would assign a higher likelihood of default as the number of bankruptcies increased to 2, 3, 4 and so on.

For models that are deemed to meet conceptual soundness requirements, the validation teams also review the model monitoring plan proposed by model developers. The plan is intended to ensure that the model is operated safely and within the firm's risk tolerance. Monitoring plans identify key risks related to the use of the proposed model, define metrics and processes for monitoring those risks, and define steps to be taken when, for example, substantial changes in performance or data or score distributions occur. Factors that are critical in the conceptual soundness review—such as the model's use case, architecture, and level of complexity—shape model monitoring frameworks and in some firms, stakeholders report that review of monitoring plans is part of the conceptual soundness review.

4.2.2.2 Relying on *Post Hoc* Explainability Techniques and Tools

Although the mathematical and theoretical underpinnings of explainability tools now in use are hardly new and approaches like PDP and ICE plots are derived from analyses used with traditional underwriting models, the prominence of explainability techniques to understand and manage machine learning underwriting models is a relatively recent development.[94] To date neither firms nor their regulators have had to articulate a consistent formal framework to govern the evaluation of these tools, although banks typically will perform some level of assessment before authorizing model development teams to use particular techniques and are likely discussing them with examiners in the context of regulatory reviews of machine learning underwriting models.

One approach to ensuring that lenders relying on *post hoc* tools are appropriately assessing and managing risks related to their use is to treat them as a separate model for MRM purposes.[95] This would require that each use of a tool—whether it be to help produce adverse action notices or illuminate the conceptual soundness of a model—be separately reviewed, validated, and monitored. In other words, users would have to demonstrate the tool's fitness for use in the proposed application by demonstrating conceptual and technical soundness, although the specific components of each inquiry and their overall rigor could be tailored as appropriate given the nature of the model at issue

BOX 4.2.2.2.1 FINREGLAB'S MODEL RISK MANAGEMENT EMPIRICAL RESULTS

FinRegLab's analysis of model diagnostic tools in the risk management context assessed the ability of the tools to identify features that described a significant part of overall model behavior, including application to data from a different time period. Our results suggested that the tools could potentially be helpful to lenders when conducting some MRM analyses.

We started by assessing the tools' fidelity, in this context whether they could reliably identify features that helped describe global model behavior. We used perturbation tests to determine whether changing the features identified by a particular tool as most important to model operations created a bigger impact on default predictions than perturbing random or closely related features. For each model type, we found that some tools beat those benchmarks, while others did not.[96]

We then evaluated consistency across tools. The tools that performed the best on the fidelity tests tended to have greater consistency in results, and grouping similar or correlated features together helped to increase the consistency of the tool outputs particularly for models involving hundreds of features.

The last analysis analyzed the tool outputs when models were applied to an out-of-time data set. We found that a variety of approaches outperformed random benchmarks in identifying features that may explain deterioration in the models' performance and assess whether such deterioration was driven by shifts in the population distribution, shifts in the underlying relationships among the features, or a combination of both factors.

and the risk rating it receives under each firm's MRM policies. This approach may have benefits in the period before regulatory expectations are formalized in guidance or rules in that it allows firms substantial leeway in formulating tests and standards for assessing *post hoc* tools, while nevertheless providing regulators with consistent, reviewable documentation of pre-deployment assessments and monitoring information. However, depending on the nature and cadence of revalidation processes, the burdens of such an approach might appear particularly onerous to smaller institutions with less developed model risk management programs or less expertise in relevant data science fields.

Whether or not policymakers require *post hoc* tools to be individually and continuously validated, they might take a further step to articulate a consistent framework for evaluating individual uses of such tools. Defining a consistent set of target characteristics for *post hoc* tools might improve the overall rigor and consistency with which firms evaluate specific deployments of such tools and help generate a range of approaches to testing for each quality or assessment characteristic. In time, standardization of practice and oversight of explainability tools might help foster trust in the use of machine learning underwriting models.

In assessing potential frameworks and guidance for assessing *post hoc* tools, two aspects of the empirical research that FinRegLab conducted with Professors Blattner and Spiess may be particularly helpful. First, one of the challenges in applying explainability tools to complex ML models is that it is often infeasible to generate a complete explanation of the model's operations in order to verify the explainability tools' performance. Despite not knowing the ground truth explanation, however, we were able to design empirical tests that allowed us to compare several explainability techniques and vendor tools to each other and to objective benchmarks.

Second, the primary qualities we studied—fidelity, consistency, and usability—may be a useful starting point for thinking about and measuring the most critical qualities of explainability techniques. We viewed fidelity and consistency as threshold technical questions about the tools' reliability, with fidelity playing the most important role. For example, if a tool cannot reliably identify features that are important to a particular aspect of a model's operation, we would not necessarily expect or care whether its results were consistent with the results of some other tool in performing the same task. Usability is also a critical quality—indeed, in some ways ultimately the most critical for judging whether the tools can be used to assess or demonstrate regulatory

compliance—but also more complicated to define and evaluate. For instance, usability results may depend not just on the general nature of the information provided by a diagnostic tool and implementation choices made in its deployment, but also on what options are available to the user in responding to the information.

Our findings demonstrate that lenders can systematically evaluate *post hoc* tools to determine their potential fitness for use. The qualities that we tested for and the techniques that we used to perform the analyses may provide a useful starting point in helping to think through important implementation choices for credit underwriting and other contexts. While the elements of our analyses can be improved and expanded over time, defining a basic framework for what qualities are important to consider in choosing among tools and for how to test those qualities could be useful to both firms and regulators in moving toward more consistent implementation.

4.2.2.3 Broader Debates

As explainability techniques evolve and our understanding of them continues to improve, broader debates about demonstrating the conceptual soundness of machine learning models are continuing. For some, conceptual soundness is inherently tied to the analysis and justification of each feature and relationship in a proposed model as a means of testing underlying economic and behavioral theories. The fact that commonly used explainability techniques cannot directly and precisely map feature interactions within the most complex ML models prompts these stakeholders to question whether expectations regarding conceptual soundness (and other regulatory requirements as discussed in subsequent sections) can be satisfied in high stakes applications such as extending credit, at least absent using up-front constraints to create more transparency.[97] Others argue that the composite picture of model behavior derived by using new *post hoc* analyses, tools, and information can be sufficient to inform the responsible use of models using larger numbers of features and more complex architectures.[98]

This debate has its roots in a broader set of disagreements about the extent to which predictive models used for critical financial applications should be driven by human-led deductive processes or by patterns that are identified through inductive data analysis:

> For economists, few sins are more heinous than data-mining. It is the last resort of a scoundrel to engage in "regression-hunting"—reporting only those regression results which best fit the hypothesis the researcher first set out to test. It is what puts the "con" into econometrics. For most economists, such data-mining has unfortunate similarities with oil-drilling—a dirty, extractive business which comes with big health warnings.
>
> For data scientists, the situation could not be more different. For them, the mining of data is a means of extracting valuable new resources and putting them to use. It enables new insights to be gained, new products to be created, new connections to be made, new technologies to be promoted. It provides the raw material for a new wave of productivity and innovation, an embryonic Fourth Industrial Revolution.[99]

The evolution in conceptual soundness practices reflects similar tensions as the relative emphasis shifts as between interrogating the decisions that led developers to include particular features in the first instance and unpacking information about how the model behaves in practice, including identifying how input features affect the model's predictions. Both processes generally involve extensive analyses of data and alternative options, but in the ML context much of the critical information to describe the behavior of machine learning models is derived after the model is built. The shift affects not just what techniques are used at what times, but also potentially the relative emphasis placed

on understanding *how* versus *why* a model behaves as it does. Proponents of traditional approaches argue that analyzing the role of each individual feature and relationship in the model is preferable in part because it facilitates greater testing of potential scenarios and issues beyond the bounds of historical data.

Existing guidance does not specifically address these conceptual questions in the context of machine learning or *post hoc* explainability techniques, but it emphasizes more broadly that all models are by definition "imperfect representations of reality that all involve varying degrees of uncertainty and inaccuracy" and acknowledges it may not always be practicable for statistical tests to unambiguously reject or accept hypotheses or to quantify the degree of uncertainty or inaccuracy presented by a given model. The guidance does not foreclose use of models in such circumstances, but rather emphasizes the importance of applying a variety of tests and analyses throughout the model development life cycle and of tailoring risk management strategies with models that present greater uncertainty, for instance by building in conservative assumptions, using supplemental models or approaches, and/or increasing loan loss reserves.[100]

The guidance thus emphasizes both front-end and back-end analyses as critical to risk management. In practice, stakeholders emphasize that the approach for demonstrating the conceptual soundness of AI and machine learning models during initial development and validation will necessarily be specific to individual model types and that acceptance of approaches to evaluating some types of models may be closer than others. For example, a consensus as to ensembles of tree-based models may be closer than one for neural networks, given that tree-based models have been more widely adopted in the lending context and can be somewhat less complex that neural networks.

4.2.3 Governance Standards for Nonbanks

Stakeholders often praise the general utility of the MRM framework in fostering a range of responsible model development practices and rigor in the approval and monitoring of models.[101] The transition to wider use of machine learning underwriting models has focused attention on a common thread in those conversations: the lack of formal model governance expectations for nonbank financial institutions. Some stakeholders suggest that the complexity involved in developing and operating machine learning underwriting models increases risks related to lack of parity between bank and nonbank requirements and heightens the urgency of the need to formalize expectations and oversight mechanisms for nonbanks regarding managing model-related risks through activities like pre-deployment model review and model monitoring.

This gap in regulatory requirements does not mean that lenders outside the banking sector operate without safeguards regarding model-related risks, especially for models like underwriting models that expose the firm directly to financial losses and reputational risk. Indeed, core business incentives may lead nonbank lenders to adopt a range of model development and oversight practices that approximate some components of a bank model risk management program. Many nonbank lenders also enter into contracts with investors, bank partners, or others that require them to adopt bank-like model risk management practices.

Nevertheless, given the high stakes of credit underwriting and complicated issues that transitioning to ML underwriting models entails, a diverse group of stakeholders has suggested that amending existing law to impose basic governance expectations on nonbank adopters could be beneficial to borrowers, lenders, and the broader ecosystem. For example, they argue that such a change would both help to ensure that nonbanks are managing their models for a consistent range of risks in developing and deploying ML underwriting models and to level the playing field.[102] It is

BOX 4.2.3.1 LEGISLATIVE INTEREST IN ALGORITHMIC IMPACT ASSESSMENTS

Legislators in both the European Union and the United States are looking to impact assessments as a core procedural safeguard when deploying machine learning and artificial intelligence in high-risk activities such as credit underwriting. Somewhat similar to environmental impact assessments, the basic idea is to require a detailed analysis of the potential impacts and risks of proposed AI/ML applications, develop detailed harm mitigation plans, consult with stakeholders before adoption, and engage in periodic monitoring.[103]

The EU Artificial Intelligence Act, which is expected to be finalized in 2024, requires a "fundamental rights risk assessment" for high-risk activities that will assess potential effects on individuals' fundamental rights, on marginalized and vulnerable populations, and on the environment. AI/ML users must also articulate detailed harm mitigation plans and governance processes and give relevant agencies and stakeholders notice and at least six weeks to provide input before finalizing and implementing the plans. If a detailed mitigation plan cannot be developed, the deployer is not permitted to implement the AI/ML application.[104]

In the U.S., the Algorithmic Accountability Act of 2022 would impose similar procedural requirements on "automated decision systems" that are used by large companies to make critical decisions including financial services. The legislation would also require benchmarking against previous decision-making processes. The bill has been introduced in both houses of Congress,[105] although it was not specifically referenced in a high-level framework for AI legislation released by Senate Majority Leader Chuck Schumer in June 2023. A series of bipartisan forums on risk management and other AI topics were held in fall 2023.[106]

also possible that emerging efforts to legislate safeguards on the use of artificial intelligence and machine learning models society-wide could expose both banks and nonbanks to new, broadly applicable procedural requirements such as algorithmic impact assessments (see Box 4.2.2.2.1). It is not yet clear how such broadly applicable requirements would relate to existing laws governing financial institution model risk management.

4.2.4 Validating Vendor-Provided Models and Model Management Tools

Stakeholder interviews suggest that addressing the governance challenges in working with vendor-provided models and tools could also have a critical effect on ML adoption for credit underwriting, particularly among smaller banks. When banks outsource certain functions to outside vendors, they are still accountable for compliance with substantive requirements that would apply if they were to conduct the vendor's activity directly.[107] These expectations require additional oversight activities to monitor the vendor's adherence to those standards and respond as needed to potential issues related to the vendor's performance.[108] Prudential regulators have also issued more specific guidance with regard to technology vendors and information security risks.[109] Given these expectations, financial institutions typically create risk management programs to conduct due diligence and ongoing monitoring of vendor relationships.

Although vendor risk management is typically treated as a separate risk management discipline, financial institutions that use underwriting models or supplementary tools provided by third parties may encounter additional challenges in model risk management. These issues take on heightened prominence in discussions about the adoption of ML underwriting models. Use of such models tends to be concentrated among the very largest banks and certain nonbank lenders with substantial technology resources. While many factors may account for the reluctance of lenders outside those groups to use machine learning underwriting models, their inability to support the expertise needed to develop and operate such models is likely a significant factor. Indeed, resource constraints have led many smaller lenders—including small banks and credit unions—to rely on vendors even for logistic regression underwriting models.[110]

In response to these resource limitations, a number of companies have begun offering model development and support services. These range from large established firms such as credit score providers to smaller firms created in response to growing interest in using AI and machine learning in financial services and other sectors. Although these firms take different technological approaches and have different products and strategies,[111] these businesses can help lenders overcome expertise gaps in developing models and support their responsible use by providing model management tools, consulting services, or some combination of the two. Some vendors also focus on direct provision of ML underwriting models to clients.

However, model validation under bank regulatory expectations can be more challenging for lenders that rely on vendor-provided machine learning underwriting models and/or model management tools. Information-sharing practices can vary widely, with some vendors offering their clients high-level descriptions of their approach to protect their intellectual property and competitive interests, as has been the accepted practice for national credit score providers,[112] and others offering extensive documentation that has been designed to help smaller lenders fulfill their model risk management obligations. Although regulatory guidance recognizes challenges relating to financial institutions' access to information about vendor-provided models,[113] ensuring proper transparency about model development activities—including how the models were produced, the data and technologies used in development, and the judgments made—is essential since standards for reviewing and validating an underwriting model do not change just because the model was developed by a vendor. Other challenging aspects can include:

- **Underlying Technologies:** Absent access to vendors' source code, lenders may lack insight into and/or familiarity with technologies used to develop vendor-provided models to evaluate whether the choices the vendor made in implementing the selected technologies are consistent with their needs and obligations.[114] This may go beyond the choice of basic model architecture or open-source model development packages to include questions such as what techniques were used to debias models.
- **Conceptual Soundness and Other Aspects of Risk Management:** Lenders using vendor-provided models need to be able to assess whether the models meet the lenders' standards for performance and credit risk tolerance, as well as to defend the underlying analytical framework used by the model to generate default predictions. This requires the generation of substantial information about the estimated relationships in the model, including potential use of *post hoc* tools as discussed above.
- **Consumer Compliance:** Lenders using vendor-provided models need to be able to assess whether the model as a whole and individual data fields used in the model comport with their consumer compliance obligations. For example, lenders need to be comfortable that appropriate adverse action notices can be generated and that vendors have taken appropriate steps to mitigate disparities across protected class groups throughout the model development process.
- **Data Use:** Vendors may develop models using data aggregated from a variety of sources and analyses. Users of vendor models need to be confident that such data is accurate, authorized for use, and does not violate privacy or other requirements. Where vendors use "big data" strategies to support model development, they themselves may rely on externally provided data sets to leverage socio-economic, behavioral, geographic or transaction-based information. In extreme cases, vendors may be subject to limitations on their ability to share details of data and analyses procured from other parties.[115]

In this context, the process of conducting vendor diligence often requires the type of expertise the lack of which may have led the financial institution to seek outside support in the first place.[116]

In the case of developing machine learning underwriting models, this gap in expertise can be prohibitive since the financial institution's analytics staff may have little experience with developing and operating machine learning models, using the range of common techniques and tools for monitoring and managing such models, or interpreting and explaining their outputs.

As substantive regulatory expectations clarify for ML models, some stakeholders have noted that increases in direct supervision of vendors by federal regulators could both ease burdens on small firms and help regulators determine whether particular practices have the potential to transmit risk across the market via use of common proprietary modeling techniques or tools.[117] While examinations are generally confidential, vendor supervision could help raise the floor on vendors' compliance and risk management practices and drive them to enhance their support of potential clients' compliance and risk management obligations. Further, agency examiners may face fewer constraints in reviewing competitively sensitive information than potential clients. A more ambitious option– which some vendors support—would be to create a certification regime to be administered by a standard setting organization so that vendors' technologies, governance, and controls receive periodic review without requiring each client to conduct full independent assessments.[118]

Federal banking regulators released updated joint third-party risk management guidance in June 2023 that acknowledged limitations that banks may face in obtaining desired due diligence from vendors and noting that alternative risk management strategies include obtaining information from alternative sources, implementing additional monitoring or controls, or considering the use of other vendors. The guidance also noted that banks may use the services of industry utilities or consortiums, consult with other organizations, or engage in other supplemental joint diligence efforts, consistent with antitrust law. However, it emphasized that the conclusions from such supplemental activities must be considered in light of the individual bank's situation and that understanding the nature of the supplemental activities itself should be treated as a risk management activity, since use of such external parties "does not abrogate the responsibility" of each bank to manage its third-party relationships. The guidance indicates that the agencies intend to produce additional resources to assist community banks.[119]

5. ADVERSE ACTION NOTICES

Adverse action notices are the most direct and concrete transparency requirement for underwriting models in federal consumer financial law, requiring lenders who reject credit applications, take other types of "adverse action," or charge higher prices based on credit report information to provide individualized explanations to the affected applicants.[120] The laws were primarily intended was to discourage discrimination, enable error correction, and educate borrowers about the basis for credit decisions that impacted them, but the disclosures also serve broader "sunshine" and procedural fairness goals. In recent years, financial services stakeholders have debated whether and how the notices could be changed to make them more practically actionable in helping applicants identify steps that they can take to increase their chances of obtaining credit or better terms in the future.

The challenge of accurately explaining complex models to individual applicants has been at the forefront of debates about whether and how machine learning can be fairly and responsibly used in the context of extending consumer credit. This section begins by summarizing the regulatory and operational context for adverse action requirements before addressing the following topics:

- Producing reliable descriptions of the behavior of machine learning models;
- Providing more information about how and why particular features affected the credit decision;
- Identifying plausible paths for individual consumers to increase their chances of credit approval; and
- Incorporating non-traditional underwriting data.

5.1 Regulatory and Operational Context

As part of consumer credit reforms in the 1970s, the Equal Credit Opportunity Act (ECOA) mandated that lenders provide disclosures that state their "principal reasons" for denying applications or taking other types of "adverse action" such as reducing the credit line of an existing borrower.[121] The Fair Credit Reporting Act also imposed disclosure requirements for adverse actions taken based on information in credit reports or obtained from other third parties,[122] and Congress later amended the law further to require lenders that charge higher prices based on credit report information to disclose the "key factors" that are negatively affecting the credit scores of the affected consumers.[123] This section summarizes (a) the core policy purposes that motivated the requirements; (b) current law and guidance; (c) compliance practices; and (d) key issues or risks in adverse action compliance when machine learning models estimate applicants' likelihood of default.

5.1.1 Policy Motivations

Contemporaneous accounts reflect three primary goals for the disclosure requirements:[124]

- **Discouraging discrimination:** The disclosures are intended to dissuade discrimination by requiring lenders to articulate the bases on which they are making credit decisions. As a secondary matter, the notices can facilitate review by disclosure recipients, public interest groups, regulators, and others to help identify where further investigation into potential fair lending violations is warranted.
- **Enabling error correction:** Describing the primary bases for an adverse decision and key factors that are negatively affecting an applicant's credit scores can help the recipient detect certain types of errors and seek corrective action. This is especially true where the disclosure highlights errors in a credit report—such as listing a prior student loan default by a consumer who never had such a loan—but may also occur with information that comes from other sources.[125]
- **Educating and empowering consumers in managing their finances:** Providing specific, point-in-time information about why a lender concluded that an applicant's default risk warranted declining an application or charging higher prices may promote self-improvement by helping recipients understand how past financial behavior or their current financial position is affecting their access to credit. More recent policy debates about adverse action notices have increasingly focused on a stronger form of this concept—specifically designing the disclosures to provide recipients with tailored, forward-looking information about how to adapt their financial behavior to increase their chances of accessing credit in the future.[126]

In practice these goals have never been well defined and can at times suggest different policy directions, which leaves room for debates about how the adverse action notice requirements can or should be implemented. For example, disclosures that are "true to the model" (as discussed in Box 2.3.2.1) and that provide more specific descriptions of the features that drove a particular credit rejection may be most helpful for purposes of discouraging discrimination and enabling error correction, while disclosures that are "true to the data" and that provide simpler feature descriptions may be more effective in helping consumers understand and take action to improve their chances of future approvals by a broad range of lenders over time. Such tensions are not a function of the type of model used to estimate an applicant's likelihood of default, but questions about the explainability of machine learning underwriting models have increased attention to broader debates about the purposes and effectiveness of current requirements, as discussed in Section 5.2.

5.1.2 Existing Law and Guidance

Unlike model risk management, adverse action requirements apply broadly to both bank and nonbank lenders, although regulatory supervision levels can vary in practice.[127] ECOA's disclosure provisions apply broadly to all adverse credit decisions, whereas FCRA requirements apply where lenders have based their decisions in whole or in part on information from a source other than the applicant or its own files.[128] In initial implementation, lenders voiced significant concerns about revealing competitively sensitive information and the burdens of generating individualized disclosures. Over time, the framework evolved as regulators issued guidance to address issues related to the use of automated underwriting systems and third-party credit scores. This section summarizes the requirements relevant to providing compliant disclosures to recipients of adverse action notices in the context of machine learning underwriting models.

ECOA and its implementing regulations require lenders to provide a specific and accurate description of their "principal reason(s)" for making an adverse decision.[129] The specificity requirement would not be satisfied, for example, by stating only that the applicant did not satisfy the lender's internal standards or have a sufficient credit score. With regard to accuracy, the content of the required disclosures must "relate to and accurately describe the factors actually considered or scored by" lenders.[130] However, existing law does not set out metrics or thresholds for evaluating compliance with these standards.

Regulations and guidance also require only a general description of *what* factors affected the decisions, rather than *how* or *why* the disclosed reason mattered in their overall analysis. Lenders are not required to describe the direction or magnitude of the factor in question, such as whether a particular factor was too high or too low or which specific metrics were used to measure it (e.g. 30- vs. 90-day delinquencies).[131] Although sample reason codes that are provided in an appendix to ECOA's regulations provide information about how and why some factors affect consumers negatively, other codes are less clear (See Box 5.1.2.1). Regulatory guidance also specifically emphasizes that users of credit scoring systems must report the actual reason for the credit decision "even if the relationship of that factor to predicting creditworthiness may not be clear to the applicant."[132]

For lenders that use credit scoring systems, supplementary guidance identifies three acceptable ways to identify the principal bases of the adverse credit decision with regard to an individual applicant:

- benchmarking the individual against applicants whose total score was at or slightly above the minimum passing score and disclosing the factors for which the individual was furthest below the average for that comparison group;
- benchmarking the individual against the average for all applicants and disclosing factors on which the individual performed least well as compared to that average; or
- "[a]ny other method that produces results substantially similar to either of these methods."[133]

Thus, the approach gives lenders latitude as to how they produce information about an adverse credit decision. A third common benchmark is to focus on the factors for which the applicant fell furthest below the maximum achievable score under the model.[134]

The guidance generally disfavors providing more than four or five reasons to explain the adverse action or pricing decision.[135] The adverse action notice must include any factor that required an automatic denial under the lender's policies, such as a prior bankruptcy or the fact that the applicant is a minor.[136]

The Consumer Financial Protection Bureau issued compliance circulars in 2022 and 2023 that focused on the use of machine learning underwriting models, *post hoc* explainability techniques, and non-traditional data sources in consumer credit.[137] The first circular emphasized that supervised entities must validate the accuracy of their chosen methodology for producing adverse action notices. It suggested that validation "may not be possible with less interpretable models," but did not discuss any particular methodologies or thresholds for determining the accuracy of individual *post hoc* explainability techniques.[138] The second circular warned against misuse of the model reason codes and emphasized that specificity in adverse action disclosures "is particularly important when creditors utilize complex algorithms" that rely upon data gathered outside of consumers' applications or credit files.[139]

BOX 5.1.2.1 SAMPLE REASON CODES

The sample reason codes that are provided in an appendix to ECOA's implementing regulation vary as to how much explanation they provide about how and why a factor contributed to an adverse action:

- Credit application incomplete
- Insufficient number of credit references provided
- Unacceptable type of credit references provided
- Unable to verify credit references
- Temporary or irregular employment
- Unable to verify employment
- Length of employment
- Income insufficient for amount of credit requested
- Excessive obligations in relation to income
- Unable to verify income
- Length of residence
- Temporary residence
- Unable to verify residence
- No credit file
- Limited credit experience
- Poor credit performance with us
- Delinquent past or present credit obligations with others
- Collection action or judgment
- Garnishment or attachment
- Foreclosure or repossession
- Bankruptcy
- Number of recent inquiries on credit bureau report
- Value or type of collateral not sufficient
- Other, specify: _____

5.1.3 Compliance Practices

Given the flexibility provided by the existing regulations in determining which features were "principal" for a particular applicant and how to describe those features in the disclosure, lenders must make a series of discretionary judgments in building their adverse action compliance processes and systems.

Lenders, credit scoring developers, and other companies that assist in disclosure production often differentiate between reasons that are based on simple categorical rules (often called "strategic reason codes") and those that reflect how a scoring or underwriting model predicted the applicant's individualized default risk ("model-based reason codes"). Users of traditional underwriting models may generate model-level reason codes in a variety of ways. Firms that use relatively simple underwriting approaches, such as applying a score cutoff such that anyone below FICO 650 will be denied, are likely to rely on adverse action reason codes provided by their score provider (which in turn often result from use of a scorecard system).[140] Firms that rely on proprietary regression models may base model-level reason codes on the model coefficients or use scorecard methodologies to clarify the specific contribution of features to the score generated for a particular applicant.

Lenders often aggregate the individual strategic and model-level reason codes that are identified in their internal processes into higher level categories that are described on the actual disclosures. These taxonomies can potentially facilitate the use of simpler and more intuitive language in the final disclosures, increase the consistency of terminology across different models and scores, reduce concerns about disclosure of competitively sensitive information, and facilitate automation processes. In some cases, the higher level descriptors may be based on the list of sample reasons that is provided in an appendix to ECOA's implementing regulations (see Box 5.1.2.1) or on versions that are developed by the compliance and legal teams of individual lenders, particularly in larger institutions.

At the same time, some lenders are also choosing to provide more detailed descriptions of important features than is required under existing regulation in an effort to increase their educational value. For instance, some lenders may view stating that an applicant had "too many credit card payments more than 30 days late" as more useful in understanding the adverse decision and helping consumers alter their financial behavior than simply listing "delinquent past or present credit obligations with others."

5.1.4 Key Risks and Compliance Issues for Adverse Action Notices

The growing use of machine learning underwriting models has focused attention on certain key issues and risks related to adverse action compliance. These issues and risks are not necessarily unique to ML underwriting models, but they may be further accentuated as model complexity increases. They include:

- **Operational Complexity and Model Transparency:** As described in Section 2, concerns about the transparency of machine learning models derive from their reliance on hundreds or thousands of features (sometimes including "latent features" generated by the ML algorithm), complex architectures, and more complex relationships within the data. Responsible design, development, and use of machine learning models requires an array of technical and operational decisions, including whether and how to use *post hoc* explainability techniques.[141] These approaches are relatively new when compared to the processes they replaced, and are evolving rapidly. Issues related to model transparency are context specific depending on the developer's choice of machine learning models and strategies for managing explainability concerns.
- **Accuracy:** Lenders have a general obligation to provide accurate adverse action disclosures, although how accuracy should be measured or what level of accuracy is required has not been clearly articulated. The transition to machine learning underwriting models has focused attention on whether available methods for describing model behavior are sufficiently reliable to provide content for adverse action notices that satisfy regulatory expectations.
- **Reduced Utility:** The shift to using models with more variables and more complex features may reduce the degree to which identifying four or five individual factors explains the adverse credit decision. Aggregating model-level information to group related features into broader categories can potentially overcome this concern but can bring its own challenges as discussed further below.
- **Risk of Information Compression and Obfuscation:** Aggregation of model-level codes into broader categories can help to make adverse action notices for ML models more meaningful, but the process can also be more challenging to execute effectively since expressing the key drivers of an adverse decision from a model with 15 input features requires less compression of information than one from a model with 1,500 input features. Use of models with substantially more complexity may also provide more opportunities for lenders to conceal information they would prefer not to disclose because it might attract unwanted scrutiny or present competitive or other risks if known to the public.[142]

5.2 Key Policy Issues for Adverse Action Notices and Machine Learning Models

As the adoption of machine learning models accelerates, stakeholders are debating both the use and sufficiency of *post hoc* explainability tools in the adverse action context and the potential value

of changing current market practice and regulatory frameworks to make disclosures more useful to consumers. This section considers key policy issues concerning producing reliable descriptions of the behavior of machine learning models, providing more information about how and why particular features affected the credit decision, identifying plausible paths for applicants to increase their chances of approval, and incorporating non-traditional underwriting data.

5.2.1 Producing Reliable Descriptions of the Behavior of Machine Learning Models

Whether they use interpretable or "black box" structures as discussed in Section 2.3, most lenders that have adopted ML underwriting models are likely to deploy *post hoc* explainability techniques to help generate the adverse action disclosures. The CFPB's 2022 circular highlighted the importance of validating the accuracy of *post hoc* tools for this purpose but did not discuss specific explainability techniques, validation methodologies, or thresholds for accuracy. Accordingly, firms are using their best judgment in determining which techniques and implementations are sufficiently reliable for compliance purposes.[143]

For the same reasons discussed in Section 4.2.2.2, our empirical framework and methodologies may be helpful to stakeholders in assessing the fidelity and consistency of particular *post hoc* techniques in explaining individual underwriting decisions. We applied two tests to assess the fidelity of various techniques when applied to a sample of consumers whose default risks were predicted to exceed the thresholds for approval. First, we compared the effect of perturbing the values of features that were identified as most important to the consumers' individual default predictions relative to perturbing sets of random or closely correlated features. Second, we used a "nearest neighbor" test to assess whether other consumers who were similarly situated based on the important features had similar default risk predictions. We also evaluated the extent to which different tools rank ordered the same four features as most important to the individual consumers' default predictions, with or without grouping similar or correlated features together.[144]

While the elements of our analyses can be improved and expanded over time, our substantive results were encouraging. We found that some but not all tools reliably identified features that were important to different models' risk predictions for individual consumers. The differences in performance on the fidelity tests tended to be largest when the tools were applied to complex models. On consistency, the highest fidelity tools tended to identify more of the same features as important than the tools that performed poorly on fidelity tests, although there were some differences even among the high fidelity tools, especially when they were applied to more complex models. Consistency improved substantially once we accounted for broader feature families and correlations.

At the same time, the results underscore the importance of making thoughtful choices in applying and deploying particular explainability tools to complex models, since there were variations in performance even within different implementations of a particular technique. For example, while many tools relying on SHAP feature importance measurements performed well, some did not. These results align with other research suggesting that some explainability tools and implementations tend to perform better than others, in part depending on the use case and in part on execution details.[145]

The findings also highlight the importance of focusing on broader relationships in the underlying data when applying explainability techniques and interpreting their outputs. Because features that a particular tool identifies as "important" serve as approximations for patterns in model behavior that are linked to both the identified features and other features that are correlated with them, other features may also be making important contributions to model outcomes. Thus, assuming a single feature within a correlated cluster is the sole driver of model behavior is likely incomplete. This speaks to the importance of lenders having a strong understanding of the data that are being

used to build, train, and deploy ML models for credit underwriting decisions, and of continuing to refine approaches to addressing differences between explanations for the same model produced by different explainability tools.[146]

Further research would be helpful to inform market practice as academics and private sector stakeholders continue to iterate on existing options and develop new approaches, many of which are focused on attempting to account for correlations among features in more nuanced ways.[147] It is also important to note that some degree of variation in fidelity and consistency even among high performing tools is to be expected in light of the nature of machine learning models and the range of decisions that must be made in the course of deploying particular explainability techniques. Reasonable differences in choices about which implementation of SHAP to apply to a particular model type, the size of comparison samples, which benchmark to use, and other technical details may introduce a certain amount of inconsistency relative to results from companies that made different decisions.[148] This is particularly true where ML models involve substantially larger numbers of variables, since the impact of any one feature is likely to be smaller than in the context of a traditional regression model. The existing regulatory guidance on adverse action notices already contemplates similar reasonable variations by recognizing the legitimacy of using different comparison groups as a benchmark to determine which features played the biggest role in shaping individual applicants' risk predictions.

Our findings also suggest that the common practice of grouping related features together to produce higher level adverse action reason codes could be particularly important in the machine learning context, both as a means of addressing some of the technical challenges created by the presence of large numbers of correlated features and as a strategy for explaining more of machine learning models' overall operations to consumers. Using the analogy of machine learning models to a box of 128 Crayons helps to illustrate the potential impact. Assume that three lenders make reasonable but slightly different implementation choices in using an explainability technique to identify the primary bases of a credit decision for the same applicant from the same machine learning model, and receive results that produce slightly different rank orders listing three closely linked or correlated features (crimson, cherry, and ruby) as most important to an individual consumer's risk prediction. Assume a fourth lender uses an aggregation process to group together related features in generating the explanation, yielding an explanation of red and leaving more room to highlight the role of additional factors such as blue and green. While all of the methodologies may be reasonably accurate, the aggregation approach can potentially produce more consistent and meaningful disclosures for consumers that convey more information about the model's operation overall.

This analogy also helps to highlight that technical accuracy is not the only consideration in making adverse action disclosures more understandable and actionable for applicants, particularly for purposes of identifying errors in credit report data and making changes to their finances to improve their chances of accessing lower cost credit over time. The next section discusses other debates concerning the optimal level of precision that have particular implications for machine learning underwriting models.

5.2.2 Providing More Information about How and Why Particular Features Affected the Credit Decision

Beyond answering methodological questions about the reliability of information generated by the current generation of *post hoc* explainability techniques, the transition to machine learning models is focusing renewed attention on questions about the importance and feasibility of disclosing granular information about individual models' operation. Many of these issues are rooted in the existing regulatory guidance, but the nature of ML models and current explainability techniques raise additional questions about what level of specificity is both required for legal compliance and optimal from a policy perspective.

As described above, the existing adverse action regime ensures that rejected applicants receive basic disclosures about what input features played a principal role in their default prediction, but does not require an explanation of what specific metrics were used (e.g., 30-day versus 90-day delinquencies), how various features interacted with each other within the model, or why a particular feature is predictive of higher default risk. Some stakeholders argue that such details could be useful to consumers both in identifying and correcting errors in historical credit data that was used to evaluate their applications and in better understanding what behavioral and financial changes might help to increase their chances of accessing affordable credit over time.[149] The CFPB's 2023 circular warned against using "reasons that are overly broad, vague, or otherwise fail to inform the applicant of the specific and principal reason(s) for an adverse action," particularly with regard to data sources that consumers are unaware of being used for underwriting purposes, but did not provide a detailed discussion of technical and policy questions beyond potential obfuscation concerns.[150]

The potential value of more detailed explanations could be heightened in the machine learning context to explain the operation of more complicated models, yet also harder to achieve. For example, because machine learning models may consider non-monotonic and non-linear relationships, they may treat certain features such as credit utilization as increasing risk of default in some situations and reducing it in others.[151] An adverse action notice that simply indicates that "use of available credit over time" was a principal factor in rejecting a particular consumer would not help the applicant understand whether reducing or increasing utilization would strengthen their application. However, lenders would potentially have to develop a more complicated and tailored menu of reason codes to match different consumers' situations.

Another example is the way that machine learning models' predictions may be driven by latent features created by the learning algorithm or by interactions of features within the model, some of which may be non-intuitive. Increasing the complexity of the model—the range of relationships used to predict someone's default risk—may make what happens within the model relatively more important to serving certain goals of the adverse action regime, such as conveying accurately why the lender made a particular decision and enabling consumers to adjust future behavior accordingly. However, while feature importance measures such as SHAP are designed to measure the cumulative importance of input features, they may still have difficulty pinpointing the precise feature interaction within the model that was critical for an individual consumer's prediction.[152]

Take the example of a machine learning algorithm that determined that late payments on a mortgage loan are associated with much greater default risk where mortgage loan balances are higher (e.g., $200,000 rather than $50,000). Listing mortgage loan delinquencies and balances as separate principal factors may not fully convey that the *combination* of the two drove a particular loan rejection, such that addressing both components may be necessary to substantially reduce predicted risk levels. If fully conveying the feature interaction is considered critical for compliance, this raises questions about the precision of particular techniques in being able to identify and measure such interactions and of the need for more complex and varied reason codes.[153]

These considerations as well as a broader interest in promoting consumer financial well- being are fueling interest in increasing the specificity of adverse action disclosures, either as a matter of market practice or regulatory requirement. However, defining and achieving the optimal balance of specificity for *both* traditional and machine learning models is challenging due to both policy and practical considerations. For example, as discussed in the previous section, there are valid policy and technical arguments for aggregating related features into somewhat higher level categories to provide more consistent messaging and convey more of the model's overall operations, particularly in the context of machine learning models. As noted in Section 5.1.1, the level of specificity that will best help applicants identify whether there may be errors in the underlying data used by the lender

may also be different than the level of specificity that best educates applicants about ways to improve odds of acceptance going forward.[154]

Policy judgments about the optimal level of specificity to inform applicants' future actions are particularly complicated. Narrower, more specific information may help recipients better understand how to improve their chances of approval in some situations but at times may also increase the risk that disclosure recipients miscalibrate their future activities. In the example with the model that rejected particular applicants because of delinquencies on a mortgage loan with a high balance, consumers who read the disclosure to imply that they should prioritize paying down their mortgage balance even at the expense of incurring delinquencies on other loan types might still find that they struggle to access credit going forward. Moreover, by the time the consumer seeks credit again, the lender's model may have changed or the consumer may apply to a different company. Such factors may counsel toward crafting disclosures that provide broader, more generalized information that is "true to the data" rather than "true to the model."[155] The next section discusses more radical suggestions to add or shift the adverse action regime specifically to focus on producing actionable, forward-looking advice rather than concentrating on explaining the lender's most recent underwriting decision.

5.2.3 Identifying Plausible Paths for Applicants to Increase Their Chances of Approval

Discussions of how to comply with adverse action requirements in the machine learning context have intersected and overlapped with broader debates about whether to shift the focus of the disclosures from retrospective documentation of the lender's recent underwriting decision to providing forward-looking, actionable information about how to improve the applicants' odds of accessing affordable credit in the future. Such ideas gathered significant interest in a tech sprint that the Consumer Financial Protection Bureau organized in 2021.[156] Yet while intuitively appealing, supplementing the existing disclosures or replacing them with forward-looking advisory material raises a number of complex issues for both traditional and ML models.

At a conceptual level, providing advice about how to make changes that would result in an approval or lower pricing requires answering a different question than the one posed by existing requirements. The current system asks what factors played a principal role in the model's prediction that the applicant presented high levels of default risk. An advisory regime would ask what factors the applicant can most reasonably improve to reduce their predicted default risk at some point in the future. The answers to these questions are not necessarily the same, as illustrated by the example of a lender that weighs prior bankruptcies heavily in its credit decisioning models: The fact of a prior bankruptcy may carry the most weight in the model overall, but since bankruptcies remain on credit reports for seven to 10 years by law, an advisory regime would likely focus on a different set of factors. The analysis required is thus far more complex than under current regulatory requirements because it must not only account for the individual and collective weights of the features but also their susceptibility to future change over particular time horizons.

Moreover, depending on the model and the time frame selected, there may be no set of factors that would be sufficient to offset the principal factors (e.g., the recent bankruptcy) or the consumer might be required to make a large number of marginal changes across multiple factors to produce a sufficient offset, particularly in connection with machine learning models given their size and structure. (See Box 5.2.3.1). Thus, the number of consumers for which plausible paths are available and the number of plausible paths identified can vary significantly depending on the underlying model, the consumer's circumstances, and the parameters set for the analysis. Shorter time frames may be more consistent with consumers' desire to obtain credit and more likely to spur them to action where plausible paths are identified, yet they are likely to reduce the number of plausible paths

BOX 5.2.3.1 FINREGLAB'S EMPIRICAL TESTING

To test the use of explainability techniques in identifying actionable paths to acceptance in our empirical study, we asked the participating companies to recommend a small set of actions (preferably no more than four) that would cause each of the consumers in the sample used for the adverse action analysis to reduce the likelihood of default sufficiently to meet approval thresholds within one year.[157]

The recommended number of changes was higher than expected, averaging about eight changes per consumer even for the relatively simple models. For more complex models involving hundreds of features, producing a large drop in the predicted probability of default required changing large numbers of features and/or making changes that were large in magnitude relative to the variations observed in the data.

The number of changes required might have been smaller if the inquiry had been restricted only to credit applicants who were "near misses." However, the results underscore the importance in the machine learning context of accounting for correlations among features when identifying and describing feasible strategies for consumers to increase their chance of future loan acceptance rather than focusing on a few features in isolation.

because the number of features that can be changed quickly is lower. Using different parameters for different loan products and sizes may be logical, but could potentially increase the complexity of the analyses and disclosures.

Shifting the focus of the disclosures also raises a number of policy issues regarding the presentation of the advisory information, starting with whether to replace or supplement the current disclosures. Doubling the amount of information that is presented on adverse action notices may risk confusion, overload, and disengagement,[158] yet narrowing the focus solely to forward-looking information may undercut other policy goals of the existing regime (such as error identification and correction) and raise questions about what information to provide to consumers where paths to acceptance within the specified time period are highly complex or unlikely.

Consumer user testing could be instrumental to evaluating different disclosure options, particularly with regard to static disclosures (whether delivered on paper or electronically). In the CFPB tech sprint, some participants explored the potential for dynamic online disclosures that would allow consumers the opportunity to understand how different combinations of changes over different time periods could potentially affect their risk assessments. Such information may be accessed by fewer consumers but provide substantially more useful information.

A final consideration is how to account for the fact that lenders' credit criteria may change over time due to a broad range of factors (including refinements to underwriting models, changes in economic conditions, and shifts in business and product strategies) and that consumers may ultimately apply to a different company that uses a different model. While advisory disclosures could potentially include warnings to this effect, lenders would likely seek assurances against liability in the event that applicants are later rejected a second time.[159] More broadly, as discussed in Box 5.2.3.1 and Section 5.2.2, such factors may counsel toward crafting any forward-looking disclosures to provide broader, more generalized information that is "true to the data" rather than "true to the model."

5.2.4 Incorporating Non-Traditional Data

Similar to Section 2.2.3's discussion of the implications for fairness and inclusion of building machine learning underwriting models to use non-traditional data sources, it is important to note that debates about applying adverse action disclosure requirements to ML underwriting models are heightened where the models incorporate new data sources.

As discussed in that section, potential non-traditional sources of underwriting data can range from financial information that is not captured in traditional credit bureau records (such as cash-flow data

from a person's bank accounts or rent and utility payment history) to behavioral data from a wide variety of activities (such as an applicant's interaction with the lender and digital footprint). When deciding whether to use such information, lenders assess a broad range of factors, including predictiveness, fairness and inclusion effects, and various other business and risk considerations. The fact that lenders must also disclose information about the data sources in adverse action notices can also play a significant role, not just with regard to technical compliance but also the broader reputational, regulatory, and competitive considerations involved in disclosing the use of particular data elements.

For example, while the sample reason codes that are provided in ECOA's implementing regulations may be relatively easy to adapt to some new sources of financial information,[160] lenders who decide to use behavioral information may need to craft new disclosure language. In addition to considering the costs and risks of such processes, lenders may shy away from using particular features if they believe there is a significant risk that disclosure recipients would feel that the data elements were unfair, irrelevant to assessments of creditworthiness, or violative of privacy norms, or that the elements would attract scrutiny regarding discriminatory practices or other regulatory concerns.[161] For example, some lenders have reported that they have chosen not to consider data from cookies or the channel by which an application was received in their underwriting models due to these broader considerations, even where analysis showed such information was predictive of default risk.

In light of these considerations, the debates about the specificity of adverse action disclosures discussed above can take on heightened policy significance in the context of data sources that are unfamiliar, unintuitive, or raise broader reputational concerns. Even where applicants themselves are supplying or authorizing access to particular data sources, they may not understand what aspects are being used by the lender, what errors might have a material effect on their assessment, or what changes in behavior might increase their chances of obtaining credit in the future. General educational materials may also be less available than in the context of traditional scoring and underwriting systems that rely on traditional credit bureau data. Thus, lenders' decisions about how broadly or narrowly to describe particular features can have a significant effect on consumer understanding.

For example, lenders that use educational attainment or professional information as a means of forecasting future income have historically faced a range of potential disclosure options, ranging from "educational attainment," to "insufficient income," to a hybrid indicating than the first was used to forecast the second. They may have considered a broad range of factors in deciding where to land on this spectrum, including the likelihood of triggering debates over the fairness of the factors,[162] the potential utility to consumers of being more or less specific, and other considerations. The CFPB's 2023 Circular recently indicated that simply stating "insufficient projected income" or "income insufficient for amount of credit requested" would likely not satisfy the lender's obligation to provide specific reasons for the adverse action, but did not provide an illustration of what would be adequate.[163] Some stakeholders believe that updating the list of sample reason codes provided in ECOA's implementing regulations or providing additional regulatory guidance about how to describe input features would be helpful to both lenders and borrowers in navigating these issues.

* * *

As reflected in this section, the debates about how to better effectuate the policy goals behind adverse action disclosures pre-date the adoption of machine learning underwriting models. However, the nature of machine learning models and current *post hoc* explainability tools do present certain technical challenges and increase the stakes of certain policy debates. While research on the reliability of explainability tools is encouraging, determining if, how, and when adverse action notices can be made more generally useful, and even actionable, for consumers requires substantially broader analyses and mediation between multiple policy goals.

6. FAIR LENDING

Although they do not fully encompass all of the notions of fairness discussed in Section 2.2, anti-discrimination requirements under the Equal Credit Opportunity Act and other federal laws provide important conceptual and procedural frameworks for assessing the fairness of inputs, processes, and outputs in credit underwriting. These requirements have shaped machine learning adoption for credit underwriting from the very beginning, requiring stakeholders to grapple with concerns about fairness impacts at a far earlier stage than in many other sectors.[164] As ML adoption has accelerated and the broader data science community has focused increased attention on fairness issues, financial services stakeholders are debating the extent to which traditional concepts and practices should shift to account for ML models.

Two sets of data science techniques are potentially relevant in this context. First, some lenders are using *post hoc* explainability tools to analyze which features play a particularly important role where ML models produce disparities in predicted default risks for different demographic groups. Historically, such analyses have been an important precursor to adjusting models to address fair lending concerns. Second, debiasing techniques and other general development tools are helping developers act more efficiently to produce a range of ML models to choose from in balancing fairness, performance, and other concerns.[165]

This section begins by summarizing policy and operational context relevant to fair lending compliance for machine learning models,[166] including an overview of debiasing techniques. The second part provides more detailed analyses of the following policy topics:

- Concerns that machine learning models may rely upon relationships that are proxies for race or other protected characteristics;
- How to measure fairness for disparate impact purposes;
- Whether particular debiasing techniques are permissible in light of how they use data about protected characteristics; and
- Clarifying expectations regarding searches for less discriminatory alternative models.

6.1 Regulatory and Operational Context

Congress adopted a series of laws in the 1960s and 1970s to prohibit discrimination and address various other types of fairness concerns in lending. The broadest of these is the Equal Credit Opportunity Act, which prohibits discrimination against "any applicant, with respect to any aspect" of

a consumer or commercial credit transaction "on the basis of" race, color, national origin, sex, and various other protected characteristics.[167] The Fair Housing Act (FHA) prohibits discrimination in residential mortgage lending on many of the same bases, as well as familial status and disability.[168]

These anti-discrimination laws have evolved to focus on two primary doctrines, disparate treatment and disparate impact.[169] Disparate treatment generally prohibits consideration of race, gender, or other protected characteristics in underwriting and scoring models. Disparate impact, in contrast, prohibits the use of facially neutral practices that have a disproportionate adverse impact on protected classes, unless the practices serve legitimate business needs that cannot be reasonably met through less impactful means.

This section summarizes (a) policy motivations; (b) current law and guidance; (c) traditional compliance practices and evolving debiasing techniques; and (d) key issues or risks in fair lending compliance when machine learning models estimate applicants' likelihood of default.

6.1.1 Policy Motivations

As discussed in Section 2.2, debates about fairness in financial services and other sectors often invoke a broad range of conceptions of fairness and inclusion, including such notions as equal treatment, equity in outcomes, and consistency in prediction accuracy, as well as broader concerns about increasing access by historically excluded groups. All of these concepts are relevant to federal fair lending law, with the first two playing a particularly important role in shaping legal analyses and compliance frameworks:

- **Equal treatment:** This conception of fairness requires that individuals be subject to the same criteria regardless of demographic characteristics, and in a more affirmative formulation that similarly situated individuals receive similar treatment.[170] Through the disparate treatment doctrine, this principle is invoked in the credit context to prohibit different treatment because of applicants' race/ethnicity, gender, or other protected characteristics.
- **Equity:** This conception of fairness focuses on the extent to which individuals in different demographic groups receive equal outcomes even if they are not similarly situated in certain other respects.[171] Its most direct use in fair lending compliance occurs at the first stage of disparate impact analysis, which evaluates whether facially neutral practices are creating disparities in approval rates and pricing among demographic groups without accounting for differences in financial situations.
- **Consistency of predictive accuracy:** Prediction errors can create an additional source of fairness concerns, particularly where they are more likely to occur among particular demographic groups.[172] In the credit context, an underwriting model that generates disproportionate numbers of false positive and false negative predictions for particular subgroups will result in more denials of creditworthy group members and approvals of group members who will in fact struggle to repay. In light of this, lenders may consider the accuracy of models in predicting default risk across protected classes as part of their disparate impact analyses.

Different notions of fairness can complement and reinforce each other in some circumstances and create substantial tensions in others. For example, as discussed in Section 2.2, improving prediction accuracy for consumers who are hard to evaluate using traditional methods and data could also advance equal treatment (consistency of outcomes based on actual default risk), equity (consistency of outcomes among different demographic groups), and inclusion (participation by historically excluded groups). However, data science research has shown that as a matter of mathematics, it is often impossible to satisfy multiple quantitative definitions of fairness at the same time.[173]

BOX 6.1.1.1 SOURCES OF BIAS AND DEBIASING TECHNIQUES

Particularly in the fair lending context, stakeholders often use terms such as "fairness," "unfairness," and "bias" interchangeably to discuss demographic disparities in model inputs and outputs that raise policy and legal concerns. However, statisticians and data scientists often use the term "bias" more broadly to include a wide range of variances between a model's predictions and actual outcomes. Using this more technical definition of "bias," the consistency of predictive accuracy is a bias issue, while notions of equity are best characterized as fairness issues.

In the context of using machine learning models to assess credit risk, biases can originate from both the data used to train the model as well as choices made during model development. For example,

- Bias can be introduced where training data are inaccurate, reflect past discriminatory practices, omit key variables, or lack representation for certain groups.
- Demographic disparities can also occur where features used by the model are correlated with protected or sensitive features.
- Choices in whether and how to optimize a model for larger populations or for different goals may affect its predictiveness with regard to particular populations.

"Model debiasing" refers to a range of methods to increase the accuracy and fairness of a model's predictions, for instance by transforming the input data, building a debiasing function into model training, or transforming a model's output. The machine learning debiasing methods discussed below involve building a debiasing function into model training. For additional detail about sources of bias and mitigation approaches, see our Market & Data Science report.[174]

In credit underwriting, there are particular concerns that "fairness through unawareness" under disparate treatment law is replicating or even exacerbating decades of past discrimination because underwriting relies so heavily on data sources that reflect deep financial disparities produced by decades of discrimination across a wide variety of economic sectors, lack of geographic access to banks, and targeting by high-cost lenders.[175] In the face of such "traumatized data," stakeholders are debating the efficacy of existing doctrines and compliance approaches for achieving broader fair lending goals, including the potential advantages of using protected class data to support more aggressive debiasing activities or to include it directly in credit underwriting models to account for underlying differences in the size and financial situations of different populations.[176] While these debates are not unique to machine learning models, those models' ability to detect more subtle relationships in underlying data and the new techniques being developed to manage such models are raising both hopes and fears about our ability to achieve broader fairness goals.

6.1.2 Existing Law and Guidance

Unlike model risk management, fair lending requirements apply broadly to both bank and nonbank lenders, although regulatory supervision levels can vary in practice.[177] The regulatory frameworks are similar to those developed under federal laws governing employment. This section summarizes the requirements relevant to the context of machine learning underwriting models.

6.1.2.1 Disparate Treatment

The disparate treatment doctrine focuses on whether creditors have treated applicants differently based on protected characteristics, and with limited exceptions prohibits consideration of race, gender, or other protected characteristics in underwriting and scoring models.[178] It is reinforced by other federal law that generally prohibits lenders from collecting data about protected characteristics for fear that it will be used for credit decisioning.[179]

While disparate treatment focuses generally on intentional discrimination, there is no requirement to show animus or a conscious intent to discriminate against a protected class, only that the

act of differential treatment was intentional. For example, except for narrowly prescribed special purpose credit programs,[180] disparate treatment would bar a lender from accepting loan applications only from women without regard to whether the lender has an animus towards men. The doctrine has also been applied to situations in which lenders incorporate factors that are closely correlated to a protected characteristic as a pretext for intentional discrimination. Examples of features that might raise statistical proxy risks are census block identifiers and certain magazine subscriptions or occupational classifications, among other features.

6.1.2.2 Disparate Impact

Disparate impact prohibits lenders from using facially neutral practices that have a disproportionately negative effect on protected groups, unless those practices meet a legitimate business need that cannot reasonably be achieved as well through less discriminatory alternatives.[181] The disparate impact doctrine is sometimes described as an "effects test" because it focuses on the effects of a process rather than its intent. Examples of underwriting features that might trigger disparate impact scrutiny include income and length of credit history, since they tend to be correlated with demographic characteristics such as race/ethnicity and age, respectively, and thus might create disparities in default predictions among different demographic groups that lead to disparities in loan approvals or pricing.

However, analyzing outcomes is only the first stage of a disparate impact analysis, which then shifts to assessing whether the practice furthers a legitimate business need (such as predicting default risk) and whether there are alternative criteria or processes that would reasonably achieve the same goal while producing fewer disparities.[182] In a courtroom setting, the burden is on the challenger to show the existence of less discriminatory alternatives, although lenders may perform all three analyses as part of their compliance programs.[183]

Federal laws and regulatory guidance do not specify thresholds for what level of disparity in outcomes requires an inquiry into legitimate business needs or searches for less discriminatory alternatives, although in the employment context the Equal Employment Opportunity Commission has used 80% as a "rule of thumb" at times.[184] Similarly, there are no defined qualitative or quantitative standards for determining what models constitute "less discriminatory alternatives" (LDAs). At 2023 conferences, CFPB officials have described "rigorous searches" for LDAs as "a critical component of fair lending compliance management" and expressed concern that lenders may tend to shortchange this aspect of compliance. However, the agency has not issued formal guidance on LDA topics.[185]

6.1.3 Operational Context

Over time, lenders have developed a range of operational processes both to manage potential biases in general and to meet fair lending compliance expectations specifically. Historical methods have tended to rely heavily on analyzing which input features are most closely correlated with protected class as a first step to determining whether and what types of back-end adjustments to the model may be warranted. However, new methods that have evolved in the machine learning context are focusing on earlier and more rapid iteration strategies to identify alternative models with lower levels of disparities. This section summarizes both traditional compliance approaches and the new debiasing techniques.

6.1.3.1 Access to Protected Class Information

Although protected class information is critical to assessing and managing fair lending risks, federal laws generally prohibit collection of information about protected characteristics except for residential

mortgages and small business loans.[186] For categories of loans for which demographic data are not available, compliance staff and agency examiners typically rely on a methodology called Bayesian Improved Surname Geocoding ("BISG") to impute race and ethnicity probabilities and similar techniques for gender probabilities based primarily on names, addresses, and birth years where available.[187] Some stakeholders have noted that the methodology is less reliable for some subpopulations than others and may become less effective over time as residential, marriage, and name patterns shift.[188]

In light of the restrictions on data collection and broader prohibitions on disparate treatment, firms typically designate separate compliance teams to be responsible for fair lending oversight and restrict access to information about protected characteristics to those teams as a matter of policy. This approach provides structural safeguards against misuse of this sensitive information, yet complete separation is not always practicable (particularly for smaller firms) and has potential implications for the efficiency and efficacy of modeling and debiasing processes.[189]

6.1.3.2 Manual Feature Reviews

Firms typically conduct manual feature reviews during the initial model development process to ensure that protected class characteristics have not been included and to exclude features that do not have a clear nexus with creditworthiness or that based on the firm's general knowledge and experience are likely to drive unjustifiable disparities in the model's decisions.[190] By excluding certain variables up front, model developers can reduce the number of iterations required to develop a model that minimizes disparate treatment and disparate impact risk. For this reason, many lenders report that features in their logistic regression underwriting models have been relatively stable over time, as developers draw on knowledge gained in past rounds of testing and validation when selecting features for updated models.

6.1.3.3 Statistical Testing

Statistical testing is an important component of fair lending risk management for both disparate treatment and disparate impact. Regardless of the type of underwriting model being used, fair lending testing often begins once the model development team has submitted the model for review and validation, and is conducted by a separate compliance team with access to actual or imputed data about protected classes.

Disparate Treatment

Statistical testing is used in the disparate treatment context to evaluate whether input features are functioning as impermissible proxies, although federal regulators have not specified a specific analytical test or threshold for what level of correlation is considered impermissible.

The most commonly used analysis calculates the level of correlation for each input feature with both protected class on one side and model performance on the other.[191] This is relatively easy to calculate for any type of underwriting model, but some stakeholders suggest that is too simplistic, in that it would be hard to argue that a variable with modest correlations to both protected class and model performance is a proxy simply because the protected class correlation is slightly higher.

Some other firms deploy alternative tests particularly in the machine learning context that evaluate the impact of variables that are highly correlated with protected characteristics within individual protected class groups, either by training separate models on control and test populations or by looking at performance contributions within specific protected classes. Proxy risk might be heightened if a variable that is highly correlated with a protected class also does not contribute

to performance within that protected class, despite contributions to performance in underwriting other groups or across the broader model as a whole.[192]

Disparate Impact

Disparate impact compliance also relies heavily on statistical analyses to determine the extent to which particular input features are producing disproportionate adverse impacts that require potential mitigation measures.[193] Much as in the disparate treatment context, regulatory guidance does not specify the methodology to be used or thresholds for what level of disparities trigger follow up activity.

In the absence of a clear directive, consumer lenders often use the adverse impact ratio ("AIR") to test for the presence of statistical disparities in decisions to approve or deny loan applications and standardized mean differences ("SMD") to test for pricing disparities.[194] For instance, AIR is often used to assess the extent to which differences in a model's default predictions across different protected class groups (e.g., men and women or Black and White applicants), will cause disparities in approval rates at given risk thresholds. Specifically, AIR is calculated as the ratio of the approval rate for a protected class group to the approval rate for a control group (often White applicants or White male applicants). For example, if 30% of Black applicants are approved for a loan and 40% of non-Hispanic White applicants are approved for a loan, then AIR is equal to 30% / 40%, or 0.75.

Where a lender selects AIR to evaluate potential new underwriting models, the next step is to evaluate whether each individual consumer in the testing population would have been approved or denied under each model and approval threshold being considered,[195] and then use actual or imputed protected class characteristics to calculate AIRs for each protected class for each alternative under consideration. Lenders may also generate accuracy statistics for the model as a whole and for each demographic group to assess whether prediction errors are disproportionately concentrated among particular populations.

Some lenders also apply a second cutoff to determine whether adverse impacts are sufficiently large to warrant a search for less discriminatory alternatives or other follow ups. These practical significance standards help lenders focus resources on disparities that pose the greatest regulatory and reputational risk.[196] However, as discussed in Section 6.1.2, case law and regulatory guidance do not provide precise mathematical thresholds for determining the level of problematic disparities, as the 80% threshold that is sometimes used in the employment context has not been formally recognized in financial services.[197] As a result, firms establish internal thresholds based on a variety of factors related to their business and operations, such as customer base, business strategy, prior experience with comparable models, and prior regulatory interactions.

6.1.3.4 Techniques for Debiasing Underwriting Models

Where statistical tests raise substantial disparate impact concerns, firms' mitigation processes typically focus on (1) identifying potential changes to the model's specifications or alternative models that could potentially reduce correlations with protected characteristics and disparities in outcomes across protected class groups; and (2) testing whether those alternatives do in fact reduce fair lending risk while achieving substantially the same business outcome as the original model.[198] The first step generally does not require access to protected class characteristics, although some firms may direct their compliance staffs to use protected class information to conduct additional manual feature reviews to remove certain features or to identify key features on which to focus and others may opt to expose algorithms used for debiasing models directly to protected class information.[199] In contrast, the second step always requires protected class information in order to test the revised model(s). Since firms typically limit access to protected class information to fair lending teams, searches for less discriminatory alternative models often occur relatively late in the process of model development.[200]

The adoption of machine learning models has expanded available methodologies for conducting searches for less discriminatory alternatives, although many of these new methods raise important policy questions. This section provides an overview of both traditional and machine learning debiasing methods as a background to the subsequent policy analysis in Section 6.2.

6.1.3.4.1 *Traditional Debiasing Methods*

In relatively simple logistic regression models, lenders rely on features' coefficients to indicate the importance of specific data inputs on a particular credit decision and in driving aggregate disparities across protected class groups. To determine whether a particular input feature was driving disparities in the model's predictions, developers could simply omit the feature and observe how predictions and disparities change when the feature was omitted. This method is colloquially known as "drop one out." The feature's contribution to a difference in predicted default between applicants in various groups—for example, the difference between Hispanic and non-Hispanic White applicants—could be measured by evaluating both populations with and without the feature included, applying an approve/deny threshold, and computing the difference in approval rates and accuracy levels between populations.

Where a particular model is associated with a disparity of sufficient size to trigger further investigation under the lender's policies, lenders may evaluate the tradeoffs of omitting individual features, adjusting them in various ways to reduce disparities in the model, and substituting other features or data. In traditional models with relatively few features, omitting features entirely may help to reduce disparities, but often comes with a significant cost in the model's predictive performance. Lenders' willingness to adopt alternative models vary in such circumstances based on a variety of factors, including general risk tolerance, the purposes for which the model will be used, the confidence in the model's accuracy metrics, and concerns about the distribution of increased defaults.

To mitigate the loss of performance that comes with "dropping one out," lenders can search for ways to adjust the weights of individual features or transform individual features to reduce the disparities in the model's predictions without incurring the full performance cost of omitting the feature altogether. For example, considering income or length of credit history up to a certain cap or using debt-to-income ratios instead of income by itself may help to preserve much of the predictive value while reducing disparities.

Another option for mitigating performance deterioration from "drop one out" strategies is to add data to enrich the context in which the model is making predictions. For instance, accounting for evidence of the fact that some applicants may not have ready access to banks or other amenities might help to clarify predictive signals and reduce disparities created by inputs concerning their prior borrowing or financial history. However, geographic information at times can create risks of proxy discrimination, as discussed in Section 6.1.2. Accordingly, each decision about whether to add data and if so what type of data to use needs to balance potential benefits and risks.

6.1.3.4.2 *Techniques for Debiasing ML Models*

Adoption of machine learning underwriting models has opened the door to using a variety of more sophisticated debiasing methods that rely on algorithmic optimization rather than manual feature or model modifications to reduce disparities in credit decisions. The two most common examples of these methods are joint optimization and adversarial debiasing.[201] Unlike the traditional fair lending compliance strategies described above, these machine learning methods typically work on a model holistically—producing new model specifications or "drafts" of the model, rather than trying to edit individual features in the model and assess the marginal effects of each change.

These machine learning debiasing approaches harness the ability of the learning algorithms to identify highly accurate models and to generate iterative sets of model specifications relatively quickly and efficiently.

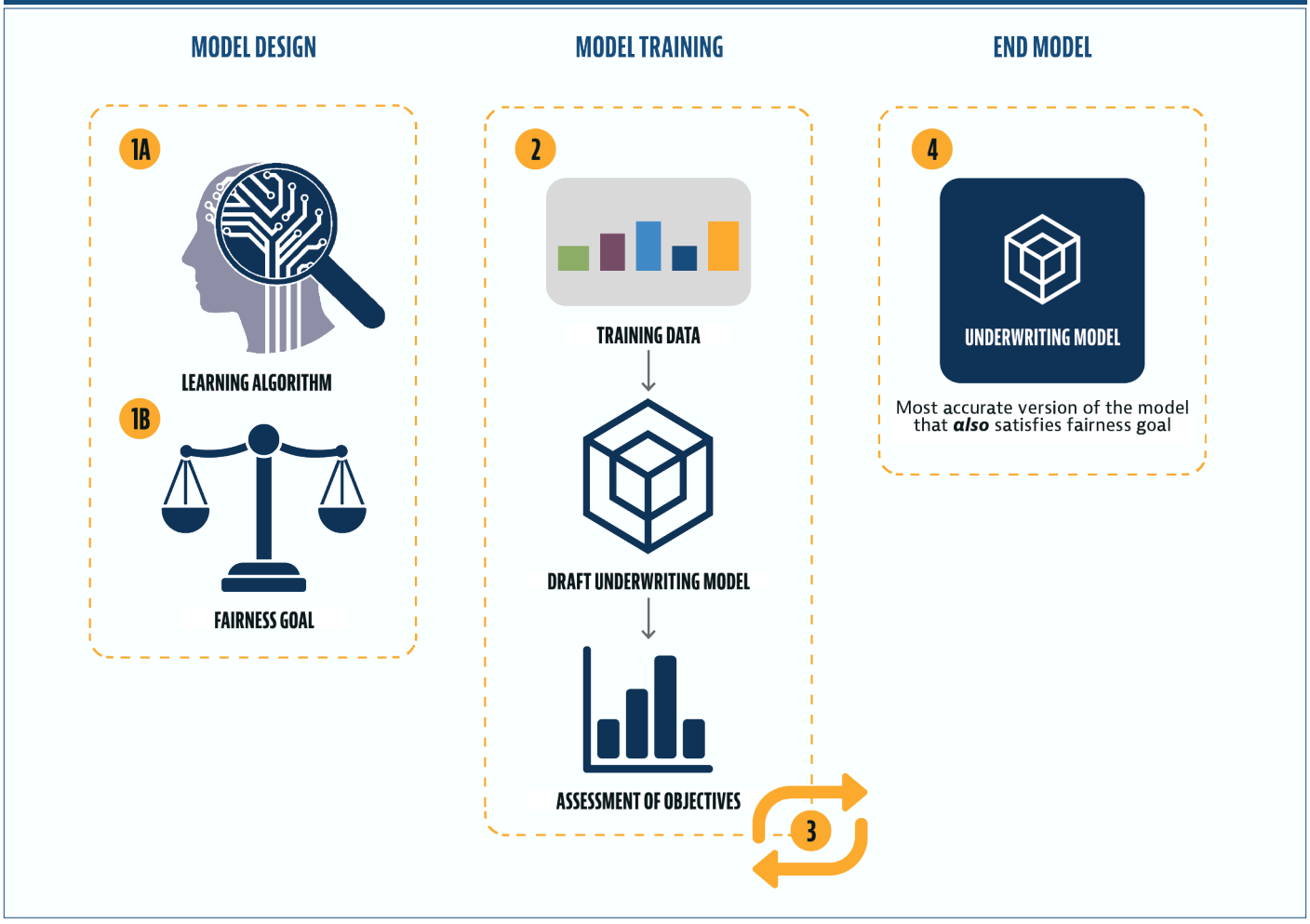
These methods typically require use of protected class information to evaluate the fairness of each draft or iteration of an underwriting model. While traditional methods also use protected class information to evaluate model disparities and the effect of mitigation attempts, iterative processes in the machine learning context are faster and more dynamic. The debiasing techniques do not use protected class information to make individual credit decisions, and the use of protected class information occurs during the training process when working to evaluate and mitigate disparities in the model. However, some stakeholders view this indirect structure as still raising substantial disparate treatment risk.

Joint Optimization

In a joint optimization approach the developer instructs the learning algorithm to simultaneously optimize two objectives as it builds successive iterations of the model, rather than simply identifying the model with the highest predictive accuracy. When used to debias machine learning underwriting models, the second objective is typically a fairness metric; for example, the learning algorithm could try to improve AIR while minimizing accuracy losses. A relative weight is assigned to each objective in the model's blended objective function, which dictates the trade-off the model developer is willing to accept when one objective must be sacrificed for another.

The diagram below shows the process of joint optimization:

JOINT OPTIMIZATION PROCESS



In the first step of joint optimization, the user selects a learning algorithm and identifies its objectives, among other things. In model training, the algorithm will first find a preliminary model based on the training data. That preliminary model specification—a first draft of the underwriting model—will then be assessed for fairness, among other objectives. This assessment provides a direction in which the algorithm can proceed to find a model that can better meet the blended objective. Real or imputed protected class data is required by the learning algorithm to run a statistical assessment of disparities.

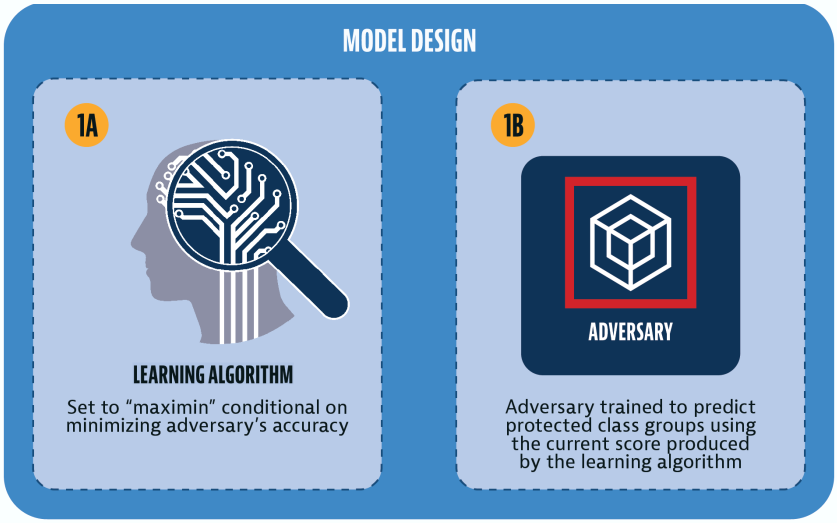
As the learning algorithm iterates additional drafts of the model (stage 3 on the diagram), each draft improves the model's accuracy or fairness to better satisfy the blended objective. This process repeats until gains in fairness necessitate unacceptable losses to accuracy, and vice versa.

Adversarial Debiasing

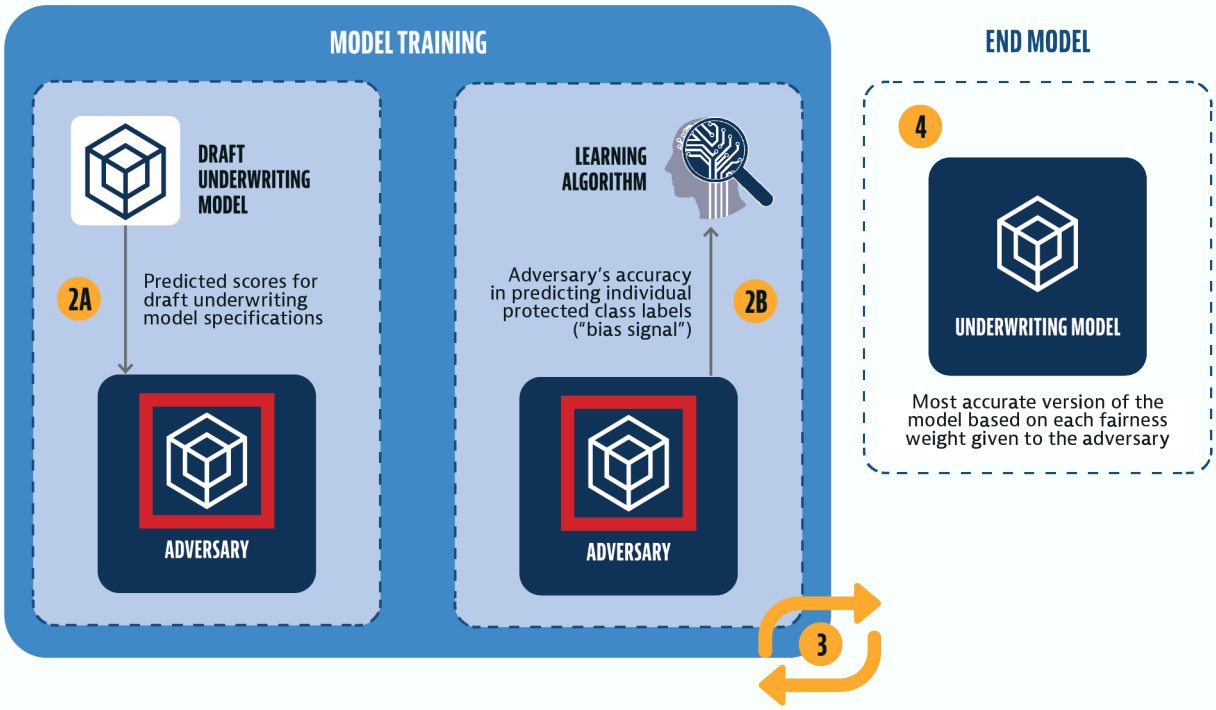
Adversarial debiasing is a method in which a second model—an "adversary"—is used to estimate the distribution of default predictions for various protected class groups for each iteration or draft of the underwriting model being developed.[202] The adversary is designed to predict protected class status, not default risk, based on the underwriting model's default predictions, and its learning algorithm has access to protected class information. At the outset of the process, the adversary makes a random guess about the protected class status associated with each score. The adversary's learning algorithm then assesses the accuracy of the adversary's prediction of the protected class characteristics of the applicant associated with each score. The underwriting model's learning algorithm takes this feedback and tries to minimize the adversary's accuracy while improving the accuracy of default predictions. The underwriting model learns to produce default predictions that are less correlated with protected class, and, as a result, the adversary becomes less accurate. The adversary does not make individual credit decisions, nor does its use typically extend beyond the portion of time in which users are searching for less discriminatory alternatives.

The diagram below shows the process of adversarial debiasing:

ADVERSIAL DEBIASING PROCESS



ADVERSIAL DEBIASING PROCESS



As in joint optimization, the user selects a learning algorithm that will develop an underwriting model and defines its objective as predicting the risk of default (step 1a in the graphic above). Unlike joint optimization, where the fairness component is calculated using protected class labels directly, adversarial debiasing takes a different approach. In the next step of preparation (step 1b in the diagram), the user sets up a second model to be trained—the adversary. This model is designed to estimate the protected class status of the borrower using the predictions of the underwriting model as inputs. The accuracy of the adversary is generally evaluated by comparing its outputs to protected class information. After initial training, the learning algorithm for the underwriting model is also given an additional objective of minimizing unfairness by minimizing the accuracy of the adversarial model. The underwriting model's learning algorithm aims to minimize its prediction error on default risks while trying to produce scores that prevent the adversary from accurately predicting the sensitive features. Thus, while joint optimization directly engages with fairness metrics, adversarial debiasing works by diminishing the ability of an adversary to infer sensitive features. This type of conditional objective is described as a "minimax" outcome.

The developer then assigns a fairness weight for the underwriting model that dictates the balance between the model's primary goal (predicting defaults) and secondary goal (reducing bias by minimizing the accuracy of the adversary model), and then uses the learning algorithms to train successive version of the underwriting model and the adversary as reflected in steps 2a and 2b. In each iteration, the underwriting model's learning algorithm is provided feedback based on the adversary's success rate. If the adversary can accurately assign protected class characteristics using the model's predictions, it indicates potential unfairness. The underwriting model's learning algorithm responds by adjusting its parameters to produce predictions that are harder for the adversary to use for guessing protected features. The adversary similarly trained on each iteration of the underwriting model's predictions.

As iterations continue, both the underwriting model and the adversary adjust. The underwriting model never directly receives information about protected status; instead, this information influences the model indirectly through its objective of minimizing the adversary's accuracy. If the adversary can effectively predict protected characteristics, the underwriting model will pivot, potentially placing more importance on features that are less tied to protected class or even using negative correlations to mislead the adversary. The end goal is to reach a model that balances predictive accuracy with fairness.

This process is often repeated using different fairness weights, resulting in additional model candidates. Each fairness weight represents a different balance between predictive accuracy and fairness. Once a firm has completed iterations of this process for the desired range of fairness weights, the set of model candidates can then be evaluated to determine which option will become the final model.

6.1.4 Key Issues and Risks

The transition to more widespread use of machine learning underwriting models and availability of new debiasing approaches have focused attention on certain key issues and risks related to fair lending compliance. Many of these issues and risks are not unique to machine learning underwriting models, but many are more prominent in debates about the responsible use of those models. Those include:

- **Amplifying Effects of Past Discrimination:** Practically all underwriting models rely on past lending data to predict future lending behavior and are therefore prone to transmitting biases embedded in historical credit data due to past discrimination in various forms.[203] However, stakeholders have expressed concern that the powerful machine learning algorithms that detect patterns related to default risk in that data may build models that increase those biases, just as they are able to increase model accuracy.
- **Reliance on Proxies:** Use of machine learning underwriting models produces heightened concern about whether learning algorithms can and do infer protected class characteristics from training data that does not include such information.[204] If an algorithm can infer that information, the resulting underwriting model may be using protected class information as the algorithm identifies relationships to be used in the underwriting model. Similarly, machine learning models may be relying on features and relationships among features that are statistical proxies for protected class characteristics and/or return predictions of default risk with large disparities across protected class groups.
- **Model Transparency:** Traditional fair lending practice has relied heavily on identifying, analyzing, and manipulating individual input features to mitigate fair lending risks. Complex models that may involve hundreds of input features and thousands of interactions between features in the model call into question whether and how to apply traditional methods for identifying, assessing, and reducing disparities. It may not be clear to compliance staff whether a machine learning model is relying on a proxy or has reverse engineered protected class status, and simply omitting or transforming initial inputs may not have the same effect as in a traditional regression model.
- **Regulatory and Business Uncertainty:** Lenders have a strong incentive to maintain the *status quo* in light of regulatory uncertainty and risks raised by fair lending enforcement actions. New methods to find less discriminatory alternatives that make significant gains in fairness and accuracy relative to traditional approaches may nonetheless be viewed by institutions as having unacceptable regulatory risk, for instance due to uncertainty with

regard to how they use protected class information. Uncertainty about how the methods produce their fairness gains and whether the alternative models will hold up in changing data conditions may also make lenders cautious about adoption.

6.2 Key Policy Issues

As the adoption of machine learning models accelerates, stakeholders are debating both the utility of traditional and new debiasing approaches and the potential value of updating regulatory guidance on fair lending compliance.[205] This section describes key debates, focusing first on disparate treatment and then on disparate impact. The disparate impact issues include measuring fairness; use of protected class information and debiasing techniques; and expectations for identifying less discriminatory alternatives, including both when to search for them and how to evaluate alternative models.

6.2.1 Disparate Treatment

As discussed in Section 6.1.3, the most commonly used test for evaluating whether input features are impermissible proxies calculates the level of correlation for each input feature with both protected class on one side and model performance on the other, but some firms go further by assessing the impact of variables on each protected class in isolation.

In the context of machine learning models, proxy analysis is complicated both by the potential for nonlinear and non-monotonic relationships and by complex interactions between features within the model. This challenge is prompting some stakeholders to turn to interpretable models, using architecture constraints to limit the creation of latent features and restrict the operation of such features if they have different distributions for different protected classes.[206] Others may start disparate treatment analyses for more complex models by building single-variable machine learning models or by using surrogate models to evaluate which sets of input features are most predictive of protected class characteristics,[207] and then training separate underwriting models for different subgroups to test the performance effects of excluding the variables of concern.[208]

Yet while these approaches and explainability techniques more generally can be used to evaluate the importance of individual input features to model operations, the most commonly used techniques cannot directly and precisely map feature interactions within more complex ML models.[209] Thus, similar to the model risk management and adverse action notice contexts, stakeholders are mulling the importance of pinpointing specific feature interactions to assess whether those interactions might be reverse engineering protected class status or might be considered statistical proxies for protected characteristics.

Beyond the technical debates about the application of current explainability techniques, this issue raises important conceptual issues. A learning algorithm that has been directed to find the most predictive underwriting model it can is not acting with intent to circumvent prohibitions on the use of protected class characteristics or relying on features that are a pretext for an applicant's race or gender. The model's learning algorithm is not seeking to identify protected characteristics or to find relationships that favor or punish members of protected classes. Moreover, recent research has highlighted some of the limitations of traditional approaches that rely on excluding input variables that are closely correlated with protected characteristics to reduce disparities produced by machine learning models, while suggesting that more automated approaches to debiasing models hold more promise (See Box 6.2.1.1 and Section 6.2.2.2).

BOX 6.2.1.1 LIMITATIONS ON THE EFFECTIVENESS OF EXCLUDING INPUT FEATURES

In addition to FinRegLab's disparate impact research as discussed in Section 6.2.2, empirical research by Talia Gillis of Columbia Law School has explored the limitations of excluding protected class and inputs that are statistical proxies for protected characteristics as a means of reducing disparities in underwriting models.[210] Using the Boston Federal Reserve Bank's Home Mortgage Disclosure Act (HMDA) dataset, Gillis constructed two models to predict default risk, one of which included all of the HMDA data elements except race and another that excluded both race and the ten features most strongly correlated with race.[211] Excluding the additional variables had only a modest effect on disparities in predicted default risk between White and non-White borrowers.[212] Further restricting the list of inputs produced smaller disparities but also reduced accuracy.[213]

Gillis also constructed models to predict whether borrowers were Black using only the HMDA data elements and only a form of zip code information, which is typically excluded from underwriting models due to concern that it operates as an impermissible proxy for race and ethnicity.[214] The model trained on traditional underwriting inputs was more accurate in predicting protected class than zip code alone, emphasizing the extent to which intuitive notions of which variables may be correlated with protected class may be incomplete.[215] Similarly, the HMDA data could be used to construct models to predict age and marital status.[216]

The article concludes that while intuitively appealing, practical challenges to defining and detecting proxies are endemic. Rather than focusing on fairness strategies that reduce predictive accuracy—which may also have disproportionately negative impacts on protected classes—the analysis suggests that fair lending assessments in the machine learning context should shift from managing individual inputs toward expanding analyses of outcomes.[217]

These considerations raise important questions about whether the disparate impact framework is both more appropriate conceptually and more effective practically for evaluating and mitigating potential concerns about ML models' fairness. As discussed further below, the availability of a more effective debiasing toolkit for managing disparate impact risks may provide a compelling counterweight to concerns about potential proxies in feature interactions constructed within the model. These considerations further underscore the importance of additional research into the effectiveness and limitations of machine learning debiasing techniques and of clarifying expectations around searches for less discriminatory alternative models.

6.2.2 Disparate Impact

As stakeholders deepen their understanding of various debiasing tools and implementation choices, public policy questions regarding disparate impact compliance have taken on additional urgency in light of the adoption of ML models. Additional regulatory guidance on these issues could help to determine the extent to which ML models—particularly when combined with more inclusive data sources—meaningfully increase access to credit.

6.2.2.1 Measuring Fairness

As discussed in Section 6.1.3, lenders who are vetting models for potential disparate impact risk typically use AIR to assess disparities in decisions to approve or deny credit and SMD in the context of pricing disparities.[218] Both metrics compare disparities in the default predictions between demographic groups, but do not account for differences in the applicants' financial circumstances or the general accuracy of the model's predictions with regard to different demographic groups. As a result, it is technically possible for a lender to achieve a perfect AIR score of 1 if it is willing to approve loans to applicants who are not likely to be able to repay them, even though such actions would raise serious concerns about both predatory lending and the lender's safety and soundness. In light of these limitations, some stakeholders have suggested that using the traditional metrics in

isolation creates tension with other regulatory regimes.[219] In practice some lenders use additional performance metrics to place AIR/SMD in context.

However, finding replacement measurements of fairness that can serve ECOA's various legal requirements is challenging. While the broader data science community has engaged in a robust dialogue over the last decade regarding how best to measure and achieve model fairness, some options are precluded in the lending context due to data limitations and the constraints imposed by the disparate treatment doctrine.[220] Stakeholders report substantial variance in the extent to which alternative fairness metrics such as predictive accuracy by group are being used today, with some lenders only deploying it in limited or exploratory contexts if at all and others treating it as a core element of compliance reviews.[221]

Within the credit ecosystem, stakeholders appear most interested in the potential adoption of fairness metrics that identify whether a model provides default predictions with the same level of accuracy across different populations. For instance, one option would be to apply the same general performance metrics that are typically used in model development processes[222] across different populations to assess whether the model has the same predictive accuracy for different demographic groups and for subgroups with similar risk profiles.[223] While some lenders may perform such analyses as part of their internal compliance programs, proponents of making them a specific regulatory focus argue that such measures are more appropriate and realistic because they hold lenders accountable for disparities in predictive accuracy rather than outcomes that are shaped by entrenched societal inequality.[224]

6.2.2.2 Use of Protected Class Information and Debiasing Techniques

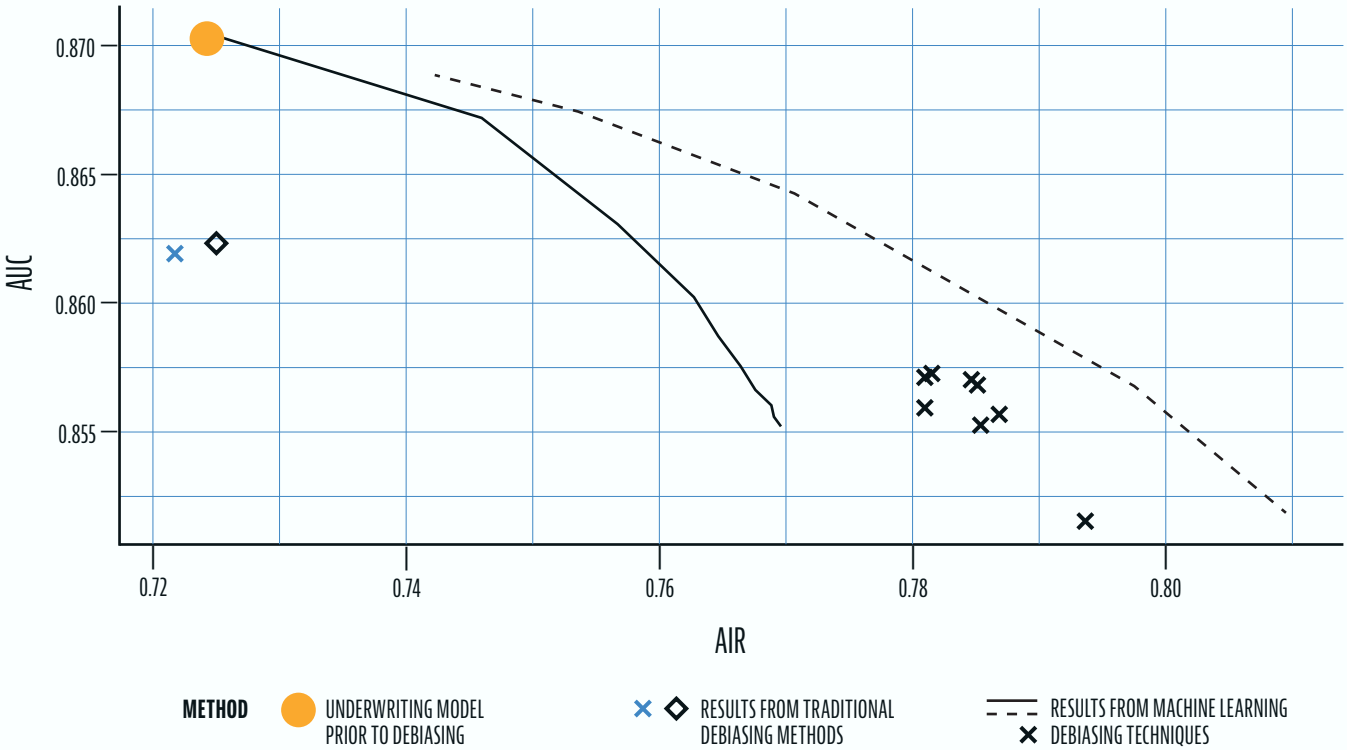
A second threshold question concerns whether the use of protected class information and newer debiasing techniques is permissible under the disparate treatment doctrine to the extent that the techniques use data about protected class membership in different ways than traditional mitigation approaches.[225]

As described in Section 6.1.3, techniques such as joint optimization and adversarial debiasing do not use protected class data to make individual credit decisions, although they do use it during the training process for assessing model prediction disparities and making a series of rapid adjustments and iterations to search for alternative models. In addition to these techniques, many vendors provide general platforms or other services that facilitate the rapid iteration of models through assigning weights, changing model constraints, and other adjustments. These various automated processes are more dynamic than traditional methods, and in some but not all cases involve exposing the learning algorithm directly to protected class information so that it can determine which combinations of input features maximize predictiveness and minimize disparities. Concerns about violating prohibitions on disparate treatment have slowed the initial adoption of joint optimization and adversarial debiasing in the credit context relative to their use in some other sectors.

Our research suggests that new automated tools can be quite powerful in using machine learning techniques to develop models to reduce disparities. Where we tested approaches that relied on traditional mitigation strategies focusing on a narrow subset of features, model performance declined with little to no improvement in fairness. But more automated approaches—which include a range of strategies including but not limited to joint optimization and adversarial debiasing—were able to produce a menu of options that provided larger fairness benefits and smaller accuracy tradeoffs. While we did not test the full spectrum of approaches or fully evaluate each individual alternative identified, our findings illustrate the more powerful toolkit that combining machine learning with *post hoc* tools can provide in searching efficiently for fairer models.

The graph below illustrates some of the results from different debiasing methods. Accuracy is represented by the area under the curve (AUC), a commonly used measure of predictiveness, while fairness is represented by the adverse impact ratio (AIR), a measure of the disparities in the selection rate between minority and non-minority consumers. As reflected in the graphic, the traditional debiasing methods (blue X and black diamond) were significantly less predictive than the baseline model (orange dot), but did not significantly improve fairness. The automated approaches (solid line, dashed line, black Xes) substantially improved fairness, with varying changes in predictive accuracy.[226]

FAIRNESS-PERFORMANCE CHARACTERISTICS OF LESS DISCRIMINATORY MODELS IDENTIFIED BY THE MODEL DIAGNOSTIC TOOLS



Note: AUC is a commonly used measure of predictiveness. AIR is a measure of disparities in the selection rate between minority and non-minority consumers.

It is not surprising that modifying or dropping a handful of features may have relatively marginal effects on model disparities where models involve larger numbers of features. As discussed in Section 6.2.1, academic research also suggests that simply dropping features that are closely correlated with demographics may not be an effective strategy for reducing disparities in machine learning models due to the presence of correlated features.[227] Some stakeholders report they are assessing whether evaluating groups of correlated features rather than individual features can strengthen the effectiveness of traditional strategies, and others are focusing on using architecture constraints to limit and manage the use of latent features that raise potential disparate impact concerns. However, these factors have also increased interest in automated debiasing tools as potentially more effective and efficient strategies for managing fairness concerns in the ML context.

At the same time, skeptics have questioned whether some alternatives generated by debiasing techniques may be achieving greater fairness by altering the weights of features in ways that could be difficult to justify under model risk management guidance or that may not hold up when there are

shifts in economic conditions or other data distributions. Research into specific debiasing approaches could thus be helpful to illuminate the most promising methodologies and specific implementation choices that lenders face when deploying these techniques.[228] It could also be helpful to probe the alternative models generated by such tools, for instance to understand the extent to which any declines in accuracy tend to be concentrated among different subgroups, how well the models perform in general risk management validation processes, and the extent to which fairness improvements remain robust in changing data conditions.[229] Thus, while the initial results are promising, additional public research could give lenders and regulators more confidence in selecting both specific debiasing approaches and from among the range of models that they generate.

Regulatory guidance could also be helpful. In recent years, lenders who are adopting machine learning models appear to have become increasingly comfortable with authorizing their fair lending compliance teams to deploy debiasing techniques during searches for less discriminatory alternatives, while prohibiting their use by business units in earlier development stages. In practice, this means machine learning debiasing techniques are used as part of reviewing a new underwriting model proposed for use, monitoring an underwriting model already in use, and assessing updates to an underwriting model in use or changes in factors that affect disparities in credit decisions, such as credit score cutoffs. This bifurcation is consistent with historical fair lending compliance practice and guards against the risk of misuse of protected class information by initial development teams. However, depending on how lenders sequence their overall model development process, bifurcation may lengthen overall timelines for validation and deployment.

Proponents of using such techniques during the initial development process point out that it can present a stronger fair lending baseline at the outset of compliance review processes and may be particularly advantageous for smaller firms that may struggle to attract and maintain equal levels of technical expertise in both their business and fair lending units. For firms that conduct conceptual soundness and general model validation before fair lending review, earlier deployment of the techniques could increase the chances that the lender does not have to repeat such processes a second time. However, enabling model development teams to do this work earlier in the process would emphasize the importance of establishing internal controls and oversight mechanisms to define and police what uses are permissible. Some stakeholders also suggest that having independent general validation processes as part of model risk management could be an important governance mechanism particularly at a time when stakeholders are still learning about the potential utility and limitations of more automated techniques for identifying less discriminatory alternatives.

Moreover, absent further regulatory guidance, some lenders remain reluctant to authorize the use of the techniques even by traditional compliance teams or vendors, separate from the initial model development process. Particularly in the absence of greater clarity about when they are obligated to search for less discriminatory alternative models in the first instance, they are reluctant to take on the technical and compliance questions involved in managing both machine learning models and new approaches to LDA discovery.

6.2.2.3 Identifying Less Discriminatory Alternatives

The transition to machine learning underwriting models has also highlighted policy questions about regulators' expectations for lenders in searching for and identifying alternative models that reasonably meet the lender's legitimate business need to predict default risk while producing less disparity in predicted outcomes among protected groups. Many lenders today do not invest substantial resources in searching for LDAs, particularly where they are relying on traditional techniques and data sources and not making significant changes to their existing underwriting systems.[230]

As discussed further below, questions about the broader search for less discriminatory alternative models include:

- Do regulators expect lenders always to search for LDAs during the model development process, or only in certain circumstances? What range of options should be considered during a search process?
- To the extent that alternative models involve some reduction in predictive accuracy, is there a threshold past which such models should not be considered LDAs because the performance losses are too large?
- If an alternative model reduces disparities for one group but increases them for another or hinges upon relationships that raise other policy or regulatory concerns, should it be considered an LDA?

As noted in Section 6.1.2, CFPB officials speaking at conferences in 2023 described "rigorous searches for less discriminatory alternatives" as "a critical component of fair lending compliance management" and expressed concern that lenders may tend to shortchange this aspect of compliance.[231] However, the agency has not issued formal guidance on LDA topics to date despite urging by advocates and some industry stakeholders.[232]

When Searches for Less Discriminatory Models Are Required

As described in Section 6.1.3, lenders often rely on two tests—statistical significance and practical significance—to determine whether a disparity warrants further investigation and searching for LDAs. Statistical significance tests whether a disparity for protected class groups can be featured to operation of a facially neutral practice or policy. Practical significance helps lenders assess whether a particular disparity puts them at risk to be found to have committed a legal violation of the ECOA.

Lenders which have adopted practical significance standards will generally not conduct searches for less discriminatory alternative models unless an adverse impact is beyond the thresholds that they have adopted for both statistical and practical significance.[233] Lenders report that practical significance standards allow them to focus investigation resources efficiently on disparities most likely to lead to legal violations.[234] However, to date, neither court decisions nor regulatory guidance have recognized use of practical significance standards to determine when searches for alternative model specifications are required in the context of credit underwriting.[235] In the absence of regulatory guidance providing thresholds and in light of varied industry practice, one recent monitorship report using the following as practical significance cutoffs: AIR less than 0.90 (where a lower AIR means greater disparities) makes an approval/denial disparity practically significantly and an SMD greater than 0.30 (where a higher SMD means greater disparities) makes an APR disparity practically significant.[236]

Defining Which Models Reasonably Meet Legitimate Business Needs

Where machine learning debiasing methods are used, lenders now have the ability to generate a spectrum of model specifications that reduce protected class disparities in the model's predictions. Where options with smaller disparities come with reductions in model accuracy, lenders are faced with a difficult choice: what constitutes a valid less discriminatory model specification? In this context, understanding better when an alternative model has to be adopted under the disparate impact requirements may help lenders define the scope and intensity of their efforts to identify alternative models.

In general, the ECOA seeks "to promote the availability of credit to all creditworthy applicants" without regard to enumerated protected characteristics such as race, religion, national origin, sex,

marital status, or age.[237] Lenders may be at risk for a violation where they failed to adopt an "alternative [model] that is *approximately equally effective* is available [and] that would cause less severe adverse impact."[238] However, regulators have not generally specified how lenders can or should specifically assess when a proposed alternative model has sufficiently close predictive accuracy to the original model to require adoption.

Many stakeholders accept the notion that lenders are not generally required to accept a less discriminatory model that meaningfully reduces the predictive accuracy of their underwriting model—in part because they are concerned that increased defaults may be concentrated among traditionally underserved groups.[239] However, these stakeholders also urge lenders and regulators should adopt greater rigor in examining whether and in what circumstances a proposed alternative offers sufficient performance to warrant adoption.

In this area, some stakeholders have suggested that thresholds and ranges that lenders use for general performance testing could potentially be adapted to screen for LDAs, suggesting that such metrics can be viewed as defining what range of performance meets the firm's business needs.[240] For example, if the original model was validated on the basis of a performance variation of +/- 0.01% of AUC, they argue there is a compelling case that a less discriminatory alternative model that has performance within that range and reduces disparities should be adopted.[241] Proponents suggest that one of the advantages of this approach is that it reflects firm-specific conditions, since the level of performance variation that a firm recognizes in model review processes reflects a range of business factors including customer base, past performance history, and risk tolerances. However, implementing such an approach may be challenging, insofar as firms need to account for the confidence intervals (or the uncertainty level) associated with each option that they identify.[242]

A second approach to defining LDAs advanced by a group of consumer advocates and other stakeholders puts more emphasis on the role of public oversight.[243] They call on the CFPB to conduct its own searches for less discriminatory alternatives during fair lending examinations. One proposal suggests that the agency should also make public its methodology as a lever to improving and standardizing industry practice.[244] Another proposal focuses on the agency screening machine learning models in underwriting, marketing, and collections using a large data set maintained by the agency and conducting its own searches for less discriminatory alternative models where lender models exceed a designated risk threshold.[245] Beyond potential questions about a public sector actor selecting a lender's underwriting models, both proposals may present practical challenges in terms of the agency staffing and would require applicable thresholds and standards for generating LDAs. The second proposal also raises questions about whether the agency can establish an appropriate data set for this purpose and the feasibility of using it to generate alternative specifications for models trained on entirely different data.

Validating Alternative Models Against Other Criteria

Where one or more alternative models have been determined to fall within an acceptable performance range, stakeholders note that lenders must still engage in general validation testing of the alternatives and assess other compliance considerations. For example, many stakeholders question whether alternative models that offer improved fairness in aggregate but which affect different protected class groups differently qualify as an LDA.[246] In practice, some stakeholders report that such conflicts occur with some regularity for individual protected class groups and compound groups. For example, an alternative model that produces smaller disparities for Black applicants may produce larger disparities for Asian applicants just as reducing disparities for Black men can coincide with

increasing them for White women or vice versa. In general, stakeholders report that lenders will not adopt a model as a less discriminatory alternative that harms either group.[247]

Three reports from Relman, Colfax PLCC describe their methodology for assessing disparate impact risk in the context of monitoring a fintech lender's operations, including searching for less discriminatory alternative models, and in so doing point to a potential framework for future guidance on this topic.[248] The reports set forth several basic assumptions about lenders' decisions about which less discriminatory model to adopt. In general, a viable less discriminatory alternative model will produce a meaningful improvement in the relevant fairness metric and will not:

- Require that the lender accept meaningful reductions in a model's performance.
- Violate other requirements applicable to underwriting models, such as validation criteria under prudential model risk management requirements and the firm's policies.
- Enhance existing disparities that were present in the original model.
- Introduce new disparities that were not present in the original model.
- Reduce disparities for one protected class group while introducing new adverse effects—such as reducing the model's accuracy—for another protected class group.

Further analyses of alternative models are necessary to better understand the specific implications of each of these principles and apply them in the context of a lender's model development processes.

One advocacy organization has proposed a quantitative approach to identifying LDAs as "a systematic way of requiring lenders to place fair lending concerns on equal grounds with repayment risk."[249] Specifically, the National Community Reinvestment Coalition has proposed that the CFPB require lenders to calculate the ratio of each alternative model's changes in fairness to changes in performance. Under this approach, an "accommodation ratio" of 5 means that a one percentage point decrease in a model's predictive accuracy would correspond to a 5% reduction in disparities. The organization calls for the agency to require that lenders adopt models with ratios of 2.0 or greater and to study whether to set the threshold lower, arguing that "some accommodation should be made to preference gains in fairness—even if the new iteration forces lender[s] to make a modest concession to model quality."[250] However, other stakeholders report concern that market-wide quantitative standards may not adequately account for variations in market dynamics for firms that serve different customer segments or for the populations most affected by decreases in the new model's predictive accuracy.

7. CONCLUSION

Our research suggests that the thoughtful deployment of evolving explainability and debiasing techniques can help to manage concerns about the transparency and fairness of ML underwriting models, but that broader evolutions in market practices and public policy are also critical to address fundamental questions about our ability to trust more complex models. At this relatively early phase when ML technologies and our understanding of them are changing rapidly, defining certain basic concepts and expectations could be particularly helpful to encourage responsible use.

Our empirical analyses found that some *post hoc* explainability tools produced reliable information about various aspects of model operations, but that tool choice, implementation details, and interpreting the results in light of the underlying data are important. We found that debiasing techniques and other automated approaches produced a range of model alternatives with greater predictive accuracy and smaller demographic disparities than traditional fair lending strategies, although lenders still face important choices in selecting among alternatives and uncertainty about underlying regulatory expectations. More broadly, the fact that the most commonly used explainability techniques today cannot directly and precisely map feature interactions within the most complex ML models is raising questions across several different regulatory areas about whether it is critical to be able to perform such analyses in order to ensure the fair and responsible use of such models.

In the face of these technical and broader conceptual questions, some lenders and model builders are deploying these new explainability and debiasing techniques in various ways, others are primarily emphasizing architectural constraints to produce interpretable ML models, and many are choosing not to proceed with ML adoption. However, competitive pressures and tighter and more uncertain economic conditions may create stronger incentives to seek greater predictive power, thereby increasing the urgency of addressing both technical and broader policy questions about the responsible use of machine learning, explainability, and debiasing techniques.

Publicly available research will be critically important to deepen our understanding of current techniques and continuing efforts to produce better tools and approaches. As outlined in Box 7.1, topics include the general predictiveness and inclusion effects of machine learning models (particularly in combination with new data sources) as well as additional details about different approaches and methods for managing explainability and fairness concerns.

Conversations and collective learning within and across different stakeholder groups will also be critical to building shared understandings about the trustworthy deployment of ML models and explainability and debiasing techniques. Dialogue is critical not only across the credit ecosystem, but also with other sectors that are also working to manage the deployment of AI/ML models across other high-risk use cases.

BOX 7.1 KEY AREAS FOR ADDITIONAL RESEARCH

Potential topics include:

- Further evaluation of the predictiveness, inclusion, and fairness effects of using ML underwriting models with or without non-traditional data sources such as cash-flow information as compared to traditional techniques and data sources.
- Assessing with rigor the transparency costs related to the use of more complex machine learning underwriting models and the benefits and tradeoffs of using up-front constraints on model complexity to manage transparency and other compliance concerns.
- Deeper evaluation of specific implementation choices that affect the performance of different explainability tools for different tasks, such as identifying whether and how the definition of the baseline set to which a rejected applicant is compared affects the quality of information given to consumers on an adverse action notice.
- Additional analysis of the extent to which different explainability tools disagree about important factors after accounting for correlations to classify the types of disagreements that persist and consider whether those types of disagreements have a material effect on the regulatory compliance tasks considered in this evaluation.
- Continuing refinement of assessment frameworks for evaluating *post hoc* explainiability tools, including consideration of whether different or additional qualities can help to identify when information from such tools can be trusted and used in high-stakes contexts.
- Deeper evaluation of specific debiasing approaches to illuminate the most promising methodologies for mitigating bias in machine learning underwriting models and the specific choices that lenders make when deploying those methods, including whether and how protected class characteristics (whether actual or imputed) can be responsibly used to improve the fairness of credit decisions.
- Deeper evaluation of the performance-fairness tradeoffs identified by debiasing tools that generate a range of alternative models, for instance to confirm whether there is a band in which lenders can improve the fairness of models without incurring significant loss of performance and how potential performance tradeoffs distribute across populations of interest.

Even as regulators are continuing to deepen their knowledge of critical issues, there are steps that they could take to encourage the development and adoption of responsible implementation practices:

- Updating governance frameworks, including potentially articulating the qualities of trustworthy AI/ML models similar to the ones described in Section 2.4 could encourage lenders to begin methodically evaluating and testing their systems and processes to address those core components. Such principles-based approaches can be especially helpful at early stages of evolution across diverse stakeholders, markets, circumstances, and technologies.
- In a similar vein, articulating the key qualities for explainability and diagnostic tools would also help lenders begin to manage for a consistent set of questions and concerns, even as the technologies and assessment processes continue to evolve.
- Given current variations in whether and how lenders search for less discriminatory alternatives to baseline underwriting models, providing greater clarity on what constitutes an LDA and on regulators' expectations for search processes could significantly increase consistency in the market.

The current moment presents both significant risk (as millions of credit applications are being decided based on firms' best judgments as to regulatory compliance and *post hoc* tool use) and significant opportunity (as policymakers have a unique moment in which they can affect the broad direction of evolution, before developing more calibrated and binding standards as the innovation lifecycle progresses). It also presents an opportunity to re-think and improve upon prior generations

of automated underwriting and the extent to which they have left substantial numbers of people behind and replicated historical disparities. The coming years could offer the most fundamental reset of lending practices in several decades. Whether and to what extent those new systems prioritize responsible, fair, and inclusive use of ML models and explainability tools will ultimately depend not just on technology issues but on business and policy decisions. Rigorous research, thoughtful deployment, and proactive regulatory engagement are critical to ensuring that any new technology must ultimately benefit borrowers and financial service providers alike.

APPENDIX A

FinRegLab Policy Working Groups and Advisory Board

FinRegLab convened more than 130 representatives of lenders, banks, fintechs, advocacy organizations, researchers, and other stakeholders to participate in the Advisory Board and the Policy Working Groups for this project. These stakeholders engaged in an extended dialogue about the challenges and opportunities of adoption of machine learning techniques in credit underwriting. Representatives of several federal and state agencies attended the sessions in an observer capacity.

This report was informed by the feedback of these and other stakeholders but represents FinRegLab's independent analysis in all respects. It does not necessarily accord with the views of the individual participants or their employers.

Affirm, Inc.
Aire
American Civil Liberties Union
American Express
Amruta, Inc.
Bank of America
Bates White Economic Consulting
Beneficial State Foundation
Bill & Melinda Gates Foundation
Capital One
Center for Responsible Lending
Charles River Associates
Citi
Columbia University
Consumer Financial Protection Bureau*
Consumer Reports
Discover
Federal Deposit Insurance Corporation*
Federal Reserve Board*
FICO
Goldman Sachs
Harvard Belfer Center
Harvard Law School
Hudson Cook, LLP
JPMorgan Chase & Co.
KeyBank
Luminos.Law
Mastercard
Microsoft
National Consumer Law Center
National Credit Union Administration*
National Fair Housing Alliance
National Institute of Standards and Technology*
Navy Federal Credit Union
New York State Department of Financial Services*
Nyca Partners
Office of the Comptroller of the Currency*
Oliver Wyman
PayPal
Petal
Public Democracy America
Regions Bank
Relman Colfax PLLC
Rensselaer Polytechnic Institute
Skoll Foundation
SmileDirectClub
Stanford Graduate School of Business
Synchrony Financial
TechEquity Collaborative
The Alan Turing Institute
The Brookings Institution
TransUnion
TruEra
University of Virginia
UpStart
Wells Fargo
World Economic Forum

APPENDIX B

Key Terms

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.

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.

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.

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.

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.

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 modeling 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.

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.

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 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

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 features 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 explanations: Global explanations refer to explanations of 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.

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 features 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 explanations: Local explanations 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.

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.

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.

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 can provide local explanations. This method measures feature importance by removing features from a data point and quantifying how much that affects the model's output.

Surrogate models: Surrogate models refer to interpretable models that mimic and explain the behavior of more complex models.

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.

Extreme Gradient Boosting (XGBoost): Extreme Gradient Boosting 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 C

Recent Research on Related Topics

This appendix summarizes recent empirical research that is particularly relevant to the issues considered in this report. Some of the studies explore activities that may not be required by or consistent with existing law or practices–for example by exploring the feasibility of generating notices designed to inform applicants of feasible paths to loan approval within a year or exploring new approaches to using protected class information to debias models–but may be informative to stakeholders in considering options for future evolution. The following research is summarized in this section:

- *The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective,* Satyapriya Krishna, *et al.*
- *The Input Fallacy,* Talia Gillis
- *The Time is Now: Advancing Fairness in Lending Through Machine Learning,* Vitaly Meursault, *et al.*

C.1 *The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective*

Authors: Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju

Publishing information: Pre-print May 2023

A team of data scientists qualitatively and quantitatively investigated an increasingly common real-world problem: disagreement among multiple *post hoc* explainability techniques being deployed simultaneously to help model developers obtain a richer and more nuanced understanding of model behavior. This study used four real-world data sets, a suite of popular "black box" machine learning model types, and a set of widely used open source explainability techniques.[251] Although one of the information sources was a German credit data set, the study did not focus on the production of adverse action notices or similar higher-level disclosures to credit applicants,[252] but rather on use of explainability tools by data scientists. The authors find that in practice disagreements among explainability techniques are common and that model developers often lack a principled framework for resolving such disagreements.

Methodology. The authors interviewed data scientists to identify what constitutes a disagreement in explanations provided by *post hoc* explainability methods and formalized their understanding in a quantitative framework designed to document and measure the extent of disagreement in explanations by popular explainability methods using a set of new metrics. Alongside their empirical

research, the authors conducted an online user study designed to identify whether and how data scientists using *post hoc* explainability methods resolve explanation disagreements.

To conduct the empirical study, the authors trained the following series of models for use with tabular data (including German credit data): a logistic regression model, a gradient boosted tree mode, a random forest model, and a neural network with 4 hidden layers. The explanation methods under evaluation were LIME, KernelShap, and four gradient-based explainability methods.[253]

In the qualitative portion of the study, practitioners focused on two considerations in identifying disagreements: (1) the extent to which techniques differ in identifying different top features of interest and the signs or directions of contribution of top features and (2) the extent to which the techniques differ in the relative ordering of certain features. To measure disagreements across various explainability methods, the authors propose six metrics:

- **Feature agreement:** whether the tools identified the same features
- **Rank agreement:** whether the tools rank ordered the identified features the same way
- **Sign agreement:** whether the identified key features have the same signs and direction of contribution
- **Signed rank agreement:** combines the foregoing to assess whether different tools identify the same top features, with same sign and direction of contribution, and in the same order
- **Rank correlation:** whether tools produce the same feature rankings for a set of features identified as particularly important to end users
- **Pairwise rank agreement:** whether different tools rank order the same features in the same way (if A is more important than B for one tool, is the same true for another tool?)

Key Findings. The authors find that the explainability techniques "often disagree" when explaining the same model. In particular, they find:

- As the number of key features considered increases, rank agreement and signed rank agreement generally decrease.
- Within the set of gradient-based explainability methods, some groups show strong internal consistency but disagree with the others.
- In terms of model complexity, the results on tabular data show greater disagreement among explainability tools when applied to neural network models than to a logistic regression model. This suggests that the weakness of some explanation strategies—such as LIME's estimation of a simpler model to approximate the "black box"—may cause greater inconsistency in information produced by the various tools.

Based on surveys of data scientists in academia and different industries, the researchers further identify the absence of "principled, well-established approaches" to reconciling those disagreements as a particular risk where these tools are used in high stakes contexts.

Implications and Further Research. The authors cite numerous opportunities for further research, including investigation of a systematic way to classify disagreements and their causes, developing relevant metrics, and educating practitioners about resolving disagreements. They also point to deeper potential for refining explainability techniques to avoid certain disagreements.

C.2 *The Input Fallacy*

Authors: Talia B. Gillis

Publishing information: Published April 2022 by the Minnesota Law Review

The author investigates whether new approaches to fair lending testing might better serve non-discrimination and inclusion purposes in the context of machine learning underwriting models so that gains in predictive accuracy achieved with modern modeling techniques also increase the availability of credit for historically disadvantaged groups.[254] The author advances two primary arguments to this end: (1) machine learning has undermined the efficacy of traditional fair lending approaches to managing disparities in underwriting models that focus on addressing fairness considerations via inputs to the credit decision, and (2) the advent of machine learning underwriting should enable exploration of new approaches to identifying and mitigating algorithmic discrimination, including emphasizing the use of empirical testing of algorithmic outcomes.

Methodology. The empirical component of this study consists of a simulation that evaluates the effects of three different approaches to managing disparities based on inputs: excluding protected class information, excluding proxies for protected class information, and restricting inputs to pre-approved features. The author uses the Boston Federal Reserve Bank's Home Mortgage Disclosure Act (HMDA) dataset containing data about residential mortgage loans to develop models for the simulation. The author fit a random forest machine learning model and a LASSO regression machine learning model on a random sample of 2,000 borrowers with more than 40 variables each to predict whether a hypothetical lender would reject an application for credit.[255] The author then calibrated the models' rejection rates to publicly available statistics on defaults.

Key Findings. Tracking the three main approaches to managing disparities based on inputs, this study finds the following:

- Excluding protected class information: The study questions whether formally excluding protected class information is effective from a fairness perspective because machine learning models can predict applicants' protected class information with relative accuracy based on other inputs. To show this, the author constructs models with additional features from the HMDA dataset to predict applicant age and marital status, which are protected characteristics under fair lending law. The model predicting age has an accuracy of 0.84, and the model predicting marital status has an accuracy of 0.90.[256]

  Further, excluding protected class information can actually increase disparities in models by failing to account for existing disparities among different demographic groups. To demonstrate this, the author constructs two models to predict default rates based on college attendance. One model is "race blind," while the other considers race. The model that considers race actually produces fewer disparities because it effectively predicts default rates based on college attendance within each subgroup, thereby factoring in variations in college attendance among White and non-White borrowers. The "race blind" model produces greater disparities because it generalizes across the entire population as a single group despite the underlying attendance disparities.

- Excluding proxies for protected class information: The author argues that excluding variables that are more closely correlated to protected class status than to the target variable (such as default) is similarly ineffective in reducing disparities in machine learning models. The author first notes that defining and identifying proxies is difficult in practice. Many

credit pricing inputs have a strong correlation with protected class information, and even a classic proxy like zip code likely contains information unrelated to protected class membership yet relevant to default risk. Secondly, the author shows that a model with traditional credit pricing inputs can better predict an applicant's race than zip code data—suggesting that classic examples of proxies for race, like zip codes, may be less indicative of race than other variables widely used by lenders. Next, Gillis constructs a model excluding race and its top 10 correlates from its inputs. This model excluding proxies produced less disparities between White and non-White borrowers than a model just excluding race, although non-White borrowers still had much higher credit risk in the proxy-excluding model. She therefore contends that excluding characteristics based on their individual correlations to protected characteristics is insufficient to ensure fairness in that it does not fully capture how variables correlate and interact in machine learning models.

- **Restricting inputs to pre-approved features:** The author compares two random forest models, one using the full set of inputs other than race, and another limited to a small subset of variables including variables traditionally used in credit pricing (e.g., income, debt-income ratio, and characteristics of loans). The model using a small subset of variables has a lesser disparity between Black and White borrowers, but it has significantly worse predictive accuracy as measured with AUC (0.86 vs 0.77).

  Given this, the author argues that restricting credit models to traditional credit pricing inputs like income and credit scores risks "entrenching disadvantage"[257] in underserved groups because lenders using these limited models struggle to price lending risk accurately, and therefore they are likely to increase the price of credit and decrease total credit extended. However, Gillis hypothesizes that using additional data, such as payment and education data, may mitigate harm from these biased inputs, especially for those in underserved groups.

Implications and Future Research. The author theorizes that an outcomes-focused testing method might prove a better response to concerns about algorithmic discrimination. Such a regime could answer two key questions: whether a model treats borrowers who are similar yet have different protected characteristics the same, and whether the model and pricing rule increase or decrease disparities compared to some baseline.

C.3 *The Time is Now: Advancing Fairness in Lending Through Machine Learning*

Authors: Vitaly Meursault, Daniel Moulton, Larry Santucci, and Nathan Schor

Publishing information: Federal Reserve Bank of Philadelphia Working Paper 22-39, updated June 2023

The authors explore whether the use of machine learning underwriting models and adopting group-specific fairness constraints can improve both the fairness and profitability of credit decisions. Group-specific fairness constraints set different fairness targets for different groups of applicants. In particular, they apply thresholds that treat census tracts that are considered low or moderate income (LMI) for Community Reinvestment Act (CRA) purposes[258] as underserved.[259] Although there is a tradeoff between lenders' profits and fairness in using group-specific thresholds, the authors suggest that adopting more sophisticated machine learning underwriting models while applying group-specific thresholds for underserved communities can actually increase both fairness and profitability overall.

Methodology. This study compares the effects of applying group-specific fairness constraints on two kinds of models that predict credit scores: a logistic regression model and an eXtreme Gradient Boosting (XGBoost) model. Neither model was trained on data containing third-party scores, and the former used binned features to approximate pre-machine learning industry practice. The models were trained on eight quarters of credit report data in order to predict whether a consumer would default in the next two years, with a credit score of zero indicating complete certainty that the consumer will default and a credit score of 100 indicating complete certainty that the consumer will not default. These scores were then used to make out-of-sample predictions on another eight quarters of data from a random sample of 1% of consumers in the Federal Reserve Bank of New York Consumer Credit Panel/Equifax data (CCP) from Q1 2000 to Q4 2019.[260] The logit model was 0.007 less accurate than the XGBoost model in terms of ROC AUC overall and 0.009 less accurate for LMI tracts.[261]

The authors created lending thresholds by normalizing the credit scores from the models into percentiles decreasing by probability of default and simulating the effects if a hypothetical lender chose to lend to consumers with scores above various potential cutoffs. The authors find that when all applicants are subject to a single profit-maximizing threshold, creditworthy LMI tract applicants are about six percentage points less likely to be correctly classified as creditworthy than non-LMI tract applicants by the XGBoost model. A similar difference in predictive power across population groups has also been documented by Fuster et al. (2022) and Blattner and Nelson (2021).[262]

To investigate the effects of group specific thresholds, the authors added a "strong" fairness constraint. This constraint equalizes true positive rates (TPR)—the percentage of truly creditworthy consumers correctly labeled as creditworthy by the model—among LMI and non-LMI tracts by setting different minimum score thresholds for each group. They also applied "medium" and "weak" fairness constraints that instead reduce disparities in TPR between LMI and non-LMI tract populations by 66% and 33%, respectively.[263]

The paper focuses its reporting of results on fairness-profitability tradeoffs to better situate the findings in the context of lenders' incentives. They assume that each false positive—approvals of applicants who are not in fact creditworthy—costs the lender four times what it will earn with each performing loan.[264]

Key Findings. The authors find that improvements in modeling technology improve overall default prediction, but the improvements in accuracy affect different groups differently, which is a common result both in machine learning and in credit scoring more broadly. Both the logistic regression and XGBoost models in the evaluation perform better for non-LMI applicants, despite the fact that geography and applicants' race, gender, and other similar characteristics are not considered by the model. For example, although geography and proxies for protected class information are absent in the models, a creditworthy LMI tract consumer was 6 percentage points less likely to be classified as creditworthy by the XGBoost model than a creditworthy non-LMI tract consumer.[265]

The authors find that using group-specific thresholds to tailor default predictions for LMI populations can reduce the gap in true positive rates between LMI and non-LMI borrowers, although the false positive rate would increase for LMI populations. A strong group-specific fairness constraint—eliminating entirely the difference in true positive rates for LMI and non-LMI populations—imposes the largest cost in the form of false positive increases.

The authors also conduct a pricing analysis and find that, because the XGBoost model predicts a lower probability of default than the logistic model for a larger proportion of individuals in the non-LMI group (68 percent) than the LMI group (61 percent), the non-LMI tract borrowers are expected to benefit more in terms of loan pricing from ML model adoption. However, the XGBoost model

would slightly reduce prices overall for both groups—a finding that contrasts with findings from Fuster et al. (2022) that ML adoption can slightly increase mortgage prices for minorities.[266] Fairness constraints also only have a minor effect on pricing in the study.

The authors find a potential win-win scenario by considering what happens when lenders shift from logistic regression models to XGBoost models and apply group-specific thresholds. A lender applying a group-specific threshold in the absence of a significant improvement in model quality would reduce its profitability. However, a lender using more sophisticated models with group-specific thresholds can increase both the fairness and profitability of credit decisions. For example, an XGBoost model used with a strong fairness constraint generates more profit for the lender than does a logistic model without fairness constraints—that is, the increase in profitability from using machine learning models more than makes up for any additional losses associated with equalizing true positive for LMI and non-LMI borrowers.[267]

Implications and Future Research. The authors contend that machine learning when used with group-specific fairness constraints can simultaneously improve the fairness and profitability of credit decisions. Notwithstanding legal doubt about the use of group-specific thresholds under the ECOA,[268] the "special purpose credit programs" sanctioned by that statute may be an attractive vehicle for implementing and learning more about this approach.

Endnotes

1 For simplicity, this paper uses "underwriting" as a broad term that generally includes predictive models and processes used to help make decisions about loan approvals, pricing, and credit limits.

2 Large language and image recognition models are leveraging "big data" across the Internet to enable content generation, raising substantial questions about accuracy, reliability, bias, intellectual property, and other topics. For further discussion of the differences between these and ML underwriting models, see Box 2.1.2.

3 Consumer Financial Protection Bureau, Data Point, Credit Invisibles 4-6, 17 (2015); Mike Hepinstall et al., Financial Inclusion and Access to Credit, Oliver Wyman (2022). See Section 2.2.1 and Section 2.2.2.

4 FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context §§ 1, 2.2, 3 (2021) (hereinafter FinRegLab, Machine Learning Market & Data Science Context).

5 We define transparency as the ability of various stakeholders to access information they need related to a model's design, use, and performance. Some use terms such as interpretability or explainability to express similar concepts. See Section 2.3; FinRegLab, Machine Learning Market & Data Science Context §§ 2-3 & Appendix A.

6 FinRegLab, Machine Learning Market & Data Science Context §§ 3 and 5. We use the term "regulatory" in this report to include supervisory expectations that are established through guidance as well to specific legal requirements that are established by regulation.

7 FinRegLab, Machine Learning Market & Data Science Context; FinRegLab, Laura Blattner, & Jann Spiess, Machine Learning Explainability & Fairness: Insights from Consumer Lending (updated June 2023) (hereinafter FinRegLab et al., Empirical White Paper); FinRegLab, Explainability & Fairness in Machine Learning for Credit Underwriting: Policy and Empirical Findings Overview (2023).

8 Logistic regression is a statistical technique that is frequently used to predict a binary (or categorical) dependent variable based on one or more independent variables (e.g., default/not default), while linear regressions are frequently used to predict numeric (or continuous) dependent variables (e.g., time to repayment or default). Both generally assume linear, monotonic relationships among the relevant variables, as discussed further below. For discussion of these and other techniques, see World Bank Group & International Committee on Credit Reporting, Credit Scoring Approaches Guidelines (2019). For additional background on the historical transition to automated systems and ongoing concerns, see FinRegLab, The Use of Cash-Flow Data for Credit Underwriting: Market Context & Policy Analysis § 2 (2020) (hereinafter FinRegLab, Cash-Flow Market Context & Policy Analysis).

9 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 and O-2 to O-4 (2007), 32-49; 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).

10 Consumer Financial Protection Bureau, Data Point, Credit Invisibles, 4-6 and 17; Hepinstall et al.; see also Oportun, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, 2 (July 1, 2021) (estimating based on internal analyses that 55 million consumers with limited credit histories may be mis-scored); Laura Blattner & Scott Nelson, How Costly Is Noise? Data and Disparities in Consumer Credit (June 2022) (finding that the risk of errors in predicting default risk appears to be higher for consumers with relatively thin credit files).

11 See our description of "overfitting" in Appendix B.

12 Additional information on non-traditional data sources can be found in past FinRegLab reports. See, e.g., FinRegLab, Cash-Flow Market Context & Policy Analysis.

13 Scorecards separately assess risk for different segments of the overall applicant population, such as consumers with little credit history or those with a substantial history of delinquent payments. Stacking or ensemble models involve the use of submodels to generate more complex features that are then used as inputs in a final underwriting model. See, e.g., Leo Breiman, Stacked Regressions, 24 Machine Learning 49 (1996). Some pricing models may also predict other factors than default, such as the likelihood that an applicant will revolve balances or pay off a loan early. FinRegLab, Machine Learning Market & Data Science Context §§ 2.4 and 4.3.2.

14 We use the term "feature" to refer to the variables in a dataset used to predict a target variable (such as default). This term is often used synonymously with input variable or independent variable.

15 For more detailed discussions of applicable regulatory frameworks, see FinRegLab, Cash-Flow Market Context & Policy Analysis, Appendix B and Sections 4-6 below.

16 Other firms use machine learning techniques in a more limited way during feature engineering and selection, where they apply the techniques to large datasets to identify particular features or feature relationships that they then use in logistic regression models. This tends to provide less predictive power but may be easier to manage for other purposes. FinRegLab, Machine Learning Market & Data Science Context, 23-24.

17 Tree-based models use a hierarchical structure of "if-then" nodes to generate predictions of the likelihood of default. For example, an initial node might separate consumers based on whether they had previously filed for bankruptcy, and then subsequent nodes on each branch would further separate the relevant group based on current balances or other criteria. XGBoost methods generate multiple tree-based models, each of which are based on the prediction error of the prior model, and then create a final prediction based on the weighted average of the prior models. This is more complex than a single decision tree but leads to lower prediction error rates and better predictive power. FinRegLab, Machine Learning Market & Data Science Context § 4.1.2.1.1.

18 Neural networks take the initial input features and then generate one or more rounds of "latent features" that help to improve predictions of default risk. This structure can boost the predictiveness of the models compared to other machine learning techniques, in part because it can help to identify relationships that are non-linear and non-monotonic in the data as described in note 19 and accompanying text. FinRegLab, Machine Learning Market & Data Science Context § 4.1.2.1.3.

19 The use of salt in cooking illustrates both concepts. The first increment of salt added to a dish may improve the flavor by a different amount than the second increment, and at some point, additional increments will start to make the dish taste worse. Such a relationship is neither linear nor monotonic.

20 The presence of highly correlated features increases the variance of regression coefficients. This complicates the task of assessing the statistical and economic significance of features' influence on the regression's dependent variable (e.g., default risk). For this reason, some stakeholders choose to use scorecards and other model structures besides regression to deal with correlations.

21 See, e.g., Financial Stability Board, Artificial Intelligence and Machine Learning in Financial Ser-vices (2017); Ting Huang et al., The History of Artificial Intelligence, University of Washington (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); Michael Jordan & Tom Mitchell, Machine Learning: Trends, Perspectives, and Prospects, 349 Science 255-260 (2015).

22 For a discussion of generative AI adoption in the financial services industry, see Melissa Koide, Written Testimony on "Artificial Intelligence in Financial Services" to the Senate Committee on Banking, Housing, and Urban Affairs (2023) § 3.b. The content creation process in generative AI relies on models that predict words or images based on patterns learned in large amounts of sequential data. For instance, auto-fill functions are a low-level version of generative AI that predict the most likely letters or phrases that follow the initial content. See Mark Riedl, A Very Gentle Introduction to Large Language Models without the Hype, Medium (2023).

23 See generally Aylin Caliskan, Detecting and Mitigating Bias in Natural Language Processing, Brookings Institution (2021); Prepare for Truly Useful Large Language Models, Editorial, 7 Na-ture Biomedical Engineering 85-86 (2023); The Politics of AI: ChatGPT and Political Bias, Jeremy Baum & John Villasenor, Brookings Institution (2023); Yogesh K. Dwivedi et al., So what if ChatGPT wrote it? Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy, 71 International Journal of Information Management (2023); Shira Ovide, Your Selfies Are Helping AI Learn. You Did Not Consent to This, Washington Post (Dec. 9, 2022).

24 The Executive Order encourages the Director of the CFPB to evaluate underwriting models for "bias or disparities affecting protected groups" and issue additional guidance on fairness considerations regarding the advertising of credit. United States, Executive Office of the President [Joseph Biden], Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Federal Register 75191 (November 1, 2023).

25 Several studies have found significant predictive accuracy gains in moving from incumbent logistic regression models to machine learning models in lending contexts. See, e.g., Niklas Bussmann et al., Explainable Machine Learning in Credit Risk Management, 57 Computational Economics 203-216 (2021); Eilif de Lange et al., Explainable AI for Credit Assessment of Banks, 15 Journal of Risk and Financial Management 556 (2022); Joao A. Bastos & Sara M. Matos, Explainable Models of Credit Losses, 301-1 European Journal of Operational Research 386-394 (2022); 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); Vitaly Meursault et al., The Time Is Now: Advancing Fairness in Lending Through Machine Learning, Federal Reserve Bank of Philadelphia Working Paper 22-39 (updated June 15, 2023). Interviews with some industry stakeholders support these findings.

26 FinRegLab, Machine Learning Market & Data Science Context § 2.1.

27 On the eve of the pandemic, for instance, median Black and Hispanic households had less than 75 % of the income and 20 % of the assets of median White households. Jessica Semega et al., Income and Poverty in the United States: 2019, U.S. Census Bureau (revised Sept. 2021); Neil Bhutta et al., Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances, FEDS Notes (2020); Ana Hernandez-Kent & Lowell R. Rickets, Wealth Gaps Between White, Black and Hispanic Families in 2019, On the Economy Blog, Federal Reserve Bank of St. Louis (2021).

28 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); Patrick Bayer et al., What Drives Racial and Ethnic Differences in High-Cost Mortgages? The Role of High-Risk Lenders, 31 Review of Financial Studies 175 (2018); Jacob S. Rugh et al., Race, Space, and Cumulative Disadvantage: A Case Study of the Subprime Lending Collapse, 62 Social Problems 186-218 (2015); Prosperity Now, Forced to Walk a Dangerous Line: The Causes and Consequences of Debt in Black Communities (2018); Jim Hawkins & Tiffany C. Penner, Advertising Injustices: Marketing Race and Credit in America, 70 Emory L.J. 1619 (2021).

29 Howell Jackson & Timothy Massad, The Treasury Option: How the US Can Achieve the Financial Inclusion Benefits of a CBDC Now, Brookings Institution (2022); Federal Deposit Insurance Corporation, 2021 National Survey of Unbanked and Underbanked Households (2023), 81; Ying Lei Toh, Promoting Payment Inclusion in the United States, Federal Reserve Bank of Kansas City (2022); Michael A. Stegman, Savings for the Poor: The Hidden Benefits of Electronic Banking (1999).

30 FinRegLab, Cash-Flow Market Context & Policy Analysis; FinRegLab, The Use of Cash-Flow Data in Credit Underwriting: Empirical Research Findings (2019).

31 VantageScore, VantageScore 4.0 Fact Sheet, https://www.vantagescore.com/lenders/why-vantagescore/our-models/ (October 30, 2023).

32 Upstart, Upstart by the Numbers, Blog (Sept. 30, 2023) (reporting the results of an internal study indicating that the company was able to offer 103% more loans to consumers with an income less than $50,000 than an industry benchmark); Upstart, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning; Oportun, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine

Learning, 2-3 (estimating that it has assisted more than 900,000 consumers who lacked FICO scores begin to build credit history since its inception and reporting that it has developed machine learning models based on alternative data, credit bureau records, and proprietary historical data that can score 100% of applicants).

33 Section 6 contains a broader discussion of this topic. See also Florian Ostmann & Cosmina Dorobantu, AI in Financial Services, The Alan Turing Institute 37 (2021).

34 BLDS, LLC et al., Machine Learning: Considerations for Fairly and Transparently Expanding Access to Credit (2020), 6 and 22.

35 See FinRegLab, Machine Learning Market & Data Science Context § 5.2.

36 See FinRegLab, Machine Learning Market & Data Science Context § 2.4. Community banks sometimes rely on third party software/service providers for ML credit models, as they generally lack personnel with the expertise required to develop and deploy these models independently. Independent Community Bankers of America, Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (July 1, 2021), 5-6.

37 See Andreas Fuster et al., Predictably Unequal? The Effects of Machine Learning on Credit Markets, 77 J. of Finance 5 (2022). For more on the use of risk-based pricing in the context of machine learning underwriting models, see FinRegLab, Machine Learning Market & Data Science Context, Box 2.1.1.

38 Racial disparities in credit are far greater than for transaction accounts. For example, a 2017 Federal Deposit Insurance Corporation survey found that about 10% of Black and Hispanic households lacked bank and/or prepaid accounts (compared to the 4% national average), 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 (compared to the 20% national average). FinRegLab, Market Context & Policy Analysis § 2.2; Federal Deposit Insurance Corporation, 2017 National Survey of Unbanked and Underbanked Households (2018). Subsequent surveys have not repeated the question about mainstream credit. Federal Deposit Insurance Corporation, 2021 National Survey of Unbanked and Underbanked Households (2022).

39 For more information, see FinRegLab, Cash-Flow Market Context & Policy Analysis § 4.2.1.1.

40 FinRegLab, The Use of Cash-Flow Data in Underwriting Credit: Empirical Research Findings. A 2010 study found that a machine learning model constructed using both credit bureau and transaction data from a large consumer bank would have resulted in a reduction of losses by between 6% and 25% through adjusting credit lines based on the new model's predictions. See Amir E. Khandani et al., Consumer Credit-Risk Models Via Machine-Learning Algorithms, 34 Journal of Banking & Finance 2767-2787 (2010).

41 Agarwal et al; Asli Demirgüç-Kunt et al., The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution, World Bank Group (2018).

42 The use of larger, more varied 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).

43 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 internet 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. See Agarwal et al.; see also Berg et al.

44 Payment of rent, utility, and telecommunications bills is not typically reflected in traditional credit reports, except perhaps where consumers are significantly delinquent. The data could be especially valuable for first-time mortgage applicants and consumers with little traditional credit history. Kelly Thompson Cochran et al., Utility, Telecommunications, and Rental Data in Underwriting Credit, Urban Institute & FinRegLab (updated March 2022).

45 Sumit Agarwal et al., Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech, Indian School of Business (2021); Tobias Berg et al., On the Rise of the FinTechs: Credit Scoring Using Digital Footprints, 33 Rev. of Fin. Studies 2845–2897 (2020); Daniel Bjorkegren & Darrell Grissen, Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment, 34 The World Bank Economic Review 618-634 (2020); FinRegLab, Cash-Flow Market Context & Policy Analysis; Cochran et al.

46 BLDS, LLC et al., 6.

47 See Section 2.3; FinRegLab, Machine Learning Market & Data Science Context §§ 2-3 and Appendix A.

48 As discussed further in Section 6, additional fairness testing is typically performed by a separate compliance team that has access to data about protected class status.

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

50 Jonathan Johnson, Blog, Interpretability vs. Explainability: The Black Box of Machine Learning, BMC (July 16, 2020); Leilani Gilpin et al., Explaining Explanations: An Overview of Interpretability of Machine Learning, arXiv:1806.00069v3 (Feb. 3, 2019).

51 Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Machine Intelligence 206-215 (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., Unwrapping the Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification (2020).

52 Agus Sudjianto & Aijun Zhang, Designing Inherently Interpretable Machine Learning Models, ACM ICAIF 2021 Workshop on Explainable AI in Finance (Nov. 3, 2021); Alejandro Barredo Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, 58 Information Fusion 82 (2020); 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).

53 E.g., Branka Hadji Misheva et al., Explainable AI in Credit Risk Management, arXiv:2103.00949 (Mar. 1, 2021) finds SHAP and LIME can produce explanations for complex ML credit models that are consistent and "in line with financial logic." See also Ian Hardy, Robust Explainability in AI Models, Zest White Paper (2020). Complex machine learning models used in various fields have also 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 J. 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); Alonso & Carbó.

54 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).

55 For more detailed discussions of these and other techniques, see FinRegLab, Machine Learning Market & Data Science Context § 3. For other comparisons of the two techniques in the credit underwriting context, see Alex Gramegna & Paolo Giudici, SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk, Frontiers in Artificial Intelligence (2021); Misheva et al., Explainable AI in Credit Risk Management (2021).

56 See, e.g., Alexey Miroshnikov et al., Mutual Information-Based Group Explainers with Coalition Structure for Machine Learning Model Explanations, arXiv:2102.10878 (updated Sept. 28, 2022); Muhammad Faaiz Taufiq et al., Manifold Restricted Interventional Shapley Values, arXiv:2301.04041 (updated Feb. 25, 2023); Scott Zoldi, Responsible AI in Credit Risk: FICO Insights at Edinburgh Conference 2023 (Aug. 29, 2023); Kjersti Aas, Martin Jullum, & Anders Loland, Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values, 298 Artificial Intelligence 103502 (2021).

57 FinRegLab, Machine Learning Market & Data Science Context § 3.4.2.2.

58 Many older SHAP implementations assume features are not correlated, but some newer variations are being developed to account for correlations. See note 56.

59 Feature interactions can be quantified using Shapley values, but few practitioners have adopted this approach due to its computational intensity. See Scott M. Lundberg et al., From Local Explanations to Global Understanding with Explainable AI for Trees, 2 Nature Machine Intelligence 56-67 (2020); Katsushige Fujimoto et al., Axiomatic Characterizations of Probabilistic and Cardinal-probabilistic Interaction Indices, 55-1 Games and Economic Behavior 72-99 (2006). Some stakeholders use model architecture limitations to make latent features easier to explain and manage. See, e.g., Scott Zoldi, Building Responsible AI for Credible Machine Learning, Medium (Mar. 6, 2023) and Sudjianto & Zhang.

60 See The Royal Society, Explainable AI: The Basics (2019).

61 See *id.*

62 See Hugh Chen et al., True to the Model or True to the Data?, arXiv:2006.16234 (June 29, 2020).

63 In this context, parsimonious is used to describe a model that achieves a desired level of predictiveness using as few variables as possible.

64 For example, surrogate models are also sometimes used in the credit context for disparate treatment analysis or for purposes of performing sensitivity analyses and are deployed in other sectors for impact audits. See, e.g., Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model: Second Report of the Independent Monitor (2021) (hereinafter Relman Colfax, Upstart Second Report); Mohsen Zaker Esteghamati & Madeline M. Flint, Developing Data-Driven Surrogate Models for Holistic Performance-Based Assessment of Mid-Rise RC Frame Buildings at Early Design, Engineering Structures 245 (2021).

65 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 (2020). A data point in this context refers to the feature values for a single applicant.

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

67 European Parliament, Artificial Intelligence Act (2023) (adopting a negotiating position ahead of final talks with member nations); European Commission, Proposal for a Regulation Laying Down Harmonized Rules on Artificial Intelligence (2021); European Commission, Building Trust in Human Centric Artificial Intelligence (2019).

68 National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework, AI RMF 1.0 (2023); Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence (2019); Ostmann & Dorobantu; see also Brian Stanton & Theodore Jensen, Trust and Artificial Intelligence, National Institute of Standards and Technology (Dec. 2020) (discussing nine qualities: accuracy, reliability, resiliency, objectivity, security, explainability, safety, accountability, and privacy).

69 See Appendix A for additional information about the makeup of the advisory board and policy working groups.

70 FinRegLab et al., Empirical White Paper.

71 FinRegLab, Machine Learning Market & Data Science Context; FinRegLab, Explainability & Fairness in Machine Learning for Credit Underwriting: Policy and Empirical Findings Overview (2023).

72 See, e.g., Golnoosh Babaei et al., Explainable FinTech Lending, 125-126 Journal of Economics and Business 106126 (2023); Bussmann et al.; Eilif de Lange et al.; Bastos & Matos; Misheva et al., Explainable AI in Credit Risk Management (2021).

73 The models built by the research team were a logistic regression model and a simple neural network trained on about 45 features and an XGBoost model and neural network model trained on about 650 features. Our data provider required the masking of certain feature descriptions, which limited the companies' ability to conduct qualitative feature reviews or create features manually in the course of model development.

74 We applied additional tests for some topics. For adverse action purposes, we also used a "nearest neighbor" test to identify applicants who were similarly situated to each other with regard to the features identified as important and analyze whether their default predictions were similar. For fair lending, we used oversampling to create a data set in which distributions of "important" features were equalized to assess whether disparity levels dropped substantially.

75 We also examined the extent to which the same tool identified the same features as important for a particular regulatory purpose across different underwriting models.

76 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 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). MRM expectations for credit unions are generally only focused on interest rate risk. National Credit Union Administration, Interest Rate Risk Measurement Systems, Model Risk (2016).

77 See, e.g., FRB, SR 11-7.

78 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.* 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.* 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").

79 Richard R. Pace, Model Risk Management in the Age of AI: A Primer for Risk Managers, Pace Analytics Consulting LLC (2021), 10. Conceptual soundness analysis may also include scenario testing that can reveal potential limitations of the proposed model that historical backtesting alone will not surface. For example, analysis will often be done to evaluate how an underwriting model's default rate predictions respond to an environment of elevated and increasing interest rates. In reviewing these effects, financial institutions can also assess whether the direction and magnitude of the predicted changes from an interest rate increase make sense based on historical experience or business expectations as well as the relevant theoretical frameworks used to develop the model.

80 Metrics such as accuracy, precision, recall, F1 score, Area Under the Receiver Characteristic Curve (AUC-ROC) are commonly used for classification problems, or Mean Squared Error (MSE) for regression problems, among others.

81 Office of the Comptroller of the Currency, Comptroller's Handbook, Model Risk Management: Version 1.0 (2021), 40.

82 See, e.g., National Institute of Standards and Technology, NIST AI Risk Management Framework Playbook (2023); U.S. Chamber of Commerce Technology Engagement Center, Comment on Artificial Intelligence Risk Management Framework Request for Information to NIST (Sept. 15, 2021).

83 Office of the Comptroller of the Currency et al., Request for Information on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, 86 Fed. Reg. 16,837 (Mar. 31, 2021).

84 See, e.g., FRB, SR 11-7 at 3 ("The use of models invariably presents model risk, which is the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports. Model risk can lead to financial loss, poor business and strategic decision making, or damage to a bank's reputation.").

85 Congress adopted fair lending laws in the 1960s and 1970s and some privacy and information security requirements for consumer financial data in 1999. For a detailed discussion of these regimes and their administration, see Financial Health Network, Flourish Ventures, FinRegLab & Mitchell Sandler, Consumer Financial Data: Legal and Regulatory Landscape §§ 3 and 6 (2020). Guidance issued by the Consumer Financial Protection Bureau in 2022 indicating that discrimination in the provision of non-credit consumer financial products and services may constitute an unfair act or practice is being challenged in federal court. Consumer Financial Protection Bureau, CFPB Targets Unfair Discrimination in Consumer Finance, Press Release (March 16, 2022); Chamber of Commerce vs. Consumer Financial Protection Bureau, No. 6:22-cv-00381, Opinion and Order (E.D. Tex. Sept. 8, 2023).

86 A number of private sources such as consultant organizations and vendors have suggested lists of principles for financial services. See, e.g., Mohan Jayaraman et al., Responsible by Design: Five Principles for Generative AI in Financial Services, Bain & Co. (July 21, 2023); Usama Fayyad, Responsible AI: A Mandate in Finance and Insurance, Forbes Technology Council (July 6, 2023); Daragh Morrissey & Nick Lewins, Microsoft's Perspective on Responsible AI in Financial Services (2019).

87 See, e.g., FICO & Corinium, State of Responsible AI in Financial Services (2023); MIT Technology Review Insights & JPMorgan Chase & Co., Deploying a Multidisciplinary Strategy with Embedded Responsible AI (Feb. 14, 2023); Daragh Morrissey & Nick Lewins, Responsible AI in Financial Services: Governance & Risk Management (2019); see also Anand Rao & Bret Greenstein, 2022 PwC AI Business Survey (reporting results of broader survey on adoption of responsible AI components across multiple sectors); Elizabeth M. Renieris et al., To Be a Responsible AI Leader, Be Responsible, MIT Sloan Management Review & BCG (Sept. 19, 2022) (same).

88 See, e.g., Scott Zoldi, AI Governance: How Blockchain Can Build Accountability and Trust, EnterpriseAI News (Dec. 1, 2022).

89 Intuitive justifications are often based on firm experience and broader economic or behavioral theories. Pace, Model Risk Management in the Age of AI, 9-10.

90 *Id.* For a contrasting perspective see Leo Breiman, Statistical Modeling: The Two Cultures, 16 Statistical Science 199-231 (2001).

91 A modeling target that is too narrowly defined will not contain enough instances of default or delinquency to be able to reliably predict the outcome. A performance window that is too long—for example, 36 or 48 months—may cause the model development data to be too old to be representative of the environment in which a new model is intended to operate, although what is appropriate can vary based on the product being underwritten and other factors. In underwriting models, developers often use reject inference modeling to infer outcomes for applicants in training data sets who were denied loans or who declined offers of credit. See FinRegLab, Machine Learning Market & Data Science Context § 4.3.1.

92 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. In the context of logistic regression models, there are analogous processes to hyperparameter reviews, and these have taken on increased prominence in the context of modern regularization methods. For example, the type of regularization used to limit the number of covariates is one hyperparameter of the logistic regression method, and defining the convergence criteria for regularization is another. In practice, statisticians may refer to these simply as choices made during the model development process that must be documented and justified rather than calling them hyperparameters.

93 For additional background see Section 2.1 and FinRegLab, Machine Learning Market & Data Science Context § 3.4.2.2.

94 Lenders may use separate models to generate adverse action notices for credit decisions that reflect input from various traditional models, scores, and analyses, but these adverse action engines did not necessarily use newer explainability techniques.

95 The guidance generally defines "model" to include quantitative methods, systems, or approaches that apply statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates. The definition also covers situations involving inputs that are partially or wholly qualitative or based on expert judgment, provided that the output is quantitative in nature. FRB, SR 11-7 at 3.

96 Among several types of diagnostic tools applied to logistic regression, XGBoost, and a neural net model, the tools that used SHAP applications were more likely to beat the random and correlated benchmarks than other tool types, although a few SHAP applications did not. Overall, our results suggest that different SHAP implementations in conjunction with different sampling methods could lead to variations in responses, so it is important to evaluate a specific implementation before deploying a new SHAP-based tool for model risk management.

97 See, e.g., Pace, Model Risk Management in the Age of AI, 11-12; Sudjianto & Zhang; see also Zoldi, Building Responsible AI for Credible Machine Learning.

98 See, e.g., Zest AI, Here's How ML Underwriting Fits Within Federal Model Risk Management Guidelines (2019); See also Bastos & Matos.

99 Andrew G. Haldane, Speech, Will Big Data Keep its Promise?, Bank of England (April 30, 2018). See also Breiman, Statistical Modeling.

100 FRB, SR 11-7, 3 and 6-8.

101 See, e.g.: Bank Policy Institute., Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, 5-8; Wells Fargo & Co., Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (July 1, 2021), 2.

102 For a more detailed discussion of ML underwriting adoption across banks and nonbank lenders, see FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context § 2.4.1.

103 See, e.g., Bernd Carsten Stahl et al., A Systematic Review of Artificial Intelligence Impact Assessments, 56 Artificial Intelligence Review 12799 (2023).

104 European Parliament, Artificial Intelligence Act, art. 29(a) (2023).

105 S. 3572, 117th Cong. (2022); H.R. 6580, 117th Cong. (2022).

106 Majority Leader Schumer Delivers Remarks to Launch SAFE Innovation Framework for Artificial Intelligence at CSIS (June 21, 2023).

107 Nonbanks that are examined by the CFPB are also subject to third party service provider expectations with regard to compliance with certain consumer protection laws. See FinRegLab, Machine Learning Market & Data Science Context § 2.3.

108 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).

109 See generally Board of Governors of the Federal Reserve System et al., Conducting Due Diligence on Financial Technology Companies: A Guide for Community Banks (2021).

110 As discussed in note 76, model risk guidance is more narrowly focused for credit unions than for small banks. Although large banks are less likely to rely on vendor-provided models and tools for underwriting specifically, they may also encounter vendor-management challenges for other types of ML applications, such as fraud and use of automated valuation models.

111 For more context, see FinRegLab, Machine Learning Market & Data Science Context § 3.5.

112 See Pace, Model Risk Management in the Age of AI, 35.

113 See, e.g., Office of the Comptroller of the Currency, Comptroller's Handbook, Model Risk Management, 29.

114 The use of open-source tools as the basis of many proprietary offerings or on their own can help address verifiability concerns but may introduce other broader concerns and risks. See, e.g., Alex Engler, How Open-Source Software Shapes AI Policy, The Brookings Institution (2021); Frank Nagle, Strengthening Digital Infrastructure: A Policy Agenda for Free and Open-Source Software, The Brookings Institution (2022).

115 See Pace, Model Risk Management in the Age of AI, 35-36.

116 See generally Board of Governors of the Federal Reserve System et al., Conducting Due Diligence on Financial Technology Companies: A Guide for Community Banks (2021).

117 Prudential regulators have authority to conduct such examinations under the Bank Service Company Act. The CFPB has authority to examine service providers to a substantial number of banks with less than $10 billion in assets. 12 U.S.C. § 5516(e). Service provider examinations are limited to the scope of activities performed on behalf of supervised entities.

118 The FDIC sought comment on the use of a standard setting organization for this purpose in 2020. 85 Fed. Reg. 44890 (2020). The CFPB is proposing to rely on a standard setting organization to implement data sharing protocols. 88 Fed. Reg. 74796 (Oct. 31, 2023).

119 Board of Governors of the Federal Reserve System et al., Interagency Guidance on Third-Party Relationships: Risk Management, 88 Fed. Reg. 37,920 (June 9, 2023).

120 ECOA's adverse action notice requirements apply to both business and consumer lending; although mandatory disclosures for business lending are less extensive, and lenders have more flexibility in providing notices orally to businesses. 12 CFR § 1002.9. FCRA's disclosure requirements have been interpreted to apply only to consumer credit transactions. 12 CFR § 222.70(a).

121 15 U.S.C. § 1691(d)(6); 12 CFR § 1002.9.

122 15 U.S.C. § 1681m(a)(1), (b).

123 15 U.S.C. § 1681m(h); 12 CFR § 1022.72.

124 See, e.g., S. Rep. 94-589, 94th Cong., 2d Sess., at 4, reprinted in 1976 U.S.S.C.A.N. 403, 406; 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).

125 An example of the latter might be a disclosure that indicates the applicant was rejected because the loan-to-value ratio was too high when in fact the appraised value far exceeds the size of the requested loan.

126 Consumer Financial Protection Bureau, Tech Sprint on Electronic Disclosures of Adverse Action Notices (2020).

127 See FinRegLab, Cash-Flow Market Context & Policy Analysis, Box 4.1.2.1.

128 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, para. 9(b)(2)-9.

129 15 U.S.C. § 1691(d); 12 CFR § 1002.9(b)(2).

130 12 C.F.R. § 1002.9.

131 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, para. 9(b)(2)-3.

132 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, para. 9(b)(2)-4.

133 Official interpretation of 12 CFR § 1002.9(b)(2), paragraph 5.

134 Patrice Alexander Ficklin, Tom Pahl, & Paul Watkins, Blog, Innovation spotlight: Providing Adverse Action Notices When Using AI/ML Models, Consumer Financial Protection Bureau (2020); Richard Pace, Blog, Using Explainable AI to Produce ECOA Adverse Action Reasons: What Are the Risks?, Pace Analytics Consulting LLC (Sept. 15, 2022).

135 15 U.S.C. §§ 1681g(f)(1), (9); 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, para. 9(b)(2)-1. Where the number of inquiries is a key factor, the adverse action notice may state up to five principal bases of the credit decision. See 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, para. 9(b)(2)-9.

136 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, para. 9(b)(2)-8.

137 Consumer Financial Protection Bureau, Consumer Financial Protection Circular 2022-03 (2022); Consumer Financial Protection Bureau, Consumer Financial Protection Circular 2023-03 (2023).

138 CFPB, Consumer Financial Protection Circular 2022-03 (2022), n.1.

139 CFPB, Consumer Financial Protection Circular 2023-03. The circular also emphasized that creditors do not comply with the law by simply selecting the closest sample reason from the regulatory appendix if that reason is "nevertheless inaccurate" as to the principal reasons for the adverse action.

140 For more on scorecard methodologies, see note 13 and FinRegLab, Machine Learning Market & Data Science Context § 4.3.2.

141 See Section 2 for more background.

142 Some stakeholders note that while aggregation of model features into adverse action codes provides one opportunity for obfuscation, the input features in a scorecard or logistic regression model are effectively aggregated as well because developers choose among correlated features to decide which ones to include in the model. In this case, the opportunity for obfuscation may simply occur earlier in the development process for more traditional models due to their size and scope limitations.

143 In some cases, firms report that they have compared the distribution of adverse action reason codes from traditional and machine learning models as one way of assessing the performance of the new techniques. However, to the extent that the two types of models are focusing on somewhat different patterns in the underlying data, a different distribution may not necessarily be cause for concern.

144 Some of these tests were unique to the analysis of adverse action notices. For additional discussion of the methodologies and results, see FinRegLab et al., Empirical White Paper § 4.

145 See, e.g., Gramegna & Giudici and Hugues Turbé et al., Evaluation of Post-hoc Interpretability Methods in Time-series Classification, 5 Nature Machine Intelligence 250-260 (2023).

146 This research is summarized in Appendix C.1.

147 See, e.g., Aas et al.; Miroshnikov et al; Taufiq et al.; Scott Zoldi, Responsible AI in Credit Risk; Satyapriya Krishna et al., The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective (2023) (calling for better frameworks for evaluating and comparing the consistency of explanations based on interviews with data scientists across a number of sectors).

148 For a discussion of considerations about different benchmark choices in the ML context, see Pace, Using Explainable AI to Produce ECOA Adverse Action Reasons.

149 See, e.g., Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 California Law Review 671 (2016); National Fair Housing Alliance, Response to Request for Information on the Equal Credit Opportunity Act and Regulation B (Sept. 8, 2020), 42-46.

150 CFPB, Consumer Financial Protection Circular 2023-03. The circular suggested that disclosing a denial based on an applicant's chosen profession as "insufficient projected income" or a credit line reduction based on the type of establishment at which a consumer shops as "disfavored business patronage" were likely to fail the legal standard without additional detail. While it stated that "[s]pecificity is particularly important when creditors utilize complex algorithms," it repeatedly referenced use of data that are not typically found in a credit application or credit file.

151 For instance, consistently low utilization may provide too little information about repayment behavior, whereas consistently high utilization suggests high ongoing payment obligations that may impact the borrower's ability to repay a new obligation, and a decline from high to low utilization may indicate increased capacity.

152 See Section 2.3.2.

153 Notably, it is not clear that if both features were assigned separate Shapley scores, that either or both would necessarily be among the reasons provided on the adverse action disclosure. Consider a hypothetical example: an applicant made one late mortgage payment when the loan balance was $225,000 and three others once the balance was below $200,000. The applicant's large outstanding mortgage balance may affect numerous other features in the model. The Shapley value for that balance would therefore reflect a significant contribution for the feature that considered it in the context of late payments for that product, as well as contributions from every other feature with which it was correlated. By contrast, the single late payment may derive its Shapley value primarily from its relationship with the feature for outstanding mortgage balance and have a lower overall Shapley value. If this holds, the late payment may be dominated by the Shapley values of other features and therefore not be included in the features with highest Shapley values, which ultimately populate the disclosure.

154 A third consideration is lenders' desire to protect proprietary information about their systems.

155 In a model with only 15 features, explaining both may not only be less technologically demanding, but the difference between the two types of information may be far smaller, given the reduced opportunity to generate predictive relationships inside the model. This is likely the case because the individual features in a 15-feature model are more generic, in effect summarizing all of the interrelated features that appear in a model with 150 or 1,500 features.

156 CFPB, Tech Sprint on Electronic Disclosures of Adverse Action Notices.

157 FinRegLab et al., Empirical White Paper § 4.7.

158 For more about factors affecting the efficacy of legal disclosures, see George Loewenstein et al., Disclosure: Psychology Changes Everything, 6 Annual Review of Economics 391-419 (2014).

159 Some lenders have raised concerns that providing more advisory information could create risk of litigation by disclosure recipients who adapt their behavior as advised, reapply for credit, and receive another denial or higher pricing. Lenders would thus likely seek assurances that providing advisory disclosures does not trigger coverage under the Credit Repair Organizations Act and is not unfair, deceptive, or abusive under state and federal law where the lender subsequently evaluates the recipient under revised criteria and policies.

160 For example, a lender using utility or telecom data in assessing repayment risk might be able to use the following reason codes with little or no alteration: "Delinquent past or present credit obligations with others," and "Collection action or judgment." See 12 CFR § 1002, Appendix C, Form C-1.

161 In this way, the "sunshine" effect of the adverse action regime shapes market practice and discourages discrimination, separate from helping to facilitate comparisons of content on disclosures for similarly situated borrowers.

162 See, e.g., U.S. Senate Banking Committee, Review: Use of Educational Data to Make Credit Determinations (2020). More broadly, there is a deeper debate on future forecasting versus collateral coverage that has transformed lending in China. It turns on whether credit underwriting is a pure modeling exercise or should adhere to the traditional approach of considering the applicant's capacity, collateral, and character. See generally Leonardo Gambacorta et al., Data Versus Collateral, 27 Review of Finance, 369-398 (2022).

163 CFPB, Consumer Financial Protection Circular 2023-03.

164 See, e.g., National Institute of Standards and Technology, NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software (2019); Jeffrey Dastin, Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women, Reuters (Oct. 10, 2018); Ziad Obermeyer et al., Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, 366 Science 447-453 (2019).

165 In the data science context, debiasing techniques are used to address many types of bias in models, not limited to fair lending concerns. See Box 6.1.1.1. for further discussion.

166 For purposes of this section, the phrase underwriting model encompasses usage of second-look models. Many of the questions and issues discussed herein apply with full force to those models, although in practice lenders' risk tolerance may differ when making specific risk management decisions about second-chance models because of their perception that they can add value by enabling applicants who would be turned down for credit based on the primary review to receive offers.

167 15 U.S.C. § 1961(a) (prohibiting discrimination on the basis of race, color, national origin, religion, sex, marital status, or age or because of the receipt of public assistance or the good faith exercise of certain rights under federal consumer financial law).

168 42 U.S.C. § 3605 (prohibiting discrimination on the basis of race, color, national origin, religion, sex, familial status or disability).

169 Overt discrimination is sometimes broken out as a third category, distinct from disparate treatment. CFPB, Examination Procedures, ECOA, Baseline Review Modules (2019). Although the Supreme Court has confirmed that these doctrines are all available under the Fair Housing Act and that overt discrimination and disparate treatment are covered by ECOA, it 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); 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.6, para. 6(a)-2. The Supreme Court has recognized the disparate impact theory under the Fair Housing Act but not yet ruled on its status under the Equal Credit Opportunity Act. For a general overview of disparate treatment and disparate impact and the ways that they overlap, see Carol A. Evans, Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks, Consumer Compliance Outlook 1-9 (Second Issue 2017).

170 The first concept is sometimes described in data science contexts as fairness through unawareness, and the second as conditional statistical parity.

171 In data science, this option is often referred to as demographic or statistical parity.

172 In data science, predictive parity and equalized odds are two commonly used measures to evaluate the distribution of different types of predictive errors.

173 See generally Jon Kleinberg et al., Inherent Trade-Offs in the Fair Determination of Risk Scores, Innovations in Theoretical Computer Science Conference (2017); see also Nicholas Schmidt & Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination (Nov. 8, 2019), arXiv:1911.05755. However, new research has shown that fairness along multiple criteria can be achieved if one allows a small margin of error in fairness metrics. See Andrew Bell et al., The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice, arXiv:2302.06347 (2023).

174 FinRegLab, Machine Learning Market & Data Science Context § 5.

175 See, e.g., National Consumer Law Center, Past Imperfect.

176 Nicol Turner Lee & Samantha Lai, The U.S. Can Improve Its AI Governance Strategy by Addressing Online Biases (2022); Barocas & Selbst, 684-687; Talia B. Gillis, The Input Fallacy, 106 Minn. L. Rev. (2022), 1192-1197. However, the latter approach would likely require Congressional action and would raise tension with recent Supreme Court opinions on consideration of demographic information in college admissions. Students for Fair Admissions, Inc. v. President & Fellows of Harvard College, 600 U.S. __ (2023).

177 See FinRegLab, Cash-Flow Market Context & Policy Analysis, Box 4.1.2.1.

178 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. 15 U.S.C. § 1691(b)(3); 12 C.F.R. § 1002.6(b)(2).

179 12 C.F.R. §1002.5. Congress has created two major exceptions to the general rule for residential mortgages under the Home Mortgage Disclosure Act of 1975 and for small business loans under ECOA, although the latter is still being implemented. 12 U.S.C. §§ 2801-2811; 88 Fed. Reg. 35150 (May 31, 2023). In some circumstances, protected class data may be collected and used subject to the requirements of special purpose credit programs. 12 C.F.R. §1002.5(a)(3).

180 Special purpose credit programs address the needs of individuals who would otherwise be declined credit or offered credit on less favorable terms without the program. In this situation, creditors may be permitted to obtain information that would otherwise be prohibited. For example, if financial need is a criterion of a special purpose program targeting low-to-moderate-income households, the creditor could review information concerning the marital status of the applicant, such as alimony payments, child support, and the spouse's income. See 12 C.F.R. § 1002.8.

181 As discussed in note 169, the Supreme Court has recognized the disparate impact theory under the Fair Housing Act but not yet ruled on its status under the Equal Credit Opportunity Act. For a general overview of disparate treatment and disparate impact and the ways that they overlap, see Carol A. Evans, Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks, Consumer Compliance Outlook 1-9 (Second Issue 2017).

182 Consumer Financial Protection Bureau, Supervision and Examination Manual: Equal Credit Opportunity Act Baseline Modules (2019), 2; see also Patrice Alexander Ficklin, Fair Notice on Fair Lending, Consumer Finance Protection Bureau (2012); Department of Housing & Urban Development et al., Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18,266 (Apr. 15, 1994).

183 See, e.g., 12 C.F.R. Pt. 1002, Supp. I (CFPB official commentary to Regulation B, the regulation implementing ECOA); Consumer Financial Protection Bureau, Supervision and Examination Manual: Equal Credit Opportunity Act Baseline Modules, 2; see also Ficklin, Fair Notice on Fair Lending; Federal Deposit Insurance Corporation, Consumer Compliance Examination Manual IV-1.3 (2021); Department of Housing & Urban Development et al., Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18,266. The analysis is derived from employment law. See Griggs v. Duke Power Co., 401 U.S. 424, 439 (1971); Reyes v. Waples Mobile Home Park Ltd. P'ship, 903 F.3d 415, 424 (4th Cir. 2018).

184 See, e.g., Equal Employment Opportunity Coalition et al., Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11,996, 11,999 (1979) (discussing the threshold, but also noting that disparities above it may not be meaningful in small sample sizes and that disparities below it may be "practically significant" in large sample sizes). See also 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.6, para. 6(a)-2; see also Office of the Comptroller of the Currency, Credit Scoring Models: Examination Guidance, Bulletin 1997-24 (May 20, 1997) (focusing on whether credit scoring variables are statistically related to loan performance and have an understandable relationship to creditworthiness).

185 Practice Law Finance, CFPB Clarifies Duty to Perform Fairness Testing on Lending Models, Westlaw Today (Apr. 23, 2023); Brad Blower, CFPB Puts Lenders & FinTechs on Notice: Their Models Must Search for Less Discriminatory Alternatives or Face Fair Lending Non-Compliance Risk, National Community Reinvestment Coalition Blog (Apr. 5, 2023).

186 Even in the mortgage industry, despite HMDA, a significant number of applicants decline to disclose demographics. See, e.g., Jason Richardson, NCRC's HMDA 2018 methodology: How to calculate race and ethnicity (2019).

187 Consumer Financial Protection Bureau, Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity (2014); 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 (Sept. 2008); 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); Blattner & Nelson, How Costly Is Noise?. Specifically, each applicant in the model development data set is assigned a separate probability that he or she is in a particular demographic category. These probabilities can later be used to aggregate summary outcomes. For example, when using a continuous BISG probability methodology, if the applicant is 80% likely to be Black, and 20% likely to be Asian American or Pacific Islander, and that applicant was approved under a proportional method of estimation the approval would count as 0.8 of a Black approval and 0.2 of an AAPI approval. In other instances, thresholds may be used. If the threshold was 70%, in this example it would be counted as a Black approval only.

188 CFPB, Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity; Arthur P. Baines & Marsha J. Courchane, Fair Lending: Implications for the Indirect Auto Finance Market, Charles River Associates (2014); Kasey Matthews, Improving This Algorithm Can Make Lending A Lot Less Racist, Zest AI (2020); Richard Pace, Blog, BISG Proxy Bias Redux: The Impact of the 2020 U.S. Census Data, Pace Analytics Consulting LLC (Aug. 8, 2023).

189 Smaller firms may have difficulty attracting and maintaining equal levels of expertise to their modeling and compliance teams. As a result, model developers and statisticians from business units may provide limited support for certain compliance activities and may have limited access to protected class information in the course of implementing less discriminatory alternatives.

190 See Evans, Keeping Fintech Fair.

191 Firms use varied statistical measures, such as Pearson correlation, to test for correlations when applying this test.

192 See, e.g., Relman Colfax, Upstart Second Report,24-30.

193 For a general overview of statistical testing for disparate impact, see Charles River Associates, What Is Disparate Impact Testing? (2023).

194 Schmidt & Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination (adverse impact ratio is "a common measure of disparate impact"); Patrick Hall et al., A United States Fair Lending Perspective on Machine Learning, Front. Artif. Intell. (2021) ("Other techniques, such as the Adverse Impact Ratio (AIR) and the standardized mean difference (SMD, which is also known as "Cohen's d"), which have a long history of use in employment discrimination analyses, can also be used for measuring disparate impact in lending").

195 In practice, this can be challenging at the model testing phase, since it is not always clear at the model development stage what the approval thresholds will be, and the model is often only one component of the underwriting decision. Further, changes to all of these factors can occur once the model is deployed.

196 Practical significance concepts are also used as a defense in disparate treatment cases where there is statistical significance in the finding that race or gender played a role, but the sample size, slightness of the significance, or other factors arguably indicate that the findings were not the result of (intentional) discrimination. See, e.g., Kevin Tobia, Disparate Statistics, 126-8 The Yale Law Journal 2260-2449 (2017).

197 See, e.g., 44 Fed. Reg. 11996 (1979); see also 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.6, para. 6(a)-2; see also Office of the Comptroller of the Currency, Credit Scoring Models: Examination Guidance, Bulletin 1997-24 (May 20, 1997) (focusing on whether credit scoring variables are statistically related to loan performance and have an understandable relationship to creditworthiness).

198 In effect, these steps also serve to evaluate the strength of the relationship between the features or practices and default risk or other legitimate business needs, which is the second component of disparate impact analysis. Similar processes may be used to search for alternative features to replace inputs that could be viewed as impermissible proxies under disparate treatment if dropping the features substantially reduces model performance.

199 As noted above, manual feature reviews during initial model development are based on general institutional knowledge, since the developers do not have access to protected class information.

200 Some firms, especially smaller ones, only conduct fair lending testing after a model has been put into use.

201 As noted in Box 6.1.1.1, a range of debiasing techniques can be applied at different times, for instance by transforming the input data, building a debiasing function into model training, or transforming a model's output. The machine learning debiasing methods discussed here involve building a debiasing function into model training, which is sometimes called "in-processing." In-processing techniques are the most relevant in the lending context because they (1) may withstand disparate treatment scrutiny in that, unlike post-processing techniques, they use protected class information in model training rather than to transform model predictions directly and (2) can typically manage any potential fairness-accuracy tradeoff in LDA search with more precision and efficiency than pre-processing methods. For additional detail about sources of bias and mitigation approaches, see FinRegLab, Market & Data Science Report.

202 See generally Gilles Louppe, Michael Kagan, & Kyle Cranmer, Learning to Pivot with Adversarial Networks (2017).

203 For more on historical bias and its importance, see Box 2.1.1 and National Consumer Law Center.

204 See Gabbrielle M. Johnson, Proxies Aren't Intentional, They're Intentional (2021), 2-4 (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, The Input Fallacy.

205 The discussion excludes doctrinal fair lending without specific implications for the fair and responsible use of machine learning underwriting models, such as what constitutes a legitimate business need in the second prong of the disparate impact evaluation. Those include efforts to recognize considerations beyond default risk as a legitimate business need and to clarify whether the model needs to be the best predictor of default risk to constitute a legitimate business need.

206 See, e.g., Scott Zoldi, Blog, Fighting Bias: How Interpretable Latent Features Remove Bias in Neural Networks, fico.com (Oct. 27, 2021).

207 See Section 2.3.2 for discussion of surrogate models such as LIME, which create simpler and more explainable models that are designed to approximate the full machine learning model.

208 If the group of features under investigation does not contribute to the performance of either of the separate models, then there is additional basis for concluding that the impact of these features is attributable to correlations with protected class characteristics. See Relman Colfax, Upstart Second Report at 24-30; Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model: Third Report of the Independent Monitor (2022), 30-36 (hereinafter Relman Colfax, Upstart Third Report).

209 Relman Colfax, Upstart Second Report, 25. See also Section 2.3.2.

210 See Appendix C for a more complete summary of the method-ology and findings of this re-search.

211 Gillis, 1220-1230. HMDA information contains actual demo-graphics. Id. 1258-1259.

212 *Id.*, 1232-1235. Figure 6 shows the two distributions of predicted risk. The author notes that using fewer inputs may account for some of the disparity between these distributions. The author does not report performance metrics for these models to help contextualize the importance of these differences.

213 *Id.* 1240-1241.

214 Specifically, the study used zip code tabulation areas that are generated by the U.S. Census to account for situations in which zip codes do not follow the borders of census tracts, block groups, or other jurisdictional boundaries. Id. 1225-1226.

215 *Id.*

216 *Id.* 1223-1224.

217 *Id.* 1237 and 1245-1256.

218 Schmidt & Stephens (adverse impact ratio is "a common measure of disparate impact"); Hall et al., A United States Fair Lending Perspective on Machine Learning ("Other techniques, such as the Adverse Impact Ratio (AIR) and the standardized mean difference (SMD, which is also known as "Cohen's d"), which have a long history of use in employment discrimination analyses, can also be used for measuring disparate impact in lending").

219 See Richard Pace, Fool's Gold? Assessing the Case for Algorithmic De-Biasing (2021), Pace Analytics Consulting LLC (discussing outcome-based fairness metrics in light of MRM and ability-to-repay requirements).

220 See generally FinRegLab, Machine Learning Market & Data Science Context § 5.2.1 and Appendix C; Deborah Hellman, Measuring Algorithmic Fairness, 108 Virginia Law Review 811-866 (2020). In addition to limitations on collecting and using data on protected class, another challenge is assessing model accuracy with regard to applicants who are rejected or simply do not take out a loan. Some information can be purchased by reporting agencies or imputed by various statistical methods, but such options are subject to various constraints.

221 See Upstart, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, 17-19.

222 Examples include area under the receiver operating characteristics curve (AUC), which is a common performance metric for classification problems like default risk estimation, and Kolmogorov-Smirnov (KS), which refers to the separation between the positive and negative distributions on Kolmogorov-Smirnov charts.

223 Depending on data availability, other options might include using underlying default rates to compare the percentage of true positives (i.e., the number of approved applicants who in fact repay their loans) across groups and subgroups, or comparing the ratio of false positives (approved applicants who fail to repay) to false negatives (rejected applicants who could have repaid) within different demographic groups. See Hellman Part III.

224 Pace, Fool's Gold; Hellman, Measuring Algorithmic Fairness. As described in Appendix C.3, one study by researchers at the Federal Reserve Bank of Philadelphia assessed the implications of setting different approval thresholds in different geographies to ensure that a lender accepts the same percentage of truly creditworthy applicants in each area ("true positive rate"), despite evidence that traditional credit report information tends to be more sparse and noisy for residents of low- to moderate-income neighborhoods. The study found that the differentiated thresholds would also tend to increase the approvals of "false positives" who are not likely to repay the loan, but that those losses could potentially be offset by greater predictive accuracy from ML models. See Meursault et al.

225 As noted in Section 6.1.1, some stakeholders are looking beyond questions about debiasing strategies to question whether direct use of protected class information in making credit decisions would actually result in fairer outcomes, particularly in connection with machine learning models. However, such approaches would likely require Congressional action and would raise tension with recent Supreme Court opinions on consideration of demographic information in college admissions. Students for Fair Admissions, Inc. v. President & Fellows of Harvard College, 600 U.S. __ (2023).

226 For further discussion, see FinRegLab et al., Empirical White Paper § 5.7.2.

227 See Gillis, 1222-1230.

228 Areas for research could also include the effect of using imputed protected class information rather than actual protected class information in the model development process.

229 In selecting among alternatives, lenders are still subject to general model risk management requirements for banks (including conceptual soundness analyses) and federal consumer protection regulations that require all credit card and mortgage lenders to assess borrowers' ability to repay. See Pace, Fool's Gold, Point 1; Section 6.2.2.3 regarding the definition of less discriminatory alternatives.

230 Part of this may be due to resource limitations and risk calculations where incumbent models are not changing significantly over time. Lenders are also divided over whether documentation of searches for LDAs will help to defend their models from potential criticisms or whether regulators and plaintiffs' attorneys will draw adverse inferences from records about alternate models that were not ultimately selected. Zest AI, Why ZAML Makes Your ML Platform Better (2019).

231 Practice Law Finance, CFPB Clarifies Duty to Perform Fairness Testing on Lending Models; Brad Blower, CFPB Puts Lenders & FinTechs on Notice.

232 Joint Letter from NCRC, Zest and Upturn calling on CFPB To Encourage Lenders to Look for Less Discriminatory Alternatives (2022); National Fair Housing Alliance, Response to Request for Information on the Equal Credit Opportunity Act and Regulation B (2020), 6-7 (calls on the CFPB to "inform financial institutions that they are expected to conduct a rigorous LDA search as part of a robust compliance management system, and to advance the policy goals of furthering financial inclusion and racial equity"); National Community Reinvestment Coalition, Comment on the CFPB's RFI on the Equal Credit Opportunity Act (2020).

233 A recent application of this approach used the following standards for practical significance: APR disparities were deemed practically significant where SMD was greater than 0.30 (where a higher SMD means greater disparities), and approval/denial disparity were deemed practically significant where AIR less than 0.90 (where a lower AIR means greater disparities). See Relman Colfax, Upstart Third Report at 8.

234 Relman Colfax, Upstart Third Report at 12 (citing Jones v. City of Bos., 752 F.3d 38, 52 (1st Cir. 2014): "[A practical significance standard] may serve important needs in guiding the exercise of agency discretion, or in serving as a helpful rule of thumb for [institutions] not wanting to perform more expansive statistical examinations.")

235 *Id.* at 7-9 and 11-12.

236 Relman Colfax, Upstart Second Report at 14-15.

237 12 C.F.R. 1002.1(b); *see also* Public Law 93-495, tit. V, § 502, 88 Stat. 1500, 1521 (1974); see also 86 Fed. Reg. 56356, 56371 (CFPB Small Business Data Collection Proposed Rule); Consumer Financial Protection Bureau, Supervision and Examination Manual: Equal Credit Opportunity Act Baseline Modules; see also 12 C.F.R. 1002.1(a) ("The purpose of this part is to promote the availability of credit to all creditworthy applicants without regard to 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. The regulation prohibits creditor practices that discriminate on the basis of any of these factors").

238 Federal Agencies, Interagency Fair Lending Examination Procedures (2009), 27 (emphasis added). Guidance on credit scoring systems also contemplates some performance variation in less discriminatory models that might be required for adoption: if "the business necessity can be achieved by substituting a comparably predictive variable that will allow the credit scoring system to continue to be validated, but also operate with a less discriminatory result." See Office of the Comptroller of the Currency, Credit Scoring Models: Examination Guidance, Bulletin 1997-24 (May 20, 1997), Appendix page 11 (emphasis added); see also Federal Housing Finance Agency, AB 2021-04 (2021).

239 Some stakeholders point out that this assumption may not be correct with regard to models that are not only predicting default but also predicting likelihood of prepayment or other outcomes. In such models, changes in performance may fall just as much or more on the group affected because of prepayment risk. See generally Relman Colfax, Upstart Third Report at 26-27.

240 See generally *id.*, 12-23. For example, if a developer sees that a particular iteration results in a performance deterioration of more than a designated percentage when compared to the results of the prior iteration in the same test, firm policy may preclude acceptance of the

model with lower performance and direct the developer to focus instead on the prior iteration. In addition, once a model is submitted for validation and testing against out-of-time data samples, its performance is often described not by a single numerical target, but rather by a target level of performance with some probability of variation.

241 *Id.*, 17.

242 Within a confidence interval, performance at levels that are further away from the original model's "expected" performance are progressively less likely to be observed. Lenders factor in these probabilities in defining what range is acceptable. Accordingly, an alternative model with an "expected" performance at the lower bound of the confidence interval has a much higher probability of actually obtaining the lower level of performance.

243 NCRC, Comment on the CFPB's RFI on the Equal Credit Opportunity Act, question 9; Michael Akinwumi et al., An AI Fair Lending Policy Agenda for the Federal Financial Regulators, Brookings Institution (2021).

244 Akinwumi et al.

245 NCRC, Comment on the CFPB's RFI on the Equal Credit Opportunity Act, Question 9.

246 Relatedly, as most fairness metrics and debiasing techniques consider only one protected class at a time, without evaluating effects on intersectional subgroups (e.g., categories along the axes of both race and gender such as Black women). Techniques to better evaluate and mitigate these disparities are a promising avenue of future research. See Michael Kearns et al., Preventing Fairness Gerrymandering: Auditing and learning for Subgroup Fairness, Proceedings of the 35th International Conference on Machine Learning (2018).

247 One possible exception relates to a scenario in which one group's AIR is above the relevant threshold for both the initial model and the alternative model candidate, but the latter has an AIR closer to the threshold. In this case, reducing the amount by which the group's AIR exceeds the threshold will not generally be seen as a practically significant harm if the alternative model candidate also improves AIR for the second group under consideration. By contrast, if the AIR for both groups is below the relevant threshold for the initial model, the alternative model candidate will likely be rejected if it improves AIR for one group but decreases it for another.

248 See generally Relman Colfax, Upstart Third Report, 10-30.

249 NCRC, Comment on the CFPB's RFI on the Equal Credit Opportunity Act, question 9.

250 The concept of working across a fairness-accuracy frontier is well established across the broader machine learning literature. See, e.g., Kit T. Rodolfa, Hemank Lamba, & Rayid Ghani, Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy, 3 Nature Machine Intelligence 896-904 (2021).

251 Although the study considered text and image processing models, this summary primarily focuses on the study's finding as to models classifying tabular data given their particular relevance to machine learning underwriting models.

252 For example, the authors did not evaluate the effect of aggregating explainability tool outputs into higher-level explanations. See Section 5 for further discussion.

253 These gradient-based methods are Vanilla Gradient, Gradient times Input, Integrated Gradients, and SmoothGrad, and they were only tested on the neural network and logistic regression models.

254 This summary focuses primarily on the paper's empirical component. In addition, the author surveys how machine learning underwriting models can increase and decrease disparities; how certain kinds of data inputs can introduce disparities; and current legal and regulatory requirements.

255 The 40 variables include more types of variables than mortgage originators typically use in traditional lending, though it does not include nontraditional data being considered in some forms of lending. See Section 2.2.3.

256 Gillis assessed area under the receiver operating characteristics curve ("AUC"), which is a common performance metric for classification problems like default risk estimation.

257 Gillis, 1238.

258 In keeping with CRA practice, the authors classify LMI census tracts as those with a median income less than 80% of its metropolitan statistical area or metropolitan division income (MSA/MD). For tracts outside of MSA/MD, the authors use statewide income.

259 While there is uncertainty about whether group-specific thresholds are permitted under fair housing laws, the paper discusses special purpose credit programs as a possible pathway for implementing such thresholds for LMI and non-LMI tracts. Meursault et al., 44-46; note 180. For a discussion of whether other group-specific approaches to addressing accuracy disparities may be legally permissible, see Hellman.

260 The CCP is an anonymized, consumer-level dataset of quarterly credit bureau records for a five percent, nationally representative random sample of individuals with a credit file. The authors also use demographic data from the U.S. Census Bureau to determine the CRA status of consumers in their sample of the CCP dataset based on the income of the consumer's census tract. The sample excluded consumers that are currently at least 90 days delinquent in keeping with industry practice.

261 Meursault et al, 47.

262 Meursault et al., 5.

263 Meursault et al., 24-25.

264 Meursault et al., 20-21.

265 Meursault et al., Figure 3.

266 Pricing is assumed to be a function of default risk in the analysis by Meursault et al. See Meursault et al., 38-39.

267 See Meursault et al., Figures 5-6 and pages 25-27.

268 See Gillis, 1196-1204 and Hellman, Part III.

Bibliography

Kjersti Aas, Martin Jullum, & Anders Loland, Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values, 298 Artificial Intelligence 103502 (2021).

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 (2021).

Michael Akinwumi, Jogn Merrill, Lisa Rice, Kareem Saleh, & Maureen Yap, An AI Fair Lending Policy Agenda for the Federal Financial Regulators, Brookings Institution (2021).

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).

Alejandro Barredo Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, 58 Information Fusion 82 (2020).

Golnoosh Babaei, Paolo Giudici, & Emanuela Raffinetti, Explainable FinTech Lending, 125-126 Journal of Economics and Business 106126 (2023).

Arthur P. Baines & Marsha J. Courchane, Fair Lending: Implications for the Indirect Auto Finance Market, Charles River Associates (2014).

Bank Policy Institute, Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (June 25, 2021).

Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 California Law Review 671 (2016).

Joao A. Bastos & Sara M. Matos, Explainable Models of Credit Losses, 301-1 European Journal of Operational Research 386-394 (2022).

Jeremy Baum & John Villasenor, The Politics of AI: ChatGPT and Political Bias, Brookings Institution (2023).

Patrick Bayer, Fernando Ferreira, & Stephen L. Ross, What Drives Racial and Ethnic Differences in High-Cost Mortgages? The Role of High-Risk Lenders, 31 Review of Financial Studies 175-205 (2018).

Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Herasymova, Lucas Rosenblatt, & Julia Stoyanovich, The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice, arXiv:2302.06347 (2023).

Tobias Berg, Valentin Burg, Ana Gombovic, & Manju Puri, On the Rise of the FinTechs: Credit Scoring Using Digital Footprints, 33 Rev. of Fin. Studies 2845–2897 (2020).

Allen N. Berger & W. Scott Frame, Small Business Credit Scoring and Credit Availability, 47 J. of Small Bus. Mgmt. 5-22 (2007).

Neil Bhutta, Andrew C. Chang, Lisa J. Dettling, & Joanne W. Hsu, Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances, FEDS Notes (2020)

Daniel Bjorkegren & Darrell Grissen, Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment, 34 The World Bank Economic Review 618-634 (2020).

Laura Blattner & Scott Nelson, How Costly Is Noise? Data and Disparities in Consumer Credit (2022).

BLDS, LLC, Discover Financial Services & H2O.ai, Machine Learning: Considerations for Fairly and Transparently Expanding Access to Credit (2020).

Brad Blower, CFPB Puts Lenders & FinTechs on Notice: Their Models Must Search for Less Discriminatory Alternatives or Face Fair Lending Non-Compliance Risk, National Community Reinvestment Coalition Blog (Apr. 5, 2023).

Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, & Office of the Comptroller of the Currency, Conducting Due Diligence on Financial Technology Companies: A Guide for Community Banks (2021).

Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, & the Office of the Comptroller of the Currency, Interagency Guidance on Third-Party Relationships: Risk Management, 88 Fed. Reg. 37,920 (2023).

Board of Governors of the Federal Reserve System, Report to Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit (2007).

Board of Governors of the Federal Reserve System, Supervisory & Regulation Letter 11-7: Guidance on Model Risk Management (2011).

Board of Governors of the Federal Reserve System, Supervisory & Regulation Letter 13-19 (2013).

Leo Breiman, Stacked Regressions, 24 Machine Learning 49 (1996).

Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, & Jochen Papenbrock, Explainable Machine Learning in Credit Risk Management, 57 Computational Economics 203-216 (2021).

Aylin Caliskan, Detecting and Mitigating Bias in Natural Language Processing, Brookings Institution (2021).

Charles River Associates, What Is Disparate Impact Testing? (2023).

Hugh Chen, Joseph D. Janizek, Scott Lundberg, & Su-In Lee, True to the Model or True to the Data?, arXiv:2006.16234 (2020).

Kelly Thompson Cochran, Michael Stegman, & Colin Foos, Utility, Telecommunications, and Rental Data in Underwriting Credit, Urban Institute & FinRegLab (updated March 2022).

Consumer Financial Protection Bureau, CFPB Targets Unfair Discrimination in Consumer Finance, Press Release (March 16, 2022).

Consumer Financial Protection Bureau, Compliance Bulletin and Policy Guidance 2016-02, 81 Fed. Reg. 74410 (2016).

Consumer Financial Protection Bureau, Consumer Financial Protection Circular 2022-03 (2022).

Consumer Financial Protection Bureau, Consumer Financial Protection Circular 2023-03 (2023).

Consumer Financial Protection Bureau, Data Point, Credit Invisibles (2015).

Consumer Financial Protection Bureau, Examination Procedures, ECOA, Baseline Review Modules (2019).

Consumer Financial Protection Bureau, Required Rulemaking on Personal Financial Data Rights, 88 Fed. Reg. 74796 (2023).

Consumer Financial Protection Bureau, Supervision and Examination Manual: Equal Credit Opportunity Act Baseline Modules (2019).

Consumer Financial Protection Bureau, Tech Sprint on Electronic Disclosures of Adverse Action Notices (2020).

Consumer Financial Protection Bureau, Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity (2014).

Jeffrey Dastin, Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women, Reuters (Oct. 10, 2018).

Petter Eilif de Lange, Borger Melsom, Christian Bakke Vennerod, & Sjur Westgaard, Explainable AI for Credit Assessment of Banks, 15 Journal of Risk and Financial Management 556 (2022).

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).

Department of Housing & Urban Development & Federal Agencies, Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18,266 (Apr. 15, 1994).

Rishi J. Desai, Shirley V. 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 (2020).

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

Yogesh K. Dwivedi et al., So what if ChatGPT wrote it? Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy, 71 International Journal of Information Management (2023).

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 (2008).

Alex Engler, How Open-Source Software Shapes AI Policy, The Brookings Institution (2021); Frank Nagle, Strengthening Digital Infrastructure: A Policy Agenda for Free and Open-Source Software, The Brookings Institution (2022).

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. 11,996, 11,999 (1979).

European Commission, Building Trust in Human Centric Artificial Intelligence (2019).

European Commission, Proposal for a Regulation Laying Down Harmonized Rules on Artificial Intelligence (2021).

European Parliament, Artificial Intelligence Act (2023).

Carol A. Evans, Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks, Consumer Compliance Outlook 1-9 (Second Issue 2017).

Muhammad Faaiz Taufiq, Patrick Blobaum, & Lenon Minorics, Manifold Restricted Interventional Shapley Values, arXiv:2301.04041 (updated Feb. 25, 2023)

Usama Fayyad, Responsible AI: A Mandate in Finance and Insurance, Forbes Technology Council (July 6, 2023).

Federal Deposit Insurance Corporation, 2017 National Survey of Unbanked and Underbanked Households (2018).

Federal Deposit Insurance Corporation, 2021 National Survey of Unbanked and Underbanked Households (2023).

Federal Deposit Insurance Corporation, Consumer Compliance Examination Manual IV-1.3 (2021).

Federal Deposit Insurance Corporation, Financial Institution Letter 22-2017: Adoption of Supervisory Guidance on Model Risk Management (Jun. 7, 2017).

Federal Deposit Insurance Corporation, Financial Institution Letter 44-2008 (2008).

Federal Deposit Insurance Corporation, Financial Institution Letter 19-2019 (2019).

Federal Deposit Insurance Corporation, Request for Information on Standard Setting and Voluntary Certification for Models and Third-Party Providers of Technology and Other Services, 85 Fed. Reg. 44890 (2020).

Federal Housing Finance Agency, AB 2021-04 (2021).

Federal Trade Commission, Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues (2016).

Patrice Alexander Ficklin, Fair Notice on Fair Lending, Consumer Finance Protection Bureau (2012).

Patrice Alexander Ficklin, Tom Pahl, & Paul Watkins, Blog, Innovation spotlight: Providing Adverse Action Notices When Using AI/ML Models, Consumer Financial Protection Bureau (2020).

FICO & Corinium, State of Responsible AI in Financial Services (2023).

Financial Health Network, Flourish Ventures, FinRegLab, & Mitchell Sandler, Consumer Financial Data: Legal and Regulatory Landscape (2020).

Financial Stability Board, Artificial Intelligence and Machine Learning in Financial Services (2017).

FinRegLab, Explainability & Fairness in Machine Learning for Credit Underwriting: Policy and Empirical Findings Overview (2023).

FinRegLab, Laura Blattner, & Jann Spiess, Machine Learning Explainability & Fairness: Insights from Consumer Lending (updated June 2023).

FinRegLab, The Use of Cash-Flow Data in Credit Underwriting: Empirical Research Findings (2019).

FinRegLab, The Use of Cash-Flow Data for Credit Underwriting: Market Context & Policy Analysis (2020).

FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context (2021).

Katsushige Fujimoto, Ivan Kojadinovic, & Jean-Luc Marichal, Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices, 55-1 Games and Economic Behavior 72-99 (2006).

Andreas Fuster, Paul S. Goldsmith-Pinkham, Tarun Ramadorai, & Ansgar Walther, Predictably Unequal? The Effects of Machine Learning on Credit Markets, 77 J. of Finance 5 (2022).

Andrew G. Haldane, Will Big Data Keep its Promise?, Speech, Bank of England (April 30, 2018).

Leonardo Gambacorta, Yiping Huang, Zhenhua Li, Han Qiu, & Shu Chen., Data Versus Collateral, 27 Review of Finance, 369-398 (2022).

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).

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 (2020).

Talia B. Gillis, The Input Fallacy, 106 Minn. L. Rev. (2022).

Leilani Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, & Lalana Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, arXiv:1806.00069v3 (2019).

Alex Gramegna & Paolo Giudici, SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk, Frontiers in Artificial Intelligence (2021).

Griggs v. Duke Power Co., 401 U.S. 424, 439 (1971); Reyes v. Waples Mobile Home Park Ltd. P'ship, 903 F.3d 415, 424 (4th Cir. 2018).

H.R. 6580, 117th Cong. (2022).

Patrick Hall, Benjamin Cox, Steven Dickerson, Arjun Ravi Kannan, Raghu Kulkarni, & Nicholas Schmidt, A United States Fair Lending Perspective on Machine Learning, Front. Artif. Intell. (2021).

Patrick Hall, Navdeep Gill, & Nicholas Schmidt, Proposed Guidelines for the Responsible Use of Explainable Machine Learning, arXiv:1906.03533v3 (2019).

David J. Hand, Classifier Technology and the Illusion of Progress, 21 Statistical Science 1-15 (2006).

Ian Hardy, Robust Explainability in AI Models, Zest White Paper (2020).

Jim Hawkins & Tiffany C. Penner, Advertising Injustices: Marketing Race and Credit in America, 70 Emory L.J. 1619 (2021).

Deborah Hellman, Measuring Algorithmic Fairness, 108 Virginia Law Review 811-866 (2020).

Mike Hepinstall, Peter Carroll, Nick Dykstra, & Yigit Ulucay, Financial Inclusion and Access to Credit, Oliver Wyman (2022).

Ana Hernandez-Kent & Lowell R. Rickets, Wealth Gaps Between White, Black and Hispanic Families in 2019, On the Economy Blog, Federal Reserve Bank of St. Louis (2021).

David C. Hsia, Credit Scoring and the Equal Credit Opportunity Act, 30 Hastings L. J. 371 (1978).

Ting Huang, Chris Smith, Brian McGuire, & Gary Yang, The History of Artificial Intelligence, University of Washington (2006).

Independent Community Bankers of America, Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (July 1, 2021).

Howell Jackson & Timothy Massad, The Treasury Option: How the US Can Achieve the Financial Inclusion Benefits of a CBDC Now, Brookings Institution (2022).

Mohan Jayaraman, Philipp Rindler, Velu Sinha, & Marie Teresa Tejada, Responsible by Design: Five Principles for Generative AI in Financial Services, Bain & Co. (July 21, 2023).

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).

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 Blog (July 16, 2020).

Joint Letter from NCRC, Zest and Upturn calling on CFPB To Encourage Lenders to Look for Less Discriminatory Alternatives (2022).

Michael Jordan & Tom Mitchell, Machine Learning: Trends, Perspectives, and Prospects, 349 Science 255-260 (2015).

Michael Kearns, Seth Neel, Aaron Roth, & Zhiwei Steven Wu, Preventing Fairness Gerrymandering: Auditing and learning for Subgroup Fairness, Proceedings of the 35th International Conference on Machine Learning (2018).

Amir E. Khandani, Adlar J. Kim, & Andrew Lo, Consumer Credit-Risk Models Via Machine-Learning Algorithms, 34 Journal of Banking & Finance 2767-2787 (2010).

Jon Kleinberg, Sendhil Mullainathan, & Manish Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, Innovations in Theoretical Computer Science Conference (2017).

Melissa Koide, Written Testimony on "Artificial Intelligence in Financial Services" to the Senate Committee on Banking, Housing, and Urban Affairs (2023).

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, & Himabindu Lakkaraju, The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective (2023).

Nicol Turner Lee & Samantha Lai, The U.S. Can Improve Its AI Governance Strategy by Addressing Online Biases (2022).

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 (2017).

Zachary C. Lipton, The Mythos of Model Interpretability, arXiv:1606.03490v3 (2017).

George Loewenstein, Cass R. Sunstein, & Russell Golman, Disclosure: Psychology Changes Everything, 6 Annual Review of Economics 391-419 (2014).

Gilles Louppe, Michael Kagan, & Kyle Cranmer, Learning to Pivot with Adversarial Networks (2017).

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, & Su-In Lee, From Local Explanations to Global Understanding with Explainable AI for Trees, 2 Nature Machine Intelligence 56-67 (2020).

Kasey Matthews, Improving This Algorithm Can Make Lending A Lot Less Racist, Zest AI (2020).

Majority Leader Schumer Delivers Remarks to Launch SAFE Innovation Framework for Artificial Intelligence at CSIS, CSIS (June 21, 2023).

Vitaly Meursault, Daniel Moulton, Larry Santucci, & Nathan Schor, The Time is Now: Advancing Fairness in Lending Through Machine Learning, Federal Reserve Bank of Philadelphia Working Paper 22-39 (2023).

Alexey Miroshnikov, Konstandinos Kotsiopoulos, Khashayar Filom, & Arjun Ravi Kannan, Mutual Information-Based Group Explainers with Coalition Structure for Machine Learning Model Explanations, arXiv:2102.10878 (updated Sept. 28, 2022).

Branka Hadji Misheva, Joerg Osterrieder, Ali Hirsa, Onkar Kulkarni, & Stephen Fung Lin, Explainable AI in Credit Risk Management, arXiv:2103.00949 (2021).

MIT Technology Review Insights & JPMorgan Chase & Co., Deploying a Multidisciplinary Strategy with Embedded Responsible AI (Feb. 14, 2023).

Tom Mitchell, Machine Learning (1997).

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

Daragh Morrissey & Nick Lewins, Microsoft's Perspective on Responsible AI in Financial Services (2019).

Daragh Morrissey & Nick Lewins, Responsible AI in Financial Services: Governance & Risk Management (2019).

National Community Reinvestment Coalition, Comment on the CFPB's RFI on the Equal Credit Opportunity Act (2020).

National Consumer Law Center, Past Imperfect: How Credit Scores and Other Analytics "Bake In" and Perpetuate Past Discrimination (2016).

National Credit Union Administration, Interest Rate Risk Measurement Systems, Model Risk (2016).

National Fair Housing Alliance, Response to Request for Information on the Equal Credit Opportunity Act and Regulation B (Sept. 8, 2020).

National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework, AI RMF 1.0 (2023).

National Institute of Standards and Technology, NIST AI Risk Management Framework Playbook (2023).

National Institute of Standards and Technology, NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software (2019).

Ziad Obermeyer, Brian Powers, Christine Vogeli, & Sendhil Mullainathan, Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, 366 Science 447-453 (2019).

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).

Office of the Comptroller of the Currency, Bulletin 2013-29 (2013).

Office of the Comptroller of the Currency, Bulletin 2020-10 (2020).

Office of the Comptroller of the Currency, Comptroller's Handbook, Model Risk Management: Version 1.0 (2021).

Office of the Comptroller of the Currency, Credit Scoring Models: Examination Guidance, Bulletin 1997-24 (May 20, 1997).

Office of the Comptroller of the Currency, The Federal Reserve System, the Federal Deposit Insurance Corporation, the Consumer Financial Protection Bureau, & the National Credit Union Administration, Request for Information on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, 86 Fed. Reg. 16,837 (Mar. 31, 2021).

Oportun, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, (July 1, 2021).

Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence (2019).

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

Shira Ovide, Your Selfies Are Helping AI Learn. You Did Not Consent to This, Washington Post (Dec. 9, 2022).

Richard Pace, BISG Proxy Bias Redux: The Impact of the 2020 U.S. Census Data, Blog, Pace Analytics Consulting LLC (2023).

Richard Pace, Fool's Gold? Assessing the Case for Algorithmic De-Biasing (2021), Pace Analytics Consulting LLC.

Richard Pace, Using Explainable AI to Produce ECOA Adverse Action Reasons: What Are the Risks?, Blog, Pace Analytics Consulting LLC (2022).

Richard Pace, Model Risk Management in the Age of AI: A Primer for Risk Managers, Pace Analytics Consulting LLC (2021).

Practice Law Finance, CFPB Clarifies Duty to Perform Fairness Testing on Lending Models, Westlaw Today (Apr. 23, 2023).

Prepare for Truly Useful Large Language Models, Editorial, 7 Nature Biomedical Engineering 85-86 (2023).

Prosperity Now, Forced to Walk a Dangerous Line: The Causes and Consequences of Debt in Black Communities (2018).

Anand Rao & Bret Greenstein, 2022 PwC AI Business Survey (2022).

Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model: Second Report of the Independent Monitor (2021).

Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model: Third Report of the Independent Monitor (2022).

Elizabeth M. Renieris, David Kiron, & Steven Mills, To Be a Responsible AI Leader, Be Responsible, MIT Sloan Management Review & BCG (Sept. 19, 2022).

S. Rep. 94-589, 94th Cong., 2d Sess., at 4, reprinted in 1976 U.S.S.C.A.N.

Lisa Rice & Deidre Swesnik, Discriminatory Effects of Credit Scoring on Communities of Color, 46 Suffolk L. Rev. 936-966 (2013).

Jason Richardson, NCRC's HMDA 2018 methodology: How to calculate race and ethnicity (2019).

Mark Riedl, A Very Gentle Introduction to Large Language Models without the Hype, Medium (2023).

Kit T. Rodolfa, Hemank Lamba, & Rayid Ghani, Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy, 3-10 Nature Machine Intelligence 896-904 (2021).

Ralph J. Rohner, Equal Credit Opportunity Act, 34 Bus. Law. 1423 (1979).

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).

Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Machine Intelligence 206-215 (2019).

Jacob S. Rugh, Len Albright, & Douglas S. Massey, Race, Space, and Cumulative Disadvantage: A Case Study of the Subprime Lending Collapse, 62 Social Problems 186-218 (2015).

S. 3572, 117th Cong. (2022).

Arthur L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, 3 IBM J. of Research & Development 211-229 (1959).

Nicholas Schmidt & Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination (2019).

Jessica Semega, Melissa Kollar, Emily A. Shrider, & John Creamer Income and Poverty in the United States: 2019, U.S. Census Bureau (revised Sept. 2021).

Yan-yan Song & Ying Lu, Decision Tree Methods: Applications for Classification and Prediction, 27 Shanghai Archives of Psychiatry 130-135 (2015).

Bernd Carsten Stahl et al., A Systematic Review of Artificial Intelligence Impact Assessments, 56 Artificial Intelligence Review 12799 (2023).

Brian Stanton & Theodore Jensen, Trust and Artificial Intelligence, National Institute of Standards and Technology (Dec. 2020).

Michael A. Stegman, Savings for the Poor: The Hidden Benefits of Electronic Banking (1999).

Students for Fair Admissions, Inc. v. President & Fellows of Harvard College, 600 U.S. __ (2023).

Agus Sudjianto & Aijun Zhang, Designing Inherently Interpretable Machine Learning Models, ACM ICAIF 2021 Workshop on Explainable AI in Finance (2021).

Agus Sudjianto, William Knauth, Rahul Singh, Zebin Yang, & Aijun Zhang, Unwrapping the Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification (2020).

Winnie F. Taylor, Meeting the Equal Credit Opportunity Act's Specificity Requirement: Judgmental and Statistical Scoring Systems, 29 Buff. L. Rev. 73 (1980).

Texas Department of Housing & Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015).

Kevin Tobia, Disparate Statistics, 126-8 The Yale Law Journal 2260-2449 (2017).

The Royal Society, Explainable AI: The Basics (2019).

Ying Lei Toh, Promoting Payment Inclusion in the United States, Federal Reserve Bank of Kansas City (2022).

Hugues Turbé, Mina Bjelogrlic, Christian Lovis, & Gianmarco Mengaldo, Evaluation of post-hoc interpretability methods in time-series classification, 5 Nature Machine Intelligence 250-260 (2023).

United States, Executive Office of the President [Joseph Biden], Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Federal Register 75191 (November 1, 2023).

Upstart, Upstart by the Numbers, Blog (Aug. 9, 2023).

Upstart, Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, (July 1, 2021).

U.S. Chamber of Commerce Technology Engagement Center, Comment on Artificial Intelligence Risk Management Framework Request for Information to NIST (Sept. 15, 2021).

U.S. Chamber of Commerce vs. Consumer Financial Protection Bureau, No. 6:22-cv-00381, Opinion and Order (E.D. Tex. Sept. 8, 2023).

U.S. Senate Banking Committee, Review: Use of Educational Data to Make Credit Determinations (2020).

VantageScore, VantageScore 4.0 Fact Sheet, https://www.vantagescore.com/lenders/why-vantagescore/our-models/ (accessed November 9, 2023).

Wells Fargo & Co., Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning (July 1, 2021).

World Bank Group & International Committee on Credit Reporting, Credit Scoring Approaches Guidelines (2019).

Mohsen Zaker Esteghamati & Madeline M. Flint, Developing Data-Driven Surrogate Models for Holistic Performance-Based Assessment of Mid-Rise RC Frame Buildings at Early Design, Engineering Structures 245 (2021).

Zest AI, Here's How ML Underwriting Fits Within Federal Model Risk Management Guidelines (2019).

Zest AI, Why ZAML Makes Your ML Platform Better (2019).

Scott Zoldi, AI Governance: How Blockchain Can Build Accountability and Trust, EnterpriseAI News (Dec. 1, 2022).

Scott Zoldi, Building Responsible AI for Credible Machine Learning, Medium (Mar. 6, 2023).

Scott Zoldi, Fighting Bias: How Interpretable Latent Features Remove Bias in Neural Networks, Blog, FICO (2021).

Scott Zoldi, Not All Explainable AI is Created Equal, Retail Banker International (Oct. 9, 2019).

Scott Zoldi, Responsible AI in Credit Risk: FICO Insights at Edinburgh Conference 2023 (Aug. 29, 2023).

Additional Acknowledgments

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

Michael Akinwumi, National Fair Housing Alliance; Brad Blower, Inclusive-Partners LLC; Nick Bourke; Jay Budzik; Marsha Courchane, Charles River Associates; Steve Dickerson; Delicia Hand, Consumer Reports; Stephen Hayes, Ken Scott, and Eric Sublett, Relman Colfax PLLC; Jeremy Hochberg; Irene Meyer, Craig Warrington, and Kristin Williams, Upstart; John Morgan, Capital One; David Moskowitz, Burning Tree Advisors; Anthony Penta; Conrod Robinson; David Silberman; Bryce Stephens; Michael Umlauf and Gene Volcheck, TransUnion; Stephen Van Meter; Scott Zoldi.

We would also like to acknowledge the FinRegLab team who worked on various elements of the research project:

Natalia Bailey, Alex Bloomfield, Kelly Thompson Cochran, Sarah Davies, Colin Foos, Saurab Guatam, Hilary Griggs, Gillous Harris, Tess Johnson, Mashrur Khan, Duncan McElfresh, Kerrigan Molland, P-R Stark, YaYa Sun, Sormeh Yazdi, and Zishun Zhao.

With support from:

Center for Inclusive Growth

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. The Center has provided funding to support this research.

JPMorgan Chase & Co.

JPMorgan Chase is committed to advancing an inclusive economy and racial equity. The firm uses its expertise in business, public policy and philanthropy, as well as its global presence, expertise and resources, to focus on four areas to drive opportunity: careers & skills, financial health and wealth creation, business growth & entrepreneurship, and community development.

flourish

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.

Digital version available at finreglab.org

Published by FinRegLab, Inc.

1701 K Street NW, Suite 1150
Washington, DC 20006
United States