FinRegLab

Framework for Managing Machine Learning Models in Consumer Credit Underwriting

APRIL 2026

About FinRegLab

FinRegLab is a nonprofit, nonpartisan innovation center that tests new technologies and data to increase access to responsible financial services that help drive long-term economic security for people and small businesses. With our research insights, we facilitate discourse across the financial ecosystem to inform market practices and policy solutions.

Acknowledgments

This framework has been developed through discussions with banks that participated in a Technology Working Group convened under the Office of the Comptroller of the Currency's Project REACh (Roundtable for Economic Access and Change). It is intended to foster greater understanding among financial institutions, agency staff, and other stakeholders about how machine learning technologies are being used responsibly to facilitate innovation in credit underwriting.

FinRegLab served as the co-chair of the working group and facilitator of the framework drafting process. FinRegLab's other work relating to machine learning underwriting models and other uses of artificial intelligence and data in financial services is available at our website, www.finreglab.org.

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

Machine learning (ML) and other forms of artificial intelligence (AI) are increasingly transforming industries across the world, including U.S. financial services markets. Although U.S. financial institutions began adopting ML models several decades ago in contexts such as fraud detection, use of ML techniques in credit underwriting has tended to move more slowly particularly among banks due to a mix of regulatory and technical considerations. However, as data science tools for building and managing these models and access to digital information sources continue to improve in quality and cost, research shows the value of this technology is significant[1] and banks are increasingly motivated to use ML models to increase the accuracy of their credit risk predictions and expand their customer bases. A number of banks have begun adjusting their modeling and compliance practices to support adoption of ML models, with or without incorporating new data sources into their underwriting analyses. Nevertheless, adopting these innovations has been made more uncertain and difficult by the fact that familiarity with ML models has varied widely among different lenders, groups of employees, regulatory staff, and other key stakeholders.

This document describes key considerations and approaches for managing machine learning underwriting models. It summarizes the models' potential benefits and describes how industry practices are evolving, while recognizing that data science techniques are continuing to progress and that lenders vary in their strategies and technical implementation. The broader goal of the document is to provide a practical framework outlining risk management principles, processes, and issues that are helpful to consider in adopting machine learning models.

The framework has been developed through discussions by banks that participated in a Technology Working Group convened under the Office of the Comptroller of the Currency's Project REACh (Roundtable for Economic Access and Change).[2] It is intended to foster greater understanding among financial institutions, agency staff, and other stakeholders about how ML technologies are being used responsibly to facilitate innovation in credit underwriting. Institutions who have participated in these dialogues have different modeling methodologies and their own validation and testing frameworks. Their participation in this process and paper does not necessarily equate to endorsing every specific technique and practice discussed in the document.[3]

It is our hope that this framework will help facilitate more efficient and productive conversations among a wide range of stakeholders, including different industry segments, examiners and regulators, researchers, advocates, and other critical constituencies. The contents reflect the experiences of large banks that are building and managing ML underwriting models in house, but may be helpful for other stakeholders to understand how ML models differ from prior approaches and how systems

are shifting in response. While technological tools, business practices, and policy frameworks will continue to evolve, distilling bank practitioners' learnings and practices can help to facilitate financial innovation and greater competition through the responsible implementation of ML techniques in credit risk modeling, even as some elements continue to change over time.

The document starts by providing background about the use of predictive models in credit underwriting, the transition to machine learning models, and relevant laws, regulations, and risk categories applicable to consumer lending. Section 3 briefly discusses initial decisions about model type and architecture, data choice, and talent and resource considerations. Section 4, Section 5, and Section 6 address how banks are managing various compliance processes and considerations for ML models, including adverse action disclosures, model risk management, vendor management, and second look programs. (The current edition of the framework does not discuss fair lending requirements in depth as they are currently being amended by federal regulators.) The appendices provide background information such as brief descriptions of common ML algorithms for use in credit risk models and common explainability and diagnostic tools.

2. BACKGROUND

Although U.S. lenders began adopting automated underwriting models decades ago and have implemented machine learning and other types of artificial intelligence for other use cases, the growing use of machine learning algorithms for credit underwriting is a substantial advancement with potential for significantly improved predictive accuracy. ML techniques can be applied to traditional credit bureau data as well as to new data sources such as bank account feeds and other cash-flow information to detect more nuanced predictive patterns and better assess subpopulations of applicants who may otherwise be masked by larger populations. These advanced techniques show particular promise for expanding credit access for customers who can be difficult to evaluate using traditional models and data sources.

Transitioning to ML models involves adopting new data science techniques, analytic tools, and compliance processes to manage risks and meet expectations with regard to model robustness, explainability, and fairness. While lenders and regulators have become accustomed to particular ways of managing traditional underwriting models, these approaches do not always translate effectively to the machine learning context. ML models can also be more complex than traditional underwriting methods, particularly when they are structured to detect more nuanced relationships between inputs and to ingest larger and more diverse data sources. Particularly for banks, which are subject to the most robust supervision and regulatory expectations, adjusting processes for model risk management, production of adverse action notices, and other compliance requirements can be a significant consideration in deciding when and how to adopt ML models.

The ability to understand how ML models derive their predictions has become an important area of focus as lenders work to satisfy specific regulatory requirements and manage risks more generally. While some commonly used techniques for assessing and managing complex models are well accepted, academic researchers continue to explore and develop new methodologies for managing explainability, fairness, and other concerns. Decisions about tool choice, implementation, and validation have become an important set of risk management questions for lenders who deploy ML models.

This framework is designed as a practitioner-oriented document to outline risk modeling principles, processes, and issues that are helpful to consider in adopting ML models that benefit both lenders and customers. These issues merit careful thought and management, but they are not insurmountable. Many of the issues are not unique to machine learning models, but in fact have also arisen in somewhat differing degrees and forms in prior generations of predictive modeling methods and automated decision-making. ML models often offer greater flexibility in managing

these issues, as well as options for overcoming some of the drawbacks and tradeoffs of more traditional modeling approaches, whether judgmental underwriting or more traditional forms of automated underwriting.[4] Indeed, developing and articulating frameworks for responsible use of ML models has the potential to help facilitate broader improvements in compliance procedures for underwriting models more generally by providing fresh perspectives and new tools and approaches for managing predictive performance, transparency, bias and fairness, and other topics.

This section briefly summarizes key background topics, including how predictive models are used in credit underwriting, basic definitions, key technical differences between ML underwriting models and prior generations of automated scoring systems, factors that are driving bank adoption of ML models and a brief overview of key regulations, guidance, and risks that lenders must account for in implementing ML underwriting models.

2.1 Use of predictive models in credit underwriting

Lenders may use predictive models for multiple purposes at different stages of the credit life cycle as detailed in Appendix B. The primary focus of this framework is the use of such models to help make decisions about whether to approve applications for credit and on what terms. This section briefly describes how such models are constructed and used in these underwriting processes.

2.1.1 Construction of scoring models

Scoring models play a critical role in the consumer credit underwriting process by providing an objective, data-driven way to assess a borrower's likelihood of default. These models use a variety of financial and behavioral data points to predict the likelihood that a consumer will repay their debt on time. Scoring models are integral to decision-making in consumer lending but have limitations that lenders must be aware of when using them as part of the overall underwriting process.

Drawing on records from traditional credit bureaus, scoring models are most often built using the following data elements:

- **Payment history:** Has the applicant made payments on time for existing debts?
- **Credit utilization:** How much of the applicant's available credit is the applicant already using?
- **Length of credit history:** How long have the applicant had active credit accounts?
- **Credit mix:** Has the applicant proven an ability to manage a variety of credit accounts (e.g., credit cards, mortgages, auto loans)?
- **New credit:** How many new credit accounts has the applicant applied for and opened recently?

The models apply quantitative methods such as logistic regression or machine learning techniques to rank order applicants based on predictions of how likely they are to default on additional credit within a certain amount of time.[5] While the actual likelihood of default will change with broader economic conditions, the scoring systems provide relative risk rankings for different groups of consumers.[6] FICO and VantageScore, which are two of the most widely used general third-party scoring models, use a scoring range from 300 to 850, with higher scores representing lower credit risk. For example, consumers with a score of 750 have a much lower chance of defaulting compared to consumers with score of 580, based on historical data about similar borrowers. Individual lenders

and third-party vendors may build other models that use different scales or are oriented so that higher scores represent greater risk, but for simplicity this framework generally assumes that higher scores represent lower risk.

Use of scoring models has spread widely over the last several decades because they offer several advantages over subjective assessments by individual employees:

- **Objectivity:** Scoring models remove some of the subjectivity from the credit underwriting process by providing a standardized way to evaluate consumers.
- **Efficiency:** Automated scoring models allow lenders to quickly assess a large volume of applications. This makes it possible to process loan applications much faster than relying on human evaluation alone.
- **Consistency:** By using standardized algorithms, scoring models ensure that all applicants are evaluated based on the same criteria, which helps maintain fairness in the lending process.
- **Ability to make more nuanced decisions:** Scoring models' ability to factor in multiple inputs in a mathematically consistent way also allows them to make more nuanced decisions. Without such models, lenders often rely on binary rules that can make it difficult for applicants to qualify for credit or may lead to higher losses because rigid rules miss circumstances that would have revealed the applicant would or would not be able to repay.
- **Predictive power:** Scoring models are often calibrated using millions of consumer data points, giving them strong predictive capabilities regarding consumer behavior.

At the same time, it is important to recognize that scoring models are only predictions—not guarantees—and that they have meaningful limitations. These include:

- **Data limitations:** Scoring models can only predict risk based on the historical data they are fed, which traditionally has come from consumer credit reports based largely on current and prior experiences with other loans. Traditional models do not typically use emerging data from new financial products (e.g., buy now pay later products) or other non-traditional data sources (e.g., utility and rental payments). This limits their ability to evaluate new and larger evidence about how people manage their finances responsibly outside of traditional credit channels.
- **Changes in circumstances:** Model performance may deteriorate rapidly where an individual consumer experiences job loss or other financial shocks, during anomalous events like a global pandemic or other broad-based shifts in economic patterns, or where consumers shift their behaviors for non-financial reasons.
- **Bias and fairness concerns:** Models reduce certain types of bias risk relative to more judgmental underwriting methods but can still present risk of inaccuracies with regard to certain populations:
  - **Perpetuating biases in existing data:** Scoring models can unintentionally embed biases if the data used to train them reflects historical gaps or biases. For example, communities with less access to credit or lower incomes may, on average, have lower credit scores, which can perpetuate cycles of financial exclusion. These data are then used to train new models, and the new models can therefore perpetuate the cycle.
  - **Over-reliance on credit history:** Approximately one quarter of U.S. adults have little or no traditional credit history, including about 12.5% of U.S. adults who were estimated to lack sufficient traditional data to generate scores under the most prevalent

third-party credit scoring models as of 2020.[7] These include disproportionate numbers of low-income consumers, younger adults, and households of color.[8] Because assessing these thin or no file consumers can be challenging using traditional models and data sources, they may be rejected, charged higher rates, or face other obstacles to accessing mainstream credit.

2.1.2 Use of scoring models in different credit decisions

Within underwriting processes, scoring models may be used for several different purposes, including deciding whether to approve an application, what price and loan amount to offer, and whether to make subsequent adjustments in the account terms:

- **Initial credit decision:** Scoring models are often the first line of assessment in the underwriting process. They provide a quick way to determine whether an applicant meets the basic risk thresholds for a loan. For example, a lender may set a policy to automatically approve applicants with a credit score of 700 or higher and reject those with a score below 600. For those in between, additional underwriting analysis may be necessary.
- **Risk-based pricing:** Scoring models help lenders adjust the terms of credit based on the predicted risk of default. Consumers with higher credit scores may qualify for lower interest rates, while those with lower scores may be offered credit at a higher rate to compensate lenders for the increased risk of loss. For example, a borrower with a credit score of 760 may be offered a mortgage at 6% interest, while a borrower with a score of 660 might be offered a mortgage at 6.7% to help the lender cover anticipated losses in lending to customers who present higher default risk.
- **Credit limit determination:** Scoring models can help determine how much credit to extend to a borrower. For example, a consumer with a high score might be offered a $10,000 credit line on a card account, while a lower-scoring consumer might only be approved for $1,000.
- **Portfolio monitoring and subsequent account adjustments:** Even after credit has been extended, scoring models are used to monitor the credit health of borrowers over time and sometimes to make adjustments in the terms of open end credit products. For example, where borrowers' credit scores drop, a lender might reduce their credit card limit, change the terms of their loan, or even close accounts to mitigate risk. These actions are subject to various compliance obligations, such as adverse action disclosures under the Equal Credit Opportunity Act.

In making these various decisions, lenders may rely solely on the scoring model outcomes or may apply rules-based criteria and/or human evaluations in addition to considering the model predictions. While the primary focus of this document is ML model development, testing, and deployment, it is helpful to consider the role that these other criteria, inputs, and processes may play when designing broader testing and validation programs. Focusing on scoring models in isolation will not address the extent to which the models' impacts are limited or overridden by other processes and criteria in particular circumstances.

In summary, scoring models are an essential tool in consumer credit underwriting. They allow lenders to quickly and efficiently assess the risk of lending to individual applicants by analyzing their past credit behavior. However, while these models are powerful, they have limitations and are often only one part of a broader, more comprehensive underwriting process that may take into account additional factors or financial context in evaluating the applicant.

2.2 Definition of terms: Artificial Intelligence vs Machine Learning

In considering ML models' advantages for credit underwriting relative to traditional quantitative techniques, it is helpful to define basic terms. While artificial intelligence (AI) and machine learning (ML) are used interchangeably in some settings, financial services practitioners frequently distinguish between the two terms:

- **AI** is a broader concept that refers to the development of computer systems capable of performing tasks that typically require human intelligence. These tasks include problem-solving, planning, understanding natural language, speech recognition, visual perception, and decision-making. AI aims to create machines that can mimic human cognitive functions. AI can be categorized into two types:
  - **Narrow AI (Weak AI):** Systems designed and trained for a particular task, such as the underwriting models we are studying in this document, spell checkers, or image recognition.
  - **General AI (Strong AI):** Hypothetical AI that possesses the ability to understand, learn, and apply knowledge across a wide range of tasks, like human intelligence. General AI remains a goal and has not been fully realized.
    - Generative AI, which became a major focus of AI policy discussions after the advent of Chat GPT in November 2022, is not in fact general AI but involves models that can create new content (text and images) based on learned patterns in extremely large amounts of training data that is often scraped off the internet. The outputs can look and sound quite convincing at times, but they are also subject to errors (often called hallucinations) and inconsistencies and raise heightened concerns due to their complexity and the nature of the data on which they were trained.
- **ML** is a subset of AI that focuses on deploying techniques that enable computers to learn from data and make predictions or decisions without being explicitly programmed. ML algorithms learn patterns and relationships within data and use this knowledge to make informed predictions. ML can be divided into three main types.
  - **Supervised Learning:** The algorithm is trained on labeled data, where the input and corresponding output are provided, and the goal is to learn the mapping between them.
  - **Unsupervised Learning:** The algorithm is given unlabeled data and must discover patterns or relationships on its own.
  - **Reinforcement Learning:** The algorithm learns by interacting with an environment, receiving feedback in the form of rewards or penalties, and adjusting its actions to maximize cumulative reward.

In addition to the AI/ML distinction, models can also be categorized based upon their method of training, online or offline learning:

- **Online learning** occurs on an ongoing basis as new data becomes available, creating dynamically updating models. This process of constantly learning through updating the parameters makes online machine learning adaptable to different situations that may occur in a dynamic environment over time. However, ensuring transparency and structuring compliance and validation processes are more complicated in light of the models' continual evolution as they operate in production.

› **Offline or batch learning** refers to processes where learning over all the available observations in a dataset occurs at one time, producing a static model that does not change until the developers manually launch an update that incorporates additional datasets and decide to deploy the revised model. This sequence produces less frequent updates, but creates a natural cadence for analyzing what changes have occurred in the newer datasets and for performing compliance and validation activities before an updated model is implemented in production. This is the paradigm used in the development of bank underwriting models.

This framework document focuses on supervised learning models that are used for credit underwriting decisions regarding approvals, amount, and pricing, where the model parameters are calibrated using offline learning. While other types of artificial intelligence models such as generative AI models and online (dynamically updating) machine learning models receive significant public attention in broader debates about responsible AI, and may have legitimate uses elsewhere, they are outside the scope of this document as they are not widely used by banks in the credit underwriting context because of concerns that they present risks that cannot be effectively managed for this use case.[9]

2.3 Technical differences between ML and traditional models

As noted in the introduction, machine learning scoring models introduce greater technical complexity than traditional predictive models, necessitating careful development and oversight. However, many of the challenges associated with ML models also arise in traditional approaches, and ML models vary significantly in how they balance predictive power with risk and compliance considerations. As financial institutions gain experience with these methods and analytic tools continue to evolve, a larger number of banks are becoming increasingly comfortable with incorporating ML techniques into their underwriting processes.

While traditional logistic regression models are widely used for credit underwriting, they impose structural constraints that can limit their predictive accuracy and flexibility. The models generally assume relatively consistent and stable relationships between input variables and predicted outcomes. Developers also generally strive to select inputs that are not highly correlated to avoid uncertainty in attributing predictive power to individual variables. This is difficult in the credit context because traditional inputs are often strongly correlated, creating a tension between the need for regulatory transparency and the desire to maximize predictive accuracy.[10] As a result, traditional models are typically designed to be "parsimonious" by selecting input variables that add significant independent predictive power, while excluding those that may primarily be useful in specific cases or for specific subpopulations.

Model developers can use specialized techniques to manage some of these limitations to varying degrees, for example by building separate scorecards (or mini-models) for certain segments of applicants[11] or by using various feature engineering techniques to transform relatively simple inputs into more mathematically sophisticated metrics.[12] However, these techniques take expertise and resources to deploy, and in some cases require significant manual effort by developers to determine what works for a particular data set in particular circumstances. Manually constructing a segmented model can be error prone and exploration of segmentation strategies and feature combinations to employ in the various segments is often allocated limited time. This results in only a partial exploration of the potential models. For example, most segmented models have less than a dozen segments.

In contrast, machine learning techniques automatically and systematically explore combinations of variables and segmentation schemes. They provide more flexible and automated approaches to modeling complex relationships within the data. Machine learning is able to capture relationships that are context-specific (where a variable influences a predicted outcome only in circumstances where another variable falls within a specific range of values), nonlinear (where increasing an input feature does not change the likelihood of default by a consistent amount), and even non-monotonic (where increasing an input feature may decrease or increase the likelihood of default in different circumstances), as described further in the chart below. Human model developers would have a difficult time manually exploring all of these combinations, so machine learning algorithms often generate models that yield better predictive performance than more manual methods. These models are often more complicated, however; instead of a handful of segments, hundreds of model variations may be explored and selected for the final model, reflecting the more thorough exploration of the model design space made possible by modern learning algorithms.

FEATURE RELATIONSHIP	DESCRIPTION	EXAMPLE
CONDITIONAL / CONTEXT SPECIFIC	Degree to which one variable influences a prediction depends on a different variable taking on a specific range of values	A recent delinquency on a revolving account that is rarely used by the consumer may be substantially less indicative of default risk than a recent delinquency on a revolving account that is frequently used.
NON-LINEAR	Increasing an input feature does not change the likelihood of default by a consistent amount	Debt-to-income ratios might show gradual increases in risk until a certain threshold of indebtedness is reached, after which borrowers become substantially more susceptible to default risk.
NON-MONOTONIC	Increasing an input feature may decrease or increase the likelihood of default in different circumstances	In assessing the relationship between credit utilization rates and default risk, risk levels may be lower for consumers with average utilization rates than for consumers who almost never use credit or consumers who are heavy users and are nearly maxed out on available accounts.

The ways in which machine learning models memorialize these more nuanced relationships vary. For example, tree-based models such as XGBoost identify key variables that define meaningful subgroups, refining predictions for each subgroup through a series of iterative splits. This structure allows features that are similar or correlated to contribute uniquely by capturing distinct patterns across different subpopulations. Using ensembles of tree-based models further enhance model reliability by combining multiple predictions to minimize errors and improve robustness. Another group of models are structured as artificial neural networks, which generate one or more layers of "latent features" that represent complex interactions between input variables and enable deeper insights into borrower risk characteristics. Both types of ML models are more dynamic in the way that they segment data and subpopulations to share learnings and signals across subgroups while also capturing more nuanced and specialized insights.

Model developers also have somewhat more flexibility in deciding how many features to incorporate into ML underwriting models than traditional logistic regression approaches, including features that may have relatively high degrees of correlation.[13] Although both feature correlations and model architecture choices can make it more complicated to understand the role of particular inputs in ML

models' predictions, model developers have a range of options for managing these concerns. For example, in addition to deciding how many inputs to include, developers can constrain the depth of trees or number of latent layers in their neural networks and place restraints on how many input variables can be used to generate a subsequent latent feature. Many lenders have also become somewhat less concerned about correlation patterns in the context of ML models such as tree-based methods because features that may be similar or correlated can be used in very different ways within separate branches that are focused on different customer segments.

Some lenders choose to manage concerns about ML model complexity and explainability largely through constraints on architecture, relationships, and input feature selections. Proponents of this approach sometimes describe these models as "inherently interpretable," although they can still involve complex mathematics that require careful validation and testing. Other lenders use a combination of architecture constraints and secondary techniques and tools (often called "post hoc explainability techniques," since they are applied to models without changing their basic operations) to analyze the operation of their models and the roles that individual input features are playing within them. Selecting and validating post hoc explainability techniques is also a critical risk management exercise for lenders who choose this approach.

Due diligence in testing and monitoring models for robustness and stability is also a significant focus in transitioning to more complex ML models. As discussed further in Section 6, these processes are designed to evaluate whether models are so precisely fitted to their training data that their performance will tend to deteriorate when exposed to new data and to monitor models that are in production for situations in which shifts in the economy, applicant populations, or other conditions warrant adjustments. Concerns about overfitting and data drift are not unique to machine learning models, but warrant careful consideration particularly since the algorithms are more precise in fitting the data and detecting nuanced interactions compared to traditional models.

2.4 Factors that are driving banks to adopt ML models

Lenders and credit score developers are increasingly drawn to using ML techniques and models because of their potential to predict credit risk more accurately, with or without new data sources.[14] More powerful predictive analytics can potentially benefit borrowers, firms, policymakers, and investors alike by:

- Expanding access to more borrowers who are creditworthy;
- Decreasing the frequency with which consumers are offered loans they cannot repay;
- Reducing portfolio default rates and credit losses;
- Reducing mispricing based on inaccurate estimation of the likelihood of default and improving terms at which credit is offered to some applicants; and
- Facilitating less costly and faster model generation and updating.

To expand credit access to consumers who often struggle to obtain mainstream loans, lenders need to improve their predictions of probability of default (and rank ordering of the probability of default) as they assess applicants that are different from those who were historically approved. These applicants may be further down in the credit spectrum as defined by traditional third party credit scores, but still be a good credit risk. This is especially true for banks that wish to lend to consumers with thin credit files. Research shows that predicting the performance of consumers with little traditional credit history or with periods of past repayment difficulties is often more challenging

than for consumers with thick files and strong credit scores.[15] Increasing lenders' confidence that they can identify which applicants can repay in these cohorts is critical to influencing their decisions about how to define (and expand) their credit boxes. ML models' ability to factor in additional variables and detect more nuanced relationships in the data both help to increase their predictive power.

Use of machine learning techniques can also improve the speed and efficiency of model development in a variety of respects. For example, it facilitates generation and evaluation of a larger set of alternative models than is feasible with conventional approaches, potentially helping lenders to identify and deploy models that have greater accuracy and fairness.[16] ML models may also reduce the complexity of credit policies—sometimes called overlays—because of their ability to assess a greater number of features and to model more complex relationships among features that are difficult to incorporate directly into traditional models, such that lenders that use logistic regression will instead layer on additional credit policy rules to account for those factors in light of their risk appetites. Finally, machine learning methods can sometimes be faster than traditional approaches in updating models in response to sudden shocks or other developments, though they still require accumulation of significant amounts of data, careful review, and analysis.

Finally, a number of lenders are also interested in incorporating new forms of credit information—particularly alternative financial information such as cash-flow information from transaction accounts—to further increase the accuracy and inclusiveness of their credit models, particularly for applicants who are difficult to assess using traditional credit history alone. The "open banking" infrastructure for customer-permissioned access to transaction account data is continuing to expand, despite some regulatory uncertainties.[17] Creating widespread access to this data has the potential to enable development of even more predictive models that combine credit and deposit data, creating opportunities to expand credit access as well as to create significant competitive advantages for early adopters. However, while machine learning models' ability to ingest large amounts of information can be helpful when starting to use new data sources, it is important to note that the two innovations are not dependent on each other. Particularly given the need to adjust business and compliance practices, for example, some lenders may choose as a first step to develop traditional regression models that include non-traditional data. Others may choose to adopt machine learning models that rely solely on traditional credit bureau inputs, even if they plan to explore additional innovations over time.

2.5 Relevant regulations, guidance, and risk categories

As lenders implement machine learning underwriting models, they must be cognizant of various federal consumer protection laws as well as broader prudential guidance expectations regarding managing model risk and third party vendors where applicable. These include:

- **Adverse action disclosures:** The Equal Credit Opportunity Act requires lenders to provide individualized disclosures to credit applicants of the "principal reasons" for rejecting an application, terminating a credit line, or taking certain other adverse actions.[18] Where lenders rely on information from credit reports, the Fair Credit Reporting Act similarly requires them to provide consumers with a list of "key factors" that are negatively affecting their credit scores if the score was a factor in an adverse action or prompted the lender to charge higher prices.[19] These requirements (which we refer to collectively as adverse action requirements) force lenders to analyze which input variables play the biggest role in generating predictions for individual applicants.
- **Fair lending compliance:** The Equal Credit Opportunity Act (and Fair Housing Act in the mortgage context) prohibits discrimination on the basis of race, ethnicity, gender, and other

protected characteristics in credit decisions.[20] Fair lending laws have historically been interpreted not only to prohibit intentional, overt discrimination, but other forms of inconsistent treatment even if not motivated by animus ("disparate treatment") and the use of facially neutral criteria that have a disproportionately adverse impact on the basis of protected characteristics, unless the criteria further a legitimate business need that cannot reasonably be achieved through less impactful means ("disparate impact"). The Trump Administration announced in late April 2025 that it intends to eliminate disparate impact liability, although some state laws also incorporate the concept.[21] The Consumer Financial Protection Bureau has proposed to modify implementing regulations under ECOA to state that it does not provide for disparate impact liability.[22] The Department of Housing and Urban Development has also proposed to eliminate its regulations on disparate impact under the Fair Housing Act, leaving application to the courts under prior Supreme Court precedent.[23]

- **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 underwriting and other models. These model risk management (MRM) processes include analyzing whether the models are relying on relationships in the data that are "conceptually sound" and assessing models' stability in changing data conditions.[24] Risk management often conceptualizes three lines of defense, with the first line consisting of business/operational staff, who are generally responsible for the model risk associated with their business strategies; the second line of compliance/risk-control staff, who manage independent validation, measurement, and monitoring; and third line of internal audit staff, who assess the efficacy of the overall framework and the other lines' processes.[25] Trump Administration officials indicated in March 2026 that banking agencies may issue additional model risk management guidance with regard to the use of artificial intelligence.[26]
- **Third party risk management:** Where banks rely on a third party to design, develop, or operate tools or processes that are part of their lending operations, they are generally responsible for overseeing the vendor's compliance with applicable regulations. Similar expectations apply to financial services providers that are supervised by the Consumer Financial Protection Bureau. Guidance emphasizes the importance of due diligence in initial selection as well as ongoing monitoring of performance.[27]

Of these various requirements, model risk management is a particularly broad and flexible set of expectations. In banking contexts, "risk" is sometimes used to describe the possibility of events negatively affecting a bank's current or projected financial condition. Other definitions focus on a broader range of potential negative outcomes, including those that violate law, raise public policy concerns, or harm customers. The OCC identifies seven specific risk categories for bank supervision: credit, interest rate, liquidity, price, operational, compliance, and strategic.[28] These risks are interconnected, and any product or service can expose a bank to multiple risks.

Use of quantitative models can also impact risk across all of the categories, with the potential to increase or decrease the organization's cumulative risk based on the particular model's purpose and use and the effectiveness of the bank's model risk management practices. These risks arise from two main sources: fundamental errors in the model and incorrect use. Errors in design, implementation, input data, or assumptions can lead to inaccurate outputs. In addition, even a sound model may pose high risk if used incorrectly, especially outside its intended environment, given the inherent simplifications that are built into quantitative models. Lenders need to understand the limitations of their models—both due to shortcomings, approximations, and uncertainties in their models and to

assumptions that may restrict models to a particular set of circumstances and situations—to avoid deploying them in ways that are not consistent with original intent.

The degree of risk generally increases to the extent that models are more complex, their inputs involve greater uncertainty, and their deployment involves larger numbers of customers or more impactful decisions. Managing model risk hinges in part on understanding its source and magnitude. Because using machine learning models in credit underwriting can introduce risks across all of the categories, a robust validation framework is essential to address inherent risks in each category effectively.

3. INITIAL MODEL DESIGN CONSIDERATIONS

Developing machine learning models involves multiple steps, including algorithm selection, data selection and preparation, model training, validation and testing against hold out datasets, and "tuning" of certain parameters to optimize performance. In addition to this core sequence, lenders conduct additional validation and testing processes to satisfy specific compliance requirements, ensure the safety and soundness of the model, and to confirm the desired business outcomes will be achieved before moving a model into production.

The decisions that shape these steps are driven by a wide range of technical, business, legal, and operational considerations, and it is beyond the scope of this document to provide a full discussion of all potential topics or choices. However, this section highlights a few model design decisions that have important implications for compliance and help to provide background for subsequent discussions on more specific compliance regimes. Above all, banks that are implementing machine learning models emphasize the importance of creating a thoughtful process that documents that lenders have made informed, intentional decisions about structuring model development, compliance processes, and the deployment of human resources and vendors to give confidence to both internal and external stakeholders that the resulting models are responsible and reliable.

3.1 Choice of algorithm and model architecture

Model developers consider several factors in choosing which machine learning techniques to use in building underwriting and scoring models, including predictive accuracy, explainability, and the level of existing expertise and familiarity with particular approaches. Although tree-based models are often viewed as more explainable than neural networks because they use relatively intuitive logic to subdivide populations and build predictions, complexity and explainability depend not just on the technique chosen but on secondary decisions such as the number of inputs (discussed more in the next section), "hyperparameters" for the model such as the depth of the trees or number of layers permitted in the model, and how many models are combined into ensembles to reach final predictions. For example, the XGBoost learning algorithm can generate thousands of tree-based models, each of which are based on the prediction error of the prior model in the sequence. It then creates a final prediction based on the sum of the predictions from the various trees. This is more complex than a single decision tree but leads to lower prediction error rates and better predictive power that generalizes well to datasets not used in the initial model development process.

Some lenders and other model developers (such as vendors and companies that sell third party credit scores) have a policy of severely constraining the complexity of their models to promote greater explainability, while others may take an empirical approach during the model development process to test how much different constraints affect explainability as compared to accuracy. In addition to limiting tree depth or the number of network layers, for example, lenders may restrict how many input features can contribute to any given latent feature or impose constraints that generally require models to structure data relationships as linear and monotonic while allowing exceptions in limited cases where the relationships are particularly strong and intuitive. Where lenders decide to allow more complexity, they typically use post hoc explainability tools to assess the role that different variables are playing in the model and other aspects of its operations to ensure the relationships the model learned are intuitive and easily justified. These tools have no impact on model predictiveness, as architecture constraints do, but they require technical capacity to use them effectively and interpret the results correctly. While some of these techniques have been used by modelers for some time, others are relatively new, and they are taking on increased significance in the context of more complex ML modeling methods.

Most lenders use a combination of both architecture constraints and post hoc tools, but the relative mix varies along a significant spectrum. In addition to weighing tradeoffs with regard to explainability and predictive accuracy, additional considerations may include computing capacity, compliance staff resources, and technical expertise. Practices are widely expected to continue evolving as techniques and thinking about both building inherently interpretable models and deploying post hoc tools continue to improve.

Focus on explainability issues in managing ML models is prompting a broader recognition of the fact that even traditional models and methods present their own transparency tradeoffs and challenges. For example, the choices practitioners made when they decided to include or exclude a variable in a traditional model may not always be that transparent, and almost certainly fail to fully capture underlying causal relationships among correlated variables. However, because the metrics and processes have become so familiar and widely accepted over time, stakeholders may not fully recognize those underlying challenges with traditional approaches.

To illustrate, assume that model developers are working with a dataset that has several similar metrics:

- **Element #1:** Maximum single limit on bank card account, including accounts for which the consumer is an authorized user
- **Element #2:** Maximum single limit on bank card account, including only accounts for which the consumer is directly liable as an accountholder
- **Element #3:** Maximum single limit on retail card account
- **Element #4:** Maximum single limit on revolving credit accounts other than home equity lines

Based upon a traditional variable selection process, developers may select only one or two of these metrics for including in a logistic regression model, such as Element #2 & #3. However, they may do so without really understanding what risk is being distinctly captured by these two elements or how they implicitly interact in a logistic regression model. Similarly, they may not know the circumstances when Element #1 is more predictive than Element #2, or the distribution of authorized users for the other elements. While documentation of what is included is often robust, documentation of what is excluded is generally not, and a thorough evaluation of more than a small number of alternative models is seldom pursued.

Greater recognition of this context can be helpful in weighing the pros and cons of different approaches for both traditional and machine learning models. It is also important to recognize that there is no one definition or metric for transparency that is applicable across all circumstances. A wide variety of stakeholders—including different types of bank staff, regulators, and borrowers themselves—require different types of information about model operation at different times. For this information to be effective, it must not only be technically reliable but also conveyed in ways that are actionable for the particular audience. This requires consideration of communication framing, wording, and delivery that are helpful to consider for particular use cases. Subsequent sections discuss different approaches to transparency and explainability for specific types of compliance and business purposes.

BOX 1 ADVANTAGES AND DISADVANTAGES OF NON-LINEAR ML ALGORITHMS

In deciding how broadly to allow machine learning models to recognize relationships that are non-monotonic or non-linear, lenders face a number of considerations:

Advantages of Non-Linear ML Algorithms:

- Typically provide high predictive accuracy, especially where there are complex, non-linear relationships within the data.
- Can effectively capture non-linear patterns and interactions among features, making ML techniques suitable for a wide range of datasets.
- Include established, automated regularization techniques that can be tuned to manage the risk of overfitting, making models robust and capable of generalizing well to new data.
- Most algorithms provide automated insights into feature importance, helping to identify which variables contribute the most to the model's predictions.
- Once input variables are set, the approach to building the model requires fewer decisions by the model developer and the decisions made are easier to document and validate (e.g. by employing systematic hyperparameter tuning).

Disadvantages of Non-Linear ML Algorithms:

- Training the model can be computationally intensive and time consuming, particularly with large datasets and complex models.
- Model development requires careful hyperparameter tuning from a knowledgeable developer, and improper tuning may lead to overfitting or underfitting issues (including sensitivity to noisy data, outliers, or anomalies), which may impact its performance.
- The models are more complex and often have more variables than linear models, which makes them more challenging to interpret and often requires different and more sophisticated approaches and tools to understand and manage them for purposes of risk management and compliance.

3.2 Choice of data

All modeling work starts with data, and in the context of machine learning models it is important that modelers select data that is reliable and representative of the customer segments they expect the model to predict (e.g., thin file, different risk bands, etc.). A model with sufficient statistical support in all relevant segments will be more accurate and robust than a model built on limited data. Model developers should take adequate time to establish the validity of the data and ensure all relevant segments are well represented. In the context of underwriting models, positive and negative outcomes are imbalanced: Fewer than 10% of borrowers go delinquent on their loans globally, and most banks target much lower default rates. Lenders' data about their booked loans is already significantly

limited, since their existing underwriting practices already select for default risk. Acquiring additional data can provide a more holistic picture of applicants that an institution would historically not accept, but who may still apply. Likewise given the low incidence of deep delinquencies, it is important to ensure the delinquencies are well distributed within segments of interest.

A thorough exploration of the data should thus include an analysis of the target rate (i.e., the ratio between defaulters and non-defaulters) by segment and over time, to ensure the model will have sufficient statistical support to underwrite all of the applicants it will be expected to underwrite. Otherwise, the model may be dominated by larger segments, creating greater variance and reducing performance as to smaller populations. Subject to certain limitations, developers can reduce the impacts of imbalanced samples by adjusting training datasets through over or undersampling, using certain ensemble methodologies, and various other modeling techniques. However, such strategies do not provide as robust a solution as obtaining representative data in the first instance and need to be deployed carefully to ensure that they operate as intended and do not instead introduce additional bias and predictiveness issues. Ensuring that initial data is as representative as possible is therefore a threshold concern across lenders and model developers.

The data exploration stage is also where developers make decisions about whether and how to deploy reject inference techniques to build models that can predict risk for populations who fall below a lender's existing credit score cutoffs or other criteria. Such techniques can be complex to deploy but are important to help lenders potentially expand beyond their existing models to reach additional creditworthy customers.[29]

Another important threshold activity is evaluating particular variables and cleaning the associated data. At a high level, this includes consideration of the lender's legal ability to use the data by law or contract and of privacy considerations. At a more granular level, data cleaning can include removing low variance or constant values (which will not help a model make predictions), removing variables that have a high missing rate, removing highly correlated variables, and removing or treating outliers. Depending on the choice of algorithm, variables may be further treated to accommodate learning algorithm assumptions and limitations. While machine learning algorithms are generally more robust to noisy data, data cleaning can still result in a better model.

The number of input variables used in machine learning models is another topic on which there tends to be a spectrum of approaches among lenders and model builders, with some models incorporating thousands of variables and others incorporating dozens or hundreds. Among the latter group, developers may sometimes start the process with a much larger number of raw inputs and engage in substantial testing and feature engineering to focus their ultimate models on a smaller number of variables that have already been screened and in some cases structured to determine their relative predictive contributions.[30]

Statistical methods for selecting features are a primary determinant of the ultimate number of features included in a model. But broader considerations can also play a role, including predictive value, whether the relationship between the feature and the modeled outcome is intuitive, whether the definition of the feature is easy to explain, and how difficult it is to operationalize a model that includes a particular feature. Compared to nonbanks that may be more likely to adopt models with thousands of inputs, banks tend to favor models that are parsimonious (simpler models with fewer variables and parameters), in part due to concerns about overfitting to training data, especially when the number of features approaches or exceeds the number of observations. Banks also frequently emphasize that large numbers of features tend to result in models that are often less interpretable and require more resources to train and deploy. These resources include controls to assess and maintain the data quality of each input feature for both development and live production deployment.

From this perspective, lenders may conclude that incorporating many highly correlated variables or variables that only marginally improve performance unnecessarily increases model complexity and operational costs without adding commensurate predictive value.

To assess the potential tradeoffs, lenders often conduct analyses to optimize the number of input features. Key strategies include feature importance analysis using post hoc secondary techniques (as discussed in more detail in subsequent sections) and marginal utility analysis, which evaluates the incremental benefit of adding features by assessing the change in performance metrics (e.g., ROC-AUC, KS statistic[31]) as features are added and enables the developer to stop when the improvement becomes negligible. Documenting the reasoning for feature inclusion/exclusion is important to justify the selection with regard to the basic business justifications for including particular features as relating to credit risk and to meet risk management and compliance obligations. Care should be taken to ensure a balance of explanatory power and diversity among the variables selected, such that the model is not over-reliant on a small set of variables or close variations, which can lead to instability over time.

Because there is no universal formula for the optimal number of features using a particular modeling algorithm, lenders that adopt ML models still face important decisions in determining how many inputs to feed into their systems and how much latitude to provide for the creation of latent features within the model. They may also make decisions about which individual variables to include based on considerations about providing adverse action disclosures to individual consumers, model risk management analyses, comparative studies to justify complexity, the reliability and stability of particular modeling methods, and fairness considerations. Subsequent sections discuss these compliance areas and considerations in more detail.

3.3 Talent and resource considerations

Talent and resource considerations are also foundational in transitioning to machine learning models because they affect the quality of models constructed, the effectiveness of risk management controls, the costs of implementation, and the ease of deploying ML models to production.

As an initial matter, the decision to build or buy ML underwriting models from vendors depends on the lender's size and the technical sophistication of its team. Third party providers, including credit score providers and technology firms, offer various solutions to address the resource challenges faced by many financial services companies. At the same time, providing appropriate oversight also requires knowledgeable staff to analyze and challenge choices made during the model development process and to monitor ongoing performance as discussed further in Section 6 and Section 7. The availability of open source software has also reduced barriers to entry for lenders that decide to build models in-house. Access to powerful open-source libraries, documentation, and datasets including pre-trained models can accelerate development and enhance efficiency, although use of such materials also requires thoughtful management of governance, security, and compliance.

Careful attention to the human teams that are responsible for developing, deploying, and managing the risks of machine learning models is also critical. As discussed further in Section 6 and in interagency guidance on model risk management, using experienced, interdisciplinary teams that were not involved in initial model development to provide effective challenges by testing and validating models is essential. Ongoing training for a broad range of staff who work with and on ML models is important as technologies and practices continue to evolve, and clear role definition and processes for oversight and feedback functions are important to ensure effective risk management. Ensuring open communication between data scientists, second-line risk management staff,

IT/engineers, and operations teams is particularly critical to ensure effective and efficient deployment and ongoing risk management functions.

A final set of considerations relates to how ML models will be deployed in production and operated over time. For example, lenders may find it advantageous to modernize their tech stacks to ensure that ML models can be consistently and easily deployed without disruption, while also maintaining strong governance and oversight. Modernizing technologies can also make the effort of developing, deploying, and operating ML models less burdensome and less error prone. For example, legacy infrastructure usually requires manual re-coding of business logic (such as model scoring logic) as it moves between systems and is implemented in an operational environment. Containers, ML operations platforms, and consistent use of application programming interfaces (APIs) can help to make the transfer of models and data more streamlined,[32] though encryption, access controls, and other data security measures may also need to be considered in refining model deployment and monitoring processes. Vendors' technical expertise and constraints can also sometimes be an important factor in both design and deployment decisions.

4. ADVERSE ACTION DISCLOSURES

Adverse action disclosures are a critical component of credit decisions, mandated by the Equal Credit Opportunity Act (ECOA)/Regulation B and the Fair Credit Reporting Act (FCRA)/Regulation V. These laws are designed to ensure that credit decisions are fair, transparent, and nondiscriminatory, and that individuals are informed of the reasons behind adverse decisions such as application denials or account terminations.

Under ECOA, the disclosures must indicate the principal reason(s) why the adverse action was taken. The reasons must be specific and must relate to and accurately describe the factors considered by the lender or used by an underwriting model. Additionally, no factor that was a principal reason may be excluded from disclosure. The model reasons must not be overly broad or vague to the extent that they obscure the specific and accurate drivers used by the model. Model reasons must be provided even if the relationship between a factor and its ability to predict creditworthiness may not be clear to the applicant.[33] Under FCRA, the lender must also disclose certain information where an adverse action was based in whole or in part on information in the consumer's credit report or a consumer's credit score was used as a factor in risk based pricing, including key factors that adversely affected any credit score used in the adverse action or pricing decision.[34]

Adverse action notice requirements promote fairness and equal opportunity for consumers engaged in credit transactions by serving as a tool to prevent and identify discrimination because creditors must affirmatively explain certain types of decisions to consumers. Adverse action notices also help consumers identify and correct errors in credit reports and other data used to evaluate their applications. In addition, such notices can educate consumers and help them understand the reasons for a creditor's decision so they can take steps to improve the factors that impact their credit status or use the information to seek a reconsideration or exception.

Because machine learning algorithms are often trained to reflect higher order interactions between credit attributes, the individual contribution of a single attribute in adversely impacting consumer's score may not be easy to isolate. Accordingly, producing compliant adverse action notices for these models can require more effort than for traditional models. For practitioners, this compliance requirement fundamentally presents an issue of ML explainability.

4.1 Adverse action regulatory compliance and policy objectives

4.1.1 Regulatory requirements

Credit scoring models used by U.S. lenders are required to meet the standards of both ECOA and FCRA. ECOA mandates that lenders provide disclosures to consumers that state their "principal reasons" for denying applications, reducing existing credit lines, or taking other adverse actions.[35] Where lenders take negative action based on information obtained from credit reports or other third parties, FCRA imposes additional requirements including the disclosure of "key factors" that have adversely affected consumers' credit scores if lenders rely on the scores in taking adverse actions or in implementing risk-based pricing.[36]

An adverse action is specifically defined as a refusal to grant credit in the amount or on the terms requested in an application; a termination of an account or an unfavorable change in the terms; or a refusal to increase the amount of credit available to an applicant who has made an application for an increase.[37] When an adverse action occurs, specific reasons for the decision must be provided to the consumer in a timely manner. The specific reasons disclosed must relate to and accurately describe the factors that were the principal reasons for the decision.[38]

Neither ECOA nor FCRA dictate specific methodologies for determining which credit attributes should be considered "principal" or "key" for purposes of the disclosures, though regulatory guidance provides some guideposts. The adverse action notice must include any factor that required an automatic denial under the lender's policies, such as where lenders reject all applicants with a bankruptcy in the past two years. For lenders that use credit scoring systems, guidance describes two potential benchmarks to compare an individual consumer to other applicants for purposes of determining which attributes should be considered the "principal" basis for an adverse action, but also permits other methods that produce substantially similar results. Generally, providing more than four or five reasons to explain the adverse action or pricing decision is disfavored due to concerns about overloading consumers with too much information.[39]

ECOA and its implementing regulations also require the description of principal reasons be specific and accurate, but they do not set out metrics or thresholds for evaluating these qualities beyond directing that the information "relate[s] to and accurately describe[s] the factors actually considered or scored."[40] Simply stating that the lender's internal standards were not met or that the applicant does not have a sufficient credit score does not satisfy the specificity requirement, but guidance also emphasizes that disclosures need not explain how or why the identified factors mattered in lenders' overall analyses. A list of sample reason codes in the appendix to Regulation B provides very simple examples, such as "length of residence" and "poor credit performance with us," but guidance emphasizes that lenders should not simply pick the closest factor listed if it was not actually used to make the adverse decision.[41]

4.1.2 Broader policy considerations: Fairness, error correction, education, and useability

Legislative history and policy debates have identified three broader policy objectives supporting the provision of adverse action and risk-based pricing notices to consumers:

- **Discouraging 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.

- **Helping applicants detect errors and seek corrective action** by describing the primary bases for an adverse decision and key factors that are negatively affecting their credit scores. This is especially true where the disclosure highlights data errors—such as a default on a student loan when the consumer never had such a loan. Where information is derived from a consumer report, regulations require notices to include the source of the data and contact information so the consumer can file a dispute if there is an error.
- **Promoting education and self-improvement** by 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. This helps notice recipients understand how past financial behavior or their current financial position is affecting their access to credit.

Although neither "useability" nor "actionability" are specifically referred to in federal regulatory guidance, some lenders view these concepts as important considerations in building their adverse action programs. They may use the terms to refer both to whether a disclosure provides recipients with sufficient information to identify data errors that may need correction and to whether the notices can help consumers understand how to improve their chances regarding future credit decisions. When emphasizing these concepts, lenders are often motivated both by a desire to make the information they are providing more useful to customers and to confirm whether disclosures are clear enough to satisfy legal requirements about accurately disclosing the specific principal reasons or key factors that drove a particular decision. For example, if a reason description is so vague that a recipient would have a hard time figuring out whether an error had occurred, examiners may also be likely to find that it is out of compliance with regulatory standards.

This framework generally describes the underlying concept as "useability," since "actionability" can sometimes be read to imply that there should be a specific action that a denied applicant could or should take in response to every reason code. However, other than taking action to correct errors if there is a mistake in the applicant's credit record, there are some factors for which there is no immediate action that the applicant can practicably take, such as the existence of a past bankruptcy, which remains on credit reports for several years. Nevertheless, such reasons cannot be omitted from the disclosures if they are principal reasons or key factors, and they still have utility to recipients for general knowledge and error correction purposes.

Lenders may make different judgments as to what kind of language and detail foster useability, where they consider it as part of their disclosure design processes. For example, listing "90-day delinquencies on mortgage loans" may be more helpful for error correction purposes to an applicant that has in fact never had such a delinquency, but might prompt some recipients to prioritize their mortgage loans at the expense of other types of credit accounts when in fact serious delinquencies on other types of credit products would be likely to result in similar scoring outcomes. Accordingly, some lenders choose to use slightly higher level descriptions (such as "recent delinquencies on credit accounts"), while still distinguishing between delinquencies in the past year as compared to older behavior. As described below, these kinds of considerations can take on additional significance in the context of machine learning models that incorporate larger numbers of similar variables or that recognize non-monotonic and conditional relationships.

4.1.3 Secondary benefits to lenders

Beyond the direct benefits of adverse action notices for individual recipients, some lenders find it helpful to analyze the distribution of adverse action reasons for other compliance purposes such as model risk management as discussed in Section 6. Particularly for models that are already in

production, for example, analyzing the distribution of adverse action reasons might be one way of assessing the extent to which a model is heavily reliant on a particular subset of variables or understanding differences in how two models are assessing the same applicant pool. After a model has been deployed, a substantial shift in the distribution of reason codes may be one way of detecting that the applicant pool is changing or that the performance of the model is shifting in response to change in other external conditions. However, other lenders rely on other mechanisms to monitor for these kinds of risks and changes as discussed in subsequent sections.

4.2 Overview of choices and challenges with machine learning models

Machine learning models often involve more complex feature interactions and sometimes incorporate more input variables and/or less traditional data sources than traditional credit scoring models. Because of this, compliance with adverse action notice requirements is a significant focus for lenders when adopting ML models. In addition to the choices discussed in Section 3 about machine learning techniques and data sources during the initial model development process, lenders who decide to adopt more complex machine learning models must also make decisions about the post hoc explainability methodologies used to assess which variables drove decisions for individual applicants. Lenders also face decisions about whether and how to aggregate similar variables into broader categories and what language to use in describing the reasons on the disclosure notices. Validating methodologies and processes used to generate the disclosures is a significant component of transitioning to machine learning models.

This section provides a brief overview of key decisions and concepts, with subsequent sections providing more detailed discussions of particular considerations and lender practices.

4.2.1 Feature importance methodologies, including post hoc explainability techniques

Lenders' methodologies and processes for generating adverse action notices vary to some degree even in the context of linear and logistic regression models. But compliance in these more traditional contexts is facilitated by the fact that the models specify the weights of individual features and there are already commonly accepted statistical methods to measure the importance of different features for business and regulatory purposes. Because machine learning models are relatively new, and often more complex in their notation and function than logistic regression models, lenders are having to adjust their processes. Early on, some companies tried various methods that proved inadequate. For example, one early method of generating adverse action notices was to build a second "proxy" model that relied on linear regression techniques to predict the outcomes of the machine learning model and then explain that proxy model using familiar techniques. However, it is widely recognized today that methods based on proxy models do not generate sufficiently accurate explanations of how a machine learning model is operating. These methods cannot accurately quantify the importance of individual variables within that model and should not be used for generating adverse action notices.

As discussed in Section 3, some lenders rely primarily on extensive constraints to create what are often described as inherently interpretable models, while others are adopting more complex ML models and using post hoc explainability techniques to assess feature importance for the purpose of generating adverse action notices and other required outputs. The latter approach often involves selecting from among multiple explainability tools, many of which are continuing to evolve as data science techniques improve. For example, lenders have increasingly shifted to approaches such as Shapley values (including multiple Shapley value estimators provided in a popular open-source software library called "SHAP") to assess the importance of particular features. They often also

use various graphical analyses such as partial dependence plots (PDPs) and individual conditional expectation (ICE) plots to understand whether the associations between input features and default risk are intuitive and easily justified. See Appendix D for brief overviews of these methodologies.

Similar to the choices faced in deciding which machine learning technique to use in developing a new model, deciding which post hoc explainability tools to use and how to deploy them depends on a range of considerations about the accuracy of the method, various technical considerations regarding its application to the specific type of model, the lender's general technical expertise and computing capacity, and other operational factors. Previous research by FinRegLab with Stanford economists suggests that there can be significant differences in performance, and that some but not all tools can reliably identify features that are important to different models' risk predictions for individual consumers.[42] For example, while many tools relying on SHAP feature importance measurements performed well in the study, other tools 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.[43]

Making informed and reasonable choices about these issues is a key element of compliance, but it is important to understand that some differences in accuracy and consistency are normal, even with top-performing tools. This happens because all models, both traditional and machine learning, that seek to extract relationships from data will produce slightly different results depending upon the specific data samples, feature reduction techniques, hyperparameter optimization approaches, and more. Similarly, different choices must be made when using explainability methods. Factors such as the type of model, the size of the data sample used for comparison, the group chosen as a benchmark, and other technical details can result in some variation in outputs between models and lenders. In some cases variation is to be expected because of the type of model used (for instance, gradient boosted trees are often configured to randomly select a subset of the variables that will be used in initial rounds of model estimation). Likewise, the impact of any one feature is likely to be smaller than in the context of a traditional regression model, in which the number of features is typically limited. Because machine learning models may include many variations on a higher level concept, the impact of such a concept may be spread across several variables, further leading to inconsistencies when results are only considered at the most granular, individual feature level.

4.2.2 Aggregation processes

Even with traditional models and data sources, some lenders engage in aggregation processes where they group specific individual data attributes into broader categories or typologies for disclosure to applicants at a less technical level that may have greater utility to consumers. For example, if a determination is made that consumer understanding will be enhanced by combining "delinquency on all trades in last 6 months" and "delinquency on all trades currently" into "recent delinquencies on all trades," then the individual factors would be summed up accordingly as part of the aggregation process.

In the machine learning context, data science research is continuing into the extent to which grouping similar variables together can also increase the consistency of post hoc explanatory tools and make their outputs easier to interpret. As the technology continues to evolve, lenders vary somewhat in their thinking about the use of aggregation as a strategy for dealing with the challenges of many specific, similar variables in more complex models. Different lenders take different approaches, depending on how they interpret legal requirements regarding specificity and accuracy and whether and how they weigh useability in crafting particular reason descriptions. Wherever lenders land in making these decisions, however, there is broad acknowledgment that irresponsible use of aggregation risks obscuring information about the overall operation of the model as well as compliance concerns.

Accordingly, it is important to be thoughtful in choosing and validating the taxonomy of reason codes and the aggregation processes used to populate them when implementing machine learning models. Doing so allows lenders to address compliance concerns and to make the disclosures more useful to recipients. For example, if a lender decides that it is more useful to applicants to differentiate between recent and older credit delinquencies or delinquencies on different credit products, this can inform how they construct their taxonomy of reason codes.

4.2.3 Disclosure crafting

Disclosure language is often crafted simultaneously with the aggregation process. At a bare minimum, this generally involves translating raw feature labels used in the modeling process into more everyday language, but it can also involve more complicated decisions about whether and how to provide more useful insights to consumers. As noted above, longstanding regulatory guidance emphasizes disclosures need not explain why or how the particular factor affected the lender's decision-making, but adding such language can be helpful to making the disclosure more helpful to consumers as they seek to identify and correct errors in their credit reports or increase the likelihood of future approvals. As an added benefit, clarity and useability in notice language can reduce the number of complaints that result from adverse actions.

To provide more helpful disclosures, some lenders create variations of input variables that in turn yield more specific adverse action reasons. For example, they may create separate variables to differentiate between consumers with no balances versus low balances on particular kinds of credit accounts (e.g., differentiating between bankcard accounts as compared to all credit accounts), or create reasons that provide directional information (e.g., too few credit lines vs. too many credit lines). Such information can be particularly helpful to customers when lenders allow their machine learning models to recognize nonlinear or nonmonotonic relationships, since simply saying that "number of credit lines" was a principal reason does not give a consumer a directional sense of how to improve their likelihood of approval in the future. However, providing more tailored descriptions to subgroups of customers requires careful attention to internal processes to ensure that the correct language is being delivered to the relevant applicants.

4.2.4 Testing and validation programs

Because of the complexity of providing adverse action notices associated with adopting ML models in underwriting, it is important to consider how processes will change as the initial model development process is playing out and to develop a comprehensive testing program to validate and monitor disclosure production processes over time.

Such programs typically include both pre-production and post-production elements. On the front end, for example, lenders frequently validate feature importance methodologies and aggregation processes considering accuracy and specificity requirements. As discussed below, FinRegLab's research in collaboration with Stanford economists is one source that analyzes several explainability techniques using strategies such as perturbing the identified input data and measuring the direction and scale of impact on consumers' credit scores to gauge the methodologies' fidelity (accuracy), consistency, and usability when applied to a range of model types for generating adverse action disclosures and other purposes. However, other lenders may use other approaches. Some lenders may also engage in edge-case simulations and test specific disclosure language with consumers before deploying a new ML model or finalizing changes to their adverse action compliance processes.

Methods of post-production monitoring also vary between institutions. Some institutions implement monitoring processes that may evaluate denial data to determine whether similar applicants are receiving the same reasons for similar denials or whether there are shifts in denial reasons due to marketing programs, other changes in the population of applicants, changes in credit policy, or model changes.

4.3 Methodologies and practices for generating adverse action reasons

Since lenders vary widely in their business, modeling, and compliance practices, there is no one-size-fits-all approach to responsibly implementing adverse action notices. However, implementing such methodologies in the context of machine learning models generally includes the following steps:

1. During the ML model development process, select the explainability method best suited to the selected model type, select the reference group of applicants that will be used for benchmarking purposes when applying the explainability method, and determine aggregation approaches including mapping individual input variables to draft reason codes.
2. Apply the explainability methodologies, aggregation approaches, and other supplemental analytic techniques (such as partial dependence plots and ICE plots), to samples used during the model development and validation processes to analyze model behavior and to generate test disclosures for individual sample applicants.
3. Analyze the model analyses and sample notices to assess accuracy and specificity, as well as broader useability (where it is considered). Review the results of the analysis with consumer compliance and/or legal. Make adjustments to the model, explainability method, aggregation method, or other elements of the adverse action program as warranted before final implementation.

These steps are discussed in greater detail below, with the primary focus on elements that are different or more complicated for machine learning models as compared to traditional regression models. Lenders may make a variety of choices as to which models they develop (e.g., XGBoost, neural networks), what explainability tools they use to analyze the models, and how they validate the use of those tools. What is most important is to adopt a methodical approach to analysis and systems implementation. Section 4.4 contains an additional discussion of issues concerning the specificity and useability of the language ultimately provided to disclosure recipients.

4.3.1 Choice of explainability methodology and reference groups

For lenders that choose to adopt more complex machine learning models and rely on post hoc explainability techniques to help generate their adverse action notices, a critical first question is what technique to use. Other lenders that rely primarily on architecture constraints may still choose to use explainability techniques for some analytical purposes during the development process.

Although some computer science model explainability researchers have pursued general purpose, model-agnostic explainability methods,[44] practitioners often choose explainability techniques that are tailored to the particular machine learning algorithm and model form selected. For example, a SHAP variation called the Interventional Tree Explainer methodology is frequently applied to XGBoost models because research suggests that it produces more accurate explanations for tree-based models than computationally faster default methodologies.[45] Some lenders have also begun using Owen values (coalitional Shapley values) as a way to address the presence of similar or correlated variables in

machine learning models.[46] Practitioners that choose to develop neural network models tend to use the Integrated Gradients method to quantify the marginal contribution of an input feature. Integrated Gradients estimates Aumann-Shapley values, which are more appropriate for smooth functions like neural networks than the original Shapley values.[47]

Continuing to survey the literature for emerging methods that may be found to perform even better helps lenders stay abreast of new developments in data science that can enhance the accuracy of explainability outputs.[48] However, internal validation processes are also critically important, since even the most widely used tools such as Shapley values include a broad range of options that are not equally suited for all models and given tasks, and must be interpreted in specific context using domain expertise.[49] It is important to consider how the combination of model, explainability method, and mapping exercises work together in addressing concerns about the accuracy and specificity of the resulting notices. Computational constraints may also play a role in the process of selecting an appropriate method. By thoroughly testing the selected approach, lenders can empirically justify the choice of explainability method and prove that it produces accurate outputs.

With regard to selecting the reference population, explainability methods in the adverse action context are used to compute the average marginal contribution of a variable to a difference in model output between the reference population and a given individual consumer. Some explainability methods, such as Integrated Gradients for neural networks, require the explicit selection of a reference population. However, many SHAP explainers rely by default on reference samples from the model development population that are random or unweighted. Because adverse action reasons are meant to explain why individual applications were denied, many lenders override the default and select the reference population for their explainability tools from the population of approved applicants. However, there is some diversity in approaches, including using a 'through the door' population, and lenders may use other reference groups when using explainability tools for model risk analyses or other purposes as discussed elsewhere in this framework.

Having to designate a reference population for purposes of adverse action compliance is not unique to machine learning models, and as noted above, existing guidance provides flexibility on this point by listing two sample methodologies that benchmark an individual applicant against either (1) applicants whose total score was at or slightly above the minimum passing score to identity and disclose the factors for which the individual was the furthest below the average for that comparison group; or (2) the average for all applicants and disclosing the factors on which the individual performed least well as compared to that average. The guidance also states that any similar method may be acceptable.[50] A third approach used by some lenders is to focus on the factors for which the applicant fell furthest below the maximum achievable score under the model, which can be advantageous because it does not require adjustment in the event that score cut offs change in response to macroeconomic or competitive conditions.

4.3.2 Deciding aggregation approaches

The most common approach to aggregation appears to be to group input variables into broader categories after applying feature importance tools to underwriting models to measure marginal contributions at the individual feature level. However, whenever aggregation occurs in the process it is important to use care in mapping all of the input variables used in the model to the hierarchy of broader groupings. With very rare exceptions,[51] each variable must map to no more than one reason, but many variables could map to the same reason depending on how lenders decide to group similar or correlated features and what communications principles they establish regarding the useability of particular information to consumers (where it is considered). The mapping process is not

unique to machine learning models but for the reasons discussed in Section 4.2, it can have significant implications for the amount and nature of the information conveyed to disclosure recipients about ML models that involve large numbers of features since only the top 4 to 5 reasons typically appear on the actual disclosure.

Deciding whether and when to group features that are variations on a given concept or theme or that may be correlated or interact with other features requires balancing clarity, specificity, and consumer understanding to determine what level of detail is required by law and useful to consumers. The following steps can be helpful in this process:

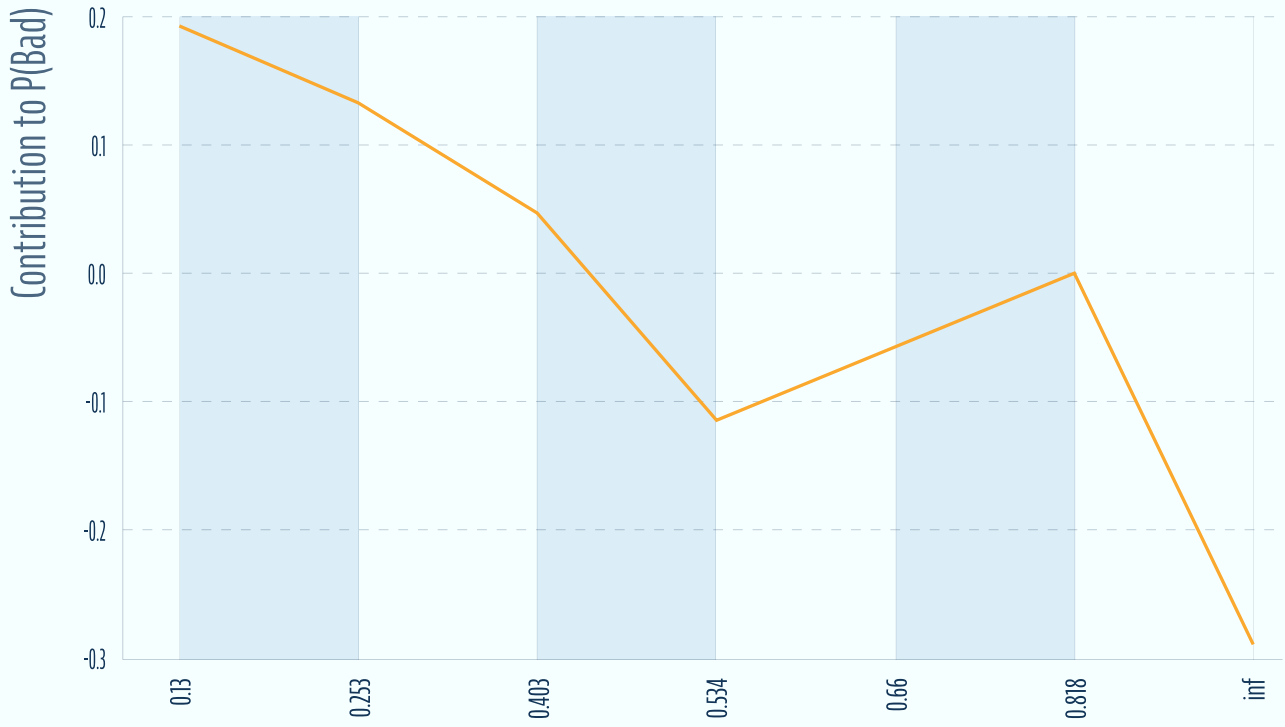
- First, review the feature engineering logic that developers have made in building the model to determine logical groupings. Many modelers generate second-order features, such as time-based lags and other statistics related to a given primary input feature, which some lenders may consider good candidates for grouping. For example, where a model considers several metrics concerning the consumer's average, maximum, and minimum credit card balances, lenders may use a more general descriptive category such as "high credit card balances." Similarly, measures of credit card balance values for various time periods could all be grouped together into a single key factor, "recent credit card balances," although as discussed above lenders may make different judgments as to what level of granularity with regard to time periods is most useful to consumers.
- Next, for those variables that remain ungrouped, some lenders identify correlated features using metrics such as the Pearson correlation as a first step to considering whether similar or highly correlated features may be logical to aggregate. However, others do not use this step, and those who do emphasize that it is only a catalyst for discussion and further analysis to determine where aggregating features is consistent with (and conducive) to legal compliance and promoting broader useability for consumers. The mere fact that something is correlated is not sufficient to justify aggregation.
- Many lenders also find it useful to articulate standing principles about what level of detail is likely to be helpful to consumers to assist in guiding decisions about aggregation, along with domain knowledge and business logic. Some stakeholders describe the process of making aggregation decisions as essentially working backward from such broader communications principles to determine where it is appropriate to group specific variables under a common reason code.

Depending on how lenders evaluate these various considerations, some may group features that are deemed to represent the same underlying issue (e.g., multiple features related to credit usage) or one feature is derived from another (e.g., variance or rate of change of balances). It is important to consider whether the selected reason category and associated wording are labeled in a way that reflects what is actually going on in the model. This is especially a concern when substantially different features are being combined and when the components of the combined variable come from different sources. For example, although debt-to-income ratios are derived from two separate features and are used in some traditional underwriting models where there is a strong emphasis on parsimonious models, in the machine learning context lenders may decide to avoid creating such compound features to simplify the disclosure process and differentiate between data coming from different sources (e.g., credit bureau records for debt and tax records or bank account statements for income).

4.3.3 Analyzing whether adjustments are warranted

The next stage is to use the sample population to determine whether refinements to the model, methodologies, aggregation groupings, or reason descriptions are warranted. One method of analyzing model behavior is to use partial dependence plots to show the relationship between individual input variable values and the model prediction. For this specific context, calculating PDPs using actual observations and the model explainability outputs is generally recommended to observe the direction of the trend of the model prediction associated with a given variable.[52] The figure below shows such a plot.

FIGURE 1 SAMPLE PARTIAL DEPENDENCE PLOT



Here, the partial dependence plot shows that the model assigns a higher probability of default ("bad") to lower values of the variable, while higher values are associated with lower default risk. A non-monotonic trend is observed, i.e., it seems that some higher values (in the 0.818 bin) are neutral not negative. Identifying this pattern allows lenders to evaluate whether the observed relationship is intuitive, whether the draft reason language should be adjusted to address accuracy, specificity, or useability concerns, or whether to adjust the model so that it treats the relationship as monotonic. Reviewing PDPs, ICE plots, or using other techniques to assess how the model behaves for values of each variable helps to ensure that issues are considered, resolved, and documented systematically.

When a non-monotonic trend is observed some lenders opt to further treat the model so that a directional reason can be provided. For example, in the above plot, a monotonic constraint could be applied so that larger values of the variable always yield a lower probability of default. The ability to impose monotonic constraints can help ensure the model outputs are intuitive and that the adverse action reasons are accurate and specific. In other cases, e.g., when the observed nonmonotonic trend is intuitive, the variable can be treated to generate more specific and accurate reasons. Consider, for

example, a model that treats both too little credit and too much credit as indicative of higher risk. In this case, a U-shaped trend will be observed in the partial dependence plot. The relationship between the variable total credit and risk is intuitive—too little credit may indicate someone with less history and therefore higher risk, but too much credit can also be high risk especially when combined with high utilization rates. In this case imposing a monotonic constraint could degrade the predictive performance of the model by missing nuances in one of those two groups. Depending on the dominant relationship for the variable, developers may use another benefitting variable to capture the opposite trend so that directional reasons can be provided to consumers in both situations.

While directional explanations can often promote useability, however, they are not required and it is important if they are provided that they are accurate. This can be more complicated with non-monotonic, multivariate, and other more complex relationships. For example, models often learn from the data that low credit usage is also associated with higher default risk, and that the lowest risk borrowers have credit usage in a middle "goldilocks" zone. As such, the natural language explanation "credit usage too high" would be inaccurate for applicants for which the model assigned greater incremental risk because of their low utilization. By carefully reviewing model behavior, grouping features, and crafting specific and accurate reasons, lenders can provide accurate and meaningful feedback to consumers while maintaining regulatory compliance. Factors concerning the wording of the disclosures are discussed further in Section 4.4 below.

4.3.4 Generating individual notices

Once the major parameters such as explainability methods, reference sets, and variable aggregation methods have been set, and initial assessments have been conducted to determine whether they require adjustment, lenders may generate individual notices for members of the training or hold out samples for purposes of another round of validation and calibration as discussed in Section 4.3.5.

One approach for the individual disclosure generation is the following: For each observation, use the explainer to compute the average marginal contribution of each variable in the model to the difference in score between the observation and the reference population. Where aggregation rules have been developed, apply them as warranted to group by reason and sum the contributions by reason (note: this summing step is valid because the Shapley values have the mathematical property that they can be summed and retain their meaning; other explainability methods may not have such a property). Then sort and select the top 5 reasons.

For example, consider an applicant below where the following attributes are identified by a Shapley values-based tool as having the biggest marginal contribution to the difference in score between the applicant and the reference group of approved borrowers.

ATTRIBUTE NAME	ATTRIBUTE VALUE	SHAPLEY VALUE	ADVERSE ACTION REASON
DELINQUENCIES ON BANK CARDS IN LAST 30 DAYS	2	0.2	Too many recent delinquencies on credit card accounts
DELINQUENCIES ON RETAIL CARDS IN THE LAST 30 DAYS	1	0.2	Too many recent delinquencies on credit card accounts
DELINQUENCIES ON BANK CARDS IN THE LAST 90 DAYS	6	0.15	Too many recent delinquencies on credit card accounts
DELINQUENCIES ON RETAIL CARDS IN THE LAST 90 DAYS	3	0.15	Too many recent delinquencies on credit card accounts
REVOLVING UTILIZATION TODAY	101%	0.1	Credit utilization too high
TOTAL OUTSTANDING UNSECURED CREDIT	$124,057.99	0.1	Unsecured balances too high
BANKRUPTCIES IN THE LAST 5 YEARS	1	0.1	Recent bankruptcies
AVERAGE REVOLVING UTILIZATION LAST 6 MONTHS	99%	0.1	Credit utilization too high

If the above output were not grouped or aggregated, the 4 reasons provided on the adverse notice would all be variations on the "credit card delinquencies" theme. The consumer would not be alerted to the other factors affecting their denial.

Instead, if the delinquency-related features are grouped and the Shapley values are summed, the results change to this:

ADVERSE ACTION REASON	SUM OF SHAPLEY VALUE WITHIN THE GROUP
TOO MANY RECENT DELINQUENCIES ON CREDIT CARD ACCOUNTS	0.7
CREDIT UTILIZATION TOO HIGH	0.2
UNSECURED BALANCES TOO HIGH	0.1
RECENT BANKRUPTCIES	0.1

Consistent with compliance considerations and useability principles, many lenders use this or other reasonable aggregation approaches in an effort to provide notices that help consumers understand more of the factors that influenced the underwriting process and the follow up actions that could improve their chances of future approval.

4.3.5 Testing and validating the adverse action reason generation process

At a high level, testing is designed to determine whether changes to the identified features have the expected effect on risk predictions for individual applicants, both in terms of direction and magnitude. Such tests can be helpful to assess the accuracy of the overall disclosure generation process, and can be performed with sample members in the initial pre-implementation phase as well as on a periodic basis after implementation to confirm that systems are continuing to work as expected.

One potential process for conducting these kinds of assessments is inspired by the method outlined in the FinRegLab et al. study, revised to account for the grouping process described above that is more typical of industry practice.

First, for each observation, generate the individual's model score and reason codes using the method described above. Then select the sample members that fall below the approval threshold. These are the applicants that would be denied. Next, for each denied applicant, and for each adverse action reason, perturb or modify the variables corresponding to the adverse action reason in ways that would be expected to lower default risk. To do this, select a random applicant from the sample of best applicants. Then replace the values associated with the adverse action reason from the denied applicant with the values from the selected best applicant. Next, re-score the modified applicant, verify the model score improved, and that the reason code corresponding to the modified variables lowered in rank. Multiple good applicants can be sampled and an average difference associated with multiple substitutions may be recorded to provide a more robust measurement of the score difference, with a confidence interval. The difference in score due to the perturbations (and the confidence interval) should be recorded for later comparison.

Note that where a lender has aggregated related variables under a single reason code as described above, all of the associated values are modified at the same time. This is a more realistic analysis than modifying one variable at a time in isolation, since it helps to account for correlated or logically related variables. In addition, substitutions are drawn from an actual observation (a real applicant), which helps to ensure that relationships between the variables (e.g., total debt, income, and debt-to-income ratio) are preserved during the testing process.

For example, consider the same applicant described on the previous page, who has substantial credit card delinquencies, high utilization rates, and a recent bankruptcy. The perturbation test would create a synthetic version of this applicant to verify that if they had fewer delinquencies, that reason would no longer show up on the applicant's notice. Using data from a randomly selected "good" applicant that has no delinquencies, the synthetic results would be the following:

ATTRIBUTE NAME	ATTRIBUTE VALUE	SHAPLEY VALUE	ADVERSE ACTION REASON
DELINQUENCIES ON BANK CARDS IN LAST 30 DAYS	0	0	Too many recent delinquencies on credit card accounts
DELINQUENCIES ON RETAIL CARDS IN THE LAST 30 DAYS	0	0	Too many recent delinquencies on credit card accounts
DELINQUENCIES ON BANK CARDS IN THE LAST 90 DAYS	0	0	Too many recent delinquencies on credit card accounts
DELINQUENCIES ON RETAIL CARDS IN THE LAST 90 DAYS	0	0	Too many recent delinquencies on credit card accounts
REVOLVING UTILIZATION TODAY	101%	0.13	Credit utilization too high
TOTAL OUTSTANDING UNSECURED CREDIT	$124,057.99	0.12	Unsecured balances too high
BANKRUPTCIES IN THE LAST 5 YEARS	1	0.11	Recent bankruptcies
AVERAGE REVOLVING UTILIZATION LAST 6 MONTHS	99%	0.1	Credit utilization too high
(OTHER ATTRIBUTES IN DESCENDING ORDER)			

The Shapley values estimated for "Too many recent delinquencies on credit card accounts" are all 0, and as such, that reason would not appear on the adverse action notice. Instead, the remaining values would be left with the following reasons.

ADVERSE ACTION REASON	SUM OF SHAPLEY VALUE WITHIN THE GROUP
CREDIT UTILIZATION TOO HIGH	0.23
UNSECURED BALANCES TOO HIGH	0.12
RECENT BANKRUPTCIES	0.11
(OTHER ATTRIBUTES IN DESCENDING ORDER)	

Because the "too many recent delinquencies …" reason disappeared from the notice and because the overall default prediction decreased after substituting in the synthetic data, this round of the perturbation test helps to validate the "too many delinquencies on credit card accounts" reason. This process would continue for each of the reasons that appear in the test disclosures. This process would then be applied to a sufficiently large sample of applicants. In this way, lenders can test the accuracy of the selected model explainability method and adverse action processes. While this method does not provide perfect ground truth since higher risk levels may be expressed across multiple variables at the same time and variables can be related even when they are not grouped, such tests can help to provide greater confidence where the results behave as expected.

Initial analyses of this type help to validate that the provided reasons correspond to factors that improve an applicant's score. The perturbation method can also be used to probe whether there are other reasons that would have improved the applicant's score even more than the ones identified as having the biggest marginal impact by the explainability tool. To do this, perform perturbations associated with reasons lower down on the list of reasons (not the top 4 or 5 provided by the chosen adverse action method, but other reasons that are ranked lower). Assess whether the average score improvement associated with perturbations corresponding to those other reasons is lower than the score improvement associated with perturbations corresponding to the top 4 or 5 reasons.

Lenders may engage in additional tests either prior to deployment or periodically after deployment, such as simulating edge cases, analyzing denial data to assess whether similar applicants are receiving similar reason codes for similar denials, and establishing processes for exception reporting and remediation processes where problems are identified during internal review or by reviewing complaints from customers.

4.4 Wording for specificity and useability

In addition to the general process steps outlined above, many lenders are thinking more deeply about specificity and useability of their disclosure language in light of the transition to machine learning models and the adoption of new data sources where applicable. Improving disclosure language to make it more useful to recipients is not unique to the ML context, but both the transition to ML models and shifts in underlying data elements have focused attention on the purposes and value of the disclosure regime and the processes by which lenders structure notice publication and compliance checks.

Lenders often start their analyses by focusing on the level of detail provided by the reason codes. Although broader descriptions may be helpful for consumers who lack much credit history, using generic language such as "unacceptable/poor credit history" or "insufficient account activity" are less useful for consumers with more substantial credit histories in understanding what factors in their credit reports may be in error or how their past behavior has raised concerns (e.g., delinquencies, utilization, length of credit history, or whether the account activity in question concerns credit accounts or other types of accounts). At the same time, using jargon or highlighting multiple specific reasons that are closely related can make the disclosures less understandable and useful overall (e.g., "Score component PQR below threshold" or listing four minor variations relating to credit delinquencies—delinquencies in the last 30 days, delinquencies in the last 90 days, delinquencies in the last 120 days, delinquencies in the last year—all relating to a single delinquency 14 days ago).

However, there is little affirmative guidance to help lenders determine how best to calibrate specificity and useability, particularly when considering not only what information is useful for error correction but also for broader customer education goals. The decades-old list of sample reasons provided in an appendix to Regulation B provides limited guidance because it is quite short and quite high level, which can lead to consumer confusion and increased incidence of complaints. Some lenders may try to benchmark their practices against peer institutions or to conduct customer testing, though such programs appear to be relatively rare. Over the last several years, some lenders have urged the Consumer Financial Protection Bureau to expand the list of sample reasons, both to address more types of model inputs and to provide more positive benchmarks with regard to specificity, useability, and general good practice. Others emphasize the need for continuing flexibility, given that lenders use different models, data sources and individual attributes, and methodologies for generating adverse action disclosures. Given the continuing evolution of relevant data sources and other aspects of credit underwriting, lenders do not want standardization initiatives to chill their ability to innovate particularly in adopting new data sources that might improve the predictiveness and inclusiveness of their models.

Lenders who choose to build useability into their program goals and/or processes are using a variety of strategies to increase the clarity, relevance, and usefulness of their disclosures, including:

- **Using customer-friendly language:** Reviewing the disclosure language to screen for technical jargon and the risk of confusion can help to increase useability to recipients. For instance, some lenders consciously seek to replace more technical terms (e.g., "high credit

utilization") with explanations like "You are using too much of your available credit." Others favor terminology that customers are likely to be familiar with from other financial contexts.

- **Considering the wording of reason codes that involve compound elements:** Where a lender chooses to use reason codes that focus on a ratio or the interaction between multiple features, it can be helpful to consider whether to describe which specific elements are having the largest impact. For example, if an applicant's debt-to-income ratio is a principal reason for rejection, it may be more useful to disclosure recipients to indicate whether high debts or low income levels are primarily driving the result than to simply report "unacceptable debt-to-income ratio" as the reason description. More specificity can also potentially be helpful for features that combine information sources (e.g., credit bureau data and other sources of information about income or assets), since recipients may need to follow up separately about any error corrections. However, as noted above some lenders decide not to create compound features to simplify disclosure processes.
- **Considering whether to provide directional context for particular reason codes particularly if allowing nonlinear or nonmonotonic relationships to be captured by a model:** On a similar note, where the lender is using a model that treats a particular variable as adding to or reducing default risk depending on its value, it may be more helpful to recipients to know if their level is too high or too low. However, for the reasons discussed in Section 4.3.3, more tailored disclosures require careful thought to ensure that the relevant customers are getting disclosures that correspond to their circumstances. Again, some lenders decide to create multiple variables or not to allow non-monotonic relationships in their models as a way to simplify disclosure processes.
- **Assessing whether customers will be able to understand from the draft language how to improve their chances of approval in the future:** As discussed in Section 4.1.2, some factors that constitute a principal reason for an individual applicant cannot be substantially changed through action, at least in the short term. These include items such as the industry of a small business applicant, the existence of a recent bankruptcy, or the length of a consumer's credit history. But even in those cases it can still be helpful to ask whether the draft reason code is clear enough that recipients would be able to identify if there is an error in the underlying data that may warrant follow up and understand that continuing to repeat a behavior that is associated with higher credit defaults is likely to worsen the likelihood of success. Lenders see this as a key tool to help manage the volume of complaints.
- **Include information about the underlying data sources and contact information for assistance:** Some of this information is specifically required by regulation particularly in the context of the FCRA's disclosure requirements for decisions based on credit bureau data. However, even where source and contact information are not specifically mandated, such elements can be helpful for customers who want to follow up on potential data corrections or ask follow up questions. Examples of such information could include referencing the data source and date (e.g., application submitted by the customer, credit bureaus, etc.), information on reaching customer service through multiple channels and on accessing credit counseling, and general encouragement to review credit reports periodically.
- **Regularly test and update reasons:** Periodic reviews of both customer feedback and existing disclosure language can be helpful to identify shifts in models, data, or the efficacy of particular wording over time. Again, while not common, affirmatively using focus groups or other testing could be one way to seek feedback.

5. OTHER MODEL RISK MANAGEMENT CONSIDERATIONS

The model risk management framework provides a flexible, principles-based framework to evaluate the fitness for use of quantitative models—including but not limited to those used for underwriting and those built with machine learning techniques—and whether model risk is being managed appropriately in light of the size and complexity of the bank, the importance of the use case, the potential effects of model malfunction, and other considerations. It addresses both risk assessment and validation prior to adopting new models and the implementation of controls and monitoring plans during and after deployment.[53]

Federal banking regulators' foundational MRM guidance dates from 2011 to 2017 and does not expressly address the use of machine learning or other forms of artificial intelligence.[54] Because some modeling elements take on greater importance in machine learning, banks that are using such models have adjusted their practices over time to assess and manage various potential risks in the machine learning context using the guidance as a framework. For credit underwriting models, these practices focus on such topics as data selection and management, model transparency and complexity, managing risk of overfitting, and consumer protection compliance (including the requirements addressed in prior sections). These practices also include the validation of the modeling elements described in Section 3. The appropriate selection and deployment of post hoc explainability tools has important implications for many of these other topics, to the extent that a few banks treat tool selection itself as subject to MRM processes. Others do not, but still emphasize the importance of thoughtful choice and deployment in the course of conducting other MRM reviews.

A detailed treatment of MRM processes for credit underwriting models is beyond the scope of this framework. After a brief high-level overview, this section highlights issues and process adjustments that have evolved more recently in the context of adopting machine learning models. It draws on the experiences of large institutions—both in how they have built out model risk management functions over time and in how they have approached building and managing ML underwriting models in house—but may be helpful to other stakeholders to understand how issues, practices, and systems are evolving and to inform conversations about vendor practices. Managing risk with regard to third party models is discussed further in Section 6.

5.1 Model risk compliance and policy objectives

5.1.1 Guidance expectations

The MRM guidance is designed to protect the safety and soundness of the banking system. The guidance defines a model to include any "quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates."[55] Model risk is defined as "the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports," such as financial loss or legal liability.[56]

Banks are expected to identify potential sources of such risk, assess their magnitude, and mitigate appropriately. Expectations are calibrated to the degree of risk posed by the particular use case to the particular institution, with credit underwriting often considered to be among the highest risk activities. The guidance emphasizes both the importance of "effective challenge" of models by independent model validators and ongoing monitoring to help ensure that the institution is not exposing itself to unnecessary risk because of errors in the initial development process or changes in model use or the external environment over time. Where there is higher uncertainty about a particular model, the model is relatively complex, or the model is particularly important to a bank's portfolio, risk management programs will generally require heightened scrutiny before approving that model for use and more vigorous oversight of its ongoing performance.[57]

5.1.2 Sequencing of compliance processes

Choreography between first line development teams and second line teams that are responsible for risk management validation are critical to MRM processes. Model developers structure their processes from the beginning with an eye toward managing model risk concerns, but independent testing and validation of their activities and judgments is also considered an essential component of MRM practice for any models used in impactful business decisions. These processes are essential both because the risk taken with models is often large relative to the company's risk appetite and because complex systems often fail in unexpected ways. Independent review is a proven tactic for increasing the resilience of complex systems, similar to red teaming in software engineering and cyber security.

Accordingly, documentation of the activities and decisions made by model developers in exploratory data analysis (EDA) and model training and testing is used to facilitate second-line review once a "champion" model has been selected to proceed to MRM validation before potential deployment. As discussed in further detail in Section 5.3.3, steps that are important for MRM validation also intersect with and help to facilitate compliance processes for specific consumer protection requirements such as generating adverse action notices.

Model validation formally begins once a model has been developed and is nearing or has reached its completion stage. Significant validation findings identified at this stage, e.g. disagreement from model validator on the choice of modeling data, will require major rework by model developers and cause significant set-backs on the model deployment timeline. Therefore many banks adopt some form of pre-validation review process to allow model validators to provide early validation feedbacks on key model design choices by model developers, while the model is being developed. This proactive approach helps identify and mitigate potential issues early, making the overall process more efficient and reducing the likelihood of significant rework. While early interaction is valuable, it is crucial to maintain the independence of the formal validation process.

MRM processes also include review of monitoring plans and their execution. Practical considerations for monitoring plans of ML models are discussed in Section 5.4.2.

5.2 Overview of choices and challenges with machine learning models

In many aspects applying MRM principles to machine learning underwriting models is not fundamentally different from applying them to more traditional regression approaches, since the framework applies equally to both contexts to guard against the same foundational concerns. However, the methodologies used in MRM analyses for traditional models have become widely familiar and accepted over time, while adoption of machine learning models typically requires lenders to take a fresh look at their processes and make decisions about how to approach the choice of the machine learning technique, model training protocols, diagnostic tools, explainability, and other issues.

Beyond the specific compliance issues discussed in previous sections, MRM processes as applied to machine learning models tend to focus in particular on probing the potential tradeoffs between model performance, stability, complexity, and explainability. To the extent that developers decide to allow their ML models to map more complex relationships in the data, incorporate larger numbers of data elements, or vary from traditional models in other ways that tend to increase explainability or operational challenges, MRM processes frequently probe the justifications and tradeoffs that come with those development decisions. Some banks also require the development of more traditional models to use as benchmarks for the ML champion model to assess these tradeoffs during validation.[58] Much of the focus is thus on whether the selection of more complex ML models is justified by benefits over simpler approaches.

5.3 Validation processes for new models

MRM validation processes generally are divided into two primary areas of focus: (1) evaluating a model's conceptual soundness and fitness for use; and (2) conducting outcomes assessments to evaluate the model's performance under various scenarios to evaluate its likely accuracy, stability, and robustness under real-world conditions. While the latter analyses tend to be largely quantitative in nature, conceptual soundness analyses may involve both qualitative and quantitative activities. Both are described further below.

MRM also places heavy emphasis on data governance and quality across the development lifecycle and on ongoing monitoring and evaluation. For the latter topic, second line reviewers typically evaluate model monitoring plans either at the same time that they are evaluating the champion model itself or in the interim period between approval and actual deployment. The plans are intended to ensure that the model is operated safely and within the firm's risk tolerance by articulating key risks, defining metrics and processes for monitoring those risks, and describing steps to be taken when, for example, substantial changes in performance or data or score distributions occur. Issues that are helpful to consider in building monitoring plans for ML models are discussed in further depth in Section 5.4.2.

5.3.1 Conceptual soundness and fitness for use

At a broad level, conceptual soundness analyses focus on assessing the quality of the model's design and construction in light of its intended use and business objectives. The assessment process involves detailed review of developers' choices concerning data sources, individual inputs elements,

modeling methodologies, and application of diagnostic tools to determine whether exercises of judgment were "well informed, carefully considered, and consistent with published research and with sound industry practice."[59] Developers need to adequately document this information, which is submitted for validation, and typically need to be available to answer subsequent questions that arise during the validation process.

Conceptual soundness reviews for machine learning models cover similar terrain to reviews of traditional models—including selection and treatment of the data, articulation of the outcome the model is designed to predict and its relationship to that data, and model reproducibility. In the context of machine learning models, additional scrutiny and care may be warranted, especially for higher risk use cases like underwriting. Best practices specific to ML models must be adopted to demonstrate to MRM reviewers that model risks have been managed. Typically, the steps described in Section 3 are subject to validation; examples include the following inquiries:

- **Suitability of the data for modeling:** This step is similar to what is done for traditional models but can be more extensive because machine learning models tend to use more data. In addition, while ML models can be good at handling noisy data, care must be taken to ensure the model is not overfit (as discussed further in Section 5.3.2), that the model algorithm is handling data in a desirable manner, and that there is sufficient data both for optimization tasks during model development such as hyperparameter tuning and for the various tests required to validate the model after it is estimated. As such, practitioners must take care that there is enough representative data to create the required out-of-time and out-of-sample evaluation datasets at the start of the process, and that the samples are suitable for evaluation This can be particularly challenging when exogenous factors such as a global pandemic or economic shifts cause abnormalities in circumstances and behaviors that show up in the data. Understanding the degree to which these shifts are occurring helps inform the frequency and character of model monitoring required to manage model risk. Moreover, where lenders have chosen to incorporate more data into ML models—particularly data that goes beyond traditional credit report elements such as information about the applicant's income, cash flows, deal terms, and collateral value throughout the loan term—more attributes require more scrutiny.
- **Suitability of the input features and target variable:** Conceptual soundness analyses focus on whether and how key drivers of the predicted outcome are accounted for in the model inputs. For underwriting models, it is important to consider what information will be available to the model versus used in other aspects of the lender's credit policy and whether there is sufficient data to make reliable predictions. Conceptual soundness analyses also include consideration of whether the estimated relationships between the input features and the target variable reflects an underlying business/economic relationship and whether and how the shape of such relationship is restricted ex ante during the estimation process or evaluated post hoc.

  Many banks apply post hoc explanatory tools and a range of other data science techniques to scrutinize the relationships that the champion model has learned between the input features and the target variable, particularly where ML models deploy more complex architectures and/or larger numbers of features. Commonly used tools include Shapley values, partial dependence plots, and individual conditional expectation plots, as described in Appendix D. Lenders may also review the choice and application of tools to ensure that they are being appropriately applied in particular contexts in light of the nature of the data, model, and assumptions on which the tools operate with regard to topics such as correlations between features.

Many banks' conceptual soundness processes also include analyses of whether models are sufficiently parsimonious. Toward that end, various automated processes can be used to assess which variables are the most predictive, their degree of correlation, and whether including the variables in the model is worth tradeoffs as to complexity, stability and other factors. Lenders may also compare ML models to simpler benchmark models to assist in this analysis.

- **Model reproducibility:** A first step in model validation processes is to recreate the model from the data. While this is a basic task, it can be more complicated in the ML context. Machine learning modeling techniques rely even more heavily on stochastic methods than traditional models. As such, model developers must take care to ensure models can be reproduced from the data. In addition, model inference pipelines may rely on open source tools that are being actively developed and frequently updated, and the tools run on open source operating systems that may also require updates to patch security issues or to gain incremental functionality. Interdependence between tools can create complex upgrade scenarios impacting model outputs in production. Installing new tools or upgrading to new versions during model development or after model deployment can create conflicts that require additional updates that could make outputs generated in earlier model development steps impossible to reproduce, or outputs to shift in ways that impact business metrics. All of this complexity requires that care be taken by model developers to ensure models are reproducible prior to submitting them to MRM teams for validation. This requires careful crafting of policies and procedures governing coordination between data scientists and the IT teams that support them.
- **Suitability of modeling method and hyperparameter choices:** This step analyzes the developers' choices regarding model architecture, learning algorithm, and hyperparameter choices that impact the complexity and performance of the champion model. In the ML context, this may involve comparisons of the potential tradeoffs between the champion model and simpler or more traditional options. A more complex ML model will generally be proposed for use only where it yields greater predictive performance than the less complex model, although individual banks may make different judgments about the size of performance improvement necessary to justify use of the more complex model. This step also typically includes justifying the choice of hyperparameters for the selected ML model, for example, the depth and number of trees used in an XGBoost. Systematic exploration of hyperparameter choice and justification of the selected parameters are required.

5.3.2 Robustness, stability, and outcomes testing

The second major component of model validation focuses on various forms of quantitative testing to evaluate such risks as whether the model has been overfitted to the training data, how much performance changes if the model's underlying assumptions are altered, how well its performance will hold up in the face of changing external conditions, and other forms of outcomes testing. Much of this work focuses on gauging whether the model is stable as well as predictive, so that small changes of input data or configuration do not produce wildly different outcomes. This section does not provide a comprehensive overview of all aspects of these processes, but rather focuses on topics where the use of ML models has increased attention to certain risks and/or prompted shifts in lender practices.

5.3.2.1 Managing overfit and evaluating the model

Gauging whether a predictive model has been underfit or overfit to its training data is an important consideration for all models and validation processes, since both situations will lead to

poor performance in deployment. In the first, the model is so simplified that it is failing to pick up key predictive relationships. In the latter, the model has so closely mapped to the training data that it has incorporated "noisy" relationships that are not in fact predictive of the main target variable in deployment and will therefore reduce accuracy when the model is exposed to new data involving other consumers, time periods, or circumstances. These problems are manifestations of a broader challenge described as "generalization error," which focuses on whether a model is able to predict outcomes when it evaluates data beyond its training sample.

Because underfitting tends to lead to poor performance on both the initial dataset and separate samples of new data, it is generally easier to detect. Thus, overfitting tends to be a more significant concern during both model development and MRM validation. While overfitting can occur with both traditional and machine learning methods, it has historically been viewed as a particular concern for the latter because ML models can be so much more flexible and detailed in mapping nuanced relationships in the data. Over time, practitioners have built out a range of strategies to guard against and assess the risk of overfitting in ML models. These developments have led to increased confidence.

Validation analyses frequently involve applying the champion model to data that it has not previously encountered, from a time period that is subsequent to the period from which the model training data was derived. Performance evaluation on an out of time and out of sample test dataset is the gold standard for determining if the model is overfit. However, if overfitting is detected at the validation stage, it is often difficult and expensive to correct, since model developers tend to want to use as much data as possible from the most recent time period when estimating the model. Therefore, model developers typically use a range of approaches earlier in the development process to prevent overfitting on the test dataset:

- **Evaluating sample size and makeup:** Overfitting is more likely to occur where models are trained on relatively small or nonrepresentative datasets. While developers generally prefer to work with as much data as they can practicably secure and manage, analysis of the constraints of the resulting dataset may be helpful as noted above to identify potential limitations and risks.
- **Use of additional hold out samples:** Splitting the data used during the development process into separate training validation and test datasets allows developers to better manage overfitting risk. The test dataset is typically pulled from a different time period and is reserved for the end of the process to test how well the model performs on entirely unseen data. The validation dataset, in comparison, is used during model development to select hyperparameters and make choices about model architecture. When significant volumes of data are available, model developers have the opportunity during development and testing phases to use multiple tests using both hold out samples from the same time period and other time periods to gauge how well the model performs in many varied conditions. Model developers often perform checks to ensure the initial data splits have consistent distributions. This can help determine where to make the split between training, testing and validation samples, and also inform downstream process such as sample weighting.
- **Cross validation:** K-fold cross validation is similar in concept to using a hold out sample, in that it splits the training data into k number of equal buckets and then conducts the training process over successive rounds in which each bucket is used once as a testing sample. (For instance, the bucket that is used for testing in the first round becomes part of the training sample for the other rounds.) This takes more computational resources than holding out a single testing sample but allows all of the data to be used eventually to train the model.

- **Reducing inputs or layers and using regularization techniques:** Reducing the number of inputs or layers in a machine learning model tends to reduce complexity and the risk of overfitting, so model developers may experiment with a range of simpler options to evaluate the potential tradeoffs between model accuracy and complexity. Regularization techniques that effectively penalize models for being overly sensitive to outlier inputs can also reduce the risk of overfitting. Many machine learning models (XGBoost included) have regularization built in to their default loss functions, and the optimal degree of regularization is determined in the process of tuning hyperparameters.
- **Leaky feature detection:** Examination of the input data to ensure that it does not include information that would not be available to the model during deployment (sometimes called "leakage") is an important step to guarding against overfitting, since the model will not be able to detect those sets of relationships when exposed to real-world data. Flawed data processing can also result in leakage, and is particularly challenging if it occurs before splitting an initial dataset into separate training, testing, and validation sets because gauging overfitting risks becomes more difficult if all the samples are affected by the same flaw.

5.3.2.2 Other quantitative testing

Another overlapping area of focus in quantitative testing of champion models focuses on their likely behavior in the face of changes in the external environment after deployment, such as shifts in economic conditions, applicant pools, or consumer behavior. Using out-of-time validation samples is one way to test for such risks in addition to probing for overfitting, but lenders also frequently perturb the data in various ways to evaluate the model's sensitivity to particular changes in input features and scenarios and to identify what types of changes in the underlying data will cause the model to make an incorrect prediction.

Some lenders also engage in "swapset" analyses as part of model risk management validation to better understand the profiles of consumers that would be approved or rejected under a new champion model as compared to a baseline or incumbent models.

These analyses are not fundamentally different in the context of ML models, but some lenders use post hoc tools to analyze model outputs and why the champion model is treating particular populations differently. Thus, validation of post hoc techniques for particular purposes has implications for these tasks as discussed in the next subsection.

5.3.3 Interplay between model risk and other types of compliance

As reflected in this and prior sections, because MRM processes are principles based and comprehensive, they address model development, testing, and deployment processes and often intersect with processes for complying with other regulatory requirements and expectations. This section briefly highlights a few examples of the ways that these various compliance processes can intersect and reenforce each other.

A first key example is the testing, selection, and deployment of post hoc explainability techniques for multiple purposes over the model development and implementation life cycle. This is not only important to ensure reliable model validation under MRM guidance, but can also play important roles in other compliance processes, such as generating adverse action disclosures as discussed above. A few banks have treated explainability technique selection as subject to MRM processes in its own right, and many more have engaged in substantial testing and analysis to inform their decisions about which tools to deploy for particular models and particular compliance processes.

Some banks have also implemented automated validation systems to check the accuracy of model explainability tools for individual models in response to past agency guidance on topics such as adverse action notice generation.

As banks get further up the learning and implementation curve in working with ML models, the process of explainability technique selection becomes increasingly standardized because of cumulative experience and subject matter expertise. Thus, the choice of techniques that will be applied to a particular model that is under development is often determined at the time that other decisions are being made about the type of learning algorithm and model structure, although implementation details must also be documented and reviewed as part of MRM validation and other compliance processes.[60] Moreover, as data science continues to evolve, new developments often trigger additional follow up and testing to help banks determine whether further adjustments in their menus of approved techniques are warranted.

Another example of intersections between MRM and other compliance processes is the sequencing of different types of compliance reviews. For example, work to build and validate the processes for generating adverse action notices is generally more efficient if it is undertaken after a champion model has been identified and approved. However, there are times when those processes may prompt follow up work and changes to the champion model, as well as instances in which MRM validation processes prompt lenders to change their approaches to particular variables or other aspects of model operations (e.g., in working with a variable with unstable performance). Making these kinds of changes will reset or trigger supplemental MRM review, so it can become an iterative process.

5.4 Steps during and after deployment

5.4.1 Implementation processes and testing

Prior to beginning to use a model for actual production, all model implementations are tested according to plans that are typically required as part of model risk management. As discussed briefly in Section 3, a common practice for traditional models has been for developers to hand off documentation to information technology staff, who re-code the input variables and the model for the production environment. As implementation complexity can increase with the adoption of machine learning, some banks are streamlining processes by using feature calculations directly rather than requiring re-coding.

Even with this change, additional complexities at the software layer can cause implementation challenges unless they are deliberately managed. Machine learning models are often built using open source tools and libraries. Libraries are interdependent with each other and with operating system versions, and discrepancies in the underlying operating environment can inadvertently cause model outputs to shift. Banks report using a variety of methods, including containers and other ML operations technologies often managed through "MLOps" processses, to ensure consistency between development and production so that model outputs match expectations and remain stable over time. Lenders are increasingly favoring these methods because the ability to apply standard software development lifecycle controls to predictive modeling software increases the reliability and accuracy of model deployment activities. Consideration must be made to ensure information security and IT processes are not in conflict with practices intended to manage model risk, and that the appropriate balance is struck between information security risk and model risk.

5.4.2 Monitoring after deployment

As noted above, periodic monitoring is a key component of MRM processes to assess the extent to which external conditions or internal usage of a model have evolved in ways that pose different risks to the bank or may merit mitigation measures. However, while the basic concept is universal, banks' monitoring systems vary substantially depending on the nature and scale of the underwriting model, what risk is being monitored for (e.g., general model performance vs. other compliance requirements), and a variety of other considerations.

With regard to general model performance and risk management, the practices discussed in this section are not used by all institutions or for all models. However, where they are used, lenders have noted that certain adjustments can be helpful when working with ML models. Examples of such adjustments can include:

- **Univariate input variable monitoring:** Some banks historically have depended on population stability index (PSI) metrics to monitor for changes in individual input variables that may indicate that conditions, applicant populations, or other external factors are beginning to shift in ways that could affect model performance. However, where an ML model uses hundreds or thousands of variables that may have very large differences in their contributions to overall predictiveness, monitoring PSI metrics for every single variable may trigger substantial internal follow up in response to changes that end up having very little impact on model performance. In an effort to better calibrate risk management, banks may choose methods that weigh the most predictive variables higher than less predictive variables in informing model performance assessments.
- **Multivariate input variable monitoring:** Some banks that have historically relied on univariate monitoring approaches have also decided to implement multivariate outlier detection techniques in the ML context. Particularly where models are constructed to rely on feature interactions to generate predictions, multivariate monitoring approaches may be helpful to measure impacts that univariate input monitoring methods may not fully capture.
- **Reason code review:** Separate from processes focused on assessing the functioning of lenders' methods for generating adverse action notices, as noted in Section 4 some banks may review the distribution of adverse action reason codes for models in production to consider whether shifts over time indicate a shift in the distribution of applicants that warrants further attention from a broader MRM perspective.
- **Calibration monitoring and refitting:** Credit policies are designed to generate specific profit and loss outcomes based on an underwriting model score and overlays. Due to concerns about ML model stability, some banks report periodically providing new odds tables to inform credit risk partners in the first and second lines of defense about when it may be necessary to recalibrate a model or adjust credit policy thresholds. Some banks even periodically refit the model to ensure model performance could not be improved by re-estimating the model with new observations. However, refitting of ML models typically occurs on a faster cadence in the marketing and fraud contexts than for underwriting, given that marketing materials and learnings change frequently and that fraud models must be adjusted frequently to detect changes in bad actor strategies. Many institutions will establish parameters for these processes to provide for refit without having to go through the full validation process applied to new models, while still providing for appropriate testing.

More broadly, institutions typically define up front various types of "significant" changes that will trigger additional validation activities or other compliance reviews. Significance is often viewed at a conceptual level in this context as having a substantial impact on the model's outputs, but different lenders may define triggers somewhat differently. Examples may include deploying an existing model for a new use or meaningful changes in inputs.

6. SPECIAL CONSIDERATIONS FOR PARTICULAR MODEL TYPES

This section briefly discusses additional considerations that come into play when banks decide to deploy machine learning underwriting models that are (1) proprietary algorithms developed by third party vendors; and/or (2) are deployed in a "second look" capacity to evaluate only those applicants that have already been declined under a bank's more traditional primary model. Each section briefly describes why the use of each type of model is potentially appealing to banks and additional practical and regulatory considerations that are triggered by using such ML models.

6.1 Vendor models

Resource constraints prompt many smaller banks to rely on vendors in various capacities even when developing and deploying traditional logistic regression underwriting models.[61] Financial services companies of all sizes use credit scores like VantageScore and FICO for certain benchmarking purposes, and even very large banks may use vendors in certain contexts such as fraud because they may have access to data across multiple institutions that would not otherwise be available to the bank working only with its own information. In the machine learning context, vendors can be even more critical in helping to fill gaps in banks' infrastructure, expertise, and data sources by providing development platforms and tools, consulting services and advice, and customized model development services. In some cases, banks may decide to rely on a proprietary, "off the shelf" third party ML model for credit scoring or broader underwriting functions. Such arrangements potentially allow smaller institutions to leverage more powerful models than they would be able to develop internally based solely on their own data and resources.

However, relying on a vendor's ML proprietary models adds another layer of considerations to managing compliance with the various regulatory expectations discussed in preceding sections for at least two reasons. First, the same internal resource and technology limitations that prompt smaller institutions to turn to vendors rather than developing models internally need to be considered in determining how to conduct meaningful third party oversight. In addition, vendors vary widely in how much information they are willing to share about their models. While some vendors routinely provide detailed reports about attribute and model definitions, structures, and the results of their internal compliance and validation testing, others provide much more limited details due to competitive and intellectual property concerns.

This is potentially thorny because banks are ultimately responsible for compliance with federal law regardless of whether a particular function is outsourced or performed internally. Indeed,

working with vendors triggers its own compliance obligations. Under prudential guidance, banks are expected not only to engage in initial due diligence when selecting vendors, but to monitor their performance on an ongoing basis and take appropriate remedial action if problems develop.[62] Bank staff are also still answerable to their examiners for model risk governance and compliance with consumer protection laws, regardless of whether particular compliance and validation functions are performed by the vendor or the bank.[63]

It is thus critical for banks to determine what information and expertise they will need to screen and monitor model vendors on an ongoing basis and to structure their screening processes to yield a clear picture of what level of information vendor candidates are willing to share about their models. The following list of potential information types can be a helpful starting point for screening potential vendors and facilitating internal deliberations about what information the bank needs to answer examiner inquiries and its own internal risk tolerances about ML models:

- The sources of data used in the model for purposes of evaluating that such data is accurate, authorized for use, and does not violate privacy or other requirements.
- Information that helps to evaluate the representativeness of the samples used for various model development and validation processes.
- Information about the attributes, independent variables, or features used to make predictions, so that a clear nexus with the credit decision can be established and that disparate treatment risk can be evaluated.
- The model type and basic structural information to help understand its level of complexity.
- What post hoc tools the vendor uses in performing various compliance functions.
- Whether and how the vendor conducted validation testing of its methodology and processes for generating adverse action notices, and what results were generated.
- The vendor's methodologies for fair lending testing and what results were generated.
- What level of information the vendor is willing to share to help the bank conduct its own conceptual soundness and other model risk management analyses.
- The vendor's internal methodologies for evaluating conceptual soundness, overfitting risk, and other model risk management concerns, as well as the results generated.
- What ongoing processes the vendor has implemented for monitoring purposes and how the results of such monitoring get communicated to clients.
- How the vendor manages communications with its clients in connection with model updates or other systems changes.

Federal banking regulators have acknowledged the challenges that banks may face in obtaining due diligence from vendors and have noted that alternative strategies for risk management can include obtaining information from alternative sources, implementing additional monitoring or controls, or considering the use of other vendors. Industry utilities or consortiums, consulting with other organizations, or engaging in other supplemental joint diligence efforts are also permissible (consistent with antitrust law), though the agencies have emphasized that each bank retains ultimate responsibility for managing its own third party relationships.[64]

Processes for validating a vendor model to assess its performance for a bank's applicant pool and for ongoing monitoring of performance after implementation do not generally differ significantly depending on whether the model involves machine learning. Accordingly, they are not discussed in detail here.

6.2 Second look models

Banks sometimes decide to deploy models that use more advanced analytics or alternative data in a limited "second look" capacity, where they are used only to evaluate applicants who have been rejected under the bank's traditional primary models. This structure can be appealing as a relatively constrained way to pilot more innovative models, since it uses the models to increase access to credit for consumers who would otherwise be rejected but does not use them to make negative decisions and thus reduces the risk of unexpected negative consequences and complaints or criticism while banks are evaluating the models' performance in deployment.

However, it can be important to make both strategic and compliance decisions when moving to a two-model system. At a broad level, it may be helpful to establish plans for monitoring the second look model's performance and for when to assess whether and how to expand its use (assuming strong performance). This is important because despite the factors that make the second look approach appealing for an initial pilot program, restricting a more predictive ML model only to second look uses means that the more powerful model cannot screen out consumers who are unlikely to succeed in repaying a new loan. Second look models also generally are not used to improve the accuracy of risk-based pricing decisions. The second look structure thus limits benefits for lenders and borrowers alike.

Second look systems also require certain specific decisions regarding compliance procedures. One such question is how to handle adverse action notices. Many lenders do not factor the second look model into adverse action generation at all, reasoning that because negative decisions are made under the primary/ traditional model that operation of the second model is irrelevant to the disclosure process. The logic is similar to situations in which a lender may apply initial screens that decline applicants who have recent bankruptcies or credit scores below minimum cut-offs, without ever being assessed by the lender's internal underwriting model. In such cases, the adverse action notices focus on the bankruptcy factor or credit score rather than on criteria that would not have changed the outcome. Other lenders see a distinction, reasoning that in situations where the second look model is actually applied, it is more logical to base the adverse action notice on the criteria that caused the consumer not to be approved by that model rather than the criteria that caused the application to be rejected by the primary/traditional model. It may be technically possible to create an ensemble system to account for the impact of different factors across the two scoring systems, but such approaches are so complicated to calibrate that they are rarely adopted in practice.

7. CONCLUSION

Transitioning to ML models holds substantial potential for increasing the accuracy of credit underwriting, particularly in evaluating populations who are difficult to assess using traditional techniques and data sources. While the resulting models can be more complex and may require updating banks' data science techniques, analytic tools, and compliance processes, these challenges are not insurmountable. Many of the issues are not unique to machine learning models, but in fact have also arisen in somewhat differing degrees and forms in prior generations of predictive modeling methods and automated decision-making. ML models and recent evolutions in practice often offer greater flexibility and new resources for managing these issues. Indeed, developing and articulating frameworks for responsible use of ML models has the potential to help facilitate broader improvements in compliance procedures for underwriting models more generally by providing fresh perspectives and new tools and approaches for managing predictive performance, transparency, bias and fairness, and other topics.

This framework has been designed as a practitioner-oriented discussion of core risk modeling principles, processes, and issues that are important to consider in adopting ML models that benefit both lenders and customers. We hope that it will facilitate more efficient and productive conversations among a wide range of stakeholders, including different industry segments, examiners and regulators, researchers, advocates, and other critical constituencies. While technological tools, business practices, and policy frameworks will continue to evolve, we believe that the learnings and practices discussed above can help to facilitate the responsible implementation of ML techniques in credit risk modeling even as some elements continue to change over time.

APPENDIX A

Working Group Members

Institutions that participated in working group discussions that shaped this framework include the following. As noted in the introduction, participating institutions have different modeling methodologies and their own validation and testing frameworks. Thus, participation in this process and paper does not necessarily equate to endorsing every specific technique and practice discussed in the document.

- Citibank
- Fifth Third Bank
- JP Morgan Chase & Co.
- Synchrony

APPENDIX B

Use of Predictive Models Over the Lifecycle of Consumer Credit Products

While this framework has focused on the use of machine learning models for credit approvals, pricing, loan size, and related decisions in the core of the underwriting process, it is worth noting that similar types of predictive models can play critical roles throughout the lifecycle of a consumer credit product. Each stage involves specific activities and key decisions that can be enhanced by data-driven models to improve marketing effectiveness, underwriting, risk management, customer engagement, and portfolio performance. Some of the considerations and techniques discussed in the main text may also be useful in managing models for these other purposes.

B.1 Pre-marketing and product design

- **Objective:** Identify target segments, design appropriate product features, and outline marketing strategies.
- **Activities:** Research customer demographics, credit needs, and spending behaviors to define product terms (interest rate, credit limits, fees, and rewards) that appeal to target markets.
- **Potential Model Types and Uses:**
  - **Propensity Models:** Use propensity-to-apply models to predict which consumers are most likely to apply for the product based on demographic and behavioral data.
  - **Market Segmentation Models:** Segment consumers based on likelihood of need (e.g., revolving credit needs, high-ticket purchases) to tailor marketing messages.
  - **Product Fit Models:** Ensure the product aligns with customer needs and risk tolerance by using analytics that evaluate potential customer responses to various product features.

B.2 Marketing and customer acquisition

- **Objective:** Attract qualified applicants and encourage applications from targeted segments.
- **Activities:** Launch multi-channel marketing campaigns (digital ads, direct mail, email, in-branch promotions) to reach potential customers. Design offers that are appealing and compliant with fair lending practices.
- **Potential Model Types and Uses:**
  - **Response Models:** Predict the likelihood that a consumer will respond to marketing efforts, optimizing ad placements and reducing costs by focusing on high-response segments.

- **Credit Eligibility Models:** Evaluate pre-screened lists for potential credit eligibility, reducing the likelihood of declines during underwriting and enhancing consumer experience.

B.3 Application and underwriting

- **Objective:** Assess applicants' creditworthiness, make approval or decline decisions, and determine credit limits and interest rates for approved applicants.
- **Activities:** Gather application information, pull credit bureau data, and validate identity and income. Run applications through underwriting models.
- **Potential Model Types and Uses:**
  - **Credit Risk Models:** Predict likelihood of default based on credit history, income, and other factors to make accurate approval/decline decisions.
  - **Fraud Detection Models:** Identify potentially fraudulent applications by comparing patterns with known fraudulent behaviors.
  - **Income and Stability Models:** Estimate income or employment stability when direct income verification isn't available, supporting creditworthiness assessments.

B.4 Account setup and onboarding

- **Objective:** Establish the account, confirm terms, and engage new customers to build initial loyalty.
- **Activities:** Set up credit limits, finalize terms, communicate product benefits, and encourage first-time use (e.g., via incentives or account welcome materials).
- **Potential Model Types and Uses:**
  - **Activation Models:** Predict likelihood of customer engagement post-activation to tailor messaging that encourages early account usage.
  - **Retention Models:** Identify customers at risk of low engagement and use personalized onboarding efforts (such as credit counseling) to build loyalty and long-term usage.

B.5 Account management and servicing

- **Objective:** Support account activity, monitor ongoing creditworthiness, manage risk exposure, and foster customer loyalty.
- **Activities:** Manage customer inquiries, address disputes, process payments, and offer credit limit adjustments or promotional offers.
- **Potential Model Types and Uses:**
  - **Behavioral Scoring Models:** Regularly assess a customer's credit risk based on real-time account behavior (e.g., spending patterns, payment timeliness).
  - **Cross-Sell/Upsell Models:** Identify opportunities for additional products (e.g., credit line increases or supplementary credit offers) based on spending habits and credit risk.
  - **Limit Management Models:** Adjust credit limits based on updated risk assessments, allowing responsible borrowers more access while minimizing exposure on higher-risk accounts.

B.6 Early risk detection and proactive risk management

- **Objective:** Identify early signs of delinquency or overextension and take preventative actions to mitigate risk.
- **Activities:** Use monitoring systems to flag high-risk behaviors (e.g., late payments, high utilization) and intervene through automated or manual outreach.
- **Potential Model Types and Uses:**
  - **Early Delinquency Models:** Predict likelihood of payment issues before they occur, allowing the bank to send reminders or offer hardship assistance.
  - **Attrition Models:** Forecast which customers are at risk of leaving due to payment struggles or lack of account engagement, helping guide retention strategies or credit counseling.
  - **Stress Testing Models:** Assess potential impacts of economic downturns or personal events (e.g., job loss) on customer portfolios, helping banks manage exposure preemptively.

B.7 Collections and recovery

- **Objective:** Recover unpaid balances on overdue accounts while maintaining customer goodwill and minimizing losses.
- **Activities:** Initiate collections processes, from soft reminders to formal collections actions, and offer repayment solutions or hardship assistance.
- **Potential Model Types and Uses:**
  - **Collections Scoring Models:** Predict the likelihood of successful recovery for each delinquent account, segmenting cases for targeted collections strategies.
  - **Recovery Optimization Models:** Optimize collections by identifying the most effective collections strategy (e.g., payment plans or settlement offers) based on customer behavior and payment history.
  - **Write-Off and Charge-Off Models:** Forecast likelihood of uncollectible debt to manage reserves and assess the timing for charge-offs, improving financial transparency and regulatory compliance.

B.8 Account closure and post-cycle analysis

- **Objective:** Close accounts either due to customer decision, account inactivity, or charge-off, and analyze performance data to inform future product improvements.
- **Activities:** Settle outstanding balances, officially close accounts, and conduct analyses to understand lifecycle performance, customer behaviors, and overall portfolio health.
- **Potential Model Types and Uses:**
  - **Attrition and Lifetime Value Models:** Assess lifetime value and reasons for account closure, identifying patterns and enhancing future product retention strategies.

- **Performance Analysis Models:** Evaluate overall product performance, analyzing the quality of origination, effectiveness of scoring models, and long-term customer behaviors to optimize future programs.
- **Portfolio Risk and Profitability Models:** Aggregate lifecycle data to measure portfolio risk and profitability, guiding product redesigns and strategic decisions.

APPENDIX C

Common Machine Learning Algorithms for Use in Credit Risk Models

This section briefly describes the differences between supervised and unsupervised machine learning and between two of the most popular supervised ML techniques for building credit models, gradient boosting trees and neural networks. For more detailed descriptions of these and other types of machine learning and artificial intelligence, see FinRegLab, Machine Learning Market & Data Science Context.

C.1 Supervised versus unsupervised machine learning

Supervised learning refers to models that are trained using both input variables and a target (outcome) variable. The purpose of supervised machine learning models is to learn to predict the target variable from input variables. Supervised learning models are almost always used when lenders are estimating the probability of default by a potential borrower, and are commonly used for credit scoring and stress testing. Information on past borrowers, including their repayment behavior, can be used to train a supervised machine learning model that predicts the repayment behavior of new borrowers. Once a new application is made for a loan, supervised learning models can then predict the likelihood of default by analyzing the applicant's information in the context of past borrowers' repayment behavior and outcomes.

Unsupervised learning refers to models that detect patterns or clusters in a dataset without using a target variable. Data without a target variable are referred to as unlabeled. Although unsupervised learning models are not used directly to predict the probability of default, they can be used to find similarities or associations between the features and characteristics of individuals included in a dataset and to label cases of interest for a supervised model. Such clustering of borrowers is widely used for customer segmentation in marketing and in models used for image recognition. In lending, it is sometimes used to optimize credit lines.

C.2 Gradient boosting trees

Among machine learning models, tree-based models are often used in credit underwriting because they offer an attractive balance of predictive power and operational efficiency. They come in various forms and degrees of complexity but all are built using traditional "if-then" logic to break the estimation of the target variable into a series of discrete, binary analyses (for example, an initial branch in the tree might be based on whether consumers have experienced a bankruptcy within a certain amount of time, while subsequent "nodes" in the branches might consider debt loads or other factors separately for each branch).

One of the most popular tree-based methods is called gradient-boosted decision trees, which are ensemble models that generate a series of individual decision trees that each focus on the prediction error of the prior model and then generate a weighted sum of the predictions of all the component trees. (One particularly popular open-source version is called Extreme Gradient Boosting (XGBoost), which combines gradient-boosted trees' structure with other analytical enhancements.[65] Such a system is more complex than relying on a single decision tree or so called random forest ensembles that do not factor in prediction errors in generating the series of individual trees. However, gradient-boosted decision trees tend to have lower prediction error rates and better predictive power than other types of tree-based models.

C.3 Neural networks

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

Neural networks have been used extensively in fraud analytics, where the accuracy and higher predictiveness of these models allow financial institutions to better understand and detect fraud patterns in extremely large volumes of transaction data. Similarly, neural networks are used by some lenders for credit underwriting, especially where lenders are working with large-scale, diverse datasets for which neural networks' capacity to recognize complicated patterns is particularly valuable. The non-linear nature of these models may be particularly valuable after an economic shock like the onset of the pandemic because they are better able to identify and assess dynamic relationships between features and the target variable. However, this increased predictiveness can come at some cost with regard to computational power and complexity/explainability. Lenders may limit the number of layers in the network or use other structural constraints can improve the model's transparency.[66]

APPENDIX D

Common Explainability and Model Diagnostic Tools

In the last fifteen years, data scientists have made considerable strides in developing supplemental methods to analyze complex machine learning models to better explain and understand their predictions. These post hoc explainability techniques and taxonomies for categorizing them are continuing to evolve as more evidence is gathered about their capabilities and performance in the context of specific applications. This section describes particular techniques across three categories: surrogate models, feature importance explainability methods, and example-based explainability methods. For more detailed descriptions of these techniques and debates about explainability in ML credit underwriting models, see FinRegLab, Machine Learning Market & Data Science Context.

D.1 Surrogate models

Surrogate models are typically small and interpretable models, such as shallow decision trees, rule sets, or regression models, that are trained on the predictions of a highly complex model to explain its functioning.

D.1.1 Local Interpretable Model-Agnostic Explanations (LIME)

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

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

LIME historically was used both to explain the model's behavior around individual data points and to quantify feature importance for the overall model. Its usage has diminished as new techniques have been developed but is sometimes used today as a baseline to compare the outputs and performance of other explainability tools against or to generate insight into feature importance as discussed further below.

The primary challenge for LIME is derived from the inherent difficulty of using a simplified model to explain a much more complicated model. This challenge is more acute when the surrogate is a linear model, since the surrogate in this instance may not do well in mimicking the effect of non-linear relationships and feature interactions in the underlying model. Some of LIME's work arounds significantly reduces its computational speed. In addition, when LIME is used to understand the importance of specific features in a model, changes to the number of samples used or other parameters can also produce large changes in results. Computationally, LIME requires a way of telling how "similar" two given points are, and this must be supplied by the model developer. LIME's explanations are relatively sensitive to this weighting function, which is based on the distances between a sample point and the particular point of interest. In practice, choosing a distance function that produces useful explanations can be challenging.

D.2 Feature importance explainability methods

Feature importance techniques evaluate how much individual variables contribute to a model's prediction. In these methods, data are usually perturbed or permuted—meaning they are purposefully distorted or altered in a variety of ways—to determine how those changes affect the model's predictions. The aggregated effect of those changes speak to how much a variable affects the model's predictions. Feature importance or variable importance scores can be presented in charts with associated predictions, or they can be aggregated together to describe the importance of a variable on the model's predictions overall, and graphed for comparison. Feature importance explainability methods include Shapley Additive Explanation (SHAP), integrated gradients, partial dependence plots, individual conditional expectation plots, and accumulated local effects plots.

D.2.1 Shapley Additive Explanations (SHAP)

SHAP uses mathematical methods derived from a body of cooperative game theory research to analyze and explain the contributions of particular features to a model's prediction. The concept of the Shapley value method is as follows: In a cooperative game with N players and a function that values how much total output is generated if all the players contribute together, the Shapley value is a method that attempts to measure the individual contribution of each player to the output generated by the cooperation of all players. If the features are the players in a given complex model, from an economic standpoint, it can be interpreted as a weighted average of a feature's marginal contribution to every possible subset of grouped features.

Similar to LIME, SHAP explains how a model behaves locally. In the context of credit underwriting, local predictions can be helpful for generating adverse action notices for individuals who are denied credit. However, unlike LIME, SHAP measures feature importance by averaging the contribution of a feature across all possible combinations of features, creating more robustness and consistency than LIME's surrogate models. Averages may be marginal or conditional depending on the particular type of SHAP used. Although some versions of SHAP are model-agnostic, specialized variants such as tree SHAP and linear SHAP have emerged to be used with particular model types and typically operate faster and produce more reliable outputs than generic implementations.

SHAP is attractive to many practitioners because it is available as an open-source tool. Furthermore, several model-specific versions of SHAP can be faster to calculate than other explainability techniques. However, there are several criticisms of SHAP. First, many machine learning models cannot naturally handle "missing" features, as required by SHAP. This means that "missingness" is typically achieved by replacing a feature with a nominal value (such as the average over a chosen

sample). It is unclear whether the theoretical benefits of SHAP hold up under these approximations. Second, like many other explanation methods including LIME, some versions of SHAP make the unrealistic assumption that features are uncorrelated. This assumption glosses over real-world nuance present in real datasets including those used in financial services. Finally, calculating exact SHAP values can require significant time and computational resources even where model-specific versions of SHAP are used, so approximation or sampling methods are often used instead with some corresponding tradeoff in the quality of explanations. If too few samples are used, moreover, the resulting SHAP values can be noisy, and not reflective of actual model behavior.

D.2.2 Integrated gradients

Integrated gradients were developed to explain outputs from a differentiable model—that is, a model where the change (or derivative) in model output can be easily calculated. Many popular machine learning models are differentiable, including many neural networks. Integrated gradients work by summing the change in model output with respect to each feature or dimension along a straight-line path between two model inputs. The result measures the marginal contribution of each feature to the difference in model score between the two observations. To determine the change in score along each dimension along the path, the gradient of the model is calculated, which represents the rate of change associated with the model in each dimension. Values of the gradient of the model are summed along the path between an observation to be explained (x_1) and some reference point (x_0) that can be chosen based on the specific application. When evaluating the conceptual soundness of an underwriting model, x_0 might be repeatedly sampled from a broad selection of applicants, and the outputs averaged to calculate the average marginal contribution of each feature to the model score. When generating adverse action notices, x_0 might be repeatedly sampled from a collection of approved applicants, and the output averaged to determine which features contributed the most to a denial of an individual applicant. The resulting feature-level values can then be fed into the grouping process described in main text to compute higher-level adverse action reasons.

Integrated gradients are attractive for a variety of reasons: they are intuitive, straightforward to implement, and available in several well-maintained open-source packages. Although the IG method does not assume feature independence like LIME or TreeSHAP, the straight-line path it relies on may pass through areas of the data manifold that are not likely to occur in reality. It is highly sensitive to the choice of reference x_0, so care must be taken in making that choice. Integrated gradients are defined only for continuous models, although some extensions have been proposed for discontinuous models such as tree ensembles and other piecewise continuous functions.

D.2.3 Partial dependence plots

Partial dependence plots (PD plots or PDPs) are common visualization methods that depict how an individual feature interacts with the model's predictions. For each value of a given feature, the PD plot shows the average predicted outcome. Consider as an example an underwriting model that analyzes the following features: number of prior loans taken, number of past defaults, and number of outstanding loans. If the model developer is interested in how the number of past defaults affects the model's prediction of the likelihood of default, a PD plot feeds the underlying model every possible value for the number of past defaults for each possible combination of features in order to understand how the model works. For a single value of the number of past defaults, it will average all those possible combinations and plot the average, then will do the same for all values of the number of past defaults. This means that for every data point, the PD plot will replace the number of defaults with zero, feed a new set of features into the underlying or original model, take

the average of the resulting predicted scores, and plot them as a single point on the PD plot. This process is repeated for every value observed in the dataset for the number of past defaults. This ultimately creates a plot of averaged predicted estimates of default probability against all possible numbers for the number of defaults. This analysis can be done for any feature. The user can then see whether or not this relationship is linear, or if there is any value that is particularly surprising that might indicate inaccuracies in the dataset, in the model, or a novel relationship worth analyzing.

PD plots are easy to understand and make identifying any unexpected behavior in the relationship between a single feature and the model's prediction easy to detect and intuitive. They are designed to represent the relationship between features and the outcome at a global level, which makes PD plots suitable for model development, but not for generating individual applicant explanations. Additionally, they can be used with any type of machine learning model. However, PD plots rely on individual data points that do not exist in the original dataset. The method replaces real instances found in the dataset with synthetic feature pairings in order to make its averaged predictions over a larger set. Where the synthetic data points do not represent well the actual dataset, this can introduce bias and lead to inaccurate estimates of the effect of the feature on the results. Further, PD plots assume that each feature is independent of each other, which may often not be the case. This assumption may limit the utility of this approach in helping understand how feature interactions and correlated features affect the predictions produced by complex models. Figure 2 shows PD plots compared to the following methods.

D.2.4 Individual Conditional Expectation plots

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

For example, in a sample ICE plot such as the one depicted in Figure 2, if "x_1" axis represents the number of past defaults and "partial yhat" represents the predicted probability of default, then the curves would show the change in the predicted probability of default as the number of past defaults varies. These plots also provide insight into how the number of past defaults interacts with, for example, the number of months since the last credit card was opened. If the number of past defaults and number of months since the last credit card opened were to show some interaction on the predicted probability of default, then the curve of the lines for instances where the individual has opened a credit card one month ago for those who opened a card 24 months ago would have different shapes/slopes. In this hypothetical example depicted in Figure 2, lines representing different amounts of elapsed time since the applicant's last card was opened do not interact in the output of predicting default since the curves all have the same parabolic shape and simply show a shift along the y-axis. This means that based on this figure, number of months since last credit card opened has no non-linear relationship with number of past defaults and so does not change its influence on predicted defaults.

ICE plots show each instance or person in the dataset as a single line, where the value of the feature of interest varies. This makes the plot more interpretable to the user who can even focus on a given line and see how changing a feature like the number of past defaults might affect the given individual.

Similar to PDPs, ICE plots are not able to generate reliable estimates with correlated features, which may create congested plots and means that they cannot fully explain relationships between features.

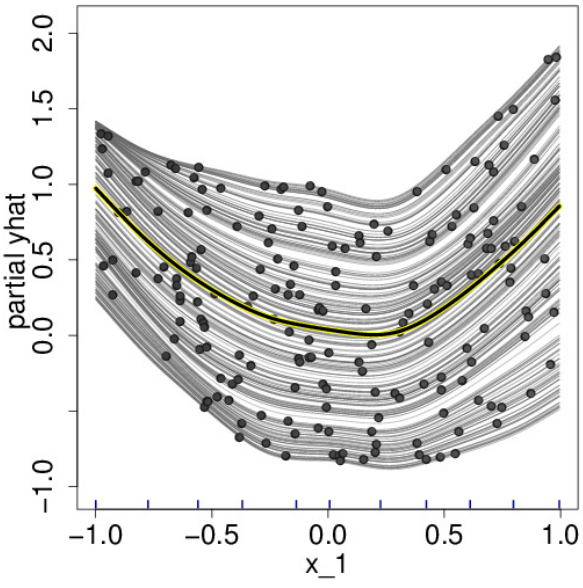
D.2.5 Accumulated Local Effects plots

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

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

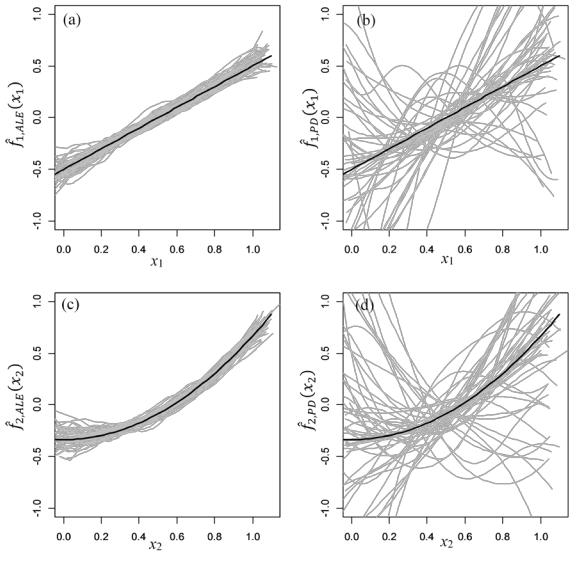
FIGURE 2 ILLUSTRATIVE ICE AND PDP PLOTS

Illustrative ICE Plot



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

Illustrative PDP Plot



ALE plots (left), and PD plots (right). The black line represents true effects in all plots.

Source: Daniel W. Apley & Jingyu Zhu, Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, arXiv:1612.08468.

APPENDIX E

Fair Lending Background

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.[67] The Fair Housing Act (FHA) prohibits discrimination in residential real estate lending on many of the same bases, as well as familial status and disability.[68] Both of these laws are enforceable by individual applicants and borrowers as well as by regulatory authorities.[69]

These anti-discrimination laws evolved historically to focus on two primary doctrines, disparate treatment and disparate impact. Disparate treatment generally prohibits consideration of race, gender, or other protected characteristics and close proxies for those characteristics in underwriting, scoring, pricing, and other models pertaining to the lending decision.[70] Disparate impact prohibits the use of facially neutral practices that have a disproportionate adverse impact on the basis of protected characteristics, unless the practices serve legitimate business needs that cannot be reasonably met through a less discriminatory alternative.[71]

As noted in Section 2, the Trump Administration announced in late April 2025 that it intends to eliminate disparate impact liability across employment, financial services, and other contexts.[72] Multiple federal banking regulators have stripped references to the doctrine from their examination manuals, stating that they will no longer examine for disparate impact and that they expect regulated institutions "to provide fair access to financial services, treat customers fairly, and comply with all applicable laws and regulations."[73] State activities relating to disparate impact in financial services have continued.[74] The Consumer Financial Protection Bureau has proposed an amendment to implementing regulations under ECOA to state that disparate impact is not cognizable under that law.[75] Under the Fair Housing Act, where the Supreme Court has held that disparate impact applies, the Department of Housing and Urban Development has proposed to eliminate its implementing regulations and leave application to the courts.[76]

This edition of the framework does not address fair lending compliance given that it is currently being revised by federal regulators.

Endnotes

1 FinRegLab, "Advancing the Credit Ecosystem: Machine Learning & Cash Flow Data in Consumer Underwriting."

2 The OCC started Project REACh in 2020. It has acted as a convener in bringing participants together for conversations on various financial services topics, but participants' subsequent work products and initiatives are their own. For another example, *see* "Reconsideration of Value (ROV) Best Practices," hosted on the American Bankers Association website. This report reflects information shared by individual participants and is not endorsed by the Federal Government, does not necessarily represent OCC views, and is not subject to Federal information quality, privacy, security, and related guidelines. It does not constitute an endorsement, recommendation, approval, or favoring of any Project REACh participant or other entity or their products, programs, initiatives, or other actions by the OCC. National banks and federal savings associations are responsible for ensuring that all actions taken related to information shared by Project REACh participants are consistent with safety and soundness, compliant with applicable laws and regulations, and protective of consumers' rights, as applicable.

3 *See* Appendix A for a list of participating institutions.

3 *See, e.g.,* Board of Governors of the Federal Reserve System, "Report to Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit"; Berger and Frame, "Small Business Credit Scoring and Credit Availability"; Gates et al., "Automated Underwriting in Mortgage Lending: Good News for the Underserved?".

5 Logistic regression is classified as machine learning in many textbooks but has been used far longer in credit underwriting than the other techniques discussed in this framework.

6 During economic downturns, for example, the same score value may be associated with a lower odds of repayment than during benign time periods. As a result, predicting actual loss rates requires additional information and calculations beyond the scoring model itself. This is one of the reasons that lenders may adjust their minimum qualifying scores in different economic conditions as a way of trying to keep loss rates consistent over time.

7 Consumer Financial Protection Bureau, "Technical Correction and Update to the CFPB's Credit Invisibles Estimate"; *see also* FinRegLab, "Advancing the Credit Ecosystem: Machine Learning & Cash Flow Data in Consumer Underwriting," 9.

8 FinRegLab, "Advancing the Credit Ecosystem: Machine Learning & Cash Flow Data in Consumer Underwriting"; Toh, "Addressing Traditional Credit Scores as a Barrier to Accessing Affordable Credit"; Cochran et al., "Utility, Telecommunications, and Rental Data in Underwriting Credit," 3-5.

9 Static supervised machine learning models can also be used in fraud and marketing screens, which help to shape the funnel as to which applicants seek credit from particular lenders, as well as in other aspects of loan origination and servicing as discussed in Appendix B. Some of the risk management practices and issues described in this framework document may apply to such use cases as well, but the data, models, and updating practices may vary and there are often additional considerations that merit separate attention in those other contexts.

10 Correlation between input variables means that as one variable shifts in value, the other tends to change as well. For example, negative payment history and amounts owed are the two most influential components of many credit scoring models, and may often tend to shift together as consumers who become over extended start to fall behind on their payments. Where features tend to change in tandem, coefficients in a traditional logistic regression model that measure the magnitude of each individual feature's impact will have greater uncertainty levels because it is unclear which feature is actually driving the change in the predicted outcome.

11 Examples include building separate scorecards for consumers who have experienced bankruptcy or who have thin credit files. A different combination of variables may be most predictive or the relative weights of common variables may be different in shaping predictions for those groups relative to applicants with extensive credit history and no history of delinquent payments, for example.

12 Examples include weight of evidence binning and related techniques, which transform input variables in ways that can capture some relationships between variables that may change in magnitude or even direction. For example, lenders can use these techniques to account for the fact that risk levels associated with credit utilization rates often follow a U shape, with the lowest risk levels among consumers who have a medium, steady level of utilization rather than very high or very low.

13 With traditional techniques, highly correlated variables can cause issues that lead to instability and unreliable results. ML models are less likely to experience these issues because of greater representational flexibility, although highly correlated variables can still complicate explainability. Thus, developers face fewer technical constraints but may still decide in some cases to omit correlated features for other reasons.

14 Attitudes and research about the potential predictiveness and access gains from applying machine learning models to traditional data sources only vary somewhat. *See* FinRegLab, "Advancing the Credit Ecosystem: Machine Learning & Cash Flow Data in Consumer Underwriting" for a literature review and quantitative analysis.

15 Blattner and Nelson, "How Costly Is Noise? Data and Disparities in Consumer Credit"; Kenneth P. Brevoort et al., "Credit Invisibles and the Unscored."

16 Albanesi and Vamossy, "Credit Scores: Performance and Equity"; FinRegLab, "Explainability & Fairness in Machine Learning for Credit Underwriting: Policy Analysis."

17 *See generally* FinRegLab, "The Next Wave Arrives: Agentic AI in Financial Services," 12-13; FinRegLab, "The Use of Cash-Flow Data in Underwriting Credit: Market Context & Policy Analysis," § 4; Consumer Financial Protection Bureau, "Required Rulemaking on Personal Financial Data Rights," Federal Register 89, no. 222 (November 18, 2024): 90838-90998; Consumer Financial Protection Bureau, "Personal Financial Data Rights Reconsideration," Federal Register 90, no. 161 (August 22, 2025): 40986-40989.

18 15 U.S.C. § 1691(d); 12 CFR § 1002.9.

19 15 U.S.C. § 1681m(a)(1), (b), (h); 12 CFR § 1022.72.

20 15 U.S.C. § 1691(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 under ECOA); 42 U.S.C. § 3605 (prohibiting discrimination on the basis of race, color, national origin, religion, sex, familial status or disability under the FHA). Regulations under ECOA also set forth standards for certain "empirically derived, demonstrably and statistically sound, credit scoring system(s)." In developing underwriting models, lenders and model developers may follow those standards even if they are not legally required to do so. 12 C.F.R. 1002.2(p), 1002.6(b)(ii).

21 Executive Office of the President. "Executive Order 14281 of April 23, 2025: Restoring Equality of Opportunity and Meritocracy." Federal Register 90, no. 80 (April 28, 2025): 17537. *See* Appendix E for more discussion.

22 Consumer Financial Protection Bureau, "Equal Credit Opportunity Act (Regulation B)," Federal Register 90, no. 217 (November 13, 2025): 50901.

23 Department of Housing and Urban Development. "HUD's Implementation of the Fair Housing Act's Disparate Impact Standard," Federal Register 91, no. 9 (January 14, 2026): 1475. *See* Appendix E for more discussion.

24 Board of Governors of the Federal Reserve System, "Guidance on Model Risk Management," Supervisory & Regulation Letter 11-7; Federal Deposit Insurance Corporation, "Adoption of Supervisory Guidance on Model Risk Management," Financial Institution Letter 22-2017; Office of the Comptroller of the Currency, "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management," Bulletin 2011-12.

25 Board of Governors of the Federal Reserve System, "Guidance on Model Risk Management," Supervisory & Regulation Letter 11-7; Federal Deposit Insurance Corporation, "Adoption of Supervisory Guidance on Model Risk Management," Financial Institution Letter 22-2017; Office of the Comptroller of the Currency, "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management," Bulletin 2011-12.

26 U.S. Treasury Department, "Remarks: A Reset on Liquidity Regulation."

27 12 U.S.C. §§ 1861-1867, 5516(e); Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency, "Interagency Guidance on Third-Party Relationships: Risk Management," Federal Register, 88, no. 111 (June 9, 2023): 37920; Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency, "Third-Party Risk Management: A Guide for Community Banks"; Consumer Financial Protection Bureau, "Service Providers," Compliance Bulletin and Policy Guidance 2016-02.

28 Office of the Comptroller of the Currency, "Comptroller's Handbook, Bank Supervision Process," 26-28. References to an eighth historical category, reputational risk, have been removed by the OCC, FDIC, and NCUA under the Trump Administration.

29 For additional background, *see* FinRegLab, "Advancing the Credit Ecosystem: Machine Learning & Cash Flow Data in Consumer Underwriting," app. A, C (online).

30 For illustration, using a debt-to-income ratio rather than separate components for existing debts and income is a simple example of feature engineering.

31 The Receiver Operating Characteristic Area Under the Curve evaluates a model's ability to distinguish between defaulters and non-defaulters across all possible classification thresholds. The Kolmogorov-Smirnov statistic measures the maximum separation between the cumulative distribution functions of predicted probabilities for defaulters and non-defaulters across the entire score range generated by the specific dataset. Both are common metrics for measuring predictiveness. For additional background, *see* FinRegLab, "Advancing the Credit Ecosystem: Machine Learning and Cash Flow Data in Consumer Underwriting," 23-24.

32 Containers help to move ML models from one environment to another, for instance from development to testing and validation to deployment. They effectively provide a system image of the critical code, model, and other elements. APIs allow different software applications to communicate with each other, and are frequently used both to transfer data between and within organizations and to integrate machine learning models into applications.

33 15 U.S.C. § 1691(d); 12 CFR § 1002.9; 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, paras. 9(b)(2)-2, -4, -5, -8.

34 15 U.S.C. § 1681m(a)(1), (b), (h); 12 CFR § 1022.72; 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, para. 9((b)(2)-9. FCRA key factors may overlap with ECOA principal factors, but are not necessarily identical. For instance, in a situation where the lender acted both based on a third party credit score generated based on credit bureau data and on a determination that the applicant's income made the likelihood of repayment too low, the two sets of disclosures would be expected to have some distinct elements.

35 15 U.S.C. § 1691(d); 12 CFR § 1002.9.

36 15 U.S.C. § 1681m(a)(1), (b), (h); 12 CFR § 1022.72.

37 15 U.S.C. § 1691(d)(6); 12 CFR § 1002.2(c).

38 12 C.F.R. § 1009(a)(1), (b)(2); 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, paras. 9(b)(2)-2, -4. Two guidance documents issued by the Consumer Financial Protection Bureau on the topic of adverse action notice compliance in the context of machine learning models and non-traditional data sources were revoked under the Trump Administration. Consumer Financial Protection Bureau, "Adverse Action Notification Requirements in Connection with Credit Decisions Based on Complex Algorithms," Consumer Financial Protection Circular 2022-03; Consumer Financial Protection Bureau, "Adverse Action Notification Requirements and the Proper Use of the CFPB's Sample Forms Provided in Regulation B," Consumer Financial Protection Circular 2023-03; Consumer Financial Protection Bureau, "Interpretive Rules, Policy Statements, and Advisory Opinions; Withdrawal," Federal Register 90, no. 90 (May 12, 2025): 20084-20087.

39 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, paras. 9(b)(2)-1, -4, -5, -8.

40 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, paras. 9(b)(2)-2.

41 12 C.F.R. § 1009(b)(2) & app. C; 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, paras. 9(b)(2)-3, -4.

42 FinRegLab et al., "Machine Learning Explainability & Fairness: Insights from Consumer Lending."

43 *See, e.g.*, Yang et al., "Survey on Explainable AI: From Approaches, Limitations and Applications Aspects"; Abusitta et al., "Survey on Explainable AI: Techniques, Challenges, and Open Issues"; Burkart and Huber, "A Survey on the Explainability of Supervised Machine Learning"; Anderson, "Testing Machine Learning Explanation Methods"; Alonso and Carbó, "Accuracy of Explanations of Machine Learning Models for Credit Decisions"; Gramegna and Giudici, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk"; Misheva et al., "Explainable AI in Credit Risk Management."

44 Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions."

45 *See e.g.*, Janzing et al., "Feature Relevance Quantification in Explainable AI: A Causal Problem."

46 *See, e.g.*, Miroshnikov et al., "Stability Theory of Game-Theoretic Group Feature Explanations for Machine Learning Models."

47 Sundararajan et al., "Axiomatic Attribution for Deep Neural Networks."

48 *See, e.g.*, Frye et al., "Shapley-Based Explainability on the Data Manifold". For recent literature reviews, *see* Yeo et al, "A Comprehensive Review on Financial Explainable AI"; Černevičienė and Kabašinskas, "Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review."

49 For example, if the SHAP direction for a feature is counter-intuitive with regard to default risk, review with domain experts, legal/compliance, and customer service teams may be needed. Edge cases where a particular applicant has very unusual combinations of attributes or significant numbers of missing features may produce SHAP values that indicate zero or near-zero contributions from all features, which is technically correct but difficult to interpret and convey to applicants. Evaluating outliers and extreme values for treatment can help to identify such situations and consider whether providing broader reasons such as "limited credit history" may be clearer than listing the raw variables identified by the SHAP package.

50 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.9, paras. 9(b)(2)-5.

51 An example of an exception could include bureau attributes that have special values that generally indicate that a particular trade type is missing or not reported. Different reason language might be used for such attributes depending on whether they have special values or regular values. In this case it is a common practice to treat the special values as new binary flags and map them like the others.

52 High levels of feature correlations can make PDPs less useful for interpretation because one of many variations of a given variable will not paint the full picture of the relationship between the concept those variables represent and the model's prediction for a given range of values. To get a sense of what is happening at the concept level (vs. the feature level) some practitioners sum the Shapley values by concept, and a summary or grouped PDP plot is produced that represents the trend for all the variables in the group. This yields a partial dependence plot that is less subject to noise at the individual feature level and thus is more useful for interpretation when features are correlated.

53 Board of Governors of the Federal Reserve System, "Guidance on Model Risk Management," Supervisory & Regulation Letter 11-7; Federal Deposit Insurance Corporation, "Adoption of Supervisory Guidance on Model Risk Management," Financial Institution Letter 22-2017; Office of the Comptroller of the Currency, "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management," Bulletin 2011-12.

54 As noted in Section 2, additional guidance regarding the use of AI may be forthcoming. U.S. Treasury Department, "Remarks: A Reset on Liquidity Regulation."

55 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. *See, e.g.*, Office of the Comptroller of the Currency, "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management," Bulletin 2011-12, attachment at 3.

56 *See, e.g.*, Office of the Comptroller of the Currency, "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management," Bulletin 2011-12, attachment at 3.

57 *See, e.g.*, Office of the Comptroller of the Currency, "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management," Bulletin 2011-12, attachment at 3-4.

58 One or more models may be selected for benchmarking purposes, either from internal or external sources (such as vendor models). If no similar models can be found, some lenders may use credit scores from a general third party model instead. When deploying benchmark models it is important to ensure as much of an apples-to-apple comparison as practicable, for instance by considering the accuracy and completeness of the data used to develop the benchmark, testing all models against the same samples (e.g., through-the-door and booked populations), and considering the extent to which the benchmark model has been built to the same standards with regard to soundness, transparency, and other qualities. A challenger model may not always have a better performance compared with a benchmark model since they may have been developed with different data and methods, but if discrepancies are beyond the expected or out of appropriate ranges they may warrant additional investigation and possibly revisions to the challenger model.

59 *See, e.g.*, Office of the Comptroller of the Currency, "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management," Bulletin 2011-12, attachment at 11.

60 For example, it is important to account for how choices in feature engineering will play out when post-hoc model explainability tools are used. If a particular special value such as "999" is used to denote a missing value or other data issue and the explainability tool treats it as an actual numeric value, this will affect the tool's outputs when used to explain feature importance and to construct adverse action reasons.

61 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.

62 Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency. "Interagency Guidance on Third-Party Relationships: Risk Management." Federal Register 88, no. 111 (June 9, 2023): 37920; Consumer Financial Protection Bureau, "Service Providers," Compliance Bulletin and Policy Guidance 2016-02. Expectations for credit unions with regard to vendor oversight are less formalized. Prudential regulators have also issued more specific guidance with regard to technology vendors and information security risks. *See generally* 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" (October 2023).

63 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, "The Use of Machine Learning for Credit Underwriting: Market & Data Science Context," § 2.3.

64 Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency. "Interagency Guidance on Third-Party Relationships: Risk Management." Federal Register 88, no. 111 (June 9, 2023): 37920.

65 For example, XGBoost uses various techniques optimization techniques such as L1 and L2 regularization, which lead to better predictive performance and speed. It also includes tree pruning, a process controlled through a hyperparameter to remove relatively irrelevant or unimportant information from the trees and manage risks related to overfitting. Further, XGBoost can recognize areas where data sparsity may affect the model's accuracy and handle missing data better by imputing values. XGBoost also includes parallelization, which sorts the data in a way that uses CPU power more efficiently, which speeds up the training process.

66 For example some lenders use a piecewise linear activation function, such as Rectified Linear Unit (ReLU), that creates a neural network consisting of many locally linear models that are each interpretable in the sense that the individual models use a linear combination of attributes to calculate an output.

67 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).

68 42 U.S.C. § 3605 (prohibiting discrimination on the basis of race, color, national origin, religion, sex, familial status or disability). The FHA applies to first mortgages, second mortgages, and home equity lines of credit.

69 Cases must generally be filed within five years of the violation under ECOA. Under the FHA, private cases must be filed in court within two years of the violation or termination of the violation. 15 U.S.C. § 1961e(f); 42 U.S.C. § 3605(a)(1)(A).

70 Overt discrimination, which includes situations in which lenders openly make decisions or distinctions on the basis of protected characteristics, is sometimes broken out as a third category, separate from broader disparate treatment. Other sources distinguish between the types of evidence that may be used, including statements or other overt evidence that protected class was considered and comparative quantitative analyses. *See, e.g.*, Office of the Comptroller of the Currency, "Comptroller's Handbook, Consumer Compliance: Fair Lending," 5-6. Examples of disparate treatment include setting different credit limits based on borrowers' ages or including factors such as gender or race in an underwriting model. Age is the only protected characteristic that is permissible to use in underwriting, subject to various safeguards. 12 C.F.R. § 1002.6(b)(2).

71 Texas Department of Housing & Community Affairs v. Inclusive Communities Project, 576 U.S. 519 (2015); 12 C.F.R. § 1002.6(a); 12 C.F.R. Pt. 1002, Supp. I, sec. 1002.6, para. 6(a)-2.

72 Executive Office of the President. "Executive Order 14281 of April 23, 2025: Restoring Equality of Opportunity and Meritocracy." Federal Register 90, no. 80 (April 28, 2025): 17537.

73 Office of the Comptroller of the Currency, "Fair Lending: Removing References to Disparate Impact," Bulletin 2025-16; Federal Deposit Insurance Corporation, "Update to the FDIC's Consumer Compliance Examination Manual," Financial Institution Letter 41-2025; National Credit Union Administration, "Removal of Disparate Impact," Letter to Credit Unions 25-04.

74 Office of the Massachusetts Attorney General, "AG Campbell Announces $2.5 Million Settlement with Student Loan Lender for Unlawful Practices Through AI Use, Other Consumer Protection Violations"; New Jersey Division on Civil Rights, "Rules Pertaining to Disparate Impact Discrimination." New Jersey Register 57, no. 24 (December 15, 2025): 2840(b) (applicable to housing related credit and financial assistance). In the employment context, private litigants have also begun seeking to intervene to continue pursuing claims after federal agencies pulled back. Ray, "Altoona Man Asserts Right to Intervene in Sheetz Employment Practices Lawsuit." The executive order directs the U.S. Attorney General to consult with federal agencies and report back with regard to potential changes in federal regulations, guidance, and other issuance and to "any appropriate measures" to address state laws.

75 Consumer Financial Protection Bureau, "Equal Credit Opportunity Act (Regulation B)," Federal Register 90, no. 217 (November 13, 2025); 50901.

76 *Texas Department of Housing & Community Affairs v. The Inclusive Communities Project, Inc.*, 576 U.S. 519 (2015); Department of Housing and Urban Development. "HUD's Implementation of the Fair Housing Act's Disparate Impact Standard," Federal Register 91, no. 9 (January 14, 2026): 1475.

Bibliography

Abusitta, Adel, Miles Q. Li, and Benjamin C.M. Fung. "Survey on Explainable AI: Techniques, Challenges, and Open Issues." Expert Systems with Applications 255, pt. C (2024): 124710.

Albanesi, Stefania and Domonkos F. Vamossy. "Credit Scores: Performance and Equity." National Bureau of Economic Research Working Paper 32917, 2024.

Alonso, Andrés and José Manuel Carbó. "Accuracy of Explanations of Machine Learning Models for Credit Decisions." Banco de España Working Paper 2222, 2022.

Anderson, Andrew A. "Testing Machine Learning Explanation Methods." Neural Computing and Applications 35, no. 24 (2023): 18073-18084.

Berger, Allen N. and W. Scott Frame, "Small Business Credit Scoring and Credit Availability." Journal of Small Business Management 47, no. 1 (2007): 5-22.

Blatner, Laura and Scott Nelson, "How Costly Is Noise? Data and Disparities in Consumer Credit." August 2024.

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. "Guidance on Model Risk Management." Supervisory & Regulation Letter 11-7, April 2011.

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." October 2023.

Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency. "Interagency Guidance on Third-Party Relationships: Risk Management." Federal Register 88, no. 111 (June 9, 2023): 37920.

Brevoort, Kenneth P., Phillip Grimm, and Michelle Kambara. "Credit Invisibles and the Unscored." Cityscape, 18, no. 2 (2016): 9-34.

Burkart, Nadia and Marco F. Huber. "A Survey on the Explainability of Supervised Machine Learning." Journal of Artificial Intelligence Research 70 (2021): 245-317.

Černevičienė, Jurgita and Audrius Kabašinskas. "Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review." Artificial Intelligence Review 57 (2024): 216.

Cochran, Kelly Thompson, Michael Stegman, and Colin Foos. "Utility, Telecommunications, and Rental Data in Underwriting Credit." Urban Institute & FinRegLab, 2021.

Consumer Financial Protection Bureau. "Service Providers." Compliance Bulletin and Policy Guidance 2016-02, October 2016.

Consumer Financial Protection Bureau. "Adverse Action Notification Requirements in Connection with Credit Decisions Based on Complex Algorithms," Consumer Financial Protection Circular 2022-03, May 2022.

Consumer Financial Protection Bureau. "Adverse Action Notification Requirements and the Proper Use of the CFPB's Sample Forms Provided in Regulation B," Consumer Financial Protection Circular 2023-03, September 2023.

Consumer Financial Protection Bureau. "Required Rulemaking on Personal Financial Data Rights,"Federal Register 89, no. 222 (November 18, 2024): 90838-90998.

Consumer Financial Protection Bureau. "Interpretive Rules, Policy Statements, and Advisory Opinions; Withdrawal," Federal Register 90, no. 90 (May 12, 2025): 20084-20087.

Consumer Financial Protection Bureau. "Technical Correction and Update to the CFPB's Credit Invisibles Estimate." 2025.

Consumer Financial Protection Bureau. "Personal Financial Data Rights Reconsideration," Federal Register 90, no. 161 (August 22, 2025): 40986-40989.

Consumer Financial Protection Bureau, "Equal Credit Opportunity Act (Regulation B)," Federal Register 90, no. 217 (November 13, 2025): 50901.

Department of Housing and Urban Development. "HUD's Implementation of the Fair Housing Act's Disparate Impact Standard," Federal Register 91, no. 9 (January 14, 2026): 1475.

Executive Office of the President. "Executive Order 14281 of April 23, 2025: Restoring Equality of Opportunity and Meritocracy." Federal Register 90, no. 80 (April 28, 2025): 17537.

Federal Deposit Insurance Corporation. "Adoption of Supervisory Guidance on Model Risk Management." Financial Institution Letter 22-2017, June 2017.

Federal Deposit Insurance Corporation. "Update to the FDIC's Consumer Compliance Examination Manual." Financial Institution Letter 41-2025, August 2025.

FinRegLab. "The Use of Cash-Flow Data in Underwriting Credit: Market Context & Policy Analysis." 2020.

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

FinRegLab. "Explainability & Fairness in Machine Learning for Credit Underwriting: Policy Analysis." 2023.

FinRegLab. "Advancing the Credit Ecosystem: Machine Learning & Cash Flow Data in Consumer Underwriting." 2025.

FinRegLab. "The Next Wave Arrives: Agentic AI in Financial Services." 2025.

FinRegLab, Laura Blatter, and Jann Spiess. "Machine Learning Explainability & Fairness: Insights from Consumer Lending." 2023.

Frye, Christopher, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. "Shapley-Based Explainability on the Data Manifold." arXiv. December 2021.

Gates, Susan Wharton, Vanessa Gail Perry, and Peter M. Zorn. "Automated Underwriting in Mortgage Lending: Good News for the Underserved?" Housing Policy Debate 13, no. 2 (2002): 369-391.

Gramegna, Alex and Paolo Giudici. "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk." Frontiers of Artificial Intelligence 4 (2021): 752558.

Janzing, Dominik, Lenon Minorics, and Patrick Bloebaum. "Feature Relevance Quantification in Explainable AI: A Causal Problem." Proceedings of Machine Learning Research 108 (2020): 2907-2916.

Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions," Neural Information Processing Systems, 30 (2017): 4768-4777.

Miroshnikov, Alexey, Konstandinos Kotsiopoulos, Khashayar Filom, and Arjun Ravi Kannan. "Stability Theory of Game-Theoretic Group Feature Explanations for Machine Learning Models." arXiv, August 2024.

Misheva, Branka Hadji, Joerg Osterrieder, Ali Hirsa, Onkar Kulkarni, and Stephen Fung Lin. "Explainable AI in Credit Risk Management." arXiv, March 2021.

National Credit Union Administration. "Removal of Disparate Impact." Letter to Credit Unions 25-04, September 2025.

New Jersey Division on Civil Rights. "Rules Pertaining to Disparate Impact Discrimination." New Jersey Register 57, no. 24 (December 15, 2025): 2840(b).

Office of the Comptroller of the Currency. "Sound Practices for Model Risk Management: Supervisory Guidance on Model Risk Management." Bulletin 2011-12, April 2011.

Office of the Comptroller of the Currency. "Comptroller's Handbook, Bank Supervision Process." v. 1.1, amended March 20, 2025.

Office of the Comptroller of the Currency. "Comptroller's Handbook, Consumer Compliance: Fair Lending." v. 1.0, amended July 14, 2025.

Office of the Comptroller of the Currency. "Fair Lending: Removing References to Disparate Impact." Bulletin 2025-16, July 2025.

Office of the Massachusetts Attorney General. "AG Campbell Announces $2.5 Million Settlement with Student Loan Lender for Unlawful Practices Through AI Use, Other Consumer Protection Violations." July 10, 2025.

Project REACh. "Reconsideration of Value (ROV) Best Practices." December 27, 2024.

Ray, Philip. "Altoona Man Asserts Right to Intervene in Sheetz Employment Practices Lawsuit." Altoona Mirror, August 15, 2025.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Neural Networks." arXiv, June 2017.

Toh, Ying Lei. "Addressing Traditional Credit Scores as a Barrier to Accessing Affordable Credit," Federal Reserve Bank of Kansas City Economic Review, Third Quarter 2023.

U.S. Treasury Department. "Remarks: A Reset on Liquidity Regulation." March 3, 2026.

Yang, Wenli, Yuchen Wei, Hanyu Wei, and Yanyu Chen. "Survey on Explainable AI: From Approaches, Limitations and Applications Aspects." Human Centric Intelligent Systems 3 (2023): 161-188.

Yeo, Wei Jie, Wihan Van Der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. "A Comprehensive Review on Financial Explainable AI." Artificial Intelligence Review 58 (2025): 189.

Digital version available at finreglab.org

Published by FinRegLab, Inc.

1701 K Street NW, Suite 1150
Washington, DC 20006
United States