FinRegLab

Explainability & Fairness in Machine Learning for Credit Underwriting

Policy & Empirical Findings Overview

JULY 2023

About FinRegLab

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

Acknowledgments

This Empirical Findings and Policy Overview is part of a broader research project on the explainability and fairness of machine learning in credit underwriting. The empirical research described herein was conducted in collaboration with Professors Laura Blattner and Jann Spiess at the Stanford Graduate School of Business. Other reports in this series are available at https://finreglab.org/ai-machine-learning/#aipublications.

FinRegLab would also like to recognize the presenters and members of our project Advisory Board and policy working groups who contributed to productive discussion of the design, execution, and interpretation of this research and the policy insights reflected in this document. The Advisory Board and policy working groups consisted of subject matter experts from computer science, economics, financial services, and regulatory backgrounds and includes representatives from approximately 45 major institutions including bank and nonbank financial institutions, technology firms, advocacy and civil society organizations, and academic institutions. State and federal regulators participated as observers in Advisory Board and policy working group meetings.

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

> Michael Akinwumi, National Fair Housing Alliance; Brad Blower; Nick Bourke; Jay Budzik; Annie Delgado, Nat Hoopes, and Irene Meyer, Upstart; John Morgan, Capital One; David Moskowitz, Burning Tree Advisors; David Silberman; Stephen Van Meter; and Michael Umlauf and Gene Volcheck, TransUnion.

We would also like to acknowledge the FinRegLab team who worked on various elements of the research project:

> Natalia Bailey, Alex Bloomfield, Kelly Thompson Cochran, Sarah Davies, Colin Foos, Saurab Guatam, Hilary Griggs, Gillous Harris, Tess Johnson, Mashrur Khan, Duncan McElfresh, Kerrigan Molland, P-R Stark, YaYa Sun, Sormeh Yazdi, and Zishun Zhao.

Support for this publication and other aspects of FinRegLab's Machine Learning in Credit Underwriting project was provided by the Mastercard Center for Inclusive Growth, JPMorgan Chase, and Flourish Ventures. Detailed information on our funders can be found on the inside back cover.

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

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

The stakes in the credit context are high. ML underwriting models' greater accuracy and capacity to analyze large datasets (particularly new, more inclusive sources of information) have the potential to increase access to credit for millions of people who are difficult to assess using traditional models and data. This underserved population includes disproportionately high numbers of Black, Hispanic, and lower-income consumers.[2] Yet the very quality that fuels ML models' greater predictive power—their ability to detect more complex data patterns than prior generations of credit algorithms—makes them more difficult to understand and increases concerns that they could exacerbate inequalities and perform poorly in changing data conditions.[3]

The complexity of many ML models has caused transparency to emerge as an urgent threshold question for both lenders and regulators in evaluating whether individual models are safe, fair, and reliable for use. For example, many stakeholders are concerned that if users cannot assess whether a model is relying on strong, intuitive, and fair relationships between an applicant's behavior and creditworthiness to predict the likelihood of default, it may be more difficult to diagnose and mitigate performance and fairness issues or to determine compliance with regulatory requirements.

New data science techniques—often themselves involving machine learning or other complex computational methodologies—have emerged both to explain ML models' operation and to manage concerns about their fairness and reliability. These include both *post hoc* explainability techniques that analyze key aspects of model behavior and debiasing techniques that can be used to reduce racial or other disparities in model predictions. Many vendors that are providing platforms and services to support the development of ML models have incorporated these techniques into their proprietary tools for diagnosing, managing, and monitoring ML models. But the techniques and tools also raise questions about whether and how to use them appropriately both to manage models and to perform regulatory compliance tasks in the credit context.[4]

To study these questions, FinRegLab has conducted extensive market context interviews, performed empirical analyses with Professors Laura Blattner and Jann Spiess of the Stanford Graduate School of Business, and convened diverse stakeholders in policy working groups and other fora. This policy overview, which we are releasing in conjunction with our updated empirical white paper and in anticipation of a longer policy analysis to be released in late summer 2023,[5] summarizes critical learnings from the broader project. As discussed further below, we find that:

- **The transition to machine learning has the potential to improve fairness and inclusion, in part by giving lenders a more robust toolkit for mitigating disparities.** Despite the focus on transparency as a threshold issue for ML models as discussed above, the most powerful approaches to managing fairness did not necessarily hinge upon explaining the inner workings of the model as an initial step. Instead, we found that automated approaches that generated a range of alternative models produced options that had greater predictive accuracy and smaller demographic disparities than traditional strategies that assessed which input features made the biggest contribution to disparities and then omitted or made narrow adjustments to those individual features.
- **Some explainability techniques provided reliable information about key aspects of model behavior, though there was no "one size fits all" technique or tool that performed the best across all regulatory tasks.** Our evaluation found that it is important to choose the right explainability tool for the particular model and task, to deploy it in a thoughtful way, and to interpret the outputs with an understanding of the underlying data. The analytical framework we developed for the project is a useful starting point for stakeholders in evaluating these tools in different settings.
- **Defining basic concepts and expectations could be a useful first step toward updating regulatory frameworks for the machine learning era.** While ML technologies and our understanding of them are evolving rapidly, regulators can take steps now to encourage responsible use. For instance, defining the key qualities of trustworthy models and explainability tools would ensure that lenders manage for a consistent set of criteria and encourage more rapid refinement of measurement tools, benchmarks, and strategies. Clarifying expectations about how and when lenders should search for fairer alternative underwriting models would also increase consistency of practice and shape how lenders use their expanded toolkits in the ML context.

As the fairness research results illustrate, adjusting market practices and regulatory expectations to account for machine learning models could provide opportunities to address longstanding concerns about prior generations of predictive credit models and the compliance frameworks that govern them. Additional public research and stakeholder dialogue will be critical to advance these efforts, not only within the credit ecosystem but also with other sectors that are grappling with the trustworthiness of AI and ML models in other sensitive use cases. At the same time, lessons from deploying machine learning and secondary tools in the credit context have the potential to inform governance activities in other sectors and the development of more effective data science techniques for understanding and managing AI and ML models.

2. POLICY BACKGROUND

Lenders' adoption of algorithms to predict the likelihood that applicants will default on loans began decades ago based largely on data from three nationwide credit bureaus and statistical techniques such as logistical regression.[6] Such models generally rely on a relatively limited number of inputs that are selected by human developers, who work to find combinations that both maximize overall predictive power and minimize correlations to simplify model operations. Developers can use coefficients generated by the regressions and other widely used metrics to measure the importance of individual features for various business and regulatory purposes:

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

With advances in computational power, some lenders have begun deploying machine learning techniques to develop underwriting models. Here, the algorithms themselves identify predictive relationships among large numbers of inputs (which may be highly correlated) while developers make critical decisions about such issues as what data the learning algorithms are trained on, how the algorithms generate underwriting models, and what techniques, tools, and strategies should be used in development and validation processes. Depending on those decisions, some ML models may not be significantly harder to understand than traditional underwriting models, while others are substantially more complex. The most complicated ML models, which are sometimes referred to as "black box" models, rely on hundreds or thousands of features (including in some cases "latent features" that are generated by the ML algorithms from the initial inputs), complicated architectures, and data relationships that may vary in magnitude and direction depending on the circumstances.[7]

This complexity can help to increase the predictive power of ML underwriting models, but also increases concerns about whether they will deteriorate in changing conditions or exacerbate existing disparities. For example, some stakeholders have raised concerns that ML models may pinpoint the financial gaps created by historical discrimination with even greater precision than current models or worsen disparities by effectively "reverse engineering" race or other demographics. The lack of transparency about how the models are generating their predictions further increases concerns about model management and regulatory compliance, especially since traditional approaches rely upon being able to distill certain information from regression models.[8]

At the same time, data science and machine learning techniques provide a range of alternative options for evaluating and managing models. For instance, lenders are managing concerns about the potential transparency of ML underwriting models by imposing up-front constraints on model complexity, applying a variety of secondary or *post hoc* methods to explain key aspects of model behavior, or combining the two approaches. Data science and machine learning techniques also provide a range of options for debiasing models. Lenders that decide to rely at least in part on these data science techniques in developing ML models may build them based on open-source code or turn to vendors that may incorporate multiple approaches when offering broader platforms and services to support model development.

Given the importance of these techniques and tools, FinRegLab has engaged in extensive qualitative, quantitative, and policy analyses to interrogate the value of these techniques and tools in managing explainability and fairness concerns in the ML underwriting context. Our empirical research tested both proprietary model diagnostic tools provided by seven vendors in the market as well as several open-source tools deployed by the research team.[9] The tools were applied to four credit card underwriting models built by the research team—ranging from a logistic regression model including more than forty features to a neural network model trained on several hundred features—to perform various model diagnostic and management tasks relating to the three regulatory compliance regimes described above.[10] The full results are available in our empirical white paper.

To complement this empirical work, FinRegLab's market and policy analyses have been informed by extensive interviews and engagement with a broad range of stakeholders, including executives from banks and fintechs, technologists, consumer advocates, academics, and regulators. In addition to convening a project advisory board, FinRegLab co-sponsored an April 2022 symposium with the

EXAMPLES OF EXPLAINABILITY AND DEBIASING TECHNIQUES

Examples of explainability techniques include:

Surrogate models: Local Interpretable Model-Agnostic Explanations (LIME) build simpler models to assess which features are most important to the prediction for an individual consumer or the model as a whole.

Feature-importance techniques: Shapley Additive Explanations (SHAP) omit individual input features over multiple iterations and analyze the resulting changes in model performance to generate a cumulative measure of the features' relative importance.

Examples of debiasing techniques include:

Joint optimization: Under this approach the learning algorithm is directed to maximize predictive accuracy at the same time that it minimizes disparities in each successive iteration of an underwriting model it builds.

Adversarial debiasing: Here, a separate model is used to analyze disparities in each successive underwriting model to provide feedback to the learning algorithm as the development process continues.

U.S. Department of Commerce, National Institute of Standards and Technology, and the Stanford Institute for Human-Centered Artificial Intelligence and organized three policy working groups in 2022 to discuss key aspects of ML adoption. FinRegLab published an initial report on the market and data science context for use of machine learning in credit underwriting in fall 2021 and will release a more detailed policy analysis in late summer 2023.[11]

This overview focuses on three topics: (1) Techniques for promoting fairness and inclusion; (2) The reliability of explainability techniques for multiple regulatory purposes; and (3) Other regulatory considerations. It concludes by discussing broader themes and next steps.

3. TECHNIQUES FOR PROMOTING FAIRNESS AND INCLUSION

Lenders that are seeking to reduce disparate impact risks have historically focused on identifying which individual features are driving any demographic disparities in a model's default predictions and assessing the potential effects of dropping or modifying those features, for instance through reweighting. However, dropping or modifying features can potentially reduce model accuracy, which may prompt lenders to continue using their baseline models. These analyses are typically conducted relatively late in the development process by separate compliance teams who are given access to actual or imputed demographic data, rather than by front-line developers.[12]

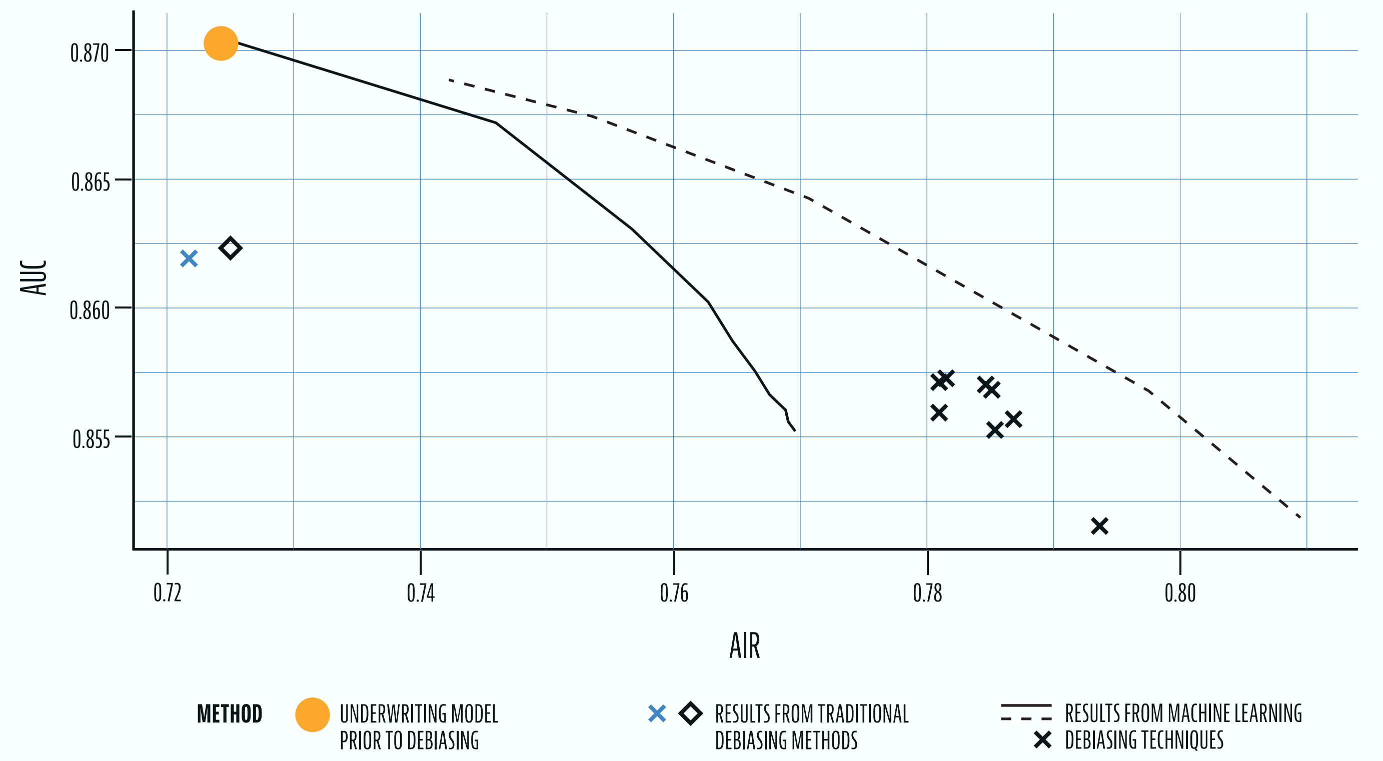
3.1 Finding Machine learning has the potential to usher in fairer credit decisions by giving lenders a more robust toolkit for mitigating disparate impacts.

Debiasing techniques give lenders a range of options that can be applied at different points in the development process for ML underwriting models, allowing them to generate a series of models to choose from that reduce the potential tradeoffs between fairness and predictive accuracy. In addition to techniques such as joint optimization and adversarial debiasing, many vendors provide general platforms or other services that facilitate the rapid iteration of models through assigning weights, changing model constraints, and other adjustments.

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

The graph below illustrates some of the results from different debiasing methods. Accuracy is represented by the area under the curve (AUC), a commonly used measure of predictiveness, while fairness is represented by the adverse impact ratio (AIR), a measure of the disparities in the selection rate between minority and non-minority consumers.[13] As reflected in the graphic, the traditional debiasing methods (blue X and black diamond) were significantly less predictive than the baseline

DEBIASING RESULTS



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

model (orange dot), but did not significantly improve fairness. The automated approaches (solid line, dashed line, black Xes) substantially improved fairness, with varying changes in predictive accuracy.

Further research into specific debiasing approaches could be helpful to illuminate the most promising methodologies and specific implementation choices that lenders face when deploying these techniques. It could also be helpful to probe the alternative models generated by such tools, for instance to understand the extent to which any declines in accuracy tend to be concentrated among different subgroups and how well the models perform in general validation processes. Thus, while the initial results are promising, additional public research could give lenders and regulators more confidence in selecting both specific debiasing approaches and from among the range of models that they generate.

3.2 Implications for public policy

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

3.2.1 Use of protected class information and specific debiasing techniques

A threshold question is whether specific debiasing techniques are permissible under fair lending laws to the extent that they use data about protected class membership in different ways than traditional mitigation approaches. While such techniques can reduce the risk of disparate impacts, concerns about violating prohibitions on disparate treatment have slowed the initial adoption of joint optimization and adversarial debiasing in the credit context relative to their use in some other sectors.

In recent years, however, lenders who are adopting machine learning models appear to have become increasingly comfortable in authorizing their fair lending compliance teams to deploy such automated debiasing techniques during searches for less discriminatory alternatives, while prohibiting their use by business units in earlier development stages. This bifurcation is consistent with historical fair lending compliance practice and guards against the risk of misuse of protected class information by the initial development team. However, depending on how lenders sequence their overall model development process, it may lengthen overall timelines for validation and deployment. Some other lenders remain reluctant to authorize the use of certain debiasing techniques by internal teams or vendors in the absence of further regulatory guidance.[14]

3.2.2 Standards in searching for and evaluating potential "less discriminatory alternatives"

A second set of policy questions concerns regulators' expectations for lenders in searching for and evaluating alternative models to determine whether they are a "less discriminatory alternative" (LDA) that reasonably meets the lender's legitimate business need to predict default risk while producing less disparity in predicted outcomes among protected groups. Many lenders today do not invest substantial resources in searching for LDAs, particularly where they are relying on traditional techniques and data sources and not making significant changes to their existing underwriting systems. Questions about the broader search for less discriminatory alternative models include:

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

At 2023 conferences, CFPB officials have described "rigorous searches for less discriminatory alternatives" as "a critical component of fair lending compliance management" and expressed concern that lenders may tend to shortchange this aspect of compliance. However, the agency has not issued formal guidance on LDA topics.[15]

One potential way to begin reducing regulatory uncertainty around these questions would be to acknowledge and build upon emerging market practices to ensure greater consistency among lenders. For example, although there are no specific thresholds articulated in existing regulatory guidance for determining whether an alternative model constitutes an LDA—either with regard to the boost in fairness or the loss of accuracy that lenders have to accept as compared to a baseline model—many lenders do set target ranges for predictive performance when they validate model performance for more general business purposes and in connection with model risk management expectations. While the ranges may vary depending on the lender, portfolio, and other circumstances, some stakeholders have suggested that such ranges could be viewed as the lender's articulation of an accepted range of performance for its business needs, such that alternative models that fall within those same ranges would constitute LDAs.

Some advocacy groups are calling on regulators to launch more ambitious initiatives, for instance by developing large datasets against which models can be tested and LDAs identified, using their examination teams to conduct searches for LDAs, and establishing specific metrics and thresholds for purposes such as determining whether an alternative model sufficiently boosts fairness relative to accuracy losses that it constitutes an LDA. However, some of these measures could be resource intensive and create substantial technical challenges, for instance in adjusting models based on data that they have not been trained upon.

4. RELIABILITY OF EXPLAINABILITY TECHNIQUES

As described above, lenders are expected to be able to explain adverse decisions to individual consumers and certain aspects of their models' overall operation to regulators. Compliance processes for meeting these various regulatory requirements have evolved in the context of regression algorithms that tend to rely on a relatively small set of input features and a widely accepted set of statistical assessments to help determine the importance of specific features. As machine learning adoption increases, stakeholders are grappling both with how to deploy *post hoc* explainability techniques most effectively and with the sufficiency of these techniques given the limitations of what such analyses reveal about the internal workings of more complex models.

4.1 Finding We can systematically evaluate the performance of explainability techniques and model diagnostic tools without having "ground truth" explanations

One of the challenges in applying diagnostic tools to complex ML models is that it is often infeasible to generate a complete explanation of the model's operations in order to verify the explainability tools' performance. Despite not knowing the ground truth explanation, we were able to design empirical tests that allowed us to compare a number of explainability techniques and vendor tools to each other and to objective benchmarks as described in Section 4.2.[16] We used these tests to analyze three primary qualities:

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

We viewed fidelity and consistency as threshold technical questions about the tools' reliability, with fidelity playing the most important role. For example, if a tool cannot reliably identify features that are important to a particular aspect of a model's operation, we would not necessarily expect or care whether its results were consistent with the results of some other tool in performing the same task. Usability is also a critical quality—indeed, in some ways ultimately the most critical for judging whether the tools can be used to assess or demonstrate regulatory compliance—but also more complicated to define and evaluate. For instance, usability results may depend not just on the general nature of the information provided by a diagnostic tool and implementation choices made in its deployment, but also on what options are available to the user in responding to the information.

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

4.2 Finding Some diagnostic tools can reliably identify features that are important for various regulatory purposes, even for complex ML models.

Our empirical analyses found that some but not all of the explainability tools we tested could reliably identify features that were important to models' behavior for particular regulatory tasks. For instance, changing the values of the features identified by the highest performing tools as important for adverse action purposes had a bigger impact on model predictions than "perturbing" features that were chosen at random or that were closely correlated to the "important" features. In contrast, changing the features identified by low performing tools sometimes caused default predictions to move in unexpected directions and had less effect than changing other features.

Similar tests applied in the fair lending and model risk management contexts also found that some tools identified features that had relatively large effects on model disparities and overall model operations, which may be useful to lenders in the course of broader activities to manage fair lending risks and meet model risk management expectations.

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

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

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

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

4.3 Implications for public policy

As stakeholders deepen their understanding of various explainability techniques and implementation choices, public policy questions regarding compliance with particular regulatory requirements have taken on additional urgency in light of the adoption of ML models. Regulatory guidance concerning what general qualities lenders should manage for in evaluating explainability technique performance or the permissibility of using specific techniques for specific tasks could be helpful to promoting more consistent practices. More broadly, stakeholders are debating whether further insight into feature interactions or other model operations beyond what current tools can provide is critical to various regulatory compliance functions.

4.3.1 Adverse action compliance

Historical regulatory guidance concerning adverse action disclosures has emphasized that lenders must provide information about the specific principal factors that shaped individual applicants' risk assessments (rather than just stating that they failed to meet the lenders' minimum credit criteria), but that guidance has allowed lenders to choose among methodologies for determining which factors to highlight for customers.[17] Though not required by regulation, concerns about protecting proprietary information and making disclosures easier to understand have prompted many lenders to adopt procedures for grouping related features together and mapping them to higher level "reason codes," rather than providing precise technical descriptions of individual features that drove a model's behavior. Regulatory guidance specifically notes that lenders need not describe how or why a feature negatively impacts predictions of default risk and discourages providing more than four to five factors.[18]

CFPB officials in 2022 issued a circular that highlighted the importance of validating any secondary tools used to explain complex models but did not discuss specific explainability techniques, validation methodologies, or thresholds for accuracy. [19]

As stakeholders consider how to adapt these historical practices and guidance to the machine learning context, our substantive results and our empirical framework and methodology could potentially help to assess the fidelity and consistency of individual model diagnostic tools and determine whether they are appropriate to rely upon in explaining individual underwriting decisions. Our results also suggest that the practice of grouping related features together to produce higher level action codes is particularly important in the machine learning context. For instance, given the large number of features in many ML underwriting models, the impact of a handful of individual input features is likely to be much less than in the context of a traditional regression model with only a few dozen inputs, and thus the disclosure of those inputs is likely to have less explanatory power as well. Aggregation processes can also help account for some of the technical challenges discussed above that are created by the presence of large numbers of correlated features.[20] Thus, thoughtful aggregation processes can potentially produce more consistent and meaningful disclosures for consumers that convey more information about the model's operation.

Beyond answering methodological questions about the responsible use and deployment of explainability techniques, stakeholders face policy questions about whether being able to pinpoint more granular information is pivotal to adverse action compliance. These issues are not unique to ML models, but they become more important in that context because specific feature interactions within ML models can drive predictions for individual consumers. For example, assume that a machine learning algorithm determined that late payments on a mortgage loan are associated with much greater default risk where mortgage loan balances are higher (e.g., $200,000 rather than $50,000). If disclosing that the *combination* of delinquencies and balance was important to an individual loan rejection is critical for compliance, it raises questions about both the precision of particular diagnostic tools and of the descriptions provided to consumers. However, there can also be potential disadvantages to requiring such specificity. For instance, consumers might read the disclosure to imply that they should prioritize paying down their mortgage balance even at the expense of incurring delinquencies on other loan types.

The transition to ML models also has implications for policy debates about the purpose of adverse action notices. These discussions include whether to mandate that the disclosures provide more information about how individual features negatively affect applicants' predicted default risk and/or tailored, forward-looking advice about how to improve chances of approval over time. Our empirical research suggests that both of these changes could be more complicated to implement for ML underwriting models due to the number of features and nature of the data relationships involved.[21] However, even for traditional models, such changes would raise a number of complicated policy decisions about what volume and specificity of information is most useful to consumers, how to determine which courses of action are feasible on what timelines, and how to account for risks that applicants may misconstrue the information. Accordingly, broader analyses would be required to determine if, how, and when adverse actions notices can be made more generally useful, and even actionable, for consumers.

4.3.2 Fair lending compliance

The results described above suggest that historical techniques for managing disparate impact risks by identifying a small number of features for targeted transformations may be supplanted by more automated strategies in the ML context. Nonetheless, lenders may still find it useful to analyze the extent to which individual input features that play a particularly important role in model operations are closely correlated with protected characteristics. For example, in the context of disparate treatment compliance, lenders will sometimes exclude features if they are so highly correlated with demographics that they might be deemed a pretext or proxy for intentional discrimination.

Yet while existing explainability techniques can be used to evaluate the significance of individual input features, they cannot directly identify individual latent features or feature interactions within more complex ML models. Thus, as in the adverse action context, stakeholders are mulling the importance of pinpointing specific feature interactions for purposes of fair lending compliance to assess whether those interactions might be considered proxies for protected class status.

However, a machine learning algorithm that has been directed to find the most predictive underwriting model it can is not acting in the same way as a human developer who decides whether to include features that could be used as a pretext or proxy for an applicant's race or gender. This raises important questions about whether the disparate impact framework is both more appropriate conceptually and more effective practically for evaluating and mitigating potential concerns about ML models' fairness. The availability of a more effective debiasing toolkit for managing disparate impact risks may provide a compelling counterweight to concerns about potential proxies in feature interactions constructed within the model. These considerations further underscore the importance of additional research into the effectiveness and limitations of machine learning debiasing techniques and of clarifying expectations around searches for less discriminatory alternative models.

4.3.3 Model risk management

Finally, as part of regulatory structures to protect the safety and soundness of depository institutions, banks are expected to implement a risk-based system of governance and monitoring that applies to all types of models used in their operations. "Model risk management" (MRM) guidance is principles-based and quite broad in nature, so that it not only prompts banks to validate and monitor their underwriting models' performance in light of changing economic conditions or other data shifts, but typically also to consider other types of reputational, legal, and business risks involved in using particular data sources and computational techniques. While non-banks are not subject to MRM guidance, bank partners and investors may impose governance and monitoring requirements by contract.

Regulators have not specified whether the explainability techniques and tools are subject to MRM governance procedures in their own right, but various notions of transparency are closely interwoven into MRM compliance. For example, documentation about data preparation and model development is considered a critical component to facilitating review and governance. Regulators expect lenders to inquire into the "conceptual soundness" of models by evaluating whether they rely on relationships in the data that are intuitive and defensible with regard to the outcomes that they are attempting to predict. And analyses of model sensitivity to changing data conditions often identify which features tend to drive the biggest changes in predictions.

Our substantive results and empirical framework and methodology could potentially help to assess the fidelity and consistency of individual model diagnostic tools and determine whether they are appropriate to rely upon in particular circumstances.[22] Defining what qualities are important to consider in choosing among tools and how to test those qualities could produce more consistent baseline practices and encourage further innovation in techniques and standards.

Beyond answering methodological questions about the responsible use and deployment of diagnostic tools, stakeholders face policy questions about how to account for differences in the type of information that diagnostic tools produce for ML models as compared to the information available for traditional regression models. These issues arise most sharply in the conceptual soundness context, where reviews have historically included assessments of whether models are relying on relationships that are empirically sound and draw on appropriate scientific, behavioral, or economic theories and industry practice. For regression models, practitioners rely on coefficients and other commonly used statistical analyses to understand the impact of each feature on the overall functioning of the model. Some firms have developed procedures that are tailored to ML models for conceptual soundness reviews—including use of various types of secondary explainability techniques to plot model relationships. In models that may involve thousands of features, they emphasize that the role of any one individual feature may be relatively limited and that it is important to distill key information about model operations overall. However, other stakeholders question at a fundamental level whether conceptual soundness expectations can be satisfied when reviewers cannot fully document and understand each individual feature and relationship within the model. Debates over the reliability of particular diagnostic tools are just one aspect of this broader issue.

5. OTHER REGULATORY CONSIDERATIONS

Although the primary focus of this project has been to interrogate the use of techniques for managing explainability and fairness concerns with ML models, we briefly note a few issues here concerning broad-based governance of ML models. These issues are most often assessed within the financial services context under model risk management frameworks that are only applicable to depository institutions, as described above, though they can be implicated under other regulatory requirements and in other sectors are often discussed under the broad rubric of "trustworthy AI."

5.1 Updating governance frameworks

Traditional guidance on model risk management is a principles-based framework for determining the responsible use of models of all types. Because it is so flexible, it can already provide a useful framework for managing the transition to ML techniques and new diagnostic and debiasing tools. At the same time, the existing guidance was drafted before ML adoption accelerated, and many stakeholders suggest it could benefit from refreshing. Prudential regulators have been studying potential updates but not yet released a specific proposal.

As this process plays out, there may be potential value in supplementing this guidance with an affirmative articulation of the qualities that make a model trustworthy, similar to frameworks that are emerging from various other sectors and jurisdictions.[23] Rather than taking as their starting point risks and mitigation processes, these broader frameworks tend to start with an articulation of the affirmative qualities that will allow humans to trust ML and AI models for sensitive and high-risk use cases. Incorporating these features into MRM frameworks could encourage lenders to develop protocols for testing machine learning models as to each quality, and regulators could evaluate those efforts.

This kind of broad-based process could be helpful to identify two areas for the evolution of policy, law, and regulation to foster fair and responsible use of machine learning in credit underwriting and financial services: (1) areas like fair lending risk management practices where existing expectations might be adapted to reflect new approaches to reducing disparities in credit decisions and (2) areas like the use of secondary data science techniques and tools where gaps in existing frameworks may need to be addressed to incorporate important elements of machine learning practice with no prior analogue. Articulating a sector-wide trustworthiness framework may also provide structure for differentiating specific risks and regulatory needs by use case, since considerations in using predictive models for risk forecasting may be different than using generative AI in customer interactions, for example.

5.2 Considering the extension of governance process expectations to nonbank institutions

Given the complicated issues that transitioning to ML underwriting models entails, a diverse group of stakeholders has suggested that amending existing law to impose basic governance expectations on nonbank adopters could be beneficial to borrowers, lenders, and the broader ecosystem. For example, they argue that such a change would both help to ensure that nonbanks are managing for a consistent range of risks in developing and deploying ML underwriting models and to level the playing field for banks.[24]

5.3 Addressing challenges in validating vendor-provided models

Finally, addressing the governance challenges in working with vendor-provided models and tools could have a critical effect on ML adoption for credit underwriting, particularly among banks with greater technological and resource limitations.[25]

While very large banks invest substantial resources in both developing proprietary underwriting models and associated model risk management programs, smaller institutions tend to be more reliant on third-party credit scores and vendors to help develop, deploy, and monitor their underwriting systems. The smaller institutions are still subject to model risk governance expectations, regardless of whether their underwriting models are in-house or outsourced, but as a practical matter gaining transparency into proprietary systems can be challenging in light of both intellectual property concerns and the same resource constraints that prompted lenders to outsource in the first place.

Thus, to the extent that MRM expectations for the adoption of machine learning underwriting models are unclear or extremely complex, this may tend to discourage adoption among such lenders because of additional concerns about managing risks in connection with vendors' models. As substantive regulatory expectations clarify for machine learning models, some stakeholders have noted that increases in direct supervision of vendors by federal regulators and/or the creation of certification programs could help to increase the consistency of compliance and promote a more efficient system for due diligence by lenders.

6. BROADER THEMES AND NEXT STEPS

The fact that commonly used explainability techniques cannot see inside the most complex ML underwriting models to directly and precisely map feature interactions raises questions across several different regulatory areas about whether it is critical to be able to perform such analyses in order to ensure the fair and responsible use of such models. While our empirical analyses found that some *post hoc* explainability tools can produce reliable information about various aspects of model operations, current tools do not produce precisely the same kinds of information that are available for traditional regression models. Thus, while research can help to better define what tools are best suited to particular tasks and best practices in their deployment, such technical information will not negate the need for broader policy dialogue and decisions about our ability to trust complex models.

At the same time, explainability and debiasing techniques can offer new strategies for managing particular policy and regulatory concerns about prior generations of predictive credit models and the compliance frameworks that govern them. These new techniques and approaches can be subject to meaningful oversight, although such functions may need to occur in different ways and at different stages of the development process relative to traditional models. The potential policy advantages and opportunities created by these developments and the fact that technologies and research are continuing to evolve also deserve serious consideration in ongoing policy debates.

As uncertain economic conditions further incentivize lenders to seek greater predictive power, both the more technical and broader policy questions are becoming more urgent. Additional research is critically important to help identify and encourage responsible, fair, and inclusive practices and to inform the evolution of regulatory frameworks to account for the increasing use of both machine learning models and secondary tools for managing explainability and fairness concerns.

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

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

» Updating governance frameworks, including potentially articulating the qualities of trustworthy AI/ML models similar to the ones described in Endnote 23, would encourage lenders to begin methodically evaluating and testing their systems and processes to address those core components. Such principles-based approaches can be especially helpful at early stages of evolution across diverse stakeholders, markets, circumstances, and technologies. In a similar vein, articulating the key qualities for explainability and diagnostic tools would

KEY AREAS FOR ADDITIONAL RESEARCH

Potential topics include:

- Deeper evaluation of specific debiasing approaches to illuminate the most promising methodologies for mitigating bias in machine learning underwriting models and the specific choices that lenders make when deploying those methods, including whether and how protected class characteristics (whether actual or imputed) can be responsibly used to improve the fairness of credit decisions.
- Deeper evaluation of the performance-fairness tradeoffs identified by debiasing tools that generate a range of alternative models, for instance to confirm whether there is a band in which lenders can improve the fairness of models without incurring significant loss of performance and how potential performance tradeoffs distribute across populations of interest.
- Evaluating whether the inclusion of additional types of underwriting data affects the fairness and inclusiveness of credit decisions as well as the performance, capabilities, and limitations of the kinds of data science techniques evaluated in the current project.
- Assessing with rigor the transparency costs related to the use of more complex machine learning underwriting models and related tradeoffs that lenders make to improve the transparency of underwriting models when they apply up-front constraints on model complexity.
- Deeper evaluation of specific implementation choices that make one model diagnostic tool higher performing on certain tasks, such as identifying whether and how the definition of the baseline set to which a rejected applicant is compared affects the quality of information given to consumers on an adverse action notice.
- Additional analysis of the extent to which different diagnostic tools disagree about important factors after accounting for correlations to classify the types of disagreements that persist and consider whether those types of disagreements have a material effect on the regulatory compliance tasks considered in this evaluation.
- Continuing refinement of assessment frameworks for evaluating secondary tools, including consideration of whether different or additional qualities can help to identify when information from such tools can be trusted and used in high-stakes contexts.

also help lenders begin to manage for a consistent set of questions and concerns, even as the technologies and assessment processes continue to evolve.

- Given current variations in whether and how lenders search for less discriminatory alternatives to baseline underwriting models, providing greater clarity on what constitutes an LDA and on regulators' expectations for search processes could significantly increase consistency in the market.

The current moment presents both significant risk (as millions of credit applications are being decided based on firms' best judgments as to regulatory compliance and secondary tool use) and significant opportunity (as policymakers have a unique moment in which they can affect the broad direction of evolution, before developing more calibrated and binding standards as the innovation lifecycle progresses). It also presents an opportunity to re-think and improve upon prior generations of automated underwriting in the extent to which they have left substantial numbers of people behind and replicated historical disparities. The coming years could offer the most fundamental reset of lending practices in several decades. Whether and to what extent those new systems prioritize responsible, fair, and inclusive use of ML models and secondary tools will ultimately depend not just on technology issues but on business and policy decisions. Rigorous research, thoughtful deployment, and proactive regulatory engagement are critical to ensuring that any new technology must ultimately benefit borrowers and financial service providers alike.

Endnotes

[1] Large language and image recognition models are leveraging "big data" across the Internet in a much more extreme way to enable interactive processing, raising substantial questions about data quality and accuracy, bias, authorship, reliability, and other topics. See, e.g. Shira Ovide, Your Selfies Are Helping AI Learn. You Did Not Consent to This, Washington Post (Dec. 9, 2022); Yogesh K. Dwivedi et al., So What If ChatGPT Wrote It? Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy, 71 International Journal of Information Management 102642 (2023). In contrast, machine learning underwriting models are trained on much smaller, curated data sets and limited in the extent to which they are allowed to update dynamically. See FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context §§ 2, 4.2 (2021); Bank Policy Institute, Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning 17 (June 25, 2021).

[2] FinRegLab, The Use of Cash-Flow Data in Credit Underwriting: Market Context & Policy Analysis § 2 (2020). For instance, about 20% of U.S. adults lack sufficient traditional credit history to generate scores under the most widely used models, with almost 30% of Black and Hispanic consumers lacking scores compared to about 16% of Whites and Asian-Americans. Consumer Financial Protection Bureau, Data Point, Credit Invisibles 4-6, 17 (2015); Mike Hepinstall et al., Financial Inclusion and Access to Credit, Oliver Wyman (2022).

[3] FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context §§ 1, 2.2, 3.

[4] FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context §§ 3, 5.

[5] FinRegLab, Laura Blattner & Jann Spiess, Machine Learning Explainability & Fairness: Insights from Consumer Lending (June 2023) (hereinafter Empirical White Paper).

[6] For more detailed discussions of the material presented in this section, see FinRegLab, The Use of Cash-Flow Data in Credit Underwriting: Market Context & Policy Analysis § 2; FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context; Empirical White Paper.

[7] For example, ML models can be used to detect relationships that may be non-monotonic (such that increasing the value of an input feature may reduce the likelihood of default in some circumstances and increase it in others) and/or non-linear (such that increasing the value of an input feature by a given amount may not change the likelihood of default by the same amount in all circumstances). FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context § 3.2.3.3.

[8] We define transparency as the ability of various stakeholders to access information they need related to a model's design, use, and performance. Some stakeholders use terms such as interpretability or explainability to express similar concepts. See FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context §§ 2-3 & Appendix B.

[9] The companies included Arthur, Fiddler, H20.ai, RelationalAI, SolasAI, Stratyfy, and Zest AI.

[10] The other models included a simple neural network trained on a small number of variables and an XGBoost model. The four models provide a spectrum from relatively simple to relatively complex based both on the architecture and number of features used. The models were built on a representative sample of data from a nationwide credit bureau from 2009 to 2017. Several participating companies also built models using the same data set and various machine learning techniques. Our data provider required the masking of certain feature descriptions, which limited the companies' ability to conduct qualitative feature reviews or create features manually in the course of model development. Empirical White Paper § 3.

[11] See https://finreglab.org/ai-machine-learning/explainability-and-fairness-of-machine-learning-in-credit-underwriting for the project reports and https://finreglab.org/podcasts/ for recordings of the April 2022 conference and other webinars and podcasts.

[12] This separation is designed to reduce risks of violating the prohibition against disparate treatment on the basis of protected characteristics. Federal laws require collection of demographic information for mortgage loans and will soon do so for small business loans, but generally prohibit its collection for other credit products. Accordingly, lenders and regulatory often impute gender, race, and ethnicity based on such factors as name and address. See FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context § 5 & Appendix B.2.

[13] Adverse impact ratio is used for evaluating disparities in a variety of contexts including credit and hiring. It is defined as the ratio of the acceptance rate for the minority group to the acceptance rate of the majority group. AIR values closer to 1 correspond to more parity. Area under the receiver operating characteristics curve provides an aggregate measure of performance across all possible classification thresholds. AUC can be interpreted as the probability that the model ranks a random positive example more highly than a random negative example. AUC values closer to 1 correspond to greater performance. Empirical White Paper Appendix A.

[14] For a more detailed discussion of model debiasing, see FinRegLab, The Use of Machine Learning for Credit Underwriting: Market & Data Science Context § 5 & Appendix B.2. For additional discussion of vendor supervision, see Section 5 below.

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

[16] For a more detailed explanation of our evaluation framework, see Empirical White Paper § 3.2.

[17] Specifically, lenders may benchmark the rejected applicant against consumers who were at the threshold for approval or against all applicants, or choose other methods that produce "substantially similar" results. 12 C.F.R Pt. 1002, Supp. I, sec. 1002.9, para. 9(b)(2)-5.

[18] 12 C.F.R Pt. 1002, Supp. I, sec. 1002.9, para. 9(b)(2)-1, -3, -4, -9.

[19] Consumer Financial Protection Bureau, Consumer Financial Protection Circular 2022-03 (May 26, 2022).

[20] In effect, aggregation could become a standardized way of reconciling inconsistencies. Broader data science research has highlighted the need for more consistent approaches where different diagnostic tools disagree as to important features. See Satyapriya Krishna et al., The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective, https://www.researchsquare.com/article/rs-2963888/v1 (May 21, 2023).

[21] See Empirical White Paper § 4.7 for a detailed discussion of our analyses of using the model diagnostic tools to identify changes in features that could improve consumers' chances of approval within one year.

[22] For a description of our tests applying model diagnostic tools to models that have been fed data from a different time era, see Empirical White Paper § 6.6.

[23] Though details vary depending on the source of the principles, transparency, reliability, fairness, privacy, and security are common elements. See, e.g., European Commission, Building Trust in Human Centric Artificial Intelligence (2019) (human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability); Brian Stanton & Theodore Jensen, Trust and Artificial Intelligence, National Institute of Standards and Technology (Dec. 2020) (accuracy, reliability, resiliency, objectivity, security, explainability, safety, accountability and privacy); Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence (2019) (inclusive growth, sustainable development and well-being; human-centered values and fairness; transparency and explainability; robustness, security and safety; and accountability); see also Florian Ostmann & Cosmina Dorobantu, AI in Financial Services, The Alan Turing Institute (2021) (fairness, sustainability, safety, accountability, and transparency).

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

[25] For a more detailed discussion of vendor management expectations, see Financial Health Network, Flourish, FinRegLab & Mitchell Sandler, Consumer Financial Data: Legal and Regulatory Landscape Section V (2020). Such expectations apply most rigorously to banks and may also shape large banks' adoption of vendor-provided models and tools for other types of ML applications, such as fraud and use of automated valuation models.

With support from:

Center for Inclusive Growth

The **Mastercard Center for Inclusive Growth** advances equitable and sustainable economic growth and financial inclusion around the world. The Center leverages the company's core assets and competencies, including data insights, expertise, and technology, while administering the philanthropic Mastercard Impact Fund, to produce independent research, scale global programs and empower a community of thinkers, leaders, and doers on the front lines of inclusive growth. The Center has provided funding to support this research.

JPMorgan Chase & Co.

JPMorgan Chase is committed to advancing an inclusive economy and racial equity. The firm uses its expertise in business, public policy and philanthropy, as well as its global presence, expertise and resources, to focus on four areas to drive opportunity: careers & skills, financial health and wealth creation, business growth & entrepreneurship, and community development.

flourish

Flourish, a venture of the Omidyar Group, has provided operating support to FinRegLab since its inception. Flourish is an evergreen fund investing in entrepreneurs whose innovations help people achieve financial health and prosperity. Established in 2019, Flourish is funded by Pam and Pierre Omidyar. Pierre is the founder of eBay. Managed by a global team, Flourish makes impact-oriented investments in challenger banks, personal finance, insurtech, regtech, and other technologies that empower people and foster a fairer, more inclusive economy.

Digital version available at finreglab.org

Published by FinRegLab, Inc.

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