August 2022

AI for good: Research insights from financial services

Melissa Koide

CEO, FinRegLab

This report is available online at: https://www.brookings.edu/center/center-on-regulation-and-markets/

B | Center on Regulation and Markets at BROOKINGS

The Center on Regulation and Markets at Brookings provides independent, non-partisan research on regulatory policy, applied broadly across microeconomic fields. It creates and promotes independent economic scholarship to inform regulatory policymaking, the regulatory process, and the efficient and equitable functioning of economic markets.

Introduction

Artificial intelligence and machine learning analyses are driving critical decisions impacting our lives and the economic structure of our society. These complex analytical techniques—powered by sophisticated math, computational power, and often vast amounts of data—are deployed in a variety of critical applications, from making healthcare decisions to evaluating job applications to informing parole and probation decisions to determining eligibility and pricing for insurance and other financial services.

The risk that these algorithms make unreliable, unfair, or exclusionary predictions is a foundational concern for a variety of highly sensitive use cases. Furthermore, it raises core questions about whether we can sufficiently understand and manage these models in the immediate and the longer term. Yet artificial intelligence (AI) and machine learning (ML), if carefully overseen and deployed with representative data, also have the potential to increase accuracy and fairness over current models by identifying data relationships that current models cannot detect. Using AI and ML techniques in ways that realize the benefits and mitigate the risks depends on how they are chosen, deployed, governed, and regulated.

Financial services is an important case study, both because of the role that credit and other financial services play in wealth creation and economic mobility and because that sector already has relatively robust regulatory and governance frameworks for managing model fairness, reliability, and transparency. There are nevertheless important questions about whether those frameworks need to evolve to calibrate to the potential benefits and risks of AI and ML adoption. Answering these questions could be instructive for other high-sensitivity AI and ML applications.

FinRegLab[1] has been conducting empirical research and creating platforms for stakeholder dialogue about critical questions concerning the adoption of AI and ML techniques and new data sources for credit underwriting. In partnership with Professors Laura Blattner and Jann Spiess at the Stanford Graduate School of Business, we are currently assessing the performance of diagnostic tools for analyzing and managing machine learning underwriting models to satisfy reliability, fairness, and transparency objectives. As explored in this paper, our findings to date suggest that these technologies hold promise. For instance, for particular tasks, automated approaches performed better than traditional methods of managing for fairness considerations. However, the results overall underscore that thoughtful human oversight at the firm and regulator level is even more critical in managing complex models than for prior generations of predictive algorithms.

Our findings lead us to call for stakeholders to engage in a dialogue centered around three core issues: the consumer experience, fairness and inclusion, and model risk management. These conversations should help to advance how public policy and market practice leverage the accuracy and fairness benefits of machine learning techniques while deploying the technology in ways that are sufficiently transparent. Research and stakeholder dialogue will help to inform a roadmap for evolving market practices and public policy to produce an era of more inclusive and fair credit underwriting.

Financial services as a critical use case for AI/ML adoption

Financial services is one of the most important use cases for AI and ML adoption because access to responsible financial products and services—from payments to credit to savings and investment—is elemental to financial stability and economic mobility. For instance, access to credit can help households bridge short-term gaps between income and expenses as well as make long-term investments in reliable transportation, homeownership, higher education, and small business formation.

Reliance on automated underwriting systems and standardized data sources is already high across the sector; over the last several decades lenders moved away from subjective assessment processes that had been particularly prone to inconsistency and bias. Yet these automated systems created their own concerns about predictiveness, inclusion, and equity, particularly given that lenders have become highly dependent

[1] "Our Process," FinRegLab, accessed July 15, 2022, https://finreglab.org/#process.

on proprietary credit scoring algorithms and standardized data from credit bureaus to assess the default risk of particular applicants.

Traditional data sources have substantial limitations, many of which are more likely to impact low-income households, recent immigrants, and applicants of color. Credit bureau records largely consist of information about how borrowers have repaid prior loans, creating a catch-22 for applicants who do not already have access to the kinds of credit products that are reported to the traditional information system. The data are reported on a lagged basis and do not provide a full picture of applicants' finances, such as inflows and their full range of recurring expenses. In light of the nation's deep and persistent disparities in income and wealth, it is little surprise that Black and Hispanic borrowers are more likely to have negative payment history and lower average scores than white borrowers.[2]

The scores and data are often used to decide which applicants get credit and to charge higher prices to applicants that are at greater risk of default. The ability to charge these prices increases lenders' willingness to provide credit to riskier borrowers, but it can also exacerbate the risk of default by imposing greater financial burdens, particularly on more vulnerable borrowers.

These factors heighten both the interest and the concern with which financial services stakeholders approach the potential adoption of machine learning models and non-traditional data sources for credit underwriting. More than 40 percent of U.S. adults may lack access or pay higher costs for credit today, including nearly 50 million who lack sufficient credit history to be scored under the most widely used models and another 57 million who are scored as subprime.[3] Black consumers are nearly twice as likely and Hispanic consumers 1.5 times as likely to be unscored or subprime as white consumers.[4] These statistics make the stakes larger than whether lenders can reduce losses due to default from more predictive credit models. The potential for model and data innovations to improve—or exacerbate—these disparities could have profound effects on millions of families and small businesses as well as the nation's broader economy.

The potential power—and peril—of machine learning for credit underwriting

Incumbent automated underwriting systems typically rely on regression techniques to identify a relatively limited number of variables with the strongest correlation to loan defaults or other outcomes. Regressions generate a coefficient—essentially a weight—for each variable in the model, which makes it easier to determine each variable's relative importance. However, machine learning techniques can find and map more complex relationships in the data, such as situations in which variables interact with each other in complex ways or do not have a consistent straight-line relationship with the predicted outcome. Machine learning models' ability to detect these complex relationships and to analyze large volumes of diverse types of data can heighten both their predictive power and their potential to better assess the credit risk of people who have been excluded under more traditional data and underwriting models.

A number of studies have found substantial predictiveness benefits and cost savings to lenders from using machine learning models relative to conventional models.[5] Some sources suggest that applying machine learning techniques to traditional data sources could have significant benefits for lenders and consumers alike. For instance, VantageScore reported that its use of machine learning models to assess consumers

[2] In 2019, for instance, the median white household's net worth was eight times larger than that of the median Black household and five times that of the median Hispanic household. Neil Bhutta et al., "Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances," *FEDS Notes*, The Federal Reserve (2020).
[3] Mike Hepinstall et al., "Financial Inclusion and Access to Credit," Experian (2022).
[4] Mike Hepinstall et al., "Financial Inclusion and Access to Credit"; Jaya Dey and Lariece M. Brown, "The Role of Credit Attributes in Explaining the Homeownership Gap Between Whites and Minorities Since the Financial Crisis, 2012–2018," *Housing Policy Debate* 32, no. 2 (March 4, 2022): 275–336, https://doi.org/10.1080/10511482.2020.1818599. The number of consumers with subprime scores was closer to 30 percent prior to the pandemic, but declined substantially in the past two years due to households using stimulus funds to pay down debt and other factors. "Did Uncle Sam's Virus Aid Help Your Credit Score? Don't Count on a Loan," *Reuters*, July 13, 2020, https://www.reuters.com/article/us-health-coronavirus-credit-idUSKCN24E19U.
[5] *The Use of Machine Learning for Credit Underwriting: Market & Data Science Context* (FinRegLab, 2021).

who are otherwise unscorable because they do not have recent credit payment history improved accuracy around 16 percent for bank card originations and 12 percent for auto loans.[6] Other studies suggest that inclusion benefits would be modest or mixed in the absence of new data sources due in part to noisiness in traditional information and the effects of risk-based pricing.[7] The combination of new sources of information about applicants' finances and machine learning techniques could be powerful. For example, one study evaluating machine learning models using both credit bureau data and cash-flow data resulted in cost savings to lenders between 6 and 25 percent of total losses.[8]

At the same time, machine learning models' greater sensitivity to subtleties in underlying data can also heighten concerns about their reliability, fairness, and transparency. Because the models are more complex, it is often more difficult for lenders and regulators to understand, adjust, and monitor for potential issues. Three specific concerns highlight this dilemma:

- Machine learning models' performance in the face of changing economic conditions or shifts in applicant pools may deteriorate rapidly because they are more prone to "overfitting" to the data used in initial development and testing of the ML model as compared to regression models. Current machine learning technologies are not generally well equipped for responding to real world data changes and may not do well in recognizing them.

- AI and machine learning models might replicate, amplify, or introduce new sources of bias. Models that rely on "latent features" identified by the learning algorithm rather than intentionally programmed into the models by developers could reverse engineer applicants' race or gender from correlations in the input data or create complex variables that have disproportionately negative effects for particular demographic groups.

- The complexity of machine learning models makes them more challenging to explain to audiences for purposes of informing their downstream activities. This dynamic affects data scientists, compliance personnel, and regulators who need to perform specific oversight functions, as well as individual credit applicants who are seeking to improve their chances of future credit approvals. Especially for non-technical audiences, explaining which features are influential to particular lending decisions can be difficult when the models rely on data relationships that are inherently complex, non-intuitive, large in number, or dependent on other variables or relationships.

Concerns about these issues are closely intertwined with concerns about the quality of the underlying data and can be heightened when machine learning techniques are applied to non-traditional data sources. Existing concerns about the comprehensiveness and timeliness of traditional credit bureau information help motivate the search for new data sources and modeling techniques. Yet continuing disparities in the traditional credit information system underscore the risks of introducing bias through sources that reflect the impact of historical discrimination or lack observations for key subgroups, as well as through human decisions about how the data is processed and used in model development and deployment.

[6] "Ushering in a New Standard for Credit Scoring" (VantageScore, 2021), https://vantagescore.com/wp-content/uploads/2022/01/4.0-Fact-Sheet-UPDATED-May-2021.pdf.content/uploads/2022/01/4.0-Fact-Sheet-UPDATED-May-2021.pdf.

[7] Laura Blattner and Scott Nelson, "How Costly Is Noise? Data and Disparities in Consumer Credit" (arXiv, May 16, 2021), https://doi.org/10.48550/arXiv.2105.07554.; Andreas Fuster et al., "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance* 77, no. 1 (2022): 5–47, https://doi.org/10.1111/jofi.13090 (finding that a random forest model outperformed linear and non-linear logistic regression models by 1.4% and 0.8% respectively in terms of AUC when using mortgage data for the United States).

[8] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo, "Consumer Credit-Risk Models via Machine-Learning Algorithms," *Journal of Banking & Finance* 34, no. 11 (November 2010): 2768, https://doi.org/10.1016/j.jbankfin.2010.06.001. FinRegLab's own work on bank account and other cash-flow data suggests that it can have substantial predictiveness and inclusion benefits, although that analysis did not involve machine learning techniques. *The Use of Cash-Flow Data in Underwriting Credit: Empirical Research Findings* (FinRegLab, 2019).

How existing regulatory and governance frameworks are shaping AI and machine learning debates and adoption in credit underwriting

While academics, general technology companies, and stakeholders in a broad range of other sectors grapple with questions about the reliability, fairness, and general trustworthiness of AI and machine learning models, financial services stakeholders approach these questions within the specific context of a relatively robust set of existing regulatory frameworks governing the deployment of predictive models in general and the credit underwriting process in particular. These frameworks give financial services stakeholders a set of mechanisms to help answer important questions about AI and ML adoption that could potentially be useful in other sectors. Three regulatory frameworks play a particularly dominant role in shaping how lenders build their governance mechanisms and their approach to adopting machine learning techniques and new data sources for underwriting:

- ***Model Risk Management:*** This regime is a component of broader regulatory requirements to protect the safety and soundness of the banking system. It does not apply to non-bank lenders, though elements may be required by funders and securitizers. Banks are expected to implement a series of governance mechanisms for development, deployment, and ongoing monitoring of models, with the most robust processes for activities that form a significant portion of their business or otherwise pose a high degree of compliance or reputational risk.[9] Many aspects of these processes hinge on different notions of transparency, including documentation of data processing and model development decisions, analyzing whether models are relying on relationships in the data that are intuitive and defensible, and assessing models for "brittleness" in the event that conditions begin to change relative to the data on which they were first trained.

- ***Fair Lending Requirements:*** Federal fair lending laws apply to all lenders, regardless of whether they are banks, and all types of underwriting models, whether they are based on subjective assessments, regression models, or machine learning algorithms. The laws prohibit both treating applicants differently on the basis of race, gender, or other protected characteristics (often called "disparate treatment") *and* using facially neutral practices that have a disproportionately adverse impact on protected classes unless the practices further a legitimate business need that cannot reasonably be achieved through less impactful means.[10] This is often referred to as "disparate impact." The disparate impact doctrine prompts analyses of whether there are significant differences in predicted outcomes among particular demographic groups and whether potential underwriting models can be adjusted in ways that preserve their overall predictive power while reducing those disparities. Historically, the latter analysis has often been conducted by examining which variables are generating the most adverse impact and dropping or transforming the variables to improve outcomes.

- ***Adverse Action Disclosures:*** When lenders reject an application or charge a higher price based on an applicant's risk profile, federal law requires that they explain to the applicant the principal reasons for that decision. These laws apply broadly to all lenders and are designed to serve several purposes, including anti-discrimination, promoting the correction of errors in credit reports, and borrower education. Although the laws and guidance give lenders some flexibility in how they determine which reasons were most important, they generally require lenders to be able to understand and explain how models differentiate between individual applicants, what is often referred to as model examination at the "local" level.

The amount of internal infrastructure that lenders devote to compliance with these regimes varies depending on whether they are subject to safety and soundness requirements, the degree to which they are subject to supervision by federal and state regulators, their size and technical sophistication, and other

[9] "Supervisory Guidance on Model Risk Management" (Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011), https://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf.
[10] 15 U.S.C. § 1691; 42 U.S.C. § 3605; 12 C.F.R. § 1002.6(a); 24 C.F.R. § 100.500.

firm-specific factors. Many of the largest banks have specialized teams with the expertise and resources to conduct comprehensive reviews of documentation submitted for model validation and to develop and test their own models from training data where warranted. In these firms, model developers can expect to defend every significant decision in model design and development prior to putting the model into use.[11] Where lenders decide to rely on vendors to develop models or perform other functions relating to deployment and monitoring, federal regulators have emphasized that the lenders are ultimately responsible for compliance and should exercise appropriate oversight.[12]

These governance and regulatory frameworks have helped to increase the financial sector's focus on data hygiene and to determine machine learning models' reliability, fairness, and transparency at the outset of development and implementation. Relative to some other sectors where the risks of machine learning models have been discovered after they have caused significant problems in deployment, adoption of machine learning models in credit underwriting has been relatively slow. Nevertheless, tens of thousands of consumers and small business owners have their applications for credit assessed and effectively decided by machine learning models each week, and interest in adoption is continuing to grow.

These developments have put financial services stakeholders in a challenging position relative to broader debates occurring in the academic and data science communities about managing the reliability, fairness, and transparency of machine learning models. Research has exploded in recent years both on explainability techniques and on defining and managing for fairness in machine learning models, yet many of these recent advances have not been specifically tailored to or tested in the lending context. For example, explainability techniques that help to explain how complex models operate at a global level may not be suited to producing local explanations of the principal reasons why an individual consumer was predicted to have a high default risk. Some fairness methods that have been raised in the larger literature involve building demographic information directly into predictive models, which raises potential questions under the "disparate treatment" prong of fair lending laws that prohibits discrimination on the basis of protected characteristics.[13]

In the absence of such information, financial services stakeholders are engaged in targeted research and dialogue efforts about various aspects of the reliability, fairness, and transparency of machine learning models and about how the existing frameworks may need to evolve to meet new challenges and opportunities. These efforts are occurring in parallel with similar work in other sectors and professional communities. While there can be substantial crossover at times, the existence of the specific regulatory regimes discussed above creates some differentiation.

Transparency as a threshold issue

Transparency is a critical issue in the adoption of machine learning credit underwriting models. An important value in its own right, it can also be instrumental to how stakeholders diagnose and manage other types of concerns. Adverse action notices are a direct transparency requirement. Historical practice in complying with fair lending and model risk management requirements has relied heavily on analyzing what particular variables are driving model predictions and disparities to help inform potential management and mitigation strategies. Without sufficient transparency, it is difficult for lenders, investors, regulators, and other stakeholders to determine whether a model is being used responsibly and fairly.

While the regulatory frameworks that apply to credit underwriting incorporate multiple types of transparency for different purposes, they do not define specific thresholds for specific settings. Even for adverse action notices, regulatory guidance emphasizes that various methods will meet the legal requirements to identify the "principal" reasons for a particular credit decision, for instance by benchmarking the individual consumer against all applicants or against applicants who just barely met approval thresholds. The guidance also

[11] *The Use of Machine Learning for Credit Underwriting: Market & Data Science Context* (FinRegLab, 2021).
[12] Board of Governors of the Federal Reserve System, "Supervisory & Regulation Letter 11-7: Supervisory Guidance on Model Risk Management," 2011, https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm.
[13] Deborah Hellman, "Measuring Algorithmic Fairness," SSRN Scholarly Paper (Rochester, NY, July 11, 2019), https://papers.ssrn.com/abstract=3418528.

emphasizes that disclosures should describe factors actually considered by a scoring model or reviewed by the human decisionmaker, but it does not specify particular thresholds for accuracy.[14]

The particular challenges posed by the prospect of explaining complex machine learning models involving hundreds of variables for multiple purposes are leading financial services stakeholders to take a fresh look at transparency requirements and recognize these challenges even with traditional models and approaches. For instance, while machine learning models tend to be more complex than traditional underwriting algorithms, there can also be transparency and model management challenges in working with logistic regression models that assess dozens of variables. Stakeholders have focused renewed attention on the fact that while adverse action notices provide some level of transparency to consumers, the law does not require them to highlight factors that consumers might be able to change to improve the likelihood of future acceptance or even to explain how or why particular variables had a negative effect.

These considerations emphasize the importance of defining and implementing transparency requirements with an eye toward the underlying purpose and audience, rather than at a generic level. Generating an accurate, detailed, and often technical explanation of how models produced particular outcomes may not always be sufficient to communicate meaning, facilitate understanding, and inform credit applicants. While financial services stakeholders have been monitoring broader computer science developments that are working to create tools to explain complex models, the lack of testing in the specific credit context is creating substantial uncertainty and stakeholder dialogue about how to meet compliance expectations and to manage for broader reliability and fairness concerns.

While these discussions and explorations play out, model developers are using a range of approaches to meet transparency needs as they work to realize the potential benefits of machine learning techniques in credit underwriting. Some are using machine learning algorithms only to identify individual features or variables that they use in more simple, traditional underwriting models, which can provide greater transparency and manageability at the sacrifice of some predictive insights. Others are using machine learning models to make underwriting decisions but choosing structures that have a higher degree of transparency by nature of their structure and design (sometimes called inherently interpretable or self-explanatory models). Some lenders are using a combination of up-front model constraints and various secondary or “post hoc” methods to assess how the models are generating their predictions. At the far end of the spectrum, some are using complex “black box” models and relying heavily on post hoc tools to assess their operations.[15]

The debate over where the market and regulatory expectations should evolve along this spectrum is heated as stakeholders disagree about whether constraints on model structure unduly reduce predictiveness and about the capabilities and performance of particular post hoc tools. In effect, the tools create a second layer of questions about whether we can rely on the tools to tell us if we can trust the underlying models.

These broad debates informed FinRegLab’s partnership with professors Laura Blattner and Jann Spiess of the Stanford Graduate School of Business to conduct threshold research into tools for managing transparency, fairness, and reliability concerns with machine learning credit underwriting models. Understanding how these tools perform on various credit-related tasks could potentially be useful in managing both traditional and next-generation underwriting models going forward. The tools could also have a substantial effect in shaping whether and how machine learning techniques are implemented in this market, with substantial implications for credit access and equity as described above.

Initial research findings and implications for public policy and market practice

Our research is designed to evaluate the performance of proprietary explainability tools provided by seven vendors in the market—Arthur AI, H20.ai, Fiddler AI, Relational AI, Solas AI, Stratyfy, and Zest AI— as well

[14] 12 C.F.R. § 1002.9(b)(2).
[15] *The Use of Machine Learning for Credit Underwriting: Market & Data Science Context* (FinRegLab, 2021).

as several open-source explainability tools built by the research team as applied to various tasks in connection with the three regulatory requirements discussed above. The goal of the research is to provide a broad sense of the range of approaches and outcomes that are possible in the context of machine learning as the financial services sector develops norms and rules to govern the responsible, fair, and inclusive use of machine learning for credit underwriting. Toward that end, the research was informed by financial services stakeholders—including executives from banks and fintechs, technologists, consumer advocates, and regulators.[16]

To conduct the analysis, the research team built four credit card underwriting models of varying degrees of complexity—a logit regression with over forty variables, a simple neural network trained on a small number of variables, an XGBoost model, and a complex neural network trained on several hundred variables—using data provided by one of the three major credit bureaus.[17] The models were applied to additional sets of credit records to assess which consumers in the sample would be accepted or rejected by the models. The explainability tools were then used to perform various model diagnostics, such as identifying which factors were important in rejecting a particular applicant or producing disparities among the default predictions for minority and non-minority candidates.[18] The research team is analyzing the tool outputs to better understand how they performed on specific tasks. We released initial results concerning use of the tools in the adverse action and disparate impact contexts in April 2022 and are working on a second report to cover model risk management and other follow up research.

Our results underscore the importance of marrying human oversight with the use of advanced tools and techniques to account for particular opportunities and challenges in working with machine learning models. We found that some explanatory tools were able to identify features that related to individual adverse credit decisions and that some tools identified features that explained a significant part of the disparities in default predictions among minority and non-minority consumers.[19]

However, other tools did not perform as well, and no single tool or approach was always the best choice across the two consumer protection requirements we analyzed.[20] The results suggest that the human choice of a particular tool for a particular task tends to be especially consequential for more complex models involving more variables.

The results underscored the importance of careful interpretation of tool outputs in the context of broader feature correlations and interdependencies. Tools that performed the best at identifying significant drivers of adverse action decisions and model disparities had the highest degree of overlap, but they did not always identify the same variables as being most important. Our initial results highlight that critical examination of the output of diagnostic tools is essential to account for the potential significance and role of correlated features for the specific tasks the users are performing.

---

[16] See finreglab.org to view FinRegLab's *Machine Learning Explainability & Fairness: Insights from Consumer Lending* empirical white paper.

[17] Several of the participating companies opted to provide additional machine learning models that trained using the same data as the baseline models developed by the research team.

[18] Because collection of applicants' demographic information is restricted by fair lending laws, race and ethnicity were imputed based on last name and geography similar to a technique that is used by regulators and lenders in compliance monitoring. "Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity" (Consumer Financial Protection Bureau, 2014), https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/. Scott Cranfill, "CFPB Proxy Methodology," GitHub repository (2017).

[19] Because collection of applicants' demographic information is restricted by fair lending laws, race and ethnicity were imputed based on last name and geography similar to a technique that is used by regulators and lenders in compliance monitoring. Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity" (Consumer Financial Protection Bureau, 2014).

[20] The research team tested the fidelity of the tools in several ways. For instance, in evaluating the results for consumers whose default predictions would have caused them to be rejected for credit, they identified "nearest neighbors" who were most similar with regard to the variables that the tools had identified as most important to the prediction to analyze whether the neighbors also would have been rejected under the model. The team also "perturbed" the values of the identified variables to see the effect on the model's prediction and compared the amount of change to what happened when they perturbed a set of randomly selected variables or a set of variables that were closely correlated to the identified features. In the disparate impact context, the team both perturbed identified drivers and created a data set with different distributions of consumers to evaluate how much impact the identified drivers had on the model outcomes.

A third piece of the analysis considered different approaches for mitigating disparities identified in the models in ways that preserve predictive accuracy but reduce disparities among minority and non-minority applicants. These approaches involved a variety of techniques and varied as to how deeply they relied on the previous diagnosis of which particular variables were driving the disparities. Among the options tested, we found that more automated approaches for generating a range of alternative models were more effective in finding options that reduced disparity levels with some modest reduction in predictive accuracy, while approaches that dropped out or reweighted particular features produced little improvement in disparities and often significant declines in accuracy. The research found that the automated approaches generated fairness improvements for the models used in the study that both increased the number of approved minority applicants and decreased the number of improperly rejected minority applicants.

The automated approaches differed in whether and how they use protected class information to search for less discriminatory models. For instance some used protected class information to evaluate for disparities after the models were trained. In other cases, protected class information was used in model development. This difference reflects a range of views among financial services stakeholders on whether fair lending laws permit lenders to use protected class information in the model development process to address potential disparities.

Our initial findings underscore the importance of implementing comprehensive governance regimes that draw on experts with significant domain knowledge about the data, models, and post hoc techniques. Human judgment is critical both in choosing the right post hoc tool for the right task and considering outputs carefully in light of data correlations, particularly when dealing with complex models. At the same time, the findings suggest that thoughtful use of automated approaches may be more effective in managing the fairness of machine learning models than more traditional manual approaches that focus on identifying and manipulating individual features in isolation.

Looking forward

A broad range of stakeholders are hungry for additional empirical research grounded in the financial services context and credit in particular. FinRegLab is following up on its initial work with additional analysis of several issues, including exploring the effect of correlated features and potentially evaluating the use of AI and ML with cash-flow data. Other academics and industry stakeholders are testing mathematical and data science techniques to address transparency needs and to explore how certain AI and machine learning techniques may help to reduce disparities among demographic groups.[21]

In addition to answering these factual and technical questions, stakeholder dialogue and education will be critical to advancing responsible and inclusive use of AI and machine learning in financial services. The evolution in mathematical techniques and data sources is helping to highlight policy tensions within the existing frameworks, as well as suggesting potential alternative processes and tools for managing some of these tensions. Toward this end, civil rights organizations, governments, and advocates are crafting principles and frameworks for auditing AI systems and the use of more representative data,[22] and lawmakers and regulators are soliciting feedback about the use of AI in the financial system.[23]

These initial conversations are encouraging, but much hard work remains to determine how to evolve existing policy and governance frameworks around three core issues: fairness and inclusion, the consumer

[21] See, e.g., Emily Black, Manish Raghavan, and Solon Barocas, "Model Multiplicity: Opportunities, Concerns, and Solutions," SSRN Scholarly Paper (Rochester, NY, January 21, 2022), https://doi.org/10.2139/ssrn.4142472.
[22] See, e.g., Michael Akinwumi, Lisa Rice, and Snigdha Sharma, "Purpose, Process, and Monitoring: A New Framework for Auditing Algorithmic Bias in Housing & Lending" (National Fair Housing Alliance, 2022), https://nationalfairhousing.org/wp-content/uploads/2022/02/PPM_Framework_02_17_2022.pdf; "AI Risk Management Framework," (National Institute of Standards and Technology, 2021) https://www.nist.gov/itl/ai-risk-management-framework.
[23] See, e.g., "Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning," Federal Register, March 31, 2021, https://www.federalregister.gov/documents/2021/03/31/2021-06607/request-for-information-and-comment-on-financial-institutions-use-of-artificial-intelligence; U.S. Congress, House, Task Force on Artificial Intelligence, "Equitable Algorithms: How Human-Centered AI Can Address Systemic Racism and Racial Justice in Housing and Financial Services," 117th Cong. (2021).

experience, and model risk management. These conversations require careful consideration of how to define and balance broad goals such as accuracy, fairness, and transparency in particular situations, as well as how to provide meaningful guidance regarding technical compliance expectations at a time when techniques and market practices are evolving. Three examples are useful to illustrate the potential need for nuanced adjustment to existing policy frameworks:

- **The quest for less discriminatory alternatives:** Our research and that of others suggest that machine learning techniques and automated approaches can produce underwriting model options with less disparity between minority and nonminority populations as compared to more manual processes for minimizing differences among demographic groups.[24] However, some of these approaches rely upon using protected class information—or a proxy for it—in different ways and stages of model development in order to find the less discriminatory alternatives. This raises important legal and regulatory questions about whether and how demographic characteristics (or other protected class information) are allowed for use in finding fairer models. Related regulatory questions have also been raised by academics, advocates, and lenders as to how to choose between options that may reduce disparities among some demographic groups but potentially worsen them for others.[25]

- **The precision of adverse action notices:** Similarly, research on adverse action compliance raises questions about how best to achieve the underlying policy goals of helping to prevent discrimination, educate borrowers, and facilitate correcting errors in credit reports. Our research and that of others suggests that post hoc diagnostic tools vary somewhat in their rankings of which features were "principal" for purposes of an individual credit decision. However, the original regulations and guidance provide lenders some latitude in choosing between different analytical methodologies and benchmarks. This raises questions going forward as to the importance of precision in ranking features, the extent to which standards can be clearly defined and enforced at a time when techniques are evolving, and the importance of pursuing greater consistency as compared to other potential policy objectives such as providing applicants with more information about how particular features affected their risk assessments and/or how to improve their chances on future applications.

- **Guardrails for data access and use:** Access to data that are representative of historically underserved populations is essential for a more inclusive credit marketplace. Federal financial services regulators issued a joint statement in 2019 that highlighted the use of alternative data—cash flow data in particular—as a more representative type of data. Even still, stakeholders would like additional guidance as to what types of data and particular variables (within data types) are acceptable to use.[26] Modernizing the rules that govern data access and data flows would meaningfully help to provide clarity and consistency to the financial sector.

The questions we are grappling within financial services may also help to inform some of the public policy and societal questions raised by AI and ML in other sensitive use cases. Efforts to define good governance of models and data in financial services will also help to inform AI and machine learning governance and data use in other sectors where important fairness and inclusion decisions are also being determined. As financial services stakeholders contend with whether the use of protected class information can be safely used to generate less disparity-inducing models, these insights may be helpful in informing how employers and others use algorithms and certain data to identify job applicants in fairer and balanced ways. And likewise, AI and ML research and dialogue insights from the healthcare, employment, and other sectors will be instructive for the public policy and empirical questions raised in financial services.

[24] See, e.g., Black, Raghavan, and Barocas, "Model Multiplicity."
[25] Talia B. Gillis, "The Input Fallacy," SSRN Scholarly Paper (Rochester, NY, February 16, 2021), https://doi.org/10.2139/ssrn.3571266.
[26] E.g.: Stephanie Wake on behalf of Bank Policy Institute members, "Re: Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning," June 25, 2021, https://www.regulations.gov/comment/OCC-2020-0049-0020, 14.

Conclusion

Now is the time to deepen our understanding of how to use AI and machine learning responsibly and inclusively. Although there are many technical and policy questions still to be answered, it is clear that rigorous governance frameworks are even more important than for prior generations of predictive algorithms. Model and data oversight by experts who know the use case, who understand the environment, and who can curate and manage the data, model selection, and deployment are essential. Independent, ongoing evaluations of models' outcomes and prompt adjustments as underlying circumstances change also reduce the risk of performance deterioration and unfair or exclusionary results. Both within firms and within regulatory agencies, deep domain expertise, a broad diversity of experience and backgrounds, and engagement throughout the organization are critical ingredients. Charting a path forward will require both rigorous research and sustained engagement from a full spectrum of stakeholders.

Such processes would be facilitated by a recognition that the technology is neither a cure-all nor a poison. This is not quick or easy work, but it is essential to producing a fairer and more inclusive future.

Center on
Regulation and Markets
at BROOKINGS

The Center on Regulation and Markets at Brookings provides independent, non-partisan research on regulatory policy, applied broadly across microeconomic fields. It creates and promotes independent economic scholarship to inform regulatory policymaking, the regulatory process, and the efficient and equitable functioning of economic markets.

Questions about the research? Email communications@brookings.edu.
Be sure to include the title of this paper in your inquiry.