The Use of Machine Learning for Credit Underwriting: Market & Data Science Context
OVERVIEW

Machine learning models are already being used to evaluate the creditworthiness of tens of thousands of U.S. consumers and small business owners each week. The models' greater accuracy and capacity to analyze large, varied forms of data create the potential to increase access to credit for millions of people – including disproportionately high numbers of Black, Hispanic, and low-income consumers – who are difficult to assess using traditional models and information.[1]

But machine learning models' greater complexity also increases concerns that they may not perform well under changing conditions and could replicate or even exacerbate past discrimination, as evidenced by experience with artificial intelligence and machine learning in other sectors.[2] Model transparency has emerged as an urgent question for assessing when machine learning can be trusted for use in such a sensitive application as credit decisioning and for facilitating sound governance of these models.

As the first step in a groundbreaking evaluation of the explainability and fairness of machine learning underwriting models, FinRegLab has released "The Use of Machine Learning for Credit Underwriting: Market & Data Science Context" to describe the current use of machine learning underwriting models and the choices firms are making in developing, implementing, and monitoring those models. Although artificial intelligence and machine learning are sometimes viewed as "black boxes" operating without human intervention or supervision, the report catalogues the variety of decisions that developers have to make even with regard to highly complex underwriting models. Effective oversight of those decisions is important for answering core questions about the models' reliability and fairness, and may require different tools and processes as compared to conventional systems.

This Overview document summarizes key findings and issues raised in the main report, as well as the broader project. FinRegLab's forthcoming empirical research with Professors Laura Blattner and Jann Spiess of the Stanford Graduate School of Business will assess the capabilities and performance of a set of proprietary and open-source model diagnostic tools in helping lenders comply with model risk management, fair lending, and adverse action reporting requirements. It is the first public research shaped by input from key financial services stakeholders – including executives from banks and fintechs, technologists, consumer advocates, and regulators – to address questions about explainability and fairness that are likely to shape the adoption of machine learning underwriting models going forward.

Market Context

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

- **Lenders are primarily attracted to machine learning models' potential to improve the accuracy of credit risk assessment**, as well as to reduce losses, streamline the process of updating and refitting models, and keep pace with market competitors. Many also cite the ability of machine learning models to leverage large, diverse datasets as a motivation. Nonbank usage is more established due to a number of factors, including reliance on digital business models, newer lending platforms, and differences in the nature and maturity of risk management and oversight processes.
- **Credit cards and unsecured personal loans are the markets in which the use of machine learning models to make credit decisions is most advanced**. This reflects the historical position of credit cards as being at the analytical forefront of consumer finance and the dominance of digital lending in unsecured personal loans. Auto lending and small business lending are also areas where machine learning underwriting models are in use.
- **Concerns about the ability to explain and understand machine learning underwriting models shape every stage of their development and use.** To facilitate management and oversight, some firms are imposing upfront constraints on their machine learning models to reduce their complexity and improve transparency. Other lenders are using *post hoc* explainability methods – supplemental models, analyses, or visualizations – to make complex or "black box" models more transparent. Explainability technologies are evolving quickly, and stakeholders are debating the tradeoffs of different approaches.
- **Firms and regulators are also focusing on whether and in what circumstances the use of machine learning can improve fair lending oversight.** Financial services stakeholders are particularly intrigued by the potential for machine learning techniques to improve available tradeoffs between performance and fairness when mitigating sources of adverse impacts in credit decisions.
- **Third-party service providers are entering the market to facilitate model development and management functions by smaller firms.** Many firms are likely to lack the resources – foremost among them personnel with appropriate data science and credit expertise – to develop and operate their own machine learning underwriting models.[3] To support those firms, a number of score providers, technology firms, and consulting firms are offering various forms of support. Some offer model diagnostic tools as a stand-alone product, while others provide those tools in the context of model development services.

Yet while interest in machine learning underwriting models is accelerating, the scope and pace of adoption going forward will depend on the extent to which various stakeholders can answer fundamental questions about the capabilities and trustworthiness of machine learning models and about how to enable necessary oversight. Concerns about the trustworthiness of machine learning models are being raised in a broad range of sectors with regard to general transparency, reliability, fairness, privacy, and security. But they are particularly pressing in credit underwriting because existing legal and regulatory frameworks force consideration of risk management questions more holistically and at an earlier stage than occurs

elsewhere. The balance of the report focuses on outlining the choices that lenders face in developing, implementing, and monitoring machine learning models and emerging developments on explainability and fairness from the broader data science community that may help to shape market and regulatory practices concerning machine learning underwriting models.

Model Transparency

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

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

Concerns about model transparency shape lenders' decisions at every stage of the process of developing, implementing, and managing machine learning underwriting models. Model developers may in effect work backwards from the transparency requirements of their use case – by designing and planning their modelling approach based on the level and type of transparency required. In practice, the developer of an underwriting model needs to be able to establish that each relationship in the model has an intuitive, defensible relationship to an applicant's likelihood of default. Further, given the need to deliver accurate adverse action notices, firms need the capacity to pinpoint the primary bases of individual credit decisions.

Model developers can use a variety of tools and techniques to build a model with the necessary transparency in whatever type of machine learning they choose for their underwriting model. They might develop an inherently interpretable model – one that can be explained and understood on its face without additional analysis. These models result from constraints on the learning algorithm that will limit the final model to considering certain kinds of relationships or ensure it has certain kinds of characteristics that improve its transparency. Examples include monotonicity[4] and linearity[5] constraints, which simplify in various ways the relationship between input variables and the predicted outcome. Alternatively, developers might build an explainable model – one that uses more complex or black box models alongside *post hoc* explainability methods or supplemental models, analyses, and techniques designed to improve

the transparency of such models.[6] Examples include machine learning techniques designed to identify the contribution that specific aspects of the model makes to its prediction such as Shapley values (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and various kinds of data visualization plots.

The choice between inherently interpretable models and models that require *post hoc* explainability methods has shaped early adoption of machine learning underwriting. Firms and researchers alike are working to understand better whether lenders should use inherently interpretable models or pair less interpretable models with supplemental explainability methods to satisfy transparency needs. Proponents of using only inherently interpretable models argue that well-designed models of this kind perform as well as more complex models and deliver the necessary transparency. Importantly, they do so without relying on secondary techniques and analyses that introduce further uncertainty and a second layer of trustworthiness questions.[7] Proponents of inherently interpretable models also commonly question whether adding a second layer of analytical complexity compounds, rather than resolves, the challenge of establishing the trustworthiness of AI and machine learning systems and can meet specific transparency requirements.[8]

Proponents of complex models that rely on *post hoc* explainability techniques argue that this approach has the potential to deliver superior predictive accuracy including for evaluating borrowers who have traditionally been excluded without limiting a lender's ability to meet model transparency needs and satisfy relevant regulatory requirements.[9] Industry proponents also argue that even so-called inherently interpretable models run the risk of being too complicated for a human to fully interpret.[10]

Fairness and Bias

Without thoughtful design and oversight, problematic biases can be built into machine learning systems and amplify the effects of discrimination in a range of everyday decisions and activities. Statistical biases come in many forms, and there is no standardized taxonomy.[11] Weaknesses in data, model design, and governance/personnel can also create feedback loops and magnification effects that make it difficult and in some cases impossible to pinpoint a single cause of bias or discrimination.[12] These biases can result from decisions made about what kind of machine learning techniques to use; how to use those techniques to develop, operate, and monitor models; what kind of data to use; and how to prepare data for use by machine learning models. Where biases occur, they can result in faulty predictions and give rise to fair lending, discrimination, and inclusion issues in various circumstances.

The shift to machine learning from incumbent underwriting models may amplify the importance of some kinds of biases, but it also presents an opportunity for practitioners and policymakers to rethink how underwriting models are developed and how new technologies and data can be used to help overcome, rather than further entrench, past patterns of bias and discrimination. For example, adoption of machine learning underwriting models may have the potential to improve identification of discrimination risks and to offer superior mitigation options when those risks are detected.[13] This may result in detecting risks not registering fully in current processes and enabling the use of models that retain the predictive power of variables and relationships causing disparities instead of having to eliminate those features entirely. Unlike incumbent underwriting models, the development of machine learning models enables

consideration of many iterations of a model, including many changes to a model's specifications, which can enhance predictive power and enable more explicit consideration of certain tradeoffs. Lenders can assess those iterations to find "less discriminatory models that maintain their predictive ability." [14]

The transition to machine learning is also inspiring consideration of how to incorporate growing sophistication in approaches to measuring algorithmic fairness in model development and oversight processes. Regulatory oversight in financial services applies well-established definitions to assess fairness in the form of disparate treatment and disparate impact requirements.[15] However, the broader community of machine learning researchers and practitioners have developed more than 20 mathematical approaches to measuring the fairness of algorithmic models. These measures are subject to a range of practical challenges with respect to use in financial services – ranging from data availability to tension with existing anti-discrimination and risk management requirements.

As a result, these metrics primarily function today as analytical tools that help firms gain insight into various aspects of a model's operations and effects in the iterative process of developing and reviewing models.[16] This means that efforts to understand what one metric might say about a model's fairness may be limited to early stages of development and occur separate and apart from traditional analyses to assess fair lending compliance prior to or after deployment. Traditional governance processes remain important, underscoring that defining and measuring fairness are deeply contextual.[17]

Data scientists have also produced a variety of methods for debiasing machine learning models that can be used at several different stages of the model development process. However, in practice, lenders face uncertainty when considering whether and how to use methods described in the report, which has substantially chilled substantial inquiry into how these methods might be used in underwriting absent clarification from regulators. Some methods require use of protected class information in ways that create tension with existing anti-discrimination laws. Other methods may undercut established risk management expectations. For example, banks' fair lending compliance teams are generally expected to conduct independent evaluations of lending decisions using real or imputed protected class information that is not available to model development teams. There is uncertainty about whether making this information available to model developers for the purpose of model debiasing is a practice that could potentially subject firms to regulatory criticism for compliance risk management weaknesses in addition to creating exposure to disparate treatment claims.

Future Research

A forthcoming evaluation of the explainability and fairness of machine learning underwriting models from FinRegLab and a team of researchers from the Stanford Graduate School of Business will assess the capabilities and performance of a set of proprietary and open-source model diagnostic tools in helping lenders comply with model risk management, fair lending, and adverse action reporting requirements. This study will assess the capabilities and performance of various model diagnostic tools designed to support responsible use of machine learning underwriting models across a variety of dimensions:

- **Type of Machine Learning Model**: Benchmark underwriting models will range from logistic regression and boosted trees to neural networks and ensemble models to identify whether the

type of underwriting model being explained affects the accuracy and utility of information produced by the model diagnostic tools;

- **Model Complexity**: Each form of machine learning being evaluated will have simple and complex forms to help us identify the tradeoffs, if any, between performance and transparency and between performance and fairness;
- **Changes in Economic Conditions**: Test datasets will simulate different economic environments, such as data from 2009-2010, to help assess whether the model diagnostic tools can help lenders identify changes in data conditions and model performance once in operation; and
- **Shifts in Applicant Distribution**: Test datasets will encompass different kinds of borrowers with respect to geographic location and socioeconomic status to help us evaluate how well these tools detect fair lending and other risks.

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

In addition to the empirical research results, FinRegLab expects to conduct in-depth analysis of the implications of that research for law, regulation, and market practices in 2022.

Endnotes

[1] For instance, more than 50 million U.S. adults lack sufficient traditional credit history to generate scores under the most widely used models, and an even larger group may struggle to access credit because they are considered "non-prime." Information limitations also make it more difficult for millions of small business owners to obtain credit. FinRegLab, The Use of Cash-Flow Data in Underwriting Credit: Market Context & Policy Analysis 12-14 (2020).

[2] *See, e.g.,* Steve Lohr, Facial Recognition Is Accurate, If You're a White Guy, N.Y. Times (Feb. 9, 2018); Ed Yong, A Popular Algorithm Is No Better at Predicting Crimes than Random People, The Atlantic (Jan. 17, 2018); Starre Vartan, Racial Bias Found in a Major Health Care Risk Algorithm, Scientific American (Oct. 24, 2019).

[3] *See, e.g.,* Cornerstone Advisors, Credit Monitoring and the Need for Speed: The Case for Advanced Technologies 4, Figure 4 (Q2 2020) (survey of 175 LendIt subscribers finding that 20% of institutions had no in-house staff for credit modelling and that even among large institutions, just 16% had four or more full time modelers).

[4] Adding salt to a savory dish presents an intuitive example of a non-monotonic relationship, which means that the relationship is not one-directional. A small amount of salt will generally make the dish taste better. However, after a certain point, adding salt will make the dish taste worse.

[5] A non-linear relationship is one in which increases or decreases in an input variable do not always produce proportionally consistent changes in the target or output variable.

[6] The terms interpretable AI and explainable AI, much like the underlying terms interpretability and explainability, are used differently among various stakeholder communities. This overview adopts usage of the main report.

[7] *See, e.g.,* Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Machine Intelligence 206-215 (May 13, 2019); Scott Zoldi, Not All Explainable AI is Created Equal, Retail Banker International (Oct. 9, 2019).

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

[9] *See, e.g.,* Weiwei Jiang & Jiayun Luo, An Evaluation of Machine Learning and Deep Learning Models for Drought Prediction Using Weather Data, preprint submitted to J. of LATEX Templates, arXiv:2107.02517v1 (2021); Rishi Desai *et al.,* Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims with Electronic Medical Records to Predict Heart Failure Outcomes, 3 JAMA Network Open (2020).

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

[11] Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 Cal. L. Rev. 677-693 (2016).

[12] *See, e.g.,* Betsy Anne Williams *et al.,* How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications, 8 J. of Information Policy 78-115 (2018); Aylin Caliskan *et al.,* Semantics Derived Automatically from Language Corpora Contain Human-Like Biases, 356 Science 183-186 (2017); Sara Hooker, Opinion, Moving Beyond "Algorithmic Bias Is a Data Problem," Patterns (Apr. 9, 2021).

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

[14] BLDS, LLC, Discover Financial Services, & H2O.ai, Machine Learning: Considerations for Fairly and Transparently Expanding Access to Credit 6, 22 (2020)

[15] The Equal Credit Opportunity Act (ECOA) prohibits discrimination in "any aspect of a credit transaction" for both consumer and commercial credit on the basis of race, color, national origin, religion, sex, marital status, age, or certain other protected characteristics, and the Fair Housing Act (FHA) prohibits discrimination on many of the same bases in connection with residential mortgage lending. See 15 U.S.C. § 1691(a); 42 U.S.C. § 3605.

[16] *See, e.g.,* Upstart, Response to Agencies' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning 17-19 (July 1, 2021).

[17] Sandra Wachter *et al.*, Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI, Computer L. & Security Rev., arXiv:2005.05906, (2021).