Explainability & Fairness in Machine Learning for Credit Underwriting

Policy & Empirical Findings

Research finds that tools for managing explainability and fairness in machine learning underwriting models hold promise and that regulatory guidance could encourage more consistent, responsible use

Machine learning (ML) underwriting models' accuracy and capacity to analyze large datasets create the potential to increase access to credit for millions of people who are difficult to assess using traditional models and information – including disproportionally high numbers of Black, Hispanic and low-income consumers. Yet the very quality that fuels machine learning models' 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.

While use of ML models is accelerating in some lender segments, market actors and policymakers are grappling with critical questions about our capacity to understand and manage these models to ensure they are safe, fair, and reliable for use in underwriting. To explore these issues, FinRegLab conducted extensive qualitative and quantitative research and policy analysis.

Empirical Research

Lenders using machine learning to extend consumer credit often rely on a range of ancillary explainability techniques and tools developed over the last decade to make what are sometimes called "black box" models more transparent. These tools are being used to provide information needed to manage model behavior in compliance with federal requirements that have long applied to credit decisions. For example, lenders need to be able to reliably identify what is driving the model's prediction *for behavior of a specific loan applicant*; what is driving the model's predictions for *particular groups of applicants*, including minorities; and what is driving the *overall behavior of the model.*

FinRegLab worked with Professors Laura Blattner and Jann Spiess of the Stanford Graduate School of Business to evaluate whether and in what circumstances the information produced by available explainability tools and techniques can help lenders manage machine learning underwriting models as required by law. The study also explored the use of different approaches to managing fair lending concerns, including use of machine learning "debiasing" techniques as well as traditional lender strategies.

The project applied model diagnostic tools from seven technology companies—Arthur, H2O.ai, Fiddler, RelationalAI, SolasAI, Stratyfy, and Zest AI—as well as several open-source tools to a spectrum of underwriting models that were built for purposes of this study. The team then analyzed the information produced by the diagnostic tools in various ways, for instance by comparing them to each other to see if the various tools produced consistent information when answering the same questions about the same underwriting models.

The team published an initial version of *Machine Learning Explainability & Fairness: Insights from Consumer Lending* in April 2022 and an updated version in July 2023. The expanded report evaluates the use of explainability techniques for model risk management expectations for banks, in addition to the earlier analyses focusing on activities relating to generating consumer disclosures and fair lending compliance. It also updates the earlier sections with additional analyses of tool consistency.

The study suggests cautious optimism about the ability of various explainability and debiasing techniques to help lenders manage machine learning underwriting models. In particular, the study found:

- Lenders can systematically evaluate the performance of explainability techniques and model diagnostic tools without having "ground truth" explanations, for instance by comparing the tool outputs to each other and to objective benchmarks.

- Even for complex machine learning models, some explainability techniques and tools reliably identified features that were important for generating consumer disclosures, fair lending analyses, and assessing the overall operation of the model for risk management purposes.

- However, there was no "one size fits all" technique or tool that performed the best across all regulatory tasks. Rather, it is important to choose the right explainability tool for the particular ML model and task, to deploy it in a thoughtful way, and to interpret the outputs with an understanding of the underlying data.

- In the fair lending context, approaches that relied on traditional mitigation strategies focusing on a narrow subset of features produced little to no improvement in fairness and substantial declines in predictiveness. But more automated approaches (including machine learning debiasing techniques) were able to produce a menu of options that provided larger fairness benefits with smaller accuracy tradeoffs.

While these results are encouraging, the study makes clear that using a machine underwriting model raises the stakes for governance decisions throughout model development, implementation, and monitoring. Lenders' decisions about what type of ML model to use, how much complexity to enable in that model, and what techniques and tools to use in the development process necessarily shape their strategies for describing, managing, and monitoring model behavior.

Policy Analysis

In addition to the empirical analyses described above, FinRegLab has conducted extensive market context interviews, convened policy working groups, and analyzed broader policy questions about adjusting market practices and regulatory frameworks to account for machine learning adoption. Recalibrating market practices and policy expectations for the use of machine learning underwriting models could provide opportunities to address longstanding concerns about prior generations of predictive credit models and the compliance frameworks that govern them.

The resulting policy paper, *Explainability & Fairness in Machine Learning for Credit Underwriting: Policy & Empirical Findings Overview*, describes emerging policy dialogues across several regulatory areas as lenders, advocates, policymakers, and stakeholders consider differences between the information and techniques used to develop, manage, and validate traditional underwriting models as compared to machine learning models. The paper highlights topics on which regulatory guidance could encourage more consistent, responsible use of machine learning models and explainability and fairness techniques, even as the underlying technologies and research are continuing to evolve. Potential areas of focus include:

- Regulatory expectations for when and how lenders search for fairer or "less discriminatory alternative" models.

- The permissibility of using specific debiasing techniques in the ML context under fair lending laws to the extent that they use demographic data in different ways than traditional mitigation approaches.

- The general qualities that lenders should manage for in evaluating explainability technique performance or the permissibility of using specific techniques for specific tasks, such as using them to help generate consumer disclosures.

- Other issues relating to general model governance, including addressing challenges in validating vendor-provided models and tools and governance expectations for nonbank lenders.

Key Considerations

The research focuses on consumer credit as a case study in part because federal regulatory frameworks force potential users of machine learning to resolve questions about model transparency earlier and more holistically than in other sectors. However, the analyses are also potentially useful to other sectors where machine learning predictive models are being used to make important decisions, such as medicine, criminal justice, and employment.

The purpose of the research was not to identify a particular "winner" among the available tools or techniques, but rather to help diverse stakeholders appreciate the range of approaches and outcomes that are possible when ML underwriting models are used. In particular, the study offers two primary contributions:

- A rigorous case study of various model diagnostic tools in the context of specific consumer protection requirements that in one way or another prompt lenders to address transparency challenges associated with ML models.

- A framework that lenders, regulators, and researchers can use as a starting point for evaluating the quality and usefulness of information about the behavior of ML underwriting models.

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 window for affecting the broad direction of evolution (before developing more calibrated and binding standards as the innovation lifecycle progresses).

This time also represents an opportunity to re-think and improve upon prior generations of automated underwriting to the extent that 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 these 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.