Advancing the Credit Ecosystem:
Machine Learning & Cash Flow Data in Consumer Underwriting

Research Measures Improvements in Model Predictiveness and Credit Access

New empirical research by FinRegLab demonstrates that adopting machine learning techniques and incorporating cash flow data into consumer underwriting can significantly improve model predictiveness and credit access—without increasing lenders' default risk. This first-of-its-kind comparative study evaluates the impact of each innovation separately and together.

Why This Research Matters

Roughly half of U.S. consumers may face challenges in accessing affordable credit, either because they have limited or no traditional credit history or because they have struggled to pay prior debts. At the same time, traditional approaches do not predict default risk equally well for all groups of consumers, which can make lenders both more vulnerable to risk and more conservative than they need to be in approving applicants who are in fact creditworthy.

Machine learning techniques can potentially detect more nuanced patterns in input data than traditional models, while cash flow data derived from checking, prepaid, and other bank accounts can provide insights into consumers' income, reserves, and expenses that do not appear in traditional credit reports. Large banks and fintech lenders are increasingly adopting one or both of these innovations, but smaller lenders' and policymakers' reactions have been mixed to date.

Methodology and Key Findings

FinRegLab built a series of models based on an anonymized dataset combining credit bureau data with electronic bank account information, and then compared the models' predictions against actual credit performance on new accounts. For four data types—credit bureau data only, cash flow data only, credit plus cash flow data, and cash flow data combined with a traditional credit score—FinRegLab built separate models using traditional logistical regression methods and XGBoost, one of the most popular machine learning techniques for credit underwriting. Machine learning is a form of artificial intelligence.

The study found meaningful impacts from both innovations:

- **Machine learning models improved predictive accuracy by up to 2% compared to logistic regression models built with the same data, which increased credit approval rates by as much as 4% in simulations with risk thresholds likely to be used by mainstream lenders.** To provide a sense of scale, in 2023 about 55 million new credit card accounts and 3.8 million new mortgages were originated using similar cutoffs. A 4% increase in borrowers who are deemed to be creditworthy at those thresholds would work out to roughly two million additional credit card accounts and 152,000 additional mortgages.

- **Adding cash flow data to credit bureau data or credit scores also produced accuracy and access benefits**. The impacts were more modest than for adopting machine learning techniques, but data limitations made it difficult to evaluate impacts on consumers who are most likely to benefit because they have little or no traditional credit history.

- **Combining both innovations produced the strongest results.** The machine learning model that combined credit bureau and cash flow data had the highest predictive accuracy and generally had the highest approval rates at various cut offs. The machine learning model that focused solely on credit bureau data frequently ranked second on both metrics.

- **Performance improvements from the two innovations benefitted all subgroups, although impacts on credit access varied somewhat based on credit history and income levels.** In the simulations, increases in credit approval rates were particularly large and consistent across models for consumers with recent derogatory history. The models' reductions in approvals of consumers who went on to default were particularly large and consistent for low-to-moderate income populations.

FinRegLab followed industry practices in various aspects of model development, but the findings are subject to certain limitations due to data and resource constraints as outlined in the report. For example, in addition to limited data on consumers with limited or no credit files, constraints due to the onset of the COVID-19 pandemic prompted the study to focus on a 12 month performance window and affected the dataset used for validation testing. The models may also differ from those produced by individual companies with regard to the nature of the cash flow features and the number of inputs used in the machine learning versions, as there is substantial variation in industry practices on both topics.

Market and Policy Implications

The study makes several contributions in light of current market practice and policy discourse over growing innovation in credit underwriting approaches:

- **The results constitute the first public research that systematically compares varying data sources and analytical techniques, separately and together.** Prior research has concentrated on one innovation or the other and reached conflicting conclusions about machine learning models' impacts on credit access. Despite the sample limitations, constructing a careful set of comparisons across data and model types is helpful to understand their respective impacts.

- **The results underscore the benefits of combining the two innovations but also explore the benefits of staging improvements for lenders who are reluctant to make both changes at the same time.** Some lenders are adopting machine learning models but are focusing only on traditional data sources because they are worried about the compliance challenges involved in managing both adjustments at once. Smaller lenders may be nervous about their ability to manage machine learning models of any type and may prefer to incorporate traditional credit scores because they are validated on large representative populations, but may be open to using cash flow variables to get a more holistic picture of how applicants manage their finances.

- **The results suggest that identifying best practices and facilitating responsible adoption of both machine learning analytics and cash flow data could have substantial benefits for consumers, lenders, and the broader national economy.** Toward that end, the report also includes a detailed appendix describing how the models were developed in light of general industry modeling practices, data management issues, compliance considerations, and other topics. This practically oriented discussion may be helpful for a broad range of audiences who want to understand more about both traditional and machine learning underwriting models and working with both credit bureau and cash flow data.