DOI: 10.59571/mpi.v3si1.3

Guest Essay

Democratising Credit Using Artificial Intelligence And Alternative Data

Anurag Gupta

Problem of practice

Access to credit is essential for economic growth. Banks have traditionally served this role, providing credit to individuals and companies. However, significant proportions in both segments remain underserved, globally – small and medium-sized enterprises (SMEs), individuals with no or short credit histories or those lacking sufficient collateral. These barriers are especially severe in emerging economies with fewer regional and community banks. Artificial Intelligence (AI) models using alternative data have the potential to materially alleviate this problem and revolutionise credit access.

Rise of non-traditional credit

Non-Bank Finance Companies (NBFCs) already own a significant share of the US consumer lending market. For example, the two largest mortgage originators are United Wholesale Mortgage and Rocket Mortgage, both NBFCs. The newest segment in consumer loans is the Buy Now Pay Later (BNPL) segment, which is growing at almost 25% annually, and is also dominated by NBFCs.[1] An estimated 45 million Americans have no credit file or insufficient data to generate traditional credit scores.[2] As some of the largest banks withdraw from consumer loans or scale back their operations in this segment, NBFCs and fintech companies are gaining ground rapidly. This is also true for lending to SMEs who face somewhat similar hurdles.

Much of this change is being driven by the increasing use of alternative data and machine learning models to make credit decisions, also called fintech lending. Alternative data includes online footprints, bill payment histories, phone and location data, employment history, educational background, transaction logs, customer and cash flow data for SMEs. Much of this data is unstructured, for which traditional statistical models are infeasible or unsuitable, thereby necessitating the use of machine learning models, including ones that use deep neural networks.

It has been well established that fintech lending improves credit access for unserved and underserved individuals and SMEs. However, this transformation of the lending landscape also raises several important questions. Can fintech lending complement or replace traditional bank lending? Is this enhanced credit access leading to a larger incidence of distressed loans? Is the cost of credit increasing inordinately? Do alternative data and machine learning models add value to credit risk evaluation? Does the use of alternative data lead to increased biases in lending? Can these new models help predict distress before it occurs, thus enabling timely corrective action? Can borrowers receive better guidance on how to improve their credit profiles? While these questions have received significant attention in the research and media in recent years, we are still at the early stages of arriving at definitive answers.

New methods and new data

Traditional credit scoring models perform well for individuals with high credit scores, but do not predict future creditworthiness for those with low credit scores

with much accuracy. Data from Upstart network (an AI lending platform that partners with banks and credit unions to provide loans to consumers) shows no relationship between the probability of default and credit scores below 700.[3] They also show that traditional lenders would have rejected a vast majority of borrowers funded by Upstart. For example, borrowers with a credit score below 640 who got loans from Upstart had a 70% chance of being rejected by traditional lenders. Traditional models were also found to generate a higher interest rate for such borrowers (lower credit scores) compared to the rates predicted by machine learning models using alternative data.

In a study, Cornelli and team used data from Funding Circle and LendingClub (fintech platforms that connect businesses seeking loans with investors looking to lend) to examine similar questions for SMEs.[4] They documented enhanced credit access for borrowers who are less likely to receive credit from traditional banks, especially those in regions with higher unemployment rates and bankruptcy filings. Hence, these fintech lending platforms were able to predict future loan performance more accurately. Equally importantly, they documented that these platforms provide credit to additional SMEs at a lower cost than traditional credit models predicted.

In light of these and other similar results, an important question that arises is whether the improvement in credit risk assessment occurs due to the use of alternative data or the use of certain machine learning models, or both. Traditional credit scoring models often use variants of logistic regression[a] to predict the probability of default. They are linear in the logarithm of the default probability, which compromises predictive accuracy. In recent years, several machine learning models have become popular for credit risk assessment. One of them is an ensemble of classification decision trees called the random forest classifier (in the category of supervised learning models). Each individual or company and its corresponding data are mapped to a default response variable. The random forest classifier merges the rules obtained from a set of such decision classification trees. With larger data sets, clustering models are often used (these are unsupervised learning models) to first group individuals or SMEs into relatively homogenous clusters. Individual risk models are then developed for each cluster using supervised learning (such as the random forest classifier). Deep learning models such as Recurrent Neural Networks (RNNs) can be used with sufficient sequential data. RNNs introduce a time dimension in the neural network, allowing for a more extended prediction memory since variable values

from prior time periods continue to affect a borrower's predicted probability of default directly. This also allows the model to monitor performance in real time and learn from changes in variables over time, which can help provide timely updates, including when there are changes in macroeconomic conditions.

Traditional models cannot work with a large amount of unstructured data. Machine learning models using only traditional data do not perform as well as those augmented with alternative data. Therefore, both are important, and both contribute to enhanced performance in predicting the default risk of borrowers and future loan performance. In addition, machine learning models can also be used to predict drawdowns and credit usage, which is not possible with traditional models. However, the use of machine learning models along with the use of alternative data does come with potential downsides and risks.

First and foremost, machine learning models such as random forest classifiers and RNNs lack interpretability. They are black box models that produce good predictions, but one cannot explain which variables contribute the most to those predictions. They do not allow the modeller to understand, let alone explain, why a particular credit decision was made. This also restricts the ability of creditors to guide borrowers regarding how to improve their credit standing in the future. Poorly understood models can not only fail but also cause harm. These models must be made explainable by either training an equivalent white box model[b] that produces similar outcomes or using a model-agnostic technique[c], such as Shapley values[d], that captures the marginal contribution of each variable to the final prediction. They are calculated by perturbing[e] each variable, one at a time and then rerunning the model for each variable perturbation combination to assess the impact on the prediction. Thus, a stepwise variable selection algorithm can be created using global Shapley values. This technique can be used to select the most appropriate model and explain the contribution of each variable to the final prediction. This must be done every single time a black box model is used.

Furthermore, machine learning models can find patterns in data that traditional models cannot. This could sometimes disadvantage borrowers from certain classes or groups, leading to biases in credit decisions that may inadvertently correlate with race or social class. The evidence on bias in lending decisions by fintechs is sparse and mixed so far. While studies by Erel and Liebersohn and Howell & team document reductions in racial disparity in credit decisions, the study by Chernenko et al.[5,6,7] observe similar racial disparities in approval rates by fintechs compared to those by banks.,, An essential exercise is to ensure that differences in credit decisions are based on credit-related data and not on proxies for race or class. An adverse impact ratio (AIR) can be created to measure the ratio of the rate at which

credit decisions were approved for a particular race or class relative to the rate of approval for others. While a ratio of 100% is almost always unattainable, the goal should be to attain a ratio as close to 100% as possible. In conversations with market participants, a minimum AIR of 80% is often stipulated as a hurdle for the model to clear. It should be noted here that any ratio below 100% will likely degrade the model performance to some extent, thus creating a trade-off between a best model that might be severely biased and a good model that would perform well enough in practice but would not

Fintech lenders have a higher capital cost relative to banks. However, that does not appear to translate to a higher cost of credit to similar-risk borrowers. They use more comprehensive data on borrowers, so they are able to get a more refined estimate of their probability of default with lower origination costs and faster speed of decisions.

incorporate any significant biases. How exactly this trade-off should be made is a matter of practical implementation, but one that must be explicitly analysed and discussed before a model is deployed in the real world. Alternatively, adversarial debiasing techniques can be used where the predicted default probabilities from a machine learning model are then used as an input in an adversarial model that attempts to predict race or class. Different combinations of alternative data variables are tested until the adversarial model cannot make such a prediction. Regardless of the method used, debiasing techniques must be used whenever alternative data is employed for credit decisions.

The share of traditional banks in lending to underserved individuals and SMEs has progressively shrunk over the years. While they use some alternative data in their credit models, they heavily rely on internal data for their own customers. They also rely to some extent on soft information gathered through interactions with existing customers. Fintech lending breaks these barriers by replacing soft information with alternative data and opening up credit access to those who would not otherwise receive credit. They can also offer credit products with novel features such as variable payment plans, different maturities and indexed APRs (Annual Percentage Rate)f. These effects have been documented in other countries as well, such as China, using data from Alibaba. There is also clear evidence of complementarity, as shown by Dolson and Jagtiani, who find that non-prime borrowers are more likely to be served by fintech lenders.[8]

Fintech rise and regulation

Fintech lenders have a higher capital cost relative to banks. However, that does not appear to translate to a higher cost of credit to similar-risk borrowers. They use more comprehensive data on borrowers, so they are able to get a more refined estimate of their probability of default with lower origination costs and faster speed of decisions. They are also relatively lightly regulated, which does raise concerns about data privacy and security, especially since a vast array of personal alternative data is often used in their models. Individuals may also not be fully aware of the data and how it is used in their credit decisions. To allay these concerns, lenders and AI model developers must ensure that data is anonymised and that individuals have the right to access, correct, or delete their data. Implementing robust cybersecurity measures is critical to protect sensitive information. Regulatory frameworks such as the General Data Protection Regulation (GDPR) in the European Union can help establish clear guidelines for how data can be collected, used, and shared.

India's consent credit stack

As regulatory frameworks evolve globally to keep pace with innovation, India is a distinctive example where policy, digital infrastructure, and fintech converge to reshape credit access at scale. India's push to democratize credit through AI and alternative data has achieved notable gains despite several challenges. Fintech has driven digital lending to an estimated market size of US$350 billion by the end of 2023, with growth at about 35% annually in 2024.[9,10], On the one hand, the Account Aggregator (AA) framework has expanded rapidly – over 112 million users have linked accounts on this consent-based data-sharing network – and the Open Credit Enablement Network (OCEN) is beginning to bridge the vast – Rs 2025 trillion credit gap for MSMEs (with less than 11% of these enterprises accessing formal credit) by facilitating small-ticket, remote lending via alternate data.[11,12], AI-driven credit solutions can help improve efficiency by personalising services, providing credit access for those who need it most, lowering operating expenses, and streamlining loan disbursals.

On the other hand, these advances come with several key concerns. First, rapidly rising unsecured lending could lead to increased household delinquency and higher personal loan non-performing assests. Second, the rise of this shadow banking system, without adequate regulation and oversight, can lead to higher rates of distress amongst fintech lenders. And last, but not least, concerns about data privacy are rising, and opaque AI credit models lack transparency, making it harder for customers or regulators to understand or contest decisions. As this sector grows and matures, regulators and policymakers need to address these concerns effectively and expeditiously.

The application of machine learning models using alternative data faces further challenges that are unique to India. With a large informal economy and workforce, obtaining reliable data for many unbanked households is harder. Large segments of the population still remain on the sidelines: over half of India's informal workforce has no meaningful access to formal credit, highlighting a

India's push to democratize credit through AI and alternative data has achieved notable gains despite several challenges. Fintech has driven digital lending to an estimated market size of US$350 billion by the end of 2023, with growth at about 35% annually in 2024.

digital divide in rural and marginalised communities.[13] In addition, extreme diversity with 22 official languages makes the training and validation of these models more challenging. Meanwhile, smaller fintech lenders are strained by rising compliance burdens. Their cost of compliance has nearly doubled (with hefty investments in technology, data protection, audits, etc.) as they strive to keep up with frequent updates in the RBI rules in an evolving regulatory climate.[14] India must continue fostering innovation in credit access while strengthening oversight and accountability, ensuring that its AI-powered credit revolution remains inclusive and responsible.

Strengthening the foundation

As alternative data increasingly drives credit decisions to meet rising demand, an important caveat is ensuring that these machine learning models using alternative data are rigorously validated and tested before deployment. Their use in credit decisions is relatively recent, mainly in the last decade. Almost all studies in this area have examined outcomes over a relatively short period, ranging from a few months to a maximum of a few years. How well these models continue to perform over longer periods of time across different macroeconomic conditions is yet to be tested. Given the large parametric size of these models, especially deep neural network models, it is always possible to fit the available dataset very well, sometimes exactly. Proper validation and testing in future data samples will be necessary to establish best practices in this space.

Traditional credit scoring and soft information have been the drivers of relationship lending in banks for a long period of time. Fintech lenders are transforming and complementing this industry by using vast amounts of unstructured alternative data in complex machine learning models to obtain better predictions of the probability of default for borrowers unserved or underserved by traditional bank lending. However, there must be adequate oversight to ensure that credit decisions are made explainable, models are debiased, and customer data is obtained and used responsibly and transparently.

Glossary of technical terms used

[a] **Logistic regression** is a statistical technique used to calculate the probability of an observation belonging to a specific category. For example, it is used to estimates a borrower's probability of defaulting on a loan.

[b] **White box model** is machine-learning model that make predictions that are interpretable and transparent, showing how the model produced an output thus, revealing the algorithm's inner logic.

[c] **Model-agnostic technique** in machine learning, explain or evaluate predictions made by a model by tweaking its inputs and observing the outputs, without relying on the underlying code. They provide clarity and explanation on the working of the models across the machine learning spectrum, ensuring analysts do not end up with unexplained decisions.

[d] **Shapley value** in game theory is a solution concept of fairly distributing both gains and costs to several actors working in a coalition.

[e] **Perturbing** in machine learning means making small changes in the parameters or inputs of the model to understand how sensitive the model is to changes in its inputs. These changes are usually done to one or two variables at a time.

[f] **Indexed APRs (Annual Percentage Rate)** is a loan or credit card interest rate that rises or falls with a public market rate plus a fixed "margin" set by the lender.

Anurag Gupta** is Professor in the Department of Banking & Finance, Weatherhead School of Management, Case Western Reserve University. You can reach out to him at **anurag.gupta@case.edu.

This article may contain links to third-party content, which we do not warrant, endorse, or assume liability for.
The authors' views are personal.

*We welcome your thoughts – drop us a note at **mpi@spjimr.org.***

REFERENCES

[1] "Buy Now Pay Later 2024 – What Comes Next?," *GlobalData* (blog), July 5, 2024, https://www.globaldata.com/store/report/buy-now-pay-later-market-analysis/.

[2] "CFPB Explores Impact of Alternative Data on Credit Access for Consumers Who Are Credit Invisible," Consumer Financial Protection Bureau, 2017, https://www.consumerfinance.gov/about-us/newsroom/cfpb-explores-impact-alternative-data-credit-access-consumers-who-are-credit-invisible/.

[3] Marco Di Maggio, Dimuthu Ratnadiwakara, and Don Carmichael, "Invisible Primes: Fintech Lending with Alternative Data" (National Bureau of Economic Research, 2022), https://www.nber.org/papers/w29840.

[4] Giulio Cornelli et al., "The Impact of Fintech Lending on Credit Access for U.S. Small Businesses," *Journal of Financial Stability* 73 (August 1, 2024): 101290, https://doi.org/10.1016/j.jfs.2024.101290.

[5] Isil Erel and Jack Liebersohn, "Can FinTech Reduce Disparities in Access to Finance? Evidence from the Paycheck Protection Program," *Journal of Financial Economics* 146, no. 1 (October 1, 2022): 90–118, https://doi.org/10.1016/j.jfineco.2022.05.004.

[6] Sabrina T. Howell et al., "Lender Automation and Racial Disparities in Credit Access," *The Journal of Finance* 79, no. 2 (December 11, 2023): 1457–1512, https://doi.org/10.1111/jofi.13303.

[7] Sergey Chernenko et al., "Applications or Approvals: What Drives Racial Disparities in the Paycheck Protection Program?," Working Paper, Working Paper Series (National Bureau of Economic Research, April 2023), https://doi.org/10.3386/w31172.

[8] Erik Dolson and Julapa Jagtiani, "Which Lenders Are More Likely to Reach Out to Underserved Consumers: Banks Versus Fintechs Versus Other Nonbanks?," *Journal of Alternative Finance* 1, no. 1 (March 1, 2024): 35–66, https://doi.org/10.1177/27533743231218019.

[9] Decentro Fintech Infrastructure, "Digital Lending Industry: Market Size and Statistics," n.d., https://decentro.tech/help/faqs/digital-lending-market-size-in-india?

[10] DD News, "India Stands Out as Global Fintech Bright Spot as Credit Demand Surges,", https://ddnews.gov.in/en/india-stands-out-as-global-fintech-bright-spot-as-credit-demand-surges/.

[11] Department of Financial Services, Ministry of Finance, Government of India, "Account Aggregator Framework," 2024, https://financialservices.gov.in/beta/en/account-aggregator-framework.

[12] OCEN, "Open Credit Enablement Network: Democratising Credit," n.d., https://ocen.dev/#:~:text=Credit%20gap%20for%20MSMEs%20is,25%20trillion.

[13] Anoop Verma, "Next Gen AI & India's Financial Leap: Building Inclusive Credit Systems for the Unbanked," *ETGovernment*, April 21, 2025, sec. news/digital-india, https://government.economictimes.indiatimes.com/news/digital-india/next-gen-ai-indias-financial-leap-building-inclusive-credit-systems-for-the-unbanked/120456758.

[14] Ajinkya Kawale, "Compliance Costs May Spur Fintech Sector Consolidation: Industry Experts," *Business Standard*, December 24, 2023, https://www.business-standard.com/economy/news/compliance-may-lead-to-consolidation-of-fintech-sector-industry-players-123122400506_1.html.

Article Information:
Date article submitted: May 27, 2025
Date article accepted: May 29, 2025
Date article published: Jun 30, 2025

Images courtesy : www.freepik.com