Financial Inclusion and Alternate Credit Scoring for the Millennials: Role of Big Data and Machine Learning in Fintech

Sumit Agarwal*Shashwat Alok†Pulak Ghosh‡Sudip Gupta§

March 30, 2020

A recent survey in the US showed that almost half of the millennials in the US feel that their credit score is holding them back[1].Younger people suffer from shorter credit history and hence are often denied credit by traditional financial institutions or are charged prohibitively high interest rates, which limits their access to credit[2]. This, in turn, exacerbates the evaluation of their creditworthiness by limiting their ability to build a good credit history. Many such individuals may actually be 'good borrowers' if their 'creditworthiness' could be evaluated using alternate data. The problem of lack of credit history for the millennials is a world-wide phenomenon and especially true for developing countries. For example, according to a recent industry report, 156 million Indians who comprise the 'urban mass' representing an annual income of USD 3000 and above have the potential of mass adoption of consumer credit. Of this 'urban mass', approximately 129 million have been mostly deprived of credit due to a lack of credit history.

*usakari@yahoo.com, National University of Singapore

†shashwat.alok@isb.edu, Indian School of Business

‡pulak.ghosh@iimb.ac.in, Indian Institute of Management, Bangalore

§sgupta24@fordham.edu, Gabelli School of Business, Fordham University

[1]Wall Street Journal Blog [Accessed on 17th October, 2019]. According to Wall Street Journal and Transunion; Around 53 million consumers are not scoreable due to lack of information at the three major credit bureaus, and this population is heavily skewed towards those under 35.

[2]2MarketWatch News Article [Accessed on 14th March, 2019].The survey looked into the credit experience of 2,000 Americans ages 18 to 34, and found that many young adults are suffering the consequences of bad credit. In fact, 24 percent of those surveyed said they never learned how to build good credit in the first place, and 15 percent reported that their level of debt is unmanageable, with 1 in 5 admitting that they don't have control over their finances.

This led to the quest for alternative data for credit scoring for the millennials. While millions across India and the world have never obtained a bank loan, they are active mobile phone users who shop online, and have a good social media presence. These traces of unstructured data that individuals leave through their online behavior and mobile phone usage can potentially be used to predict their loan behavior. Consistent with this idea, a plethora of fintech firms have mushroomed all around the world that aim to service such customers by leveraging unstructured data and big data analytics to predict their default behavior. However, thus far, there is limited evidence on whether or not "mobile footprint" of an individual can substitute for traditional credit bureau scores. In this paper, we examine whether an individual's online behavior captured from their mobile phones can be used to predict their likelihood of default.

We use data from one of the largest Fintech lending firms in India to examine the discriminatory ability of mobile footprint variables in predicting loan outcomes. Specifically, we want to understand whether and how the mobile footprint is associated with loan level outcomes such as the likelihood of loan approval and the likelihood of default. More importantly, we want to understand whether these variables can be used to predict the likelihood of default for a borrower without any credit history and, consequently, a credit bureau score.

A natural follow-up question is whether we can use the social and mobile footprint variables to come up with an alternate credit scores for borrowers who do not have traditional credit bureau scores. How many of the borrowers who are denied loans could potentially be creditworthy if their creditworthiness could be evaluated using information from their social and mobile footprints? Importantly, how would granting loans to such borrowers affect the overall default rate of the lender's loan portfolio? These counterfactual questions have significant policy implications. Importantly these questions pertain to default prediction and are not causal in nature. We use a novel machine learning procedure in addressing the policy counterfactual questions posed above.

We obtain the universe of loan applications made to one of the largest fintech lender in India, between the period of February 2016 to November 2018. Unlike prior studies, we also have access to loan applications that were denied allowing us to examine the determinants of loan approval.

Out of about 417,000 loan applications in our sample, about 272,000 were approved while rest were denied. The lender is a stereotypical mobile-only fintech lending platform

targeted towards meeting the short-term credit needs of the salaried millennial. It grants loans ranging from a minimum of INR 10,000 to a maximum of INR 200,000 for 15, 30, 90, 120, and maximum loan duration of 180 days.

To apply for a loan, an individual need to log on to the mobile application and submit regulation mandated identification and address documents, along with bank statements, and salary slip. The potential borrower authorizes the lender to use its digital mobile presence for the evaluation of her creditworthiness and research. They also provide the fintech lender data on their traditional credit score: CIBIL-Transunion credit score (if available), education, and job designation. Importantly for our study, the lender also collects detailed digital information from the individuals' mobile phone such as the mode of login (for example, Facebook and Linkedin), the various applications installed, number of calls, number of contacts on phone, number of social connections, and the kind of mobile operating system such as IOS and Android. We have access to detailed anonymized data on the kind of mobile applications that an individual uses that we club into 6 broad categories: Sales apps which includes applications for e-commerce such as Amazon, Flipkart, Snapdeal among others, Social Network apps such as Whatsapp, Twitter, Messenger services, Financial Apps such as Mobile banking and stock trading applications, Travel apps such as Airbnb, Tripadvisor, and MakeMyTrip, Mloan app which includes other mobile-based lending platforms, and Dating apps such as Tinder.

In addition, we have detailed information on call logs of individuals. For ease of reference, we categorize this digital information captured from an individual's mobile phone into three categories: (1) "social footprint" which refers to the presence of social apps, the preferred social network for logging on to the fintech lender's app, number of contacts, number of calls/sms, whether the customer was acquired through a referral (2) "deep social footprint" which captures information obtained from call logs pattern, and (3) broader "mobile footprint" which refers to the kind of applications installed, the number of applications, and the type of mobile operating system.

This kind of deep digital information on the number of social connections or kind of applications that a customer uses can potentially proxy for otherwise hard to quantify and unobservable aspects of individual behavior that is unavailable to traditional banks.

We begin by analyzing whether and how the customer characteristics, mobile footprint, and social footprint relates to loan approval decisions. As one would expect, we find that a

loan applicant with a higher credit score, salary, and education is more likely to get approved. Importantly, we find that larger is the mobile and social footprint of an individual, the higher is her likelihood of loan approval. Specifically, we find that the number of contacts, the number of apps installed, the number of calls made or received, and the presence of financial and mobile loan apps are positively associated with the loan approval. The discriminatory ability of various aspects of mobile footprints is robust to controlling for the credit bureau scores, customer's earnings, age, education, and location. This suggests that mobile footprint variables provide incremental information that is important for predicting loan outcomes beyond what is captured in the credit score.

Next, we examine the ability of mobile and social footprint variables in predicting defaults. Here, we rely on both the economic and statistical significance of individual explanatory variables as well as Area Under the Curve (AUC) - an easy and commonly used measure of the predictive power of credit scores. The area under the curve is used as a measure of the goodness of a prediction. It measures the proportion of true positives in a prediction. Higher the AUC, the higher is the prediction accuracy We first note that the AUC of the model using only the credit score for predicting defaults is 59%.

This suggests that the discriminatory ability of the credit score in predicting defaults is likely to vary across geographies and intermediaries. To the extent that mobile footprint variables complement the information content of credit score, the marginal value of such information is higher in contexts where the credit score itself has lower discriminatory power. Thus, fintech firms that rely on the mobile footprint for screening borrowers maybe even more important to expand credit access in countries with weak information environments and lower levels of financial inclusion.

The AUC of a model that relies exclusively on the mobile and social footprint to predict defaults at 60.4% is approximately 2% more than the AUC of the model using only the credit score. Our results suggest that mobile and social footprint variables may be capturing hard to quantify aspects of individuals' behavior, which has implications for the likelihood of default. For instance, customers without a financial application installed on their phones are about one and a half times more likely to default relative to those who have such an application installed. This is consistent with the idea that installing financial applications may proxy for the financial sophistication of a customer. In contrast, those with a dating application (any other social network app) are 30% (38%) more likely to default. Interestingly, customers

who log in to the application via Linked or Facebook are 24% and 9% more likely to default respectively relative to those who log-in via other means.

These results hold after controlling for customer's salary, age, and education. This is important because if mobile footprint only proxies for easily measurable financial or customer characteristics, then fintech lending firms should directly collect data on those characteristics rather than trying to infer it from the mobile footprint variables. Indeed such digital information holds more promise if it captures some soft or hard information that would be otherwise difficult to measure or verify. In such a case, mobile and social footprints can be used to improve traditional credit scoring models.

Our results suggest that mobile and social footprint captures an unobservable aspect of individuals which is not fully absorbed by earnings, education, or credit score. Importantly, the AUC of this specification is 61% , two percentage points higher than the AUC of the model using only the credit bureau score and seven percentage points higher than the model, which includes only customer characteristics. In other words, a predictive model that includes customer characteristics, social and mobile footprint performs better in predicting defaults as a model, which includes credit bureau score, and customer characteristics. Overall, these findings suggest that mobile and social footprint variables complement the credit bureau score and observable customer characteristics.

Further, we can use digital information to build credit scoring models for and make loans to individuals without credit or financial history, thereby expanding credit access. To strengthen the evidence in favor of this thesis, we examine the predictive ability of mobile and social footprint in predicting defaults for the set of customers without a credit score or history. The AUC of the mobile footprint model for this sample is 58% and comparable to the predictive performance of the credit bureau score in the primary sample for customers with a credit bureau score.

Our analysis of default prediction thus far was based on measures of mobile and social footprint such as the nature of apps installed, the number of apps installed, the number of calls, etc., to predict defaults. We now seek to understand whether we can use "deep social footprint" of customers from their call logs to improve upon the default prediction. Using various proxies based on the frequency and duration of daily incoming, outgoing, and missed calls that attempt to capture the breadth and strength of an individual's social capital, we find that these measures are strongly correlated with the likelihood of default. Specifically,

we find that defaulters are more likely to have their call concentrated over a smaller number of individuals. Consistent with this, defaulters seem to have stronger ties with individuals in their contact list as measured by the average number of calls and duration of calls per person. Delinquent customers have a smaller duration of incoming calls but have a higher duration of outgoing calls, which along with their frequency of missed calls, suggests that defaulters are less likely to respond to calls initiated by others.

Most importantly, the AUC of a model that includes call log measures along with other mobile and social footprint variables is 66%, an 8% improvement over the model with credit score alone.

We also have access to the detailed financial reports for a random subset of the borrowers in our sample. The report provides detailed financial information like the borrower's spending and income patterns, number of transactions, other borrowing information, etc., over the last three months, which we collectively refer to as 'deep financial information'.The fintech lender accessed these reports during the loan application process. We find that for the subset of the borrowers for whom we have access to this financial report, the 'deep' mobile footprint has greater predictive power for borrower's credit risk relative to the 'deep' financial information.

We next verify the predictive performance of social and mobile footprint variables using different machine learning algorithms. The problem at hand is to train the algorithms on the sample data to predict defaults "out-of-sample". Standard estimation approaches, where we use all the data to make in sample prediction, is not well suited for such analysis. The in-sample estimation approaches works on being unbiased (having the bias close to zero), thus leaving only the variance to be optimized to minimize the out of sample prediction error. Thus, standard estimation approaches does not offer joint optimality of bias and variance. Machine learning techniques are particularly useful here, which minimizes the mean squares error of the prediction by a joint minimization procedure cognizant of the bias-variance trade-off. Using various machine learning algorithms, we first show that the mobile and social footprints have significantly higher predictive power for both borrowers with and without traditional credit scores.

Next, we run a horse race between 'deep' financial information and 'deep' social footprint variables based on call logs to see if the deep mobile footprint has incremental predictive power beyond what is captured in the borrower's income and spending patterns. This is important as it can inform us regarding the nature of data that should be collected to build

alternate credit scores. First, we find that both 'deep' financial information and 'deep' mobile footprint variables have significant discriminatory ability in predicting defaults. Second, the information content of deep mobile footprint complements and exceeds the 'deep' financial variables. Specifically, the out of sample AUC of the models which includes only deep mobile footprint variable (deep financial information) is 74% (59%). Overall, we find that digital mobile footprint has significant ability in predicting defaults and the information content of these variables complements rather than substitutes for both the credit bureau score and detailed financial information regarding a customer's income and expenses.

The prediction of default risk for borrowers without a traditional credit score is useful and can be used to ask counterfactual questions such as: how many denied borrowers (perhaps due to lack of traditional credit score) would have been approved had we used the social and mobile footprint based alternate credit scores? What would have been the impact on default if we had used these scores? These counterfactual prediction policy questions are not causal in nature, as our objective is to find the best predictor of default risk of the borrower. Using our methodology, we find that even if we use a low predicted default threshold of 10% (1%) for the probability of default relative to the in-sample default rate of 12% for approving loans, about 42% (22%) borrowers who were denied credit would have been granted loans.

Overall, our study documents that mobile and social footprint variables have significant discriminatory power in both loan approvals and default prediction. Importantly, with the use of big data, fintech lenders can potentially build credit scores and can expand access to credit to even customers with little or no credit history that are underserved by the traditional banks.

Consistent with this conjecture, the average individual in our sample is a sub-prime borrower with a credit score of 641.11 Moreover, an economically significant 20% of borrowers in our sample do not have a credit score. This is in contrast to the USA, where fintech lenders primarily cater to borrowers who already have access to credit via traditional banks. However, the use of machine learning algorithms combined with big data for credit allocation decisions is not without costs.

Conclusion

In this paper, we have used a unique and proprietary dataset to analyze the impact of the mobile footprint of individual borrowers in predicting loan outcomes. Our dataset comes from a leading fintech lending company in India. We find that the mobile and social footprint

has significantly more predictive power than traditional credit score used by banks.

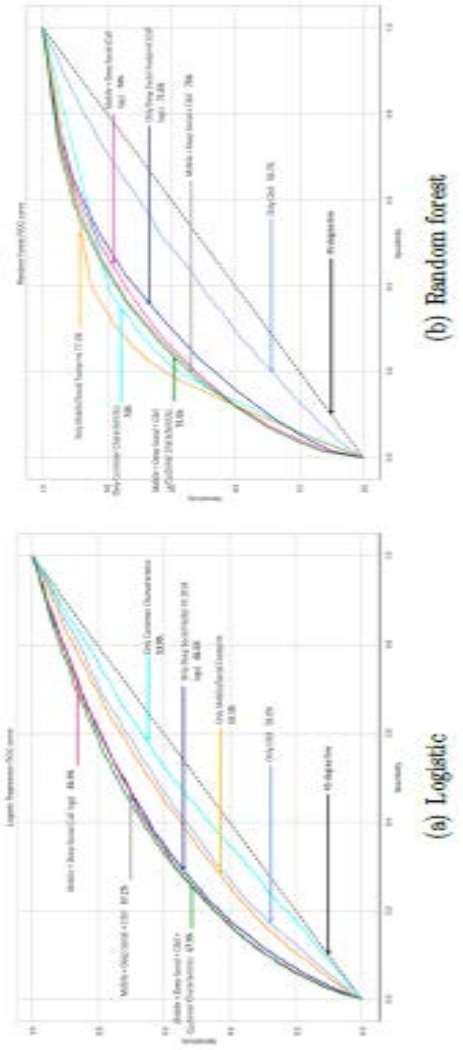
We find a number of interesting results. First, we document a statistically and economically significant role of individuals' mobile and social footprint variables in the loan approval process. In the absence of sufficient credit history and credit scores for millennial customers to judge their creditworthiness, the fintech lender uses individuals' mobile footprint as an alternative credit screening process. This is consistent with the wide use of social media-based credit scoring recently adopted by fintech companies worldwide.

We also find that a simple predictive model in which an individual's both crude mobile/social footprint and deeper social footprint based on call logs significantly outperforms a model with a credit score in predicting defaults.

We verify these results using machine learning algorithms that are especially suited for prediction and find qualitatively similar results. Importantly, our counterfactual exercise indicates that evaluating creditworthiness based on social and mobile footprints can potentially expand credit access to the financially excluded borrowers without adversely affecting loan performance.

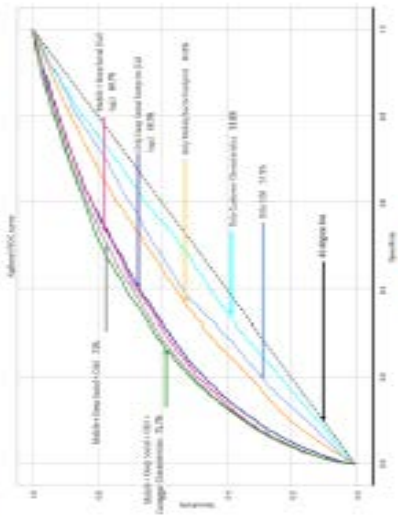
Overall, our paper underscores the importance of individuals' mobile footprint, and social footprint in predicting consumer loan approval and default prediction. These have wider policy implications as we design new modes of financial intermediation, services, and regulations in the era of 'big data.'

Figure: AUC Plots of Different Machine Learning Models



(a) Logistic

(b) Random forest



(c) XGboost

Figure: Variable Importance Factor