Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech*

Sumit Agarwal †, Shashwat Alok‡, Pulak Ghosh§, and Sudip Gupta¶

This Version: April 2023

Abstract

We use unique and proprietary data from a large Fintech lender in India combined with machine learning models to document that alternative data captured from an individual's mobile phone, such as the number and types of apps installed, measures of social connections, and borrowers' "deep social footprints" based on call logs, can substitute for traditional credit bureau scores in credit risk evaluation and improve financial inclusion. Using machine learning-based prediction counterfactual analysis, we find that alternate credit scoring based on an individual's digital presence can expand credit access to financially excluded individuals who lack credit scores without adversely impacting default outcomes. Our findings imply that alternative digital data sources have the potential to significantly improve credit risk assessment and financial inclusion in developing countries.

JEL codes: G20, G21, G29

Keywords: Fintech, Big data, Credit scores, Financial inclusion, Lending, Machine learning, Mobile footprint, Prediction Counterfactual, Social footprint, Social capital, Digital economy, Open banking, Data sharing

*We thank Tobias Berg, Souphala Chomsisengphet, Francesco D'Acunto, Andrew Hertzberg, Kose John, Venkatesh Panchapagesan, Manju Puri, Tarun Ramadorai, Raluca Roman, Amit Seru, Wenlan Qian, Michael Weber, and the participants at the NSE-NYU conference (2019), ABFER conference (2019), AFA (2020), Fifth Annual Conference on Alternative Finance, Cambridge (2020), Northern Arc Foundation - Dvara Trust field workshop on Household Finance (2020), New York Fed Fintech conference (2020), and FMA (2020) for their helpful comments. We thank Piyush Gupta for providing excellent research assistance. We acknowledge generous funding from NSE- NYU Initiative on the Study of Indian Financial Markets, NSE ISB Trading Laboratory, and SRITNE.

†Email: ushakri@yahoo.com. National University of Singapore.

‡Email: shashwat_alok@isb.edu. Indian School of Business.

§Email: pulak.ghosh@iimb.ac.in. Indian Institute of Management, Bangalore.

¶Email: sgupta24@fordham.edu. Fordham University.

I Introduction

A recent survey in the US showed that almost half of the millennials in the US feel that their credit score is precluding them from achieving their financial objectives.[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 be 'good borrowers' if their 'creditworthiness' could be evaluated using alternate data. The problem of lack of credit history is a worldwide phenomenon and is especially true for developing countries. For example, as of 2022, 160 million Indians, primarily comprising the younger population ("Millennials"), have been mostly deprived of credit due to a lack of sufficient credit history (TransUnion (2022)). This underscores the importance of using alternative data for credit scoring.

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 social media presence.[3] 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, many fintech firms have mushroomed worldwide that aim to service such customers by leveraging unstructured data and big data analytics to predict their default behavior. Thus, an important area of research is whether one can tease out better quality (low default risk) borrowers from individuals who do not have the traditional credit score based on social and mobile footprints (Das (2019)). However, thus far, there is limited evidence on whether or not an individual's "mobile footprint" can substitute for traditional credit bureau scores and improve access to credit. Our study adds to the recent but growing body of work examining the implications of increasing usage of 'fintech,' big data, and machine learning algorithms on consumer welfare.[4]

We use unique data from one of the largest Fintech lending firms in India to examine

[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 a lack of information at the three major credit bureaus, and this population is heavily skewed toward those under 35.

[2] MarketWatch 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 suffer from the consequences of poor credit history. Twenty-four 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.

[3] 97% of users access the internet in India through mobile phones. See Kantar-IMRB (2018)

[4] See Chava, Paradkar & Zhang (2017), Fuster, Goldsmith-Pinkham, Ramadorai & Walther (2018), Berg, Burg, Gombović & Puri (2020), D'Acunto, Prabhala & Rossi (2019), Rossi & Utkus (2019), D'Acunto, Rauter, Scheuch & Weber (2019), Balyuk (2019), Das (2019).

the discriminatory ability of mobile footprint variables in predicting loan outcomes. 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. Our goal is not to pin down the causal channels through which a customer's mobile footprint may affect her creditworthiness but rather to analyze the association between the mobile footprint and credit worthiness of individuals.

A natural follow-up question is whether we can use the social and mobile footprint variables to come up with alternate credit scores for borrowers who do not have traditional credit bureau scores. What fraction of the borrowers without a credit score, who were rejected, would potentially receive credit if their creditworthiness could be evaluated using alternative data? Importantly, can such alternate credit scoring methods be used to expand credit access without adversely impacting the overall default rate of the lender's loan portfolio? These counterfactual questions have significant policy implications. Notably, these questions pertain to default prediction and are not causal in nature. The focus on prediction policy counterfactual is new in economics (Kleinberg, Ludwig, Mullainathan & Obermeyer (2015), Athey (2017))). We follow Kleinberg et al. (2015) and use machine learning algorithms in addressing the counterfactual policy questions posed above.[5]

We obtained the universe of loan applications made to one of the largest fintech lenders in India between February 2016 and November 2018. Unlike prior studies, we also have access to loan applications that were rejected allowing us to examine the counterfactual approval rates using alternate credit scoring models. Out of approximately 363,000 loan applications in our sample, about 265,000 were approved, while the rest were rejected. The lender is a stereotypical mobile-only fintech lending platform targeted towards meeting the salaried millennial's short-term credit needs. It grants loans ranging from a minimum of ₹10,000 to a maximum of ₹200,000. The loans are available for a duration of 15, 30, 90, 120, and 180 days.

To apply for a loan, an individual needs to log on to the mobile application and submit regulation-mandated identification and address documents, along with bank statements and salary slips. 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), income, 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 the phone, number of social connections, and the kind of mobile operating

[5]See also Athey & Imbens (2019)

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 six 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.[6]

In addition, we have detailed information on call logs obtained from their mobile devices for a subsample of the customers in our data. For ease of reference, we categorize these alternative data captured from an individual's mobile phone into three categories: 1) broader "mobile footprint," which refers to the kind of applications installed, the number of applications, and the type of mobile operating system, 2) "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, etc., and 3) "deep social footprint" which captures information obtained from call logs pattern.[7]

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. To the best of our knowledge, our paper is the first to examine whether such deep aspects of an individual's digital presence captured from their mobile phones can be used to predict loan defaults, particularly for individuals without a traditional credit bureau score.

We begin by analyzing the ability of mobile and social footprint variables in predicting defaults using standard Logistic regressions. 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 (Iyer, Khwaja, Luttmer & Shue (2015)).[8] We first note that the AUC of the model using only the credit score for predicting defaults is 55% and significantly different from flipping a coin (AUC of 50%).

The AUC of a model that relies exclusively on the mobile and social footprint to predict defaults at 63% is approximately 8% more than the model with a credit bureau score. Our results suggest that mobile and social footprint capture hard-to-quantify aspects of individuals' behavior, which has implications for the likelihood of default. For instance,

[6]We use apps and applications interchangeably throughout the paper.

[7]Throughout the paper we often use the mobile footprint to collectively refer to both social footprint and broader mobile footprint variables.

[8]Section A.III in Appendix A provides a detailed explanation of AUC and other prediction performance measures.

customers without a financial application installed on their phones are about 25% more likely to default than those with such an application installed. This is consistent with the idea that installing financial applications may proxy for a customer's financial sophistication. In contrast, those with a mobile loan app are 26% more likely to default. Interestingly, customers who log in to the application via Linkedin are 32% less 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, fintech lending firms should directly collect data on those characteristics rather than 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 thus far suggest that mobile and social footprint captures an unobservable aspect of individuals not fully absorbed by earnings, education, or credit score. Importantly, the AUC of the alternate data model is nine percentage points higher than the traditional credit screening model, which includes credit score and customer characteristics.

Next, we verify the predictive performance of social and mobile footprint variables using two machine learning algorithms: random forest and XGBoost (See Athey & Imbens (2019)). Our primary objective is to train the algorithms on the sample data to predict defaults "out-of-sample". Standard estimation techniques, such as Ordinary Least Squares (OLS), which utilize all available data to make in-sample predictions, are unsuitable for this analysis. The in-sample estimation methods are unbiased, leaving only the variance to be optimized in order to minimize the out-of-sample prediction error. Thus, OLS does not allow for joint optimality of bias and variance. Machine learning techniques account for the bias-variance trade-off through a joint optimization procedure and are particularly suitable for our study. Moreover, our subsample analysis based on standard logistic regressions highlights significant variation in the predictive ability of different digital variables based on the nature of the borrower. For instance, the presence of a social network app is negatively related to defaults for borrowers located in regions with low levels of financial inclusion but positively related to defaults for those in regions with high levels of financial inclusion. These results provide a compelling motivation to use machine learning methods which are better suited to account for non-linear relationships between explanatory and outcome variables (Bali, Beckmeyer, Moerke & Weigert (2022)).

Our primary goal is to examine whether alternate data can be used to assess credit risk of individuals who lack a traditional credit score. To this end, we compare the discriminatory

ability of the mobile and social footprint variables with that of conventional credit screening variables that include credit scores and customer characteristics. Using the Random Forest algorithm, we find that the out-of-sample AUC of the model, which only includes mobile/social footprint, is 71%, approximately twenty percentage points higher than the model, which only includes credit score and equal to the conventional model, which includes both credit score and customer characteristics. The results are qualitatively similar based on other measures of prediction performance, such as accuracy and precision. In subsequent analysis, we use the borrowers with the CIBIL score as a training sample and treat those without the credit score as the testing (hold-out) sample. We use our training sample data to train our model and select optimal features using Random Forest and XGBoost algorithms. We then use these features to predict the default probability of the sample without a credit score. The mobile and social footprint variables do a good job of predicting defaults for the hold-out sample, with an AUC of 65% (59%) based on the Random Forest (XGBoost) algorithm. The predictive performance improves if we combine these variables with information regarding customer characteristics, such as age, salary, and education, which is also available to the fintech lender. In further robustness tests, we find that these results hold across different subsamples on income and education levels and across regions with varying levels of financial inclusion.

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. We find that both 'deep' financial information and 'deep' mobile footprint variables significantly improve default prediction. Second, the information content of deep mobile footprints complements and exceeds the 'deep' financial variables. Specifically, the out-of-sample AUC of the models, which includes only the deep mobile footprint variables, is fifteen percentage points higher than the AUC of the model with only 'deep' financial information.

Overall, we find that the mobile/social footprint has a significant ability to predict defaults, and the information content of these variables complements rather than substitutes both the credit bureau score and detailed financial information regarding a customer's income and expenses.

Our next set of analyses is based on the idea that Fintech lenders are likely to use historical data to train the models and then use the model to predict defaults for new loan applicants. To mimic their strategy, we use the loans that originated in a quarter (T) as the training sample to predict defaults for loans that originated in the subsequent quarter (T+1). We find

that the discriminatory ability of digital mobile and social variables is consistently higher than that of the CIBIL score throughout the sample period. The analysis implies that our results are robust to different time periods and unlikely to be driven by some time-specific shock.

Finally, we use the optimal features generated by the machine learning algorithms for default prediction to answer two counterfactual policy questions. 1) What proportion of the borrowers who lack traditional credit scores could be given loans if we were to access their creditworthiness using social and mobile footprints? 2) Can we potentially expand credit access without any adverse impact on loan portfolio performance?[9] We follow the methods outlined in Athey (2017) and Kleinberg et al. (2015) to address the aforementioned prediction counterfactual questions. Using our methodology, we find that if we use a predicted default threshold of 4% (similar to the in-sample default rate) for approving loans, about 59% of the borrowers without a credit score could be granted a loan. This represents a 22% improvement over the in-sample approval rate of 37%. This is because about 17% of such borrowers who were denied credit would be approved under our alternate credit scoring model. Additionally, we find that the estimated counterfactual approval rate for the full sample based on a credit scoring model that combines mobile/social footprint is 82%, nine percentage points higher than the in-sample approval rate of 73%.

In addition, subsample tests reveal that the potential benefit of using alternative data to evaluate creditworthiness is significantly greater for customers with low incomes and those residing in regions with low financial inclusion. 56% of customers without a credit score and earning less than the median salary level would be approved for a loan with a predicted default threshold of 5%. Given that the sample approval rate for loan applicants without a credit score and earning less than the median income is 29%, our analysis suggests that approximately 27% more consumers could be approved for a loan. Similarly, based on our alternative credit scoring model, approximately 15–19% more consumers living in relatively financially excluded areas could be granted credit. In conclusion, our counterfactual analysis indicates that deploying credit scoring models based on alternative data captured from mobile devices could expand credit access without negatively affecting default rates. Extrapolating these findings suggests that digital inclusion can result in financial inclusion (Jia & Kanagaretnam (2022)).

Overall, our study documents that mobile and social footprint variables have significant discriminatory power in 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

[9]These counterfactual prediction policy questions are not causal in nature, as our objective is to find the best predictor of the borrower's default risk.

with little or no credit history who are under-served by traditional banks (Chava et al. (2017)). Consistent with this conjecture, the average individual in our sample is a sub-prime borrower with a credit score of 633.[10] Moreover, an economically significant 37% 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 (Buchak, Matvos, Piskorski & Seru (2018), Tang (2019), Di Maggio & Yao (2019)). However, the use of machine learning algorithms combined with big data for credit allocation decisions is not without costs. Fuster et al. (2018) show that while the use of machine learning for evaluating creditworthiness can expand credit access for some borrowers, it can also exacerbate racial disparity in credit access and the interest rate charged to borrowers.

Our study is most closely related to Berg et al. (2020), who also examine the use of digital footprints for default predictions. Using data covering approximately 250,000 purchases from an E-Commerce company located in Germany, Berg et al. (2020) document that the digital footprint complements rather than substitutes for credit bureau information and is informative even for customers who do not have credit bureau scores. While related, our paper further builds on and complements their findings. First, our data is from a stereotypical fintech lender operating in a developing country and covers all kinds of loans and not just those for e-commerce purchases. Thus, our setting allows us to extrapolate the importance of mobile footprints in measuring creditworthiness for loans taken for different purposes, not just e-commerce purchases.

Second, our study underscores the usefulness of alternate data that capture deeper aspects of the digital footprint from the mobile phones of customers to expand credit access in a setting with low levels of financial development. This is important given that globally, about 50% of the users access the internet through mobile phones and 5% through tablets. This is particularly true in a developing country setting. For instance, 97% of the Internet access time in India is through mobile phones (Kantar-IMRB (2018)). Moreover, even in developed countries like the UK, the USA, and Germany, the fraction of users that access the internet primarily through mobile phones is increasing.

Moreover, fintech adoption is notably higher among the younger generation, who primarily access the internet via mobile phones (Frost, Gambacorta, Huang, Shin & Zbinden (2019)). Thus, given the mobile-based digital footprints and the developing country setting, our findings are potentially generalizable to other developing countries and the millennial generation. To the extent that mobile footprint variables complement the information content of credit score, the marginal value of such information is likely to be higher in contexts

[10] The credit scores and associated risk tiers in India are: 801–900 (*Prime plus*), 751-800 (*Prime*), 651–750 (*Near prime*), and 300–650 (*Subprime*).

with weak information environments. Thus, fintech firms that rely on the mobile footprint for screening borrowers maybe even more important in expanding credit access in countries with weak information environments and lower levels of financial inclusion (Hau, Huang, Shan & Sheng (2019)).

Finally, and most importantly, we use novel machine learning algorithms to predict default and conduct counterfactual policy experiments to examine what proportion of loan applicants would be given loans if one were to rely on social/mobile footprints to access their creditworthiness. This is uniquely possible in our setting because we have information on the full sample of loan applicants, those who were approved and denied credit. We conclude that using alternate credit scoring based on mobile and social footprints can expand credit access without adversely affecting the overall default rate. This has significant policy implications for financial inclusion. Ours is the first paper to conduct such an alternative data-based counterfactual analysis for lending.

Further, because we have data on the salary, education, job, and detailed income and expense of the customers, we can disentangle whether digital footprint only proxies for these characteristics or provides incremental information. For instance, we find that owning an IOS device has predictive power even after controlling for earnings. Importantly, for a subsample of our customers for whom we have deep financial information, such as a borrower's income and spending patterns. We run a horse race between the "deep financial" variables with the borrowers' mobile/social footprints to evaluate their credit risk. This comparison helps us get important insights about the relative role of different kinds of information in traditional versus fintech lending.

Overall, our study contributes to the recent but growing body of work examining the implications of increasing usage of financial technology, big data, and machine learning algorithms on consumer welfare (Chava et al. (2017), Fuster et al. (2018), Fuster, Plosser, Schnabl & Vickery (2019), D'Acunto et al. (2019), Rossi & Utkus (2019), Tang (2019), Balyuk (2019)) and the broader economy (Philippon (2016), Buchak et al. (2018), Chen, Wu & Yang (2019)). Our study also has implications for further innovations in open banking. Open banking allows customers to securely share their financial data with different financial intermediaries securely, thus promoting competition among lenders, lowering transaction costs, and expanding credit access (Rishabh (2022), He, Huang & Zhou (2023)). Our findings suggest that incorporating mobile/social footprints within the open banking data-sharing framework can further improve credit screening technology and expand credit access.[11]

[11] See Thakor (2019) and Berg, Fuster & Puri (2022) for a survey of the literature on Fintech and Fintech lending.

II Data and summary statistics

We obtain proprietary data on about 363,165 loan applicants from a mobile-only Fintech lending platform operating in India since 2016. The lender aims to provide short-term credit to young salaried professionals by using their mobile footprints, and social footprint to determine their creditworthiness even when a credit history may not be available. The fintech lender provides loans of amount ranging from a minimum of ₹10,000 ($135) to ₹200,000 ($2696).[12] The loan duration ranges from a minimum of 15 days to a maximum of 180 days. We obtained data from the lending firm for all loans granted from February 2016 to November 2018.

A total of ₹6500 million ($88 million) worth of loans have been disbursed since its inception in 2016. To get a loan, a customer has to download the lending app and submit all the requisite details and documentation. The borrower allows the lender to gather additional information on the mode of login, the various apps installed, the number of calls and SMSes, the number of contacts on the phone, the number of social connections, and the kind of mobile operating systems such as IOS and Android. Table B1 of appendix C provides detailed description of the variables used in our study.

II.A Summary statistics: loan and financial Variables

Table 1 reports the summary statistics.[13] Out of the 363,165 loan applications in our sample, 265,007 were approved, while 98,158 were rejected. The default rate in our full sample is approximately 4.5%[14]. The average loan size for the set of approved customers is ₹22,188 ($299) age of the customer is 32, consistent with the idea that the lending firm target segment is a young salaried customer. The average credit score is 633 and is obtained from TransUnion CIBIL. The average interest rate charged on a loan is 25% (log value of 1.4). Thus, the lender caters to relatively higher-income customers. The application process also records the purpose of the loan: Medical, Travel, EMI, Purchases, Loan Repayment, and Others. Among the sample of approved loans, 8% were taken for travel, 9% for EMI, 13 % for purchasing a good, about 8% for repaying a loan, 22% for medical expenditure, and the rest are uncategorized.

[12] Based on the nominal exchange rate of $1=₹74.16 as of June 11th 2021.

[13] In section A.I of Appendix A, we discuss the results based on univariate analysis to examine how different variables related to loan approval and defaults. We also discuss the differences in characteristics of the customers with and without a CIBIL score.

[14] $\frac{12{,}008 \text{ defaults}}{265{,}007 \text{ approvals}}$

II.A.1 Sample representativeness

On average, a customer in our sample earns ₹37,709 ($508) per month or $6101 per annum. Our sample is representative of the lower to upper middle income population in India. While there is no clear consensus on the definition of middle and high-income groups among policymakers, academics, or the public at large (Atkinson & Brandolini (2013), Gornick & Jäntti (2014)), several studies define those a daily income in the range of $10-$50 (2011 PPP) per person as belonging to the middle class (Meyer & Birdsall (2012), Bank (2015), Birdsall (2015), Kochhar (2020)). Consistent with our sample comprising of the middle-income group in India, the average monthly salary is ₹37,709 ($ 508).[15] A monthly salary of ₹37,709. translates into approximately $12 per day for a household of 4 and $48 for a single person in 2011 PPP. The average consumer in our sample is 32 years old.

While there is no publicly available data on the demographics of the average retail borrower in India, we obtained data on a random sample of customers who received credit cards from one of the largest publicly listed traditional banks in India for comparison. This data provides suggestive evidence regarding the external validity of our study. The average borrower in the traditional banks earns ₹38, 057, comparable to the income (₹37,709) of the average borrower in the fintech sample. However, the average consumer in the traditional bank is older at 39 years relative to the average age of 32 years in the fintech sample. The average age of an Indian citizen is 29. The lower age in our fintech sample is not surprising given that prior literature highlights that fintech adoption is higher among the young as they are more comfortable with new technologies (Frost et al. (2019)). Further, fintech lenders in developing economies are likely to cater to new to credit, tech-savvy customers who were likely underserved by traditional banks (Hau et al. (2019)). Notably, a 2022 report by Transunion estimates that about 160 million potentially creditworthy underserved customers, primarily in the age group of 18-33, can be provided credit if their default risk could be determined using alternate means (TransUnion (2022)). According to the study, this is also the group of customers actively seeking credit. Consistent with this thesis, 37% of the customers in our fintech sample lack credit scores, suggesting that our sample is representative of the new to credit, younger, and traditionally underserved customers.

II.B Summary statistics: mobile and social footprint variables

In addition to the credit bureau score and other customer-level variables, the lender also captures information on the various kinds of mobile applications installed on the user's phone: such as Facebook, Linkedin, financial apps, dating apps, e-commerce apps, and travel apps.

[15] Exchange rate of $ 1= ₹74.16 on June 11th 2021.

The app also collects data on other variables that may capture the social behavior and status of the customer, such as the number of calls, the number of SMSes, the number of contacts on the phone, the number of social media connections, and the kind of mobile operating systems such as IOS and Android. Facebook (Linkedin) dummy variables identify customers that logged in to the app using Facebook (Linkedin). About 27% of customers logged in to the app using Facebook, while 2% used Linkedin. On average, 68% of the customers have a banking or stock trading app. About 43% of customers have installed another mobile-loan application suggesting that they look for loans on other platforms as well, while 12% of the customers own an Apple phone (IOS dummy).

III Results

III.A Simple Logit regressions

Given the ease of interpretability, we report the simple logit analysis first as they help us uncover the marginal effects of each of the digital variables we used to predict default. We focus on the out-of-sample tests using machine learning algorithms in section III.B. Formally, we run logit regressions of loan outcome measures on, mobile and social footprint, and customer characteristics:

$$\begin{aligned}\text{Default}_{ilt} = \beta_0 + \sum_{j=1}^{M} \beta_j \text{Loan Characteristics}_{lt} + \sum_{j=1}^{N} \beta_j \text{Customer financials}_{it} \\ + \sum_{j=1}^{O} \beta_j \text{Customer mobile/social footprint}_{it} + \varepsilon_{ilt}\end{aligned} \tag{1}$$

where i identifies a unique customer, l identifies a unique loan, and t refers to a year-month. *Default* is a dummy variable that identifies loans in default. Customer financial refers to customer age, salary, education, and job designation. Customer mobile/social footprint refers to all the variables summarized and discussed in the previous section.

In this section, we focus on analyzing the relationship between mobile/social footprint variables, credit score, customer characteristics, and default. The dependent variable in these tests is a dummy variable that takes the value one for delinquent loans. Panel A of Table 2 reports the results from these tests. Focusing on columns 1–6, we first analyze the subsample of loan applicants with non-missing values of all digital mobile footprint variables and customer characteristics.

Column 1 reports the results using only the credit bureau score (*CIBIL*) as the explana-

tory variable for the sample of approved loans. Not surprisingly, a higher credit bureau score is associated with a significantly lower likelihood of default. The AUC of credit score in our sample at 55%, while significantly different from chance (AUC of 50%), is lower than 62% reported by Iyer et al. (2015) based on a sample of loans from the peer-to-peer lending platform, "Propser.com" but comparable to the AUC of 59.8% using U.S. credit scores from Lending Club reported in Berg et al. (2020). Fintech lenders in developing economies are likely to cater to new to credit, tech-savvy customers who were likely underserved by traditional banks (Hau et al. (2019)). Using a simple theoretical model, Hau et al. (2019) conjecture that fintech expands the extensive margin of credit to customers with poor credit history by implementing alternative and better credit screening technology. This conjecture is borne out in a comprehensive dataset of 28.67 million credit offers in China. The information asymmetry regarding such "new to credit" customers is likely to be high, and consequently, conventional credit scores may be poor at assessing default risk. Thus, the discriminatory ability of the credit score in predicting defaults is likely to vary across countries and types of financial intermediaries depending on the information asymmetry regarding the borrowers. In column 2, we include customer characteristics, excluding mobile footprint variables. Focusing on individual explanatory variables, we find that salary, age, and education are negatively related to defaults.

In column 3, we report the results for mobile and social footprint variables. The AUC of this specification is 62% and 7% more than the AUC of the model with just the credit bureau score. Focusing on the individual variables, we find that mobile and social footprint variables may proxy for hard to quantify aspects of individual behavior, which has implications for the likelihood of default. We find that individuals that have a financial app installed on their phones have a significantly lower likelihood of default. The odds ratio of *Finsavvy* dummy is 0.75, implying that individuals without a financial app are about 25% more likely to default than those with such an app installed. This suggests that *Finsavvy* dummy may be correlated with the financial sophistication of a customer. Interestingly, customers with a mobile loan app are about 26% more likely to default. Finally, those who log in to the application via Linkedin are 32% less likely to default relative to those who login via other means. Focusing on the IOS dummy, we find that borrowers with IOS operating system (Apple) are significantly less likely to default than those with the Android operating system. The odds ratio of *IOS* dummy is 0.577, implying that those with an android phone are twice as likely to default as those with an Apple phone.

As mentioned before, it is difficult to pin down the channel through which these variables may be affecting the likelihood of default. However, to the extent that the objective in a credit scoring exercise is to increase the precision of predicting default, these results indicate

that the nature of the apps installed on the phone has significant discriminatory power in default prediction.

Since the primary aim of our study is to examine if digital information can be used to assess the credit risk of a borrower without a conventional credit score, we compare the predictive performance of traditional screening technology (Credit Score + Customer characteristics) with that of social and mobile footprints. In column 4, we examine default predictions based on a model that combines both credit score and customer characteristics. The AUC of the traditional model at 53.6% is 8.5% lower than a model with just social and mobile footprints. To the extent that even fintech lenders may have access to details regarding customer characteristics, in column 5, we evaluate the prediction performance of a model that combines customer characteristics with social and mobile footprints. We find this model does marginally better than the model with just digital variables reported in column 3. In column 6, adding credit score to a model with customer characteristics and social/mobile footprint doesn't significantly improve predictive performance.

Overall, we document that mobile and social footprint variables can be used to predict the likelihood of default and can perform at least as well as the credit score. These findings have implications for expanding credit access to those without a credit history and, consequently, a credit score so long as we can capture enough aspects of their mobile footprint. To further strengthen this thesis, in the next section, we focus on predicting defaults using mobile and social footprints for borrowers without a credit score.[16]

III.A.1 Predicting defaults for customers without credit score

Our analysis so far suggests that the digital mobile footprint has incremental explanatory power for predicting defaults. However, customers who lack credit history and credit score may be very different from the set of customers with a credit bureau score. There are 50,551 borrowers in our sample who were approved for loans but did not have a credit score.[17] To examine if these results can be generalized to this set of customers, in columns 7–9 of Table 2, we focus on the set of customers without a credit score and examine whether and how digital mobile footprints perform in default prediction for this subsample. In column 7, we only include customer characteristics and find that these have significant discriminatory power with an AUC of 57%. In column 8, we include only mobile and social footprint variables and find that the AUC of the model is 55% and comparable to the credit bureau score's predictive

[16] To further strengthen the evidence regarding the discriminatory ability of digital mobile footprint variables in predicting defaults, in Table A2 of Appendix A, we repeat our baseline models on subsamples based on credit score, age, salary, and job designation terciles. We find that the digital variables retain their discriminatory abilities across such subsamples.

[17] Table **??** in Appendix A reports the summary statistics and univariate analysis for this set of customers.

performance in column 1. Importantly, in column 9, we include customer characteristics and mobile footprint variables together to examine if mobile footprint variables have incremental explanatory power over customer and loan characteristics. Compared to column 7, including digital variables improves the AUC by 1.6%, which is considered a significant improvement.[18]

These findings suggest that digital mobile footprints have significant discriminatory power to predict defaults and can be used to score customers without a conventional credit score.

III.B Machine Learning Models for Default Prediction and Credit Scoring

III.B.1 Motivation

Our results thus far document a strong relationship between social and mobile footprints and defaults. In this section, we examine whether we can use the social and mobile imprints to create an "alternate credit score", which can be used to give loans to borrowers without credit history or traditional credit score.

Therefore, the problem at hand is to assess whether social and mobile footprints predict loan default. This is essentially a prediction problem, where we want to use the sample data to predict the risk of defaults "out-of-sample". Standard estimation approaches like OLS or Logit, where we use all the data to make "in-sample prediction," is not well suited for such analysis. The in-sample estimation approach first minimizes bias and then the variance of the estimator, which in turn ignores the bias-variance trade-off in minimizing the out-of-sample prediction error. In contrast, machine learning techniques minimize the mean squared error of the prediction by a joint minimization procedure cognizant of the bias-variance trade-off and, thus, better suited to address our research question.

Importantly, comparing the coefficients in Table 2 and Table A2 (reported in Appendix A), we find that customer characteristics and digital footprint variables are differentially related to defaults, depending on the type of consumers. For instance, the coefficient on *Socialconnect app* dummy is negatively related to defaults for borrowers located in regions with low levels of financial inclusion but positively related to defaults for those in regions with high levels of financial inclusion. Similarly, the coefficients on *Linkedin status* and *IOS dummy* vary significantly across sub-samples. These results provide a compelling motivation to use machine learning methods which are better suited to account for non-linear relationships between explanatory and outcome variables (Bali et al. (2022)).

Finally, machine learning methods are appropriately suitable to ask counterfactual questions such as: how many rejected borrowers (perhaps due to lack of traditional credit score)

[18] See, for instance, Iyer et al. (2015).

would have been approved had we used the social and mobile footprint-based alternate credit scores? What would have been the impact on default rates 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 the borrower's default risk. We follow the methodology outlined in Athey (2017) and Kleinberg et al. (2015) to address the counterfactual prediction questions in this section. Using different machine learning algorithms, we start by verifying the predictive power of social and mobile footprint variables for defaults.[19] Subsequently, we conduct the counterfactual prediction exercise.

III.B.2 Data Preparation

III.B.2.i Training versus Testing Split

Due to their high degree of non-linearity, machine learning models often have a very high "in-sample" predictive power due to over-fitting. It is, therefore, essential to evaluate the predictive power of the machine learning model "out-of-sample" or the testing set. For our analysis, following standard procedure, we split the data into three groups: the training set (60% of the sample), the validation set (20% of the sample), and the hold-out sample (20% of the sample). We report all our prediction performance measures based on the hold-out sample. Section A.III in Appendix A provides details of the various measures used in our study to evaluate the performance of a particular machine learning model.

III.B.2.ii Balancing the Data

One common issue in machine learning-based classification problems like default prediction is the skewness in the data because the outcome has a low likelihood of occurrence. For example, the proportion of default in our dataset is a little over 4.5%. Thus, the training dataset is heavily populated by one outcome (in our case, non-default). In such a situation, machine learning prediction algorithms are most likely to predict non-default. It is advisable to balance the training sample data by increasing the default sub-population representation to mitigate such concerns. There are various ways to deal with unbalanced data issues, like under-sampling the majority (non-default) group, oversampling the minority (default group), or generating synthetic data from the minority class (SMOTE).[20] In our analysis, we have used SMOTE followed by Edited Nearest Neighbor (ENN) to deal with the unbalanced data problem.[21] Figure A1 in Appendix A reports a representative example of the raw unbalanced

[19] Section A.II in Appendix A describes the three machine learning models used in our study.

[20] see Chawla, Bowyer, Hall & Kegelmeyer (2002) for a detailed explanation of SMOTE. The survey paper More (2016) covers a variety of techniques to deal with unbalanced data.

[21] Our results are robust to various other techniques to deal with unbalanced data. Our results are qualitatively similar if we apply the ML algorithms on unbalanced data without SMOTE and ENN. Specifically,

and balanced sample.

While we use a balanced training dataset for estimation purposes, we use actual sample data for out-of-sample predictions (testing sample). Therefore, all the reported out-of-sample prediction performance measures are based on actual observed loan outcomes.

III.B.3 Feature Scaling

We scale all explanatory variables (features) are scaled by their means and standard deviations, i.e., for all variables $x_i, i = 1, .., k$, the normalized data z_i is used as the feature.

$$z_i = \frac{x_i - \mu_x}{\sigma_x}$$

III.C Comparison of social and mobile footprints

In this section, we compare the machine learning algorithms and evaluate the performance of the mobile and social footprint variables relative to the traditional variables like the credit scores and other customer characteristics using machine learning models.[22] Panel A of Table 3 reports these results. We note that the mobile and social footprint variables alone have a much higher AUC score in predicting the probability of default relative to the borrower's credit score (CIBIL) across both the methods of machine learning algorithms. Specifically, focusing on panel A, we find that under the Random Forest algorithm, the model with only CIBIL score has an AUC (Accuracy) of about 51% (76.5%) while the models with digital mobile/social footprint variables have an AUC of about 71% (84%). Further, the AUC and accuracy of the alternate model with digital mobile/social footprint variable is comparable to the traditional model which includes both credit score and customer characteristics. The results are qualitatively similar using the XGBoost algorithm.[23]

Figure 2 plots the AUC curves for the two machine learning models (Random Forest and XGBoost) corresponding to different feature groups. The $X - axis$ of the plot measures the

we note that even without SMOTE and ENN, the predictive performance of digital variables is significantly better than the conventional model that includes credit score and customer characteristics. However, the performance of ML algorithms significantly improves with SMOTE-ENN. For example, in Random Forest models with SMOTE-ENN balancing, the AUC is about 75% for the low-education borrowers with all feature groups. This drops to about 68% in the testing sample without SMOTE-ENN. This is because applying ML to unbalanced data with few defaults may result in overfitting in the training sample resulting in lower out-of-sample performance.

[22] Section A.II in Appendix A provides additional details regarding the machine learning estimation procedure.

[23] In Tables A3 reported in Appendix A, we find qualitatively similar results if we use Logistics Regression for default prediction. Since the tree-based machine learning models perform better with underlying heterogeneity and clustering in the data (Efron & Haste (2018)), for brevity, we only report results using Random Forest and XGboost in rest of the paper.

false positive rate and the $Y - axis$ measures the true positive rate. Section A.III in the appendix explains these metrics in detail. The AUC of the alternative data based features like the digital information is more than the traditional features like the borrower's credit score and customer characteristics.

Figure 3 shows the feature importance values of the top ten features for predicting defaults using the Random Forest model. Feature importance measures the contribution of each feature in a Random Forest model. The feature importance is determined by examining the decrease in accuracy of the model when a particular feature is randomly permuted. If a feature is important for making predictions, then permuting its values should significantly reduce the model's accuracy. In figure 3, we report the top ten features from the Random Forest model. Seven of the top ten features are related to the digital variables. While credit score (CIBIL) is one of the top ten features, it is not the most important. Digital variables such as the number of contacts, number of apps, and the natural log of call logs are top features along with the borrower's salary.

Figure A2 in Appendix A reports the prediction's confusion matrix using all features. Panel (a) of Figure 3 reports the confusion matrix from Random Forest specification with a default threshold of 50%. We show the prediction performance in the testing dataset for actual versus predicted default (and actual non-default versus predicted non-default) in a 2×2 contingency table. For the full sample, the actual default rate was close to 4.5%. The lower left entry corresponds to the non-defaults. Out of about 95.5% (83.1% + 12.4%) actual non-defaults in the testing sample, the model correctly predicts about 83.1% non-defaults in the testing sample. The upper-right-hand entry shows that out of 4.5% (2.9% + 1.6%) of actual defaults in the testing sample, the model correctly predicts 1.6% of defaults. Hence the model's accuracy for correct prediction is close to 84.7%(83.1% + 1.6%). Note that the confusion matrix's lower-left entry is also referred to as the Type I error. Our objective is to evaluate if we can use alternative data to evaluate the creditworthiness of potentially good borrowers. Thus, identifying the non-defaulters as good borrowers is the more relevant metric in our context. On that metric, our Random Forest model with alternative data can correctly identify 83.1% of the 95.5% of actual good borrowers in the testing sample. These results are qualitatively similar using the XGboost specification in panel (b) of figure A2.

III.C.1 Borrower heterogeneity based on credit scores

We next turn to address the issue of whether social and mobile footprint variables have different levels of predictive power for borrowers based on their credit score heterogeneity. The underlying idea is that social and mobile footprint variables may be particularly useful in predicting defaults for customers with low or no credit scores as there is likely to be greater

information asymmetry regarding such customers. Panels B and C of the Table 3 reports the results for the subsample of customers with low and no credit score respectively.

III.C.1.i Subsample with low credit scores: role of the social and mobile footprints

Focusing on panel B of Table 3, we find that the social and mobile footprint variables have higher predictive power for default than the CIBIL score for the borrowers in the bottom 25% of the CIBIL score distribution. Figure A3 in Appendix A plots the AUC curves for the two machine learning models. Using the Random Forest method, the AUC of the model with only the mobile/social footprint variables is 76%, whereas that of the model with the CIBIL score is 53%. Moreover, the AUC of the mobile/social footprint model is higher for the borrowers in the bottom 25% of the CIBIL (AUC of 76%) score relative to the full sample of customers with a credit score in panel A (71%).

Overall, we conclude that mobile and social footprint variables have greater predictive power as compared to the CIBIL score for all customers and especially so for customers with low credit scores.

III.C.1.ii Subsample with no credit scores: role of the social and mobile footprints

As mentioned before, a large number of potential borrowers around the world lack credit score and consequently access to credit. Alternate credit scoring mechanisms would be especially useful if it can be used for default prediction and consequently to expand access to credit for these set of individuals. In this section, we focus on the set of borrowers without a credit score and use a sample splitting technique to evaluate the performance of the social and mobile footprint variables as a measure of alternative credit score in predicting the default for these group of borrowers.

We use the borrowers with the CIBIL score as a training sample and treat the borrowers without the CIBIL score as the testing sample. We use our training sample data to train our model and select the optimal features using Random Forest, and XGBoost algorithms. We then use the features to predict the default probability of the hold-out sample: the set of borrowers who were approved without the CIBIL score. We report the performance of these alternative measure based credit scores in panel C of Table 3. We find again that the mobile and social footprint variables together do a remarkable job in predicting defaults even for the testing sample (without credit scores) with AUCs in the range of 65%.

III.C.2 Borrower heterogeneity based on demographics

One may be concerned that the performance of the mobile/social footprint variables documented thus far may not hold for some subsamples. In particular, our interest is in examining whether these alternate data features can predict defaults for the relatively marginalized population, such as those with low income or education levels. In Table 4, we repeat our baseline machine learning analysis on subsamples based on income and education. Columns 1-5 of Panel A reports the results for the subsample of customers with below median income level. Consistent with our baseline results reported in Table 3, we find that the mobile/social footprint variables have significant discriminatory ability for default prediction. The AUC of the Random Forest model with only mobile/social footprint is 75%, approximately 28% more than the AUC of the model based on credit score alone, and 3% more than the traditional model which includes CIBIL score and customer characteristics. The results are qualitatively similar for the high income group.

In Panel B of Table 4, we repeat the tests with subsamples based on education. The key insights remain same.

III.C.3 Regional heterogeneity based on ex-ante financial inclusion

In this section, we examine the performance of the mobile/social footprint variables for subsamples of customers based on the level of financial inclusion in their district of residence. The underlying idea is to examine if such variables' discriminatory ability in terms of default prediction holds for customers located in areas with low levels of financial inclusion. In particular, we construct two ex-ante measures that capture different dimensions of financial inclusion. Our first measure is the fraction of households without bank accounts. Second, we use a comprehensive district-level measure of financial inclusion, which is annually released by CRISIL. It combines three critical parameters of basic financial services: bank branch penetration, deposit penetration, and credit penetration into one metric in the form of an index. It is a relative index with a scale of 0 to 100. We invert the index for ease of interpretation so that higher values of both our measures imply lower levels of financial inclusion.[24]

Table 5 reports the results of these analyses. Panel A, first set of columns reports the results for the subsample of customers in regions with the above-median value of the fraction

[24]In unreported tests, we also use two alternate measures of financial inclusion: a) a proxy for bank branch penetration that captures the average number of adults serviced by one bank branch in an area (Adults per Unit Bank Branch in a district) and b) the percentage of state-owned bank branches in a district. This second measure is based on the idea is that private banks are less likely to expand in financially excluded lower-income areas. In contrast, given their mandate to promote social welfare, state-owned branches are more likely to open branches in such areas. The results are qualitatively similar using these alternate measures.

of unbanked households. Again, we find that the mobile/social footprint variables have a significant discriminatory ability for default prediction. The AUC of the model with mobile/social footprint is 73%, approximately 28% more than the AUC of the model based on credit score alone. The results are qualitatively similar for the customers located in regions with a low fraction of the unbanked population, reported in the next set of columns. For these set of customers, we again find that the CIBIL score has lower AUC than the model with mobile/social information variables. The borrower characteristics together with CIBIL score has similar AUC as the mobile/social footprint variables alone. This again suggests that the use of alternative data for credit scores is more promising for relatively financially excluded individuals.

In Panel B of Table 5, we repeat the tests with subsamples based on the comprehensive financial inclusion index discussed above. Again, our main thesis holds for customers in regions with low and high level of financial inclusion.

III.D Comparison of deep social and deep financial variables

Thus far, we have relied on rudimentary 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 to improve upon the default prediction. For instance, if the presence of a financial app on a customer's phone can predict defaults, it would not be unreasonable to conjecture that the duration of time spent across different kinds of apps, time spent on social media, nature and time of online searches etc; could have incremental explanatory power for default prediction. Unfortunately, we do not have detailed information regarding the customer's usage of different installed applications. We do, however, have detailed call logs for a large subsample of borrowers in the data. Prior literature highlights that call log patterns can be used to infer an individual's social capital (Singh & Ghosh (2017), Wiese, Min, Hong & Zimmerman (2014)), which is known to be an important predictor of loan defaults (Karlan (2005), Karlan, Mobius, Rosenblat & Szeidl (2009)).

Following prior literature, we create two kinds of proxies using call logs that attempt to capture the breadth and strength of an individual's social capital. We proxy for breadth using total frequency and duration of daily incoming, outgoing, and missed calls. Singh & Ghosh (2017) find that the frequency of missed calls and duration of incoming vs. outgoing calls is also related to reciprocity– the propensity of an individual to respond to and engage in calls associated by others. We proxy for the strength of an individual's social connections using the average number and duration of calls per person. The underlying idea is that an

individual is likely to make a greater number of calls or longer duration calls to people with whom they have stronger ties. Finally, we create a Herfindahl index, which captures whether the calls of an individual are concentrated over a few connections or spread across multiple contacts. These measures are constructed both ex-ante based on the call logs information available prior to loan approval, and ex-post based on the call logs information available in the first 15 days after loan approval. Table B1 of Appendix B provides the details of how we construct these measures. Panel A of Table A4 and A5 report the univariate summary statistics for the subsamples with and without credit score respectively..

Focusing on the total and the average number of missed calls per person, we see that defaulters, on average, are less likely to accept calls initiated by others. Defaulters are also more likely to have their calls concentrated over a smaller number of individuals, as evidenced by the HHI index for all measures of incoming/outgoing calls. 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, suggest that defaulters are less likely to respond to calls initiated by others. These patterns are consistent across ex-ante and ex-post call logs based measures.

For a subset of customers in our sample, we also have detailed information regarding their financial transactions, income, expenditure, investments, account balance before and after salary, etc. A detailed description of the 73 "deep financial" variables are available in Table B2 of Appendix B. 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.

In Table 6, we compare the discriminatory ability of digital footprint variables relative to deep financial variables. We find that both simple mobile footprint variables and deep social footprint variables have a greater discriminatory ability in predicting defaults relative to deep financial variables. Focusing on the Random Forest model, we find that the AUC of the model with the deep social footprint is 83%, about 15% more than the AUC of the model with only deep financial information. Moreover, a model that includes the mobile footprint and deep social variables performs significantly better in predicting default out of sample as compared to a model with deep financial information and CIBIL score.[25]

[25]The AUC of the model with mobile and deep social variables at 90.9% is 25% higher than AUC of a model with deep financial variables and credit score.

We conclude that both mobile and deep social footprint variables have 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.

III.D.1 Dynamic quarter-ahead analysis

We also report our results for different sub-periods, wherein we train the model on the sample of approved loan applicants in a quarter T and use it to predict defaults for the set of loans granted in quarter T+1. For example, suppose we have 50,000 data on loans that originated in 2017 Q1, and 52,000 loans originated in 2017 Q2. We split this data into 80:20 for training and validation, i.e., we use 40,000 loans originated in 2017 Q1 as the training sample and 10,000 as a validation sample. We then use the entire 2017 Q2 data (52,000 loans) as the testing sample to evaluate our model's prediction performance.

Our quarter-ahead analysis serves three purposes. First, prior studies highlight that credit scoring models used by Fintech lenders may improve over time as the consumer base expands. For instance, Balyuk & Davydenko (2019) reports that the AUC of Prosper's credit score in predicting defaults shows a linearly increasing trend between 2013 and 2018. Prediction based on data from the entire sample might get averaged in such a setting, thus reducing the prediction accuracy. Second, we want to make sure that our finding that mobile and social footprints have a higher predictive ability relative to conventional credit scores is robust to different time-periods. For instance, one concern with alternate data models is that customers may learn about these models and potentially alter their digital behavior to get a higher score. This implies that some of the mobile/social footprint variables may lose their predictive ability over time. Finally, fintech lenders are likely to use historical data to train the models and then use the model to predict defaults for new loan applicants. Our quarter-ahead analysis is consistent with this idea. We discuss the quarter-ahead analysis in detail in section A.IV. Reassuringly, our quarter-ahead analysis confirms the higher discriminatory ability of social and mobile footprint variables over the conventional credit score. Importantly, the predictive ability remains stable over time.

IV Counterfactual policy experiment for loan approval

Our results thus far show that social and mobile footprints have high predictive power for assessing borrower credit risk. The predictive power outweighs that of the traditional variables like the credit score or other customer characteristics. A natural follow-up question is whether we can use the social and mobile footprint variables for accessing creditworthiness

of borrowers who do not have traditional credit scores. Specifically, we seek answers to the following counterfactual questions: 1) What proportion of the borrowers would have been given loans if we had relied on accessing their creditworthiness using social and mobile footprints? 2) Can we potentially expand credit access without any adverse impact on default rates if we were to screen borrowers who were not approved (possibly because of a lack of credit score), based on social and mobile footprints?

We follow Kleinberg et al. (2015) in addressing the counterfactual policy questions posed above.[26] Our algorithm proceeds in the following steps:

1. We first split the sample of all borrowers who were approved into a training and testing sample. We then use the different machine learning algorithms to estimate the model parameters. Since a relatively small portion of the approved borrowers eventually defaulted, we use SMOTE and ENN methods described in section III.B.2.i to balance the training sample. We then use a cross-validation procedure to minimize the error term to choose the best model. We then use the testing sample to evaluate the prediction of the default risk of the model.

2. We use the predicted model from step 1 and apply it to the borrowers with and without credit score who were not approved for a loan to predict their probability of default. Next, we use a default thresholds of 5% for the predicted probability of default to evaluate how many borrowers who were not approved would have been approved based on alternate credit screening technology.[27]

Since Random Forest emerges as the better prediction model in both the baseline as well as in quarter-ahead analysis, we report results using the Random Forest model as the first step in the prediction counterfactual analysis.[28]

In Table 7, we report the results of our counterfactual exercise, examining the fraction of borrowers who were denied credit but would have been approved for the full sample of customers. We begin by comparing counterfactual approval rates using the traditional screening technology (only customer characteristics and the credit score) with the counterfactual approval rates based on the model with only mobile and social footprint variables. We find that the counterfactual approval rate is 68% based on the traditional model. The counterfactual approval rate is 9% higher if we were to use an alternate credit scoring model

[26] The focus on prediction policy counterfactual rather than causal questions is relatively new in economics (Athey (2017), Kleinberg et al. (2015)).

[27] These results are qualitatively similar at different default thresholds. We choose a default threshold of 5% to be consistent with the in-sample default rate of 4.5%.

[28] In unreported tests, we find that the results are qualitatively similar if we use other machine learning models.

based on mobile/social footprint. Moreover, the fraction of rejected customers who would be approved under the alternate scoring model is also higher at 18% compared to 13% for the traditional screening technology. Finally, we use a model that combines Mobile/Social footprint with customer characteristics and find that about 82% of the loan applicants would be approved under this alternate model, a 9% improvement over the in-sample approval rate of 73%. Moreover, our counterfactual estimates, not reported here for brevity, suggest that we could maintain a higher approval rate of 78% even at a conservative predicted default threshold of two percentage points.

We next provide a back of the envelope calculation of the lender's revenue increase due to improved credit access using our alternate credit scoring model. There is one-to-one mapping between interest rate and loan duration in our setting. Thus, the interest rate is completely determined by the loan duration. Now, the average loan amount for the set of approved customers is ₹22,467. The average interest rate and loan duration for the set of approved customers is 1.75% per month and 2.7 months. The total potential interest earning over a 2.7 month period is ₹1,062 (=0.0175*22467*2.7) from one loan applicant. The total number of approved loan applicants in our sample is 265,007. We assume that the entire interest is forfeited in the event of a default. Thus, assuming a default rate of 5%, the total potential interest earnings from the set of approved customers is approximately ₹267 million (=1062*265007*0.95).

The average loan duration for the rejected applicants is approximately 2 months and the corresponding interest rate is 1.5% per month. The average loan amount for the rejected applicants is ₹19,447. The total potential interest earning over a 2 month period is 0.015*19447*2=₹583 from one loan applicant. The total number of rejected loan applicants in our sample our 98,158. At a predicted default threshold of 5%, the counterfactual number of rejected customers who would have been approved using our alternate credit scoring technology is about 24.5% (per the counterfactual analysis reported in Table 7). Thus, assuming a default rate of 5%, the total potential interest earnings from the set of denied customers who would be approved under the counterfactual model is approximately ₹13 million (=583*0.245*98158*0.95). Therefore, if the lender were to use our alternate credit scoring for loan approval, its revenues would increase by 5% ($\frac{13*100}{267}$).

IV.A Counterfactual policy experiment for customers without credit score

Panel A of Table 8 reports the counterfactual analysis for the borrowers without a CIBIL score who were rejected. In these experiments, we again combine Mobile/Social footprint

variables with customer characteristics. At the ex-ante predicted default threshold of 5%, about 59% of the borrowers without a credit score would have been approved. Note that the approval rate for customers without a credit score in our sample is 37% (See Panel (a) of Figure 1). Our counterfactual exercise suggests that using an alternate credit score model, we could expand credit access for about 22% more borrowers without a credit score, even at a conservative predicted default threshold of five percentage points. Importantly. 17% of borrowers who were rejected would be approved if their creditworthiness were to be evaluated using our model based on the mobile/social footprint variables.

Overall, these results indicate that evaluating creditworthiness based on social and mobile footprints can potentially expand credit access to financially excluded borrowers without adversely affecting loan performance.

IV.B Counterfactual policy experiment with demographic and regional heterogeneity

In this section, we investigate heterogeneity in the potential effects of the alternative credit scoring method by conducting our counterfactual policy experiment on different subsamples. In Table A7 of Appendix A, we report the actual approval rates for various subsamples as a benchmark for the counterfactual analysis.

In panels B and C of Table 8, we examine the counterfactual approval rates for income- and education-based subsamples. Focusing on Panel B, we find that about 56% of customers without a credit score and below the median salary level would be approved for a loan at a predicted default threshold of 5%. Given that the approval rate of such loan applicants in our sample is 29% (See Column 1 of Table A7), our analysis suggests that about 27% more such customers could potentially receive a loan. At the same (5%) predicted default rate, about 21% of the rejected borrowers would have been approved. The results for the high-income group are qualitatively similar, albeit with a relatively higher approval rate of 62%. 15% of the rejected borrowers in the high-income cohort would have been approved. Overall, these results imply that the benefits of using alternative credit scoring methodologies based on mobile/social footprint are likely to be greater for lower-income consumers.

In Panel C, we repeat the analysis with subsamples based on education level and without a credit score. We find that the marginal benefit of alternate credit scoring methods is significantly higher for applicants with higher education attainment levels.

Finally, in panels C and D of Table 8, we repeat the counterfactual exercise for subsamples of customers based on the level of financial inclusion in their district of residence. We document an overall increase in approval rates for customers without a credit bureau score

across regions with low and high levels of financial inclusion. Importantly, the marginal increase in approval rates is greater for customers located in regions with lower levels of financial inclusion.

In summary, we conclude that alternative credit scoring methods based on mobile/social footprints have the potential to increase credit access for the relatively marginalised population that lacks access to a traditional credit bureau score, particularly those residing in ex-ante financially excluded regions.

V Conclusion

We examine the discriminatory ability of individual borrowers' mobile and social footprints in predicting loan defaults using a unique and proprietary dataset from a top fintech lending company in India. We show that a predictive model based on an individual's mobile/social footprint and deeper social footprint based on call logs outperforms a traditional model that relies on credit score and customer characteristics in predicting defaults, using both simple Logit regressions and advanced machine learning algorithms. This is consistent with the growing usage of alternative credit scoring technologies by fintech companies around the world.

Overall, our study documents that alternate data such as mobile and social footprint variables have significant discriminatory power in evaluating credit risk. Importantly, our counterfactual prediction exercise indicates that with the use of such alternate data, fintech lenders can potentially build credit scores and can expand access to credit to even customers with little or no credit history who are underserved by the traditional banks. Our study has broader policy implications as the world is developing new modes of financial intermediation, such as open banking to expand credit access to traditionally underserved customers suggesting that digital inclusion can result in financial inclusion.

However, a caveat with credit scores relying on alternate data is that customers may learn about the underlying variables through information leakage or the experience of other customers. Thus, customers may change their digital behavior to get a higher score rendering some digital variables unsuitable for the lending decision. In contrast, traditional credit scores and customer characteristics are less vulnerable to manipulation. Our quarter-ahead analysis indicates that the predictive power of the digital variables remains stable, at least during the sample period of our study. Nonetheless, the potential for manipulation is an important concern for alternate credit score models, and fintech lenders using alternate data must continuously evaluate their credit risk models to mitigate the adverse consequences of such manipulation.

References

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483–485.

Athey, S. & Imbens, G. (2019). Machine learning methods economists should know about. *Working paper*.

Atkinson, A. B. & Brandolini, A. (2013). On the identification of the middle class. *Income inequality: Economic disparities and the middle class in affluent countries*, 77–100.

Bali, T. G., Beckmeyer, H., Moerke, M., & Weigert, F. (2022). Option return predictability with machine learning and big data. *Review of Financial Studies, Forthcoming*.

Balyuk, T. (2019). Financial innovation and borrowers: Evidence from peer-to-peer lending. *Working paper*, (2802220).

Balyuk, T. & Davydenko, S. A. (2019). Reintermediation in fintech: Evidence from online lending.

Bank, W. (2015). *A measured approach to ending poverty and boosting shared prosperity: Concepts, data, and the twin goals*. The World Bank.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies*, *33*(7), 2845–2897.

Berg, T., Fuster, A., & Puri, M. (2022). Fintech lending. *Annual Review of Financial Economics*, *14*, 187–207.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063–1095.

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, (pp. 81–88).

Birdsall, N. (2015). Does the rise of the middle class lock in good government in the developing world? *The European Journal of Development Research*, *27*(2), 217–229.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Buchak, G., Matvos, G., Piskorski, T., & Seru, A. (2018). Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics*, *130*(3), 453–483.

Burks, S. V., Cowgill, B., Hoffman, M., & Housman, M. (2015). The value of hiring through employee referrals. *The Quarterly Journal of Economics*, *130*(2), 805–839.

Chava, S., Paradkar, N., & Zhang, Y. (2017). Winners and losers of marketplace lending: evidence from borrower credit dynamics. *Working paper*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, M. A., Wu, Q., & Yang, B. (2019). How valuable is fintech innovation? *The Review of Financial Studies*, *32*(5), 2062–2106.

D'Acunto, F., Rauter, T., Scheuch, C., & Weber, M. (2019). Perceived precautionary savings motives: Evidence from fintech. *Working paper*.

Das, S. R. (2019). The future of fintech. *Financial Management*, *48*(4), 981–1007.

Di Maggio, M. & Yao, V. W. (2019). Fintech borrowers: Lax-screening or cream-skimming. *Working paper*.

D'Acunto, F., Prabhala, N., & Rossi, A. G. (2019). The promises and pitfalls of robo-advising. *The Review of Financial Studies*, *32*(5), 1983–2020.

Efron, B. & Haste, T. (2018). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press.

Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., & Zbinden, P. (2019). Bigtech and the changing structure of financial intermediation. *Economic Policy*, *34*(100), 761–799.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2018). Predictably unequal? the effects of machine learning on credit markets. *Working paper*.

Fuster, A., Plosser, M., Schnabl, P., & Vickery, J. (2019). The role of technology in mortgage lending. *The Review of Financial Studies*, *32*(5), 1854–1899.

Gornick, J. C. & Jäntti, M. (2014). *Income inequality: Economic disparities and the middle class in affluent countries*. Stanford University Press.

Hau, H., Huang, Y., Shan, H., & Sheng, Z. (2019). How fintech enters china's credit market. In *AEA Papers and Proceedings*, volume 109, (pp. 60–64).

He, Z., Huang, J., & Zhou, J. (2023). Open banking: Credit market competition when borrowers own the data. *Journal of Financial Economics*, *147*(2), 449–474.

Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2015). Screening peers softly: Inferring the quality of small borrowers. *Management Science*, *62*(6), 1554–1577.

Jia, X. & Kanagaretnam, K. G. (2022). Does digital inclusion relate to financial inclusion? further evidence from peer-to-peer lending. *Working paper*.

Kantar-IMRB (2018). Twenty first edition of icube report. Technical report.

Karlan, D., Mobius, M., Rosenblat, T., & Szeidl, A. (2009). Trust and social collateral. *The Quarterly Journal of Economics*, *124*(3), 1307–1361.

Karlan, D. S. (2005). Using experimental economics to measure social capital and predict financial decisions. *American Economic Review*, *95*(5), 1688–1699.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–95.

Kochhar, R. (2020). A global middle class is more promise than reality. In *The middle class in world society* (pp. 15–48). Routledge India.

Meyer, C. & Birdsall, N. (2012). New estimates of india's middle class. *CGD Note, Center for Global Development, Washington, DC*.

More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.

Philippon, T. (2016). The fintech opportunity. *Working paper*.

Rishabh, K. (2022). Can open banking substitute credit bureaus.

Rossi, A. & Utkus, S. (2019). Who benefits from robo-advising. *Working paper*.

Schmitt, P., Skiera, B., & Van den Bulte, C. (2011). Referral programs and customer value. *Journal of marketing*, *75*(1), 46–59.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, *90*(2), 227–244.

Singh, V. K. & Ghosh, I. (2017). Inferring individual social capital automatically via phone logs. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 95.

Tang, H. (2019). Peer-to-peer lenders versus banks: substitutes or complements? *The Review of Financial Studies*, *32*(5), 1900–1938.

Thakor, A. V. (2019). Fintech and banking: What do we know? *Journal of Financial Intermediation*, 100833.

TransUnion (2022). Empowering credit inclusion: A deeper perspective on credit underserved and unserved consumer. Technical report.

Wiese, J., Min, J.-K., Hong, J. I., & Zimmerman, J. (2014). Assessing call and sms logs as an indication of tie strength. *Working paper*.

TABLE 1: Summary statistics of customer and loan characteristics

This table reports summary statistics on the customer characteristics, loan characteristics and mobile/social footprints. Columns 1-3 compares these characteristics for loan applications that were approved and those that were denied. Columns 4-6 compares these characteristics for approved and disbursed loans that were in default and those that were not in default. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

	Approved (1)	Not Approved (2)	Difference (3)	Default (4)	Not Default (5)	Difference (6)
Loan Amount	22188.50	19063.97	3124.53***	41039.372	21293.330	19746.042***
Loanpurpose Medical	0.213	0.096	0.118***	0.221	0.213	0.008**
Loanpurpose Travel	0.082	0.027	0.055***	0.071	0.082	-0.012***
Loanpurpose EMI	0.086	0.066	0.020***	0.071	0.087	-0.015***
Loanpurpose purchase	0.132	0.065	0.066***	0.128	0.132	-0.004
Loanpurpose Loanrepayment	0.081	0.041	0.040***	0.073	0.081	-0.008***
Loanpurpose Other	0.407	0.215	0.192***	0.436	0.405	0.031***
Age	31.90	30.15	1.75***	31.824	31.905	-0.081
Salary	37709.58	31868.34	5841.24***	37527.046	37717.727	-190.681
CIBIL (>0, N=219k & 16k)	632.63	512.66	119.97***	593.510	634.189	-40.678***
Facebook Status	0.271	0.282	-0.010***	0.261	0.272	-0.010**
Linkedin Status	0.022	0.016	0.006***	0.017	0.022	-0.006***
Googleplus_status	1.70	1.70	0.004*	1.722	1.706	0.016***
Referral	0.118	0.044	0.073***	0.095	0.119	-0.023***
Sales App	0.195	0.197	-0.002	0.186	0.195	-0.009**
Dating App	0.029	0.027	0.002	0.026	0.029	-0.004**
Finsavy app	0.683	0.047	0.636***	0.500	0.691	-0.192***
Socialconnect app	0.717	0.050	0.668***	0.544	0.726	-0.182***
Travel app	0.575	0.061	0.514***	0.430	0.582	-0.152***
Mloan app	0.427	0.028	0.399***	0.326	0.432	-0.106***
Referrer	0.236	0.044	0.192***	0.130	0.241	-0.111***
# of SMS	2470.71	1256.43	1214.28***	1762.993	2503.389	-740.396***
# of Apps	54.46	42.10	12.35***	42.254	50.751	-8.497***
# of Contacts	842.68	717.82	124.86***	803.118	844.565	-41.447***
# of Connections	524.34	401.17	123.16***	425.641	528.447	-102.807***
# of Calls	3062.72	2058.95	1003.77***	2153.926	3104.617	-950.691***
IOS	0.121	0.086	0.035***	0.121	0.121	0.000
Education						
<High School	0.113	0.275	-0.162***	0.119	0.112	0.007**
High School	0.647	0.568	0.079***	0.661	0.646	0.015***
College	0.240	0.157	0.083***	0.220	0.241	-0.021***
Job Designation						
Worker	0.363	0.384	-0.020***	0.347	0.364	-0.017***
Supervisor	0.245	0.260	-0.015***	0.241	0.245	-0.005
Manager	0.391	0.355	0.036***	0.412	0.390	0.021***
N	265,007	98,158		12,008	2,52,999	

TABLE 2: Predicting loan defaults using mobile and social footprint

This table reports the estimates from our logit regressions examining the relationship between mobile/social footprint variables, customer characteristics and likelihood of default for customers with credit bureau score. The dependent variable, Default takes the value one for loans that are delinquent and zero otherwise. Columns 1–6 report the results for the sample of customers with CIBIL score (credit bureau score). The specification in Column (1) includes only the (Log of CIBIL). Column (2) includes only customer characteristics. Column (3) includes only mobile/social footprint variables excluding IOS dummy. Column (4) includes the CIBIL score and customer characteristics. Column (5) includes only mobile/social footprint variables and customer characteristics but not the CIBIL score. Column (6) includes all variables including the CIBIL score. Columns 7–9 report the results for the sample of customers without the CIBIL score. Column (7) includes only customer characteristics. Column (8) includes only mobile/social footprint variables excluding IOS dummy. Column (9) includes the CIBIL score and customer characteristics. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

	Customers With CIBIL						Customers Without CIBIL		
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log of cibil	0.922***			0.925***		0.935***			
	(0.000)			(0.000)		(0.000)			
Log of Salary		0.950*		0.950*	1.072**	1.081***	1.385***		1.431***
		(0.065)		(0.065)	(0.015)	(0.006)	(0.000)		(0.000)
Log Age		0.810**		0.810**	0.526***	0.549***	1.174		0.970
		(0.022)		(0.022)	(0.000)	(0.000)	(0.192)		(0.809)
High School Dummy		0.945		0.945	0.990	0.989	0.937		0.963
		(0.173)		(0.173)	(0.815)	(0.791)	(0.277)		(0.536)
College Dummy		0.797***		0.797***	0.837***	0.836***	0.805***		0.824***
		(0.000)		(0.000)	(0.000)	(0.000)	(0.001)		(0.005)
Supervisor Dummy		1.122***		1.122***	1.142***	1.146***	1.015		1.007
		(0.001)		(0.001)	(0.000)	(0.000)	(0.768)		(0.884)
Manager Dummy		1.190***		1.190***	1.252***	1.255***	1.214***		1.223***
		(0.000)		(0.000)	(0.000)	(0.000)	(0.000)		(0.000)
Log no of SMS			0.951***		0.947***	0.947***		0.979**	0.978**
			(0.000)		(0.000)	(0.000)		(0.023)	(0.017)
Log No of Contacts			0.940***		0.935***	0.936***		1.001	0.962*
			(0.000)		(0.000)	(0.000)		(0.964)	(0.093)
Log no of Apps			0.658***		0.649***	0.651***		0.896***	0.874***
			(0.000)		(0.000)	(0.000)		(0.000)	(0.000)
Log Callog			0.908***		0.908***	0.908***		0.948***	0.957***
			(0.000)		(0.000)	(0.000)		(0.000)	(0.001)
Finsavy App			0.750***		0.743***	0.746***		0.441***	0.479***
			(0.000)		(0.000)	(0.000)		(0.000)	(0.001)
Socialconnect App			0.958		0.988	0.984		1.186	1.203
			(0.593)		(0.883)	(0.846)		(0.421)	(0.391)
Travel App			1.009		0.995	0.990		0.925	0.893
			(0.781)		(0.873)	(0.767)		(0.263)	(0.109)
Mloan App			1.261***		1.258***	1.258***		1.114	1.100
			(0.000)		(0.000)	(0.000)		(0.448)	(0.502)
Facebook status			1.015		1.019	1.017		0.884***	0.882***
			(0.607)		(0.534)	(0.579)		(0.008)	(0.007)
Linkedin status			0.677***		0.671***	0.669***		1.010	0.913
			(0.000)		(0.000)	(0.000)		(0.942)	(0.509)
IOS Dummy			0.578***		0.555***	0.555***		1.031	0.984
			(0.000)		(0.000)	(0.000)		(0.818)	(0.905)
Observations	180,381	180,381	180,381	180,381	180,381	180,381	42,963	42,757	42,722
Pseudo R-squared	0.00107	0.00136	0.0194	0.00234	0.0215	0.0223	0.00582	0.00583	0.0114
AUC	0.551	0.531	0.621	0.536	0.627	0.628	0.566	0.553	0.582

TABLE 3: Predicting Defaults using machine learning

This table reports results for different machine learning models to evaluate the default prediction performance of mobile and social footprint variables relative to traditional credit scores and other customer characteristics. Specifically, we compare three groups of variables a) CIBIL score b) Customer characteristics c) Mobile/Social Footprints. Panel A reports the results for all customers with CIBIL score. Panel B shows results for the subsample of borrowers with credit score in bottom 25% of the distribution. Panel C reports the results for the subsample of borrowers with no CIBIL score. For each of the sub-sample analyses, we report AUC, Accuracy, Precision, Recall and F1 score measures based on the out-of-sample tests.

Panel A: Performance Evaluation using customers with CIBIL

Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)
Random Forest	Only Cibil	0.509	0.765	0.050	0.227	0.082
	Only Mobile/Social Footprint	0.710	0.839	0.093	0.301	0.143
	Only Customer Characteristics	0.668	0.816	0.079	0.288	0.125
	Cibil + Customer Characteristics	0.713	0.829	0.092	0.310	0.142
	Cibil + Mobile/Social Footprint	0.730	0.843	0.109	0.338	0.164
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.738	0.847	0.113	0.349	0.171
Xgboost	Only Cibil	0.472	0.783	0.051	0.213	0.083
	Only Mobile/Social Footprint	0.627	0.734	0.075	0.439	0.128
	Only Customer Characteristics	0.554	0.678	0.055	0.378	0.096
	Cibil + Customer Characteristics	0.626	0.716	0.075	0.463	0.130
	Cibil + Mobile/Social Footprint	0.671	0.672	0.078	0.572	0.138
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.671	0.702	0.081	0.536	0.141

Panel B: Performance Evaluation using customers with Bottom 25% CIBIL

Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)
Random Forest	Only Cibil	0.534	0.415	0.056	0.669	0.103
	Only Mobile/Social Footprint	0.760	0.858	0.141	0.374	0.205
	Only Customer Characteristics	0.710	0.833	0.102	0.311	0.154
	Cibil + Customer Characteristics	0.740	0.840	0.118	0.340	0.176
	Cibil + Mobile/Social Footprint	0.789	0.858	0.154	0.438	0.228
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.800	0.864	0.178	0.479	0.260
Xgboost	Only Cibil	0.516	0.412	0.056	0.670	0.103
	Only Mobile/Social Footprint	0.689	0.746	0.096	0.494	0.160
	Only Customer Characteristics	0.588	0.710	0.069	0.393	0.117
	Cibil + Customer Characteristics	0.575	0.687	0.066	0.402	0.114
	Cibil + Mobile/Social Footprint	0.714	0.741	0.100	0.547	0.169
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.708	0.748	0.106	0.542	0.177

Panel C: Performance Evaluation using customers with no CIBIL

Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)
Random Forest	Only Mobile/Social Footprint	0.649	0.762	0.116	0.321	0.170
	Only Customer Characteristics	0.614	0.732	0.100	0.308	0.151
	Mobile/Social Footprint + Customer Characteristics	0.674	0.747	0.115	0.356	0.174
Xgboost	Only Mobile/Social Footprint	0.593	0.668	0.103	0.438	0.167
	Only Customer Characteristics	0.576	0.599	0.094	0.482	0.157
	Mobile/Social Footprint + Customer Characteristics	0.596	0.617	0.098	0.503	0.164

TABLE 4: Predicting defaults using Machine Learning with demographic heterogeneity

This table reports results for different machine learning models to evaluate the default prediction performance of mobile and social footprint variables relative to traditional credit scores and other customer characteristics for different demographic subsamples. Specifically, we compare three groups of variables a) CIBIL score b) Customer characteristics c) Mobile/Social Footprints. Panel A, Columns (1), (2), (3), (4) and (5) ((6), (7), (8), (9) and (10)) report results for the subsample of customers with below (above) median income. Panel B, Columns (1), (2), (3), (4) and (5) ((6), (7), (8), (9) and (10)) report results for the subsample of customers with below (above) median education level. For each of the sub-samples analyses, we report AUC, Accuracy, Precision, Recall and F1 score measures based on the out-of-sample tests.

Panel A: Income Level

		Low					High				
Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)	AUC (6)	Accuracy (7)	Precision (8)	Recall (9)	F1 (10)
Random Forest	Only Cibil	0.479	0.525	0.036	0.430	0.067	0.459	0.512	0.036	0.402	0.066
	Only Mobile/Social Footprint	0.753	0.874	0.092	0.274	0.137	0.695	0.842	0.091	0.321	0.142
	Only Customer Characteristics	0.689	0.853	0.080	0.302	0.127	0.648	0.804	0.067	0.276	0.108
	Cibil + Customer Characteristics	0.728	0.861	0.098	0.293	0.146	0.688	0.818	0.077	0.317	0.124
	Cibil + Mobile/Social Footprint	0.761	0.882	0.117	0.334	0.173	0.715	0.839	0.099	0.349	0.154
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.770	0.887	0.112	0.302	0.163	0.703	0.840	0.089	0.304	0.137
Xgboost	Only Cibil	0.455	0.682	0.038	0.287	0.067	0.434	0.721	0.040	0.235	0.068
	Only Mobile/Social Footprint	0.646	0.685	0.060	0.521	0.108	0.635	0.736	0.071	0.453	0.122
	Only Customer Characteristics	0.561	0.788	0.048	0.266	0.081	0.526	0.850	0.053	0.150	0.079
	Cibil + Customer Characteristics	0.596	0.666	0.060	0.489	0.106	0.645	0.760	0.080	0.467	0.137
	Cibil + Mobile/Social Footprint	0.666	0.689	0.063	0.530	0.112	0.697	0.740	0.086	0.541	0.148
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.640	0.702	0.058	0.465	0.103	0.666	0.737	0.082	0.516	0.141

Panel B: Education Level

		Low					High				
Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)	AUC (6)	Accuracy (7)	Precision (8)	Recall (9)	F1 (10)
Random Forest	Only Cibil	0.490	0.572	0.042	0.390	0.076	0.455	0.554	0.034	0.348	0.061
	Only Mobile/Social Footprint	0.746	0.842	0.114	0.343	0.171	0.697	0.845	0.089	0.297	0.138
	Only Customer Characteristics	0.702	0.840	0.090	0.288	0.137	0.673	0.817	0.066	0.257	0.105
	Cibil + Customer Characteristics	0.723	0.826	0.094	0.279	0.140	0.697	0.829	0.082	0.324	0.131
	Cibil + Mobile/Social Footprint	0.748	0.838	0.110	0.358	0.168	0.710	0.837	0.088	0.325	0.138
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.746	0.835	0.112	0.321	0.166	0.746	0.846	0.097	0.323	0.149
Xgboost	Only Cibil	0.499	0.728	0.055	0.312	0.094	0.426	0.725	0.034	0.207	0.059
	Only Mobile/Social Footprint	0.642	0.713	0.076	0.456	0.131	0.615	0.700	0.064	0.458	0.112
	Only Customer Characteristics	0.580	0.744	0.061	0.333	0.103	0.559	0.696	0.049	0.337	0.085
	Cibil + Customer Characteristics	0.595	0.685	0.076	0.466	0.131	0.632	0.731	0.072	0.478	0.125
	Cibil + Mobile/Social Footprint	0.648	0.699	0.073	0.478	0.127	0.662	0.698	0.071	0.542	0.126
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.644	0.711	0.084	0.472	0.143	0.678	0.714	0.077	0.534	0.135

TABLE 5: Predicting defaults using machine learning with regional heterogeneity

This table reports results for different machine learning models to evaluate the default prediction performance of mobile and social footprint variables relative to traditional credit scores and other customer characteristics for customers in regions with low/high levels of financial inclusion. Specifically, we compare three groups of variables a) CIBIL score b) Customer characteristics c) Mobile/Social Footprints. Panel A, Columns (1), (2), (3), (4) and (5) ((6), (7), (8), (9) and (10)) report results for the subsample of customers in regions with above (below) median value of the fraction of households without bank accounts. Panel B, Columns (1), (2), (3), (4) and (5) ((6), (7), (8), (9) and (10)) report results for the subsample of customers with below (above) median value of a comprehensive financial inclusions. For each of the sub-samples analyses, we report AUC, Accuracy, Precision, Recall and F1 score measures based on the out-of-sample tests.

Panel A: Fraction of Households Without Bank Accounts

		High					Low				
Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)	AUC (6)	Accuracy (7)	Precision (8)	Recall (9)	F1 (10)
Random Forest	Only Cibil	0.447	0.539	0.031	0.348	0.058	0.479	0.574	0.040	0.375	0.072
	Only Mobile/Social Footprint	0.733	0.845	0.092	0.304	0.141	0.703	0.844	0.081	0.265	0.124
	Only Customer Characteristics	0.686	0.829	0.077	0.292	0.121	0.675	0.818	0.076	0.284	0.120
	Cibil + Customer Characteristics	0.740	0.834	0.087	0.328	0.137	0.693	0.829	0.080	0.277	0.124
	Cibil + Mobile/Social Footprint	0.744	0.849	0.091	0.302	0.139	0.719	0.833	0.094	0.342	0.147
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.738	0.851	0.095	0.310	0.146	0.723	0.845	0.095	0.332	0.148
Xgboost	Only Cibil	0.405	0.809	0.038	0.152	0.061	0.434	0.773	0.042	0.190	0.069
	Only Mobile/Social Footprint	0.575	0.688	0.056	0.408	0.099	0.617	0.690	0.063	0.464	0.111
	Only Customer Characteristics	0.565	0.704	0.051	0.355	0.089	0.549	0.692	0.056	0.374	0.097
	Cibil + Customer Characteristics	0.629	0.725	0.071	0.485	0.124	0.605	0.731	0.071	0.428	0.122
	Cibil + Mobile/Social Footprint	0.647	0.701	0.068	0.501	0.119	0.654	0.644	0.066	0.563	0.117
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.659	0.712	0.072	0.511	0.127	0.638	0.693	0.065	0.487	0.115

Panel B: Financial Exclusion Index

		High					Low				
Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)	AUC (6)	Accuracy (7)	Precision (8)	Recall (9)	F1 (10)
Random Forest	Only Cibil	0.450	0.560	0.029	0.330	0.054	0.460	0.530	0.035	0.384	0.065
	Only Mobile/Social Footprint	0.699	0.858	0.082	0.254	0.125	0.719	0.846	0.074	0.273	0.116
	Only Customer Characteristics	0.675	0.830	0.064	0.254	0.102	0.677	0.827	0.076	0.292	0.120
	Cibil + Customer Characteristics	0.711	0.833	0.081	0.277	0.126	0.706	0.832	0.081	0.283	0.126
	Cibil + Mobile/Social Footprint	0.739	0.848	0.090	0.356	0.144	0.720	0.835	0.081	0.286	0.126
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.742	0.859	0.093	0.298	0.141	0.743	0.848	0.105	0.328	0.160
Xgboost	Only Cibil	0.412	0.707	0.030	0.219	0.053	0.426	0.748	0.036	0.191	0.060
	Only Mobile/Social Footprint	0.578	0.667	0.052	0.430	0.093	0.595	0.696	0.054	0.436	0.096
	Only Customer Characteristics	0.518	0.747	0.042	0.259	0.072	0.549	0.700	0.045	0.315	0.079
	Cibil + Customer Characteristics	0.621	0.752	0.075	0.418	0.128	0.596	0.744	0.071	0.413	0.121
	Cibil + Mobile/Social Footprint	0.664	0.671	0.060	0.554	0.108	0.647	0.659	0.066	0.545	0.117
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.644	0.696	0.062	0.485	0.111	0.634	0.704	0.072	0.480	0.125

TABLE 6: Deep Social vs. Deep financial information

This table reports results for different machine learning models to evaluate the default prediction performance of 'deep' social footprint and 'deep financial' variables relative to traditional credit scores and other customer characteristics. For each of the sub-samples analyses, we report AUC, Accuracy, Precision, Recall, F1 score, Realize and Predict measures based on the out-of-sample tests.

Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5))
Random Forest	All Features	0.901	0.981	0.169	0.272	0.208
	Only Cibil	0.511	0.485	0.010	0.537	0.020
	Only Customer Characteristics	0.768	0.940	0.078	0.411	0.131
	Only Deep Social	0.830	0.987	0.397	0.552	0.462
	Only Deep Finance	0.678	0.962	0.020	0.065	0.031
	Deep Finance + Cibil	0.655	0.963	0.012	0.031	0.017
	Deep Finance + Customer Characteristics	0.662	0.969	0.018	0.040	0.025
	Only Mobile/Social Footprint	0.776	0.939	0.078	0.476	0.134 0
	Mobile/Social Footprint + Deep Social	0.909	0.982	0.319	0.638	0.425
	Mobile/Social Footprint + Deep Social + Cibil	0.923	0.982	0.308	0.602	0.408 0
	Mobile/Social Footprint + Deep Social + Cibil + Customer Characteristics	0.904	0.984	0.337	0.560	0.420
Xgboost	All Features	0.922	0.960	0.097	0.401	0.157
	Only Cibil	0.504	0.711	0.009	0.270	0.018
	Only Customer Characteristics	0.652	0.618	0.017	0.582	0.032
	Only Deep Social	0.726	0.926	0.086	0.669	0.152
	Only Deep Finance	0.572	0.837	0.012	0.212	0.023
	Deep Finance + Cibil	0.588	0.858	0.015	0.197	0.027
	Deep Finance + Customer Characteristics	0.617	0.974	0.029	0.048	0.036
	Only Mobile/Social Footprint	0.623	0.665	0.014	0.480	0.027
	Mobile/Social Footprint + Deep Social	0.913	0.851	0.053	0.805	0.100
	Mobile/Social Footprint + Deep Social + Cibil	0.926	0.896	0.073	0.784	0.134
	Mobile/Social Footprint + Deep Social + Cibil + Customer Characteristics	0.927	0.921	0.086	0.714	0.154

TABLE 7: Counterfactual Policy experiment with all customers

This table reports results on the overall counterfactual approval rate and fraction of denied customers who would be approved at a predicted default likelihood of 5%.

All Customers		
Feature Groups	Counterfactual Approval Rate (1)	Fraction of Denied Customers who would be Approved (2)
Cibil +Customer Characteristics	0.680	0.134
Only Mobile/Social Footprint	0.773	0.177
Mobile/Social Footprint +Customer Characteristics	0.817	0.245

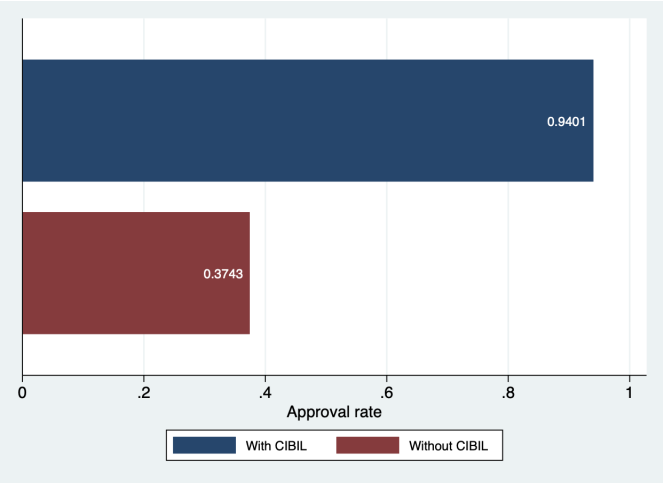
TABLE 8: Counterfactual Policy experiment for customers without credit score

This table reports results on the overall counterfactual approval rate and fraction of denied customers without a CIBIL score who would be approved at a predicted default likelihood of 5%.

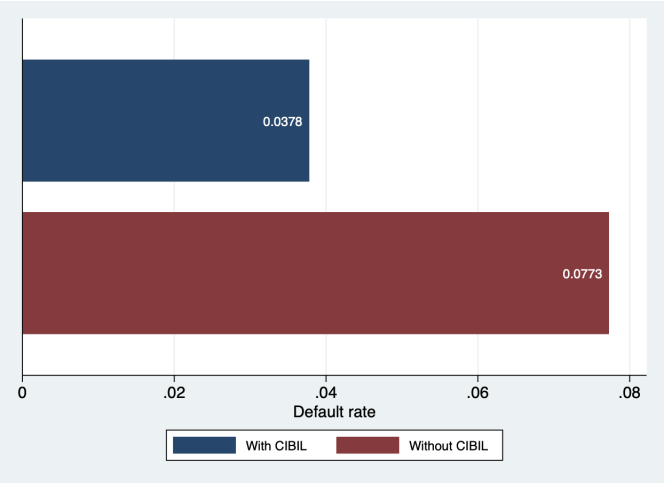
Customers without CIBIL Score		
Subsample	Counterfactual Approval Rate (1)	Fraction of Denied Customers who would be Approved (2)
Panel A: Full Sample of Customers without CIBIL Score		
All customers	0.589	0.168
Panel B: Heterogeneity based on Income Level		
High Income	0.623	0.153
Low Income	0.556	0.210
Panel C: Heterogeneity based on Education Level		
High Education	0.671	0.199
Low Education	0.400	0.131
Panel D: Heterogeneity based on Fraction of Households Without Bank Accounts		
High Fraction	0.602	0.179
Low Fraction	0.529	0.153
Panel E: Heterogeneity based on Financial Exclusion Index		
High Exclusion	0.623	0.192
Low Exclusion	0.502	0.149

Figure 1: Customers with and without credit score

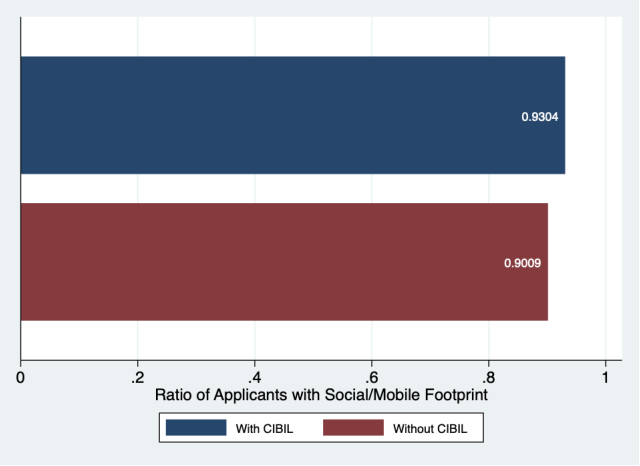
This figure shows the a few key statistics for customers with CIBIL and customers without CIBIL: Approval rate for credit score and no credit score customers (Panel A), Default rate for credit score and no credit score (Panel B), Digital footprint distribution for customers with and without credit scores (Panel C).



(a) Approval rate for credit score and no credit score customers

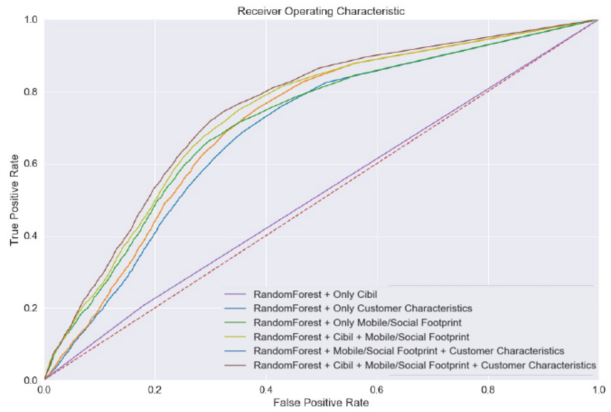


(b) % Default rate for credit score and no credit score customers

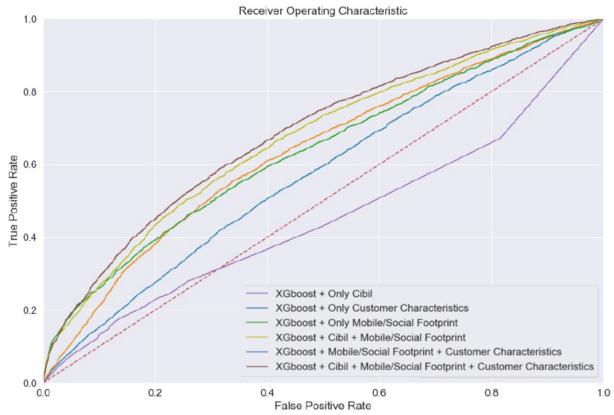


(c) Digital footprint distribution for customers with and without credit scores

Figure 2: AUC Plots for machine learning models (Full Sample of Customers with CIBIL Score)
This figure plots the AUC curves for default prediction based on the two machine learning models for the sample of customers with credit score. Panel A reports the figure based on Random forest and Panel B reports based on Xgboost.

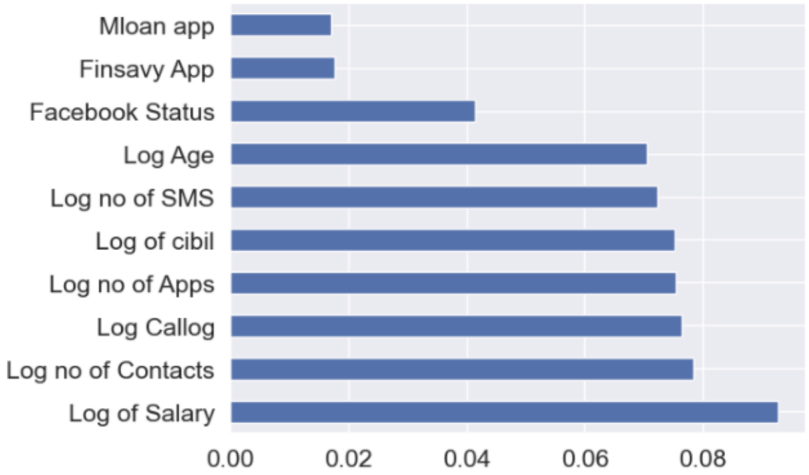


(a) Random forest



(b) XGboost

Figure 3: Feature Importance
This figure reports the top ten features for default prediction using the Random Forest model.



Appendix A

This appendix reports the set of additional results that are referenced in the text.

A.I Univariate analysis

In columns 1-3 of Table 1, we compare the customer and loan characteristics of loans that were approved and those that were denied. Surprisingly, the average size of the loan demanded is about 16% higher for loan applications that were approved.[29] Consistent with conventional wisdom, we also find that customers with a higher salary, credit score, and older customers have a higher likelihood of approval. Focusing on the mobile and social footprint variables, we find that, approved customers are more (less) likely to log in through Linkedin or Google (Facebook). Approved customers are also significantly more likely to have installed a financial app (banking apps, mutual fund apps, and stock tracking apps), social networking app (Facebook, Twitter, Whatsapp, and other chat apps). Whether or not the customer installs a dating app or an e-commerce application (such as Amazon and Flipkart captured in the *Sales* dummy) does not seem to be associated with the likelihood of loan approval. Customers that have either been referred by others (*Referral* dummy) and those who have referred others (*Referrer* dummy) are also more likely to be approved. On average, approved customers have a higher number of apps, send and receive a greater number of SMSes and calls, have a higher number of contacts but fewer connections on a social platform. Approved customers are also 3.5% more likely to own an iPhone (*IOS* dummy).

In columns 4-6 of Table 1, we analyze the customer and loan characteristics that can potentially predict the likelihood of default. Customers who default on average borrow 92% more than those who don't.[30]. Customers who default on average are charged a higher interest rate ex-ante, consistent with such customers being riskier. Not surprisingly, customers who default have lower credit scores.

Focusing on the social and mobile footprint variables, we find that customers who default are less likely to have logged in through either Facebook or Linkedin. This suggests that the mode of login has predictive power for the likelihood of default. Further, delinquent customers are more likely to have installed a social network app. We also find that other social footprint variables that capture various aspects of social behavior have a bearing on the likelihood of default. For instance, customers who were referred by others, and those who refer others are less likely to default.[31] Our finding suggests that social ties may have positive spillover effects on the customer's attitude towards default to the extent that the likelihood of referring or being referred is associated with the strength of an individual's social connections. Consistent with this idea that customers who do not default, send and

[29] $\frac{(22188.50-19063.97)*100}{19063.97}$

[30] $\frac{(41039-21293)*100}{21293}$

[31] This is consistent with the marketing and economics literature that finds that customers or employees acquired through referrals have a stronger sense of commitment and attachment to the firm (Schmitt, Skiera & Van den Bulte (2011), Burks, Cowgill, Hoffman & Housman (2015)). Using data on referred customers of a German bank, Schmitt et al. (2011) find that such customers have a higher retention rate and are more valuable in both the long and short term. Similarly, Burks et al. (2015) find that referred workers yield substantially higher profits per worker than non-referred workers.

receive a greater number of SMSes and calls have a higher number of contacts but fewer connections on a social platform. These variables again potentially capture the strength of the social ties of a customer. The number of apps also seem to have a discriminatory ability to predict defaults as defaulting customers have fewer apps.

In Table A1, we compare the customer characteristics, loan characteristics, and mobile/social footprints for the sub-sample of loan applicants with and without a credit bureau score. Customers without a credit score apply for smaller loans, are younger, have a lower income, and are less likely to be referred. We observe that while customers without a credit score also have a significant digital presence, their mobile footprint is relatively smaller than those with a credit score. Specifically, customers without a credit score have 13% fewer apps installed on their phone, send fewer SMS, and have fewer contacts and connections. Such customers are also less likely to have installed a financial, mobile-loan, social, or travel app.[32] Figure 1 reports the fraction of customers with and without credit score for which mobile/social footprint data is available. From Panel A of Figure 1, we learn that the availability of data on mobile/social footprint is similar across both sets of customers. Specifically, such data is available for about 93% and 90% of the customers with and without credit scores.

A.II Machine learning models

We use two widely used classification techniques namely Random Forest classification and XGBoost models to evaluate the predictive power of the mobile and social footprint variables relative to the traditional variables like the credit scores and other customer characteristics. The default prediction problem is called classification problem in machine learning, where the state of default is assigned a value 1 and the state of non-default is assigned a value 0. We briefly describe below various models.

A.II.1 Random Forest:

Random forest is a tree-based classification procedure to evaluate the default probability. In a tree-based classification algorithm, the dependent or outcome variable is discrete, like default. The feature set or X variables are divided into various sub-groups. The dependent variable's average value for a sub-group is the best predictor for each sub-group conditioned on X. For example, suppose there is only one feature variable – age. If we find that the average default rate is 5% for people above age 25 and 8% for people below age 25, then 5% and 8% are the best prediction of default rates for the two age-subgroups of the populations. The final outcomes in the two groups are called the *leaves* of the tree. The cutoffs of the age-based sub-groups are chosen by minimizing the error rate of prediction through a procedure called *pruning*. In the *pruning* procedure, first, a large tree with lots of sub-groups is created. The large tree is subsequently *pruned* by cost complexity pruning. Where for each value of the regularization parameter α, the following term is minimized:

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

[32] On average, customers without a credit score have 47 apps installed on their mobile device compared to 54 apps for those with a credit score. This translates into 13%=$\frac{(54-47)*100}{54}$ fewer apps.

where y_i is the actual outcome and $\hat{y}_{R_m}$ is the predicted outcome in the m^{th} terminal node. The tree procedure follows a cross-validation procedure to estimate the optimal α.

While the tree is straightforward to explain, it is typically non-robust and does not have the same level of predictive accuracy (in the testing sample) as in some other methods. These problems are overcome in a random forest by drawing multiple bootstrap samples of size Z^* and fitting a tree for each such sample. The average outcome over the bootstrap sample is the predicted value for each tree.[33] In each bootstrap sample, the set of features are randomly drawn to avoid a dominant feature affecting the prediction. (see Breiman (2001) and Biau (2012)).

A.II.2 Boosting regression trees (XGBoost)

In boosting, regression trees are expanded sequentially using information from previous trees. This is a slow learning approach where residuals from the current tree are used to improve the model. In the XGboost algorithm, the tree is updated using a gradient boosting method. The boosting has three parameters: the number of trees B, the regularization parameter α and the number of splits in each tree.

A.II.3 Logistic regression:

In a logistic regression the default probability is modeled as a logistic link function,

$$\Pr(Y = 1|X) = \frac{e^{\beta_0+\beta X}}{1 + e^{\beta_0+\beta X}}$$

where Y takes value 1 if the borrower defaults and 0 otherwise, X is termed as a set of features or explanatory variables. The estimation procedure follows by maximizing a likelihood function

$$l(\beta_0, \beta) = \Pi_{i:y_i=1} p(x_i) \Pi_{i:y_i=0} (1 - p(x_i))$$

The estimation procedure using machine learning follows a different procedure as compared to the standard logistic regression problem in traditional econometric analysis. In a standard logistic regression, we generally use the entire data set to estimate the coefficients. This procedure may result in overfitting if we have a large set of features with some features having few observations. The machine learning approach overcomes this issue by first splitting the dataset into training and testing samples. The training sample is used to fit the model, while the testing sample is used to evaluate the prediction of the model. The estimation procedure in the training sample follows a procedure called the minimization of the cross-validation errors to estimate the optimal parameters. In the cross-validation procedure, the training sample is further divided into k sub-groups, and the estimation procedure is performed in one sub-group and evaluated in the other sub-group to generate cross-validation errors.

[33] For a classification problem; typically, a majority vote is taken over the bootstrap sample.

A.III Model selection: prediction performance

There are various measures to evaluate the performance of a particular machine learning model. Area Under the Curve (AUC) and Recall are two widely used model evaluation criteria. The prediction performance is based on comparing the predicted default rate relative to the actual default rate. Default prediction is a classification problem where the state of default y_i for customer i is assigned a value 1 for default and 0 otherwise. For each of the machine learning models described above, we estimate a predicted probability of default conditional on the set of feature vector x_i. If the predicted probability is above a certain threshold λ (say $\lambda = 50\%$) then the predicted outcome is assigned as default. True positives (TP) is defined as the number of observations that the model correctly identifies non-defaulters. Similarly, True Negatives (TN) is the number of observations that are correctly classified as default. Conversely, False Positive (FP) is the number of times an actual default is incorrectly classified as non-default by the machine learning model. Similarly, False Negative (FN) is the number of times an actual non-default is incorrectly classified as a default.

The above definitions of TP, FP, FN and TN help us define the following performance criterion to evaluate the performance of a machine learning model.

$$\text{True Negative Rate} = \frac{TN}{TN + FP}$$

$$\text{False Positive Rate} = \frac{FP}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = 2 * \frac{\text{Pr ecision x Recall}}{\text{Precision+Recall}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{ROC AUC} = \int_{-\infty}^{\infty} TPR(\alpha) FPR'(\alpha) d\alpha$$

The area under the ROC curve is often used as a measure of the goodness of a prediction. For our purpose of predicting default, it captures the diagnostic ability of the default prediction model as the threshold of discrimination (α) changes. The ROC curve plots the True Positive rate (TPR) against the False Positive rate (FPR). The best possible prediction will be yield a point in the upper left corner or coordinate $(0, 1)$ of the ROC space.

The area under the ROC curve (AUC) measures the model's ability to classify actual default as default and actual non-default as not default. A higher AUC measure implies a better predictive performance of the underlying model. The AUC measure by construction

ranges between $[0, 1]$. An AUC measure close to 1 therefore, means the model is good in separating actual default as default and actual non-default as non-default. An AUC value equals 0.5 implies that the model has no separation power.

Accuracy, defined as the proportion of correct predictions out of total predictions, is also another measure of checking the performance of the machine learning model. While AUC and accuracy are often reported as a measure of the prediction performance of a machine learning model, the precision, recall and confusion matrices are also important other metrics. The precision, computed as the ratio between true positive relative to all positive predictions, measures the number of correctly predicted non-defaults relative to all predicted non-defaults. Precision, therefore is more like the type I error.

Recall, defined as the fraction of correctly identifying good borrowers (non-defaults (TP)), relative to all good borrowers (TP+FN). Recall, therefore measures how accurately the predictive model identify the good borrowers in the data, and can be thought of more like the type II error. Importantly, while both recall and precision are measures to evaluate the effectiveness of the model, they are often in conflict. Improving recall for instance, typically reduces precision. Since the focus of our paper is financial inclusion, whose objective is to correctly identify good borrowers, recall is more important than precision.

The $F1$ score, more like a harmonic mean, measures a balance between precision and recall. Finally, the confusion matrix depicts the distribution of correct and incorrect predictions in a 2 x2 matrix.

In all our tests, we report Accuracy and AUC based on out-of-sample (hold-out or test sample) tests. While we report all relevant metrics discussed above, in our analysis, we primarily focus on AUC as a measure of the prediction performance. This is because AUC has a few desirable properties. First, it is scale-invariant. Second, AUC is invariant to classification-threshold. It measures the quality of prediction irrespective of the classification threshold chosen.

A.IV Dynamic Quarter-ahead prediction

For this analysis, we use the lending data originated from the prior quarter T as the training and validation sample and used the following quarter data $T + 1$ as the testing sample to evaluate our prediction and analyze any dynamic patterns in prediction.[34] As expected for a new venture starting in a new market, we have very few observations in the early stages of the operation to conduct any meaningful analysis. The number of loans originating in a quarter steadily increases over time. Figure A4 plots the number of loans originating in each quarter over the entire sample period. We, therefore, start our quarter-ahead using 2016 Q4 as our training and validation sample and 2017 Q1 as the corresponding testing sample. The actual default rates for loans originating in each quarter are reported in Figure A5. Reassuringly, we note from the figure, the actual default rates for the samples starting from 2016 Q4 through 2018 are closely equal to each other and low without any specific temporal pattern.

In Figure A6, we depict the temporal patterns of AUC of the model with only CIBIL

[34] In unreported tests, we also use the entire sample of observations from the quarters 0–T as the training to sample, and the following quarter data $T + 1$ as the testing sample.

score along with the AUC of model with only the digital information and for the model with all the features. The AUC of the model with only the CIBIL score lies below throughout the sample period for each quarter.

The actual values of the prediction performance measures are reported in Table A6. One notable point that makes the quarter ahead analysis different from the entire sample is the intertemporal increase in the testing sample size. Specifically, given that the number of loans originating had a steady increase over time, the sample size at quarter $T+1$ is always larger than the sample size at quarter T. This is in sharp contrast to our baseline analysis, where we used the entire sample data from 2016 to 2018 as one sample and divided it into training, validation, and test samples in 60:20:20 proportions. Thus test samples for the entire sample were way smaller than the training samples. Therefore, the increasing number of observations in the testing sample for the quarter ahead predictions may not result in high predictability due to two reasons (a) the heterogeneity in the test sample is more than in the training sample and (b) since new customers are monotonically increasingly in each quarter, there is a potential under-representation of the training data, and any change in the covariates/features in the test sample might not be picked up while training the model. The resulting discrepancy between training and testing distributions, then, may lead to poor predictions (Bickel, Brückner & Scheffer (2007), Shimodaira (2000)).

However, reassuringly we find that the prediction performance measures are consistently high and steady in the quarter ahead prediction model throughout the sample period. Specifically, note from Table A6 and Figure A6 that the prediction performance of the digital variables are consistently higher than that of the CIBIL score.

Overall, the quarter-ahead analysis confirm the higher discriminatory ability of social and mobile footprint variables over the conventional credit score.

TABLE A1: Summary statistics: comparison of customers with and without credit score

This table reports summary statistics on the customer characteristics, loan characteristics and mobile/social footprints. Columns 1-3 compares these characteristics for customers with credit score and without credit score. (∗ ∗ ∗), (∗∗), (∗) denote statistical significance at 1%, 5%, and 10% levels respectively.

	With CIBIL (1)	Without CIBIL (2)	Difference=(2)-(1) (3)
Loan Amount	22466.97	19447.59	3019.37***
Log Interest Rate	1.432	1.237	0.195***
Loanpurpose Medical	0.216	0.123	0.093***
Loanpurpose Travel	0.083	0.040	0.043***
Loanpurpose EMI	0.087	0.069	0.018***
Loanpurpose purchase	0.132	0.082	0.049***
Loanpurpose Loanrepayment	0.082	0.050	0.031***
Loanpurpose Other	0.400	0.278	0.122***
Age	32.04	30.38	1.65***
Salary	38055.41	32878.72	5176.69***
Facebook Status	0.281	0.251	0.030***
Linkedin Status	0.023	0.017	0.006***
Googleplus_status	1.696	1.733	-0.037***
Referral	0.124	0.058	0.066***
Sales App	0.195	0.192	0.003
Dating App	0.029	0.031	-0.002
Finsavy app	0.791	0.038	0.753***
Socialconnect app	0.831	0.040	0.791***
Travel app	0.654	0.068	0.586***
Mloan app	0.496	0.022	0.474***
Referrer	0.250	0.088	0.163***
# of SMS	2417.59	1678.71	738.88***
# of Apps	53.65	46.80	6.85***
# of Contacts	845.19	747.74	97.44***
# of Connections	511.42	461.34	50.07***
# of Calls	2955.17	2518.80	436.36***
IOS	0.117	0.101	0.016***
Education			
<High School	0.117	0.223	-0.106***
High School	0.646	0.590	0.056***
College	0.236	0.186	0.050***
Job Designation			
Worker	0.365	0.375	-0.010***
Supervisor	0.244	0.258	-0.014***
Manager	0.391	0.366	0.024***
N	228,116	135,049	

TABLE A2: Heterogeneity in default prediction

This table reports the estimates from our logit regressions examining the relationship between customer characteristics, mobile/social footprint variables and likelihood of default for different sub-samples of customers in based on income, education, and financial inclusion levels in the customer's state. The dependent variable, Default takes the value one for loans that are delinquent and zero otherwise. The specification includes all variables including the credit score, customer characteristics, and mobile/social footprint variables. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

	Income Level		Education Level		Households without bank accounts		Financial Exclusion	
	Low	High	Low	High	High	Low	High	Low
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log of cibil	0.916***	0.857***	0.911***	0.968	0.931***	0.897***	0.923***	0.962***
	(0.000)	(0.000)	(0.000)	(0.212)	(0.000)	(0.000)	(0.000)	(0.008)
Log of Salary			0.881	1.201***	1.107**	1.150***	1.152***	1.032
			(0.239)	(0.001)	(0.026)	(0.009)	(0.007)	(0.404)
Log Age	0.351***	0.612**	0.803	0.382***	0.600***	0.588***	0.533***	0.579***
	(0.000)	(0.010)	(0.353)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)
High School Dummy	0.903	1.211			0.931	0.945	0.965	1.009
	(0.242)	(0.124)			(0.312)	(0.408)	(0.624)	(0.878)
College Dummy	0.541***	1.113			0.712***	0.816**	0.880	0.808***
	(0.000)	(0.409)			(0.000)	(0.018)	(0.132)	(0.002)
Supervisor Dummy	0.969	1.512***	1.026	1.239**	1.070	1.125*	0.972	1.256***
	(0.710)	(0.000)	(0.781)	(0.019)	(0.260)	(0.067)	(0.659)	(0.000)
Manager Dummy	1.080	1.587***	0.954	1.470***	1.137**	1.236***	1.073	1.352***
	(0.398)	(0.000)	(0.661)	(0.000)	(0.017)	(0.000)	(0.221)	(0.000)
Log no of SMS	0.945***	0.943***	0.961**	0.951***	0.932***	0.951***	0.940***	0.949***
	(0.001)	(0.000)	(0.031)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log No of Contacts	0.992	0.922***	0.854***	0.954	0.962	0.896***	0.948*	0.938***
	(0.862)	(0.009)	(0.002)	(0.166)	(0.155)	(0.000)	(0.051)	(0.007)
Log no of Apps	0.642***	0.647***	0.742***	0.641***	0.662***	0.675***	0.679***	0.636***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Callog	0.882***	0.893***	0.926***	0.895***	0.896***	0.919***	0.896***	0.925***
	(0.000)	(0.000)	(0.006)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Finsavy App	0.868	0.850	0.647***	0.705***	0.613***	0.844*	0.697***	0.769***
	(0.279)	(0.130)	(0.001)	(0.001)	(0.000)	(0.060)	(0.000)	(0.000)
Socialconnect App	1.163	0.938	0.990	0.975	0.786**	1.374**	0.655***	1.342**
	(0.531)	(0.685)	(0.964)	(0.881)	(0.033)	(0.046)	(0.000)	(0.020)
Travel App	0.957	0.867**	1.056	1.056	0.986	0.959	1.027	0.982
	(0.573)	(0.046)	(0.545)	(0.465)	(0.779)	(0.481)	(0.623)	(0.710)
Mloan App	1.258***	1.325***	1.347***	1.240***	1.307***	1.143***	1.211***	1.318***
	(0.002)	(0.000)	(0.000)	(0.001)	(0.000)	(0.008)	(0.000)	(0.000)
Facebook status	1.157*	0.896*	0.900	1.029	1.018	0.963	0.978	0.989
	(0.063)	(0.070)	(0.231)	(0.660)	(0.723)	(0.475)	(0.666)	(0.785)
Linkedin status	0.331*	0.465***	0.651	0.542***	0.541***	0.690**	0.734*	0.599***
	(0.060)	(0.000)	(0.269)	(0.005)	(0.005)	(0.045)	(0.071)	(0.002)
IOS Dummy	0.153***	0.741**	0.720	0.774	0.455***	0.647**	0.484***	0.587***
	(0.000)	(0.041)	(0.226)	(0.153)	(0.000)	(0.012)	(0.000)	(0.000)
Observations	51,308	26,580	42,621	20,833	65,388	56,678	58,672	94,865
Pseudo R-squared	0.0317	0.0342	0.0239	0.0227	0.0290	0.0206	0.0243	0.0222
AUC	0.666	0.640	0.629	0.629	0.644	0.621	0.630	0.629

TABLE A3: Predicting defaults using machine learning - Logistics Regressions

This table reports out-of-sample test results using Logistic regressions to evaluate the default prediction performance of mobile and social footprint variables relative to traditional credit scores and other customer characteristics. Specifically, we compare three groups of variables a) CIBIL score b) Customer characteristics c) Mobile/Social Footprints. Panel A reports the results for all customers with CIBIL score. Panel B shows results for the subsample of borrowers with credit score in bottom 25% of the distribution. Panel C reports the results for the subsample of borrowers with no CIBIL score. For each of the sub-sample analyses, we report AUC, Accuracy, Precision, Recall and F1 score measures based on the out-of-sample tests.

Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)
Logistic regression	Only Cibil	0.64416156	0.58508376	0.6870797	0.38440304	0.49299072
	Only Mobile/Social Footprint	0.70435592	0.60575013	0.62686701	0.61891754	0.62286691
	Only Customer Characteristics	0.65466246	0.64916236	0.65942122	0.68368962	0.67133617
	Cibil + Customer Characteristics	0.706499	0.66549452	0.67340075	0.70130073	0.68706762
	Cibil + Mobile/Social Footprint	0.70537927	0.65012033	0.66410794	0.68057525	0.67224077
	Cibil + Mobile/Social Footprint + Customer Characteristics	0.73228729	0.66630061	0.61339264	0.99495419	0.7589129

TABLE A4: Summary statistics of deep social footprint based on call logs (customers with credit score)

This table reports summary statistics on the various call log variables (*Deep Social Footprint*) for the set of approved customers with a CIBIL Score. Columns 1-3 compares these characteristics for approved loans that were in default and those that were not in default. Panel A and B reports the statistics for the measures based on pre-approval call logs and post-approval call logs respectively. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

Panel A: Call log Metrics (Pre Loan Approval)			
Call Log Metric	Default (1)	Not Default (2)	Difference (3)
Past days: Per day Per person Avg No. of Incoming calls	1.61	1.53	0.08***
Past days: Per day Per person Avg No. of Outgoing calls	2.29	2.11	0.19***
Past days: Per day Per person Avg No. of Missed calls	1.67	1.52	0.15***
Past days: Per day Per person Avg Duration of Incoming calls	166.03	518.05	-352.03
Past days: Per day Per person Avg Duration of Outgoing calls	160.66	77.21	83.45**
Past days: Per day No. of persons called	16.08	14.39	1.69***
Past days: Per day Total No. of Incoming calls	12.06	10.51	1.55***
Past days: Per day Total No. of Outgoing calls	23.88	19.29	4.59***
Past days: Per day Total No. of Missed calls	8.52	6.3	2.23***
Past days: Per day Total Duration of Incoming calls	1124.52	1352.97	-228.45
Past days: Per day Total Duration of Outgoing calls	1523.25	489.38	1033.87***
Past days: HHI of No. of Incoming calls	199.86	134.09	65.77***
Past days: HHI of No. of Outgoing calls	181.2	140.54	40.67***
Past days: HHI of Total Duration of Incoming calls	433.05	329.46	103.59***
Past days: HHI of Total Duration of Outgoing calls	450.16	372.51	77.65***
Past days:HHI of No. of Missed calls	266.29	190.06	76.22***
Panel B: Call log Metrics (Post Loan Approval)			
First 15 days: Per day Per person Avg No. of Incoming calls	1.6	1.49	0.11***
First 15 days: Per day Per person Avg No. of Outgoing calls	2.27	2.04	0.24***
First 15 days: Per day Per person Avg No. of Missed calls	1.68	1.49	0.20***
First 15 days: Per day Per person Avg Duration of Incoming calls	155.87	157.19	-1.32
First 15 days: Per day Per person Avg Duration of Outgoing calls	145.8	-48.19	193.99
First 15 days: Per day No. of persons called	16.08	13.69	2.39***
First 15 days: Per day Total No. of Incoming calls	11.9	9.82	2.08***
First 15 days: Per day Total No. of Outgoing calls	24.02	17.72	6.30***
First 15 days: Per day Total No. of Missed calls	8.91	5.94	2.96***
First 15 days: Per day Total Duration of Incoming calls	1049.47	948.84	100.63***
First 15 days: Per day Total Duration of Outgoing calls	1420.66	-2167.68	3588.34
First 15 days: HHI of No. of Incoming calls	1150.62	906.41	244.21***
First 15 days: HHI of No. of Outgoing calls	1021.27	854.52	166.75***
First 15 days: HHI of Total Duration of Incoming calls	1833.4	1616.99	216.41***
First 15 days: HHI of Total Duration of Outgoing calls	1829.28	1701.19	128.10**
First 15 days:HHI of No. of Missed calls	1489.66	1281.65	208.00***
N	5,627	151,360	

TABLE A5: Summary statistics of deep social footprint based on call logs (customers without credit score)

This table reports summary statistics on the various call log variables (Deep Social Footprint) for the set of approved customers without a CIBIL Score. Columns 1-3 compares these characteristics for approved and disbursed loans that were in default and those that were not in default. Panel A and B reports the statistics for the measures based on pre-approval call logs and post-approval call logs respectively. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

Panel A: Call log Metrics (Pre Loan Approval)			
Call Log Metric	Default (1)	Not Default (2)	Difference (3)
Past days: Per day Per person Avg No. of Incoming calls	1.728	1.603	0.125***
Past days: Per day Per person Avg No. of Outgoing calls	2.506	2.285	0.221***
Past days: Per day Per person Avg No. of Missed calls	1.792	1.596	0.196***
Past days: Per day Per person Avg Duration of Incoming calls	184.812	178.677	6.135
Past days: Per day Per person Avg Duration of Outgoing calls	158.307	186.413	-28.107***
Past days: Per day No. of persons called	16.944	14.459	2.485***
Past days: Per day Total No. of Incoming calls	12.707	10.693	2.014***
Past days: Per day Total No. of Outgoing calls	28.668	21.72	6.948***
Past days: Per day Total No. of Missed calls	9.518	6.769	2.749***
Past days: Per day Total Duration of Incoming calls	1190.832	1058.944	131.888**
Past days: Per day Total Duration of Outgoing calls	1631.473	1545.418	86.056
Past days: HHI of No. of Incoming calls	359.218	188.842	170.375**
Past days: HHI of No. of Outgoing calls	244.706	207.947	36.759
Past days: HHI of Total Duration of Incoming calls	625.214	429.688	195.526**
Past days: HHI of Total Duration of Outgoing calls	644.53	473.003	171.527*
Past days:HHI of No. of Missed calls	410.622	272.132	138.491*
Panel B: Call log Metrics (Post Loan Approval)			
First 15 days: Per day Per person Avg No. of Incoming calls	1.695	1.588	0.107**
First 15 days: Per day Per person Avg No. of Outgoing calls	2.452	2.25	0.202**
First 15 days: Per day Per person Avg No. of Missed calls	1.871	1.581	0.290**
First 15 days: Per day Per person Avg Duration of Incoming calls	160.995	163.796	-2.8
First 15 days: Per day Per person Avg Duration of Outgoing calls	136	172.028	-36.028***
First 15 days: Per day No. of persons called	16.014	13.815	2.199***
First 15 days: Per day Total No. of Incoming calls	12.304	10.29	2.014***
First 15 days: Per day Total No. of Outgoing calls	27.061	20.159	6.902***
First 15 days: Per day Total No. of Missed calls	9.248	6.42	2.828***
First 15 days: Per day Total Duration of Incoming calls	1074.002	978.093	95.909
First 15 days: Per day Total Duration of Outgoing calls	1465.966	1388.045	77.921
First 15 days: HHI of No. of Incoming calls	1226.355	849.715	376.640**
First 15 days: HHI of No. of Outgoing calls	913.671	820.746	92.925
First 15 days: HHI of Total Duration of Incoming calls	1815.054	1530.397	284.656
First 15 days: HHI of Total Duration of Outgoing calls	1582.718	1633.848	-51.13
First 15 days:HHI of No. of Missed calls	1476.673	1248.335	228.339
N	215	3,149	

TABLE A6: Quarterly prediction performance

This table reports results for different machine learning models for each quarter to evaluate the default prediction performance for various feature groups. Columns 1 through 5 report results for random forest model and columns 6 through 10 report results for XGBoost model. In each model, the previous quarter data is used as training and validation sample and the following quarter is used as the testing sample.For each we report AUC, Accuracy, Precision, Recall and F1 score measures based on the testing sample.

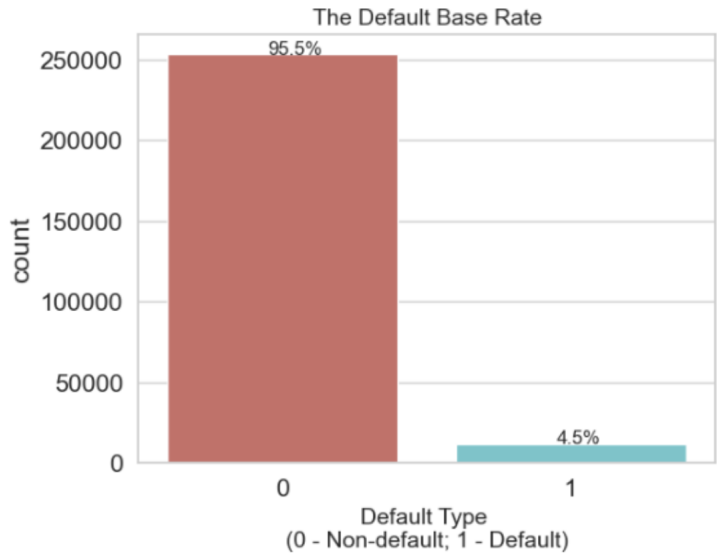
		Random Forest					Xgboost				
Model	Feature Groups	AUC (1)	Accuracy (2)	Precision (3)	Recall (4)	F1 (5)	AUC (6)	Accuracy (7)	Precision (8)	Recall (9)	F1 (10)
2017Q1	Only Cibil	0.575	0.603	0.115	0.523	0.188	0.670	0.696	0.152	0.534	0.236
2017Q2	Only Cibil	0.556	0.539	0.049	0.571	0.090	0.599	0.704	0.058	0.426	0.103
2017Q3	Only Cibil	0.503	0.469	0.043	0.539	0.080	0.515	0.614	0.047	0.418	0.085
2017Q4	Only Cibil	0.516	0.408	0.043	0.629	0.080	0.553	0.547	0.048	0.534	0.088
2018Q1	Only Cibil	0.532	0.439	0.024	0.630	0.046	0.547	0.610	0.026	0.465	0.049
2018Q2	Only Cibil	0.537	0.377	0.018	0.703	0.034	0.569	0.393	0.018	0.698	0.035
2018Q3	Only Cibil	0.489	0.752	0.027	0.211	0.048	0.456	0.798	0.027	0.168	0.047
2017Q1	Only Mobile/Social Footprint	0.836	0.865	0.337	0.547	0.417	0.885	0.853	0.350	0.782	0.484
2017Q2	Only Mobile/Social Footprint	0.668	0.903	0.094	0.165	0.119	0.691	0.901	0.104	0.195	0.136
2017Q3	Only Mobile/Social Footprint	0.662	0.836	0.082	0.279	0.127	0.664	0.673	0.070	0.542	0.124
2017Q4	Only Mobile/Social Footprint	0.659	0.812	0.073	0.306	0.117	0.699	0.563	0.065	0.730	0.120
2018Q1	Only Mobile/Social Footprint	0.609	0.832	0.036	0.262	0.063	0.670	0.621	0.034	0.604	0.064
2018Q2	Only Mobile/Social Footprint	0.602	0.887	0.028	0.181	0.048	0.641	0.619	0.024	0.580	0.046
2018Q3	Only Mobile/Social Footprint	0.583	0.898	0.043	0.115	0.063	0.600	0.595	0.038	0.525	0.071
2017Q1	Only Customer Characteristics	0.574	0.705	0.112	0.339	0.168	0.559	0.632	0.109	0.442	0.175
2017Q2	Only Customer Characteristics	0.557	0.684	0.047	0.365	0.084	0.584	0.562	0.051	0.571	0.094
2017Q3	Only Customer Characteristics	0.577	0.765	0.052	0.260	0.086	0.566	0.658	0.053	0.417	0.094
2017Q4	Only Customer Characteristics	0.569	0.759	0.050	0.271	0.084	0.565	0.596	0.049	0.479	0.088
2018Q1	Only Customer Characteristics	0.563	0.774	0.028	0.276	0.050	0.576	0.627	0.028	0.476	0.052
2018Q2	Only Customer Characteristics	0.540	0.786	0.017	0.221	0.031	0.583	0.611	0.021	0.506	0.039
2018Q3	Only Customer Characteristics	0.550	0.827	0.034	0.176	0.057	0.542	0.636	0.034	0.411	0.063
2017Q1	Cibil + Customer Characteristics	0.698	0.740	0.169	0.497	0.253	0.677	0.675	0.156	0.610	0.249
2017Q2	Cibil + Customer Characteristics	0.619	0.745	0.063	0.389	0.108	0.638	0.688	0.066	0.523	0.118
2017Q3	Cibil + Customer Characteristics	0.620	0.836	0.062	0.202	0.095	0.591	0.710	0.060	0.393	0.104
2017Q4	Cibil + Customer Characteristics	0.597	0.803	0.054	0.233	0.088	0.567	0.598	0.049	0.484	0.090
2018Q1	Cibil + Customer Characteristics	0.584	0.841	0.028	0.192	0.049	0.584	0.670	0.029	0.443	0.055
2018Q2	Cibil + Customer Characteristics	0.542	0.850	0.021	0.190	0.038	0.579	0.580	0.020	0.535	0.039
2018Q3	Cibil + Customer Characteristics	0.566	0.852	0.037	0.159	0.060	0.557	0.521	0.035	0.575	0.066
2017Q1	Cibil + Mobile/Social Footprint	0.879	0.866	0.350	0.605	0.443	0.899	0.851	0.347	0.786	0.482
2017Q2	Cibil + Mobile/Social Footprint	0.681	0.922	0.125	0.158	0.139	0.689	0.922	0.128	0.165	0.144
2017Q3	Cibil + Mobile/Social Footprint	0.687	0.860	0.091	0.255	0.135	0.675	0.700	0.076	0.537	0.133
2018Q1	Cibil + Mobile/Social Footprint	0.665	0.873	0.047	0.253	0.079	0.688	0.694	0.039	0.554	0.072
2018Q2	Cibil + Mobile/Social Footprint	0.621	0.917	0.027	0.125	0.045	0.635	0.669	0.024	0.509	0.046
2018Q3	Cibil + Mobile/Social Footprint	0.601	0.916	0.049	0.101	0.066	0.605	0.627	0.040	0.510	0.075
2017Q1	Cibil + Mobile/Social Footprint + Customer Characteristics	0.881	0.868	0.351	0.592	0.441	0.887	0.844	0.329	0.741	0.456
2017Q2	Cibil + Mobile/Social Footprint + Customer Characteristics	0.696	0.924	0.120	0.143	0.130	0.692	0.915	0.112	0.165	0.134
2017Q3	Cibil + Mobile/Social Footprint + Customer Characteristics	0.700	0.875	0.106	0.259	0.150	0.673	0.714	0.075	0.500	0.130
2017Q4	Cibil + Mobile/Social Footprint + Customer Characteristics	0.696	0.844	0.083	0.282	0.129	0.678	0.635	0.067	0.615	0.121
2018Q1	Cibil + Mobile/Social Footprint + Customer Characteristics	0.675	0.881	0.052	0.264	0.087	0.694	0.688	0.041	0.596	0.076
2018Q2	Cibil + Mobile/Social Footprint + Customer Characteristics	0.631	0.937	0.031	0.099	0.047	0.645	0.699	0.025	0.486	0.048
2018Q3	Cibil + Mobile/Social Footprint + Customer Characteristics	0.610	0.934	0.059	0.083	0.069	0.611	0.659	0.041	0.473	0.076

TABLE A7: Summary statistics on approval rate and default rate for different subsamples

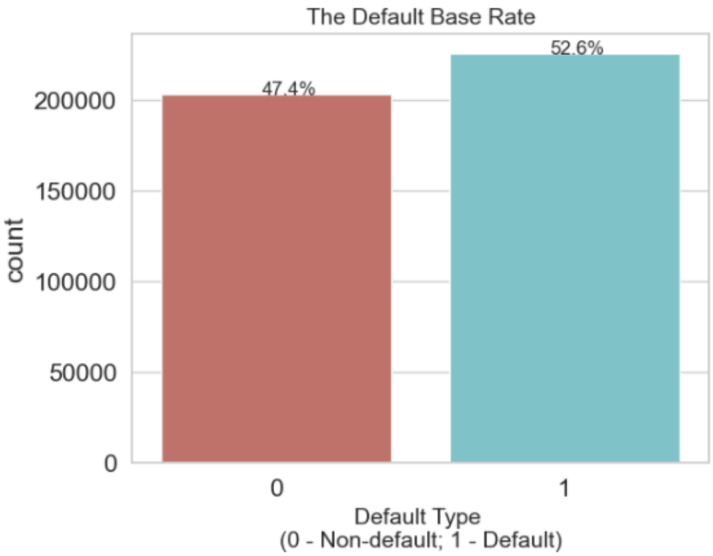
This table reports summary statistics on the approval rate and default rate of customers without a credit score for subsamples based on income, education, and regional level of financial inclusion. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

Without CIBIL			
Panel A: Income Level			
	(1)	(2)	(3)
	High	Low	Difference
Default rate	0.091	0.055	-0.035***
Approval rate	0.454	0.293	-0.161***
Panel B: Education Level			
	High	Low	Difference
Default rate	0.076	0.075	-0.001
Approval rate	0.497	0.187	-0.310***
Panel C: Fraction of Households Without Bank Accounts			
	High	Low	Difference
Default rate	0.081	0.073	-0.008***
Approval rate	0.342	0.385	0.043***
Panel D: Financial Exclusion Index			
	High	Low	Difference
Default rate	0.082	0.075	-0.007***
Approval rate	0.300	0.463	0.163***

Figure A1: Balancing of Data (representative graph)



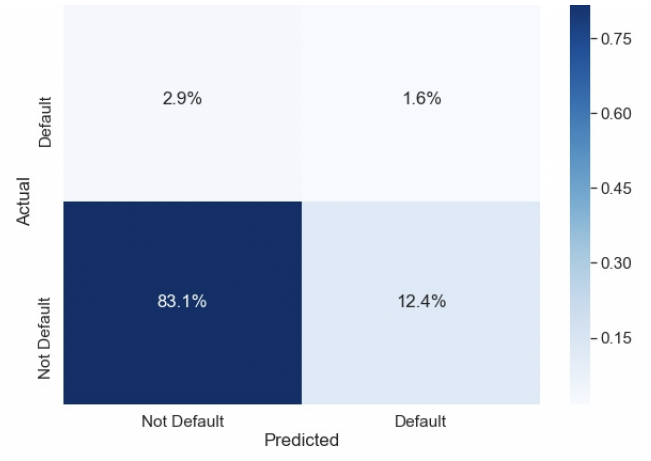
(a) Before Data Balancing



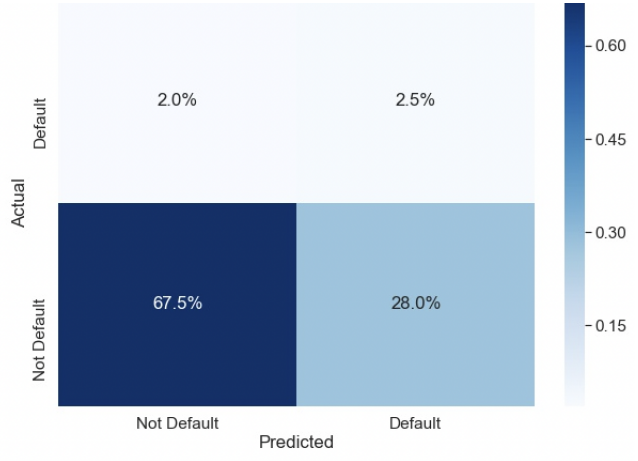
(b) % After Data Balancing

Figure A2: Confusion matrix

The following figures represents the confusion matrix from the random forest (panel A) model with all features from the testing sample. The X-axis represents the predicted loan outcome and the Y-axis represents the actual loan outcome. The bottom left hand corner cell, for example, represents the percentage of good loans (loans which did not default) in the testing sample which were predicted to not default. Other cells have similar interpretation.

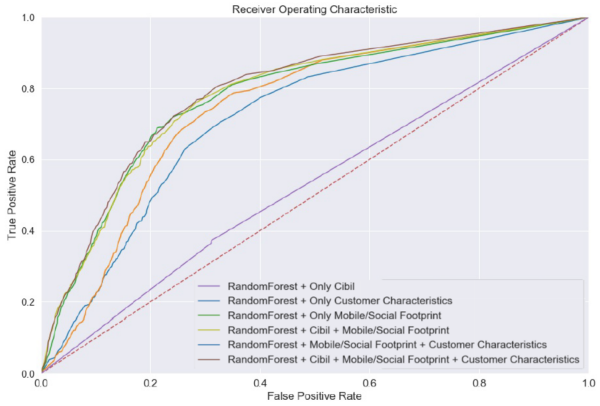


(a) Random forest

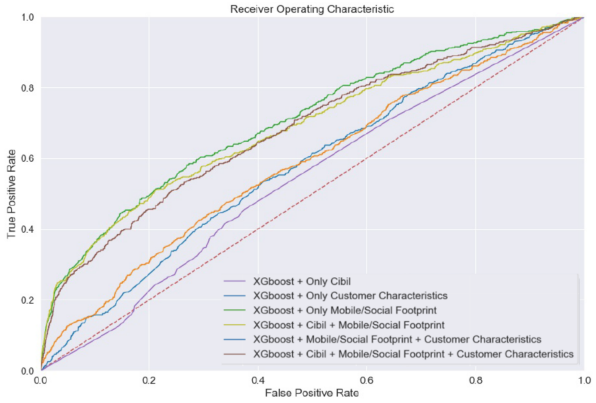


(b) XGboost

Figure A3: AUC plots for machine learning models (Customers with credit score in Bottom 25%)
This figure plots the AUC curves for default prediction based on the two machine learning models for the sample of customers with a credit bureau score in Bottom 25%. Panel A reports the figure based on Random forest and Panel B reports based on Xgboost.



(a) Random forest



(b) XGboost

Figure A4: Loans originated by quarter
The following graph represents the number of loans that originated each quarter in our sample period.

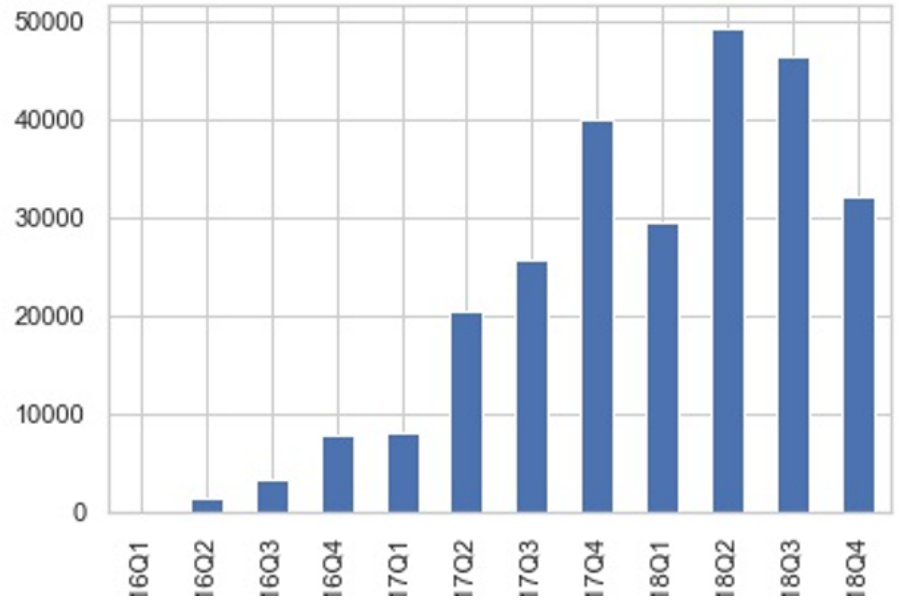
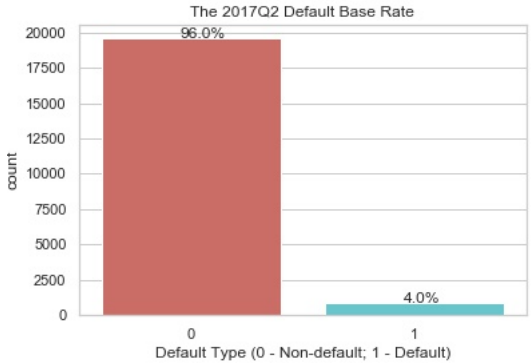
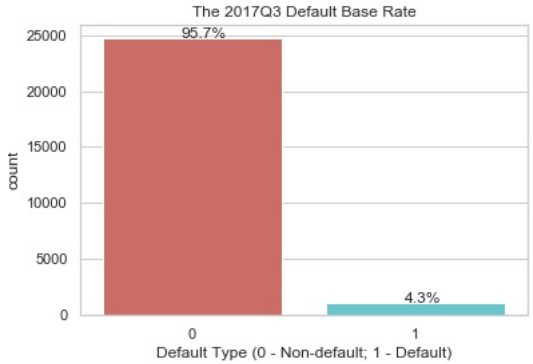


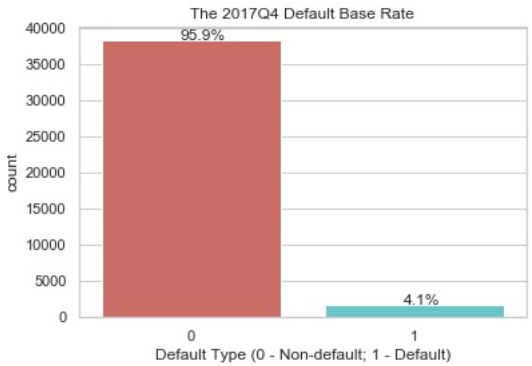
Figure A5: Quarterly actual default Rate
The following set of graphs represents the quarterly actual default rates in our sample



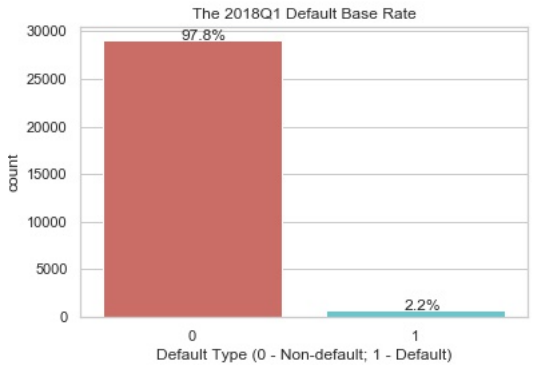
(a) 2017Q2



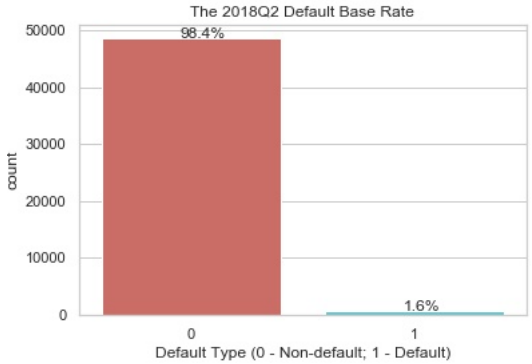
(b) 2017Q3



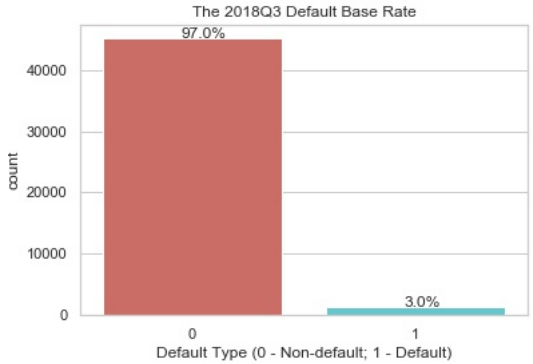
(c) 2017Q4



(d) 2018Q1



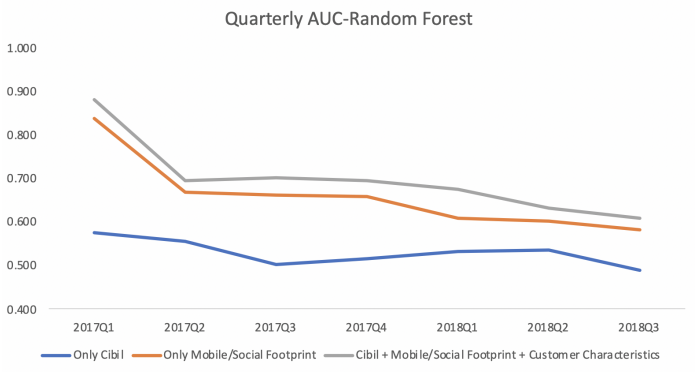
(e) 2018Q2

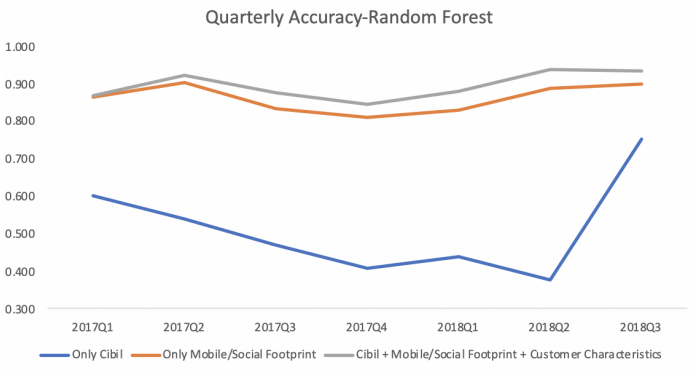


(f) 2018Q3

Figure A6: Quarterly prediction performance

These graphs represents the time series pattern of predictive performances. The first graphs represents the AUC scores from the random forest model for each quarter from our quarter ahead prediction. The second graph represents the accuracy scores from the random forest model for each quarter of the testing sample in our quarter ahead prediction. The X-axis represents quarter and the Y-axis represents prediction performance scores. In these models the previous quarter data is used for training and validation sample and the next quarter data is used as the testing sample. The prediction performance from the testing sample for various models (only CIIBIL, only digital and all variables) are represented below.





Appendix B: Variable Definitions

TABLE B1: Variable Definitions

This table provides the description of the variables used in our baseline analysis.

SNo.	Variable name	Variable definition
		Credit Score
1	Log of cibil	Log of Credit bureau score
		Customer Characteristics
2	Log of Salary	Log of customer's salary
3	Log Age	Log of customer's age.
4	High School Dummy	Dummy takes value 1 if customer's highest qualification is High School.
5	College Dummy	Dummy takes value 1 if customer's highest qualification is College.
6	Supervisor Dummy	Dummy takes value 1 if customer's designation falls in the supervisor category.
7	Manager Dummy	Dummy takes value 1 if customer's designation falls in the manager category.
		Loan Characteristics
8	Log Loan Amount	Log of Loan Amount of the loan.
9	Travel.purpose cashe	Dummy takes value 1 if purpose of loan is travel.
10	EMI.purpose cashe	Dummy takes value 1 if purpose of loan is to pay EMI.
11	Loan repayment.purpose cashe	Dummy takes 1 if purpose of loan is to pay another loan.
12	Other purpose.purpose cashe	Dummy takes 1 if purpose of loan is other than travel, EMI, loan repayment and medical.
		Alternative Data: Mobile and Social Footprint Variables
13	Log no of SMS	Log of Total No. of SMS.
14	Log no of Contacts	Log of No. of people in contact list.
15	Log no of Apps	Log of no. of applications in phone.
16	Log Callog	Log of Total No. of calls.
17	Dating App	Dummy takes 1 if customer has a dating app.
18	Finsavy App	Dummy takes 1 if customer has a financial services app (stocks, banking, payment and wallet).
19	Socialconnect App	Dummy takes 1 if customer has a social connect app (messaging app, video streaming app, music streaming app, social network app, dating app, video call app).
20	Travel App	Dummy takes 1 if customer has a Travel app.
21	Mloan App	Dummy takes 1 if customer has another loan app.
22	Facebook Status	Dummy takes 1 if customer logged into Cashe app using Facebook.
23	Linkedin Status	Dummy takes 1 if customer logged into Cashe app using Linkedin.
24	IOS Dummy	Dummy takes 1 if customer has an Apple phone.
		Alternative Data: Deep Social Footprint Variables (based on Call Logs)

25	Per day Per person Avg No. of Incoming calls	No. of incoming calls received from a person on average in a day.
26	Per day Per person Avg No. of Outgoing calls	No. of outgoing calls made to a person on average in a day.
27	Per day Per person Avg No. of Missed calls	No. of missed calls received from a person on average in a day.
28	Per day Per person Avg Duration of Incoming calls	Duration of incoming calls with a person on average in a day.
29	Per day Per person Avg Duration of Outgoing calls	Duration of outgoing calls with a person on average in a day.
30	Per day No. of persons called	No. of persons called (includes incoming, outgoing and missed) in a day.
31	Log of Per day Total Duration of Incoming calls	Total Talk time of incoming calls in a day.
32	Per day Total No. of Incoming calls	No. of incoming calls in a day.
33	Per day Total No. of Outgoing calls	No. of outgoing calls in a day.
34	Per day Total Duration of Outgoing calls	Total Talk time of outgoing calls in a day.
35	Per day Total No. of Missed calls	No. of missed calls in a day.
36	HHI of No. of Incoming calls	Herfindahl-Hirschman index of incoming calls. To compute this measure, we first calculate the no. of calls received from a person for every day (for a customer). We then take average across all days to get the no. of calls received from the person per day. We then assign share of calls to every person and compute HHI for the customer.
37	HHI of No. of Outgoing calls	Herfindahl-Hirschman index of outgoing calls. To compute this measure, we first calculate the no. of calls made to a person for every day (for a customer). We then take average across all days to get the no. of calls made to the person per day. We then assign share of calls to every person and compute HHI for the customer.
38	HHI of Total Duration of Incoming calls	Herfindahl-Hirschman index of duration of incoming calls. To compute this measure, we first calculate the duration of calls received from a person for every day (for a customer). We then take average across all days to get duration of calls per day. We then assign share of durations to every person and compute HHI for the customer.
39	HHI of Total Duration of Outgoing calls	Herfindahl-Hirschman index of duration of outgoing calls. To compute this measure, we first calculate the duration of calls made to a person for every day (for a customer). We then take average across all days to get duration of calls per day. We then assign share of durations to every person and compute HHI for the customer.
40	HHI of No. of Missed calls	Herfindahl-Hirschman index of missed calls. To compute this measure, we first calculate the no. of missed calls received from a person for every day (for a customer). We then take average across all days to get the no. of missed calls received from the person per day. We then assign share of missed calls to every person and compute HHI for the customer.

TABLE B2: Deep Financial Variable Definitions

The following deep financial variables were populated in the data fields and were used in the ML estimations.

SNo.	Variable name	Variable definition
1	debits_to_credits_ratio	Ratio of total debit to total credit in 3-month window before start of loan.
2	transactions_number	No of transactions in 3-month window before start of loan.
3	log_exp_to_inc_ratio	Log of ratio of Expense to Income for the 3-month window before start of loan.
4	avg_2_month_depreciation	Increase in account balance between account snapshots in the 2 months before start of loan. Data consists of snapshots spaced out at 10 day gaps. Formula: (rate_of_dep_stage5 + rate_of_dep_stage6 + rate_of_dep_stage7 + rate_of_dep_stage8)
5	transactions_minbal	Minimum account balance in the 3-month period.
6	transactions_maxbal	Maximum account balance in the 3-month period.
7	transactions_averag	Average account balance.
8	transactions_mindeb	Minimum debit amount in the 3-month period.
9	transactions_maxdeb	Maximum debit amount in the 3-month period.
10	ransactions_mindep	Minimum deposit amount in the 3-month period.
11	transactions_maxdep	Maximum deposit amount in the 3-month period.
12	netsavings_highest	Highest saving in the 3-month period.
13	netsavings_lowest	Lowest saving in the 3-month period.
14	netsavings_lastmonth	Net saving in the last month
15	avg_income	Average income in the 3-month period.
16	avg_expense	Average expense in the 3-month period.
17	avg_surplus	Average surplus in the 3-month period.
18	exp_to_inc_ratio	Expense to income ratio for the 3-month period.
19	avg_inc_exc_inv	Monthly average income excluding income from investments in the 3-month period.
20	avg_exp_exc_inv	Monthly average expense excluding expense put into investments in the 3-month period.
21	avg_surplus_exc_inv	Monthly average surplus excluding surplus from investments in the 3-month period.
22	avg_inc_exc_transfers	Monthly average income excluding income from transfers in the 3-month period.
23	avg_exp_exc_transfers	Monthly average expense excluding expense from transfers in the 3-month period.
24	avg_surplus_exc_transfers	Monthly average surplus excluding surplus from transfers in the 3-month period.
25	avg_inc_exc_both	Monthly average income excluding income from investments and transfers in the 3-month period.
26	avg_exp_exc_both	Monthly average expense excluding expense in investments and transfers in the 3-month period.
27	avg_surplus_exc_both	Monthly average surplus excluding surplus from transfers and investments in the 3-month period.
28	expens_debits_highest	Highest debit amount in the 3-month period.
29	expens_debits_lowest	Lowest debit amount in the 3-month period.
30	expens_debits_lastmonth	Total debit amount of the last month.

31	income_credits_highest	Highest credit amount in the 3-month period.
32	income_credits_lowest	Lowest credit amount in the 3-month period.
33	income_credits_lastmonth	Total credit amount of the last month.
34	inv_inflow_highest	Highest investment inflow in the 3-month period.
35	inv_inflow_lowest	Lowest investment inflow in the 3-month period.
36	inv_inflow_lastmonth	Total investment inflow amount of the last month.
37	salary_month_1	Salary in the 1st month.
38	bal_on_payday_month_1	Account balance at end of salary pay date (usually 1st of the month) of the 1st month.
39	bal_10days_bef_month_1	Account balance 10 days before salary pay date (usually 1st of the month) of the 1st month.
40	bal_20daysbef_month_1	Account balance 20 days before salary pay date (usually 1st of the month) of the 1st month.
41	salary_month_2	Salary in the 2nd month.
42	bal_on_payday_month_2	Account balance at end of salary pay date (usually 1st of the month) of the 2nd month.
43	bal_10daysbef_month_2	Account balance 10 days before salary pay date (usually 1st of the month) of the 2nd month.
44	bal_20daysbef_month_2	Account balance 20 days before salary pay date (usually 1st of the month) of the 2nd month.
45	salary_month_3	Salary in the 3rd month.
46	bal_on_payday_month_3	Account balance at end of salary pay date (usually 1st of the month) of the 3rd month.
47	bal_10daysbef_month_3	Account balance 10 days before salary pay date (usually 1st of the month) of the 3rd month.
48	bal_20daysbef_month_3	Account balance 20 days before salary pay date (usually 1st of the month) of the 3rd month.
49	investment_in	Average income - Average income excluding investment
50	investment_out	Average expense - Average expense excluding investment
51	net_loss	investment_in - investment_out
52	rate_of_dep_stage1	Bal_on_payday_month_1 - salary_month_1
53	rate_of_dep_stage2	Bal_20daysbef_month_2 - Bal_on_payday_month_1
54	rate_of_dep_stage3	Bal_10daysbef_month_2 - Bal_20daysbef_month_2
55	rate_of_dep_stage4	Salary_month_2 - Bal_10daysbef_month_2
56	rate_of_dep_stage5	Bal_on_payday_month_2 - Salary_month_2
57	rate_of_dep_stage6	Bal_20daysbef_month_3 - Bal_on_payday_month_2
58	rate_of_dep_stage7	Bal_10daysbef_month_3 - Bal_20daysbef_month_3
59	rate_of_dep_stage8	Salary_month_3 - Bal_10daysbef_month_3
60	balonpaydayaverage	Average of Balance at end of pay day.
61	bal10daysbefaverage	Monthly average of Balance 10 days before salary pay date.
62	bal20daysbefaverage	Monthly average of Balance 20 days before salary pay date
63	salaryaverage	Monthly salary average.
64	credits_sum	Sum of credits.
65	debits_sum	Sum of debits.

66	exptoincratioexcludinginvestment	Expense to income ratio after excluding investments.
67	exptoincratioexcludingtransfers	Expense to income ratio after excluding transfers.
68	exptoincratioexcludingboth	Expense to income ratio after excluding transfers and investments.
69	investmentinflow_categoryhighest	Highest category investment inflow- Fixed Deposit, MF Redemption, Interest, etc
70	investmentinflow_categorylowesti	Lowest category investment inflow- Fixed Deposit, MF Redemption, Interest, etc
71	averagemonth_1	Average account balance during the 1st month.
72	averagemonth_2	Average account balance during the 2nd month.
73	averagemonth_3	Average account balance during the 3rd month.