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

This Version: February 2020

Abstract

Using a unique and proprietary loan-level data from a large Fintech lending firm in India, we analyze whether unstructured data pertaining to a consumer's social and mobile footprint can act as a substitute for traditional credit bureau scores. We find that the mobile footprint of an individual outperforms the credit score in predicting loan approvals and defaults. Importantly, including measures of borrower's "deep social footprints" based on call logs significantly improves default prediction. We use machine learning-based prediction counterfactual analysis to predict the loan outcome for borrowers who were denied credit, perhaps due to the lack of traditional credit scores. We show that using alternate credit scoring using the mobile and social footprints can expand credit as well as reduce the overall default rate. Our study has implications for expanding access to credit to those who do not have a credit history but who leave a large trace of unstructured information on their mobile phones that can be used to predict loan outcomes.

JEL codes: G20, G21, G29

Keywords: Fintech, Big data, Credit scores, Financial inclusion, Lending, Machine Learning, Mobile footprint, Prediction Counterfactual, Social footprint, Social capital

-

1 Introduction

A recent survey in the US showed that almost half of the millennials in the US feel that their credit score is holding them back.[1] Younger people suffer from shorter credit history and hence are often denied credit by traditional financial institutions or are charged prohibitively high interest rates, which limits their access to credit.[2] This, in turn, exacerbates the evaluation of their creditworthiness by limiting their ability to build a good credit history. Many such individuals may actually be 'good borrowers' if their 'creditworthiness' could be evaluated using alternate data. The problem of lack of credit history for the millennials is a world-wide phenomenon and especially true for developing countries. For example, according to a recent industry report, 156 million Indians who comprise the 'urban mass' representing an annual income of USD 3000 and above have the potential of mass adoption of consumer credit. Of this 'urban mass', approximately 129 million have been mostly deprived of credit due to a lack of credit history.[3] This led to the quest for alternative data for credit scoring for the millennials.

While millions across India and the world have never obtained a bank loan, they are active mobile phone users who shop online, and have a good social media presence.[4] These traces of unstructured data that individuals leave through their online behavior and mobile phone usage can potentially be used to predict their loan behavior. Consistent with this idea, a plethora of fintech firms have mushroomed all around the world that aim to service such customers by leveraging unstructured data and big data analytics to predict their default behavior. However, thus far, there is limited evidence on whether or not "mobile footprint" of an individual can substitute for traditional credit bureau scores. To the best of our knowledge, ours is the first study to examine whether an individual's online behavior captured from their mobile phones can be used to predict their likelihood of default. 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 (Chava, Paradkar & Zhang (2017), Berg, Burg, Gombović & Puri (2019), Fuster, Goldsmith-Pinkham, Ramadorai & Walther (2018), D'Acunto, Prabhala & Rossi (2019), Rossi & Utkus (2019), D'Acunto, Rauter, Scheuch & Weber (2019), Balyuk (2019)).

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

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

[2] 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 are suffering the consequences of bad credit. In fact, 24 percent of those surveyed said they never learned how to build good credit in the first place, and 15 percent reported that their level of debt is unmanageable, with 1 in 5 admitting that they don't have control over their finances.

[3] Financial Expressed News Article [Accessed on 14th March, 2019]

[4] 97% of users access internet in India through mobile phones. See Kantar-IMRB (2018)

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, rather analyze the association between the mobile footprint and credit worthiness of individuals.[5]

A natural follow-up question is whether we can use the social and mobile footprint variables to come up with an alternate credit scores for borrowers who do not have traditional credit bureau scores. How many of the borrowers who are denied loans could potentially be creditworthy if their creditworthiness could be evaluated using information from their social and mobile footprints? Importantly, how would granting loans to such borrowers affect the overall default rate of the lender's loan portfolio? These counterfactual questions have significant policy implications. Importantly these questions pertain to default prediction and are not causal in nature. 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 policy counterfactual questions posed above.[6]

We obtain the universe of loan applications made to one of the largest fintech lender in India, between the period of February 2016 to November 2018. Unlike prior studies, we also have access to loan applications that were denied allowing us to examine the determinants of loan approval. Out of about 417,000 loan applications in our sample, about 272,000 were approved while rest were denied. The lender is a stereotypical mobile-only fintech lending platform targeted towards meeting the short-term credit needs of the salaried millennial. It grants loans ranging from a minimum of ₹10,000 to a maximum of ₹200,000 for 15, 30, 90, 120, and maximum loan duration of 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 slip. The potential borrower authorizes the lender to use its digital mobile presence for the evaluation of her creditworthiness and research. They also provide the fintech lender data on their traditional credit score: CIBIL−Transunion credit score (if available), education, and job designation. Importantly for our study, the lender also collects detailed digital information from the individuals' mobile phone such as the mode of login (for example, Facebook and Linkedin), the various applications installed, number of calls, number of contacts on phone, number of social connections, and the kind of mobile operating system such as IOS and Android. We have access to detailed anonymized data on the kind of mobile applications[7] that an individual uses that we club into 6 broad categories: *Sales apps* which includes applications for e-commerce such as Amazon, Flipkart, Snapdeal among others, *Social Network apps* such as Whatsapp, Twitter, Messenger services, *Financial Apps* such as Mobile banking and stock trading applications, *Travel apps* such as Airbnb, Tripadvisor, and MakeMyTrip, *Mloan app* which includes other mobile-based lending platforms, and *Dating apps* such as Tinder.

In addition, we have detailed information on call logs of individuals. For ease of reference,

[5] Berg et al. (2019) also examine the discriminatory ability of an individuals' online presence for default behavior using data from an e-commerce website. We discuss our marginal contribution relative to Berg et al. (2019) in detail on page 7.

[6] See also Athey & Imbens (2019)

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

we categorize these digital information captured from an individual's mobile phone into three categories: (1) "social footprint" which refers to the presence of social apps, the preferred social network for logging on to the fintech lender's app, number of contacts, number of calls/sms, whether the customer was acquired through a referral (2) "deep social footprint" which captures information obtained from call logs pattern, and (3) broader "mobile footprint" which refers to the kind of applications installed, the number of applications, and the type of mobile operating system.[8]

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 individuals digital presence captured from their mobile phones can be used to predict loan approvals and defaults.

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

Next, we examine the ability of mobile and social footprint variables in predicting defaults. Here, we rely on both the economic and statistical significance of individual explanatory variables as well as Area Under the Curve (AUC) - an easy and commonly used measure of the predictive power of credit scores (Iyer, Khwaja, Luttmer & Shue (2015)). We first note that the AUC of the model using only the credit score for predicting defaults is 59%. The AUC of credit score in our sample, is significantly different from flipping a coin (AUC of 50%) is lower than 62% reported by Iyer et al. (2015) based on a sample of loans from peer to peer lending platform, "Propser.com", and comparable to the AUC of 59.8% using U.S. credit scores from Lending Club reported in Berg et al. (2019).

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

The AUC of a model that relies exclusively on the mobile and social footprint to predict defaults at 60.4% is approximately 2% more than the AUC of the model using only the credit score. Our

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

results suggest that mobile and social footprint variables may be capturing hard to quantify aspects of individuals' behavior, which has implications for the likelihood of default. For instance, customers without a financial application installed on their phones are about one and a half times more likely to default relative to those who have such an application installed. This is consistent with the idea that installing financial applications may proxy for the financial sophistication of a customer. In contrast, those with a dating application (any other social network app) are 30% (38%) more likely to default. Interestingly, customers who log in to the application via Linked or Facebook are 24% and 9% more likely to default respectively relative to those who log-in via other means.

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

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

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

Our analysis of default prediction thus far was based on measures of mobile and social footprint such as the nature of apps installed, the number of apps installed, the number of calls, etc., to predict defaults. We now seek to understand whether we can use "deep social footprint" of customers from their call logs to improve upon the default prediction. Using various proxies based on the frequency and duration of daily incoming, outgoing, and missed calls that attempt to capture the breadth and strength of an individual's social capital, we find that these measures are strongly correlated with the likelihood of default.[9] Specifically, we find that defaulters are more likely to have their call concentrated over a smaller number of individuals. Consistent with this, defaulters seem to have stronger ties with individuals in their contact list as measured by the average number of calls and

[9]The underlying idea behind these tests builds on prior work which documents that call log patterns can be used to infer an individual's social capital (Singh & Ghosh (2017), Wiese, Min, Hong & Zimmerman (2014)), which is an important predictor of loan defaults (Karlan (2005)).

duration of calls per person. Delinquent customers have a smaller duration of incoming calls but have a higher duration of outgoing calls, which along with their frequency of missed calls, suggests that defaulters are less likely to respond to calls initiated by others.

Most importantly, the AUC of a model that includes call log measures along with other mobile and social footprint variables is 66%, an 8% improvement over the model with credit score alone. This is better than the 5.7 percentage points AUC improvement reported in Iyer et al. (2015) who compare the AUC using the Experian credit score to the AUC in a setting where, in addition to the credit score, lenders have access to a large set of borrower financial information as well and comparable to the improvement in the AUC by +8.8 percentage points reported by Berg, Puri & Rocholl (2017) in a consumer loan sample of a large German bank in a setting where, in addition to the credit score, lenders have access to account data, as well as socio-demographic data and income information.

We also have access to the detailed financial reports for a random subset of the borrowers in our sample. The report provides detailed financial information like the borrower's spending and income patterns, number of transactions, other borrowing information, etc; over the last three months, which we collectively refer to as 'deep financial information'.[10] The fintech lender accessed these reports during the loan application process. We find that for the subset of the borrowers for whom we have access to this financial report, the 'deep' mobile footprint has greater predictive power for borrower's credit risk relative to the 'deep' financial information.

We next verify the predictive performance of social and mobile footprint variables using different machine learning algorithms. The problem at hand is to train the algorithms on the sample data to predict defaults "out-of-sample". Standard estimation approaches like OLS, where we use all the data to make in sample prediction, is not well suited for such analysis. The in-sample estimation approaches works on being unbiased (having the bias close to zero), thus leaving only the variance to be optimized to minimize the out of sample prediction error. Thus, OLS does not offer joint optimality of bias and variance. Machine learning techniques are particularly useful here, which minimizes the mean squares error of the prediction by a joint minimization procedure cognizant of the bias-variance trade-off. Using various machine learning algorithms like logistic regression, random forest, and XGBoost (See Athey & Imbens (2019)), we first show that the mobile and social footprints have significantly higher predictive power for both borrowers with and without traditional credit scores.

Next, we run a horse race between 'deep' financial information and 'deep' social footprint variables based on call logs to see if the deep mobile footprint has incremental predictive power beyond what is captured in the borrower's income and spending patterns. This is important as it can inform us regarding the nature of data that should be collected to build alternate credit scores. First, we find that both 'deep' financial information and 'deep' mobile footprint variables have significant discriminatory ability in predicting defaults. Second, the information content of deep mobile footprint complements and exceeds the 'deep' financial variables. Specifically, the out of sample AUC of the models which includes only deep mobile footprint variable (deep financial information)

[10] The list of the variables available from the detailed credit report are provided in Table C2 of appendix C.

is 74% (59%). Overall, we find that digital mobile footprint has significant ability in predicting defaults and the information content of these variables complements rather than substitutes for both the credit bureau score and detailed financial information regarding a customer's income and expenses.

The prediction of default risk for borrowers without a traditional credit score is useful and can be used to ask counterfactual questions such as: how many denied borrowers (perhaps due to lack of traditional credit score) would have been approved had we used the social and mobile footprint based alternate credit scores? What would have been the impact on default if we had used these scores? These counterfactual prediction policy questions are not causal in nature, as our objective is to find the best predictor of default risk of the borrower. We follow the methods outlined in Athey (2017) and Kleinberg et al. (2015) to address these prediction counterfactual questions. Using our methodology, we find that even if we use a low predicted default threshold of 10% (1%) for the probability of default relative to the in-sample default rate of 12% for approving loans, about 42% (22%) borrowers who were denied credit would have been granted loans.

Overall, our study documents that mobile and social footprint variables have significant discriminatory power in both loan approvals and default prediction. Importantly, with the use of big data, fintech lenders can potentially build credit scores and can expand access to credit to even customers with little or no credit history that are underserved by the traditional banks (Chava et al. (2017)). Consistent with this conjecture, the average individual in our sample is a sub-prime borrower with a credit score of 641.[11] Moreover, an economically significant 20% of borrowers in our sample do not have a credit score. This is in contrast to the USA, where fintech lenders primarily cater to borrowers who already have access to credit via traditional banks (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.

The paper closest to our study is Berg et al. (2019). Using data covering approximately 250,000 purchases from an E-Commerce company located in Germany, Berg et al. (2019) 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.

Second, the large majority of customers in their sample access the digital world through desktop, while our data capture deeper aspects of the digital footprint from the mobile phones of customers. 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

[11] 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*)

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. 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.

Third, 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. Fourth, given the nature of our data, we study a richer set of loan outcomes, which includes the likelihood of approval. This allows us to document whether and how lenders use mobile footprints in their loan approval decisions. Moreover, our setting allows us to extrapolate the importance of mobile footprints in measuring creditworthiness for loans taken for different purposes and not just an e-commerce purchase.

Fourth, we find that the default prediction can be improved significantly by using proxies that capture deeper aspects ("deep social footprint") of an individual's digital presence.

Fifth, we also have access to the detailed credit reports of a random sample of borrowers. Using these reports we can construct a set of variables like various expenditure patterns, income, savings, investment etc. which represents the 'deep financial information' of the borrowers. Traditional banks generally have access to these kind of information while granting loans and use them extensively. We run a horse race between the 'deep financial" with the 'deep digital' footprints of the borrowers 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.

Finally, we use novel machine learning algorithms to predict default and conduct counterfactual policy experiments for borrowers who do not have conventional credit bureau scores and generally denied credit. We find that using alternate credit scoring using the mobile and social footprints can expand credit as well as reduce overall default rate. This has significant policy implications for financial inclusion. To the best of our knowledge, this is the first paper which conducted such counterfactual analysis for 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)).[12]

2 Data and Summary Statistics

We obtain proprietary data on about 417,578 loan applicants from a mobile-only Fintech lending platform operating in India since 2016. The lender aims to provide short-term credit to young

[12]See Thakor (2019) for a recent survey of the literature on Fintech.

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 ($141) to ₹200,000 ($2816).[13] The loan duration ranges from a minimum of 15 days to a maximum of 180 days. A total of ₹6500 million ($92 million) worth of loans have been disbursed since its inception in 2016. To get a loan, a customer has to download the lending app, submit all the requisite details and documentations. The borrower also gives permission to the lender to gather additional information on the mode of login, the various apps installed, the number of calls and SMSs, number of contacts on the phone, number of social connections, and the kind of mobile operating systems such as IOS and Android. We obtained data from the lending firm for all loans granted from February 2016 to November 2018.

Table C1 of appendix C provides detailed description of the variables used in our study.

2.1 Summary statistics: loan and financial Variables

Table 1 reports the summary statistics. Out of the 417,578 loan applications in our sample, 272,931 were approved, while 144,647 were denied. The default rate in our full sample is quite high at approximately 13.5%[14]. The average loan size is ₹22,174 ($312) age of a customer is 32 consistent with the idea that lending firm target segment is a young salaried customer.[15] The average credit score is 634 and is obtained from TransUnion CIBIL. The average interest rate charged on loan is 25% (log value of 1.4). On average, a customer earns ₹37,524 ($527) per month or $6324 per annum. Thus, the income of a customer is our sample is roughly three times the median per capita income of $2,134 in 2018. Thus, the lender caters to relatively higher-income customers. The application process also records the purpose for which loan is taken, which can be of the following: Medical, Travel, EMI, Purchases, Loan Repayment, Others. Amongst the sample of approved loans, 8% were taken for the purpose of travel, 9% for EMI, 13 % for purchasing a good, about 8% for the purpose of repaying a loan, 22% for medical expenditure, and rest is uncategorized.

2.2 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. 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 SMSs, 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 2.1% used Linkedin. On average, 68% of the customers have a banking or stock trading app. About 42%

[13] Based on the nominal exchange rate of $1=₹71 as of October 2019.

[14] $\frac{\text{32,555 defaults}}{\text{240,376 approvals}}$

[15] The average loan amount of $312 is based on the exchange rate of $1=₹71.

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).

3 Results

3.1 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 29% higher for loan applications that were approved.[16] 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 SMSs and calls, have a higher number of contacts but fewer connections on a social platform. Approved customers are also 5% more likely to own an iPhone (*IOS* dummy).

In columns 1-3 of Table 1, we analyze the customer and loan characteristics that can potentially predict the likelihood of default. Customers who default on average borrow 71% more than those who don't.[17]. Customers who default on average are charged a higher interest rate ex-ante, consistent with such customers being riskier. Surprisingly, customers who default one average are slightly older and have a greater salary as compared to customers that have not defaulted. Not surprisingly, customers who default have lower credit scores.

Focusing on the social and mobile footprint variables, we find that customers who default are more 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 less likely to have installed a financial app but more likely to have installed a social network/travel 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. 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.

16 $\frac{(22174.26-17182.04)*100}{17182.04}$

17 $\frac{(35228.33-20509.83)*100}{20509.83}$

(2011) find that such customers have a higher retention rate and are more valuable in both the long and short term. Along similar lines, Burks et al. (2015) find that referred workers yield substantially higher profits per worker than non-referred workers. To the extent that the likelihood of referring or being referred is associated with the strength of an individual's social connections, our finding suggests that social ties may have positive spillover effects on the customer's attitude towards default. Consistent with this idea that customers who do not default, send, and receive a greater number of SMSs 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. Finally, owning an Apple phone is negatively associated with the likelihood of default.

3.2 Multivariate Analysis

We now move on to the discussion of our multivariate analysis. Formally, we run a logit or multinomial logit regressions of loan outcome measures on loan and customer characteristics:

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

Where i identifies a unique customer, l identifies a unique loan, and t refers to a year-month. The *Loan outcome* is one of the following: *Approved* is a dummy variable which takes the value one for loans that were approved and zero otherwise, and *Default* which identifies loans in default. Loan Characteristics refer to loan size, and loan purpose. 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.

3.2.1 Loan approvals

We begin our multivariate analysis by examining the determinants of loan application approval. The dependent variable in these tests is a dummy variable which takes the value one for loans that were approved and zero otherwise. Table 2 reports the results of our analysis. Column (1) reports the results using only the credit bureau score (*CIBIL*) as the explanatory variable for the full sample. Not surprisingly, loan applicants with higher credit scores have a higher likelihood of getting approved. The R^2 of the regressions is 0.008, implying that credit scores explain only about 0.8% of the variation in the likelihood of loan approval. In column 2, we repeat the analysis for the subsample of loan applicants with non-missing values of all digital mobile footprint variables and customer characteristics and obtain qualitatively similar results.

In column (3), we repeat these tests after including other loan and customer characteristics. We find that customers that earn more, are older, and need smaller loans, have a higher chance of approval. Education level isn't associated with the likelihood of approval.

In column (4), we report the results for the mobile and social footprint variables. Since prior literature docuemnts that the IOS dummy has significant predictive power for loan outcomes, to make sure that our results are not just driven by the IOS variable, we do not include it in column (4). We find that the number of contacts, the number of apps installed, the presence of financial and mobile loan apps (Finsavvy and Mloan dummy variables) are positively associated with the loan approval. These results continue to hold when we include the IOS dummy in column (5). We find that customers with an IOS device have a 41% higher likelihood of approval compared to those without an IOS device. This is consistent with prior studies, which highlight that owning an IOS device is a strong predictor of higher earnings (Bertrand & Kamenica (2018)). Overall, these results indicate that social and mobile footprint variables have significant explanatory power for the likelihood of loan approval. The AUC of the model with digital mobile footprint variables at 54.2 is significantly higher than the AUC of 51.6 for the model with credit score alone in column (2). The results remain robust to including credit score in column (6).

Column (7), includes all, customer characteristics, and mobile footprint but excludes credit bureau score. Our objective here is two folds. First, we want to understand whether our results on mobile footprint continue to hold once we control for other loan level and customer level characteristics. For instance, some of the variables, such as owing an IOS device, may simply be a proxy for the income of the customer and thus may not have any independent explanatory power over the customer's salary. Second, we want to examine if observable customer, and mobile footprint characteristics can explain a higher fraction of the variation in loan approval decisions as compared to just the CIBIL score. We find that customer's salary, number of contacts, number of apps installed, finsavvy, mloan, and IOS dummies continue to be statistically significant. Further, the AUC of the model with customer, and mobile footprint characteristics is 4% more than that of the model with CIBIL score alone. This suggests that customer characteristics, and mobile footprint, have some complementary information beyond what is captured in the CIBIL score.

Finally, in columns (8) and (9), we also include CIBIL score and state fixed effects. The results remain qualitatively similar. Summarizing, the key takeaway from this section for the purpose of our study is that mobile footprint variables have significant explanatory power for loan approval decisions even in the absence of a credit bureau score.[18]

3.2.2 Defaults

In this section, we focus on analyzing the relationship between mobile/social footprint variables, loans, and customer characteristics and default. The dependent variable in these tests is a dummy variable which takes the value one for delinquent loans. Table 3 reports the results from these

[18]In Table A1 of Appendix A, we repeat these tests with the subsample of customers without a credit score. The AUC of a model using mobile footprint variables to predict loan approval is 71.2%, again supporting our claim that the fintech lender relies heavily on digital variables in the absence of a credit bureau score.

tests. Column (1) reports the results using only the credit bureau score (*CIBIL*) as the explanatory variable for the sample of approved loans. Not surprisingly, a higher credit bureau score is associated with a significantly lower likelihood of default. In column (2), we repeat the analysis for the subsample of loan applicants with non-missing values of all digital mobile footprint variables and customer characteristics. The AUC of the CIBIL score in this sample is 59%. The AUC of credit score in our sample, 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 peer to peer lending platform, "Propser.com" and 68.3% reported by Berg et al. (2019) based on a sample of purchases from a German e-retailer but comparable to the AUC of 59.8% using U.S. credit scores from Lending Club reported in Berg et al. (2019). This suggests that the discriminatory ability of the credit score in predicting defaults is likely to vary across countries and types of financial intermediary.

In column (3), 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 (4), we report the results for mobile and social footprint variables. Again, since the IOS dummy has significant predictive power for loan outcomes, to make sure that our results are not just driven by the IOS variable, we do not include it in column (4). The AUC of this specification is 60.5% and approximately 2% 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 0.68, implying that individuals without a financial app are about one and a half times more likely to default relative to those that have such an app installed. This suggests that *Finsavvy* dummy may be correlated with the financial sophistication of a customer. In contrast, those with a dating app (any other social network app) are 30% (38%) more likely to default.[19]. Interestingly, customers with a travel app are about 4% more likely to default than those without. Finally, those who log in to the application via Linked or Facebook are 24% and 9% more likely to default respectively relative to those who via other means. 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 apps installed on the phone has significant discriminatory power in default prediction.

In column (5), we also include the IOS dummy. The statistical and economic significance of other mobile footprint variables remains qualitatively similar. In line with the evidence in Berg et al. (2019), we find that borrowers with IOS operating system (Apple) are significantly less likely to default relative to the Android operating system. The odds ratio of *IOS* dummy is 0.496, implying that those with an android phone are twice as likely to default as those with an Apple phone. In column (6), we include both the CIBIL score and mobile mobile footprint variables together. We note that the AUC of this model is 60.8% and 2.2 percentage points higher than that of a model

[19]The odds ratio of *Dating* dummy is 1.30 and that of *Socialconnect* dummy is 1.38

using only the credit bureau score.

As in Table 2, column (7) includes all customer characteristics, and digital mobile footprint variables but excludes the credit bureau score. We find that the coefficients of the digital mobile footprint variables largely remain unchanged, suggesting that these variables have incremental predictive power over loan and customer characteristics. More specifically, the digital mobile footprint seems to be capturing unobservable aspects of customer behavior, which is not absorbed by education, age, salary, or job designation of the customer. Interestingly, the coefficient estimate of *IOS* dummy remains statistically significant even after controlling for the customer's monthly salary. In this respect, our study complements Berg et al. (2019) who conjecture that discriminatory ability of owning an apple device is presumably driven by its correlation with earnings. Specifically, our finding implies that owing an Apple device captures an unobservable aspect of individuals that is not fully absorbed by earnings.

Interestingly, we also find that customers who log in through Facebook are more likely to default even once we control for customer characteristics. Importantly, the AUC of this specification is 72% , three percentage points higher than the AUC of the model using only the credit bureau score and eight percentage points higher than the model, which includes customer characteristics. This suggests that mobile and social footprint variables not only complements credit bureau score but also customer characteristics.

Finally, in columns (8) and (9), we also include CIBIL score and state fixed effects for robustness. The results remain qualitatively similar.

One concern with our evidence so far could be that it is driven by a subsample of customers in our sample. For instance, mobile footprint variables may have predictive power only for customers with a high credit score or salary. This would limit the promise of using digital mobile footprints to score customers without a credit bureau score/history. To further strengthen the evidence regarding the discriminatory ability of digital mobile footprint variables in predicting defaults, in tables A2-A5 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.

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. Our 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..

3.2.3 Predicting defaults for customers without credit score

While our analysis so far suggests that the digital mobile footprint has incremental explanatory power for predicting defaults, customers who lack credit history and credit score may be very different from the set of customers with a credit bureau score. To examine if these results are

generalizable to the set of unscorable customers, in Table 4, we focus on the set of customers without a credit score and examine whether and how does the digital mobile footprint perform in default prediction for this subsample. In column (1), we only include customer characteristics and find that these have significant discriminatory power. The AUC of the model is 55%. In column (2), we include mobile and social footprint variables and find that the AUC of the model is 58% and comparable to the predictive performance of the credit bureau score in Table 3. Importantly, in column (3), we include customer characteristics, and mobile footprint variables together to examine if mobile footprint variables have incremental explanatory power over customer and loan characteristics. As compared to column (1), including digital variables improves the AUC by 5.6% which is considered to be a significant improvement.][20] Summarizing, these findings suggest that digital mobile footprints can indeed be used to score customers without a credit history and conventional credit score.

3.2.4 Predicting defaults using deep social footprints from call logs

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 et al. (2014)), which is known to be an important predictor of loan defaults (Karlan (2005)).

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 vsòutgoing 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 A6 of Appendix A provides the details of how we construct these measures, and panel A

[20] See for instance, Iyer et al. (2015) and Berg et al. (2019)

of Table 5 reports the univariate summary statistics. 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 suggesting that defaulters, 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.

In Table 6, we again use our baseline multivariate logit model to examine whether the measures based on call logs predict defaults. Given that the various call based measures are correlated with each other, it is important to note that our goal is not to understand the direction of causality but rather to understand whether a model that includes these variables does a good job of predicting loan defaults. We start by analyzing the predictive ability of the credit bureau score for the subsample of customers for whom call details are available in column (1). The AUC of the credit score at 58.4% is comparable to what we observed in the full sample in Table 3. In column (2), we include only deep social footprints based on call logs. The AUC of this model is remarkably high and 6% more than the model with credit score alone. In columns 3 and 4, we include call log measures along with other digital mobile footprint variables and credit score respectively and find that the AUC goes up to 66%, an 8% improvement over the model with credit score alone. This is better than the 5.7 percentage points AUC improvement reported in Iyer et al. (2015) who compare the AUC using the Experian credit score to the AUC in a setting where, in addition to the credit score, lenders have access to a large set of borrower financial information as well and comparable to the improvement in the AUC by 8.8 percentage points reported by Berg et al. (2017) in a consumer loan sample of a large German bank in a setting where, in addition to the credit score, lenders have access to account data, as well as socio-demographic data and income information.

We next examine whether digital mobile footprints taken together have incremental explanatory power over and above a model that includes credit score, l and customer characteristics. In column (5) of Table 6, we also include customer characteristics. The AUC of this model is 67.1%. In column (5), we include mobile footprint variables along with customer characteristics and find that the AUC of this model outperforms the model in column 1 by about 8.7%. Finally, in column (6), we include credit score, and customer characteristics, and mobile footprint variables together. We find that including credit score does not improve the AUC significantly over a model with customer characteristics, and mobile/social footprint variables.[21]

Finally, Table 7 reports the relative performance of 'deep' financial vs 'deep' digital information

[21] In Table A7 of the appendix, we also include ex-post measures based on call logs information during the first 15 days after the loan was granted and obtain similar results. In Table A8 of the appendix, we repeat these tests with the subsample of customers without a credit bureau score and obtain qualitatively similar results. We do not report these in the main tables as the sample of customers without a credit score for whom call logs information is also available is small.

for a subset of the borrowers in predicting defaults. As mentioned earlier, the 'deep' financial information like spending in last three months, other borrowing, number of transactions in the bank account etc. are found in the financial reports of the borrower accessed during the loan application process. Column (1) reports the performance of mobile and social footprint variables which has a significantly higher AUC of about 60% in predicting default. In contrast, the AUC of the model with 'deep' financial information reported in column (4) is only 54%.

4 Machine Learning Models for Default Prediction and Credit Scoring

4.1 Motivation

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

The problem at hand is, therefore, to see whether social and mobile footprints predict loan default using machine learning models. This is essentially a prediction problem, where we want to use the sample data to predict defaults "out-of-sample." Standard estimation approaches like OLS, where we use all the data to make in sample prediction is not well suited for such analysis. The in-sample estimation approaches 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 minimizes the mean squares error of the prediction by a joint minimization procedure cognizant of the bias-variance trade-off and, as such, are particularly useful to address our research question.

The prediction of default risk for borrowers without a traditional credit score is useful and can be used to ask the counterfactual questions: how many denied borrowers (perhaps due to lack of traditional credit score) would have been approved had we used the social and mobile footprint based alternate credit scores? What would have been the impact on default 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 default risk of the borrower. We follow the methodology outlined in Athey (2017) and Kleinberg et al. (2015) to address these prediction counterfactual questions in this section. We start by verifying the predictive power of social and mobile footprint variables for defaults using different machine learning algorithms. Subsequently, we conduct the counterfactual prediction exercise.

4.2 Machine learning models

We use three machine learning 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. We use Logistic regression, Random Forest classification and XGBoost models to estimate the default predictability. We briefly describe below various models.

4.2.1 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. In 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. Throughout this section, following standard practice we use a five fold cross-validation.

4.2.2 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, and the average of the outcome is taken as the best predictor for each sub-group of 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 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 cut-offs of the age-based sub-groups is 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 are 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:

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

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 very easy to explain, it is generally 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 B bootstrap samples of size Z^* and fitting a tree for each such sample and averaging[22] the outcome over the bootstrap sample as a predicted value for each tree.

4.2.3 Boosting regression trees (XGBoost)

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

4.3 Model selection: prediction performance

There are various ways that the performance of a particular model can be evaluated in machine learning. Area Under the Curve (AUC) and Recall are two widely used model selection criteria.

AUC: The area under the ROC curve is used as a measure of the goodness of a prediction. It measures the proportion of true positives in a prediction. Higher the AUC, the higher is the prediction accuracy.

Accuracy: Accuracy is another measure of prediction accuracy in machine learning models defined as the proportion of correct predictions out of total predictions.

4.4 Comparison of social and mobile footprints

In this section, we compare the three 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-level financial variables. Panel A of Table 8 reports these results. Figure 1 plots the AUC curves for different models.[23] 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 all three methods of machine learning algorithms. For example, under the Random Forest algorithm, the model with only CIBIL score has an AUC of about 0.58 while

[22] For a classification problem, generally a majority vote is taken over the bootstrap sample.

[23] Appendix B provides additional details regarding the machine learning estimation procedure.

the models with digital mobile and 'deep' social footprint variables have an AUC of about 0.77 and 0.72 respectively.

Figure 2 shows the variable importance factors of various features for predicting defaults.[24] Interestingly the social and mobile footprint variables (standardized incoming and outgoing call, log of number of apps, and ios dummy etc.) come as significant features in predicting the default; far ahead of the customer characteristics such as salary and age and the borrower credit score (CIBIL).

4.4.1 Borrower heterogeneity

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 credit scores as there is likely to be greater information asymmetry regarding such customers. Panels B and C of the Table 8 reports the results of these tests.

Focusing on panel B, we find that for the borrowers who are in the bottom 25% of the CIBIL score distribution, the social and mobile footprint variables have higher predictive power for default than that of the CIBIL score. Using the Random forest method, the AUC of the model with only the mobile/social footprint variables is 0.72, whereas that of the model with the CIBIL score is 0.55. Moreover, the AUC of the deep social footprint model is higher (0.77) for the borrowers in the bottom 25% of the CIBIL score relative to the full sample (0.72).

In panel C of Table 8, we evaluate the predictive power of mobile and social footprint variables for the borrowers who are in the top end of the spectrum of the CIBIL score (more than 750 CIBIL score). We find that social variables have higher predictive power relative to the credit score (CIBIL) even for the borrowers who belong to the higher end of the spectrum of the CIBIL score. Using Random forest algorithm, the AUC of the mobile footprint model is 0.72, whereas that of the CIBIL score is 0.53. Interestingly, the customer characteristics which includes the age, salary and designation of the borrowers have better predictability than the CIBIL scores. This is intuitive as most of our borrowers are young and hence do not have a long credit history to have a higher credit score. However their salary and designation is more informative of their credit worthiness. This reinforces our argument of non-suitability of credit ratings for millennials.

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.

[24]The importance of a particular variable (feature) in the random forest (or boosting) based classification algorithm is evaluated by its relative contribution in improving the prediction strength. For each split in each tree, the improvement in the split criterion (say Gini index) is accumulated over the entire tree for each variable and then averaged relative to the number of trees in the forest. The variable importance measure therefore effectively summarizes the importance of a particular variable in designing a random forest. We use such variable importance measures to identify the key variables in the prediction problem.

4.4.2 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 these borrowers based on an alternative credit score based on their social and mobile footprints.

Our database has about 3300 borrowers who were approved for loans.[25] We 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 Logistic, random forest, and XGBoost. We then use the predicted 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 D of Table 8. We find that the mobile and social footprint variables together do a remarkable job in predicting defaults even for the training sample with AUCs in the range of 0.64–0.77.[26]

4.5 Comparison of deep social and deep financial variables

As mentioned before, for a subset of customers in our sample, we have detailed information regarding their call logs and 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 C2 of appendix C. In Table 9, 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 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 deep social footprint is 0.744, about fifteen percentage points higher than AUC of the model with only deep financial information. Moreover a model that includes mobile footprint, deep social footprint and CIBIl score does better in predicting defaults out of sample as compared to a model with deep financial information and CIBIL score.

Finally, figure 3 shows the variable importance factors of different features for predicting defaults in a model that includes Cibil score, mobile and deep social footprint, and deep financial variables. We highlight that the deep social footprint variables are the most commonly occurring feature and trump both CIBIl score and deep financial information.

We conclude that digital footprint has significant ability in predicting defaults and the information content of these variables complements rather than substitutes for both the credit bureau score and detailed financial information regarding a customer's income and expenses.

[25] Table A9 in the appendix reports the summary statistics for these set of customers.

[26] Iyer et al. (2015) note that and AUC > 0.7 is considered desirable in informationally scarce environments.

4.6 Counterfactual policy experiment for borrowers without a credit bureau score

Our results thus far show that social and mobile footprints have higher predictive power in borrower credit risk prediction for fintech lending. 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 these borrowers who would have been given loans if we had relied only on accessing their creditworthiness using social and mobile footprints?; 2)What would be the performance of the loan outcome if we had replaced the high-risk borrowers with credit scores (who were eventually approved for a loan) by a specific group of borrowers who were not approved (perhaps because they did not have credit scores), but for whom creditworthiness can be evaluated based on social and mobile footprints?

These counterfactual questions have significant policy implications. Importantly these questions are not causal in nature. The focus on prediction policy counterfactual rather than causal questions is relatively new in economics (Athey (2017), Kleinberg et al. (2015)). We follow Kleinberg et al. (2015) in addressing the counterfactual policy questions posed above.

We have data on loan outcomes, such as whether a borrower was approved for a loan or not and whether they defaulted conditional on getting a loan. We also observe a set of common characteristics for both sets of borrowers. The common characteristics include (see Table A9) personal characteristics like age, salary, etc., and social and mobile footprints like number of SMSes, no of calls, number of contacts, number of apps installed in the mobile phone, the type of phone (Apple vs. others), whether they logged in using their Facebook or LinkedIn information etc.

Our algorithm proceeds in the following steps:

1. Split the sample of all borrowers who were approved into a training and testing sample. Use various machine learning algorithms (Logistic Regression, Random Forest, XGBoost) to estimate the model parameters. Since a relatively small portion of the approved borrowers eventually defaulted, we use a balancing method 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 without credit score who were not approved for a loan to predict their probability of default. Next, we use different thresholds for the predicted probability of default to evaluate how many borrowers who were not approved would have been approved based on their mobile and social footprint.

In Table 10, we report the results of our counterfactual exercise examining the fraction of borrowers who were denied credit but would have been approved based on the different cut-offs of the predicted default probability. Panel A of Table 10 reports the counterfactual proportion of people for the borrowers who had a CIBIL score but were denied a loan. For instance, if we had

selected a very high predicted default threshold of (say) 95% then about 94% of the borrowers would have been approved. The default rate in our sample is approximately 12%. Even if we choose a relatively conservative threshold of 10% (1%) predicted default probability, about 42% (22%) borrowers with the CIBIL score who were denied loans would have been approved.

In panel B of Table 10 we report the the counterfactual proportion of people for the borrowers who did not have a CIBIL score and were denied a loan. Here if we choose a relatively conservative threshold of 10% (1%) predicted default probability, about 36% (14%) borrowers without the CIBIL score who were denied loans would have been approved.

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

5 Conclusion

In this paper, we have used a unique and proprietary dataset to analyze the impact of the mobile footprint of individual borrowers in predicting loan outcomes. Our dataset comes from a leading fintech lending company in India. We find that the mobile and social footprint has significantly more predictive power than traditional credit score used by banks.

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

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

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

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

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*.

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

Berg, T., Burg, V., Gombović, A., & Puri, M. (2019). On the rise of fintechs–credit scoring using digital footprints. *Review of Financial Studies, Forthcoming*.

Berg, T., Puri, M., & Rocholl, J. (2017). Loan officer incentives, internal ratings and default rates. *Working paper*.

Bertrand, M. & Kamenica, E. (2018). Coming apart? cultural distances in the united states over time. *Working paper*.

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*.

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*.

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.

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.

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.

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

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.

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

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.

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.

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: **Panel A: Summary Statistics of Customer and loan characteristics**

This table reports summary statistics on the customer and loan characteristics. 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	22174.26	17182.04	-4992.22***	35228.33	20509.83	-14718.49***
Log Interest Rate	1.445	0.892	-0.552***	1.857	1.393	-0.463***
Loanpurpose Medical	0.214	0.095	-0.118***	0.247	0.209	-0.037***
Loanpurpose Travel	0.082	0.024	-0.057***	0.075	0.082	0.007***
Loanpurpose EMI	0.087	0.081	-0.005***	0.073	0.088	0.014***
Loanpurpose purchase	0.133	0.068	-0.065***	0.129	0.133	0.004***
Loanpurpose Loanrepayment	0.081	0.047	-0.034***	0.082	0.081	-0.0006
Loanpurpose Other	0.405	0.232	-0.172***	0.395	0.405	0.010***
Age	31.89	29.45	-2.44***	32.00	31.88	-0.117***
Salary	37524.53	30346.39	-7178.13***	39262.32	37342.43	-1919.89***
CIBIL (>0, N=219k & 16k)	634.40	470.82	-163.58***	602.04	639.10	37.06***
Facebook Status	0.267	0.296	0.029***	0.274	0.263	-0.011***
Linkedin Status	0.021	0.015	-0.006***	0.023	0.021	-0.002**
Googleplus_status	1.712	1.690	-0.021***	1.700	1.714	0.013***
Referral	0.116	0.039	-0.077***	0.115	0.118	0.002*
Sales App	0.195	0.198	0.003	0.188	0.196	0.007***
Dating App	0.029	0.028	-0.001	0.029	0.029	0.0006
Finsavy app	0.679	0.034	-0.645***	0.677	0.862	-0.019***
Socialconnect app	0.714	0.036	-0.677***	0.760	0.708	-0.051
Travel app	0.567	0.048	-0.518***	0.576	0.566	-0.010***
Mloan app	0.423	0.020	-0.403***	0.423	0.423	-0.0002
Referrer	0.234	0.034	-0.200***	0.167	0.243	0.075
# of SMS	2481.71	1109.00	-1372.71***	1949.25	2548.19	598.94***
# of Apps	54.53	41.26	-13.27***	47.07	55.47	8.40***
# of Contacts	844.84	683.81	-161.02***	827.64	847.03	19.38***
# of Connections	525.89	452.39	-73.50	413.23	539.15	125.92***
# of Calls	3136.50	2071.97	-1064.53***	2394.96	3229.05	834.08***
IOS	0.119	0.066	-0.053***	0.112	0.120	0.007***
			Education			
<High School	0.112	0.310	0.197***	0.119	0.112	-0.007***
High School	0.645	0.546	-0.090***	0.647	0.645	-0.002
College	0.241	0.144	-0.097***	0.233	0.243	0.010***
			Job Designation			
Worker	0.354	0.410	0.056 ***	0.347	0.354	0.007***
Supervisor	0.248	0.254	0.005***	0.245	0.248	0.003
Manager	0.398	0.335	-0.063***	0.407	0.397	-0.010***
N	272,931	144,647		32,555	240,376	

Table 2: **Approval of loans**

This table reports the estimates from our logit regressions examining the determinants of loan approval. The dependent variable, Approved takes the value one for loan applications that were approved and zero for those that were denied. The specification in column (1) only includes the credit bureau score (Log of CIBIL) as the explanatory variable with observations from the full sample. Column (2) includes the credit bureau score (Log of CIBIL) with observations from only the subsample. Column (3) includes only customer characteristics. Column (4) includes only mobile/social footprint variables excluding IOS dummy. Column (5) includes only mobile/social footprint variables along with IOS dummy. Column (6) includes only mobile/social footprint variables and CIBIL score and IOS Dummy. Column (7) includes all customer characteristics and mobile/social footprint variables but not the CIBIL score. Column (8) includes all variables including the CIBIL score. Column (9) includes all variables including the CIBIL score and state fixed effects. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)	Odds Ratio (4)	Odds Ratio (5)	Odds Ratio (6)	Odds Ratio (7)	Odds Ratio (8)	Odds Ratio (9)
Log of cibil	1.182***	1.030**				1.029**		1.034**	1.031**
	(0.000)	(0.023)				(0.031)		(0.012)	(0.024)
Log of Salary			0.822***				0.790***	0.786***	0.775***
			(0.000)				(0.000)	(0.000)	(0.000)
Log Age			1.208**				1.373***	1.292**	1.289**
			(0.044)				(0.001)	(0.011)	(0.013)
High School Dummy			1.015				1.003	0.991	1.006
			(0.753)				(0.953)	(0.855)	(0.913)
College Dummy			1.008				1.003	0.989	1.003
			(0.884)				(0.958)	(0.848)	(0.952)
Supervisor			0.905**				0.901***	0.904**	0.908**
			(0.010)				(0.008)	(0.011)	(0.017)
Manager			0.945				0.947	0.939*	0.943
			(0.117)				(0.133)	(0.088)	(0.124)
Log no of SMS				0.986*	0.986*	0.984**	0.990	0.988	0.986*
				(0.073)	(0.086)	(0.046)	(0.221)	(0.123)	(0.091)
Log No of Contacts				0.979	0.978	0.974	0.986	0.984	0.988
				(0.235)	(0.212)	(0.133)	(0.438)	(0.375)	(0.515)
Log no of Apps				1.162***	1.167***	1.170***	1.189***	1.191***	1.183***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Callog				1.037***	1.039***	1.039***	1.036***	1.036***	1.035***
				(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.003)
Dating App				0.973	0.970	1.000	0.999	1.026	1.047
				(0.752)	(0.727)	(0.999)	(0.993)	(0.772)	(0.618)
Finsavy App				1.199***	1.200***	1.175***	1.191***	1.160**	1.157**
				(0.002)	(0.002)	(0.009)	(0.004)	(0.016)	(0.021)
Socialconnect App				0.940	0.976	0.998	0.929	0.951	0.892
				(0.502)	(0.796)	(0.981)	(0.431)	(0.607)	(0.261)
Travel App				0.987	0.983	0.976	1.025	1.018	1.007
				(0.718)	(0.644)	(0.512)	(0.496)	(0.627)	(0.861)
Mloan App				1.087***	1.087***	1.088***	1.081**	1.082**	1.094***
				(0.007)	(0.007)	(0.008)	(0.013)	(0.013)	(0.005)
Facebook status				1.035	1.034	1.030	1.038	1.035	1.040
				(0.300)	(0.316)	(0.379)	(0.271)	(0.317)	(0.253)
Linkedin status				0.901	0.906	0.887	0.969	0.952	0.976
				(0.262)	(0.288)	(0.202)	(0.734)	(0.602)	(0.797)
IOS Dummy					1.407***	1.393***	1.467***	1.447***	1.422***
					(0.001)	(0.002)	(0.000)	(0.001)	(0.001)
Constant	4.821***	33.590***	165.025***	19.063***	17.765***	15.687***	65.531***	73.334***	75.718***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
State Fixed Effects	N	N	N	N	N	N	N	N	Y
Observations	235,564	189,096	194,136	194,136	194,136	189,096	194,136	189,096	185,203
Pseudo R2	0.00791	0.000113	0.00106	0.00219	0.00246	0.00250	0.00385	0.00396	0.00549
AUC	0.585	0.516	0.534	0.539	0.542	0.541	0.553	0.554	0.562

Table 3: **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. The dependent variable, Default takes the value one for loans that are delinquent and zero otherwise. The specification in column (1) only includes the credit bureau score (Log of CIBIL) as the explanatory variable with observations from the full sample. Column (2) includes the credit bureau score (Log of CIBIL) with observations from only the subsample. Column (3) includes only customer characteristics. Column (4) includes only mobile/social footprint variables excluding IOS dummy. Column (5) includes only mobile/social footprint variables along with IOS dummy. Column (6) includes only mobile/social footprint variables and CIBIL score and IOS Dummy. Column (7) includes all customer characteristics and mobile/social footprint variables but not the CIBIL score. Column (8) includes all variables including the CIBIL score. Column (9) includes all variables including the CIBIL score and state fixed effects. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Odds Ratio	Odds Ratio	Odds Ratio	Odds Ratio	Odds Ratio	Odds Ratio	Odds Ratio	Odds Ratio	Odds Ratio
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Log of cibil	0.877***	0.899***				0.906***		0.902***	0.906***
	(0.000)	(0.000)				(0.000)		(0.000)	(0.000)
Log of Salary			1.253***				1.419***	1.442***	1.419***
			(0.000)				(0.000)	(0.000)	(0.000)
Log Age			0.861***				0.581***	0.614***	0.617***
			(0.002)				(0.000)	(0.000)	(0.000)
High School Dummy			0.864***				0.898***	0.909***	0.904***
			(0.000)				(0.000)	(0.000)	(0.000)
College Dummy			0.781***				0.805***	0.820***	0.816***
			(0.000)				(0.000)	(0.000)	(0.000)
Supervisor Dummy			0.986				1.007	1.028	1.046**
			(0.468)				(0.707)	(0.148)	(0.022)
Manager Dummy			0.978				1.015	1.025	1.056***
			(0.192)				(0.404)	(0.172)	(0.003)
Log no of SMS				0.970***	0.968***	0.973***	0.962***	0.967***	0.967***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log No of Contacts				0.969***	0.972***	0.983*	0.971***	0.980**	0.980**
				(0.000)	(0.001)	(0.057)	(0.001)	(0.022)	(0.024)
Log no of Apps				0.655***	0.649***	0.651***	0.630***	0.632***	0.633***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Log Callog				0.917***	0.913***	0.916***	0.918***	0.921***	0.924***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Dating App				1.303***	1.310***	1.286***	1.233***	1.215***	1.199***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Finsavy App				0.683***	0.681***	0.723***	0.685***	0.731***	0.733***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Socialconnect App				1.383***	1.282***	1.322***	1.382***	1.423***	1.616***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Travel App				1.041**	1.048***	1.047***	0.987	0.985	0.980
				(0.019)	(0.006)	(0.009)	(0.468)	(0.384)	(0.270)
Mloan App				1.010	1.010	1.012	1.018	1.021	1.018
				(0.516)	(0.500)	(0.422)	(0.234)	(0.183)	(0.264)
Facebook status				1.092***	1.095***	1.102***	1.088***	1.094***	1.089***
				(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Linkedin status				1.237***	1.223***	1.217***	1.123**	1.110**	1.113**
				(0.000)	(0.000)	(0.000)	(0.012)	(0.026)	(0.024)
IOS Dummy					0.496***	0.511***	0.459***	0.474***	0.482***
					(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.324***	0.258***	0.026***	1.709***	1.969***	2.885***	0.412***	0.430***	0.545***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)
State Fixed Effects	N	N	N	N	N	N	N	N	Y
Observations	219,373	184,538	189,403	189,403	189,403	184,538	189,403	184,538	180,816
Pseudo R-squared	0.00422	0.00241	0.00200	0.0212	0.0227	0.0229	0.0267	0.0271	0.0286
AUC	0.601	0.586	0.538	0.605	0.608	0.608	0.620	0.620	0.622

Table 4: **Predicting loan defaults (subsample without credit score)**
This table reports the estimates from our logit regressions examining the relationship between mobile/social footprint variables, customer characteristics and likelihood of default using the sample of observations with no credit bureau score available. The dependent variable, Default takes the value one for loans that are delinquent and zero otherwise. The specification in column (1) includes customer characteristics. Column (2) includes the mobile/social footprint variables for the same sample. Column (3) includes customer characteristics with mobile/social footprint variables. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)
Log of Salary	1.286***		1.482***
	(0.000)		(0.000)
Log Age	1.200		0.928
	(0.116)		(0.505)
High School Dummy	0.902**		0.934
	(0.044)		(0.196)
College Dummy	0.746***		0.786***
	(0.000)		(0.000)
Supervisor	0.867***		0.918*
	(0.001)		(0.061)
Manager	1.026		1.164***
	(0.526)		(0.000)
Log no of SMS		0.961***	0.958***
		(0.000)	(0.000)
Log No of Contacts		0.981	0.952**
		(0.346)	(0.017)
Log no of Apps		0.841***	0.820***
		(0.000)	(0.000)
Log Callog		0.946***	0.956***
		(0.000)	(0.000)
Finsavy App		0.212***	0.220***
		(0.000)	(0.000)
Socialconnect App		10.366***	10.974***
		(0.000)	(0.000)
Travel App		0.931	0.894*
		(0.215)	(0.053)
Mloan App		0.974	0.966
		(0.745)	(0.666)
Facebook status		0.878***	0.881***
		(0.002)	(0.002)
Linkedin status		1.099	0.975
		(0.434)	(0.837)
IOS Dummy		0.829	0.779*
		(0.157)	(0.066)
Constant	0.005***	0.385***	0.012***
	(0.000)	(0.000)	(0.000)
Observations	45,648	45,385	45,350
Pseudo R2	0.00367	0.0283	0.0346
AUC	0.550	0.583	0.606

Table 5: **Summary statistics of call logs and financial transactions**
This table reports summary statistics on call log variables. Columns 1-3 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.

Panel A: Call logs Metrics

Call Log Metric	Default (1)	Not Default (2)	Difference (3)
First 15 days: Per day Per person Avg No. of Incoming calls	1.53	1.49	-0.045***
First 15 days: Per day Per person Avg No. of Outgoing calls	2.14	2.03	-0.118***
First 15 days: Per day Per person Avg No. of Missed calls	1.57	1.48	-0.095***
First 15 days: Per day Per person Avg Duration of Incoming calls	156.14	157.55	1.40
First 15 days: Per day Per person Avg Duration of Outgoing calls	154.19	154.80	0.617
First 15 days: Per day No. of persons called	15.00	13.60	-1.40***
Past days: Per day Per person Avg No. of Incoming calls	1.58	1.52	-.051***
Past days: Per day Per person Avg No. of Outgoing calls	2.22	2.10	-.124***
Past days: Per day Per person Avg No. of Missed calls	1.61	1.51	-.099***
Past days: Per day Per person Avg Duration of Incoming calls	167.05	560.02	392.96***
Past days: Per day Per person Avg Duration of Outgoing calls	194.89	167.28	-27.61***
Past days: Per day No. of persons called	15.69	14.32	-1.37***
First 15 days: Per day Total No. of Incoming calls	10.97	9.73	-1.23***
First 15 days: Per day Total No. of Outgoing calls	20.76	17.53	-3.23***
First 15 days: Per day Total No. of Missed calls	7.44	5.80	-1.63***
First 15 days: Per day Total Duration of Incoming calls	1023.86	943.72	-80.13***
First 15 days: Per day Total Duration of Outgoing calls	1346.78	1205.45	-141.32***
Past days: Per day Total No. of Incoming calls	11.61	10.44	-1.172***
Past days: Per day Total No. of Outgoing calls	22.45	19.13	-3.31***
Past days: Per day Total No. of Missed calls	7.84	6.16	-1.67***
Past days: Per day Total Duration of Incoming calls	1113.62	1415.59	301.97 ***
Past days: Per day Total Duration of Outgoing calls	1561.83	1360.44	-201.38***
First 15 days: HHI of No. of Incoming calls	1049.70	890.56	-159.14***
First 15 days: HHI of No. of Outgoing calls	965.42	835.32	-130.09***
First 15 days: HHI of Total Duration of Incoming calls	1766.61	1597.18	-169.42***
First 15 days: HHI of Total Duration of Outgoing calls	1805.54	1681.39	-124.14***
First 15 days:HHI of No. of Missed calls	1430.33	1265.01	-165.31***
Past days: HHI of No. of Incoming calls	202.09	123.67	-78.41***
Past days: HHI of No. of Outgoing calls	201.19	128.40	-72.79***
Past days: HHI of Total Duration of Incoming calls	467.19	307.44	-159.75***
Past days: HHI of Total Duration of Outgoing calls	499.17	347.80	-151.36***
Past days:HHI of No. of Missed calls	291.76	176.06	-115.70***
N	17,095	89,052	
Panel B: Financial Transaction Metrics			
Debits to credits ratio	0.699	0.707	-0.007
# of Transactions	169.09	159.65	-9.44***
Expenditure to Income ratio	101.51	101.75	-0.321
Avg 2 Month Appreciation in Balance	411.90	-855.62	-1267.22
N	1,189	15,299	

Table 6: **Predicting Loan Defaults with deep social footprint based on call logs**

This table reports the estimates from our logit regressions examining the relationship between mobile/social footprint variables, customer characteristics and call logs and likelihood of default. The dependent variable, Default takes the value one for loans that are delinquent and zero otherwise. The specification in column (1) only includes the credit bureau score (Log of CIBIL) as the explanatory variable. Column (2) includes only call log variables. Column (3) includes call logs with mobile/social footprint variables. Column (4) includes call logs, mobile/social footprints and credit bureau score. Column (5) includes all variables of customer characteristics, mobile/social footprints and call logs excluding CIBIL score. Column (6) includes all variables including the CIBIL score. Standard errors are clustered at the state level. (∗ ∗ ∗), (∗∗), (∗) denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)	Odds Ratio (4)	Odds Ratio (5)	Odds Ratio (6)
Log of cibil	0.889***			0.911***		0.905***
	(0.000)			(0.000)		(0.000)
Past days: Per day Per person Avg No. of Incoming calls		1.080***	1.089***	1.082***	1.095***	1.090***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: Per day Per person Avg No. of Outgoing calls		1.017	1.015	1.012	1.010	1.008
		(0.334)	(0.283)	(0.359)	(0.482)	(0.570)
Past days: Per day Per person Avg No. of Missed calls		0.950***	0.947***	0.945***	0.956***	0.955***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: Per day Per person Avg Duration of Incoming calls		0.000***	0.000***	0.000***	0.000***	0.000***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: Per day Per person Avg Duration of Outgoing calls		0.916	0.942	0.962	0.934	0.955
		(0.733)	(0.718)	(0.755)	(0.711)	(0.753)
Past days: Per day No. of persons called		0.828***	0.888***	0.894***	0.894***	0.900***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: Per day Total Duration of Incoming calls		0.020	0.826	0.606	0.025	0.017
		(0.328)	(0.958)	(0.889)	(0.330)	(0.271)
Past days: Per day Total No. of Incoming calls		0.990	0.983	0.990	0.991	0.998
		(0.688)	(0.455)	(0.665)	(0.687)	(0.930)
Past days: Per day Total No. of Outgoing calls		1.368***	1.282***	1.256***	1.317***	1.291***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: Per day Total Duration of Outgoing calls		0.776**	0.825**	0.858**	0.787***	0.816***
		(0.029)	(0.021)	(0.030)	(0.008)	(0.009)
Past days: Per day Total No. of Missed calls		1.380***	1.353***	1.353***	1.344***	1.344***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: HHI of No. of Incoming calls		0.820***	0.825***	0.824***	0.833***	0.832***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: HHI of No. of Outgoing calls		1.013	1.014	1.021	1.006	1.012
		(0.783)	(0.767)	(0.658)	(0.904)	(0.804)
Past days: HHI of Total Duration of Incoming calls		1.317***	1.297***	1.299***	1.289***	1.291***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: HHI of Total Duration of Outgoing calls		1.230**	1.185*	1.164*	1.180*	1.161
		(0.017)	(0.056)	(0.100)	(0.066)	(0.111)
Past days: HHI of No. of Missed calls		1.093***	1.078***	1.076***	1.077***	1.075***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.312***	0.117***	1.288**	2.073***	0.081***	0.103***
	(0.000)	(0.000)	(0.016)	(0.000)	(0.000)	(0.000)
Customer Characteristics	N	N	N	N	Y	Y
Digital Variables	N	N	Y	Y	Y	Y
Observations	144,103	147,223	147,223	144,103	147,223	144,103
Pseudo R-squared	0.00279	0.0361	0.0529	0.0521	0.0583	0.0577
AUC	0.584	0.644	0.662	0.660	0.671	0.671

Table 7: **Mobile footprint vs. financial transactions in predicting loan defaults**

This table reports the estimates from our logit regressions examining the relationship between Financial transactions, mobile/social footprint variables, customer characteristics, call logs and likelihood of default. The dependent variable, Default takes the value one for loans that are delinquent and zero otherwise. The specification in column (1) includes variables correponding to only Call logs and Digital Footprints. Column (2) includes Call logs, Digital footprints and credit bureau score. Column (3) includes the credit bureau score (Log of CIBIL) with call logs, Digital footprints, customer characteristics. Column (4) includes only the Financial Transactions. Column (5) includes Financial Transactions with call logs and mobile/social footprints. Column (6) includes Financial transactions, call logs, mobile/social footprints with credit bureau score. Column (7) includes Financial transactions, call logs, mobile/social footprint variables, credit bureau score along with customer characteristics. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)	Odds Ratio (4)	Odds Ratio (5)	Odds Ratio (6)	Odds Ratio (7)
Debit to credit ratio				1.026	1.021	1.024	1.067
				(0.571)	(0.671)	(0.614)	(0.176)
# of Transactions				1.141***	1.128***	1.129***	1.122***
				(0.000)	(0.002)	(0.002)	(0.003)
Expenditure to Income ratio				1.083*	1.073	1.074	1.055
				(0.099)	(0.182)	(0.177)	(0.305)
Avg 2 Month Appreciation in Balance				1.025	1.027	1.025	1.027
				(0.659)	(0.623)	(0.654)	(0.570)
Log of cibil		1.150**	1.143**			1.151**	1.145**
		(0.010)	(0.015)			(0.010)	(0.015)
Log of Salary			1.461***				1.493***
			(0.000)				(0.000)
Log Age			0.753				0.811
			(0.293)				(0.451)
High School Dummy			1.241				1.254
			(0.125)				(0.115)
College Dummy			1.365**				1.362**
			(0.039)				(0.045)
Supervisor Dummy			1.080				1.071
			(0.444)				(0.496)
Manager Dummy			1.169*				1.161
			(0.097)				(0.116)
Log no of SMS	0.998	1.001	1.000		0.992	0.995	0.994
	(0.903)	(0.963)	(0.979)		(0.664)	(0.794)	(0.764)
Log No of Contacts	1.003	0.996	0.976		1.001	0.993	0.971
	(0.931)	(0.914)	(0.532)		(0.987)	(0.856)	(0.462)
Log no of Apps	0.931	0.902	0.876*		0.934	0.904	0.878*
	(0.262)	(0.118)	(0.052)		(0.290)	(0.128)	(0.059)
Log Callog	0.946**	0.955*	0.962		0.945**	0.954*	0.962
	(0.048)	(0.099)	(0.164)		(0.043)	(0.090)	(0.166)
Dating App	1.456**	1.493**	1.482**		1.450**	1.486**	1.488**
	(0.029)	(0.020)	(0.024)		(0.031)	(0.022)	(0.023)
Finsavy App	0.976	1.004	1.024		0.998	1.027	1.048
	(0.891)	(0.980)	(0.895)		(0.990)	(0.886)	(0.803)
Socialconnect App	1.047	1.051	1.041		1.189	1.197	1.179
	(0.907)	(0.899)	(0.919)		(0.671)	(0.661)	(0.686)
Travel App	1.301**	1.305**	1.234*		1.224*	1.228*	1.167
	(0.014)	(0.013)	(0.052)		(0.062)	(0.059)	(0.159)
Mloan App	1.448***	1.473***	1.480***		1.417***	1.443***	1.457***
	(0.000)	(0.000)	(0.000)		(0.000)	(0.000)	(0.000)
Facebook status	1.104	1.113	1.077		1.096	1.104	1.066
	(0.205)	(0.175)	(0.348)		(0.248)	(0.215)	(0.426)
Linkedin status	0.736	0.735	0.649*		0.714	0.711	0.625*
	(0.219)	(0.217)	(0.085)		(0.189)	(0.183)	(0.068)
IOS Dummy	1.629***	1.555***	1.495**		1.558***	1.483**	1.440**
	(0.004)	(0.010)	(0.018)		(0.010)	(0.024)	(0.035)
Past days: Per day Per person Avg No. of Incoming calls	0.869	0.856	0.887		0.863	0.849	0.881
	(0.162)	(0.122)	(0.223)		(0.147)	(0.109)	(0.198)

Past days: Per day Per person Avg No. of Outgoing calls	1.039	1.056	1.056		1.065	1.087	1.098
	(0.634)	(0.490)	(0.476)		(0.477)	(0.345)	(0.283)
Past days: Per day Per person Avg No. of Missed calls	0.948	0.944	0.993		0.962	0.958	1.011
	(0.536)	(0.502)	(0.937)		(0.660)	(0.624)	(0.903)
Past days: Per day Per person Avg Duration of Incoming calls	0.049*	0.044**	0.068*		0.053**	0.047**	0.063**
	(0.062)	(0.037)	(0.052)		(0.034)	(0.021)	(0.035)
Past days: Per day Per person Avg Duration of Outgoing calls	0.384	0.372	0.382		0.406	0.395	0.428
	(0.250)	(0.245)	(0.261)		(0.281)	(0.278)	(0.322)
Past days: Per day No. of persons called	0.748*	0.762	0.786		0.807	0.828	0.868
	(0.087)	(0.109)	(0.145)		(0.234)	(0.297)	(0.429)
Past days: Log Per day Total Duration of Incoming calls	1.080	1.093	1.047		1.061	1.073	1.022
	(0.286)	(0.215)	(0.523)		(0.421)	(0.342)	(0.763)
Past days: Per day Total No. of Incoming calls	1.072	1.086	1.075		1.062	1.078	1.067
	(0.547)	(0.473)	(0.534)		(0.605)	(0.519)	(0.580)
Past days: Per day Total No. of Outgoing calls	1.235	1.205	1.237		1.168	1.133	1.150
	(0.198)	(0.262)	(0.193)		(0.363)	(0.471)	(0.411)
Past days: Per day Total Duration of Outgoing calls	1.483	1.569	1.427		1.581	1.677	1.512
	(0.428)	(0.379)	(0.483)		(0.350)	(0.306)	(0.409)
Past days: Per day Total No. of Missed calls	1.238***	1.218***	1.180**		1.212**	1.190**	1.146*
	(0.004)	(0.009)	(0.030)		(0.010)	(0.024)	(0.079)
Past days: Log HHI of No. of Incoming calls	1.031	1.053	1.088		1.067	1.092	1.129
	(0.740)	(0.580)	(0.361)		(0.487)	(0.357)	(0.194)
Past days: HHI of No. of Outgoing calls	0.684	0.706	0.742		0.660	0.685	0.731
	(0.470)	(0.507)	(0.571)		(0.443)	(0.483)	(0.560)
Past days: HHI of Total Duration of Incoming calls	0.985	0.999	0.999		1.014	1.028	1.036
	(0.907)	(0.995)	(0.996)		(0.913)	(0.824)	(0.780)
Past days: HHI of Total Duration of Outgoing calls	2.064*	2.009*	1.830		2.041*	1.983	1.795
	(0.085)	(0.097)	(0.153)		(0.091)	(0.105)	(0.168)
Past days: HHI of No. of Missed calls	1.189	1.175	1.085		1.159	1.145	1.050
	(0.217)	(0.251)	(0.576)		(0.271)	(0.317)	(0.726)
Constant	0.097***	0.042***	0.002***	0.078***	0.097***	0.042***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	11,991	11,866	11,855	12,316	11,566	11,443	11,432
Pseudo R2	0.0156	0.0171	0.0225	0.00247	0.0174	0.0191	0.0248
AUC	0.596	0.601	0.616	0.544	0.599	0.604	0.620

Table 8: **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) Digital Footprints. Panel A shows results using 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 shows results for the subsample of borrowers with credit score more than 750. Panel D shows results for subsample of borrowers with no CIBIL score. The specification in columns (1) and (2) report estimates from the Logistic regression model. Columns (3) and (4) report estimates from the Random Forest model and columns (5) and (6) report estimates from the XGBoost model.

Panel A: Performance Evaluation using customers with CIBIL

Feature Groups	Logistic Regression		Random Forest		XGBoost	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Only Cibil	0.130	0.586	0.857	0.578	0.499	0.579
Only Customer characteristics	0.216	0.539	0.787	0.740	0.498	0.536
Only Mobile/Social Footprint	0.476	0.603	0.735	0.716	0.613	0.608
Only Deep Social Footprint (Call logs)	0.623	0.665	0.790	0.771	0.621	0.689
Mobile + Deep social (Call logs)	0.605	0.669	0.770	0.740	0.664	0.697
Mobile + Deep social+Cibil	0.608	0.672	0.775	0.750	0.664	0.710
Mobile + Deep social+Cibil+ Customer characteristics	0.601	0.679	0.780	0.755	0.667	0.717

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

Feature Groups	Logistic Regression		Random Forest		XGBoost	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Only Cibil	0.180	0.521	0.807	0.550	0.470	0.518
Only Customer characteristics	0.407	0.547	0.746	0.747	0.493	0.568
Only Mobile/Social Footprint	0.470	0.602	0.721	0.724	0.593	0.621
Only Deep Social Footprint (Call logs)	0.630	0.662	0.754	0.767	0.639	0.696
Mobile + Deep social (Call logs)	0.616	0.671	0.745	0.746	0.672	0.712
Mobile + Deep social+Cibil	0.618	0.671	0.744	0.748	0.674	0.709
Mobile + Deep social+Cibil+ Customer characteristics	0.607	0.677	0.748	0.753	0.674	0.722

Panel C: Performance Evaluation using customers with CIBIL > 750

Feature Groups	Logistic Regression		Random Forest		XGBoost	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Only Cibil	0.100	0.552	0.850	0.528	0.695	0.540
Only Customer characteristics	0.233	0.486	0.815	0.775	0.632	0.621
Only Mobile/Social Footprint	0.526	0.644	0.796	0.720	0.777	0.635
Only Deep Social Footprint (Call logs)	0.660	0.668	0.820	0.758	0.706	0.706
Mobile + Deep social (Call logs)	0.678	0.684	0.830	0.729	0.774	0.719
Mobile + Deep social+Cibil	0.677	0.685	0.845	0.754	0.772	0.720
Mobile + Deep social+Cibil+ Customer characteristics	0.676	0.684	0.839	0.750	0.791	0.725

Panel D: Performance Evaluation using customers with no CIBIL

Feature Groups	Logistic Regression		Random Forest		XGBoost	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Only Customer characteristics	0.460	0.546	0.760	0.752	0.642	0.676
Only Mobile/Social Footprint	0.583	0.639	0.715	0.738	0.698	0.706
Only Deep Social Footprint (Call logs)	0.655	0.654	0.738	0.769	0.662	0.721
Mobile + Deep social (Call logs)	0.645	0.652	0.735	0.760	0.692	0.724

Table 9: **Predicting Defaults using Machine Learning (Sample with 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. The specification in columns (1) and (2) reports the estimates from the Logistic regression model. Columns (3) and (4) report estimates from the Random Forest model and columns (5) and (6) report estimates from the XGBoost model.

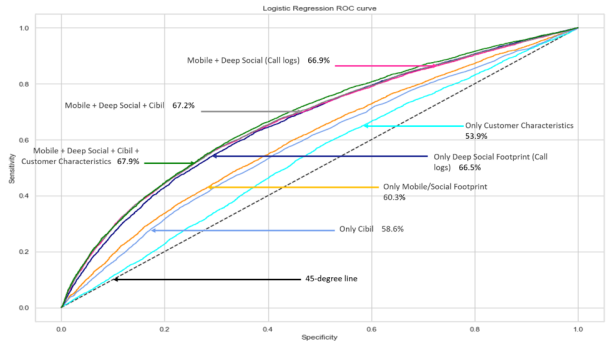
Feature Groups	Logistic Regression		Random Forest		XGBoost	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Only CIBIL	0.100	0.483	0.910	0.603	0.496	0.570
Only Cibil Score	0.100	0.483	0.910	0.603	0.496	0.570
Only Customer characteristics	0.261	0.584	0.746	0.712	0.445	0.575
Only Mobile/Social Footprint	0.201	0.529	0.817	0.721	0.765	0.530
Only Deep Social Footprint (Call logs)	0.076	0.492	0.770	0.744	0.569	0.620
Only Deep financial	0.217	0.576	0.767	0.587	0.711	0.520
Mobile + Deep social (Call logs)	0.076	0.492	0.805	0.766	0.773	0.591
Mobile + Deep social+Cibil	0.076	0.492	0.810	0.775	0.770	0.592
Mobile + Deep social+Cibil+ Customer characteristics	0.076	0.492	0.807	0.760	0.753	0.617
Deep financial + Cibil	0.217	0.576	0.766	0.594	0.712	0.527
Deep financial + Cibil +Customer characteristics	0.217	0.576	0.770	0.586	0.711	0.530

Table 10: **Policy Experiment: Alternate Credit Scoring**
This table reports results on the percentage of non-approved borrowers who would be approved for different chosen levels of the default score cut-offs. Panel A shows results using all borrowers. Panel B shows results using subsample of borrowers with CIBIL score less than 350.

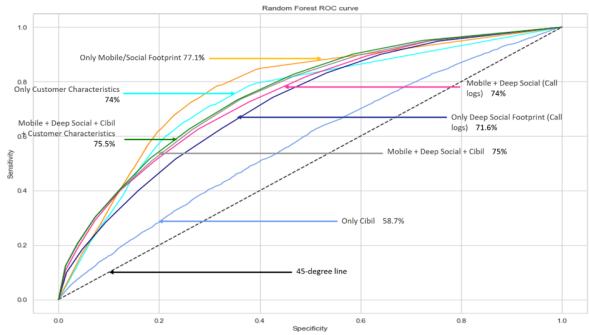
Panel A (Denied customers with CIBIL score)	
How many would have approved had we used these threshold of predicted default risk	
Predicted Default Threshold	What Proportion More Would Have Been Approved
0.95	0.949
0.9	0.948
0.8	0.903
0.7	0.848
0.6	0.799
0.5	0.747
0.4	0.693
0.3	0.626
0.2	0.537
0.1	0.421
0.05	0.227
0.01	0.226

Panel B (Denied customers without CIBIL score)	
How many would have approved had we used these threshold of predicted default risk	
Predicted Default Threshold	What Proportion More Would Have Been Approved
0.95	0.917
0.9	0.917
0.8	0.854
0.7	0.803
0.6	0.755
0.5	0.686
0.4	0.638
0.3	0.584
0.2	0.483
0.1	0.363
0.05	0.143
0.01	0.144

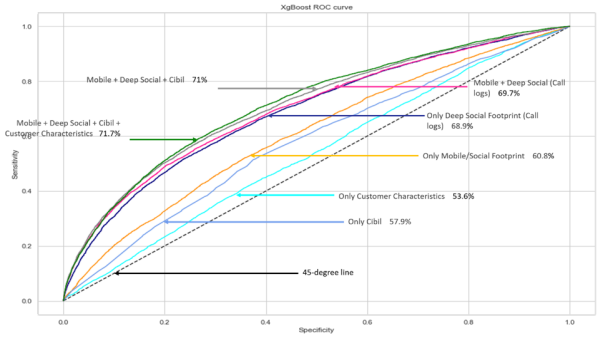
Figure 1: **AUC Plots for machine learning models (Full Sample)**
This figure plots the AUC curves based on three machine learning models: a) Logistic b) Random forest, and c) Xgboost.



(a) Logistic



(b) Random forest



(c) XGboost

Figure 2: **Variable importance factors**
This figure plots the variable importance factors in predicting defaults.

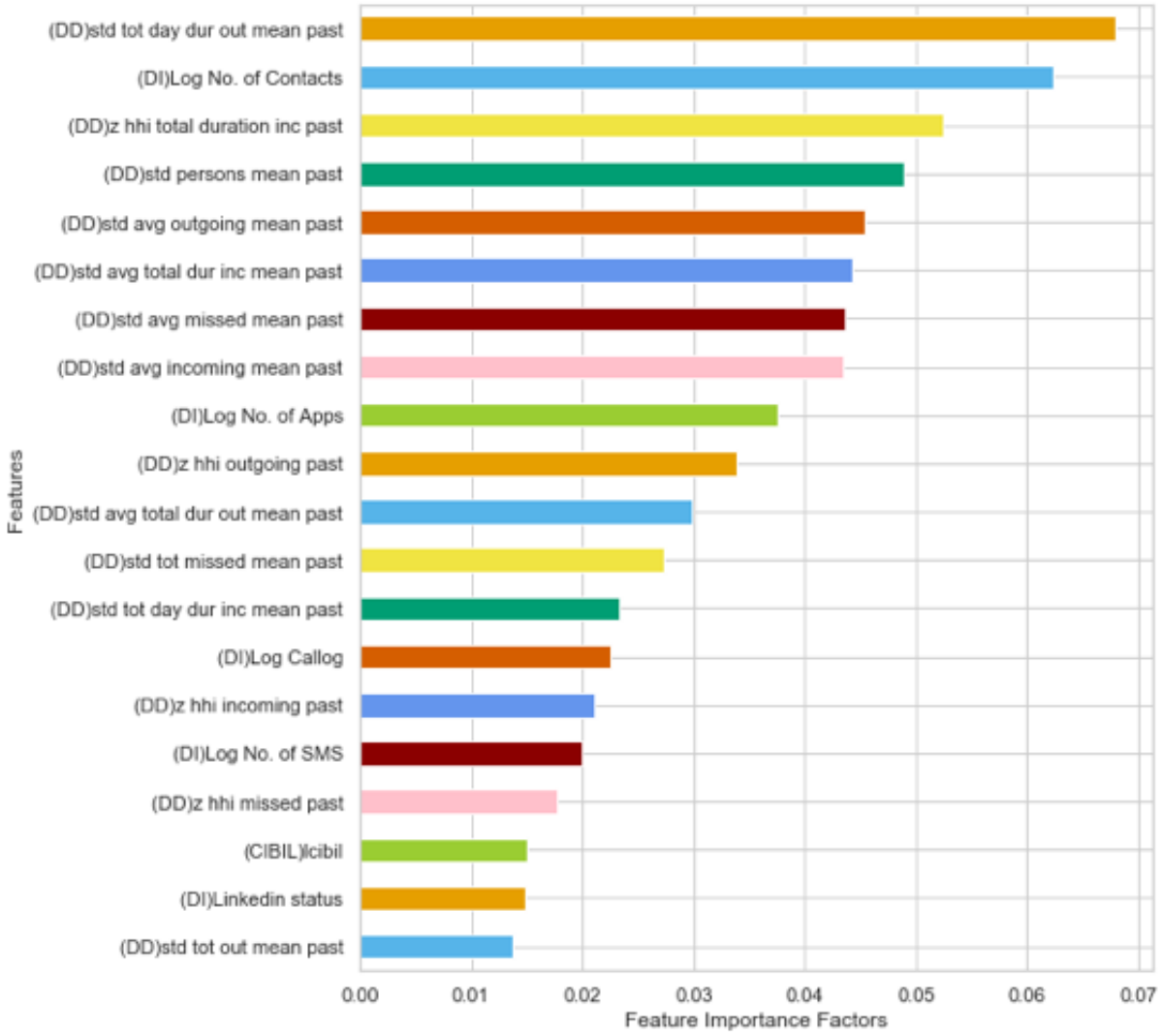
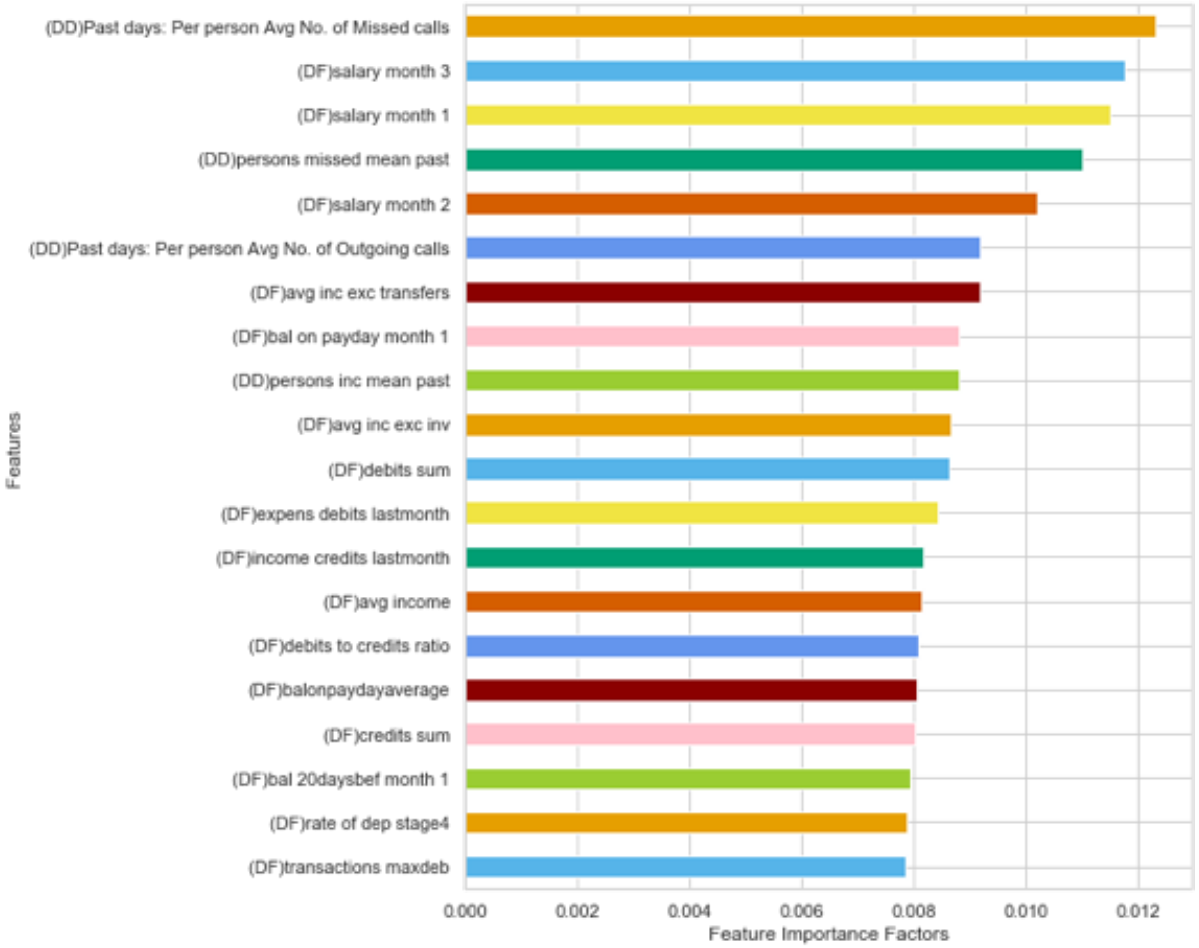


Figure 3: **Variable importance factors (When All Variables are Present)**



Appendix A

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

Table A1: **Approval of loans**

This table reports the estimates from our logit regressions examining the determinants of loan approval using the sample of observations with no credit bureau score available. The dependent variable, Approved takes the value one for loan applications that were approved and zero for those that were denied. The specification in column (1) includes customer characteristics. Column (2) includes the mobile/social footprint variables for the same sample. Column (3) includes customer characteristics with mobile/social footprint variables. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLESs	Odds Ratio	Odds Ratio	Odds Ratio
	(1)	(2)	(3)
Log of Salary	1.784***		1.554***
	(0.000)		(0.000)
Log Age	4.818***		6.739***
	(0.000)		(0.000)
High School Dummy	3.405***		3.038***
	(0.000)		(0.000)
College Dummy	4.421***		3.885***
	(0.000)		(0.000)
Supervisor	1.002		1.007
	(0.925)		(0.719)
Manager	0.954***		0.910***
	(0.005)		(0.000)
Log no of SMS		1.150***	1.155***
		(0.000)	(0.000)
Log No of Contacts		1.367***	1.214***
		(0.000)	(0.000)
Log no of Apps		1.585***	1.555***
		(0.000)	(0.000)
Log Callog		1.025***	1.034***
		(0.000)	(0.000)
Finsavy App		1.294	1.315
		(0.192)	(0.192)
Socialconnect App		26.518***	28.659***
		(0.000)	(0.000)
Travel App		1.740***	1.624***
		(0.000)	(0.000)
Mloan App		0.917	0.919
		(0.597)	(0.612)
Facebook status		0.649***	0.665***
		(0.000)	(0.000)
Linkedin status		0.829***	0.705***
		(0.001)	(0.000)
IOS Dummy		4.449***	3.964***
		(0.000)	(0.000)
Constant	0.000***	0.007***	0.000***
	(0.000)	(0.000)	(0.000)
Observations	98,498	97,727	97,689
Pseudo R2	0.0965	0.113	0.187
AUC	0.707	0.710	0.777

Table A2: **Default prediction: heterogeneity by credit score**

This table reports the estimates from our logit regressions examining the relationship between customer characteristics, mobile/social footprint variables and likelihood of default for customers in different terciles of the credit score distribution. 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 digital mobile footprint variables. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Default Regressions: Low CreditRating	Default Regressions: Medium CreditRating	Default Regressions: High CreditRating
	(1)	(2)	(3)
Log of Salary	1.500***	1.374***	1.393***
	(0.078)	(0.093)	(0.088)
Facebook status	0.985	1.220***	1.015
	(0.051)	(0.055)	(0.060)
Linkedin status	1.106	1.097	0.987
	(0.121)	(0.121)	(0.151)
Log of cibil	0.961***	0.005***	4.547**
	(0.011)	(0.005)	(3.440)
Log no of SMS	0.972***	0.963***	0.961***
	(0.010)	(0.010)	(0.010)
Log Age	0.615**	0.744**	0.477***
	(0.123)	(0.096)	(0.085)
Log No of Contacts	0.968	0.996	0.966
	(0.023)	(0.021)	(0.026)
Log no of Apps	0.675***	0.614***	0.635***
	(0.020)	(0.019)	(0.021)
Log Callog	0.910***	0.917***	0.951**
	(0.015)	(0.013)	(0.019)
Dating App	0.964	1.395***	1.437***
	(0.081)	(0.171)	(0.163)
Finsavy App	0.811***	0.655***	0.687***
	(0.060)	(0.064)	(0.072)
Socialconnect App	1.575***	1.888***	1.578***
	(0.157)	(0.253)	(0.197)
Travel App	0.955	0.977	0.938
	(0.042)	(0.059)	(0.059)
Mloan App	0.983	0.972	1.167**
	(0.037)	(0.046)	(0.075)
IOS Dummy	0.369***	0.600***	0.553***
	(0.047)	(0.069)	(0.074)
High School Dummy	0.985	0.853*	0.835**
	(0.083)	(0.071)	(0.063)
College Dummy	0.900	0.813***	0.718***
	(0.090)	(0.063)	(0.058)
Constant	0.194**	2.486e+14***	0.000**
	(0.143)	(1.572e+15)	(0.000)
Digital Variables	Y	Y	Y
Observations	62,276	66,377	52,176
Pseudo R-squared	0.0252	0.0289	0.0229
AUC	0.614	0.624	0.613

Table A3: **Default prediction: heterogeneity by age**

This table reports the estimates from our logit regressions examining the relationship between customer characteristics, mobile/social footprint variables and likelihood of default for customers in different terciles of the age distribution. 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 digital mobile footprint variables. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Default Regressions: Age 1Q	Default Regressions: Age 3Q
	(1)	(2)
Log of Salary	1.470***	1.324***
	(0.155)	(0.080)
Facebook status	1.105	1.002
	(0.089)	(0.049)
Linkedin status	0.940	1.120
	(0.158)	(0.149)
Log of cibil	0.926***	0.853***
	(0.020)	(0.025)
Log no of SMS	0.943***	0.973***
	(0.013)	(0.010)
Log Age	0.716	1.881**
	(0.288)	(0.574)
Log No of Contacts	1.055	0.969
	(0.038)	(0.027)
Log no of Apps	0.577***	0.668***
	(0.037)	(0.024)
Log Callog	0.934***	0.932***
	(0.020)	(0.015)
Dating App	1.015	1.256
	(0.074)	(0.246)
Finsavy App	0.711**	0.666***
	(0.114)	(0.055)
Socialconnect App	1.926***	1.528***
	(0.432)	(0.218)
Travel App	0.933	0.966
	(0.087)	(0.052)
Mloan App	1.116*	0.977
	(0.074)	(0.045)
IOS Dummy	0.255***	0.740*
	(0.071)	(0.114)
High School Dummy	0.949	0.918
	(0.077)	(0.054)
College Dummy	0.884	0.835**
	(0.102)	(0.060)
Constant	0.138	0.021***
	(0.255)	(0.019)
Digital Variables	Y	Y
Observations	20,853	60,847
Pseudo R-squared	0.0373	0.0284
AUC	0.636	0.615

Table A4: **Default prediction: heterogeneity by salary**

This table reports the estimates from our logit regressions examining the relationship between customer characteristics, mobile/social footprint variables and likelihood of default for customers in different terciles of the salary distribution. 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 digital mobile footprint variables. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Default Regressions: Salary 1Q	Default Regressions: Salary 3Q
	(1)	(2)
Log of Salary	0.332***	0.940
	(0.124)	(0.050)
Facebook status	1.111*	1.083
	(0.069)	(0.054)
Linkedin status	0.836	1.064
	(0.224)	(0.105)
Log of cibil	0.891***	0.867***
	(0.016)	(0.021)
Log no of SMS	0.961***	0.977**
	(0.013)	(0.011)
Log Age	0.455***	0.616***
	(0.104)	(0.098)
Log No of Contacts	0.989	1.002
	(0.041)	(0.027)
Log no of Apps	0.635***	0.628***
	(0.033)	(0.024)
Log Callog	0.900***	0.935***
	(0.028)	(0.014)
Dating App	1.277	1.127*
	(0.540)	(0.073)
Finsavy App	0.720**	0.912
	(0.098)	(0.081)
Socialconnect App	2.031***	1.426***
	(0.450)	(0.173)
Travel App	0.936	0.915
	(0.064)	(0.065)
Mloan App	0.897*	1.116***
	(0.057)	(0.043)
IOS Dummy	0.189***	0.658***
	(0.050)	(0.049)
High School Dummy	0.844*	0.976
	(0.074)	(0.078)
College Dummy	0.779**	0.866*
	(0.085)	(0.067)
Constant	2145188.988***	35.875***
	(8679837.677)	(19.994)
Digital Variables	Y	Y
Observations	26,744	51,227
Pseudo R-squared	0.0422	0.0228
AUC	0.647	0.608

Table A5: **Default prediction: heterogeneity by job designation**

This table reports the estimates from our logit regressions examining the relationship between customer characteristics, mobile/social footprint variables and likelihood of default for customers in three different employment category: workers, supervisors, and managers. 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 digital mobile footprint variables. Standard errors are clustered at the state level. $(***)$, $(**)$, $(*)$ denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Default Regressions: Worker	Default Regressions: Supervisor	Default Regressions: Manager
	(1)	(2)	(3)
Log of Salary	1.422***	1.496***	1.425***
	(0.092)	(0.079)	(0.077)
Facebook status	1.101	1.079	1.078*
	(0.064)	(0.059)	(0.048)
Linkedin status	1.116	1.034	1.142
	(0.167)	(0.151)	(0.114)
Log of cibil	0.912***	0.904***	0.894***
	(0.020)	(0.019)	(0.015)
Log no of SMS	0.956***	0.959***	0.984
	(0.009)	(0.013)	(0.011)
Log Age	0.654**	0.713	0.512***
	(0.134)	(0.216)	(0.054)
Log No of Contacts	1.002	0.951	0.967
	(0.023)	(0.034)	(0.024)
Log no of Apps	0.651***	0.667***	0.599***
	(0.020)	(0.027)	(0.021)
Log Callog	0.937***	0.902***	0.923***
	(0.012)	(0.018)	(0.014)
Dating App	1.157*	1.429**	1.156
	(0.087)	(0.210)	(0.161)
Finsavy App	0.665***	0.633***	0.857**
	(0.063)	(0.093)	(0.064)
Socialconnect App	1.844***	1.834***	1.341**
	(0.224)	(0.265)	(0.164)
Travel App	0.908*	0.980	1.075
	(0.045)	(0.072)	(0.049)
Mloan App	1.031	0.956	1.056
	(0.046)	(0.036)	(0.048)
IOS Dummy	0.391***	0.433***	0.632***
	(0.057)	(0.111)	(0.060)
High School Dummy	0.830**	0.915	1.062
	(0.073)	(0.063)	(0.089)
College Dummy	0.723***	0.835**	0.955
	(0.070)	(0.071)	(0.086)
Constant	0.265	0.196*	0.856
	(0.256)	(0.165)	(0.563)
Digital Variables	Y	Y	Y
Observations	64,879	44,457	71,493
Pseudo R-squared	0.0285	0.0291	0.0260
AUC	0.625	0.620	0.616

Table A6: Average and total call logs metrics

The below metrics have been calculated for every customer in the Database. We divide the above metrics into ex-ante (all days before start date of loan) and ex-post (first 15 days after start date of loan) call logs. $C_{i,j}$ is the total number of calls made to the person i on j^{th} day. k is the number of contacts in the customer's contact list and n_j is the number of persons contacted on the j^{th} day.

Metric	Formula
First 15 day Average: Per day Per person	
First 15 days: Per day Per person Avg No. of Incoming calls	$\frac{\sum_{j=1}^{j=15} \frac{\sum_{i=1}^{n_j} C_{i,j}}{n_j}}{15}$
First 15 days: Per day Per person Avg No. of Outgoing calls	
First 15 days: Per day Per person Avg No. of Missed calls	
First 15 days: Per day Per person Avg Duration of Incoming calls	
First 15 days: Per day Per person Avg Duration of Outgoing calls	
Past History Average: Per day Per person	
Past days: Per day Per person Avg No. of Incoming calls	$\frac{\sum_{\forall j \leq 0} \frac{\sum_{i=1}^{n_j} C_{i,j}}{n_j}}{\sum_{\forall j \leq 0} 1}$
Past days: Per day Per person Avg No. of Outgoing calls	
Past days: Per day Per person Avg No. of Missed calls	
Past days: Per day Per person Avg Duration of Incoming calls	
Past days: Per day Per person Avg Duration of Outgoing calls	
First 15 day Average: Per day	
First 15 days: Per day No. of persons called	$\frac{\sum_{j=1}^{j=15} n_j}{15}$
First 15 days: Per day Total Duration of Incoming calls	$\frac{\sum_{j=1}^{j=15} \sum_{i=1}^{n_j} C_{i,j}}{15}$
First 15 days: Per day Total No. of Incoming calls	
First 15 days: Per day Total No. of Outgoing calls	
First 15 days: Per day Total Duration of Outgoing calls	
First 15 days: Per day Total No. of Missed calls	
Past History Average: Per day	
Past days: Per day No. of persons called	$\frac{\sum_{\forall j \leq 0} n_j}{\sum_{\forall j \leq 0} 1}$
Past days: Per day Total Duration of Incoming calls	$\frac{\sum_{\forall j \leq 0} \sum_{i=1}^{n_j} C_{i,j}}{\sum_{\forall j \leq 0} 1}$
Past days: Per day Total No. of Incoming calls	
Past days: Per day Total No. of Outgoing calls	
Past days: Per day Total Duration of Outgoing calls	
Past days: Per day Total No. of Missed calls	

Table A6: Herfindahl-Hirschman call log index

The below metrics have been calculated for every customer in the Database. We divide the above metrics into ex-ante (all days before start date of loan) and ex-post (first 15 days after start date of loan) call logs. $C_{i,j}$ is the total number of calls made to the person i on j^{th} day. k is the number of contacts in the customer's contact list and n_j is the number of persons contacted on the j^{th} day.

Metric	**Formula**
First 15 days: Herfindahl-Hirschman Index	
First 15 days: HHI of No. of Incoming calls	$\sum_{i=1}^{i=k}\left[\frac{\frac{\sum_{j=1}^{j=15} C_{i,j}}{15}}{\frac{\sum_{j=1}^{j=15} C_{i,j}}{15}} \times 100\right]^2$
First 15 days: HHI of No. of Outgoing calls	
First 15 days: HHI of Total Duration of Incoming calls	
First 15 days: HHI of Total Duration of Outgoing calls	
First 15 days: HHI of No. of Missed calls	
Past History : Herfindahl-Hirschman Index	
Past days: HHI of No. of Incoming calls	$\sum_{i=1}^{i=k}\left[\frac{\frac{\sum_{\forall j\leq 0} C_{i,j}}{\sum_{\forall j\leq 0} 1}}{\frac{\sum_{\forall j\leq 0} C_{i,j}}{\sum_{\forall j\leq 0} 1}} \times 100\right]^2$
Past days: HHI of No. of Outgoing calls	
Past days: HHI of Total Duration of Incoming calls	
Past days: HHI of Total Duration of Outgoing calls	
Past days: HHI of No. of Missed calls	

Table A7: **Predicting loan defaults using deep social footprint based on call logs (with ex-post and ex-ante measures)**
This table reports the estimates from our logit regressions examining the relationship between customer characteristics, mobile/social footprint variables and call logs and likelihood of default. The dependent variable, Default takes the value one for loans that are delinquent and zero otherwise. The specification in column (1) only includes the credit bureau score (Log of CIBIL) as the explanatory variable. Column (2) includes credit bureau score with customer characteristics. Column (3) includes only call log variables. Column (4) includes call logs with credit bureau score. Column (5) includes call logs with customer characteristics. Column (6) includes call logs and mobile/social footprint variables. Column (7) includes all call variables, customer characteristics and mobile/social footprint variables but not the CIBIL score. Column (8) includes all variables including the CIBIL score. Standard errors are clustered at the state level. (∗ ∗ ∗), (∗∗), (∗) denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)	Odds Ratio (4)	Odds Ratio (5)	Odds Ratio (6)	Odds Ratio (7)	Odds Ratio (8)
Log of cibil	0.897***	0.894***		0.910***				0.913***
	(0.000)	(0.000)		(0.000)				(0.000)
First 15 days: Per day Per person Avg No. of Incoming calls			1.024	1.023	1.028*	1.023	1.026*	1.025
			(0.112)	(0.137)	(0.071)	(0.123)	(0.086)	(0.103)
First 15 days: Per day Per person Avg No. of Outgoing calls			1.009	1.004	1.008	1.005	1.004	1.000
			(0.567)	(0.792)	(0.619)	(0.763)	(0.800)	(0.985)
First 15 days: Per day Per person Avg No. of Missed calls			0.975*	0.973**	0.978*	0.972**	0.976*	0.975*
			(0.052)	(0.044)	(0.091)	(0.037)	(0.071)	(0.065)
First 15 days: Per day Per person Avg Duration of Incoming calls			0.937***	0.941***	0.934***	0.934***	0.930***	0.935***
			(0.005)	(0.010)	(0.004)	(0.003)	(0.002)	(0.005)
First 15 days: Per day Per person Avg Duration of Outgoing calls			0.953**	0.962*	0.951**	0.964*	0.963*	0.971
			(0.036)	(0.087)	(0.035)	(0.073)	(0.074)	(0.159)
First 15 days: Per day No. of persons called			0.973	0.970	0.979	0.987	0.995	0.989
			(0.372)	(0.324)	(0.480)	(0.671)	(0.871)	(0.721)
Past days: Per day Per person Avg No. of Incoming calls			1.048***	1.043**	1.049***	1.056***	1.057***	1.054***
			(0.004)	(0.010)	(0.003)	(0.001)	(0.001)	(0.001)
Past days: Per day Per person Avg No. of Outgoing calls			1.028*	1.026*	1.026	1.027*	1.023	1.021
			(0.071)	(0.100)	(0.109)	(0.075)	(0.134)	(0.169)
Past days: Per day Per person Avg No. of Missed calls			0.962***	0.961***	0.965***	0.959***	0.964***	0.963***
			(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Past days: Per day Per person Avg Duration of Incoming calls			0.000	0.000	0.000	0.000	0.000	0.000
			(0.312)	(0.278)	(0.435)	(0.195)	(0.315)	(0.284)
Past days: Per day Per person Avg Duration of Outgoing calls			1.045	1.037	1.040	1.025	1.026	1.018
			(0.637)	(0.652)	(0.710)	(0.709)	(0.727)	(0.787)
Past days: Per day No. of persons called			0.884***	0.894***	0.885***	0.926**	0.926**	0.933*
			(0.000)	(0.002)	(0.001)	(0.027)	(0.030)	(0.051)
First 15 days: Per day Total Duration of Incoming calls			1.007	1.001	1.009	1.020	1.023	1.016
			(0.782)	(0.970)	(0.725)	(0.440)	(0.369)	(0.527)
First 15 days: Per day Total No. of Incoming calls			1.016	1.020	1.013	1.004	1.000	1.003
			(0.571)	(0.497)	(0.657)	(0.885)	(0.994)	(0.911)
First 15 days: Per day Total No. of Outgoing calls			1.102***	1.112***	1.105***	1.092***	1.094***	1.104***
			(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
First 15 days: Per day Total Duration of Outgoing calls			1.024	1.009	1.023	1.016	1.014	1.000
			(0.309)	(0.707)	(0.339)	(0.472)	(0.538)	(0.994)
First 15 days: Per day Total No. of Missed calls			1.082***	1.082***	1.080***	1.075***	1.072***	1.072***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: Per day Total Duration of Incoming calls			0.003	0.005	0.000*	0.014	0.000	0.001
			(0.203)	(0.254)	(0.071)	(0.352)	(0.101)	(0.110)
Past days: Per day Total No. of Incoming calls			1.011	1.017	1.017	1.015	1.025	1.032
			(0.732)	(0.605)	(0.590)	(0.633)	(0.437)	(0.325)
Past days: Per day Total No. of Outgoing calls			1.249***	1.213***	1.269***	1.190***	1.212***	1.181***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: Per day Total Duration of Outgoing calls			0.747***	0.802***	0.732***	0.808***	0.781***	0.833**
			(0.000)	(0.003)	(0.000)	(0.003)	(0.001)	(0.012)
Past days: Per day Total No. of Missed calls			1.297***	1.300***	1.295***	1.280***	1.275***	1.278***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
First 15 days: HHI of No. of Incoming calls			1.007	0.971	1.010	1.009	1.017	0.981
			(0.771)	(0.246)	(0.690)	(0.734)	(0.499)	(0.454)
First 15 days: HHI of No. of Outgoing calls			1.001	1.002	1.000	1.003	1.002	1.003
			(0.981)	(0.923)	(0.987)	(0.896)	(0.937)	(0.890)

First 15 days: HHI of Total Duration of Incoming calls			1.039	1.044*	1.039	1.023	1.018	1.024
			(0.105)	(0.072)	(0.112)	(0.333)	(0.465)	(0.336)
First 15 days: HHI of Total Duration of Outgoing calls			1.549***	1.633***	1.520***	1.517***	1.477***	1.553***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
First 15 days: HHI of No. of Missed calls			1.065***	1.072***	1.068***	1.053***	1.057***	1.064***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: HHI of No. of Incoming calls			0.868***	0.870***	0.874***	0.868***	0.876***	0.878***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: HHI of No. of Outgoing calls			1.055	1.062	1.049	1.052	1.045	1.050
			(0.426)	(0.366)	(0.475)	(0.448)	(0.514)	(0.463)
Past days: HHI of Total Duration of Incoming calls			1.270***	1.272***	1.266***	1.259***	1.251***	1.253***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Past days: HHI of Total Duration of Outgoing calls			1.121	1.098	1.116	1.096	1.088	1.070
			(0.356)	(0.460)	(0.382)	(0.473)	(0.520)	(0.607)
Past days: HHI of No. of Missed calls			1.049***	1.043***	1.050***	1.041***	1.042***	1.037**
			(0.001)	(0.004)	(0.001)	(0.006)	(0.005)	(0.013)
Constant	0.368***	0.230***	0.163***	0.291***	0.024***	1.360**	0.188***	0.244***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.022)	(0.000)	(0.000)
Customer Characteristics	N	Y	N	N	Y	N	Y	Y
Digital Variables	N	N	N	N	N	Y	Y	Y
Observations	97,062	97,062	97,043	94,889	97,043	97,043	97,043	94,889
Pseudo R-squared	0.00257	0.00478	0.0404	0.0405	0.0430	0.0546	0.0591	0.0583
AUC	0.587	0.547	0.648	0.648	0.652	0.664	0.671	0.671

Table A8: **Predicting Loan Defaults with deep social footprint based on call logs (Subsample without credit score)**
This table reports the estimates from our logit regressions examining the relationship between customer characteristics, mobile/social footprint variables and likelihood of default using the sample of observations with no credit bureau score available. The dependent variable, Default takes the value one for loans that are delinquent and zero otherwise. The specification in column (1) includes only customer characteristics. Column (2) includes the mobile/social footprint variables for the same sample. Column (3) includes mobile/social footprint variables with call log variables. Column (4) includes customer characteristics with mobile/social footprint variables and call logs. Standard errors are clustered at the state level. (∗∗∗), (∗∗), (∗) denote statistical significance at 1%, 5%, and 10% levels respectively.

VARIABLES	Odds Ratio (1)	Odds Ratio (2)	Odds Ratio (3)	Odds Ratio (4)
Log of Salary	1.595***			2.005***
	(0.000)			(0.000)
Log Age	2.299***			1.005
	(0.005)			(0.988)
High School Dummy	0.525***			0.681***
	(0.000)			(0.009)
College Dummy	0.363***			0.441***
	(0.000)			(0.000)
Supervisor	1.120			0.986
	(0.304)			(0.912)
Manager	1.206*			1.270*
	(0.079)			(0.063)
Log no of SMS		0.898***	0.892***	0.880***
		(0.000)	(0.000)	(0.000)
Log No of Contacts		0.818***	0.815***	0.829***
		(0.001)	(0.002)	(0.005)
Log no of Apps		0.714***	0.715***	0.706***
		(0.000)	(0.000)	(0.000)
Log Callog		0.920**	0.984	0.999
		(0.048)	(0.725)	(0.988)
Finsavy App		0.281***	0.325***	0.349***
		(0.000)	(0.000)	(0.000)
Socialconnect App		0.767	0.770	0.765
		(0.445)	(0.457)	(0.486)
Travel App		1.156	1.233*	1.151
		(0.173)	(0.062)	(0.221)
Mloan App		0.858*	0.860	0.850*
		(0.091)	(0.114)	(0.092)
Facebook status		1.100	1.200	1.240*
		(0.401)	(0.124)	(0.075)
Linkedin status		1.034	1.222	0.861
		(0.913)	(0.543)	(0.654)
IOS Dummy		0.489	0.498*	0.412**
		(0.161)	(0.091)	(0.036)
Past days: Per day Per person Avg No. of Incoming calls			1.093	1.094
			(0.162)	(0.162)
Past days: Per day Per person Avg No. of Outgoing calls			1.084	1.081
			(0.204)	(0.283)
Past days: Per day Per person Avg No. of Missed calls			1.004	1.042
			(0.964)	(0.660)
Past days: Per day Per person Avg Duration of Incoming calls			0.000	0.000
			(0.313)	(0.372)
Past days: Per day Per person Avg Duration of Outgoing calls			0.336	0.301
			(0.226)	(0.185)
Past days: Per day No. of persons called			0.790	0.768*
			(0.117)	(0.098)
Past days: Per day Total Duration of Incoming calls			148,462.791	6.392
			(0.653)	(0.943)
Past days: Per day Total No. of Incoming calls			0.994	1.013
			(0.964)	(0.917)
Past days: Per day Total No. of Outgoing calls			1.545***	1.600***

			(0.001)	(0.000)
Past days: Per day Total Duration of Outgoing calls			0.358*	0.380*
			(0.050)	(0.068)
Past days: Per day Total No. of Missed calls			1.303***	1.267***
			(0.000)	(0.001)
Past days: HHI of No. of Incoming calls			0.913	0.897
			(0.495)	(0.385)
Past days: HHI of No. of Outgoing calls			0.787	0.717
			(0.223)	(0.115)
Past days: HHI of Total Duration of Incoming calls			1.206	1.232*
			(0.124)	(0.078)
Past days: HHI of Total Duration of Outgoing calls			1.593**	1.670**
			(0.027)	(0.015)
Past days: HHI of No. of Missed calls			1.103	1.094
			(0.449)	(0.432)
Constant	0.000***	44.515***	14.196***	0.015**
	(0.000)	(0.000)	(0.000)	(0.010)
Observations	3,255	3,143	3,121	3,120
Pseudo R-squared	0.0256	0.0846	0.139	0.157
AUC	0.616	0.685	0.740	0.753

Table A9: **Summary Statistics of Call logs for customers without cibil score (for customers with call log data)**

This table reports summary statistics on call logs for loans granted without cibil score for the set of customers with call log data. Columns 1-3 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.

	Default (1)	Not Default (2)	Difference (3)
Past days: Per day Per person Avg No. of Incoming calls	1.68	1.60	-0.08***
Past days: Per day Per person Avg No. of Outgoing calls	2.44	2.27	-0.17***
Past days: Per day Per person Avg No. of Missed calls	1.71	1.58	-0.12***
Past days: Per day Per person Avg Duration of Incoming calls	175.80	178.76	2.96
Past days: Per day Per person Avg Duration of Outgoing calls	163.46	189.37	25.90***
Past days: Per day No. of persons called	15.38	14.56	-0.82***
Past days: Per day Total No. of Incoming calls	11.94	10.68	-1.25***
Past days: Per day Total No. of Outgoing calls	25.15	21.67	-3.47***
Past days: Per day Total No. of Missed calls	8.36	6.65	-1.70
Past days: Per day Total Duration of Incoming calls	1111.35	1058.79	-52.56**
Past days: Per day Total Duration of Outgoing calls	1524.39	1572.14	47.74***
Past days: HHI of No. of Incoming calls	314.51	172.04	-142.46***
Past days: HHI of No. of Outgoing calls	361.49	169.07	-192.42***
Past days: HHI of Total Duration of Incoming calls	631.72	395.94	-235.78***
Past days: HHI of Total Duration of Outgoing calls	772.78	415.09	-357.69***
Past days:HHI of No. of Missed calls	483.52	230.91	-252.61***
N	776	2,595	

Appendix B: Details of the Machine Learning Estimation

In this section we describe in detail different steps for the machine learning estimation procedures. The estimation of default for the entire data set used default as a dummy variable (= 1 for loans which defaulted and 0 otherwise) and estimated the probability of default using various machine learning based classification algorithm (logistic regression, random forest and XGBoost). The algorithm uses different set of variables (feature vectors) under the sub-category customer characteristics, CIBIL score, digital variables etc. as outlined in detail in the main text of the paper as well as in the appendix.

The first issue of the estimation involves balancing the data. As described in the summary table in the paper, the proportion of defaults in the dataset is far smaller than the proportion of loans which did not default. A machine learning prediction algorithms in such situation are most likely to predict not -default in the out of sample prediction (testing sample). To avoid these kind of situation, it is advisable to balance the data in the training sample to get a more balanced set and even representation of the default sub population. There are various ways to deal with the unbalanced data issues like under sampling the majority (non-dafault) group, over sampling the minority (default group) or generating synthetic data from the minority class (SMOTE)[27]. In our analysis we have used SMOTE followed by Edited Nearest Neighbor (ENN) to deal with the unbalanced data problem.[28] The before and after data sizes to deal with the unbalanced data problem can be seen from the following graph.

While we use balancing for the training dataset for estimation purpose, 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 outcome. Following standard practice we use a five-fold cross validation procedure for each machine learning estimation procedure. Throughout the paper we use 70% training and 30% of the data for out of sample prediction.[29].

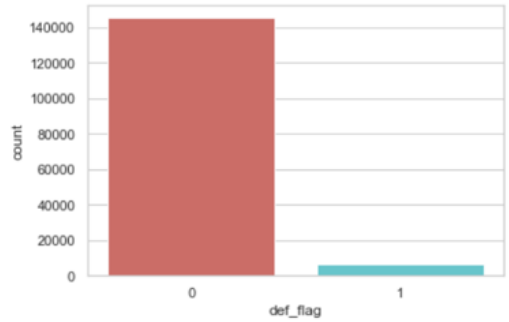
[27] see Nitesh V Chawla et al (2002): Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357 for a detail description.

Also see "Survey of resampling techniques for improving classification performance in unbalanced datasets" by More (University of Michigan) for a variety of techniques to deal with unbalanced data.

[28] Our results are robust to various other techniques to deal with unbalanced data .

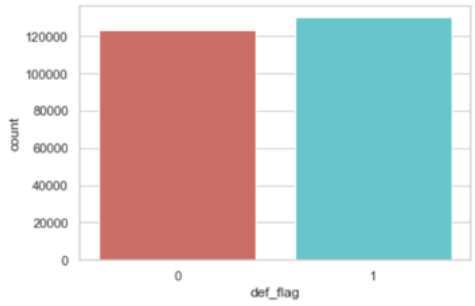
[29] Our results are qualitatively robust to a 80-20 split.

Figure B1: **Balancing of Data (representative graph)**




```
There are 151890 records in the dataset
There are 6578 records make def_flag 1 in the dataset
There are 145312 records make def_flag 0 in the dataset
```

(a) Before Data Balancing




```
There are 253828 records in the dataset
There are 130160 records make def_flag 1 in the dataset
There are 123668 records make def_flag 0 in the dataset
```

(b) After Data Balancing

Appendix C: Variable Definitions

Table C1: **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.
	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.
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 C2: **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.