Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment

Daniel Björkegren[1] and Darrell Grissen[2]

ABSTRACT

Many households in developing countries lack formal financial histories, making it difficult for firms to extend credit, and for potential borrowers to receive it. However, many of these households have mobile phones, which generate rich data about behavior. This paper shows that behavioral signatures in mobile phone data predict default, using call records matched to repayment outcomes for credit extended by a South American telecom. On a sample of individuals with (thin) financial histories, our method actually outperforms models using credit bureau information, both within time and when tested on a different time period. But our method also attains similar performance on those without financial histories, who cannot be scored using traditional methods. Individuals in the highest quintile of risk by our measure are 2.8 times more likely to default than those in the lowest quintile. The method forms the basis for new forms of credit that reach the unbanked.

This draft: October 21, 2018. First draft: January 6, 2015.

ACKNOWLEDGEMENTS

Thanks to Entrepreneurial Finance Lab and partners for data, and Jeff Berens, Nathan Eagle, Javier Frassetto, and Seema Jayachandran for helpful discussions. Darrell Grissen is a former employee of the Entrepreneurial Finance Lab, and was employed during data collection. We appreciate the feedback of audiences at the AEA Annual Meetings, NEUDC, Microsoft Research, and the AMID/BREAD Summer School in Development Economics. Daniel gratefully acknowledges the support of the W. Glenn Campbell and Rita Ricardo-Campbell National Fellowship at Stanford University.

[1] Brown University, Department of Economics. Box B, Providence, RI 02912. E-mail: danbjork@brown.edu, Web: http://dan.bjorkegren.com, Phone: 650.720.5615. (corresponding author)

[2] Independent. E-mail: dgrissen@gmail.com.

1. INTRODUCTION

Many studies have found that households and firms in developing countries have access to opportunities with high returns that, puzzlingly, remain untapped (De Mel, McKenzie, & Woodruff, 2008; McKenzie & Woodruff, 2008; A. V. Banerjee & Duflo, 2014). Part of the reason appears to be that individuals lack access credit which would make it possible to tap opportunities and smooth frictions.

Traditional approaches to improve access to credit in the developing world have done so by extending general forms physically, by replicating institutions from wealthier societies such as bank branches and credit bureaus (de Janvry, McIntosh, & Sadoulet, 2010; Luoto, McIntosh, & Wydick, 2007; World Bank, 2014), or tapping informal networks, such as peer screening in microfinance. However, it is costly to physically provide small amounts of credit, especially to remote populations. Many remain unserved: 2 billion people lack bank accounts (Demirguc-Kunt, Klapper, Singer, & Van Oudheusden, 2014). And those with access do not appear to be served particularly well by the products currently available: current microfinance models do not appear to have led to transformative effects for borrowers (A. Banerjee, Duflo, Kinnan, & Glennerster, 2014; A. Banerjee, Karlan, & Zinman, 2015; Karlan & Zinman, 2011).

However, physical interaction may no longer be required: there are over 4.5 billion mobile phone accounts in developing countries (ITU, 2011). Mobile phone networks can be used to extend credit remotely. However, it is difficult to know who would repay credit, particularly when given remotely, and few poor households have the formal records needed for traditional credit scores. However, many have maintained a rich history of interaction with a formal institution over an extended period of time—their mobile phone activity, recorded by their operator.[3] This paper proposes and demonstrates that use of a phone itself generates information that predicts who is likely to repay credit.

[3] Nonpayment is particularly problematic in developing societies, as creditors have little recourse if a borrower were to default: borrowers have little in the way of collateral, and systems for legal enforcement are limited.

We develop and assesses a low cost method to predict repayment of credit using mobile phone metadata, which are already being collected by mobile phone networks. From raw phone transaction records we extract signals plausibly related to repayment, and use a machine learning approach to combine these signals into a prediction of repayment. We show that using only minor tweaks to off-the-shelf machine learning methods, our approach achieves performance comparable to that of traditional credit bureau models, including for unbanked individuals who cannot be scored by traditional methods.

Our method consumes raw operator transaction records, which are already being collected at close to zero cost. These records can yield rich information about individuals, including mobility, consumption, and social networks (Blumenstock, Cadamuro, & On, 2015; Gonzalez, Hidalgo, & Barabasi, 2008; Lu, Wetter, Bharti, Tatem, & Bengtsson, 2013; Onnela et al., 2007; Palla, Barabási, & Vicsek, 2007; Soto, Frias-Martinez, Virseda, & Frias-Martinez, 2011). This paper shows how indicators derived from this data can predict the repayment of credit. Since this paper's proposal (Björkegren, 2010) and working paper (Bjorkegren & Grissen, 2015) were posted, this idea has received substantial attention (NPR, 2015), and is now widely used.

There are many straightforward indicators of behavior that are plausibly related to repayment of credit. For example, a responsible borrower may carefully manage their balance over time so usage is more smooth. An individual whose usage repeats on a monthly cycle may be more likely to have a salaried income. Or, an individual whose calls to others are returned may have stronger social connections that allow them to better follow through on entrepreneurial opportunities.

From raw transaction records, we extract approximately 5,500 behavioral indicators. Determining which indicators to extract is a crucial determinant of the performance of machine learning predictions ('applied machine learning is basically feature engineering' (Ng, 2011)). A brute force data mining approach such as Blumenstock et al. (2015) would algorithmically extract indicators while being agnostic towards the outcome variable. However, such approaches can pick up spurious correlations that make them unreliable in practice (Lazer, Kennedy, King, & Vespignani, 2014). With infinite data and variation, a machine learning approach would drop spurious variables on its own; but when data and variation are finite, focusing on features with intuitive or theoretical links to the outcome of interest can improve stability. Intuition can also suggest more nuanced indicators which a brute force method would neglect. Like the working paper version of this paper (Bjorkegren & Grissen, 2015), our approach generates indicators intuitively linked to repayment, which may pick up how consumers manage usage over time, social connections, and consumers' potential capacity to repay. Indicators that have an intuitive link are also more palatable to implementation partners who can be wary of 'black box' methods.

We show that the behavioral indicators we derive can reduce the uncertainty around a person's type. They can derisk interactions between large firms and the poor, at scale. The largest promise of these new technologies is not to duplicate the familiar loan offerings that evolved under the constraints of physical interaction, but to enable new types of formal lending that would not be feasible under historical constraints.

We focus on a credit problem faced by developing country mobile telecoms, which currently have the best access to the necessary data. Of the over 4.5 billion mobile phone accounts in developing countries, most are prepaid. Mobile phone consumption represents 2% of consumption in the South American country we focus on. While some consumers value the nuanced control of spending that prepaid plans afford, they entail substantial frictions: consumers must carefully manage and regularly top up balances, and telecoms must maintain an extensive physical distribution network for small scratch cards. Consumers top up phones frequently: in a sample of Caribbean mobile phone users, 27% topped up at least every other day. As incomes rise, consumers are expected to transition to postpaid accounts. The transition to a postpaid account entails an extension of credit, and thus exposes the telecom to risk. Few of these individuals have formal financial histories which would allow them to undergo the credit checks which a developed country telecom would require. This lending problem has features analogous to other credit decisions where our method could be applied, within telecom (top up advances and loans to purchase handsets), in other sectors (pay as you go solar systems), and for general digital credit sent over mobile money.

We demonstrate the method with data from a telecom in a middle income South American country that is transitioning subscribers from prepaid to postpaid plans. In this country, only 34% of adults have bank accounts but 89% of households have mobile phones. Our setting has two crucial features. First, in the exploratory phase we observe, the telecom extended credit permissively, selecting subscribers with sufficient usage from across the distribution of credit histories (including no history at all). As a result, we observe outcomes from a distribution of individuals who the telecom might conceivably wish to transition to a postpaid plan, and can evaluate the performance of any screening rule. Second, our sample includes both banked and unbanked consumers, which allows us to both benchmark our performance against credit bureau models, and also evaluate whether performance is comparable for individuals without bureau records. We observe each applicant's mobile phone transaction history prior to the extension of credit, and whether the credit was repaid on time. We predict who among these individuals ended up repaying, based on how they used their mobile phones prior to the switch, in a retrospective analysis. Our data include call and SMS metadata, but not mobile money or top up information. We expect performance to increase with richer data and larger samples, observed over longer time periods.

After developing our method, we present three main findings.

First, we show that the method has the potential to achieve useful predictive accuracy, even with standard machine learning models. We assess its performance in two steps. First, we benchmark performance in our sample of formally banked (but thin file) consumers: our method actually outperforms credit bureau models, which perform relatively poorly (we consider the industry standard measure Area Under the ROC Curve or AUC, which ranges from 0.61-0.76 for our models versus 0.51-0.57 for bureau models). Second, our method performs similarly well for unbanked consumers, who cannot be scored with traditional methods (AUC 0.63-0.77). Our models perform within the (wide) range of published estimates of traditional credit scoring in the literature (AUC 0.50-0.79). Individuals in the highest quintile of risk by our most conservative measure are 2.8 times more likely to default than those in the lowest quintile.

Second, care must be taken to ensure stability over time. In practice, a creditor would use past performance to train the model that disperses future credit. The performance of machine learning methods can deteriorate if the underlying environment shifts over time (Butler, 2013; Lazer et al., 2014). The most straightforward way to set up the prediction task can pick up coincident shocks in addition to underlying factors correlated with repayment. We develop a technique to minimize this form of intertemporal instability, by using only variation within each time period to differentiate individuals (analogous to a form of temporal fixed effects). We demonstrate that this technique improves intertemporal stability, and that our models continue to outperform bureau models in our setting when estimated and tested on different time periods.

Third, we find that information gathered by the bureau is only slightly complementary to that in our indicators. This suggests that in contexts with thin bureau files, there may be limited gains from integrating these new forms of credit with the information already in legacy traditional credit bureaus. (However, digital forms of credit can be reported to bureaus, which can improve incentives to repay as well as the amount of information in the bureau.)

We conclude with a discussion on the logistics of implementing these methods in lending decisions. While our empirical exercise investigates the usefulness of this method for a telecom-specific form of credit, mobile money dramatically lowers the cost of providing small, remote, general loans. Methods like we propose here are now in use in developing countries to screen borrowers for digital credit. There are already over 68 digital credit products with 11m borrowers (Francis, Blumenstock, & Robinson, 2017), and in Kenya more individuals have loans through these new digital platforms than through traditional banking, or microfinance (FSD, 2016). As overhead costs decline to zero, the profitability of making a loan will be defined increasingly by its risk profile (Björkegren & Grissen, 2018). The ability to screen is thus fundamental. Our paper shines light on these markets in three additional ways. First, the ability to screen determines the profitability of lenders, and thus the products and populations that will be served by the private sector, as well as the amount of elbow room that regulators have to shape lending. Second, it will shape the organization of the market: in particular, whether lending will emerge connected to existing institutions such as telecoms or banks, or through new institutions like smartphone lending apps that independently request access to data. Third, documenting the method democratizes access, and can thus have a direct effect on entry.

The paper most similar to this is Pedro, Proserpio, and Oliver (2015), which finds that among individuals with active credit cards, those who recently defaulted have different calling behavior afterwards than those who repaid successfully, using 58 indicators of behavior. However, a person is likely to alter their patterns of calling after a default. It is unclear whether the indicators the paper derives could predict default, or simply pick up shocks that are only correlated with default ex post. Additionally, the paper observes only individuals who were already screened through a traditional credit bureau, so it cannot assess performance relative to a benchmark, or whether phone indicators can be used to screen the unbanked.

Our findings suggest that nuances captured in the use of mobile phones themselves can reduce information asymmetries, and thus can form the basis of new forms of low cost lending. Together with mobile money, these tools are enabling a new ecosystem of digital financial services. This ecosystem is leading to what appears to be a revolution in access to finance in the developing world.

2. CONTEXT AND DATA

The primary organizational partner is EFL (Entrepreneurial Finance Lab), which works on alternative credit scoring methods in developing and emerging markets, with an emphasis on the underbanked.[4] EFL identified a partner that was interested in exploring alternate methods of assessing creditworthiness.

[4] From their website, "EFL Global develops credit scoring models for un-banked and thin-file consumers and MSMEs, using many types of alternative data such as psychometrics, mobile phones, social media, GIS, and traditional demographic and financial data. We work with lenders across Latin America, Africa and Asia." http://www.eflglobal.com Darrell Grissen was employed by EFL while data was collected for this paper.

As a side effect of operation, telecoms already gather rich information about subscribers' transactions. We consider one particular application. As consumers in emerging economies have become wealthier, many telecoms have begun transitioning their subscribers from prepaid plans to postpaid subscriptions. Under postpaid plans, subscribers do not face the hassle of topping up their account, and so tend to consume more, and are also less likely to switch to competitors. However, postpaid plans expose the telecom to the risk that a subscriber may run up a bill that they do not pay back. In developed countries, many telecoms check subscribers' credit bureau files before granting a postpaid account. However, in lower income countries these files are often thin, or nonexistent. We partnered with a telecom in a middle income South American country, with GDP per capita of approximately $6,000, which sought to transition a subset of its prepaid subscribers to postpaid plans.[5] It wished to expand this subset to include those with sparse or nonexistent formal financial histories. The telecom offered a preselected set of subscribers the chance to switch to a postpaid plan with lower rates, and recorded who among these subscribers paid their bills on time. Because the telecom wanted to learn about the risks of transitioning different types of users in this initial exploration, it was permissive in selecting customers to transition. It selected customers who used their phones sufficiently (who were more likely to benefit from postpaid billing) from across the distribution of credit bureau records.[6] (We assess the impact of selection on our results in the Discussion section.) Preselected subscribers received a call inviting them to transition to a postpaid plan; those who opted in were switched from prepaid to the cheapest postpaid plan. We observe 7,068 subscribers who were offered postpaid plans and opted in, which is the relevant sample for assessing performance for the telecom. The telecom was aware that paying a phone bill was new for these subscribers, so if a subscriber did not pay their postpaid bill, they were notified by SMS and other channels that their bill was soon to become overdue. If consumers were more than 15 days overdue, their service was cancelled and they were reported to the

[5] All results reported in US dollars.

[6] In the resulting sample, 15% are missing credit bureau records, 26% have perfect bureau summary scores, 41% have near-worst bureau summary scores, and the remainder have summary scores in between.

credit bureau.[7] (While this form of credit has different features from a traditional bank loan, so do many emerging forms of digital credit; for example, short term loan ladders are common: Carlson (2017).) For each subscriber, they pulled mobile phone transaction records (Call Detail Records, or CDR). In this setting, many subscribers also had formal financial histories maintained at the credit bureau; the telecom also pulled these records. Bureau records include a snapshot of the number of entities reporting, number of negative reports, balances in different accounts (including consumer revolving, consumer nonrevolving, mortgage, corporate, and tax debt), and balances in different states of payment (normal, past due, written off). It also includes the monthly history of debt payment over the past 2 years (no record, all normal, some nonpayment, significant defaults), and includes a summary score that combines these indicators according to the bureau's judgment of what factors are important (using a decision rule not trained on data). Subscribers were matched to their financial histories based on an encrypted, anonymized identifier.

The mobile phone data include metadata for each call and SMS, with identifiers for the other party, time stamps, tower locations, and durations. It does not include top-ups, balances, data access, charges, handset models used, or mobile money transactions; thus we expect our performance to be a lower estimate of the performance that can be achieved with richer data. We do not observe any information on the content of any communication.

We aim to predict default based on the information available at the time credit was granted, so we include only mobile phone transactions that precede the date of plan switching. Descriptive statistics for the sample are presented in Table 1. Although 85% of our sample has a file at the credit bureau, many of these files are thin: 59% have at least one entity currently reporting an account, 31% have at least two, and only 16% have at least three. By construction, 100% of the sample has a prepaid mobile phone account. The median individual places 26 calls per week, speaking 32 minutes, and sends 24.4 SMS. In the data we

[7] For many consumers this would be the first record in their credit history. After this point, the consumer could use a prepaid account. Because the telecom could pause service, the credit could be thought of as one with the subscriber's phone number held as collateral. However, that collateral is limited, as subscribers could open a new prepaid account with a new phone number.

obtained from the partner, we observe the median individual's phone usage for 16 weeks; an implementation that can obtain longer histories is likely to perform better.

Table 1: Description of Individuals

	Mean	SD	Median
Country GDP per capita (Approx.)	$6,000		
Borrowers			
Gender is female	39%	-	-
Age (years)	35.8	12.8	34.0
Has a mobile phone	100%	-	-
Credit bureau record	85%	-	-
Entities reporting:			
At least one	59%		
At least two	31%		
At least three	16%		
Average weekly mobile phone use			
Calls out, number	32.0	25.6	26.0
Calls out, minutes	41.6	39.9	32.0
SMS sent	31.3	26.3	24.4
Days of mobile phone data preceding plan switch	107	14	112
Credit			
Default	11%	-	-
N	7,068		

3. METHOD

Our goal is to predict the likelihood of repayment using behavioral features derived from mobile phone usage. We consider a sample of completed plan transitions, and consider whether information that was available at the time the credit was extended could have predicted its repayment. Because this sample of individuals did obtain credit, risk is reported among those who received credit based on the selection criteria at the time, which was relatively permissive and spanned the distribution of credit histories (including no history at all).

The credit data provide an indicator for whether a particular borrower repaid their obligation (we use our partner's definition: 15 days past due). From the phone data we derive various features that may be associated with repayment. In a similar exercise, Blumenstock et al. (2015) generates features from mobile phone data using a data mining approach that is agnostic about the outcome variable. Our approach is instead tailored to one outcome, repayment. As in Björkegren and Grissen (2015), we extract a set of objects that may have an intuitive relationship to repayment, and then compute features that summarize these objects. We focus on features with an intuitive relationship because implementation partners can be wary of 'black box' methods, and indicators that have a theoretical link are more likely to have a stable relationship to the outcome of interest. While our approach is likely to extract some features similar to Blumenstock et al. (2015), it will also measure more nuanced features that would not have been generated by a generic method.

Phone usage captures many behaviors that have some intuitive link to repayment. A phone account *is* a financial account, and captures a slice of a person's expenditure. Most of our indicators measure patterns in how expenses are managed, such as variation (is usage erratic?), slope (is usage growing or shrinking over time?), and periodicity (what are the temporal patterns of usage?). In particular, individuals with different income streams are likely to have different periodicities in expenditure (formal workers may be paid monthly; vendors may be paid on market days). We also capture nuances of behavior that may be indirectly linked to repayment, including usage on workdays and holidays, and patterns of geographic mobility, which can reflect information on employment. Although social network measures may be predictive (who one is connected to may reflect one's level of responsibility or ability to access resources), we include only basic social network measures that do not rely on the other party's identity (degree, and the distribution of transactions across contacts), as we are hesitant to suggest that a person's lending prospects should be affected by their contacts. While many traditional credit scoring models aim to uncover a person's fixed type (whether the person is generally a responsible borrower), the high frequency behavior we capture may also pick up features specific to the time when an individual is being evaluated for credit (a person may be likely to repay this credit, even if they are not generally responsible). Our process has three steps:

First, we identify atomic events observed in the data, each represented as a tuple *(i, t, e, X_{iet})*, where *i* represents an individual, *t* represents the timestamp, *e* represents an event type, and X_{iet} represents a vector of associated characteristics. Event types include transactions (call, SMS, or data use), device switches, and geographic movement (coordinates of current tower). Characteristics derived from the raw transaction data include variables capturing socioeconomics (the handset model, the country of the recipient), timing (time until the credit is granted, day of the week, time of day, whether it was a holiday), and management of expenses (whether the sender or receiver had pre- or post-paid account, whether the transaction occurred during a discount time band, or at the discontinuity of a time band).

Second, for each individual *i*, event type *e*, and characteristic *k*, we compute a vector with the sum of events of each potential value of the characteristic:

$$D_{iek} = \left[\sum_{t} 1\{X_{ietk} = d\}\right]_{d \in unique(X_{ek})}$$

This generates, for example, the count of top ups by time of day, data usage by days since top up, the number of minutes spoken with each contact, the number of SMS to pre- and post-paid accounts, and the total duration of calls immediately before and after the start of a discount time band.

Finally, for each vector we compute a set of summary statistics. For sequences, these include measures of centrality (mean, median, quantiles), dispersion (standard deviation, interquantile ranges), and for ordinal sequences, change (slope) and periodicity (autocorrelation of various lags, and fundamental frequencies—which correspond to the periods of the strongest repeating temporal patterns. For counts by category, we compute the fraction in each category and overall dispersion (Herfindahl-Hirschman Index).

For geographic coordinates, we compute the maximum distance between any two points, the distance from the centroid to several points of interest, and use a clustering algorithm to identify important places (Isaacman et al., 2011). We also compute statistics that summarize pairs of sequences, including correlations, ratios, and lagged correlations (e.g., the correlation of minutes spoken with SMS, which may indicate whether a person coordinates in bursts of activity).

These three steps generate various quantifications of the intuitive features presented (including strength and diversity of contacts) as well as other measures (intensity and distribution of usage over space and time, and mobility). For each feature, we also add an indicator for whether that individual is missing that feature. Altogether, we extract approximately 5,500 features with variation.

The code associated with this paper also defines other features using top up, balance, and handset information which are not present in this dataset, including how balances are depleted over time, and the degree to which a person exploits discount calling times. Those can be used in richer transaction data from other contexts.

4. PREDICTION AND RESULTS

A first question is how individual features correlate with default. Table 2 presents the single variable correlation with default.

Characteristics traditionally available to lenders are not very predictive. Demographic features (gender and age) have very low correlation with repayment (magnitudes between 0.04 and 0.07). Having a credit bureau record has a small negative correlation with repayment (-0.02). For individuals with records, the most predictive feature is the summary score (-0.072; lower is better), and the fraction of debt lost (-0.046). That individual credit bureau features are only slightly predictive suggests that predicting repayment in this setting is a difficult problem.

Individual features derived from mobile phone usage have slightly higher correlations, ranging up to 0.16. But mobile phone usage data are richer, so there are many more features of behavior that can be included in a model. Since many features measure similar concepts, we present broad categories, and the correlation of one top feature within that category. Correlated features include the periodicity of usage (top correlation -0.16), slope of usage (0.13), correlations in usage (0.11), and variance (-0.10). The table highlights particular features that perform well in isolation, including the slope of daily calls sent, and the number of important geographical location clusters where the phone is used. We next use multiple features together to predict repayment.

Table 2: Individual Features

	Correlation with repayment	t-stat	Number of Features
Demographics			**2**
Age	0.073	2.35	
Female	-0.039	-1.26	
Credit Bureau			**36**
Has a credit bureau record	-0.022	-1.89	
Summary score (lower is better)	-0.072	-6.15	
Fraction of debt lost	-0.046	-3.86	
Phone usage			**5,541**
Categories	*High performing example feature:*		
Periodicity	-0.163	-5.27	796
	SMS by day, ratio of magnitudes of first fundamental frequency to all others		
Slope	0.126	4.06	44
	Slope of daily calls out		
Correlation	0.111	3.57	224
	Correlation in SMS two months ago and duration today		
Variance	-0.104	-3.34	4,005
	Difference between 80th and 50th quantile of SMS use on days SMS is used		
Other	0.100	3.07	542
	Number of important geographical location clusters		

Predicting Repayment

We find that our features are predictive even using standard methods common in the machine learning literature. We estimate two standard machine learning models: random forests, and logistic regressions using a model selection procedure (stepwise search using the Bayesian Information Criterion or BIC), for bureau indicators and phone indicators (CDR).[8] Random forests are a generalization of decision trees designed to reduce overfitting, by combining multiple trees which each have access to a subset of the sample (Breiman, 2001).

However, these straightforward estimation routines may muddle the individual factors that explain repayment with common temporal shocks that lead to differences in the proportion of credit repaid in different time periods. High frequency indicators such as our phone indicators are particularly susceptible to picking up these shocks. For phone indicators, we develop two new models that improve intertemporal stability by basing predictions off of only within-week variation (CDR-W). We train an OLS model with week fixed effects; these absorb week-to-week variation in repayment.[9] We form predictions differently from a standard fixed effect model. A standard model would include the fixed effect for each offer week in its predicted repayment, but that is not feasible in our setting: a lender would not know the fixed effect for future weeks. Instead, we form a prediction using the average of the past weeks' fixed effects, weighted by the proportion of loans granted in that past week. We train a random forest analogously: we fit separate random forest models to each past week of data, and combine them in an ensemble. When making a prediction for an individual, we weight each submodel by the proportion of transitions granted in that past week.[10] This approach reduces the discrepancy between within-time and out-of-time performance; it may also lead to selecting indicators that are more stable over time.

[8] We initialize the stepwise search from multiple sets of starting variables, and keep the model with the highest within-fold fit. We use the randomForest R package with default tuning of 500 trees, sample sizes of 63.2% drawn for each tree, and $\sqrt{K}$ variables considered at each node (Breiman & Cutler, 2006).

[9] If few transitions are made in a week, we combine it with adjacent weeks.

[10] When giving out loans, one could upweight more recent models to capture changes in conditions.

To illustrate the features that the models select, we first estimate these models on each entire sample, and present random forest importance plots in Appendix Figure A and regression parameter estimates in Appendix Table A. Standard models tend to place substantial weight on various periodicities of behavior. While some of these patterns are related to repayment, others pick up high frequency artifacts in the data. Our within-week models place less weight on periodicities, and more weight on the fraction of duration spoken during the workday or late at night, the distance traveled, variation in usage, and correlations between calls and SMS. The OLS fixed effect model is also simpler than the logistic model, suggesting the fixed effect approach penalizes model complexity more.

Performance

Within Time

We measure how the method performs out of sample using cross validation. Following common practice in supervised machine learning, we divide the sample into *R* randomly selected folds. We cycle through each fold, estimating (training) the model on *R-1* folds, and reporting predictive performance on the *R*th omitted fold (testing). We average the results over each fold, and over multiple fold draws. Larger values of *R* exploit more of the sample for training, which tends to improve predictive performance, but increase the computational burden because the model must be estimated *R* times. For our main results, we select *R*=5, which is commonly used in the machine learning literature (Appendix Table C reports results for R=10).

As a first check, we consider how well the best model separates low and high risk borrowers. Our models generate continuous scores. Because we do not know where in the distribution of scores a lender would set the acceptance threshold, the metrics we report trace outcomes along the range of thresholds. We report results from the most conservative model, the random forest weekly ensemble. Figure 1 shows how the default rate varies with the fraction of borrowers accepted (where borrowers with lowest predicted default are accepted first). In our most conservative model, individuals with the highest quintile of risk scores are 2.8 times more likely to default than those with the lowest quintile.

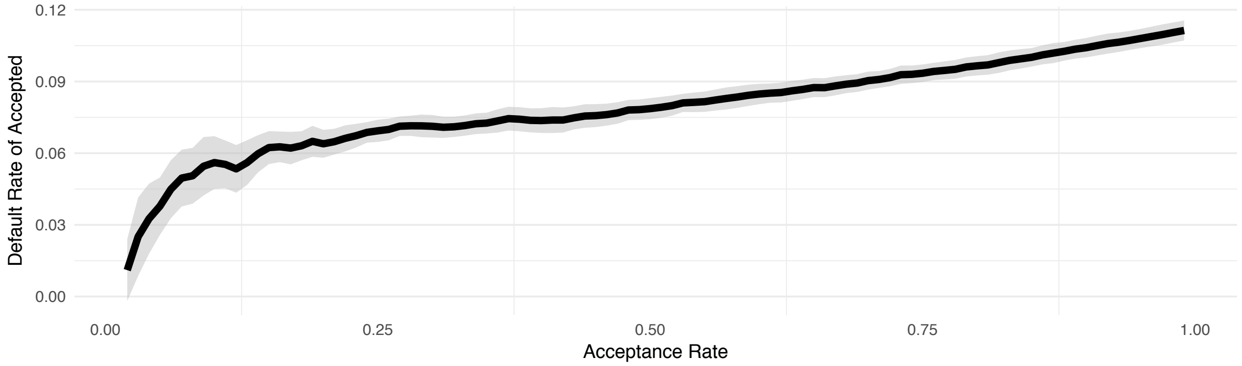


Figure 1: Default Rate by Proportion of Borrowers Accepted

Phone indicators using the conservative random forest weekly ensemble model (CDR-W). Line shows mean, and ribbon standard deviation, of results from multiple fold draws.

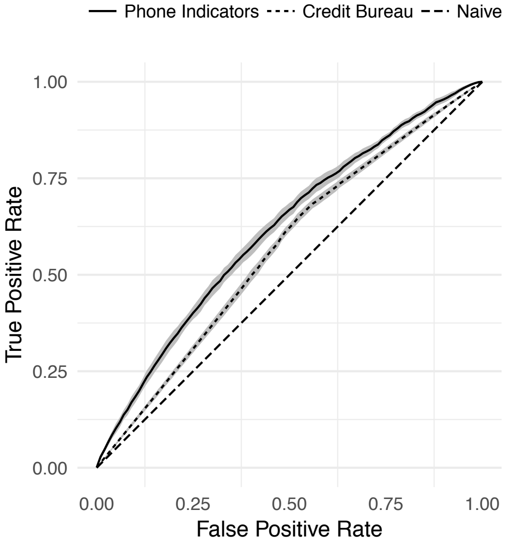


Figure 2: Receiver Operating Characteristic Curve

Credit bureau uses the highest performing stepwise logistic model. Phone indicators use the conservative random forest weekly ensemble model (CDR-W). Line shows mean, and ribbon standard deviation, of results from multiple fold draws.

The receiver operating characteristic curve (ROC) plots the true positive rate of a classifier against the false positive rate, tracing out performance as the acceptance rate is varied. Following the credit scoring literature, we report the area under this curve (AUC) to summarize performance across the range of possible acceptance thresholds.[11] A naïve classifier would generate an AUC of 0.5 and a perfect classifier would generate an AUC of 1.0. Figure 2 illustrates the ROC for the best benchmark model (stepwise logistic) and the most conservative model using indicators derived from phone data (random forest weekly ensemble).

We show results for a variety of specifications in Table 3, measuring performance with AUC. (We also assess performance with alternate metrics in Appendix Table D). We present results for the entire sample, and then split into the subsamples that do and do not have credit bureau records. In this population with thin files, credit bureau information does not perform especially well in predicting repayment (AUC 0.51-0.57). For bureau indicators, the logistic model outperforms the random forest, suggesting that the underlying relationship between those indicators and repayment is relatively linear. In contrast, standard models built on phone indicators (CDR) are predictive, reaching AUCs of 0.71-0.77 when trained and tested on the same time period. Our more conservative CDR-W models achieve lower performance when trained and tested in the same time period (AUCs 0.62-0.63), but also outperform credit bureau models. The performance of our models is also in the range of a sample of published within-time AUC estimates for traditional credit scoring on traditional loans in developed settings (0.50-0.79, shown in Appendix Table B). Our method's performance is similar overall and within each quartile of mobile phone usage, suggesting it picks up nuances in usage rather than overall usage (See Appendix Figure B and (Björkegren & Grissen, 2018)). Combining our indicators with information from the credit bureau slightly boosts performance, suggesting that the information gathered by the bureau is only slightly complementary to that collected by our approach.

In a robustness check, we train models instead with 10 fold cross validation in Appendix Table C, which exploits more of the sample but is more computationally demanding. We find that it improves performance of the CDR models in particular, which are more complex and thus data hungry. Their performance is also likely to improve with additional tuning.

[11] Credit approval decisions tend to approve applicants with scores above a threshold. AUC has two useful properties for these types of decisions: they consider only the relative ranking of observations, and they trace through the range of potential thresholds.

Our method also performs better than credit bureau models when assessed with alternate metrics of performance that are informative for our decision problem; in particular, the H-measure designed to overcome some weaknesses of AUC (Hand, 2009) and out of sample R^2 (see Appendix Table D). We also present ROC curves in Appendix Figure D and comparisons of scores to actual repayment in Appendix Figure E, for the main models.

Table 3: Model Performance

	Main Results			**Check**
Dataset:	**Standard Indicators**			**Offset Indicators**
Performance:	Out of Sample (5 fold CV)			Out of Time (train early period, test late)
Sample:	All	Has Bureau Records	No Bureau Records	All
	AUC	AUC	AUC	AUC
Baseline Model				
Credit Bureau				
Random Forest	0.516	0.509	-	0.507
Logistic, stepwise BIC	0.565	0.565	-	0.550
Our Models				
Phone indicators (CDR)				
Random Forest	0.710	0.708	0.719	0.631
Logistic, stepwise BIC	0.760	0.759	0.766	0.595
Phone indicators, within-week variation (CDR-W)				
Random Forest Weekly Ensemble	0.616	0.614	0.630	0.641
OLS FE, stepwise BIC	0.633	0.634	0.631	0.593
Combined				
Credit bureau and phone indicators				
Random Forest	0.711	0.708	-	0.642
Logistic, stepwise BIC	0.772	0.770	-	0.616
Credit bureau and phone indicators, within-week variation				
Random Forest Weekly Ensemble	0.618	0.616	-	0.639
OLS FE, stepwise BIC	0.645	0.645	-	0.586
Default Rate	11%	12%	10%	
N	7,068	6,043	1,025	6,975

Standard indicators evaluate out of sample performance using 5-fold cross validation, averaged over fold draws. Offset indicators are derived from only half of the data (the first half for early transitions; the last half for late transitions); the out of time model is estimated on the early half of transitions and tested on the late half. AUC represents the area under the receiver operating characteristic curve. For middle two columns, model is trained on all individuals except the omitted fold, and performance is reported for the given subsample within the omitted fold.

Out of Time

When implemented, a model trained on past data will be used to predict future repayment. As a robustness check, we assess the out-of-time performance of all models by training and testing on different time periods. To do this, we construct an offset version of the dataset. We split the sample of individuals into two; the early group that was transitioned before the median date, and the late group after the median. Then, we evenly divide the phone data, into an early and late period, and construct offset versions of our indicators using only transactions occurring in that half of the data (up to the date of each transition). Because these offset indicators are constructed on a shorter panel, they capture less information than our full indicators. We train the model on the early group, with phone indicators derived from the early period of phone data, and test it on the late group, with indicators derived from the late period of phone data, with results in the last column of Table 3. Because there is only one late period to test on, out of time results are exposed to much more noise than the within time results (for which we can compute performance across many fold draws). As a result, these out of time results should be viewed as only a rough check of how much we should trust the within-time results.

Models based on the bureau data, which is lower dimensional, tend to be relatively stable over time. On the other hand, standard models using phone indicators see substantial deterioration (AUC declines from 0.71 to 0.63 for Random Forest and from 0.76 to 0.60 for logistic stepwise). Our modified phone indicator models that use only within week variation are much more stable (AUC increases from 0.62 to 0.64 for Random Forest and decreases from 0.63 to 0.59 for stepwise OLS FE).[12] All phone indicator models continue to outperform models using credit bureau data on this cut of the data (AUC 0.55-0.58). Our performance also lies within the range of the one comparable published benchmark of out of time performance of traditional credit scoring we could find in the literature, from a developed setting (AUC 0.57-0.76, Appendix Table B). Those and our results suggest that bureau models can face at least slight

[12] The CDR-W Random Forest model is likely to underperform when trained on the same time period with cross validation: it learns less structure when an equivalent sample size is split across multiple time periods (as is the case with out of sample test, which trains on a random subset of loans across weeks).

deterioration when tested out of time, with the caveat that we have tested only one time period. We expect the out of time performance of our methods to improve when trained on multiple cohorts (just as credit bureaus have evolved the data they collect by observing default patterns over many cohorts).

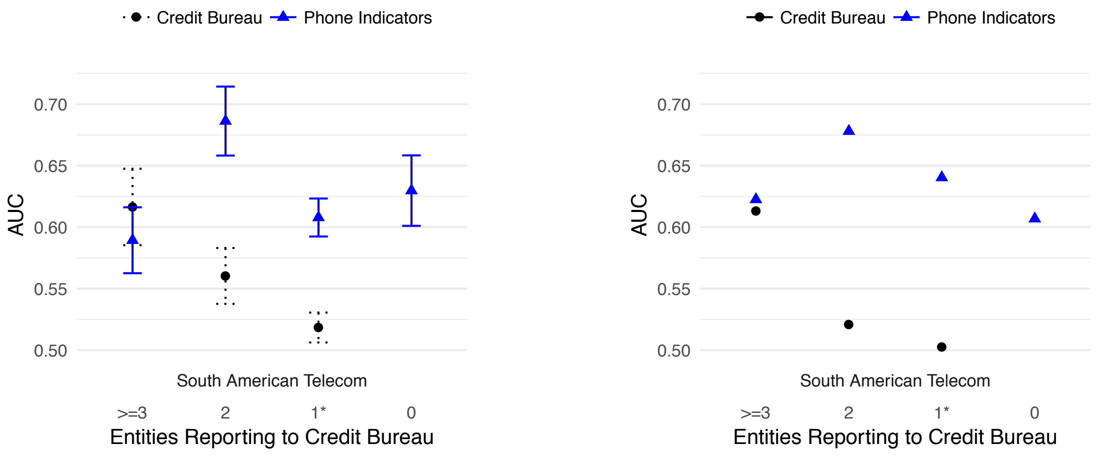
After the results from this pilot, the telecom implemented a scoring system using data and methods similar to what we suggested, suggesting they viewed it to be profitable.

5. DISCUSSION

Mobile phone data appear to quantify nuanced aspects of behavior that are typically considered soft, making these behaviors ‘hard’ and legible to formal institutions (Berger & Udell, 2006). Further, these data are already being captured. We expect that the method can assist with the provision of financial products to the poor in several ways.

Expanding lending to the unbanked

This paper studies individuals who are near the existing financial system. We summarize the performance of our method by level of formalization in Figure 4. The performance of credit bureau models deteriorates as we move from individuals with rich financial histories (3 or more entities contributing reports to the bureau) to those with sparser histories. Our method does not deteriorate across levels of formalization, and generates scores of similar performance among individuals with no bureau history, who cannot be scored with traditional methods.



(a) Within Time **(b) Out of Time**

Figure 4: Performance by Level of Formalization

1*: either one entity reporting, or has a file at the credit bureau which may include previous activity but zero entities are currently reporting. Comparison of the highest performing bureau model (logistic stepwise) and most conservative phone indicator model (CDR-W Random Forest). Models are trained on all individuals but the omitted fold and AUC is reported for the subset of individuals within the omitted fold with the given number of entities reporting to the credit bureau. For out of sample estimates, the point shows the mean, and error bars standard deviation, of results from multiple fold draws. Out of time estimates use the offset indicators and only have a single fold draw.

While our method does not require observing a traditional bureau history, it does require observing phone usage. Our partner selected users who spend more on the telephone than the average phone user in the country. We assess the extent to which performance would deteriorate among sparser users in two exercises, in (Björkegren & Grissen, 2018), reproduced and updated in Appendix Figure B. First, we compare performance between quartiles of usage within our sample; we do not find that our method performs better among users who spend more. Second, we construct a synthetic dataset by dropping transactions from our dataset to match the spending of lighter users. These synthetic datasets simulate the number of transactions that a lighter user might make in the same time span we observe. Performance begins to deteriorate near $1 in spending per month (dropping 96% of transactions). We expect performance for lighter users would improve if they were observed for longer than the median of 16 weeks that we observe. These results suggest that the method may be able to reliably score individuals through most of the distribution of phone usage.

Our approach can dramatically reduce the cost of screening individuals on the margins of the banking system. When poor individuals in the developing world are able to get loans, they are often screened through costly methods such as detailed interviews or peer groups. In contrast, our method can be implemented at extremely low cost, and can be executed over a mobile phone network without the need for physical interaction. These methods enable new forms of lending that do not require the full structure of current branch lending, such as digital credit. Digital credit can both reduce the cost of serving existing markets, and make it profitable to serve consumers outside the current financial system.

Implementation

As demonstrated, this approach can be used to extend telecom-specific credit, within the firms that already possess the necessary data.[13] However, the applications are much broader. Mobile money makes it cheap to deliver a loan and collect general payment. With regulatory approval, telecoms may connect to the banking sector, and offer loans to consumers.[14] Alternately, telecoms can package these data into a credit score that can be used by third parties, either through mobile banking platforms or an independent credit bureau.[15] A third implementation, a smartphone app, allows third parties to access usage data independently of telecom operators, and is being explored by several startups.[16] These apps ask for permission to view call history and other behavioral data, and can collect real-time data for a set period.

Our performance estimates are derived from provision of postpaid credit. If phone usage is particularly informative about repayment of this form of credit, performance could differ when predicting

[13] In addition to assisting with the transition to postpaid plans, this method can be used to extend credit for handset purcases, or to maintain consistent airtime balances. Many developing country operators offer small airtime loans like this; a scoring model could improve their provision.

[14] See for example, Jumo.

[15] See for example, Cignifi.

[16] See for example, Tala and Branch.

default of other forms of loans.[17] But we expect repayment patterns for general loans extended over mobile phones to have similarities to the postpaid credit we study. Both are extended remotely, without human interaction. In case of nonpayment of a general loan, telecoms can also report borrowers to a credit bureau, and may also be able to freeze a phone account or garnish funds from mobile money balances (as permitted by regulation).[18]

Privacy

Privacy will be a key consideration in any implementation. As demonstrated in this paper, the scoring model can be estimated with anonymous data, by anonymizing the identifier that links phone and lending data. However, to generate a prediction for a lending decision, the model must be run on that potential borrower's data. An implementation can be designed to mitigate privacy risks. It can be opt-in, so that only consumers who consent are scored with the system.[19] It can reveal to lenders only a single number summarizing default risk, rather than the underlying features describing behavior. Additionally, it can be restricted to use features that are less sensitive, such as top up behavior rather than the network structure of an individual's contacts.

Manipulation

Some indicators are 'gameable' in the sense that a subscriber may be able to manipulate their score if they knew the algorithm. The feasibility of manipulation depends on the complexity of the final model and the susceptibility of individual indicators to manipulation. Both dimensions of the model can be tailored

[17] One particularly crisp formulation of this concern would be if heavy phone users are willing to do more to avoid phone account closure, and the method simply picked up the level of phone usage. In that case, it may perform better predicting when repayment of a phone bill than a general loan. However, as mentioned above and in Appendix Figure B, we find that the method achieves similar performance in each quartile of usage, and in our sample that has a sufficient amount of usage, the correlation between airtime usage and repayment is very small (-0.03).

[18] Alternately, it could be that bureau information is less predictive for this form of credit than a general loan, which would deflate the benchmark that we compare our results against. Bureau information would represent the status quo for the extension of postpaid credit. There is not much evidence on the performance of bureau information in low income populations; the performance we observe is in the lower end of the range of published estimates for general loans from more developed settings (Appendix Table B). However, bureau information cannot be used to score our population of interest, and our method scores this population at a higher level of performance within this range.

[19] Potential borrowers who opt in may be differentially selected from the broader population, in which case a model estimated on anonymous data from the broader population may not be optimal for use in practice. After the system is operational, it can be periodically refit on outcomes from borrowers who opt in. (Thanks to an anonymous referee for this point).

to reduce the probability of manipulation. For example, it is preferable to use indicators that are less susceptible (e.g., manipulating spending or travel can be costly).

Heterogeneity in performance by subgroup

An extension to this paper evaluates performance by different subgroups (Björkegren & Grissen, 2018), replicated in Appendix Figure C. The model is trained on all individuals except the omitted fold, and performance is reported for the given subgroup (women, men, and residents in or outside the capital) within the omitted fold. Although our error bars can be wide, performance is not widely heterogeneous across groups, suggesting the method may be able to score different types of individuals.

If multiple users share each mobile phone account

In many developing countries, individuals share phones to lower expenses. When a phone account is shared among multiple people, this method will produce one score for the account. The method will still produce an unbiased predictor of the account owner's repayment if sharing practice does not differ between estimation and implementation. In that case, the method will capture both the behavior of phone owners as well as those they choose to share with (indeed the choice of who to share with may also correlate with repayment).

If each user has multiple mobile accounts

On the other hand, in competitive mobile markets each individual may use multiple accounts, to take advantage of in-network pricing across multiple networks. This practice is convenient with prepaid plans (with mainly marginal charges) on GSM phones (which allow SIM cards to be easily swapped or may have dual SIM card slots). When users split their call behavior across multiple networks, data gathered from a single operator will represent only a slice of their telephony. While this will make their data sparser, as long as the practice does not differ between estimation and implementation, it will not introduce biases into the method. If individuals use multiple accounts on a single handset (if the handset supports dual SIMs or users swap SIM cards), data gathered from that handset through an app could measure activity across all accounts.

6. CONCLUSION

This paper demonstrates a method to predict default among borrowers without formal financial histories, using behavioral patterns revealed by mobile phone usage. Our method is predictive of default in the middle income population we study, which tends to have thin or nonexistent credit bureau files. In this population our method performs better than credit bureau models. But our method can also score borrowers outside the formal financial system, who cannot be scored with traditional methods. While this paper is focused on predicting repayment, the type of data we use can reveal a much wider range of individual characteristics (Blumenstock et al., 2015), and could conceivably be used to predict other outcomes of interest—such as lifetime customer value, or the social impact of a loan.

It has been widely acknowledged that mobile phones can enable low cost money transfers and savings in developing countries (Suri, Jack, & Stoker, 2012). Our results suggest that nuances captured in the use of mobile phones themselves can alleviate information asymmetries, and thus can form the basis of new forms of low cost lending. These tools together are enabling a new ecosystem of digital financial services.

7. REFERENCES

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635.

Banerjee, A., Duflo, E., Kinnan, C., & Glennerster, R. (2014). The Miracle of Microfinance? Evidence from a Randomized Experiment.

Banerjee, A., Karlan, D., & Zinman, J. (2015). Six Randomized Evaluations of Microcredit: Introduction and Further Steps. *American Economic Journal: Applied Economics*, *7*(1), 1–21.

Banerjee, A. V., & Duflo, E. (2014). Do Firms Want to Borrow More? Testing Credit Constraints Using a Directed Lending Program. *The Review of Economic Studies*, *81*(2), 572–607.

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking & Finance*, *30*(11), 2945–2966.

Björkegren, D. (2010). “Big data” for development. Proceedings of the CEPR/AMID Summer School. Retrieved from http://dan.bjorkegren.com/files/CEPR_Bjorkegren.pdf

Bjorkegren, D., & Grissen, D. (2015). Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment. Working Paper.

Björkegren, D., & Grissen, D. (2018). The Potential of Digital Credit to Bank the Poor. *American Economic Association Papers and Proceedings*.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., & Cutler, A. (2006). *randomForest*. Retrieved from http://stat-www.berkeley.edu/users/breiman/RandomForests

Butler, D. (2013). When Google got flu wrong. *Nature News*, *494*(7436), 155. https://doi.org/10.1038/494155a

Calabrese, R., & Osmetti, S. A. (2013). Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics*, *40*(6), 1172–1188.

Carlson, S. (2017). Dynamic Incentives in Credit Markets: An Exploration of Repayment Decisions on Digital Credit in Africa. *Working Paper*.

de Janvry, A., McIntosh, C., & Sadoulet, E. (2010). The supply- and demand-side impacts of credit market information. *Journal of Development Economics*, *93*(2), 173–188. https://doi.org/10.1016/j.jdeveco.2009.09.008

De Mel, S., McKenzie, D., & Woodruff, C. (2008). Returns to Capital in Microenterprises: Evidence from a Field Experiment. *The Quarterly Journal of Economics*, *123*(4), 1329–1372.

Demirguc-Kunt, A., Klapper, L., Singer, D., & Van Oudheusden, P. (2014). The Global Findex Database. World Bank. Retrieved from http://www.worldbank.org/en/programs/globalfindex

Francis, E., Blumenstock, J., & Robinson, J. (2017). Digital Credit: A Snapshot of the Current Landscape and Open Research Questions. *CEGA White Paper*.

FSD. (2016). FinAccess Household Survey.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123.

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J., & Varshavsky, A. (2011). Identifying Important Places in People’s Lives from Cellular Network Data. In K. Lyons, J. Hightower, & E. Huang (Eds.), *Pervasive Computing* (Vol. 6696, pp. 133–151). Springer. Retrieved from http://www.springerlink.com/content/r14x8r7573738143/abstract/

ITU. (2011). *World telecommunication/ICT indicators database*. International Telecommunication Union.

Karlan, D., & Zinman, J. (2011). Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation. *Science*, *332*(6035), 1278–1284. https://doi.org/10.1126/science.1200138

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Lu, X., Wetter, E., Bharti, N., Tatem, A. J., & Bengtsson, L. (2013). Approaching the Limit of Predictability in Human Mobility. *Scientific Reports*, *3*. https://doi.org/10.1038/srep02923

Luoto, J., McIntosh, C., & Wydick, B. (2007). Credit Information Systems in Less Developed Countries: A Test with Microfinance in Guatemala. *Economic Development and Cultural Change*, *55*(2), 313–334.

McKenzie, D., & Woodruff, C. (2008). Experimental Evidence on Returns to Capital and Access to Finance in Mexico. *The World Bank Economic Review*, *22*(3), 457–482.

Ng, A. (2011). Machine Learning and AI via Brain simulations.

NPR. (2015). How Cellphone Use Can Help Determine A Person's Creditworthiness. *Morning Edition*. Retrieved from https://www.npr.org/2015/08/04/429219691/how-cellphone-usage-can-help-determine-a-person-s-credit-worthiness

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., … Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, *104*(18), 7332–7336. https://doi.org/10.1073/pnas.0610245104

Palla, G., Barabási, A.-L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, *446*(7136), 664–667. https://doi.org/10.1038/nature05670

Pedro, J. S., Proserpio, D., & Oliver, N. (2015). MobiScore: Towards Universal Credit Scoring from Mobile Phone Data. In *User Modeling, Adaptation and Personalization* (pp. 195–207). Springer, Cham. https://doi.org/10.1007/978-3-319-20267-9_16

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011). Prediction of Socioeconomic Levels Using Cell Phone Records. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), *User Modeling, Adaption and Personalization* (pp. 377–388). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22362-4_35

Suri, T., Jack, W., & Stoker, T. M. (2012). Documenting the birth of a financial economy. *Proceedings of the National Academy of Sciences*, *109*(26), 10257–10262. https://doi.org/10.1073/pnas.1115843109

Van Gool, J., Verbeke, W., Sercu, P., & Baesens, B. (2012). Credit scoring for microfinance: is it worth it? *International Journal of Finance & Economics*, *17*(2), 103–123.

World Bank. (2014). Facilitating SME Financing through Improved Credit Reporting.

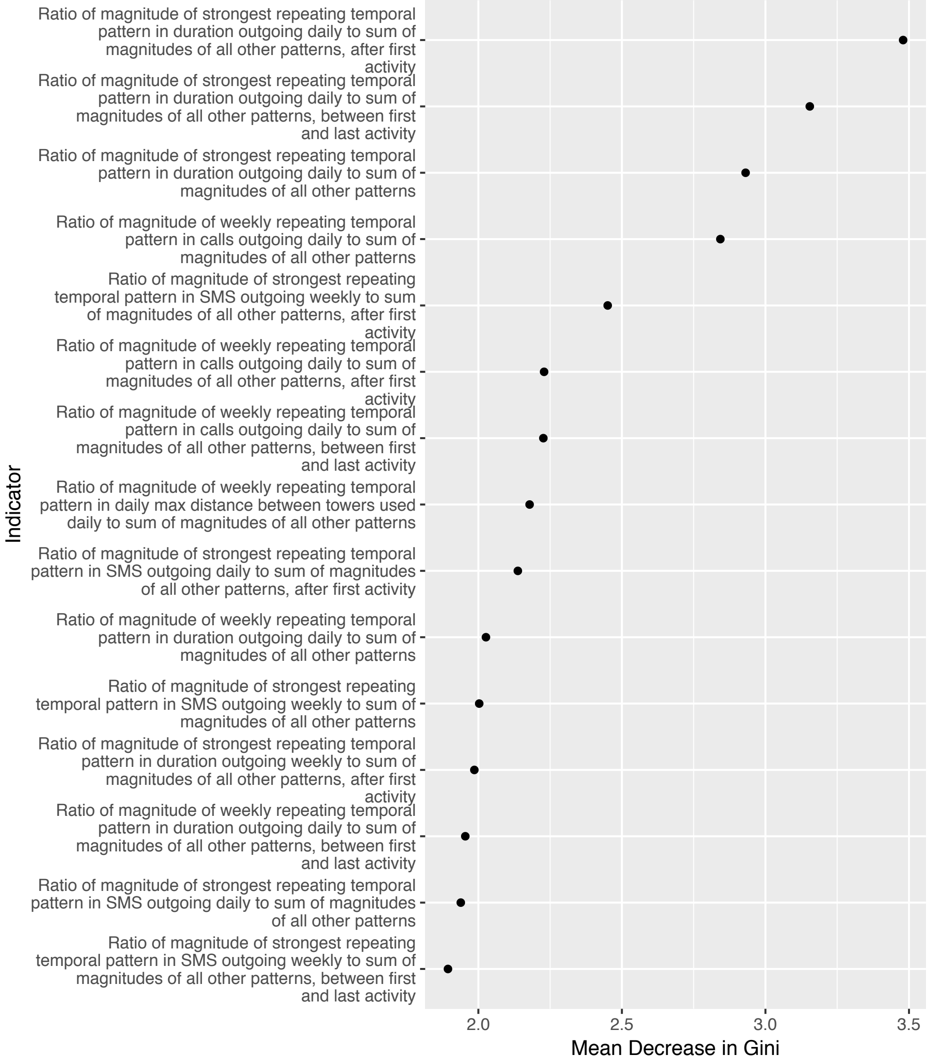
8. APPENDIX

Figure A: Random Forest Estimates

The following importance plots measure the mean decrease in the Gini coefficient of the top fifteen features, which corresponds to the marginal impact of including the variable in the model.

(a) Standard Random Forest

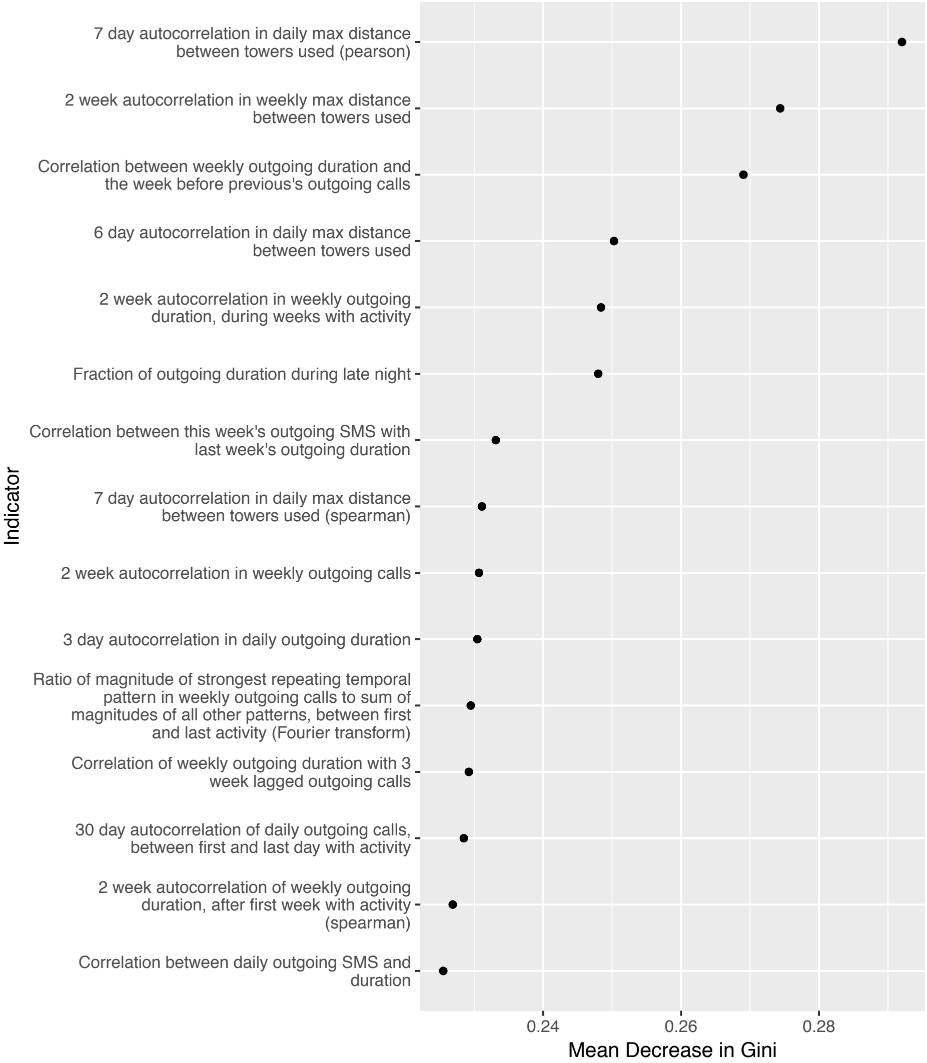
Importance Plot: Top 15 Features



Ensemble of 500 trees. Mean nodes per tree: 424.

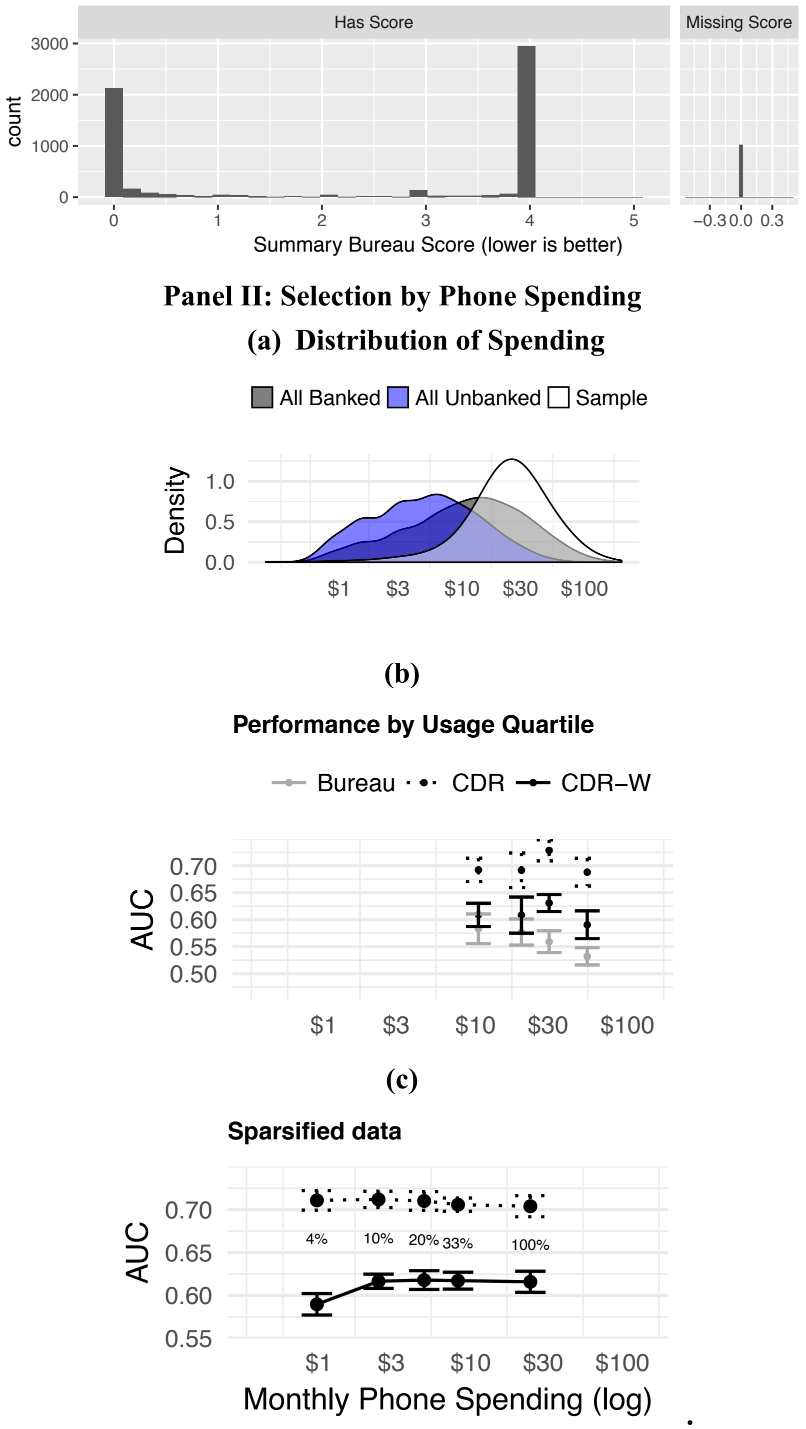
(b) Random Forest Weekly Ensemble

Importance Plot: Top 15 Features



Ensemble of 4 random forests, each of 500 trees.

Figure B: Sample Selection



Panel I: Selection by Bureau Summary Score

Panel II: Selection by Phone Spending

(a) Distribution of Spending

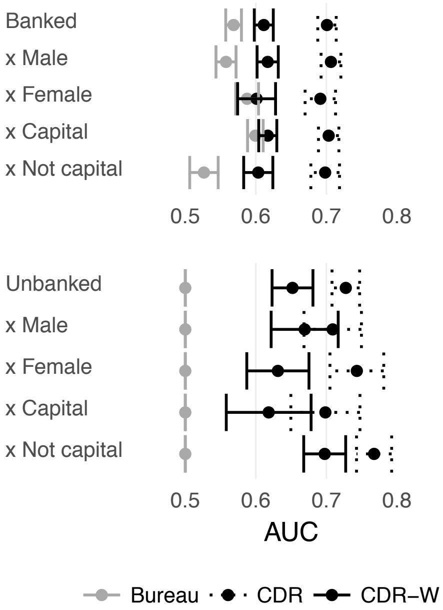
(b)

(c)

:

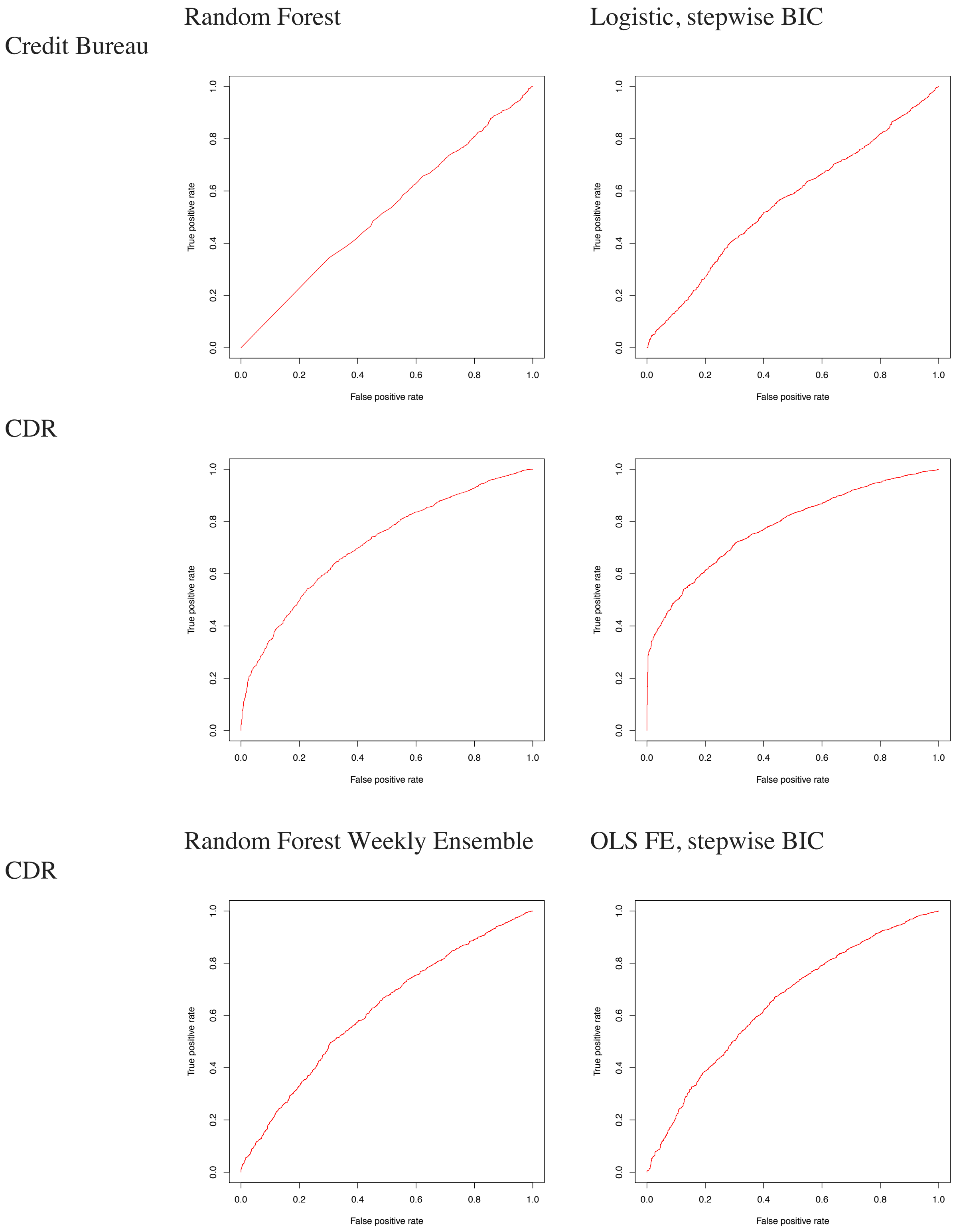
Note: CDR represents base random forest model, and CDR-W weekly ensemble model. Since logistic regression performed better with bureau data, we present that model here. Sparse CDR results derived from estimating and testing on synthetic datasets that subsample every *k*th transaction transactions. Bars represent 1 standard deviation. In our sample, spending is measured per phone account. National spending levels obtained from a household survey; that survey reports total spending per household; plot assumes one phone account per household. (Bottom two figures reproduced from (Björkegren & Grissen, 2018).)

Figure C: Performance by Subgroup



Note: CDR represents base random forest model, and CDR-W weekly ensemble model. Since logistic regression performed better with bureau data, we present that model here. Bars represent 1 standard deviation. Reproduced from (Björkegren & Grissen, 2018).

Figure D: Receiver Operating Characteristic (ROC) Curves



Random Forest

Logistic, stepwise BIC

Credit Bureau

CDR

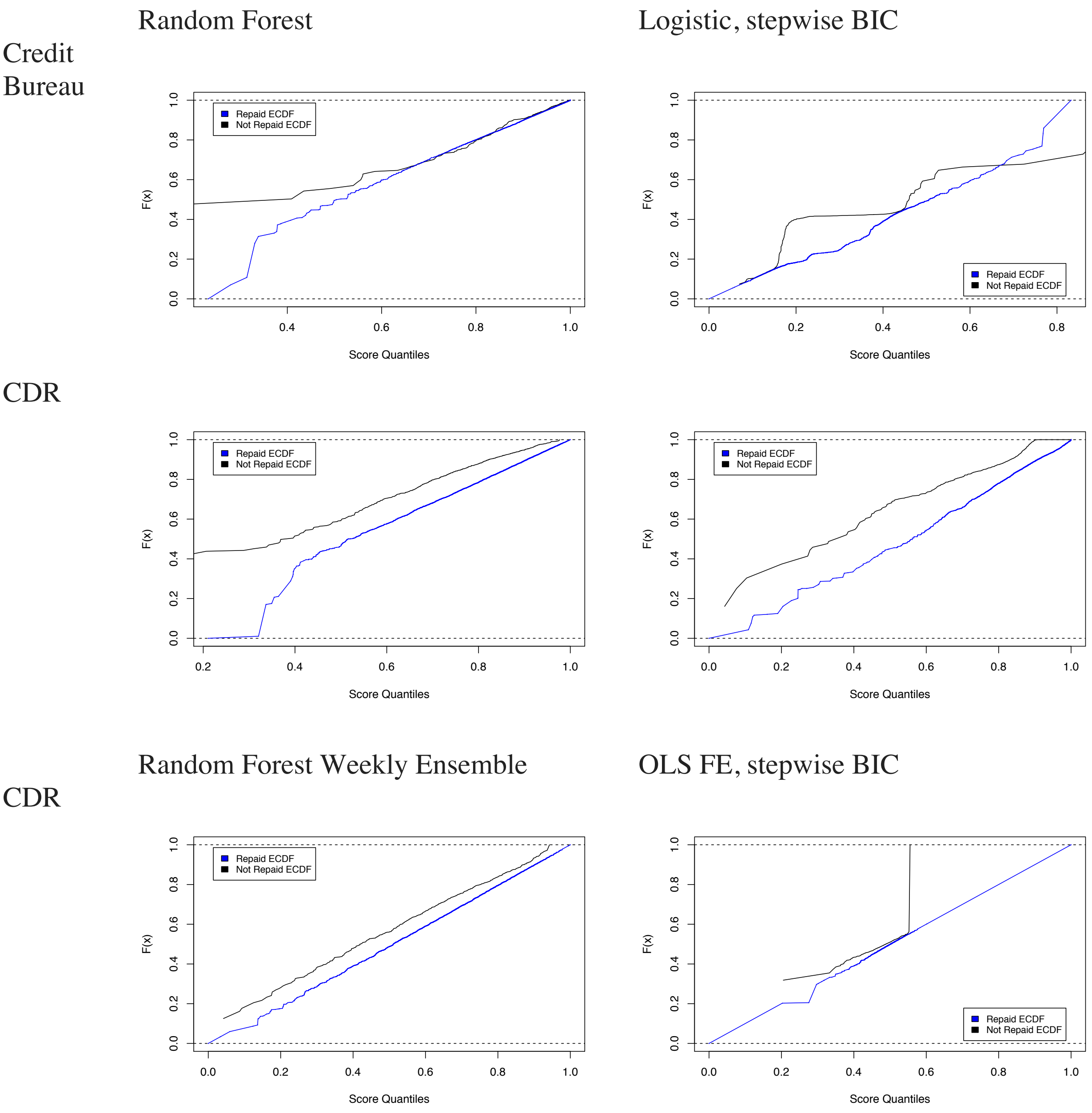
Random Forest Weekly Ensemble

OLS FE, stepwise BIC

CDR

Note: Cells computed based on out of sample predictions resulting from 5 fold cross validation.

Figure E: Empirical Cumulative Distribution Functions (CDF)



Random Forest

Logistic, stepwise BIC

Credit Bureau

CDR

Random Forest Weekly Ensemble

OLS FE, stepwise BIC

CDR

Note: Cells computed based on out of sample predictions resulting from 5 fold cross validation. X axis represents quantile in the distribution of predicted repayment scores, and Y axis represents the cumulative distribution function.

Table A: Parameter Estimates from Stepwise Regression

(a) Standard Stepwise Logistic

Stepwise Logistic

	Coefficient	SE
Number of places visited (Isaacman et al., 2011)	0.029***	(0.011)
Fraction of outgoing duration during working hours	0.959***	(0.283)
SD of daily outgoing SMS on days when it is above zero	0.791***	(0.200)
SD of daily outgoing SMS	-0.066	(0.063)
Max daily outgoing SMS	-0.004	(0.005)
Q80 of outgoing SMS on days when it is above zero	-0.038**	(0.018)
SD of daily outgoing calls	4.938***	(1.781)
SD of daily outgoing calls (between first and last days with activity)	-5.240***	(1.778)
SD of weekly outgoing duration	-16.786***	(2.761)
SD of weekly outgoing duration (between first and last weeks with activity)	16.874***	(2.765)
Q80/(Q80-Q50) of daily outgoing duration (between first and last days with activity)	0.157	(0.175)
Q50-Q20 of daily outgoing duration, on days when it is above zero	-0.001***	(0.0002)
Q60/(Q60-Q40) of daily outgoing duration (after first day with activity)	0.260**	(0.102)
Magnitude of strongest repeating temporal pattern in daily outgoing SMS (Fourier transform)	0.002	(0.003)
Magnitude of second strongest repeating temporal pattern in daily outgoing SMS (Fourier transform)	-0.002	(0.003)
Magnitude of third strongest repeating temporal pattern in daily outgoing SMS (Fourier transform)	0.006***	(0.002)
Ratio of magnitude of strongest repeating temporal pattern in daily outgoing SMS to sum of magnitudes of all other patterns (Fourier transform)	36.902**	(17.825)
Ratio of magnitude of strongest repeating temporal pattern in daily outgoing SMS to magnitudes of third strongest patterns, after first SMS (Fourier transform)	0.519**	(0.255)
Ratio of magnitude of strongest repeating temporal pattern in daily outgoing SMS to sum of magnitudes of all other patterns, after first SMS (Fourier transform)	-130.552**	(53.701)
Ratio of magnitude of strongest repeating temporal pattern in daily outgoing SMS to sum of magnitudes of all other patterns, between first and last weeks with activity (Fourier transform)	78.901	(51.873)
7 day lagged spearman autocorrelation of outgoing SMS (between first and last days with activity)	1.024***	(0.330)
Magnitude of second strongest repeating temporal pattern in weekly outgoing SMS (Fourier transform)	0.033***	(0.012)
Ratio of magnitude of strongest repeating temporal pattern in weekly outgoing SMS to sum of magnitudes of all other patterns (Fourier transform)	-0.692	(0.475)
Magnitude of strongest repeating temporal pattern in weekly outgoing SMS, between first and last days with activity (Fourier transform)	-0.004	(0.002)
Magnitude of second strongest repeating temporal pattern in weekly outgoing SMS, between first and last days with activity (Fourier transform)	-0.030**	(0.012)
2 month lagged autocorrelation of monthly outgoing SMS	14.030	(293.357)
Ratio of magnitude of strongest repeating temporal pattern in daily outgoing duration to sum of magnitudes of all other patterns (Fourier transform)	-12.447***	(3.335)
Ratio of magnitude of weekly temporal pattern in daily outgoing duration to sum of magnitudes of all other patterns (Fourier transform)	-4.377	(2.790)
Difference in magnitudes of strongest and second strongest repeating temporal pattern in weekly outgoing duration (Fourier transform)	0.00002**	(0.00001)
2 month lagged autocorrelation of monthly outgoing duration (after first month with activity)	-0.039	(1.098)
Magnitude of strongest repeating temporal pattern in daily outgoing calls (Fourier transform)	0.004***	(0.001)

Magnitude of third strongest repeating temporal pattern in weekly outgoing calls (Fourier transform)	0.007***	(0.002)
2 month lagged spearman autocorrelation of monthly outgoing calls (after first month with activity)	13.496	(325.088)
Correlation of monthly outgoing SMS with 1 month lagged outgoing duration	-5.137*	(2.899)
Correlation of monthly outgoing SMS with 2 month lagged outgoing calls	13.600	(320.260)
Missing: correlation of monthly outgoing SMS with 1 month lagged outgoing duration	-4.941*	(2.898)
Constant	6.980**	(2.922)
Observations	7,068	
Note: Standard errors computed based on the final logistic model; they do not adjust for model selection.	*p**p***p<0.01	

(b) Stepwise OLS FE

Stepwise OLS FE

	Coefficient	SE
Fraction of outgoing duration during working hours	0.098***	(0.025)
SD of daily outgoing calls	0.556	(0.453)
SD of daily outgoing calls, after first activity	-0.573	(0.449)
SD of weekly outgoing calls	0.001	(0.001)
SD of monthly calls out	-0.001***	(0.0003)
SD of daily SMS out, after first activity	-0.011***	(0.003)
Q80-Q20 of weekly calls out	-0.026***	(0.008)
Q70-Q30 of weekly outgoing calls, on weeks it is above zero	-0.001	(0.001)
Q80-Q50 of daily outgoing SMS, on days it is above zero	-0.007***	(0.002)
Median daily outgoing calls, on days it is above zero	-0.020***	(0.004)
Mean weekly outgoing calls, on weeks it is above zero	0.004***	(0.001)
Q60/(Q70-Q30) of daily duration out	0.057***	(0.015)
Magnitude of strongest repeating temporal pattern in daily outgoing SMS (Fourier transform)	0.0003***	(0.0001)
Missing: daily autocorrelation in outgoing duration	-0.411***	(0.134)
Observations	7,068	
Note: Demeaned by week prior to estimation. Standard errors computed based on the final model; they do not adjust for model selection or demeaning.	*p**p***p<0.01	

Table B: Comparison to Traditional Credit Scoring in Developed Settings

<u>Other Settings</u> **Traditional Credit Scoring**	Performance (AUC)				Default Rate	Features
	Within Time		Out of Time			
	All Models	Best Model	All Models	Best Model		
UK (Baesens et al., 2003)	0.500-0.758	0.668-0.758	-	-	10-25%	16-19 unspecified predictors
Belgium / Netherlands / Luxembourg (Baesens et al., 2003)	0.696-0.791	0.776-0.791	-	-	30-33%	33 unspecified predictors
Italian small and medium enterprises (Calabrese & Osmetti, 2013)	0.615-0.723	0.723	0.573-0.762	0.623-0.762	1-5%	Firm leverage, liquidity, profitability
Bosnia microfinance (Van Gool, Verbeke, Sercu, & Baesens, 2012)	0.679-0.707	0.707	-	-	22%	Demographics, earnings, capital, debt, loan

AUC represents the area under the receiver operating characteristic curve. Each study presents AUC estimates from multiple specifications; we present the range as well as the best out of sample AUC for each sample. These best estimates will tend to overstate performance on independent samples because they are selected based on performance on the test dataset. (Baesens et al., 2003) also reports results from publicly available Australian and German data sets, but the outcomes are not specified so they have been omitted.

Table C: Model Performance (10 Fold)

	Main Results			**Check**
Dataset:	**Standard Indicators**			**Offset Indicators**
Performance:	Out of Sample (10 fold CV)			Out of Time (train early period, test late)
Sample:	All	Has Bureau Records	No Bureau Records	All
	AUC	AUC	AUC	AUC
Baseline Model				
Credit Bureau				
Random Forest	0.516	0.510	-	0.507
Logistic, stepwise BIC	0.566	0.565	-	0.550
Our Models				
Phone indicators (CDR)				
Random Forest	0.716	0.713	0.730	0.631
Logistic, stepwise BIC	0.765	0.762	0.789	0.595
Phone indicators, within-week variation (CDR-W)				
Random Forest Weekly Ensemble	0.619	0.616	0.632	0.641
OLS FE, stepwise BIC	0.637	0.635	0.655	0.593
Combined				
Credit bureau and phone indicators				
Random Forest	0.717	0.716	-	0.642
Logistic, stepwise BIC	0.778	0.775	-	0.616
Credit bureau and phone indicators, within-week variation				
Random Forest Weekly Ensemble	0.623	0.621	-	0.639
OLS FE, stepwise BIC	0.650	0.648	-	0.586
Default Rate	11%	12%	10%	
N	7,068	6,043	1,025	6,975

Standard indicators evaluate out of sample performance using 10-fold cross validation, averaged over several fold draws. Offset indicators are derived from only half of the data (the first half for early transitions; the last half for late transitions); the out of time model is estimated on the early half of transitions and tested on the late half. AUC represents the area under the receiver operating characteristic curve. For middle two columns, model is trained on all individuals except the omitted fold, and performance is reported for the given subsample within the omitted fold.

Table D: Model Performance (alternate metrics, out of sample)

ScopeName	MethodName	AUC	H	R2.mean	R2.min	R2.max	Accuracy	MSPE
Credit Bureau	Logistic Stepwise	0.565	0.013	0.005	0.002	0.013	0.886	0.100
Credit Bureau	Random Forest	0.516	0.007	0.001	0.000	0.002	0.886	0.111
Phone Indicators	Logistic Stepwise	0.760	0.129	0.077	0.058	0.110	0.883	0.094
Phone Indicators	OLS FE	0.633	0.049	0.021	0.000	0.044	0.886	0.100
Phone Indicators	Random Forest	0.710	0.093	0.059	0.035	0.077	0.887	0.095
Phone Indicators	Random Forest Weekly Ensemble	0.616	0.038	0.018	0.008	0.028	0.886	0.099
Credit Bureau & Phone Indicators	Logistic Stepwise	0.772	0.145	0.090	0.059	0.114	0.883	0.093
Credit Bureau & Phone Indicators	OLS FE	0.645	0.054	0.025	0.010	0.046	0.886	0.099
Credit Bureau & Phone Indicators	Random Forest	0.711	0.095	0.060	0.038	0.083	0.887	0.095
Credit Bureau & Phone Indicators	Random Forest Weekly Ensemble	0.618	0.039	0.018	0.007	0.030	0.886	0.099

Note: All measures are out of sample. AUC represents the area under the ROC curve; H represents the H-measure (Hand, 2009). R2 measures aggregated over multiple fold draws. Accuracy is computed at the threshold 0.5; it is only informative of performance at that threshold while the other measures take into account performance over the range of thresholds. MSPE can be swayed by absolute differences, which are not relevant for the decision problem we consider: when a lender is setting an approval threshold only relative differences matter. An outlier observation was removed for computing the MSPE of the OLS FE models.