

Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment

Daniel Björkegren¹ and Darrell Grissen²

ABSTRACT

Many households in developing countries lack formal financial histories, making it difficult for banks to extend loans, and for potential borrowers to receive them. 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 loan default, using call records matched to loan outcomes. In a middle income South American country, individuals in the highest quintile of risk by our measure are 2.8 times more likely to default than those in the lowest quintile. On our sample of individuals with (thin) financial histories, our method outperforms models using credit bureau information, both within time and when tested on a different time period. The method forms the basis for new forms of lending that reach the unbanked.

This draft: February 16, 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. 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 small 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). One reason may be that households lack access to credit.

Traditional approaches to improving access to finance in the developing world have done so physically, either replicating institutions from wealthier societies such as bank branches and credit bureaus (World Bank, 2014), or creating new institutions such as microfinance. However, it is costly to physically provide small loans, especially to remote populations. Many remain unserved: 2 billion people lack bank accounts (Demirguc-Kunt, Klapper, Singer, & Van Oudheusden, 2014). And even 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, modern developing societies have new tools at their disposal. Mobile phones provide widespread access to digital services. Mobile money enables financial transfers at close to zero cost (Suri, Jack, & Stoker, 2012). Mobile money can be used for savings: simply keep a balance on the account. Mobile money can also be used to provide credit: just transfer the loan amount to a recipient's phone number, and ask that they repay. However, the recipient may not repay. That problem is the subject of this paper. This paper develops and assesses a low cost method to identify profitable investments, using information on potential borrowers that is already being collected by mobile phone networks.

In developed countries, banks have access to robust information on borrower reputation through credit bureaus, which aggregate information on an individual's historical management of credit. The credit bureau model has been copied in many developing countries (de Janvry, McIntosh, & Sadoulet, 2010; Luoto, McIntosh, & Wydick, 2007), but many remain sparse: few households in developing countries interact with the formal institutions that generate the necessary data. As a result, lenders have very little formal information on potential borrowers. This is particularly problematic, as lenders have little recourse

if a borrower were to default: borrowers have little in the way of collateral, and systems for legal enforcement are limited.

Although unbanked households lack the formal records needed for traditional credit scores, 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. In 2011, there were 4.5 billion mobile phone accounts in developing countries (ITU, 2011). Operator records are already being collected at close to zero cost, and 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). Björkegren (2010) proposed using indicators derived from this data to predict loan repayment. This idea has received attention following the publicity of an earlier working paper version of this paper (NPR, 2015), and this method is already forming the core of emerging digital credit products. However, this paper is the first to fully describe that insight and evaluate its performance.³

There are many straightforward indicators of behavior that are plausibly related to loan repayment. For example, a responsible borrower may keep their phone topped up to a minimum threshold so they have credit in case of emergency, whereas one prone to default may allow it to run out and depend on others to call them. Or, an individual whose calls to others are returned may have stronger social connections that allow them to better follow through on entrepreneurial opportunities.

This paper demonstrates that indicators of behavior derived from mobile phone transaction records are predictive of loan repayment. From raw transaction records, we create a method to extract approximately 5,500 behavioral indicators that have some intuitive link to repayment. Our approach thus differs from Blumenstock et al. (2015), who generate behavioral indicators from similar data with a data mining approach that is agnostic towards the outcome variable. Like an earlier working paper version of

³ Pedro, Proserpio, and Oliver (2015) find that defaulting on a loan is correlated with later calling behavior, but do not answer whether calling behavior prior to a loan can predict default.

this paper, Björkegren and Grissen (2015), we tailor our approach towards indicators intuitively linked to repayment, as implementation partners can be wary of ‘black box’ methods, and indicators with a theoretical link are more likely to have a stable relationship to repayment.

Since being proposed (Björkegren, 2010), similar methods have begun to be used 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). However, there is little evidence on how to approach this emerging revolution. 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. First, it 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.

This paper documents our method, and provides evidence on its ability to predict risk.

We evaluate the performance of the approach with data from a telecom in a middle income South American country where only 34% of adults have bank accounts but 89% of households have mobile phones. The telecom is transitioning subscribers from prepaid to postpaid plans, which entails an extension of credit. This setting has two crucial features. First, in the exploratory phase we observe, the telecom extended credit permissively, with only minimal fraud checks. As a result, we observe outcomes for the full population of individuals who might conceivably qualify, 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 differs 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 their loan, based on how they used their mobile phones before taking a loan, in a retrospective analysis. Our data includes 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. 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. Because traditional methods of assessing creditworthiness are often unavailable (bureau files are empty) or prohibitively costly (in person screening is challenging to scale), our method could be useful even if it underperformed credit bureau methods. However, we find that for our sample of formally banked but thin file consumers, credit bureau models perform poorly (Area Under the ROC Curve of 0.51-0.57), and are substantially outperformed by our models (AUC 0.61-0.76). Our method performs similarly well for unbanked consumers, who cannot be scored with traditional methods (AUC 0.63-0.77). Our models also perform within the (wide) range of published estimates of credit scoring in the literature (AUC 0.50-0.79).

Second, care must be taken to ensure stability over time. In practice, a lender would use past performance to train the model that disperses future loans. The performance of machine learning methods can deteriorate if the underlying environment shifts over time (Butler, 2013; Lazer, Kennedy, King, & Vespignani, 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 overfitting, by using only variation within each time period to differentiate loans (analogous to a form of temporal fixed effects). We demonstrate that this technique reduces intertemporal overfitting, 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 legacy traditional credit bureaus.

We conclude with a discussion on the logistics of implementing these methods in lending decisions. While our estimates are tied to telecom specific lending, this method can also be used to grant general loans. There are several avenues through which independent lenders can access this data. Transaction data can be gathered independently from a smartphone app, which scores borrowers and deliver loan; alternately, telecoms could produce credit scores and sell them through bureaus, or directly to lenders.

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.

As a side effect of operation, telecoms already gather rich information about subscribers' transactions, and thus could implement our method. We consider one particular application. As consumers in emerging economies have become wealthier, many telecoms have begun transitioning their subscribers

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>

from prepaid plans to postpaid subscriptions. 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. A telecom in a middle income South American country, with GDP per capita of approximately \$6,000, wanted to manage these risks among its prepaid subscribers.⁵ This subset of subscribers tends to have sparse formal financial histories. The telecom offered a 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 allowing customers to transition, and screened using only minimal fraud checks. If a subscriber did not pay their postpaid bill, their service was paused until it was paid, upon which point they were then transitioned back to a prepaid account.⁶ (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. 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). Our dataset is anonymized: 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

⁵ All results reported in US dollars.

⁶ Because the telecom could pause service, the loan 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.

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 a loan was granted, so we include only mobile phone transactions that precede the loan date. 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. We observe the median individual's phone usage for 16 weeks.

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 loan	107	14	112
Loan			
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 loans, and consider whether information that was available at the time of a loan 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. The telecom applied only minimal screening in offering the postpaid transition, so that we observe outcomes for the full sample of people to which the firm would conceivably consider extending credit.

The loan data provides 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) generate features from mobile phone data using an exhaustive, 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.

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, slope, and periodicity. In particular, individuals with different income streams are likely to have different periodicities in expenditure (formal workers may be paid on the first of the month; 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. Although social network measures may be predictive, 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 a loan is applied for (a person may be likely to repay this loan, 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, balance inquiry, top up,

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 loan 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:

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

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, interquartile ranges), and for ordinal sequences, change (slope) and periodicity (autocorrelation of various lags, and fundamental frequencies). 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, and the distance from the centroid to several points of interest. We also compute statistics that summarize pairs of sequences, including correlations, ratios, and lagged correlations (e.g., the correlation of minutes spoken with the previous day's balance).

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.

4. 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 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.

In contrast, many features derived from mobile phone usage have higher correlations, ranging up to 0.16. 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, 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 and loan characteristics			2
Age	0.073	2.35	
Female	-0.039	-1.26	
Loan term	-		
Loan size	-		
Credit Bureau			36
Has a credit bureau record	-0.022	-1.89	
Fraction of debt lost	-0.046	-3.86	
Phone usage			5,541
<i>Categories</i>	<i>High performing example feature:</i>		
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 80 th and 50 th quantile of SMS use on days SMS is used		
Other	0.100	3.07	542
	Number of important geographical location clusters		

Predicting Repayment

We estimate two standard models: random forests, and logistic regressions using a model selection procedure (stepwise search using the Bayesian Information Criterion), for bureau indicators and phone indicators (CDR).⁷ 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 logistic regression parameter estimates in Appendix Table A. Features used include the periodicity of usage (particularly, the

⁷ 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 (Breiman & Cutler, 2006).

magnitude of the first fundamental frequency to other frequencies), the fraction of duration spoken during the workday, variation in usage, and autocorrelation between calls and SMS.

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 loans 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.⁸ We form predictions differently from a standard fixed effect model. A standard model would include the fixed effect for each loan's 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 fixed effects, weighted by the proportion of loans granted in that week. We train a random forest analogously: we fit separate random forest models to each week of data, and combine them in an ensemble. When making a prediction for an individual, we weight each submodel by the proportion of loans granted in its week.⁹ 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.

Performance

Within Time

We measure how the method performs out of sample using 5-fold cross validation. As a first check, we consider how well the best model separates low and high risk borrowers. We report results from the most conservative model, the random forest weekly ensemble. Figure 1 shows how the default rate varies

⁸ If few loans are given in a week, we combine it with adjacent weeks.

⁹ When giving out loans, one could upweight more recent models to capture changes in conditions.

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.

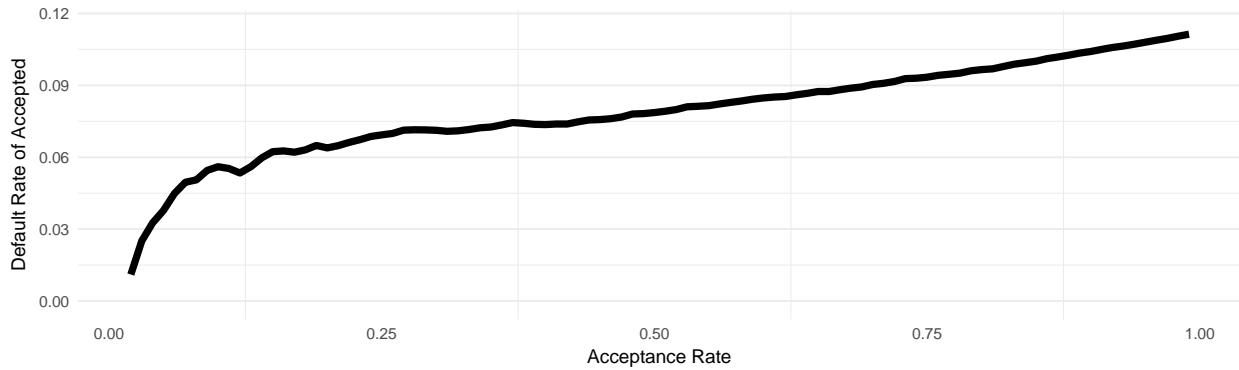


Figure 1: Default Rate by Proportion of Borrowers Accepted

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.

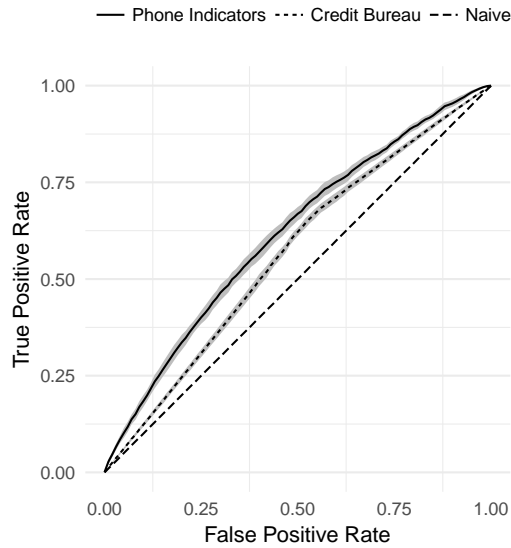


Figure 2: Receiver Operating Characteristic Curve

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; the area under this curve (AUC) is a summary of its performance. 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 present results for the entire sample, and then split into the subsamples that do and do not have credit bureau records. As suggested by the single variable correlations, in this population with thin files, credit bureau information does not perform especially well in predicting repayment (AUC 0.51-0.57). 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 (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.

Table 3: Model Performance

Dataset:	Standard Indicators			Offset
	Out of Sample (5 fold CV)			Indicators
Performance:				Out of Time (train early period, test late)
Sample:	All	Has Bureau Records	No Bureau Records	All
	AUC	AUC	AUC	AUC
<u>Baseline Model</u>				
Credit Bureau				
Random Forest	0.516	0.509	-	0.507
Logistic, stepwise BIC	0.565	0.565	-	0.550
<u>Our Models</u>				
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
<u>Combined</u>				
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 loans; the last half for late loans); the out of time model is estimated on the early half of loans 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. 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 loans into two; the early group that took out a loan 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 loan). 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.

All models apart from CDR-W Random Forest perform worse when tested on a different time period. 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). Phone indicator models using 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).¹⁰ However, all phone indicator models continue to outperform models using credit bureau data in our population (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 deterioration when tested out of time. 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).

¹⁰ 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).

5. DISCUSSION

Mobile phone data appears to quantify nuanced aspects of behavior that are typically considered soft, making these behaviors ‘hard’ and legible to formal institutions (Berger & Udell, 2006). Further, this data is 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.

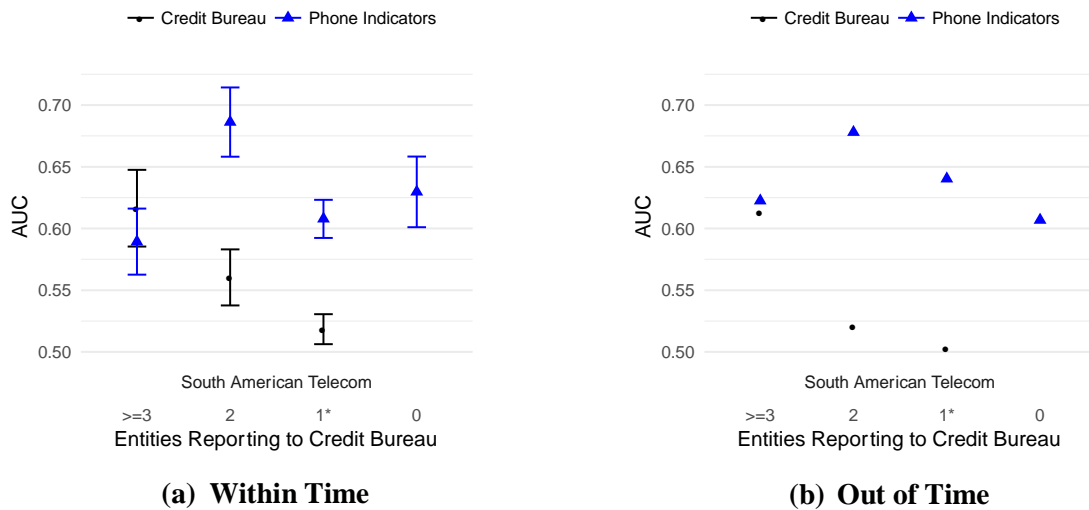


Figure 4: Performance by Level of Formalization

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.

Our approach can dramatically reduce the cost of screening individuals on the margins of the banking system. Current screening methods used in the developing world are costly, often relying on detailed interviews or peer groups, even for small loans. 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.¹¹ 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.¹² Alternately, telecoms can package this data into a credit score that can be used by third parties, either through mobile banking platforms or an independent credit bureau.¹³ A third implementation, a smartphone app, allows third parties to access usage data independently of telecom operators, and is being explored by several startups.¹⁴ These apps ask for permission to view call history and other behavioral data, and can collect real-time data for a set period.

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

¹¹ In addition to assisting with the transition to postpaid plans, this method can be used to extend credit for handset purchases, or to maintain consistent airtime balances. Many developing country operators offer small airtime loans like this; a scoring model could improve their provision.

¹² See for example, [Jumo](#).

¹³ See for example, [Cignifi](#).

¹⁴ See for example, [Tala](#) and [Branch](#).

that only consumers who consent are scored with the system.¹⁵ 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 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). Our sample of borrowers tends to use phones higher than average; we find a slight deterioration in the performance of the method when estimated on synthetic datasets that have been made sparser. We expect performance to improve when individuals are observed for longer time periods, with the richer data that can be gathered from smartphones, and as technology usage increases.

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

¹⁵ 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).

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. 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 et al., 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
- Björkegren, 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., & 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/globalindex>

- 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>
- 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.
- 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
Importance Plot: Top 20 Features



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

The above importance plot measures the mean decrease in the Gini coefficient of the top twenty features, which corresponds to the marginal impact of including the variable in the model.

Table A: Parameter Estimates from Stepwise Logistic Regression

Stepwise Logistic	Coefficient	SE
Duration.Out.Week.BetweenFirstAndLast.Mean.SD	16.874***	(2.765)
SMS.Out.Day.SD	-0.066	(0.063)
SMS.Out.Day.Max	-0.004	(0.005)
SMS.Out.Day.Periodicity.Magnitude.Rank0	0.002	(0.003)
SMS.Out.Day.Periodicity.Magnitude.Rank1	-0.002	(0.003)
SMS.Out.Day.Periodicity.Magnitude.Rank2	0.006***	(0.002)
SMS.Out.Day.Periodicity.MagnitudeRatio.Rank0_AllOtherRanks	36.902**	(17.825)
SMS.Out.Day.AfterFirst.NonZero.Mean.SD	0.791***	(0.200)
SMS.Out.Day.AfterFirst.Periodicity.MagnitudeRatio.Rank0_Rank2	0.519**	(0.255)
SMS.Out.Day.AfterFirst.Periodicity.MagnitudeRatio.Rank0_AllOtherRanks	-130.552**	(53.701)
SMS.Out.Day.BetweenFirstAndLast.Periodicity.MagnitudeRatio.Rank0_AllOtherRanks	78.901	(51.873)
SMS.Out.Day.NonZero.Q80	-0.038**	(0.018)
SMS.Out.Day.BetweenFirstAndLast.AutoCorrelation.L7.Spearman	1.024***	(0.330)
SMS.Out.Week.Periodicity.Magnitude.Rank1	0.033***	(0.012)
SMS.Out.Week.Periodicity.MagnitudeRatio.Rank0_AllOtherRanks	-0.692	(0.475)
SMS.Out.Week.BetweenFirstAndLast.Periodicity.Magnitude.Rank0	-0.004	(0.002)
SMS.Out.Week.BetweenFirstAndLast.Periodicity.Magnitude.Rank1	-0.030**	(0.012)
SMS.Out.30Day.AutoCorrelation.L2.Spearman	14.030	(293.357)
Duration.Out.Day.Periodicity.MagnitudeRatio.Rank0_AllOtherRanks	-12.447***	(3.335)
Duration.Out.Day.Periodicity.MagnitudeRatio.Weekly_AllOtherRanks	-4.377	(2.790)
Duration.Out.Day.BetweenFirstAndLast.Q80.Q80minusQ50	0.157	(0.175)
Duration.Out.Day.BetweenFirstAndLast.NonZero.Q50minusQ20	-0.001***	(0.0002)
Duration.Out.Day.AfterFirst.Q60.Q60minusQ40	0.260**	(0.102)
Duration.Out.Week.Mean.SD	-16.786***	(2.761)
Duration.Out.Week.Periodicity.MagnitudeDifference.Rank0_Rank1	0.00002**	(0.00001)
Duration.Out.30Day.AfterFirst.AutoCorrelation.L2.Pearson	-0.039	(1.098)
Calls.Out.Day.SD	4.938***	(1.781)
Calls.Out.Day.Periodicity.Magnitude.Rank0	0.004***	(0.001)
Calls.Out.Day.BetweenFirstAndLast.SD	-5.240***	(1.778)
Calls.Out.Week.Periodicity.Magnitude.Rank2	0.007***	(0.002)
Calls.Out.30Day.AfterFirst.AutoCorrelation.L2.Spearman	13.496	(325.088)
Geography.ImportantPlaces.DaysUsed.Number	0.029***	(0.011)
(SMS.Out Duration.Out) By30DayL0L1.Correlation.Pearson	-5.137*	(2.899)
(SMS.Out Calls.Out) By30DayL0L2.Correlation.Pearson	13.600	(320.260)
Duration.Out.WorkDay.Fraction	0.959***	(0.283)
(SMS.Out Duration.Out) By30DayL0L1.Correlation.Pearson.missing	-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 < 0.05 ** p < 0.01 *** p < 0.001

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		
<i>UK</i> (Baesens et al., 2003)	0.500-0.758	0.668-0.758	-	-	10-25%	16-19 unspecified predictors
<i>Belgium / Netherlands / Luxembourg</i> (Baesens et al., 2003)	0.696-0.791	0.776-0.791	-	-	30-33%	33 unspecified predictors
<i>Italian small and medium enterprises</i> (Calabrese & Osmetti, 2013)	0.615-0.723	0.723	0.573-0.762	0.623-0.762	1-5%	Firm leverage, liquidity, profitability
<i>Bosnia microfinance</i> (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.