Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment

Daniel Björkegren[1] and Darrell Grissen[2]

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. We evaluate our approach using call records matched to loan outcomes in two settings. In a middle income South American country, individuals in the highest quartile of risk by our measure are 7.4 times more likely to default than those in the lowest quartile. We benchmark performance on our sample of individuals with (thin) financial histories: our method performs no worse than models using credit bureau information. We also test our method among microfinance borrowers in a much poorer Caribbean country; our partner could have increased return on assets by at least 5% using our method. The method is already forming the basis for new forms of lending that reach the unbanked.

This draft: December 13, 2017. 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. This work was supported by the Stanford Institute for Economic Policy Research through the Shultz Fellowship in Economic Policy.

[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 these opportunities may remain untapped is if potential lenders have difficulty identifying the investments that will be profitable.

Developing country banks that would lend to poor consumers and small firms face several particular challenges. 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 developing country banks that would lend to small or informal businesses may have little recourse if a borrower were to default: borrowers have little in the way of collateral, and systems for legal enforcement are limited.

Traditional microfinance has presented one solution to the repayment problem, relying on community members to aid in selection and monitoring of loans.[3] However, it is not clear that the investments currently selected by microfinance have led to transformative effects for borrowers (A. Banerjee, Duflo, Kinnan, & Glennerster, 2014; A. Banerjee, Karlan, & Zinman, 2015; Karlan & Zinman, 2011).

This paper evaluates a new method to identify profitable investments, using information on potential borrowers that is already being collected by mobile phone networks.

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—

[3] Some researchers have also explored using the limited information already available to banks to identify borrowers; for example, (Schreiner, 2004) finds simple characteristics can complement the judgment of loan officers.

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). If indicators derived from this data are also predictive of creditworthiness, they may help banks identify profitable opportunities.

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. While Björkegren (2010) proposed using these types of behavior to predict repayment, and this insight has started to be used in industry, this is the first work to describe it in full and evaluate its performance.[4]

This paper demonstrates that indicators of behavior derived from mobile phone transaction records are predictive of loan repayment. From raw transaction records, we 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 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. We evaluate

[4] The only paper of which we are aware is an earlier, working paper version of this paper (Bjorkegren & Grissen, 2015). 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. Several companies have been testing similar methods similar to ours, including Cignifi, Tala, Branch, and Jumo.

our approach in two settings, which span the levels of development and potential implementation partners where we envision this insight being used.

Mobile telecoms already possess the data to score applicants. We first test the approach with data from a telecom in a middle income South American country. The telecom is transitioning subscribers from prepaid to postpaid plans, which entails an extension of credit. This country has a credit bureau, and 85% of applicants have a file, though as in many emerging markets these files can be sparse. Crucially, in this exploratory phase the telecom extended credit permissively, with only minimal fraud checks, so that we observe outcomes for the full population of individuals who might conceivably qualify. 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. Individuals in the highest quartile of risk by our measure are 7.4 times more likely to default than those in the lowest quartile. Our method is predictive for both banked and unbanked consumers. For unbanked consumers, who cannot be scored using traditional methods, we obtain an area under the receiver operating characteristic curve (AUC) of 0.77. For our sample of formally banked but thin file consumers, we can compare our performance to an industry benchmark: our best model obtains an AUC of 0.76, while models trained on credit bureau information obtain 0.60. We find that information gathered by the bureau is only slightly complementary to that in our indicators (performance increases from 0.76 to 0.77 when bureau information is added). In a suggestive test, our models also perform no worse than bureau models in this population when estimated and tested on different time periods.

There are also 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. We evaluate the potential of this latter approach with a microfinance institution (MFI) in a poor Caribbean country that does not have a credit bureau, by linking repayment records to transaction records from the

dominant telecom. Like many lenders operating in low information environments, the MFI screens applicants manually, which is costly for the small loans we study. In this sample, we observe borrowers' phone usage for a short time, and do not obtain the same performance as in the South American sample. However, we find that our approach has some predictive power when added to the screening already performed. A simple calibration suggests that screening borrowers with this method could have improved return on assets by at least 5%.

Our results have important implications for financial inclusion. While traditional approaches to improving access to finance in the developing world have focused on replicating institutions from wealthier societies such as bank branches and credit bureaus (World Bank, 2014), it has become widely recognized that modern developing societies have new tools at their disposal. Mobile phones provide widespread access to digital services, and the increasing prevalence of mobile money enables low cost financial transfers. Our findings suggest that nuances captured in the use of mobile phones themselves have the potential to alleviate information asymmetries, and thus can form the basis of new forms of low cost lending. These tools together enable a new ecosystem of digital financial services, which may reshape 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.[5] EFL identified two different partners that were interested in exploring alternate methods of assessing creditworthiness.

South American Telecom

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

[5] 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

in emerging economies have become wealthier, many telecoms have begun transitioning their subscribers 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. We worked with a telecom in a middle income South American country, with GDP per capita of approximately $6,000, that wanted to manage these risks among its prepaid subscribers.[6] This subset of subscribers tends to have sparse formal financial histories. It was considering predicting default using either traditional bureau data or our alternative method. As a first step, 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.[7] 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). We matched subscribers to their financial histories based on an identifier that was encrypted using a one-way hash.

[6] All results reported in US dollars.

[7] 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.

Caribbean Microfinance Institution

We also identified a local microfinance institution that provides individual loans to microentrepreneurs, in a low income Caribbean country with a GDP per capita of approximately $800. We focus on the bank's small loan product, with median size $179, repaid over a term of 5 months. To determine whether to grant a loan, loan officers typically travel to prospective clients to gather information about their businesses (balance sheet, income statement, and financial ratios) and households (number of dependents, health, education, etc.). The bank approves approximately 90% of first time applicants. After approval, borrowers make regular payments. If a borrower fails to make a payment for 90 days, they are considered to be in default. 20% of loans ended in default in our sample.

EFL obtained linked phone and outcome data for a sample of borrowers in two steps. First, anonymous loan records were obtained from the microfinance institution for a sample of borrowers. These records include basic demographics such as age and gender, the terms of the loan provided, and whether the loan was defaulted on. (The bank was unable to share other information gathered in loan officer interviews due to confidentiality.) We matched borrowers to their phone accounts based on an encrypted identifier derived from the phone number, using a one-way hash. In the second step, mobile phone transaction records (CDR) were obtained for these matched accounts from the dominant telecom in the country. Over 90% of loan accounts matched to a phone account.

In both settings, the mobile phone data include metadata for each call and SMS, with identifiers for the other party, time stamps, tower locations, and durations. The Caribbean data also includes top-ups, balances, data access, charges, and handset models used; this information is not in the South American data. Neither dataset includes 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. The South American sample is wealthier, and although 85% have 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. The median South American individual uses phones much more than in the Caribbean sample: placing 1.6 times the number of calls, 4.2 times the duration, and 81 times the number of SMS in an average week. However, variation within each sample is also high. In the Caribbean sample we observe spending patterns, and find that subscribers keep low balances: the median subscriber has an average weekly balance of $0.09 but tops up $0.90 and spends $0.95. In the Caribbean data, we face a hard test, as we observe phone behavior for only a short period (a mean of 27 days): we expect performance to improve if behavior is observed over a longer time period.

Table 1: Description of Individuals

	South American Telecom			Caribbean Microfinance		
	Mean	SD	Median	Mean	SD	Median
Country GDP per capita (Approx.)	$6,000			$800		
Borrowers						
Gender is female	39%	-	-	71%	-	-
Age (years)	35.8	12.8	34.0	34.8	9.7	34.2
Has a mobile phone	100%	-	-	100%	-	-
Credit bureau record	85%	-	-	0%	-	-
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	23.6	22.9	16.5
Calls out, minutes	41.6	39.9	32.0	11.9	13.8	7.7
SMS sent	31.3	26.3	24.4	2.5	6.1	0.3
Top-ups	-	-	-	$1.45	$1.73	$0.90
Spend	-	-	-	$1.51	$1.87	$0.95
Balance	-	-	-	$0.22	$0.76	$0.09
Days of mobile phone data preceding loan	107	14	112	27	14	27
Loan						
Size	-	-	-	$192	$83	$179
Term (days)	-	-	-	146	22	150
Default	11%	-	-	20%	-	-
N	7,068			1,302		

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. However, the microfinance institution applied their existing screening approach, so that we do not observe outcomes for the approximately 10% of individuals who they screened out. In that setting we thus face a harder test: what our method adds in addition to the existing screening.[8]

The loan data provides an indicator for whether a particular borrower repaid their obligation (we use our partners' definitions: for the telecom, 15 days past due, for the microfinance institution, 90 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 balances and expenses are managed, such as variation, slope, and periodicity. In particular, individuals with different income streams are likely to have different periodicities in balances

[8] Patterns of default may differ if credit were extended to those currently denied loans.

and 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:

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

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). 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 patterns of top-up and depletion, as well as strength and diversity of contacts) as well as other measures (intensity and distribution of usage over space and time, mobility, the pattern of handset use). For each feature, we also add an indicator for whether that individual is missing that feature. Altogether, in each setting 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 in the two settings. We start with the South American telecom.

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, despite the fact that the South American telecom data does not include balances and top ups. 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).

We also evaluate these features for the Caribbean MFI. Since that sample has already been screened by the bank's process, for that setting these correlations measure the value of information after the screening that was already performed. In addition to demographics, we also include the terms of the loan (loan size and duration), which are functions of information observed by loan officers but not by us. We find that bank features have low correlation with default (0.02-0.06), which would be consistent with the already using the information it already collects. In contrast, we also find that features derived from mobile phone usage have higher correlations (magnitudes up to 0.13), suggesting that they are valuable even in addition to the screening already performed.

The table highlights particular features that perform well, including the correlation of the daily balance with phone spending two days ago; how long the account could have gone without a top up, which captures whether subscribers maintain a balance, and the number of important geographical location clusters where the phone is used.

Table 2: Individual Features

	South American Telecom			Caribbean Microfinance		
	Correlation with repayment	t-stat	Number of Features	Correlation with repayment	t-stat	Number of Features
Demographics and loan characteristics			**2**			**4**
Age	0.073	2.35		0.063	2.28	
Female	-0.039	-1.26		0.039	1.41	
Loan term	-			0.052	1.88	
Loan size	-			0.018	0.65	
Credit Bureau			**36**			
Has a credit bureau record	-0.022	-1.89		-		
Fraction of debt lost	-0.046	-3.86		-		
Phone usage			**5,541**			**5,602**
Categories	*High performing example feature:*			*High performing example feature:*		
Periodicity	-0.163	-5.27	796	-0.134	-4.89	846
	SMS by day, ratio of magnitudes of first fundamental frequency to all others			Autocorrelation in max distance spanned, between this week and two weeks ago		
Slope	0.126	4.06	44	-0.050	-1.81	35
	Slope of daily calls out			Slope of data usage		
Correlation	0.111	3.57	224	-0.083	-3.02	502
	Correlation in SMS two months ago and duration today			Correlation of today's balance with spending two days ago		
Variance	-0.104	-3.34	4,005	0.098	3.54	3,103
	Difference between 80^{th} and 50^{th} quantile of SMS use on days SMS is used			How long could account have gone without a top up, difference between 80^{th} and 50^{th} quantile		
Other	0.100	3.07	542	0.106	3.86	1,116
	Number of important geographical location clusters			Distance from primary workday location to the capital		

Predicting Repayment

We next use multiple features together to predict repayment. We estimate random forests, and logistic regressions using a model selection procedure (stepwise search using the Bayesian Information Criterion).[9] 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

[9] 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).

estimates in Appendix Table A. The South American telecom data does not include balances and top ups; features used include the periodicity of usage (particularly, the 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. The Caribbean MFI models use features based on geography and spending, as well as use of discount timebands. We first discuss performance for the South American telecom, and then the Caribbean MFI. All performance estimates reported are out of sample.

South American Telecom

We measure how the method will perform 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, focusing on the South American sample with logistic regression. Figure 1 shows how the default rate varies with the fraction of borrowers accepted (where borrowers with lowest predicted default are accepted first). Individuals with the highest 25% of risk scores are 7.4 times more likely to default than those with the lowest 25%.

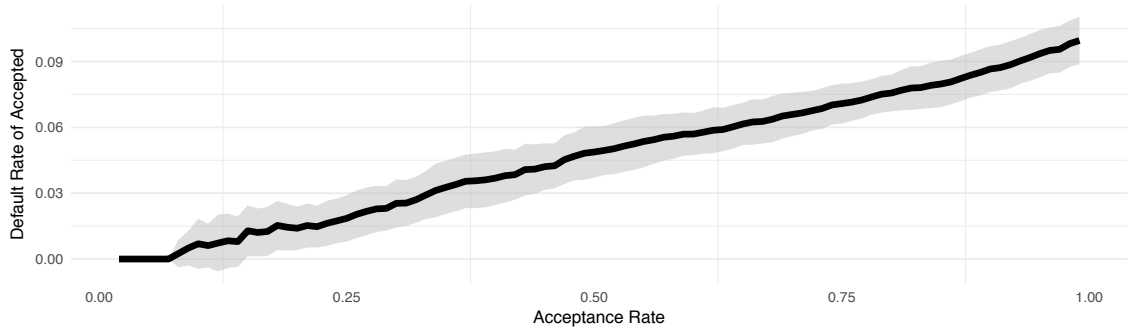


Figure 1: Default Rate by Proportion of Borrowers Accepted
South American sample with phone indicators and stepwise logistic regression. Line shows mean, and ribbon standard deviation, of results from multiple fold draws.

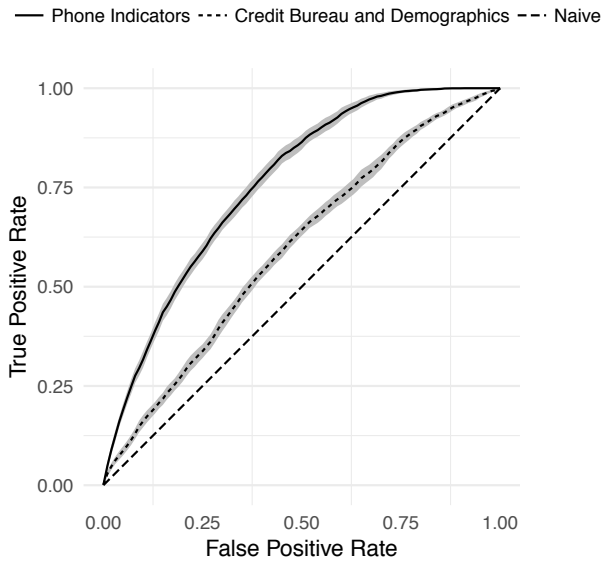


Figure 2: Receiver Operating Characteristic Curve

South American sample with stepwise logistic regression. 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 and the best model using indicators derived from phone data (stepwise logistic). We show results for a variety of specifications in Table 3, measuring performance with AUC and the H measure (Hand, 2009) which has been proposed as a more robust measure. We present results for the entire sample, and then split into the subsamples that do and do not have credit bureau records.

Benchmark models using only demographic and loan characteristics perform poorly (AUCs in the range of 0.50-0.58). As suggested by the single variable correlations, credit bureau information does not perform especially well in predicting repayment for this population (AUC 0.53-0.60). In contrast, models built on phone indicators are predictive, reaching AUCs of 0.70-0.76. This performance is also in the range of a sample of published AUC estimates for traditional credit scoring on traditional loans in more developed settings (0.67-0.79, within time estimates, shown in Appendix Table B). Combining our indicators with

information from the credit bureau slightly boosts max performance from 0.76 to 0.77, suggesting that the information gathered by the bureau is only slightly complementary to that collected by our approach.

Table 3: Model Performance

	South American Telecom						Caribbean Microfinance	
Sample:	All		Has Credit Bureau Records		No Credit Bureau Records		All	
	AUC	H-measure	AUC	H-measure	AUC	H-measure	AUC	H-measure
Baseline Models								
Demographics and loan characteristics*								
Random Forest	0.500	0.004	0.500	0.004	0.500	0.011	0.525	0.027
Logistic, stepwise BIC	0.576	0.018	0.584	0.026	0.581	0.035	0.500	0.001
Credit Bureau (all variables) and demographics								
Random Forest	0.536	0.015	0.533	0.017	-		-	
Logistic, stepwise BIC	0.595	0.028	0.597	0.031	-		-	
Our Model								
Phone indicators								
Random Forest	0.709	0.094	0.707	0.096	0.721	0.143	0.553	0.035
Logistic, stepwise BIC	0.759	0.131	0.758	0.134	0.767	0.182	0.570	0.044
Combined								
Credit bureau, phone indicators, and demographics								
Random Forest	0.709	0.093	0.707	0.095	-		-	
Logistic, stepwise BIC	0.770	0.143	0.767	0.144	-		-	
Default Rate	11%		12%		10%		20%	
N	7,068		6,043		1,025		1,302	

*: loan characteristics included only in Caribbean MFI sample. Out of sample performance is estimated using 5-fold cross validation, averaged over fold draws. AUC represents the area under the receiver operating characteristic curve; we also report the H measure (Hand, 2009). 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.

Caribbean MFI

We evaluate cross-validated performance for the Caribbean MFI in the final column of Table 3. Since that sample has already been screened by internal criteria, it omits the roughly 10% of applicants who were denied credit. In that sample, our measures reveal how predictive models are when used in addition to this existing screening. The information already available to the bank that we observe (demographics and characteristics of the loan) has limited predictive power (AUC 0.50-0.53), which would be expected if the bank took that information into account in its internal screening. Other information available to the bank but not to us may perform better on this sample. Our method does have some predictive power, though considerably less than in the telecom sample (AUC 0.55-0.57); this could result from many differences (such as substitutability with existing screening and observing individuals' phone usage for a shorter period). To gauge the usefulness of this predictive power, we evaluate its potential impacts on portfolio performance. [10]

Impact on Portfolio Performance

The key statistic for our portfolio performance estimates is the ratio of profits lost on a default to profits gained on a repaid loan. We build up this statistic using our partner bank's rates and cost benchmarks. On this basket of loans, the bank charged an annual interest rate of 60% (high interest rates are common in microfinance). Based on an industry benchmark cost of capital of 10% and operating expenses of 15%, the bank would expect to earn profits of 35% per year on a repaid loan, or 14.6% of principal for the typical 5 month loan in our sample.[11] When a loan is defaulted on, the firm not only gives up the profits from interest, but may lose part of the principal. We use an estimate from our partner that loss in case of default is 33.9%

[10] The benefits of switching a subscriber to postpaid are less straightforward to quantify, so we do not model the performance implications for the telecom.

[11] Note that many banks do not compound interest for short loans like this, so a loan of half the duration is charged half the total interest.

of principal, plus the cost of capital and expenses.[12] These figures suggest the ratio of profits lost on a default is 3.0 times the profits gained on a repaid loan.

We consider the impact of using our method on return on assets (profit divided by portfolio size), which at baseline is 3.0%. We rank borrowers in terms of risk (using the stepwise logistic model which appeared to perform best), and consider the return on assets as a function of the acceptance rate, in Figure 3. Performance increases under a wide range of acceptance rates; the profit maximizing acceptance rate of 94% would increase return on assets by 5%, to 3.2%. Being more selective would increase the return on assets further: to 3.8% if accepting 75% of loans, or to a maximum of 5.6% if accepting 34% of loans. These results hold constant the bank's existing screening methods and in that sense represent a lower bound.

Alternately, a bank can use improved screening to achieve social objectives, including expanding lending to the unbanked. Because we only observe borrowers who were accepted by the bank, we are unable to evaluate the performance of our method on this population; however, we discuss it below.

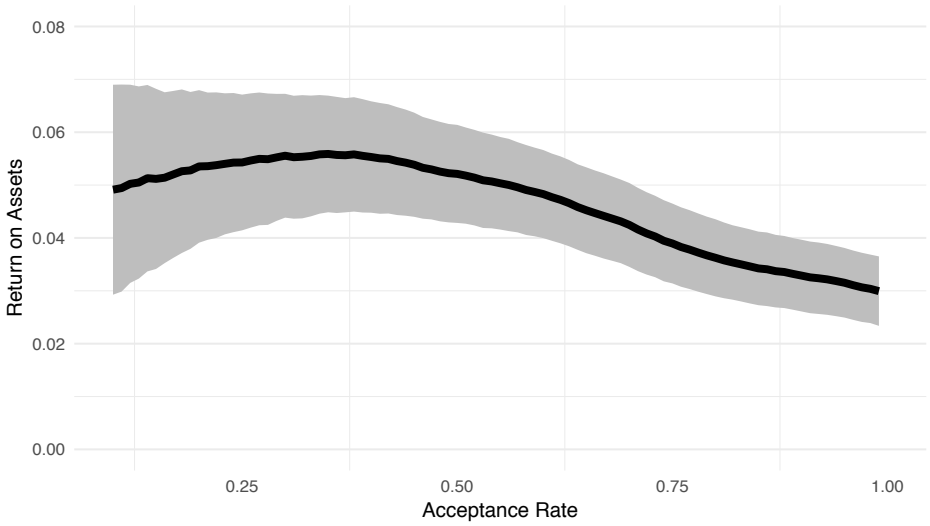


Figure 3: Estimated Return on Assets by Proportion of Loans Accepted

The black line shows the mean, and the gray ribbon traces out one standard deviation of the results from multiple fold draws

[12] This is the loss under default for this bank's larger microentrepreneur product, for which we were able to obtain financial data. For our smaller loans, we expect recovery to be lower and thus our method to improve performance more.

5. DISCUSSION

Mobile phone data quantifies 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—many of the South American clients have formal financial histories, and our Caribbean sample received loans from an MFI.

We summarize the performance of our method by level of formalization in Figure 4. The performance of credit bureau models deteriorates slightly as we move from individuals with rich financial histories (3 or more entities contributing reports to the bureau) to those with sparser histories. In contrast, our method achieves comparable performance across levels of formalization (though we obtain lower performance with the Caribbean MFI).

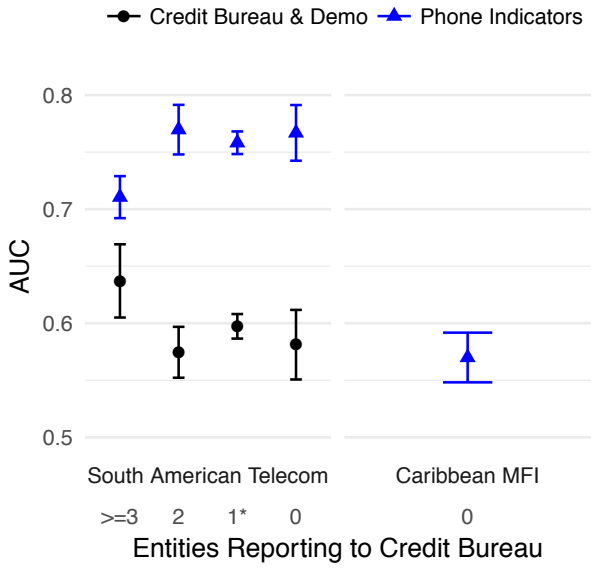


Figure 4: Performance by Level of Formalization

1*: either one entity reporting, or has a file at the credit bureau which may include previous activity but zero entities are currently reporting. South American sample with stepwise logistic regression. Model is 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. The point shows the mean, and error bars standard deviation, of results from multiple fold draws.

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 anonymously over a mobile phone network. Thus it could enable new forms of lending that do not require the full structure of current branch lending. Such implementations may reduce the cost of serving existing markets, but also make it profitable to serve consumers outside the current financial system.

Implementation

As demonstrated with the South American partner, this approach can be used to extend telecom-specific credit, within the firms that already possess the necessary data. That firm is evaluating this model to extend the credit implicit in a postpaid plan. However, many developing country households also find it difficult to save to purchase a handset or to maintain a consistent airtime balance. With a user's permission, a scoring model can be used to provide credit for purchase of a handset or airtime.[13]

However, we anticipate broader applications. Mobile money services which are increasingly popular in developing countries make it cheap to deliver a loan and collect payment. With regulatory approval, telecoms may connect to the banking sector, and offer loans to consumers.[14] 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.[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. Once data has been collected and an individual has been scored using this approach, mobile

[13] Many developing country operators already 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.

money can be used to provide credit and to collect repayment. Although smartphone use is growing even among the poor, this implementation will misses current feature phone users.

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 encrypting 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.[17] 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).

Stability Over Time

Our results are estimated out of sample within the same time period, but when implemented, a model trained on past data will be used to predict future repayment. While we lack a long panel, we perform a suggestive test of 'out of time' performance for the South American telecom sample in the Appendix (the MFI sample is not observed long enough to perform this test). We find that the performance of nearly all

[17] 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).

methods deteriorates, but in this sample of borrowers with thin financial histories our method continues to perform better than models relying on credit bureau data. (See Appendix Table C and surrounding discussion.)

Stability Across Contexts

A related question is whether the best model is portable across regions. Our contexts are very different: the South American country is 7.5 times wealthier than the Caribbean country, the types of credit are different, and the Caribbean sample has been prescreened through interviews. It would thus not be surprising if a model estimated in one context were not predictive in another. In Appendix Table D, indeed we find that our model is predictive in both settings when trained with local data, but a model trained in one context is not universal. We expect models to be more portable across more similar settings.

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. 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, which forms the basis of new forms of low cost lending. These tools together enable 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.

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

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.

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.

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.

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

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.

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

Schreiner, M. (2004). Scoring Arrears at a Microlender in Bolivia. *Journal of Microfinance / ESR Review*, *6*(2), 65–88.

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

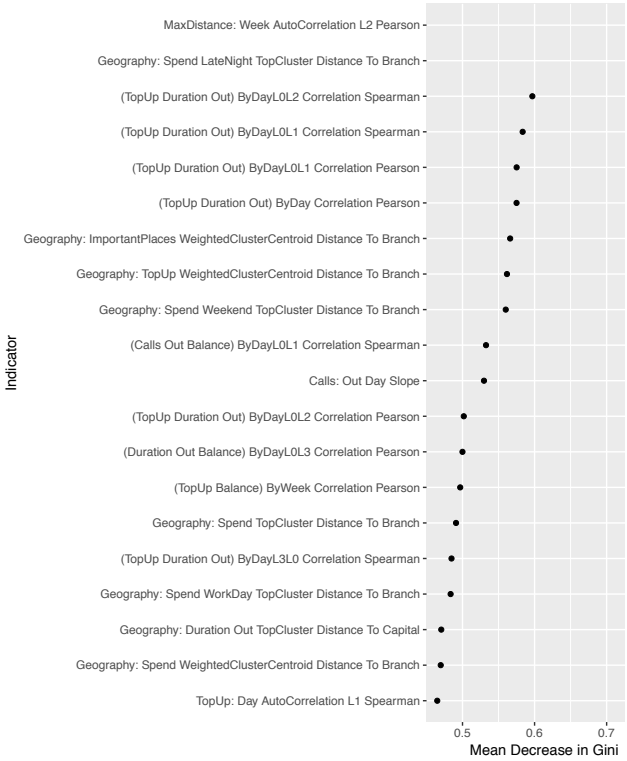
Figure A: Random Forest Estimates

South American Telecom
Importance Plot: Top 20 Features



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

Caribbean Microfinance
Importance Plot: Top 20 Features



Ensemble of 500 trees. Mean nodes per tree: 125.

The above importance plots measure 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
South American Telecom		
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	

Table A: Parameter Estimates from Stepwise Logistic Regression (Continued)

Caribbean Microfinance

Geography: Distance from Top Duration Cluster to Capital	0.005***	(0.002)
Data Usage: 30 Day Autocorrelation	4.375	(4.008)
Daily Spend: Q60/(Q80-Q20)	0.569***	(0.192)
Balance Depletion: SD(Time that could have gone without a top up)	-0.00000	(0.00000)
Balance Depletion: Time that could have gone without a top up: Q80-Q50	0.00002***	(0.00000)
Geography: 3 Week Autocorrelation in Max Distance Between Visited Locations	-1.044***	(0.253)
Correlation Between Calls Out 2 Days Ago and Balance Today	-0.961***	(0.341)
Fraction of duration within 2 minutes of a discount discontinuity X fraction of calls to prepaid accounts	-0.122***	(0.047)
Constant	0.733***	(0.142)
Observations	1,302	

Note: Standard errors computed based on the final logistic model; they do not adjust for model selection. *p**p***p<0.01

Table B: Comparison to Traditional Credit Scoring in More Developed Settings

Other Settings **Traditional Credit Scoring**	Best AUC (In time, out of sample)	Default Rate	Features
UK (Baesens et al., 2003)	0.668-0.758	10-25%	16-19 unspecified predictors
Belgium / Netherlands / Luxembourg (Baesens et al., 2003)	0.776-0.791	30-33%	33 unspecified predictors
Italian small and medium enterprises (Calabrese & Osmetti, 2013)	0.723	1-5%	Firm leverage, liquidity, profitability
Bosnia microfinance (Van Gool, Verbeke, Sercu, & Baesens, 2012)	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; to be conservative we present the best out of sample AUC for each sample. Because we select based on performance on the test dataset, these estimates will tend to overstate performance on independent samples. (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.

Stability Over Time

Our main results are estimated out of sample within the same time period, but when implemented, a model trained on past data will be used to predict future repayment. We test performance 'out of time' for the South American telecom sample (the MFI sample is not observed long enough to perform this test). To do this, we construct an offset version of the dataset in two steps. 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.

To create a baseline comparison, we train a model on half of the offset dataset selected completely at random, without regard to timing, and test on the other half (2 fold cross validation). This shows the performance of the offset dataset, in a half sample split when the algorithm is trained on data from both time periods.

We then test performance when the algorithm is trained only the early period of data. 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. Results are presented in Table C. All models apart from the complete credit bureau model (Random Forest) perform worse when tested on a different time period. Our method's AUC deteriorates from a baseline of 0.70-0.71 to 0.60-0.63 out of time, but continues to outperform models using credit bureau data (AUC 0.55-0.58).[18] Since our method picks up behaviors in high temporal frequency, it can be sensitive to temporal shocks spuriously correlated with repayment, and thus care must be taken to ensure it picks up relationships that are stable over time. We expect the out of time performance of our method to improve when trained on multiple cohorts (just as credit bureaus have evolved by observing default patterns over many cohorts).

[18] Out of time performance tests of credit scoring methods are rare in the literature; Table B shows comparison estimates of bureau scoring, but these are in time tests.

Table C: Model Performance: Out of Time

South American Telecom
Offset Indicators Dataset

	Baseline		Out of Time	
Training:	2 Fold CV		Early period	
Test:			Late period	
	AUC	H-measure	AUC	H-measure
Baseline Models				
Demographics				
Random Forest	0.500	0.007	0.500	0.005
Logistic, stepwise BIC	0.574	0.014	0.563	0.005
Credit Bureau (all variables) and demographics				
Random Forest	0.536	0.074	0.547	0.004
Logistic, stepwise BIC	0.590	0.021	0.578	0.008
Our Models				
Phone indicators				
Random Forest	0.700	0.074	0.631	0.013
Logistic, stepwise BIC	0.709	0.088	0.595	0.011

Both scenarios use phone indicators derived from only half of the data (the first half for early loans; the last half for late loans). The left columns show model performance when training and test are drawn evenly from the early and late periods. The right columns show out of time performance: model is estimated on the first half of loans and tested on the latter half. AUC represents the area under the receiver operating characteristic curve; we also report the H measure (Hand, 2009).

Stability Across Contexts

In Table D we test the stability of the models across contexts; the first two columns report the cross-validated performance from Table 3. The following columns report results from combining the two samples, using the indicators that are available in both. We omit credit bureau specifications as the Caribbean country does not have a bureau. The third column shows cross validated performance on the joint sample drawn from both contexts. The fourth column trains the model on the South American telecom loans and tests it on the Caribbean microfinance organization; the fifth column does the opposite. Our results suggest that our model is predictive in both settings when trained with local data (AUCs 0.55-0.76), but a model trained in one context is not universal (AUCs 0.50-0.56 when trained in the other context). We expect models to be more portable across more similar settings.

Table D: Model Performance across Contexts

Training dataset:	South American Telecom (5 fold CV)	Caribbean Microfinance (5 fold CV)	Joint (5 fold CV)	South American Telecom	Caribbean Microfinance
Test dataset:				Caribbean Microfinance	South American Telecom
	AUC	AUC	AUC	AUC	AUC
Baseline Models					
Demographics and loan characteristics*					
Random Forest	0.500	0.525	0.531	0.500	0.500
Logistic, stepwise BIC	0.576	0.500	0.569	0.558	0.502
Indicators	6	5	4	4	4
Our Models					
Phone indicators					
Random Forest	0.709	0.553	0.681	0.512	0.555
Logistic, stepwise BIC	0.759	0.570	0.694	0.515	0.500
Indicators	5,541	5,602	1,783	1,783	1,783

*: loan characteristics included only in Caribbean MFI sample. AUC represents the area under the receiver operating characteristic curve. Out of sample performance is estimated using 5-fold cross validation for the first three columns, and with data from the other country for the last two columns. Each test uses the set of indicators shared between the datasets.