Leverage Mobile Data in Consumer Credit Risk Modeling

Shen, Hao

UNIVERSITÉ DE GENÈVE GENEVA FINANCE RESEARCH INSTITUTE

Leverage Mobile Data in Consumer Credit Risk Modeling

Dissertation Submitted to

The University of Geneva

in partial fulfillment of the requirement

for the professional degree of

Doctorate of Advanced Professional Studies in Applied Finance, with Specialization in Wealth Management

by

Hao Shen

(FCO N° 58457)

Dissertation Supervisor: Professor Harald HAU
University of Geneva

Associate Supervisor: Professor MA Jun
Tsinghua University

April, 2021

Disclaimer

I declare that I have read the plagiarism information and prevention documents issued by the University of Geneva.

I certify that this work is the result of personal work and has been written independently. The work is the responsibility of the author, in no way does the work engage the responsibility of the University of Geneva, nor of the supervising Professors.

I declare that all sources of information used are cited in a complete and accurate manner, including sources on the Internet. Other individuals and groups that have contributed to the research work involved in this paper have been clearly identified in the paper.

I am aware that the fact of not citing a source or not quoting it correctly is plagiarism and that plagiarism is considered a serious fault within the University, punishable by penalties.

In view of the above, I declare on my honor that the present work is original.

Signature： SHEN Hao Date： 4/21

Acknowledgement

I would like to express my sincere gratitude to Professor Harald HAU and Professor MA Jun for their guidance and support.

My sincere thanks also goes to XU Huan and WU Peng in Huabo Group who provided data support to this paper.

Abstract

Credit risk modeling has been a critical and established procedure in the financial services industry. In the past 10 years, China's consumer lending market has experienced rapid growth and many alternative (non-bank) lending firms expanded into sub-urban and rural areas, against the backdrop of policies of financial inclusion. But traditional credit bureau data covers limited proportion of China's population, especially outside tier one cities. Many individuals lack credit bureau data or even banking histories, making it extremely challenging for banks and lending firms to assess individual's credit quality, and for potential consumers to receive credit at reasonable cost and speed. As of Jan 2018, the PBOC (People's Bank of China) credit bureau covered less than 400 Million individuals, which is lower than 30% of the total population. Bureau coverage has not been able to keep up with the pace of industry growth. However, most of individual consumers do have personal mobile phones, with rich behavioral data being generated continuously and with minimal cost.

This paper is to discuss a credit modeling work based on mobile behavioral data, instead of traditional structured credit bureau data. Research data are collected from a handset financing product rolled out since Aug 2017 in some provinces by one major telecom company in China. This paper shows that behaviors captured in mobile data can be used to predict consumer credit quality, using call/sms/data usage records matched to installment repayment. On a sample of individuals with no historical credit bureau data available, our analysis shows good prediction power, on both validation data set and test data set. Customers in the highest decile of risk by our measure are 8.8x – 17.6x times more likely to default on the installment payment than those in the last decile. The method discussed in this paper forms a new way to measure and quantify the credit risk for telecom related lending product.

Table of Contents

Leverage Mobile Data in Consumer Credit Risk Modeling

1. Introduction

The objective of this paper is to examine the quality of credit risk modeling based alternative mobile data (i.e. mobile data such as call log, data usage pattern, account balances, etc.) in the absence of credit bureau information in a handset financing product.

Mobile phones have spread globally and China is no exception: there are about 6.6 billion mobile-cellular telephone users in developing countries (ITU, 2019), and 1.6 billion are in China (statista.com, 2019). These mobile devices, especially those in developing markets and China, make it possible for the lenders to explore new ways of extending credit. In China, telecom companies gradually started granting postpaid credit to help smooth phone usage, or to finance handset purchase to make it more affordable. Through telecom companies' distribution channel/network, new credit product can potentially be extended directly to consumers throughout the country at minimal transaction costs. The data possessed by telecom companies, and telco's established distribution network in rural and sub-urban areas would enable a new form of credit product different from those traditional lending firms or banks provide.

But, how reliable these behavioral data will be in predicting whether the individual borrowers will pay back – this has been a fundamental question all lenders would ask. Few developing country residents have structured credit data or scores. Given the huge business potential residing in developing markets (particularly China) and lack of better alternatives, some lenders including traditional banks who used to heavily rely on structured credit bureau information have started conducting pilot programs exploring the effectiveness and relevancy of telecom company behavioral data.

We developed a method to predict repayment probability of the individual customers on a handset financing product. One key observation is that most consumers (regardless of their bureau data status) do have rich records of interactions and usage data with mobile phone. The way that individual consumers use a mobile phone can potentially predict whether they will make repayment on time. This paper develops and evaluates a relatively low cost approach to predict repayment of credit product using mobile behavioral data, which are collected via normal mobile phone usage. From basic phone usage records it derives behavioral variables plausibly related to payment likelihood, and uses a few

approaches to consolidate and leverage these variables to predict payment probability. We pick three basic machine learning methods with very minor parameter changes and demonstrates good prediction performance without using any credit bureau data.

Our method consumes raw data on mobile phone usage, which are collected via the telco's existing distribution channel/branches under formal partnership with the telecom company at very low cost. This paper shows how variables extracted and derived from this data can be used to predict the repayment behavior. Inside the data sets, there are variables indicating customer behavior which are clearly related to good repayment. For example, a responsible customer normally manages their phone billing / spend carefully so usage won't be disrupted. Also a customer who has phone calls made with larger number of contacts tend to be more socially connected (compared to customers who have few contacts), hence credit default tends to be less likely. A responsible customer wouldn't leave the phone offline for an extended period as professional/business activities would be affected during this period. There are also data points indicating likely credit product usage even though We and the telecom company wasn't able to retrieve the credit bureau information. For example, in China most financial institutions (banks and insurance companies) use phone number starting with "95" prefix for customer inquiry / services, both inbound and outbound. Hence by checking whether/how many calls have been made with "95" numbers would potentially indicate the engagement level with financial institutions. Similarly, there are also variables showing whether customer has made contacts with a few alternative money lenders which potentially can be deemed as "sub-prime" lending business.

From the raw data collected, we extract approximately 2000 indicators, and on top of that, another 500 derived variables were generated for data inputs. It is a critical step to decide what fields are the most relevant ones as data input. Though the machine learning techniques should theoretically be able to drop those variables which don't show predicting power, We still chooses to conduct basic manual screening to select those ones with theoretical link or business intuitive to predict customers' repayment, with three key objectives: 1. Efficiency - to make the machine learning calculation more efficient; 2. Stability - to avoid variables showing counter intuitive features or showing predicting power only within certain period due to data set selection constraints; 3. Explainability - to be able to explain to external parties (i.e. regulator, partner, customer) when a business decision needs to be made, given this paper is built upon a real business product and the learnings would be applied back to the product.

Based on the above considerations, three categories of data are selected:

1. Customers' demographic info, including employment status and basic salary levels. This is to assess customer's social stability;

2. Indicators on existing connection with financial institutions and alternative lenders. This is to capture and measure customer's potential debt level and capacity to repay additional credit;

3. Phone usage behavior, including number of call/sms made in past 6 months, length of duration when there is no phone calls/sms. This is to measure the engagement level of phone usage and customer loyalty to telecom company services

The business product underlying the data is a handset financing product launched by the telecom company in China back in 2017. Traditionally, the telco's customers could sign a two-year service subscription contract with commitment of using the telco's services continuously and prepaying a basic fixed monthly fee. The overall pricing under this contract would be preferable to the customers as long as the fixed monthly fees are being paid continuously. If the customer chooses to purchase a new mobile handset at the contract signing stage, a separate one-time purchase payment needs to be made by the customer.

In 2017, the telecom company decided to add on the additional handset financing plan into the existing contract scope. Instead of one-time purchase payment for the handset, the customers can choose to take on a 24-month installment payment plan[1]. The major challenge for the telecom company was the lack of credit related data for underwriting decisioning.

This product was first time launched in Jiangsu, which is one of the most developed

[1] A recurring direct debit process was introduced in this program, where a one-time deduction of both the monthly service subscription fees and the handset installment payment will happen on the 5th day of every month, and the successful payment deducted are split into subscription fees and installment repayment, going into respective parties (i.e. telecom and lending firm). From customers' view, this new program made the fee paying process and installment repayment much smooth where they don't need to manage two separate payments. In case the direct debit process fails, another attempt will be triggered on the next day till the end of the month. To be more specific, this two repayment streams are not really affecting each other except for the fact that they are being deducted together at the same time. Literally, a customer can choose to default on the handset installment repayment but can still pay the monthly subscription fee and enjoyed the discounted mobile service pricing. The only downside for the customer is he/she needs to manage the monthly payment manually through other alternative channels instead of relying on the recurring direct debit process. In a way, cost of default is very low, which is a significant challenge to the lending firm.

provinces in China. Although the credit bureau coverage is still limited in the rural and sub-urban area, the mobile penetration rate in Jiangsu is extremely high. And due to regulation, only banks are allowed to access credit bureau data during underwriting and credit review. It is not possible for telecom company or other alternative lenders to query credit bureau data. Hence the telecom company decided to use very basic criteria to select customers who are qualified for this new handset financing product – account tenure. At the beginning of product launch, customers with account tenure > 1 year would be eligible. This leads to two outcomes: first, since no credit related criterion was applied, the customers default rate tends to be higher than expected; second, good data points are being captured without material biases, which can be used to develop more sophisticated selection criteria without credit bureau data available.

Based on observation, majority customers with sufficient usage history before the credit extended showed good repayment behavior throughout 24 months. But there are also sizable customers who missed/delayed payment even in the first few months after contract start.

The method discussed in this paper (i.e. variable extraction and selection, machine learning techniques with very basic parameter tuning) shows good potential to achieve decent predictive accuracy. Performance is assessed in both validation and test data set.

- In the validation data set, this method shows AUC ranges from 0.76 - 0.78.
- In the test data set, this method shows AUC ranges from 0.76 – 0.78.

More interestingly, customers in the highest decile of risk by this method are 8.8x – 17.6x times more likely to default than those in the lowest decile. Given the funding available in this handset financing program are always limited, this method can effectively identify a group of prospects with best credit quality which will generate direct profit impact to the businesses.

It is important to call out that most of the customers covered in this analysis are unlikely having any credit bureau records[2], and at this stage we are not able to access credit bureau data due to regulation restrictions. Should credit bureau data become accessible to telecom companies, we would conduct a performance comparison on those customers who

[2] This estimate is based on the call log analysis elaborated later in this paper, where close to 70% of customers had less than 5 calls with financial institutions within 6 months prior to application.

do also have bureau records and evaluate whether mobile behavior data can be more effective in credit scoring.

2. Literature Review

There are researches done about the use of mobile phone data in credit scoring. For alternative data usage in fintech space especially related to financial inclusion, a lot of research had been conducted also by commercial firms especially consulting companies.

Sheng, Yip, Cheng from Oliver Wyman (2017) explores the underlying value drivers of the consumer finance industry in China, and discusses how big data technology and various alternative data will evolve and how banks, consumer finance service providers should respond to these trends.

Alain Shema (2019) tested using airtime recharge to build up effective credit scoring model largely due to the rising privacy concerns. The research was built up on an airtime lender in Africa that made it possible to run a side-by-side comparison of an airtime-only model against a model that also incorporated past loan data, as well as the current model used by the lender. There are learnings we incorporated into our data selection and creation processes.

Khan (2018) discussed his work on extracting features from mobile communication logs and billing data using various techniques, include brute force approach and Deterministic Finite Automata (DFA).

Björkegren and Grissen (2020) developed an approach using call records to build loan outcomes for a sample of borrowers in a Caribbean country. They demonstrated that the microfinance lender could have reduced its default rate by 41%, while only declining 25% of their existing borrowers. This paper provided us additional insights when we create additional derived variables for this study.

Lazer, Kennedy, King, Vespignani (2014) discussed how data mining without reasonable interpretation can lead to counter intuitive conclusions using the example of "Google Flu" to demonstrate unstable/unreliable correlations when models put into practical use.

Our analysis is based on a study of a credit product (handset financing product launched in year 2017) of which the risk management processes are entirely built based on

telecom data without credit bureau. On top of the learnings from other researchers, we categorized the data points collected into three broad categories – social stability, capability to payback, and existing credit product usage based on mobile usage history, and we ensure all of the variables have good explainability, based on business interpretation/experience. Potentially the variables could be predictive in assessing consumer's creditworthiness for generic credit product in the absence of credit bureau data in developing world. At this moment we haven't got opportunity to test it with other financial institutions. We would have no hesitation when such opportunity rises in the future.

3. Data and Context

We have been working with one major telecom company in China to analyze data collected from the handset financial product applicants since 2017. Those users' application information and mobile behavior data are used to generate credit prediction models to predict repayment / default behavior, based on which, a decision can be made to approve or decline a handset installment purchase plan with a 24-month payment schedule.

The telecom company offered the handset financing product to a pre-selected set of existing prepaid mobile customers with account tenure above 1 year. It is a very straightforward prospecting method but the telecom company expect a more sophisticated methodology can be established to improve the selection criteria to source more quality customers.

Raw data are primarily from the telco, which is the largest mobile service operator in China. Based on a long term agreement with necessary data privacy clearance from various regulatory authorities, we have been granted data access (in a highly controlled access environment) to a database containing information of customers who visit telecom company's distribution branches to take up this handset financing product. All data set are located within the telecom company's data center while We can access some of the raw and derived data points for this purpose without any personal identifiable information being taken outside of the data center.

The data set collected covers approximately 90,000 customers who were offered and took up the handset financing product. The repayment in this product cover both mobile monthly subscription fee (prepaid), plus the monthly installment on the handset purchasing. All of the customers have completed their first 24 months' repayment

schedule. Approximately 94% of the customers paid all months' installments while the rest 6% missed at least 1 repayment. There are also customers who missed payment due to some other reasons (e.g. service disputes, data processing error, etc.) which have been excluded from this analysis.

The mobile data include raw records for each call, SMS and data consumption activity in past 12 months when offer taken and application submitted, with date & time stamps, duration, contact number (all private mobile numbers have been scrambled). It also carries latest account balance info and past 12 months billing data. Customers' demographic data are also captured including gender, age, employment status, handset brand and model, etc. No personal identifiable information was collected such as name, IC number, address, mobile phone. And all data processing and analysis are conducted within the telco's data center and only aggregated info can be carried away.

Descriptive statistics for the data set are presented in Figure 1 – Figure 5 below.

From the demographic distribution shown in Figure 1 & 2, the main customer groups took up this handset financing product are in the age of 30 – 55 years old and majority are male. There is no specific concentration into any age groups though. There weren't any criteria set in prospecting strategy hence We interprets it that this is a natural reflection of customers who tend to respond to this new product from the telco's current customer base.

Figure 3 & 4 shows the handset related stats. On the handsets underlying the financing product, "Vivo" and "Oppo", which are the low end handset brands from local phone manufacturers, dominate. This is also reflected in the pricing distribution. Three price bands (CNY, 1k – 1.1k, 1.2k – 1.3k, 1.5-1.6k) are where Vivo and Oppo phones retail price normally fall in.

Figure 5 shows the number of calls between the customer and numbers from financial institutions (i.e. banks and insurance firms) within 12 months before product sign up. A large portion of the customers had no contact with financial institutions at all, reflecting the "thin bureau" nature of the customer base profile.

Figure 1: Customer count distribution by age range

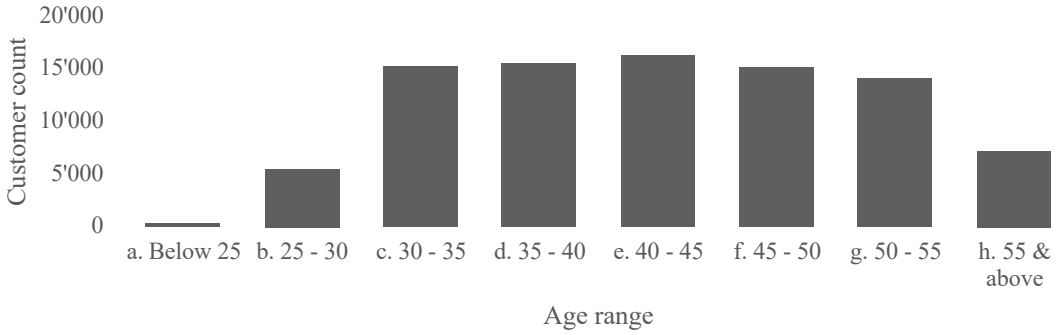


Figure 2: Customer count distribution by gender

Figure 3: Customer count distribution by handset brand

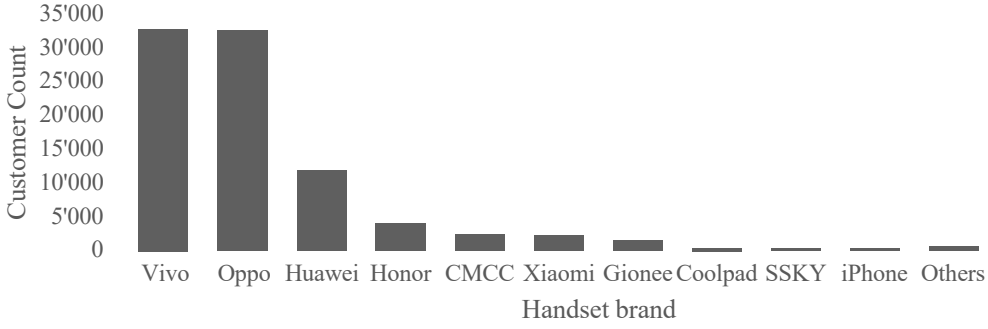


Figure 4: Customer count distribution by handset purchase price

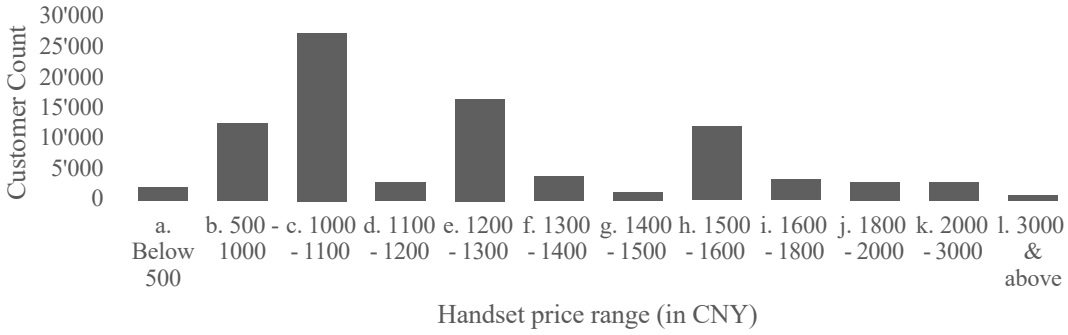
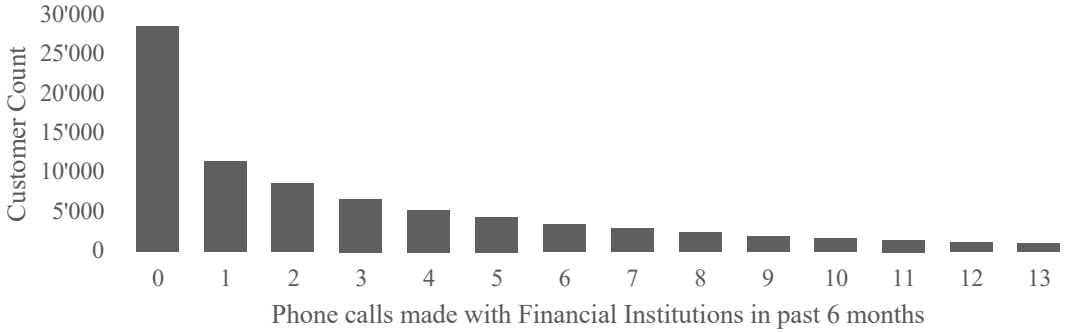


Figure 5: Customer count distribution by phone calls made with financial institutions in past 6 months



4. Methodology

The target of this research can be defined as a binary classification task - default (target = 1) or no default (target = 0). Research data are extracted from the telecom company's data center and stored in dedicated MySQL data warehouse. Data covers handset financing applications during Aug 2017 – Dec 2018 in Jiangsu, Hubei, Hebei provinces. All approved applications have accumulated 24 months' payment performance available. Data are collected at nearly to zero monetary cost.

In the data set, the customers are deemed as fulfilled their repayment obligation only when all the 24 months' installment are fully paid (target = 0). Customers missed any repayment will be labeled as default (target = 1).

We extract a set of data variables that potentially have intuitive relationships to repayment behavior, and also derived variables based on business interpretation. Since this is from a real business product and learnings will be applied back on to the same product, it is important to ensure variables have good explainability and intuitive to understand the relationship, as the selling processes are mostly conducted face to face (i.e. important to communicate reasons to customers in case of decline), and local regulators would conduct audit from time to time (i.e. ensure reasonable due diligence are used in application).

Customer demographic information captures the customer's social stability and potential capacity to repay. Such information includes marriage status, employment status, salary level, etc. But We is fully aware of the limitations as well. Many of the demographic information is collected based on customer's self-declaration (such as employment status, salary, etc.), and it is not feasible to conduct robust verification given the selling/application process and potential data privacy concerns. Technically this type of information could be manipulated easily. Hence We doesn't plan or recommend to implement those variables into the decisioning process, even if some of them can be predictive.

In the raw data records from telco, there is a specific data field indicating the contact number. Most of the financial institutions in China use a prefix "95" in their customer service contact for both inbound and outbound calls. We anticipate that any phone calls made with such numbers should indicate the customer's engagement with banks or insurance companies, which potentially can be used as an indicator on whether the customer is an existing bank loan user. And the level of calls should rise along as the level

of engagement (i.e. number of product, balance of accounts, credit payable, etc.). Similarly, there are also existing numbers publicly known belonging to alternative lending firms (non-bank players). The number of calls made with these lending firms would indicate whether/how many existing lending relationships the customer may have. We hence derived additional variables to calculate the number of calls made with these FIs or alternative lenders in different time frame (7 days/1 month/3 months/6 months/12 months prior to product take up) for assessment.

Phone usage data captures other behaviors that with intuitive link to repayment. For example, the account monthly spend can tell whether the customer has a stable usage and expenditure pattern. For example, during what period of the day are most calls made will reflect the working pattern of the customers, including holiday usage vs. workday usage. The number of contacts the customer has potentially shows the level of social connection and stability. Whether the customer has an extended period of not using the phone may indicate either the customer is not responsive to any inbound inquiries from time to time, or maybe there is another phone the customer mainly uses and the current number won't tell sufficient information for assessment.

The data processing has following steps:

- Data extraction: in total approximately 2000 raw feature/variables are extracted from the telco's data base, covering customer's past raw transactional records, account usage & billing data, basic demographic data, and application information such as handset purchase related features
- Derived variable creation: as discussed above, we applied business knowledge and intuition to create additional variables from the raw data including variables capturing usage trend, variation, and others covering social connection and existing lending product usage variables.
- Data cleaning: a basic screening has been conducted to remove those fields with insufficient observations, too many missing values or single/unary inputs. Also some data with counter intuitive relationship links are also removed to enhance stability and avoid potential data quality problem.
- Data standardization: To make machine learning / classification process works better, necessary variable conversion work has been conducted including regrouping similar categorical variables into smaller groups, changing continuous variables into intervals, and basic standardization work to reduce the variance.

In the end of this process, a table with target indicator and customer's behavioral data has been put together which contains close to 90,000 rows (observations) and approximately 120 columns (data features/variables). All the variables kept in this table are carrying meaningful / intuitive links to the repayment behavior. Full list of selected variables is presented in Table 3 in Appendix.

The tools used to conduct next step analysis is SAS Enterprise Miner, as it is easy to use and has convenient parameter setting features. We decide to use three of the existing machine learning modeling modules – Logistic Regression, Decision Trees, and Gradient Boosting.

5. Prediction and Results

Entire data set has been randomly split in to training (60%), validation (20%), and test (20%) groups. The random selection is executed using "data partition" module in SAS.

Before putting the data into the modeling modules, it is common to ask what are the most important predictive variables to predict default (target = 1). Table 1 below present results from SAS "variable selection" module with individual variables carrying highest R-Square.

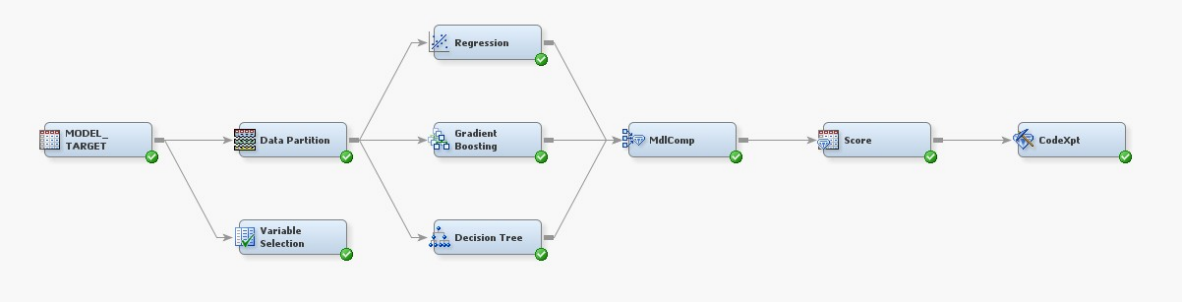
Table 1: Variables with highest R-Square

Variable	Notes	R-Square
mobile_account_age	Tenure of mobile account in months	0.017103
lending_p2p_12m	Count of calls made with p2p lending firms	0.013608
lending_installment_12m	Count of calls made with installment lending firms	0.012307
brand*province	Handset brand with customer province	0.007257
marital_status*province	Marital status with customer province	0.006526
age	Age in years	0.006421

Demographic variables tend to have low contribution/correlation with repayment. The most predictive variables are the tenure of the account, and past contact count with alternative lending firms. This is in line with writer's expectation.

The data set with selected variables are predictive using standard machine learning techniques. In order to ensure good prediction performance and good explainability can be achieved together, we chooses three commonly used modeling techniques: Logistic regression, Decision tree, and Gradient boosting. All three modeling toolkits are available in SAS Enterprise Miner, and writer only made small parameter changes and kept the rest at default setting. Figure 6 below shows the process flow implemented on SAS Enterprise Miner. Better performance can be expected with more parameter fine turning and data transformation.

Figure 6: Data processing and modeling flow

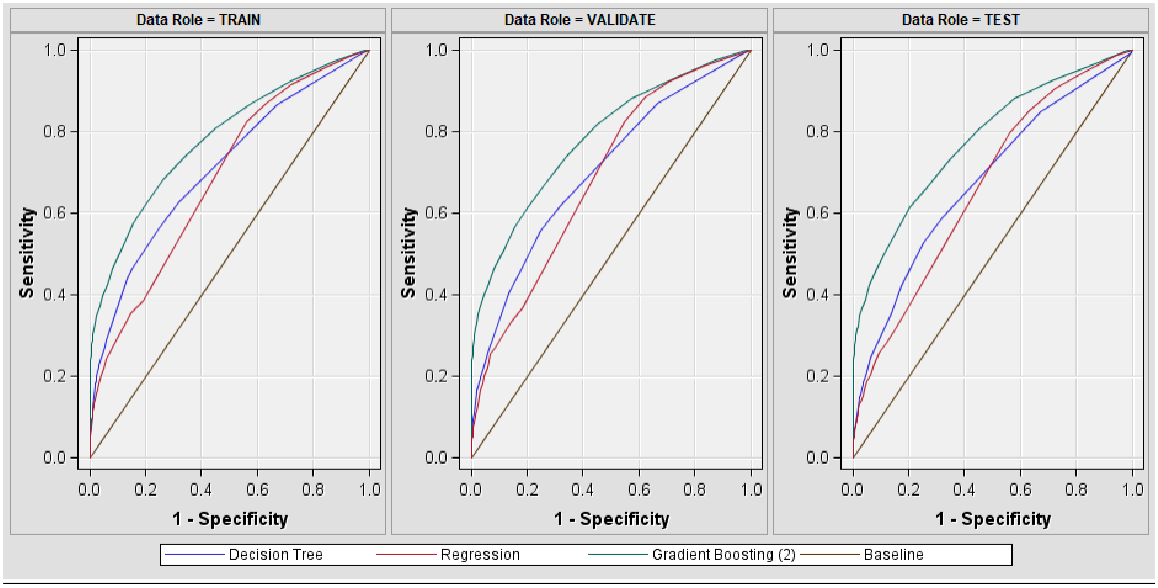


Performance

ROC (receiver operating characteristic) curve is one commonly used approach to compare model's performance (accuracy and stability). It maps the sensitivity (true positive rate) against the specificity (false positive rate), plotting performance as more cases are accepted.

Figure 7 below shows the ROC curves from training/validation/test data sets with results from all three modeling techniques. The performance has been fairly consistent across different data sets with Gradient Boosting being the best performer.

Figure 7: Receiver operating characteristic curve comparison



The area under the ROC curve (AUC) is equal to the probability that a classification model orders a baseline (randomly chosen) positive instance higher than a negative one. A baseline classification model (predicting randomly) generates AUC of 0.5 and a perfect

model will generate AUC of 1.0. Table 2 below shows the AUC across all data sets with three modeling techniques. All three models achieved AUC above 0.76 demonstrate the data set prepared has good predictive power. And very little performance variance is observed across training/validation/test data sets shows good model stability. We expect better performance can be achieved with further data and modeling parameter fine tuning.

Table 2: AUC comparison

AUC	TRAINING	VALIDATION	TEST
GRADIENT BOOSTING	0.778	0.778	0.778
DECISION TREE	0.767	0.765	0.763
REGRESSION	0.764	0.762	0.762

Putting the performance back to business perspective – all three modeling techniques would be able to rank order customers based on their probability of default. So telecom company and lending firms can decide acceptance thresholds based on targeted profitability, value proposition, and other considerations. Purely from model performance standpoint, in the most conservative model (logit regression), customers in the highest risk decile are 8.8 times more likely to default than those ones in the lowest risk decile. Should the best performing model (gradient boosting) be used, this ratio would be 17.6x.

This model is able to effectively differentiate customers with high default risk from those with good repayment behaviors by only using telecom company behavior data. And as a result, a model with nearly identical variables have been deployed by the telecom company to drive better customer prospecting and enhance product profitability.

We believe this model can be expanded to other generic credit product, especially those product carrying similar lending size and targeting rural area where the credit bureau coverage is relatively poor. At this moment we haven't got any opportunities to work with other financial institutions to test the effectiveness of this model. Should such opportunities rise, we would have no hesitation to do so.

6. Discussion

As high mobile penetration rate and low bureau coverage rate co-exist in emerging markets, it is beneficial and sensible to explore more usage cases leveraging mobile data other than telecom company related cases. There are also unique challenges in mobile data usage, and more importantly there are potential impacts to be considered carefully around upcoming legislation changes or regulatory compliance considerations in emerging markets.

Collusion risk

From observations on the repayment behavior and investigations on a few unique default cases, we recognize there are loopholes in the prospecting/selling processes after the new model deployed, and there are cases where the sales person encouraged applicants to game with the underwriting rules, especially when decline reasons are being clearly communicated and some of the key variables are relatively easy to manipulate. It is especially challenging when many customers have multiple mobile numbers in use.

The customer identity related information (e.g. demographic) are being validated through government agency's data services so data reliability has been high. Information around customer's social stability such as salary and job status are based on self-declaration. Although some variables in this category can be predictive based on model performance, we decide not to rely on them as none of them can be validated with reasonable cost.

Most of mobile behavioral data points used in this modeling exercise are directly from telecom company's database and there is no touch point where customers can change or manipulate before the application submission. But customers do have the flexibility of choosing which mobile account will be used for application underwriting. Based on local regulation, we can't query telecom company's database until the customer's consent is granted. For example, applicant may choose to use the mobile account with longer tenure or stable expenditure pattern and less or no contacts with banks or alternative lenders which would normally lead to a favorable decision outcome.

Account tenure, usage pattern, past contact history are all critical variables in this model. Removing them would lead to model performance deterioration, and jeopardize the objective of this work. In order to remediate this risk, we made recommendations to the telecom company that the lending limit should be calculated based on the data richness of

the account. For example, when a customer has two mobile accounts of which one is the main account and the other one is a backup account, the customer is more likely able to get a higher priced handset application approved if the main account is used. A lower limit would be granted if only the backup account is submitted as the data would not be as rich as what the main account can provide.

Additional rules and continuous optimization on the decisioning process are always required and can certainly remediate some of the collusion risk, but we recommend that the sales compensation mechanism is critical to minimize/prevent such activities. For example, some part of the sales incentive should be calculated based on credit performance of the customers solicited and a proportion of payout should be delayed to 3 months after approval. A continuous review and improvement process on the incentive mechanism itself would be equally important than the underwriting logic, if not more.

Financial inclusion

Using mobile data to extend credit would provide strong support to the nationwide financial inclusion policy, given majority of individuals in rural area do not possess credit bureau data and couldn't be credit assessed using traditional methods.

We recognize that many of thin bureau or no bureau individuals also lack of credit awareness. Even with mobile data available, it would be more prudent to roll out product with small credit limit and allow the credit history to be built up, and more credit awareness and knowledge can be groomed gradually. And at the same time, more comprehensive legislations/laws would be desired to encourage responsible credit usage and establish the cost of credit default.

Usage expansion

On top of telecom company specific credit product, there are also other generic financial product can potentially leverage mobile behavioral data. Traditional banks and alternative lending firms have started piloting adding mobile data into their underwriting model and extending non-telecom company related credit product to prospects. The telecom company also implemented similar models to choose quality customers on prepaid plan to convert to post-paid plan to increase profitability and enhance customer loyalty.

In both banks and telecom company post-paid usage cases, customers would bear higher

cost of credit default as the defaulter's credit bureau or the mobile usage would naturally be affected or disrupted if the repayment is not made within agreed period.

Data consent and privacy protection

Data privacy is a key consideration in mobile data usage. As illustrated in this paper, many behavioral variables required direct access and calculation on the raw calls/sms/data activities. Customers' consent must be requested and explicitly granted prior to data access, even when no personal identifiable information is used.

On legislation side, China released its draft Personal Information Protection Law, which closed seeking opinion period on November 19, 2020. While the timeline on the law's implementation has not been confirmed, there is no doubt data privacy in China will experience a dramatic enhancement once the Personal Information Protection Law comes into effect. All entities collect and possess customer information should prepare ahead to ensure their policy and processes in compliance against the backdrop of this legislation establishment.

7. Conclusion

This paper presents a method to predict credit repayment/default behavior for applicants without credit bureau history, by using mobile usage data. The method demonstrates good performance in both accuracy and stability, and can be used to cover large portion of individual customers in rural and sub urban areas of China.

The intelligent use of mobile data can effectively reduce information asymmetries between financial institutions and individual customers. With the financial inclusion policy support and low cost distribution channel, a new digital lending ecosystem expanding to the unbanked and underbanked population can be established.

Appendix

Table 3: Selected Variables

Variables	Notes/Explanation
borrower_index	Customer ID
status	Current payment status
contract_index	Contract ID
province	Customer location
age	Age
sex	Gender
is_married	Marrital Status
brand_name	Brand
handset_price	Price
funding_source	Funding source
tot_payment_made	# payment made
first_payment_start_month	First payment month
last_payment_month	Last payment month
max_late_days_f1m	Highest delinquency in days in first 1 month
max_late_days_f2m	Highest delinquency in days in first 2 month
max_late_days_f3m	Highest delinquency in days in first 3 month
max_late_days_f6m	Highest delinquency in days in first 6 month
max_late_days_f12m	Highest delinquency in days in first 12 month
max_late_days_f24m	Highest delinquency in days in first 24 month
month_repay	Monthly installment payment amount
p2p_7d	Contact history with p2p lenders in past 7 days
p2p_1m	Contact history with p2p lenders in past 30 days
p2p_3m	Contact history with p2p lenders in past 90 days
p2p_12m	Contact history with p2p lenders in past 360 days
micro_credit_7d	Contact history with micro lenders in past 7 days
micro_credit_1m	Contact history with micro lenders in past 30 days
micro_credit_3m	Contact history with micro lenders in past 90 days
micro_credit_12m	Contact history with micro lenders in past 360 days
consumption_installment_7d	Contact history with consumer financing lenders in past 7 days
consumption_installment_1m	Contact history with consumer financing lenders in past 30 days
consumption_installment_3m	Contact history with consumer financing lenders in past 90 days
consumption_installment_12m	Contact history with consumer financing lenders in past 360 days
account_balance	balance in mobile account
mobile_net_time	mobile account open date
mobile_net_age	mobile account tenure
contact_count_1month	number of contact in past 1 month
contact_count_3month	number of contact in past 3 month
contact_count_active_3month	number of contacts (calling out) in past 3 month

Variables	Notes/Explanation
contact_count_passive_3month	number of contacts (calling in) in past 3 month
contact_count_mutual_3month	number of contacts (calling out and in) in past 3 month
contact_count_call_count_over10_3month	number of contacts with > 10 calls in past 3 month
contact_count_6month	number of contact in past 6 month
contact_count_active_6month	number of contacts (calling out) in past 6 month
contact_count_passive_6month	number of contacts (calling in) in past 6 month
contact_count_mutual_6month	number of contacts (calling out and in) in past 6 month
contact_count_call_count_over10_6month	number of contacts with > 10 calls in past 6 month
contact_count_mobile_6month	number of mobile contacts in past 6 month
contact_count_telephone_6month	number of landline contacts in past 6 month
contact_count_not_mobile_telephone_6month	number of non-mobile/non-landline contacts in past 6 month
call_count_1month	number of calls in past 1 month
call_count_3month	number of calls in past 3 month
call_count_active_3month	number of outgoing calls in past 3 month
call_count_passive_3month	number of incoming calls in past 3 month
call_count_work_time_3month	number of calls made in working hours in past 3 month
call_count_offwork_time_3month	number of calls made outside working hours in past 3 month
call_count_workday_3month	number of calls made on working days in past 3 month
call_count_holiday_3month	number of calls made on holidays in past 3 month
call_count_6month	number of calls in past 6 month
call_count_active_6month	number of outgoing calls in past 6 month
call_count_passive_6month	number of incoming calls in past 6 month
call_count_work_time_6month	number of calls made in working hours in past 6 month
call_count_offwork_time_6month	number of calls made outside working hours in past 6 month
call_count_workday_6month	number of calls made on working days in past 6 month
call_count_holiday_6month	number of calls made on holidays in past 6 month
call_count_call_time_less1min_6month	number of calls less than 1 min in past 6 month
call_count_call_time_1min5min_6month	number of calls between 1 min and 5 min in past 6 month
call_count_call_time_5min10min_6month	number of calls between 5 min and 10 min in past 6 month
call_count_call_time_over10min_6month	number of calls over 10 min in past 6 month
call_time_1month	total call duration of calls in past 1 month
call_time_3month	total call duration of calls in past 3 month
call_time_active_3month	total call duration of outgoing calls in past 3 month
call_time_passive_3month	total call duration of incoming calls in past 3 month

Variables	Notes/Explanation
call_time_work_time_3month	total call duration of calls made in working hours in past 3 month
call_time_offwork_time_3month	total call duration of calls made outside working hours in past 3 month
call_time_6month	total call duration of calls in past 6 month
call_time_active_6month	total call duration of outgoing calls in past 6 month
call_time_passive_6month	total call duration of incoming calls in past 6 month
call_time_active_mobile_6month	total call duration of calls made in working hours in past 6 month
call_time_passive_mobile_6month	total call duration of calls made outside working hours in past 6 month
call_time_work_time_6month	total call duration of calls made on working days in past 6 month
call_time_offwork_time_6month	total call duration of calls made on holidays in past 6 month
msg_count_1month	sms in past 1 month
msg_count_3month	sms in past 3 month
msg_count_6month	sms in past 6 month
tot_call_contacts_95	number of 95 contacts
callout_times_95	total number of outgoing calls with 95 contacts
callin_times_95	total number of incoming calls with 95 contacts
callout_duration_95	total duration of outgoing calls with 95 contacts
callin_duration_95	total duration of incoming calls with 95 contacts
active_day_1call_3month	number of days with active calls in past 3 months
active_day_1call_6month	number of days with active calls in past 6 months
max_continue_active_day_1call_3month	highest number of continuous days with active calls in past 3 months
max_continue_active_day_1call_6month	highest number of continuous days with active calls in past 6 months
silence_day_0call_3month	number of days without any calls in past 3 months
silence_day_0call_active_3month	number of days without any outgoing calls in past 3 months
silence_day_0call_0msg_send_3month	number of days without any outgoing sms in past 3 months
silence_day_0call_6month	number of days without any calls in past 6 months
silence_day_0call_active_6month	number of days without any outgoing calls in past 6 months
silence_day_0call_0msg_send_6month	number of days without any outgoing sms in past 6 months
continue_silence_day_over3_0call_3month	count of continuous 3 days without calls in past 3 months
continue_silence_day_over15_0call_3month	count of continuous 15 days without calls in past 3 months
continue_silence_day_over3_0call_active_3month	count of continuous 3 days without outgoing calls in past 3 months
continue_silence_day_over15_0call_active_3month	count of continuous 15 days without outgoing calls in past 3 months
continue_silence_day_over3_0call_0msg_send_3month	count of continuous 3 days without sms in past 3 months

Variables	Notes/Explanation
continue_silence_day_over15_0call_0msg_send_3month	count of continuous 15 days without sms in past 3 months
continue_silence_day_over3_0call_6month	count of continuous 3 days without calls in past 6 months
continue_silence_day_over15_0call_6month	count of continuous 15 days without calls in past 6 months
continue_silence_day_over3_0call_active_6month	count of continuous 3 days without outgoing calls in past 6 months
continue_silence_day_over15_0call_active_6month	count of continuous 15 days without outgoing calls in past 6 months
continue_silence_day_over3_0call_0msg_send_6month	count of continuous 3 days without sms in past 6 months
continue_silence_day_over15_0call_0msg_send_6month	count of continuous 15 days without sms in past 6 months
max_continue_silence_day_0call_3month	highest number of days without any calls in past 3 months
max_continue_silence_day_0call_active_3month	highest number of days without any outgoing calls in past 3 months
max_continue_silence_day_0call_0msg_send_3month	highest number of days without any sms in past 3 months
max_continue_silence_day_0call_6month	highest number of days without any calls in past 6 months
max_continue_silence_day_0call_active_6month	highest number of days without any outgoing calls in past 6 months
max_continue_silence_day_0call_0msg_send_6month	highest number of days without any sms in past 6 months
gap_day_last_silence_day_0call_6month	number of days since last day without any calls in past 6 months
gap_day_last_silence_day_0call_active_6month	number of days since last day without any outgoing calls in past 6 months
gap_day_last_silence_day_0call_0msg_send_6month	number of days since last day without any sms in past 6 months
spend_avg	average spend on mobile account
spend_tot	total spend on mobile account
spend_max	highest monthly spend on mobile account
spend_min	lowest monthly spend on mobile account

References

Agarwal, Sumit, Shashwat Alok, Pulak Ghosh, and Sudip Gupta (2020). "Financial Inclusion and Alternate Credit Scoring for the Millennials: Role of Big Data and Machine Learning in Fintech." Indian School of Business.

Aitken, Rob (2017). "'All Data Is Credit Data': Constituting the Unbanked." *Competition and Change* 21(4), pp 274–300. doi: 10.1177/1024529417712830.

Anderson, Billie, and J. Michael Hardin (2014). "Credit Scoring in the Age of Big Data." *Encyclopedia of Business Analytics and Optimization 18*(1): 549–57. doi: 10.4018/978-1-4666-5202-6.ch049

Baer, Tobias, Tony Goland, and Robert Schiff (2012). "New Credit-Risk Models for the Unbanked." *McKinsey Working Papers on Risk* 30.

Björkegren, Daniel, and Darrell Grissen (2020). "Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment." *World Bank Economic Review* 34(3): 618–34. doi: 10.1093/wber/lhz006.

Carroll, Peter, and Saba Rehmani (2017). "Alternative Data and the Unbanked." *Oliver Wyman*: 1–17.

Chaffins, Jenna, and Ann Chen (2018). "Alternative Data Across the Loan Life Cycle: How FinTech and Other Lenders Use It and Why." *Experian*: 1–18.

Cooper, Cheryl R. (2020). "Alternative Data in Financial Services." *Congressional Research Service (CRS)*.

Djeundje, Viani B., Jonathan Crook, Raffaella Calabrese, and Mona Hamid (2021). "Enhancing Credit Scoring with Alternative Data." *Expert Systems with Applications* 163(August):113766. doi: 10.1016/j.eswa.2020.113766.

Experian (2018). "The State of Alternative Credit Data." *Experian White Paper* 18.

Financial Stability Board (2017). "Artificial Intelligence and Machine Learning in Financial Services - Market Developments and Financial Stability Implications." *Financial Stability Board* (November).

Gambacorta, Leonardo, Yiping Huang, Han Qiu, and Jingyi Wang (2019). "How do Machine Learning and Non-traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm." *BIS Working Papers* 834.

Gulamhuseinwala, Imran, David Wu, James Lloyd, and Vikram Kotecha (2017). "China and UK FinTech - Unlocking Opportunity." *EY Report.*

Khan, Muhammad. R. (2018). "Machine Learning for the Developing World using Mobile Communication Metadata." Ph.D. thesis, University of California, Berkley

Kumar, Harish K. D., and Vivek Iyer (2017). "Crossing the Credit Divide with Alternative Data." *White Paper*, Tata Consultancy Service.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014). "Big Data. The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176). doi: 10.1126/science.1248506

Liu, Xinhai, Ti, Wang, Wei Ding, Yanjun Liu, and Qiuyan, Xu (2017). "A Credit Scoring Model Based on Alternative Mobile Data for Financial Inclusion." *2017 Credit Scoring and Credit Control Conference* (2).

Nopper, Tamara K (2020). "Alternative Data and the Future of Credit Scoring." *Data for Progress* (August).

Schneider, Rachel, and Arjan Schuette (2007). "The Predictive Value of Alternative Credit Scores." *Center for Financial Services Innovation*.

Shema, Alain (2019). "Effective Credit Scoring Using Limited Mobile Phone Data." *ACM International Conference Proceeding Series* (February). doi: 10.1145/3287098.3287116.

Sheng, Cliff, Jasper Yip, and James Cheng (2017). "Fintech in China: Hitting the Moving Target." *Olivier Wyman*: 1–31.

TransUnion (2015). *The State of Alternative Data*. TransUnion/Versta Research: 1–20.

Turner, Michael A., Patrick Walker, and Katrina Dusek (2009). *New to Credit from Alternative Data*. Durham, NC: PERC Press.

World Bank Group (2008). "Credit Scoring." *Encyclopedia of Finance* 76–76. doi: 10.1007/0-387-26336-5_521

World Bank Group (2019). "Disruptive Technologies in the Credit Information Sharing Industry: Development and Implications." *Fintech Note* 3. doi: 10.1596/31714