

Research Article

Measuring Multi-Dimensional Mobile Behavior Effect on Inclusive Finance: Evidence from China

Chi Ming Chen^{1,*} , Geoffrey Kwok Fai Tso¹ , Kaijian He² 

¹Department of Management Sciences, City University of Hong Kong, Hong Kong, China

²Hunan Engineering Research Center for Industrial Big Data and Intelligent Decision Making, Hunan University of Science and Technology, Xiangtan, China

Abstract

Credit Invisible is one key area that many countries put much effort to solve in decades. According to the 2020 World Bank statistics, for example, there are over 500 million Chinese and 45 million American, classified as credit invisible who don't have banking and finance history in bank or credit bureau, making them difficult to borrow money from financial institution. Previous studies adopted different non-financial information to evaluate one's credit worthiness and status to address this issue. However, they provide little information about how real mobile user interactions can be used to solve this issue in inclusive finance. This paper proposes a novel data generative framework to fusion APP data, call detail record data and SMS data with a total of 4,689 attributes derived from a large-scale mobile dataset. We then construct a unique set of mobile behavior-driven credit risk factors based on statistical diversity, intensity, consistency, and regularity of mobile user behavior characterizing user preferences, attitudes, geolocation, and temporal patterns. Empirical analysis demonstrates that the newly discovered mobile behavior factors are useful as new inputs for credit scoring and proves the factors representing new source of positive and negative credit information. Decision tree analysis and Quantile regression are conducted to validate effect of these factors to credit default. It facilitates credit assessment based on non-financial data for the credit invisible people, which promoting inclusive finance to larger community in society. We also analyze implications of mobile user characterization findings in relation to credit default which helps decision makers to optimize credit policy and product design.

Keywords

Inclusive Finance, Big Data, Mobile Behavioral-based Credit Risk Factors

1. Introduction

In the credit market, one of the main reasons people can hardly get loan from banks is because of information asymmetry between customers and banks [18] Up to year 2020 according to figures from the World Bank, there are about 45 million American and 500 million Chinese people who are

credit invisible, in other words, neither has credit bureau record nor credit transaction history in bank to evaluate one's credit worthiness and status. To solve this issue, different types of data and suitable credit risk modeling techniques are required to apply in different types of lending scenarios. In

*Corresponding author: cmchen5-c@my.cityu.edu.hk (Chi Ming Chen)

Received: 8 September 2024; **Accepted:** 16 October 2024; **Published:** 29 October 2024



Copyright: © The Author(s), 2024. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

practice, there are three main categories of bank loan: collateralized loan (including mortgage), guaranteed loan and credit loan. Since security measures are included in collateralized and guaranteed loans, hence, they are usually referred as secured loans. In credit loan, it is solely based on borrower's re-payment capability without secured asset held by bank against the loan. Thus, credit risk control of credit loan needs to base on more hard information compared with collateralized and guaranteed loan, which more towards relationship type of lending [3]. This paper analyses non-financial information of borrower's characteristics, specifically mobile behavior, applicable to credit loan. While smart phone has been widely adopted in past decade, previous credit scoring methods and datasets mostly focus on financial data, which do not address how mobile application data linked to financial credit risk. The methodology proposed in this paper expands the lendable base, more customers becoming credit visible by leveraging big mobile data. Hence, it facilitates adoption of inclusive finance to the unbanked customer group.

The fast-growing adoption of mobile phone [8] has facilitated the need to study different aspects of mobile behavior implications in various areas. Smartphones provide a new 'data collection point' about mobile phone usage. The usage of smart phone in daily life enables researchers to collect new source of personal behavioral data and develop more targeted models of user behavior in temporal and social contexts [15], [16]. The analysis shows that aggregated features obtained from smartphone usage data can be indicators of personality traits, including extraversion, agreeableness, conscientiousness, neuroticism, and openness to experiences. Mobile data can infer one's daily mobility and social behavior which reflects an individual's personality [4]. Furthermore, some studies have been done to investigate how behavioral traits based on mobile data relating to self-reported personality characteristics. And the result shows mobile-data derived features can be personality characteristics' factors [6]. In addition, the individual mobile user profile can be used to predict one's credit behavior and financial wellbeing [1, 13, 17]. Using call detail records applied in credit scoring [14], the results show that combining call detail records with traditional financial data significantly increases model performance when measured in area under the curve (AUC). And regarding data privacy, mobile user self-identifiable personal information including name and address are removed and only anonymized identifiers are used for the data analysis purpose [1]. However, these datasets and analyses provide little or no information about real mobile user interactions to credit default. And above papers do not cover mobile apps and mobile web behavior, which represent most of the mobile device usage in today's mobile consumer world. Our study further identifies mobile behavior data linked to financial literacy and related factors, which prove to have a strong relationship with financial debt. The big-scale dataset with mobile behavioral features empowers research community to further explore new findings about loan applicants' personal

and social behavior characteristics such as spending and mobility patterns, social interaction, and wellbeing, which can apply in credit scoring prediction.

A novel data processing framework is proposed to consolidate three types of mobile data, APP data, call detail record data and SMS data, and classify the data into six newly discovered mobile behavior dimensions about personal spending, financial and social wellbeing characteristics. From the dimensions identified, we design a unique set of credit risk factors based on statistical diversity, intensity, consistency, and regularity of user mobile behavior which can be applied as a new source of inputs in credit decisioning.

Section 2 in this paper describes novel approach of building the large-scale mobile behavior dataset and transforming the data into unique set of mobile behavior-driven credit risk factors. Section 3 presents experimental results that quantifies the effect of newly developed factors on credit default and analyzes the new insights identified. Section 4 and 5 concludes research implications and recommended directions for future research.

2. Data & Methodology

This data construction initiative creates a unique base data with variety of data types, data scale, data coverage and temporal dimension. The considered dataset has 450,722 individual mobile users sampled from a large internet platform's database in China. For each sampled user, the usage history between September 2016 and December 2018 is collected. This amounts to 13,190,082 number of mobile apps usage from the sampled mobile users in China. Basic demographics shows gender proportion of 52% male and 48% female respectively, which is close to official statistics published from the 2020 China National Bureau of Statistics Report. In this analysis base dataset, 99,350 mobile users are identified and tagged with credit default cases (90+ days overdue), which will be used to analyze its relationship with the newly created mobile behavior factors to be described in section 2.3.

The main types of mobile applications include 1) social networks, 2) map navigation, 3) online shopping, 4) personal communication, 5) lifestyle, 6) personal assistance, 7) photo editing, 8) online video, 9) online browsing, 10) news, and 11) gaming. In the China Mobile Internet Industry Data Research from Jiguang 2019, Chinese mobile users install average 56 mobile applications. From the App Annie Company 2018 report, average monthly number of mobile applications (active definition) is 27–28. In our big dataset, average number of apps per user is 29, whereas median number of apps per user is 20. Although the statistical aggregation of application usage from different reports might be different, the overall market figures show that our sample data substantially represents recent mobile user behavior and usage in China.

Data Privacy and Confidentiality

This subsection aims to discuss the research data used in

this paper relating to anonymity, confidentiality, and data protection according to the requirements of the General Data Protection Regulation (GDPR) and the UK Data Protection Act 2018. The relevant data protection guidelines emphasize the rights of the individual whose data are being processed (the 'data subject') but also incorporate a range of exemptions from these rights when processing data for research purposes. The data used in this paper is fully anonymized in a form such that the original data subject cannot be identified by either us or the dataset. Besides, the mobile loan application has explicit legal terms stating data collected from user mobile device and other data sources solely for credit evaluation purpose. Mobile user is under full obligatory or voluntary condition to supply the data. Apart from user data authorization and consent, only mobile behavioral data points are used for model development without any personal data or any information relating to an identified or identifiable natural individual, which cannot trace back to certain individuals' personal information. Since the processing of our research data is fully anonymized, which falls outside above scope of the related regulatory requirements.

2.1. Data Attribute Extraction

The data extraction focuses on three types of data. Those attributes extracted from call, SMS and application logs are collected from all events triggered by mobile user when making or receiving calls, messaging, or using apps installed on the mobile phone. The call data log includes number of outgoing & ingoing calls, duration, number of contacts and calling status. The SMS data log includes number of received and sent SMS, timestamps and contact counts. The APP data log includes device type, mobile brand, OS type, timestamp, location, app categories for example news, streaming video, education etc. Then, the extracted attributes are transformed and derived into 4,689 attributes under the six proposed factor dimensions to be described in section 2.3.

2.2. Mobile Data Dimension Construction

This paper further identifies the credit related characteristics with using big mobile data that can be benefits on credit loan business, an online questionnaire (refer to Appendix) is conducted to a group of senior credit managers of licensed financial institutions in China. The design of questionnaire aims to identify priority areas that represent majority of loan business context in China in relation to different sources of data used during credit decision process. The questions cover four main areas: types of loan businesses that are applicable to mobile user behavior; key risk factors and criteria that credit officers use for credit approval; types of data sources that can support their business to identify the above risk factors; and corresponding data characteristics from the suggested data sources. We sent a total of thirty emails to banks, consumer finance companies and lending companies. Their businesses and branches have present nationwide in China, which provides a good representation of personal financing customer

coverage in China. Eleven responses were received, with one blank response being invalid.

We base on the respondent feedback and insights on credit risk measures to determine six new credit risk dimensions based on mobile data, as shown below. These key dimensions are personal spending, personal lifestyle, personal interests, personal credit, social interaction, and demographics (predicted). A detailed algorithm on the measures of the derived factors is also presented in a later section.

i) Personal Spending:

From mobile subscription plan used, first, it can tell whether it is a business plan, normal or low-end basic plan. Moreover, a person is relatively less likely to default if spends consistently via credit or debit card for his or her daily living consumption, including utility, food and beverage and other daily consumption items over a certain time period. This indicates an individual having actual spending needs as well as a certain income to spend. Additionally, having multiple credit cards determined by spending alert SMS indicates that an individual's credit risk should be relatively lower since his or her profile can be approved by banks. Hence, we derive attributes like mobile subscription plan, number of active bank cards, insurance payment frequency, online spending frequency, bank card spending frequency, etc.

ii) Personal Lifestyle:

The mobile app usage also indicates how an individual spends their daily time which can then determine whether the individual has some bad habits or a sudden downsize of living standard that will likely lead to a higher credit risk. For example, someone who mainly uses smartphones around midnight for gaming should receive more attention in approving his or her loan application. In addition, using multiple wealth management or securities-related apps for a certain period of time also indicates that an individual should have better experiences in financial management and investment knowledge than those who do not use any of these apps. From there, we derive other similar attributes like gaming related, online shopping, social/dating/photo, professionals / white collars, health fitness, gaming, online video, travelers, etc.

iii) Personal Interest:

By understanding how active an individual is using his or her mobile device and, more importantly, what the main usage areas are, there are different implications of the usage areas. One implication is that very low activity regardless of the calling or mobile internet or app usage for a user 20–45 years old working in a tier-one city such as Shenzhen or Beijing. One possibility is that the device may be newly activated or opened for applying for loans. Abnormally low usage can provide an alert indicator when performing loan approvals by credit officers. Another potential implication is that if most of the phone usage is only loan-related content and app usage, it is quite a strong risk signal that the mobile device is mainly used for applying for many different loan applications but not for normal daily use. Hence, we derive attributes like number of new APP installed, active APP counts, APP type counts,

APP usage days, etc.

iv) *Personal Credit*:

Bill statements for the past 3 to 6 months indicate the number of times that someone did not pay their bills, for how long and whether their mobile service provider has stopped their service plan as well. Furthermore, the number of loan-related apps an individual has installed and used indicates the number of times an individual applied for loans or credit cards, and that they may not be able to borrow money easily so they keep applying from different financial institutions. This information can be used to infer whether one person's credit status is good. In addition, recent loan collection SMS indicates an individual does not make a loan repayment or even potential fraud cases. From there, we derive attributes like, loan payment balance, active loan counts, credit limit & utilization, credit card application counts, loan overdue counts, bank card fund transfer amount, asset remaining balance, wealth management application usage, work management application usage, etc.

v) *Social Interaction*:

From the calling records, a person's social network can be understood person by analyzing the number of unique incoming and outgoing calling numbers occurring on a monthly basis. If the unique count is high, for example, over 50, and if we see that the frequency of incoming and outgoing calls is also very high, it provides strong evidence that the person has a relatively large social network and that the chance an individual will relinquish the phone number is relatively low. In contrast, time of calling and messages can also provide some clues as to whether a user has a regular 9 am to 5 pm job. Thus, we derive attributes like mobile internet usage by different time periods (12am-6am/6am-6pm/6pm-12am), top 10 calling counts, total call-in counts, total call-out counts, etc.

vi) *Demographics*:

This dimension basically aims to infer individual life-stage, frequent living or working places, etc., mobility behavior as a proxy to understand personal or family situation, which is particularly useful in online lending where applicants do not provide much detailed information. One implication is that based on the geolocations granted from user applications, it can determine an individual's frequent daytime and nighttime locations. Another implication is that if an individual recently used many baby-related apps, it can infer that an individual has a family with a child or will soon. From a credit approval perspective, if an individual has stable mobility and a family, it indicates that one should have a positive credit status in terms of willingness to repay loans. We derive attributes include location preference of 573 cities identified in China, gender, marital Status: married/single, family with kids, etc.

2.3. Credit Risk Factor Generation

After establishing the key data dimensions and categories, this section presents how to generate data factors to become the feature variables in the big mobile dataset representing the

credit risk composite measurements. The mobile behavioral factors are derived based on below four measurement types:

1. **Diversity**: this refers to the fact that spending and usage frequency may vary differently over time. In addition, his or her transactions may be in various "buckets" or "bins". The term "bins" is defined as segments over time (for example, over the last 1 or 3 months). For each bin j , the proportion of transactions f_{ij} for user i is defined. Then we derive user diversity i using the normalized entropy as follow:

$$D_i = \frac{-\sum_j^N f_{ij} \log f_{ij}}{\log P} \quad (1)$$

where N total bins of all counted transactions and P non-empty bins. The closer the value to 1, indicating higher the user's usage diversity. A high diversity user means that his or her usage spread across different period of time almost equally. The $\log P$ aims to normalize users with different level of usage, so that giving equal chance getting high value in this measurement.

2. **Interest concentration**: It measures the percentage of usage in most frequent interests among all various interests. We used the same bin definition for each user and the top three highest usage bins are considered. Given h_i the aggregated proportion for the top three highest usage bins out of total usage of user i . The interest concentration is then calculated as follows:

$$C_i = \frac{h_i}{\sum_j^N f_{ij}} \quad (2)$$

The closer the value to 1, indicating higher user's interest concentration which shows most usage concentrates among the top three highest interest areas.

3. **Consistency**: It characterizes user usage behavior's similarity level over three-month and six-month period of time. We use normalized Euclidean distances between these two time periods. The consistency value is calculated as below:

$$CI_i = 1 - \frac{\sqrt{(D_i^1 - D_i^T)^2 + (C_i^1 - C_i^T)^2}}{\sqrt{2}} \quad (3)$$

where D_i^1 is the user's diversity for the first three month, whereas D_i^T is the diversity for the whole six month. Correspondingly, C_i^1 and C_i^T are the user's interest concentration for the same time frames. The output discrepancy value CI_i is again in the range of 0 and 1, higher the values more the consistency. For example, value of one indicates the usage is identical compared the first three months and entire six months, whereas value of zero showing complete inconsistency.

4. **Overspending**: It determines whether an individual user i overspends by estimating aggregated spending amount, cc_i , from credit card transactions and projected income, I_i , as derived from finance information extracted SMS text data. The overspending ratio of user i is defined as below:

$$O_i = cc_i/I_i \quad (4)$$

For those having above ratio above value of one, indicating user spends more than their income, and higher the ratio, more likely of excess spending which probably leads to potential risk financially.

A bin is defined as an individual item interpretation for a particular time period.

- 1) Temporal-hourly: hour slots, such as 10 am -11 am as one bin and 10 pm -11 pm as another bin.
- 2) Temporal-weekly: weekend or weekday. For example, Monday to Friday is one bin and Saturday and Sunday another.

Data distribution

In previous studies, the German credit dataset, [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credi](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credi)

t+data), is commonly used for authors of credit scoring prediction and heterogeneous feature selection research. These data and other similar ones can be found from the UCI Repository of Machine Learning Databases. However, only 20 static data fields from the German credit data, and most are demographic and financial related data from credit bureau, which financial data is not readily available specifically in developing countries, with people do not have banking relationship. Although these data fields are proven to identify and differentiate likely to default cases, the Chinese big mobile dataset provides many more factors in terms of financial and nonfinancial behavior derived from mobile data. Below, Table 1 illustrates descriptive statistics of sample features generated according to the data dimension construction and factor generation procedures presented above.

Table 1. Sample descriptive statistics for the Chinese big mobile data.

Features	Statistics				
	Count	Mean	STD	Min.	Max.
Personal Spending					
E-commerce spending diversity in the last 3 months	335,934	0.3524	0.1677	0	0.8974
E-commerce spending concentration in the last 3 months	335,934	0.2839	0.1252	0	1
Mobile subscription plan amount in the last 3 months	450,016	235	54	0	>10k
Bank card spending diversity in the last 3 months	194,836	0.1973	0.1180	0	0.7316
Bank card spending concentration in the last 3 months	194,836	0.2095	0.0937	0	0.5991
Insurance payment diversity in the last 3 months	53,552	0.1253	0.0673	0	0.5273
Bank card fund transfer in amount in the last 3 months	64,576	5,754	946	0	>10 m
Number of cards in use in the last 3 months	164,931	3.175	0.4517	0	13
E-commerce spending consistency in the last 3 months	335,934	0.3817	0.1821	0	1
Overspending ratio in the last 3 months	194,836	0.6327	0.4729	0	12.3445
Personal Lifestyle					
Gaming weekdays after 6pm in the last 3 months	174,371	15.2491	3.6647	0	77
Wealth management weekdays in the last 3 months	113,953	20.8713	4.8355	0	80
Social network weekdays in past 12 months	398,642	32.2954	8.9904	2	248
Travel weekdays in past 12 months	193,825	18.0411	6.2962	0	164
News/Radio weekdays in past 12 months	273,906	22.7315	7.0243	0	198
Online Shopping weekdays in past 12 months	317,426	30.6945	9.0237	1	239
Health fitness weekdays in past 12 months	208,542	15.8630	5.6330	0	207
Online video weekdays in past 12 months	325,790	33.4418	10.9114	0	240
Personal management weekdays in past 12 months	180,274	15.6143	4.3593	0	238
Finance weekdays in past 12 months	253,709	19.5204	6.1074	0	226
Personal Interests					

Features	Statistics				
	Count	Mean	STD	Min.	Max.
Travel consistency in the last 12 months	193,825	0.3212	0.0973	0	1
Gaming consistency in the last 12 months	174,371	0.4783	0.1388	0	1
Wealth management consistency in the last 12 months	113,953	0.3713	0.1403	0	1
Online shopping consistency in the last 12 months	317,426	0.4069	0.1365	0	1
Health fitness consistency in past 12 months	208,542	0.3993	0.1620	0	1
Online video consistency in past 12 months	325,790	0.3268	0.1174	0	1
Personal management consistency in past 12 months	180,274	0.4271	0.1513	0	1
Finance consistency in the last 12 months	253,709	0.3950	0.1427	0	1
App count diversity in the last 12 months	430,845	0.2356	0.1236	0	1
App usage days diversity in the last 12 months	430,845	0.2290	0.1148	0	1
Personal Credit					
Loan application count in past 3 months	60,173	2.5330	0.8245	0	12
Loan overdue count in past 3 months	25,038	1.0283	0.1604	0	8
Credit card application count in past 3 months	52,651	1.9162	0.2839	0	11
Unique loan app counts in past 3 months	82,746	2.6703	0.7714	0	20
Social Interaction					
Total SMS sent in past 3 months	410,392	3.1573	1.1930	0	126
Top 10 calling counts in past 3 months	405,175	10.6855	3.9524	0	130
Total call-in counts in past 3 months	405,175	13.8251	4.9073	0	394
Total call-out counts in past 3 months	405,175	9.7136	3.6417	0	173
Total unique calling number count in past 3 months	405,175	12.8437	4.5770	0	153
Unique social app counts in past 3 months	230,115	3.8265	1.3094	0	9
Demographics					
Distance between frequent location and loan application location (km)	90,467	11.7186	5.9022	0.2239	482.3165
Family with kids age < 12 years	450,722	0.2076	0.0554	0	1
Marital Status: single	450,722	0.31996	0.1164	0	1

3. Experimental Results

To understand the effect of the Chinese big mobile dataset with its data dimensions and factors on credit default, three types of analysis are conducted: (1) data exploratory analysis to discover any correlation exists, and value-added insights after using the mobile data; (2) principal component analysis (PCA) and decision tree (DT) analysis to validate the data dimensions constructed in section 2 can be used as a new credit risk measures for credit scoring prediction; (3) quantile regression (QR) to analyze robustness of the proposed new

measures. Decision tree is often used in data exploratory analysis, prediction model, and other data mining practices due to its interpretability and visualization, indicating the result in simple rule logic. To validate the data dimensions, PCA is used to map the high-dimensional mobile data into lower-dimensional space without losing much information. PCA identifies and inputs important data components into decision tree analysis validating whether highly related to credit default. DT algorithm aims to split a dataset into branch-like groups by using various statistical methods. It uses a combination of rule sets to classify many individual data records into smaller number of homogeneous groups, which relates to a particular target variable, in our case, the

credit default event. Quantile regression models the relationship between exogenous variables (in this paper, refers to mobile behavior) and the conditional quantiles of endogenous variable (credit default event probability) given the existence of exogenous variables. Quantile regression can provide a more complete picture of the conditional distribution of credit default event by analyzing both lower and upper or all quantiles of mobile behavioral variables.

Samples of loan applications and default data are collected with first-, second- and third-tier cities in China to obtain a representative base for the analysis. The mobile behavior dataset is divided into two portions: one is for model training, and the other is for model validation. Within the dataset, 300,000 records are randomly selected as the training dataset,

and the remaining 150,722 are selected as the validation dataset.

Table 2 shows the top correlated features, and it shows that high-usage days in mobile loan applications have a Point-Biserial coefficient of 0.4344; similar feedback also shows in our business interview that the borrower probably lacks funds and cannot easily borrow money from finance institutions, which is why the borrower needs to look for many different lenders that may be easier relatively easier to borrow money from. Besides, proximity of mobile user's frequent location to the location they apply loan shows relative high Phi correlation coefficient of 0.3969. This circumstance needs to alert credit underwriters for further checking reason why trying to apply loan from far away.

Table 2. Top features correlated to credit default. Value towards +1 represents higher correlation to credit default, whereas value towards -1 represents higher correlation to non-credit default.

a. Phi coefficient.

Features (binary type)	Phi correlation coefficient	P-value
1. Loan APP usage was over 60% total usage of mobile APP in past 3 months	0.4021	0.0112
2. Location applying loan was different from user frequent location	0.3969	0.0365
3. More than 30 days using gaming-related APP in past 3 months	0.3698	0.0332
4. Less than 3 unique calling phone numbers in past 6 months	0.3506	0.0495
5. More than 3 unsuccessful payment transactions in past 6 months	0.3470	0.0078
6. Using more than 6 different types of mobile APP in past 3 months (use one time or more in a month)	-0.3852	0.0115
7. Over 10 days using mobile APP every month in past 12 months	-0.3801	0.0245
8. More than Rmb900 credit card monthly spend in past 6 months	-0.3590	0.0294
9. More than Rmb500 insurance monthly payment is in past 12 months	-0.3511	0.0129
10. Family with children below age of 12	-0.3351	0.0433

b. Point-Biserial coefficient.

Features (continuous type)	Point-Biserial coefficient	P-value
1. Days using loan APP in past 3 months	0.4345	0.0120
2. Days using work management APP in past 6 months	-0.3178	0.0124

Furthermore, in Table 3, principal component analysis is conducted with the top five components already accounting for 78.1% of the total variance in the input variable set, which substantially represents the total variable space. The loading values in component matrix in Table 4 indicate the correlations between each component and the explained variables, with the range of value from -1 to +1. From Table 4, compo-

nent 1 is composed of personal spending variables followed by personal interest variables. Component 2 is formed by credit history and status variables followed by personal interest variables. Component 3 is composed of social networks followed by life-stage variables. Component 4 is formed by personal interest variables and personal spending variables. Component 5 is formed by life-stage variables.

Table 3. PCA Total variance explained for the Chinese big mobile data.

Component	Eigenvalues		
	Total	% of Variances	Cumulative %
1	6.249	52.1	52.1
2	1.229	10.3	62.3
3	0.719	5.9	68.3
4	0.613	5.1	73.4
5	0.562	4.6	78.1
6	0.503	4.1	82.3
7	0.471	3.9	86.2
8	0.452	3.2	89.5
9	0.392	3.0	92.5
10	0.321	2.7	95.3
11	0.259	2.2	97.9

Table 4. Component matrix.

Features *	Components				
	1	2	3	4	5
E-commerce spending diversity	-0.8631	-0.1364	-0.1035	-0.1351	-0.1500
E-commerce spending concentration	-0.8022	-0.1956	-0.3464	-0.1600	-0.2141
Mobile subscription plan amount	-0.8490	-0.2431	-0.2645	-0.2275	-0.1777
Bankcard spending diversity	-0.7713	-0.1374	-0.3085	-0.1289	-0.1829
Bankcard spending concentration	-0.7968	-0.1839	-0.3534	-0.1263	-0.2462
Number of cards in use	-0.6945	-0.2335	-0.2002	-0.1018	-0.2753
Insurance payment diversity	-0.7906	-0.2196	-0.1764	-0.2307	-0.1398
Bankcard fund transfer amounts	-0.7201	-0.1587	-0.1309	-0.2046	-0.1701
Loan application count	0.2334	0.8498	0.2512	-0.2048	0.2629
Loan overdue count	0.3122	0.7973	0.1234	0.1205	0.1633
Credit card application count	0.2022	0.6834	0.3843	0.1901	0.2774
Loan payment overdue count	0.1734	0.7507	0.1753	0.1591	0.1117
Total unique calling number count	-0.2323	0.1934	-0.7892	-0.2430	
Total call-in number counts	-0.1896	-0.1045	-0.7854	-0.2584	
Total call-out counts	-0.2670	0.1255	-0.7719	-0.2636	
Top 10 calling counts	-0.1746	0.1699	-0.7708	-0.1945	
Total SMS sent	-0.2709	-0.2578	-0.5053	0.1673	
Mobile internet usage (total minutes)	-0.1469	-0.2994	-0.4234	0.1109	0.1671
App count diversity	-0.2873	-0.4534	-0.2903	-0.9298	-0.1960

Features*	Components				
	1	2	3	4	5
App usage days diversity	-0.2433	-0.2334	-0.2612	-0.8552	0.1717
Unique app type counts	0.1753	-0.4933	0.1098	-0.7234	-0.2558
Total number of apps in use	-0.2084	-0.2356	-0.1856	-0.7055	0.2127
New app installed and in use	-0.2845	-0.3578	-0.2987	-0.6934	-0.1825
Gender: male/female			0.1134	0.2156	0.6392
Family with kids age < 12 years			0.2833	0.1456	0.6217
Marital status: married/single			0.3452	0.3007	0.5824
Location preference: 573 cities identified			0.1965	0.1743	0.4039

* Only features with a 12-month study period were selected for the PCA

Decision tree model results

The decision tree algorithm selects a component factor with the highest discriminant power as the first splitting branch. We use the approach of cross-validation with ten-fold, which our sample dataset is randomly split into ten subsets with equal size; nine of the subsets is used for decision tree model training, whereas the remaining one is kept for model valida-

tion. This process is conducted repeatedly for ten times to obtain the final average outcome. The validation subset is not used to train the model to minimize model overfitting. We analyze the frequency components shown in the first six levels and sorted components according to their degree of influence on credit default probability. Using the binary split tree setting, the experimental result is organized in Table 5.

Table 5. The decision tree structure of components.

Components	First layer	Second layer	Third layer	Fourth layer	Fifth layer	Sixth layer
Component 1	62%	8%	0%	0%	0%	0%
Component 2	38%	9%	5%	0%	0%	0%
Component 4	0%	22%	20%	14%	0%	0%
Component 3	0%	26%	25%	20%	13%	0%
Component 5	0%	20%	31%	22%	8%	1%

In Table 5, it is found that component 1 (mostly composed of personal spending variables) is the most important factor among others for predicting whether a credit default will occur since 62% of the models picked this component as the first branching factor. The second important factor is still component 2 (mainly contributed by credit history). Approximately 38% of the models selects component 2 as the first branching factor. In the second layer, in addition to component 4 (personal interest-related variables) and component 3 (social networks), component 5 (life stages) is also an important factor. Sixty-eight percent and 76% of the decision trees chose these 3 factors, among others, to predict the credit default probability in the second and third layers, respectively. The life stage indicates a relatively strong and

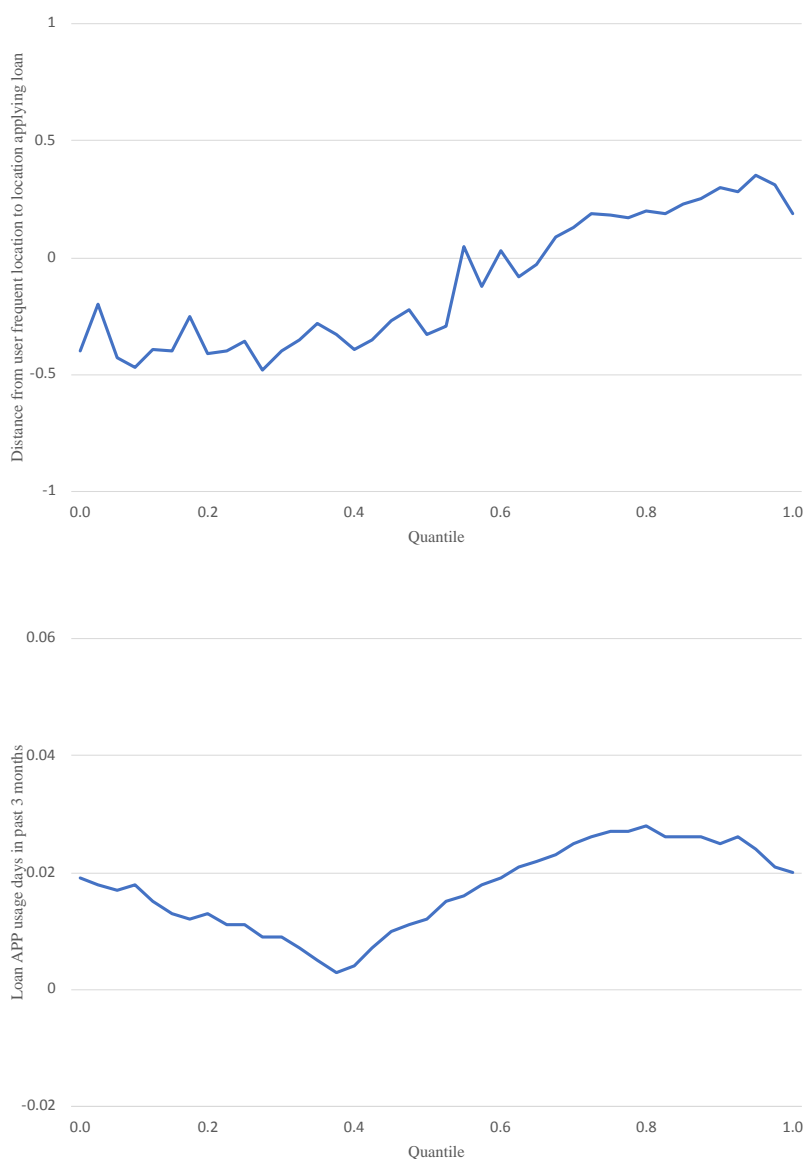
positive relationship with credit defaults, partly because the mix of life stages and demographics leads to difficulty in differentiating its relationship with credit default, and obviously different life stages exhibit different personal and related credit risk preferences. Further variable transformation and analysis is suggested to be performed when in the stage of actual predictive modeling. Going further to the fourth to sixth layers, the major splitting factors are not shown by these top fix components; the combination of these components is used as the branching criteria in the further layers down. This paper validates the applicability of the proposed data dimensions and features for credit risk profiling and prediction.

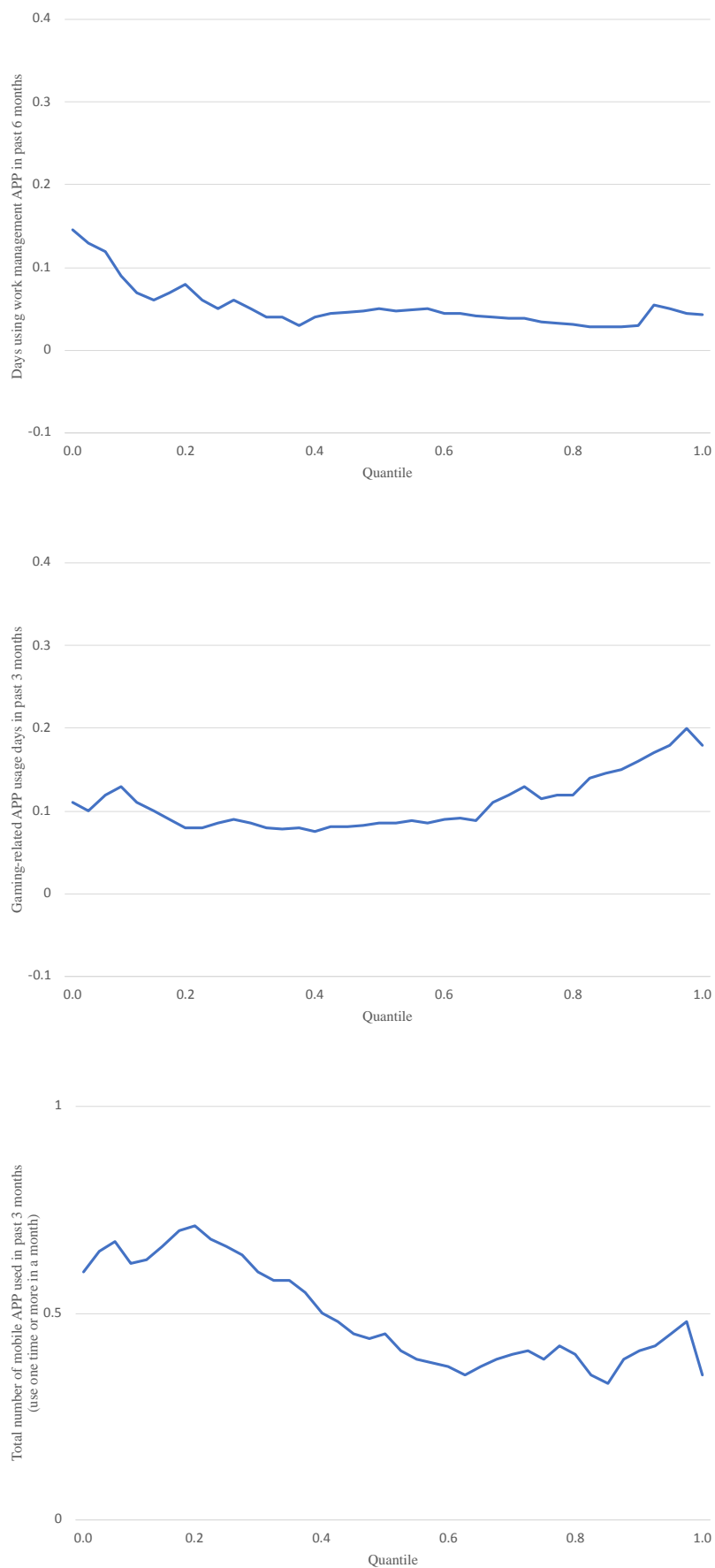
Quantile regression model results

Quantile regression extends the regression model to condi-

tional quantiles of the target variable, such as 95th percentile of credit default probability. Quantile regression is particularly useful when the rate of change in the conditional quantile, which represented by the regression coefficients, depends on the quantile. The main advantage of quantile regression over least-squares regression is its flexibility for modeling data with heterogeneous conditional distributions. Data of this type occur in many fields, including mobile usage and behavior data used in this paper, which is in different scale and distributional shape. Quantile regression provides a complete picture of the variable effect when a set of percentiles is modeled, and it makes no distributional assumption about the error term in the model. In Figure 1, a total of 6 quantile process plots of top selected mobile behavioral variables from quantile regression are computed, higher the value of quantile the higher probability of being credit default. The 95% confidence bands are shaded. The confidence bands are computed using the sparsity method with the non-independent and identically

distributed assumption. The effects of each variable are quite varied. The figure suggests similar findings from previous correlation and decision tree analysis, that strong effect of longer distance from user frequent location to location applying loan on potential credit default's likelihood. Similar finding applies for loan APP, which indicates that borrowers may be likely to apply lots of different loans through mobile applications; For gaming related APP usage, people indulge in online gaming or even gambling involved which may lead to personal financial instability and risk eventually. Whereas, average monthly APP usage days, variety of mobile application usage and days using work management APP are relatively constant mostly the entire distribution, except with a stronger effect in the lower tail, which indicates the more different types of APP or work-wise related APP usage, more likely borrower having normal work and personal living in certain extend, and hence less likely to be credit default.





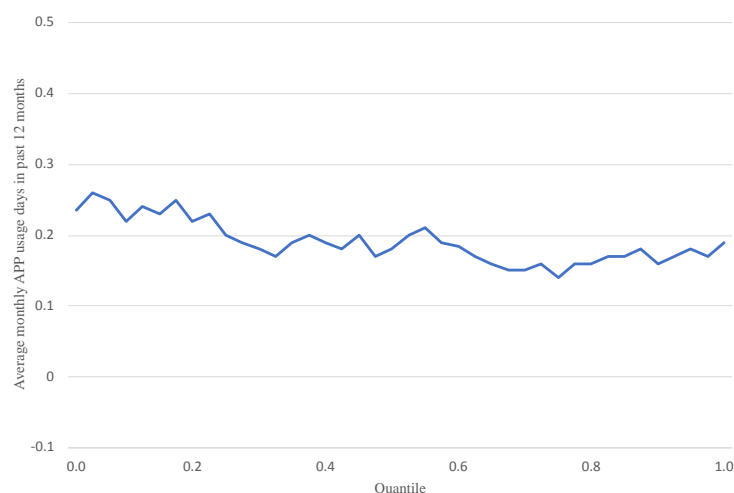


Figure 1. Selected top features from quantile regression with 95% confidence bands. (a) Distance from user frequent location to location applying loan. (b) Loan APP usage days in past 3 months. (c) Days using work management APP in past 6 months. (d) Gaming-related APP usage days in past 3 months. (e) Total number of mobile APP used in past 3 months (use one time or more in a month). (f) Average monthly APP usage days in past 12 months.

Regarding important aspect of defining the performance metrics of decision tree and quantile regression, we adopt the commonly used confusion metric with the two-class prediction problem (binary classification), in which the outcomes are labeled either as credit default or non-credit default. Based on the metric, three kinds of credit scoring performance measures are used: precision, recall and false positive (FP) metrics. Precision, also known as specificity, measures how often the prediction correctly identifies credit default cases. It is calculated as the ratio of the number of loan applicants correctly identified by the predictive model as positives (TP) to the total number of loan applicants. Recall, also known as sensitivity, measures how often the predictive model correctly finds the right class for a loan applicant. It is defined as the proportion of true positives against potential correct examples. Table 6 shows the DT and QR prediction methods have achieved similar precision, recall and false positive rates in both training and validation datasets. The higher the value of precision is, the better the distinguishing capacity of the classifier. This means that the chosen mobile behavioral variables by these two methods, are useful in identifying the behavior of an applicant to repay a loan. The recall rates of both methods are also very close to each other, which proves the chosen mobile behavioral variables can help to identify the high-risk credit default events. Future study can explore different types of modern machine learning predictive methods to further utilize this type of data, and this is beyond the analysis in this paper.

Table 6. Confusion matrix measurement result summary.

Dataset	Model	Precision	Recall	FP rate
Training	DT	0.8321	0.9074	0.1513

Dataset	Model	Precision	Recall	FP rate
Lending	QR	0.8304	0.8982	0.1534
Validation	DT	0.8402	0.8903	0.1425
Lending	QR	0.8455	0.9086	0.1417

4. Discussion of Implications

The study conducted in this paper provides a few insightful findings. We discuss the main implications of such findings as below.

4.1. Deep Mobile Data Model

Our study proposes a large-scale mobile behavioral data-processing framework. The substantial size of our Chinese credit dataset having extensive 28-months data study period as well as breadth and depth of mobile user behavior with a total of 4,689 attributes. It facilitates the construction of new set of credit risk factors based on statistical diversity, intensity, consistency, and regularity of mobile user behavior characterizing user preferences, attitudes, geolocation, and temporal patterns. The carefully designed data structure provides a solid and comprehensive foundation for future research that focuses on further leveraging the power of mobile data; specifically, 5G network facilitates a much greater variety of data dimensions to be considered in credit risk field or other applications like customer churn, customer cross-selling etc.

4.2. New Alternative Data for Credit Risk Prediction

Traditional financial credit decision models mainly rely on

explicit personal traits such as age, income, gender, job type, and marital status as well as limited personal financial information such as credit bureaus while being oblivious to mobility or the habits of an individual. With the careful design of data categorization and data logic generation specific to credit risk application, statistically significant correlation results of derived data factors to credit default show that mobile behavioral data can be used as one of the new data sources in credit risk decisioning policy. Besides, the proposed data and respective factors can be utilized with other modern machine learning algorithms to analyze additional insights when using with other types of data as well as further improvement in credit scoring predictive accuracy and robustness.

4.3. Enable More Targeted Inclusive Financing and Servicing

With the new source of proposed deep mobile data, it makes possible for credit assessment based on non-financial data possible to the credit invisible people, which promoting inclusive finance to larger community in society. In addition, it also enables studying the data for mobile users to segment them based on their mobile usage pattern and profile. It would potentially formulate a new path of in-depth customer segmentation study with developing classification model to categorize users according to their characteristics and life-stage journey. Studying the profile of users among communities not only can provide more targeted inclusive financing products like different pricing with different loan size to meet different borrowing needs, but also better understand and identify their needs other than borrowing, for example, health or car insurance, investment and other services or products that match with particular user-segment characteristics.

5. Conclusion and Future Works

In this paper, we design a large-scale heterogeneous credit dataset, which is different from previous studies that mainly use call detail records and household survey data to infer mobile activities predicting credit default. The mobile behavior data model framework proposed composes of six user profile-specific categories and four sets of data factor measurements holistically represents mobile users' financial and non-financial behavior and habits. The six behavioral data dimensions include 1) personal spending level, 2) lifestyle, 3) personal interests, 4) personal credit status, 5) social networks and

6) life stage and demographics. It observes some interesting findings regarding both credit risk factor related and usage characteristics perspectives. The findings are important for a number of reasons, and we discussed in earlier section about their implications for mobile-behavioral based data framework potentially applied in credit market and other fields, especially when 5G network is fully adopted in the future, which can generate much boarder data dimensions; the deep mobile data as a new alternative data for big data credit scoring prediction and credit risk decisioning policy; more targeted inclusive financing and other servicing to mobile users.

The exploratory data analysis as well as decision tree and quantile regression models of the Chinese big mobile dataset validate that industry partitioners' feedback on mobile usage behavior reflects the underlying user behavior and, more specifically, attitudes and behavior toward personal financial management. The derived data factors can be used for further study on enhancing predictive power in credit scoring with applying modern machine learning methods as well as extracting user segmentation insights applying in other fields.

Several possible data and analysis optimization directions as follows. (a) A mobile data processing framework with associated algorithms to generate a unique set of credit risk indicators introduced in this study can be used to further analyze the properties of mobile data and develop insights in future credit scoring and other customer behavioral types of research. (b) Other areas of credit related analytics in addition to credit assessment can be leveraged the proposed big data framework, such as fraud detection, or loan renewal prediction. (c) The proposed data framework could be expanded including spatial location-based data, which can analyze user's mobility related behavior.

Abbreviations

APP	Application (Software Application in Mobile Device)
SMS	Short Message Service
AUC	Area Under the Curve

Conflicts of Interest

The authors declare no conflicts of interest. There are no direct competing financial interests or personal relationships that could have appeared to influence our work reported in this paper.

Appendix

Questionnaires and Responses:

Question 1: Types of loan businesses	Frequency of responses
a. Personal cash loan	10
b. Installment loan	9
c. Revolving credit	7
d. Secured loan	5
e. SME loan	7
f. Corporate loan	4
Question 2: Top factors to assess loan applicant's risk level	Frequency of responses
a. Personal income and spending	10
b. Credit history and status	10
c. Socioeconomic status	8
d. Personal living status	7
e. Life stage	9
f. Academic qualification	6
g. Related / Contract person profile	5
h. Others	3
Question 3a: Types of data sources for generating personal income and spending risk factors	Frequency of responses
a. Credit bureau	10
b. Government and utilities	9
c. Corporate financials	2
d. Asset class	5
e. Mobile device	8
f. Social network and contacts	7
g. Online e-commerce	8
h. Payment gateway	7
i. Banking transaction	9
Question 3b: Types of data sources for generating credit history and status risk factors	Frequency of responses
a. Credit bureau	10
b. Government and utilities	3
c. Corporate financials	0
d. Asset class	3
e. Mobile device	4
f. Social network and contacts	2
g. Online e-commerce	1
h. Payment gateway	2
i. Banking transaction	0
Question 3c: Types of data sources for generating socioeconomic status risk factors	Frequency of responses

a. Credit bureau	1
b. Government and utilities	2
c. Corporate financials	0
d. Asset class	0
e. Mobile device	7
f. Social network and contacts	8
g. Online e-commerce	3
h. Payment gateway	2
i. Banking transaction	0
Question 3d: Types of data sources for generating personal living trait risk factors	Frequency of responses
a. Credit bureau	5
b. Government and utilities	7
c. Corporate financials	0
d. Asset class	1
e. Mobile device	7
f. Social network and contacts	6
g. Online e-commerce	8
h. Payment gateway	8
i. Banking transaction	9
Question 3e: Types of data sources for generating life-stage risk factors	Frequency of responses
a. Credit bureau	6
b. Government and utilities	1
c. Corporate financials	0
d. Asset class	0
e. Mobile device	5
f. Social network and contacts	1
g. Online e-commerce	1
h. Payment gateway	1
i. Banking transaction	1
Question 4: Major data characteristics and benefits of using above data sources	Frequency of responses
a. Data update timeliness	7
b. Unaggregated raw data for credit modeling	6
c. Data acquisition cost	6
d. Historical data (e.g., 1 year or longer)	7
e. Data coverage nationwide	5
f. Others	2

References

- [1] Agarwal, R. R., Lin, C. C., Chen, K. T., Singh, V. K. (2019). Predicting financial trouble using call data – On social capital, phone logs, and financial trouble. *Applied Soft Computing journal* 74, 26-39. <https://doi.org/10.1371/journal.pone.0191863>
- [2] Arshed, N., Nasir, S., Saeed, M. I. (2022). Impact of the External Debt on Standard of Living: A Case of Asian Countries. *Social Indicators Research*, 163, pp. 321–340. <https://doi.org/10.1007/s11205-022-02906-9>
- [3] Berger, A., Udell, G. (2002). Small Business Credit Availability And Relationship Lending: The Importance of Bank Organizational Structure. *The Economic journal (London)*, 2002, Vol. 112(477), p. F32-F53. <https://doi.org/10.2139/ssrn.285937>
- [4] Butt, S., Phillips, J. G. (2008). Personality and self reported mobile phone use. *Computers in Human Behavior*, 24: 346 360. <https://doi.org/10.1016/j.chb.2007.01.019>
- [5] Carter, S., Yeo, A. C. M. (2016). Mobile apps usage by Malaysian business undergraduates and postgraduates: Implications for consumer behavior theory and marketing practice. *Internet research*, Vol. 26(3), p. 733-757. <https://doi.org/10.1108/IntR-10-2014-0273>
- [6] Chittaranjan, G., Blom, J., Perez, D. G. (2011). Who's with big-five: Analyzing and classifying personality traits with smartphones. In *Proceedings of the 15th Annual International Symposium on Wearable Computers. IEEE*, 29–36. <https://doi.org/10.1109/ISWC.2011.29>
- [7] Chu, Z., Wang, Z. W., Xiao, J. J., Zhang, W. Q. (2017). Financial Literacy, Portfolio Choice and Financial Well-Being. *Social Indicators Research*, 132, pp. 799–820. <https://doi.org/10.1007/s11205-016-1309-2>
- [8] Church, K., et. al. (2015). Understanding the challenges of mobile phone usage data. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services ACM*, 504–514. <https://doi.org/10.1145/2785830.2785891>
- [9] Grable, J. E., Park, J. Y., Joo, S. H. (2009). Explaining Financial Management Behavior for Koreans Living in the United States Malden, USA. Wiley, *The Journal of Consumer Affairs*, Vol. 43(1), p. 80-107. <https://doi.org/10.1111/j.1745-6606.2008.01128.x>
- [10] Huston, S. J. (2010). Measuring Financial Literacy. *Oxford, UK: Wiley, The Journal of Consumer Affairs*, Vol. 44(2), p. 296-316.
- [11] Kakinuma, Y. (2022). Financial literacy and quality of life: a moderated mediation approach of fintech adoption and leisure. *International Journal of Social Economics*, Vol. 49 No. 12, pp. 1713-1726. <https://doi.org/10.1108/IJSE-10-2021-0633>
- [12] Lin, Z. X., Whinston, A., Fan, S. K. (2015). Harnessing Internet finance with innovative cyber credit management. *Financial Innovation*, Vol. 1(1), pp. 1-24. <https://doi.org/10.1186/s40854-015-0004-7>
- [13] Luo, J., Li, B. Z. (2022). Impact of Digital Financial Inclusion on Consumption Inequality in China. *Social Indicators Research*, 163, pp. 529–553. <https://doi.org/10.1007/s11205-022-02909-6>
- [14] Oskarsdottir, M., et al. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing Journal* 74, 26-39. <https://doi.org/10.1016/j.asoc.2018.10.004>
- [15] Romero, J. J. (2011). NO. 1 Smartphones. *IEEE Spectrum*, 48(1), 28–31.
- [16] Seneviratne, S. et al. (2014). Predicting user traits from a snapshot of apps installed on a smartphone. *ACM SIGMOBILE Mobile Computing and Communications Review* 18, 2, 1–8. <https://doi.org/10.1145/2636242.2636244>
- [17] Singh, V. K., Bozkaya, B., Pentland, A. (2015). Money Walks: Implicit Mobility Behavior and Financial Well-Being. *PLoS ONE*, Vol. 10(8), p.e0136628.
- [18] Stiglitz, J., Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. *The American Economic Review*, 71, 393-410.
- [19] Wang, Y., Li, S., Lin, Z. X. (2013). Revealing Key Non-Financial Factors for Online Credit-Scoring in e-Financing. *10th International Conference on Service Systems and Service Management*, pp. 547-552.