Can Small Banks Compete on Local Data in Online Retail Loans? The Example of Baotou Rural Commercial Bank

Chen, Yunxiang

Can Small Banks Compete on Local Data in Online Retail Loans?

The Example of Baotou Rural Commercial Bank

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

Chen Yunxiang

(FCO N° 58501)

Dissertation Supervisor:	Professor Harald HAU University of Geneva
Associate Supervisor:	Professor FAN Wenzhong Tsinghua University

September, 2020

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： CHEN Yunxiang Date： 09/2020

Acknowledgment

Thanks Professor Harald Hau and Professor Fan Wenzhong for their guidance on this paper.

Thanks for the support provided by Baotou Commercial Bank, especially Shanshan Wang, Bing wang, Zhilong Duan and Shi Wang.

Thanks Can Zhu for her sincere support.

Abstract

At present, more than 2000 local small- and medium-sized banks in China are facing fierce competition in terms of retail loans between online internet company and offline homogeneous banks. However, these small and medium-sized banks are not lack of opportunities, but lack of capacity. With the rapid development of Internet technology, small and medium-sized banks can gain competitive advantage by using the local big data resources which are relatively easy to obtain.

Through the application of an online ‘Citizen loan’ personal credit product of Baotou rural commercial bank (BTB), this paper deeply studies and analyzes all kinds of data fusion and modeling available to small and medium-sized banks, and establishes a localized ‘Citizen loan’ risk control model by using XGBoost and "Knowledge Graph" tools. In this practice process, this paper deeply compares and analyzes the advantages and disadvantages of the models based on public data sets and local data sets (superimposing the former), and concludes that the optimization local model based on local data sets has better performance. This ‘Citizen loan’ risk control model based on local data set was actually deployed and applied in BTB in 2019, and has greatly improved both customer benefits and bank benefits.

This result fully shows that the research results of this paper can help small and medium-sized banks to find a feasible way to realize the transformation of retail strategy through financial technology empowerment.

Key words: Citizen Loan; Credit Risk Modeling; Data Feature Engineering; XGBoost; Knowledge Graph

Table of Contents

List of Tables

List of Figures

Can Small Banks Compete on Local Data in Online Retail Loans? The Example of Baotou Rural Commercial Bank

1. Introduction

1.1. The Difficulties and Problems Solved in Baotou Rural Commercial Bank

In China, there are more than 2000 local small- and medium-sized banks similar to Baotou Rural Commercial Bank (BTB) (Li, Sun and Zhu 2018). These banks are facing fierce competition both online and offline in terms of retail loans (Sun 2019).

On the one hand, the retail loan products launched by the Internet financial institution, such as 'Jiebei loan' by Alibaba and 'Weili loan' by Tencent, based on their very large amount of transaction data (Nicoletti 2018), have the characteristics of fast loan application (generally completed in 20-30 minutes), small amount (average 5000 yuan per loan), simple procedures, good experience and so on. These online platforms cause continuous loss of traditional customers, especially young customers, for small and medium sized banks.

On the other hand, the pressure of homogenization competition among local banks is also increasing (Bian 2017). Because there is no better risk control based on big data analysis, small and medium-sized banks can only focus on high-quality people based on identity (such as high-salary people in enterprises and institutions). This traditional way cannot form unique competitive advantages. Due to the price war of retail loans, these small and medium banks cannot develop quickly. Their asset quality cannot be guaranteed as well.

However, small and medium banks could still gain predominance in the competition by applying financial technology. With the rapid development of Internet technology, small and medium-sized banks have the advantages in aspects of quicker decision process, easier effective data gathering and higher customer loyalty, which could bring opportunities for them and achieve new development. In more detail,

1. local banks could gather a large number of local public data in local public enterprises and institutions (such as social insurance, individual income tax, accumulation housing fund, communication operators, etc.);

2. Customers, especially young customers, generally use various mobile phone smart applications. These applications generate a large number of personal data that can also be fused with local data to generate new financial value;

3. There are a large number of local Internet applications, such as a hospital app, community app and local e-Business to Customers, which can provide cooperation opportunities for small and medium-sized banks, on the one hand, and also provide a large number of local data, on the other hand;

4. The actual survey shows that even if Internet finance is rapidly popularized, the penetration rate of personal credit of local customers is still less than 40%[1], local banks still have a chance to gain market share.

Therefore, this research is interested in how local small- and medium-sized banks could face the fierce online and offline competition, and how to seize opportunities to be used in the local market.

The research question of this paper is ‘Can local banks innovate and establish a localized ‘Citizen loan’ risk control model to achieve new development in the FinTech industry?’ The BTB is used as a case study to answer this question.

This paper discusses the application of an online ‘Citizen loan’ personal credit product launched by BTB in 2019, deeply studies and analyzes various data fusion and modeling available to small and medium-sized banks, and establishes a localized ‘Citizen loan’ risk control model by using XGBoost and Knowledge Graph tools.

In the process of research and practice, this paper deeply compares and analyzes the advantages and disadvantages of the models based on public data sets and local data sets (superimposing the former), and draws an exciting conclusion that the optimization local model based on local data sets has better performance. This ‘Citizen loan’ risk control model based on local data set was actually deployed and applied in BTB in 2019, and has greatly improved both customer benefits and Bank benefits.

This result fully shows that the research results of this paper can help small and medium-sized banks to find a feasible way to realize the successful transformation of retail strategy through financial technology empowerment.

1.2. Literature Review

1.2.1. Research on the Credit Risk Feature of Citizen Loan

Loan features are very important to this study. The first step of modeling is to select and optimize the features. This paper has noticed that there have been some studies on different features (Dai, 2018). Normally, the loan features can be classified into basic features, internal features and external features. The basic features of loans refer to loan amount, term, interest rates, usage of loan repayment method and so on. Internal features refer to the age, gender, housing, occupation and work of citizens. External features refer to other external factors such as external records, lending institutions, guarantees, policies and so on. Through the division of different features, the assessment of credit risk feature of Citizen loans can be more intuitive and easier to understand.

1.2.1.1. Research on Basic Features

Baptista, Ramalho and da Silva (2006) concluded that the credit risk impact indicators of farmers' loans are as follows: loan amount, loan terms, loan usage, legal records, business philosophy, and business level. Through the analysis of microfinance in Bangladesh, Rubana Mahjabeen (2008) believes that the loan risk will be affected by the total loan amount, loan cycle, and the value of durable goods owned by the lender.

Lv and Lv (2011) concluded that the loans limit significantly affected the repayment behavior of farmers' loans, and played a positive role. Zhang, Du, Jing et al. (2017) found that the correlation.

[1] According to the investigation report of BTB in local.

between credit level indicators and farmers' real default is not significant, which means that the internal credit rating of local credit institutions for loan farmers cannot effectively predict the credit risk of farmers. Micro indicators such as interest rates, gender, marital status, occupation, education has a greater impact on credit risk.

1.2.1.2. Research on Internal Features

Compared with basic features, the internal factors of citizens basically reflected some features of citizens. From the research level, it will be slightly less than the basic features. The research all over the world basically conforms to the general structure.

In the 1930s, Durand constructed a Consumer Credit Scoring System including age, gender, residential stability, occupation and job stability (Chi 2010, Cui, 2005). James Copestake (2007) conducted a questionnaire survey on financial institutions, and found that the main influencing factors of farmers' loan risk were the lender's profile, gender, age, the number of family workers, and the family's net assets. Jha and Bawa (2007) studied the case of small loans in India, and concluded that education level, family income, legal constraints, fixed assets and so on affect the loan risk.

Lianwei Chen (2008) used the decision tree algorithm to select 15 indicators, including age, gender, family income, loan history, judicial records and so on, from the natural situation, family situation and credit situation to establish the evaluation system of farmers' loans. Junyang Cui (2011) extracted 6000 samples of farmers from rural credit cooperatives in Jilin Province, and conducted binary regression analysis on the influencing factors of farmers' loan default. The results showed that gender, age, education level, family income and expenditure of household played significant roles in farmers' default behavior. Lanxiang Zheng and Xue Wan (2014) did Logit analysis by selecting 13 indicators such as labor force, age and education levels from the credit rating table of rural microfinance companies in Feixi County, Anhui Province. Before determining the optimal variable index, normal hypothesis test, heteroscedasticity test and multicollinearity test are used to determine the available indicators of the model. Meng Xia, Banghong Zhao, Junqin Wang (2015) conducted a correlation analysis study. The results showed that the education level of farmers, the number of family labor force, business status, credit status and so on have a great impact on the possibility of farmers defaults on loans.

1.2.1.3. Research on External Features

The research on the external characteristics of Citizen loans is affected by various factors, which is obviously less than the basic characteristics and internal characteristics. On the one hand, it is affected by the acquisition of extension indicators, on the other hand, it is also affected by the effectiveness of indicators.

Hartarska and Nadolnyak (2007) analyzed the situation of small loans issued by the World Bank. The results showed that the knowledge and ability of the lender and whether there was a guarantee or not it was the main influencing factors of loan risk. Zhengbo Li, Jie Gao, and Weijie Cui (2006) indicated 19 evaluation indexes including the amount of labor force, credit union reputation, service and so on. Guotai Chi and Wei Wang (2009) established a set of Scientific Development Evaluation Index System including five aspects of economy, ecology, society, human development and technology through correlation analysis, correlation analysis and cluster analysis. Qiang Wei and Diguang Wang (2016) combined the special risk factors of P2P online lending platform, and individual borrowers and small and micro enterprises, a total of 18 indicators were selected to build an online loan index system.

1.2.2. Research on Credit Risk Evaluation Method of Citizen Loan

Credit risk evaluation methods can be divided into three types, mainly according to the basic disciplines, including statistical methods, operational research methods and data mining methods. The XGBoost algorithm applied in this paper also belongs to a data mining method. XGBoost (extreme gradient boosting) algorithm is an efficient implementation version of gradient boosting algorithms. Because of its excellent effect and efficiency in application practice, XGBoost is widely praised by the industry.

XGBoost algorithm is an integrated learning model proposed by Chen and Guestrin (2016) to improve the gradient lifting decision tree model. The decision tree in the algorithm has sequential correlation. Based on the prediction error of the previous round, the prediction error of each round is used to construct the model iteratively to improve the accuracy of prediction.

At the research level of common data mining methods, Dutta and Shekhar (1988) applied neural network prediction models to bond credit rating. Tao Wen et al. (2004) established a farmer credit rating model based on BP neural network. The credit status of 150 farmers in Chongqing is studied. Jiana Xu, (2004) established the Ahp-ann model for credit risk assessment of commercial banks by combining the artificial neural network credit risk assessment technology with the analytic hierarchy process (AHP). Cai et al. (2011) and others used the decision tree method to evaluate the credit risk of farmers' micro loans in a credit cooperative. Tao Zhang (2017) established a decision tree classification model with 8 factors as independent variables, such as household features and household features, with farmers' willingness to loan as the target, so as to provide valuable classification rules for rural financial institutions. The results show that the values of 7 factors, such as household population, land area and education level of the head of the household, can form 10 classification rules with different attributes.

For other data mining methods, Wang et al. (2005) used modified ants algorithms, MAA. This paper makes an empirical study on the actual cases of commercial banks. Yanling Wei (2009) used fuzzy clustering method to divide loan farmers into groups with different credit characteristics. Rural credit cooperatives further analyzed their credit risk to provide theoretical support.

Knowledge Graph, also known as the map of scientific knowledge, is a series of different graphs showing the relationship between the development process and structure of knowledge. It describes knowledge resources and their carriers with visualization technology, and excavates, analyzes, constructs, draws and display knowledge and their interrelations. The concept evolution of Knowledge Graph has gone through the stages of semantic networks, ontology, web, semantic web, linked data and so on. It was proposed by Google in 2012. Google hopes to build the next generation of search engines through knowledge graph, so as to optimize the search results. In 2018, Fengyu Lei (2018) described the association network between credit entities through the knowledge graph efficiently and intuitively. With the help of the knowledge graph technology innovation and full-dimensional portrait of the subject, the real situation of the subject can be realized. With the help of the knowledge graph, many problems and shortcomings of the previous traditional risk management schemes can be realized, and better help financial institutions to carry out risk management. In 2018, Xiaoyang Yu (2018) proposed a knowledge graph construction scheme with entity acquisition and entity-relationship extraction as the main means through the general framework of knowledge mapping technology. This paper designs and implements an information system based on the knowledge graph of capital market figures and enterprises.

1.2.3. Summary

This chapter mainly studies and analyzes the literature related to personal loans credit risk evaluation from two aspects: the selection of credit risk evaluation features and risk evaluation methods, and has a more in-depth understanding of the current domestic and foreign research. It should be said that the previous research has made more valuable achievements and summed up a wealth of experience. It enriches the research results from different research perspectives and research methods, and plays an important reference role in the research and application of personal credit risk evaluation.

From the content of literature research, it is not difficult to find that the current research on personal risk evaluation index mainly focuses on the basic elements and internal factors of personal loans. The problems of limited sources of evaluation information, few evaluation dimensions and insufficient pertinence are common. With the continuous development of Internet technology and the continuous expansion of everyone's digital survival, the data sources for individual customers are more extensive and more complete. Therefore, in a more objective way, in a better data environment, the credit risk evaluation of personal loans can be further studied. This paper uses external big data resources and local big data resources to further study personal credit risk evaluation, which is also based on this big premise.

2. Methodology

2.1. Problems of Retail Loan of BTB before 2019

Before the practice of this paper in 2019, the main problems of BTB's retail loan are as follows:

1. The external Internet network banks began to penetrate the local market. For example, Haoqi loans (ZhaoLian Consumer Finance Company), Haoren loans (XWBank) and XinYiDai Loan (Ping An Puhui Enterprise), etc. According to market research, these products have good performance in terms of credit granting groups, approval efficiency, interest rate, average line of credit and customers' convenience, which makes local agricultural commercial banks face greater challenges in small and microcredit and personal consumption loans.

Company	Product	Credit Group	Efficiency of Approval	Annual Interest Rate	Average Line of Credit	Customer's Convenience	Strength
ZhaoLian Consumer Finance Company	Haoqi Loan	Consumers with high Zhima credit score	2 mins	15% - 36%	12,000	★ ★ ★	The application details are simple; give the credit line within 2 minutes.
WBank	Haoren Loan	A wide range of customers	2 mins	10.8% - 28%	3,500	★ ★ ★	Each step has a marketing action; application details are relatively simple, and most of them are optional options; the credit line is given within 2 minutes.
Ping An Puhui Enterprise	XinYiDai Loan	Small and micro entrepreneurs got in offline	2 mins for the first time and 30 mins for raising the credit line	7.5% - 24%	135,000	★ ★	Large credit line, wide interest rate range, customer manager service, maintenance.

Table 1: Analysis of External Internet Banking Products (2018)

2. Other local banks seized the market. At present, all kinds of local commercial banks have occupied the credit group of small and micro businesses. Due to the low capital cost and low loan interest rate to customers, most of the head customers, such as CCB Kuai loan, Postal savings bank Jisu loans, Mengshang bank Shangyingbao Loan and Jingu agriculture bank's farm loan products, have good performance in credit granting group, approval efficiency, interest rate, average line of credit and customers' convenience. As a result, the quality of customers obtained by farmers is relatively low, and the stock market is gradually eroded. Taking the credit market in rural areas as an example, the market share of BTB decreased from 65% (2015 data) in the past to 48% (data in 2019);

Company	Product	Credit Group	Efficiency of Approval	Annual Interes t rate	Average Line of Credit	Customer's Convenience	Strength
CCB	Kuai Loan	Old customers	5 mins	4.3% - 5%	5,000	★ ★ ★	The examination and approval speed are fast, the interest rate advantage is very obvious, and the operation is completed online.
Postal Savings Bank	Jisu Loan	Small and micro entrepreneurs got in offline	15 mins	7.5% - 10%	35,000	★ ★	The credit line is high, up to 200000 yuan, and the approval speed is fast.
Mengshang Bank	Shang-yingbao Loan	Small and micro entrepreneurs got in offline	7 Working Days	7.5% - 12%	155,000	★ ★	Large line of credit up to 1 million, flexible products, customer manager service, maintenance.
Jingu Bank	Farmer loan	Customers in rural areas	3 Working Days	7% - 9%	85,000	★ ★	High credit line, fast approval and low interest rate.

Table 2: Analysis of Other Local Banking Products (2018)

3. Lack of external data and the rate of NPL was very high. Before 2018, BTB did not introduce big data resources for risk assessment, relying solely on the credit reporting records of the central bank, making subjective judgments, and the standards are not uniform. As a result, the quality of credit assets is uneven, and the key influencing factors cannot be evaluated. When risks are found, effective risk control measures cannot be adopted in a consistent manner. The rate of NPL is high, and it is difficult to control effectively. Only relying on manual investigation and subjective experience approval, the rate of NPL in retail loans was 3.2%.

Personal Credit Reference in PBC (People's Bank of China)	Personal Data Owned by BTB
Cumulative overdue times in recent 3, 6 and 12 months	Customer Ranking
Having special transaction type in the past 2 years including repayment by guarantor, etc.	AUM balance, deposit balance and financial balance
Accumulated overdue times of credit card in recent 3, 6 and 12 months	Asset Information
The frequency of 1, 2, 3 + in recent 6, 12, 24 months	Liability Information
Overdue amount of current loans and credit cards	Business information
Number of queries on loan approval and credit card approval in the last 2, 3, 4 and 6 months	Revenue Information
Outstanding monthly mortgage payments	Family member information

Table 3: List of Data Used as Risk Assessment (2018)

Time	Performance
First Quarter in 2018	2.85%
Second Quarter in 2018	3 .20%
Third Quater in 2018	3.11%
Fourth Quater in 2018	3.20%

Table 4: Cumulative NPL of Retail Loans of BTB before Introducing External Data (2018)

Time	A Sub-Branch	B Sub-Branch	C Sub-Branch	D Sub-Branch	E Sub-Branch	F Sub-Branch
First Quarter in 2018	1.70%	3.45%	3.25%	2.36%	1.11%	1.89%
Second Quarter in 2018	2.28%	1.76%	1.16%	3.13%	3.32%	2.28%
Third Quater in 2018	1.14%	3.36%	1.27%	1.44%	3.17%	1.12%
Fourth Quater in 2018	2.85%	1.37%	2.54%	1.55.1%	1.77.5%	3.55%

Table 5: Cumulative NPL of Retail Loans of Different Sub-Branches of BTB (2018)

(Due to the different standards of credit assessment, the NPL performance of the same sub-branch is unstable)

4. The cost of developing customer is high and the occupation of human resources is high. Small and high-frequency loans are the main positioning of BTB's personal credit. When granting loans, they rely on labor-intensive way to operate, which results in large personal investment and high loan operating costs. At present, the average number of customer managers managing accounts have reached 300, but the per capita amount of account management is only 18.9 million, which is not conducive to the expansion of new loans and post loan management. The increase of personnel will increase the cost of exhibition industry, which will lead to a dilemma in cost control and business expansion.

Items	Investment Cost	Information Cost	Processing Cost	Other Cost (Notarization and Evaluation)	Total (100,000 yuan as an example)
Credit and Guaranteed Loan	150 yuan/person	20 yuan/person	0	0.3%*credit line	470 yuan
Mortgage	250 yuan/person	20 yuan/person	80 yuan	1.3%*credit line	1650 yuan

Table 6: The Cost of Developing Customer Analysis (2018)

Item	Account Manager Develops Customers	Time to Go through the Formalities
Credit and Guaranteed Loan	5 working days	3 working days
Mortgage	10 working days	10 working days

Table 7:Time and Cost Analysis of Development and Promotion, Comparison of Customer Manager's Time and Procedures (2018)

5. There were a lot of offline collected data, which were not fully utilized. The application form information, such as identity information, education information, work information, family information, social relations and other information are the data advantages of BTB. They are scattered in 14 different systems and have no ability to integrate and quantify. The characteristics of bad debt customers in the past have not been quantified and refined, so they do not have the guidance role for individual credit.

2.2. Available Resources

The key to solve the problem is still big data, which is the basis of all analysis and modeling. Therefore, for BTB, in addition to obtaining the same big data resources as its external competitors, it is a feasible way to excavate local big data resources and form competitive advantages locally.

Fortunately, BTB, after all, is a bank that has been deeply cultivated in local for many years. In the process of long-term cooperation with various local institutional customers (such as local social security, provident fund, communication corporations, agricultural bureau, etc.), BTB can legally use all kinds of data resources of these institutions for individual customers through technical means, and quantify and label them. These customer information scattered in various local institutions can be associated through key fields such as ID card number and mobile phone number to form local big data resources.

According to the local research, both the external competitors and the local competitors of BTB have not applied the local big data resources on a large scale.

Local big data has the following values:

First, the application of external big data with local features, such as local e-commerce transactions, social data and mobile communication information, will be more accurate in local customer groups.

Second, local banks have more advantages in obtaining big data with local features, such as local provident fund data, tax records of local small and micro enterprises, medical care, education, etc. BTB can dig customers and control risks in these scenarios, which is an advantage that online Internet credit institutions do not have.

Third, local institutions can obtain a large number of data through manual collection by taking advantage of sub branch network advantages, and these data can be precipitated, which is not available in other competitors. BTB has a long history of local operation, and has a good foundation in local customers, especially the data collection and filing of farmers are relatively complete and comprehensive. In particular, BTB has natural advantages in mastering the villagers' social relations, such as the degree of family harmony, the relationship between father and son, and brother-in-law.

At the same time, with the help of all kinds of nationwide big data companies which have developed rapidly in recent years, BTB can also fully obtain the external big data resources of local customers (such as external special loan information (customer credit information that he borrowed money from different parties), e-commerce transaction records, social information, etc.) like external Internet financial competitors.

In this way, the local big data resources + external big data resources constitute the data resources that can be used for modeling in this paper, which makes the research of this paper have a solid foundation, and on this basis, the established risk control model is continuously optimized and evaluated, and the ultimate goal is to form a competitive advantage for competitors locally.

2.3. Data Sources

The target customer of samples refers to Baotou City in Inner Mongolia Autonomous Region. This city is a typical fourth tier city in China, with good industrial and agricultural foundation. There are more than 300 similar cities in China, involving a population of 300-400 million. In this city, there are about 2.6 million permanent residents, aged from 18 to 60 years old, who have lived for more than a year, including about 400 thousand farmers. Local citizens (including farmers) are the main service objects of the BTB. The BTB, relying on their own sub branch network advantages, get a large market share in the local agricultural areas.

Due to the development of the external big data environment of the whole country and Baotou City for many years, the data sources that BTB can obtain and can continuously obtain within three years are mainly divided into the following six categories:

1. Bank internal data sources (including 114,356 individuals, whose name, ID number, telephone number, account opening information, deposit performance, loan performance, blacklist, etc.); these data have about 96,516,464 fields.

2. External big data sources form local financial institutions (including 114,356 individuals credit reference from PBC (People's Bank of China), in detail, personal credit information and debt performance in those financial institutions) these data have about 77,533,368 fields.

3. Internet external big data sources (including about 114,356 local people, whose blacklist and external credit information that he borrowed money from different parties, e-commerce information, social information, mobile communication record and application information, etc.); these data have about 552,568,192 fields.

4. Local external big data sources (including about 64,357 local people, whose customer blacklist and credit information that he borrowed money from different parties, customer credit score, social security, personal tax, provident fund, communication records, civil affairs, etc.); these data have about 2,188,138,000 fields.

5. Local scene big data sources (including about 59,453 local people, whose medical, education, transportation, business, etc.); these data have about 1,367,419 fields.

6. Big data sources of farmers (including 12,468 local people, whose family relations, neighborhood relations, land acre, annual income, etc.); these data have about 10,522,992 fields.

Item	Overall Data			BTB Data		
	Persons	**Information Volume of Single Person**	**Total Number of Information**	**Persons**	**Information Volume of Single Persons**	**Total Number of Information**
Internal Data	612,543	18	11,025,774	114,356	844	96,516,464
External Financial Data (Credit Reference)	825,436	156	128,768,016	114,356	678	77,533,368
Internet Big Data	1,414,356	4,832	6,834,168,192	114,356	4,832	552,568,192
Local External Big Data	2,403,552	34,000	81,720,768,000	64,357	34,000	2,188,138,000
Local Scene Big Data (POS e-Commerce Data)	1,825,088	23	41,977,024	59,453	23	1,367,419
Big Data of Farmers	63,564	156	9,915,984	12,468	844	10,522,992
Total			88,746,622,990			2,926,646,435

Table 8: Quantity Analysis of Various Big Data (2018)

(Note: the overall data refers to the total amount of the data based on the local market, while the BTB data refers to the data that can be directly obtained based on the stock customer data of BTB.)

According to the needs of modeling in this paper, the above six parts of data are mainly divided into two parts:

Data Source 1 (Public Big Data): big data inside the bank; big data from external financial institutions; big data from the Internet;

This part of the data source is available to any competitor of an external bank, which is relatively a public resource; the main external competitors of BTB can obtain such data.

No.	Data Source 1	Field Description
1	Internal Data	Account opening date, deposit performance, business data... A total of 55 fields.
2	External Financial Data (Credit Reference in PBC)	In the past 3, 6, 12 months, the cumulative number of overdue; hit the special transaction type in the past 2 years, there are guarantor repayment and repayment with assets; credit card overdue cumulative times in the past 3, 6, 12 months; 1, 2, 3 + times in the past 6, 12, 24 months; current loan, credit card overdue amount; loan approval, credit card approval query times, the amount of outstanding monthly mortgage, the number and amount of unsettled loans and the monthly payment amount of credit cards…, which have 156 fields.
3	Big Data from the Internet	There are 80 fields in total, including the blacklist of court executors through ID card, the high risk of non-banks (including all non-bank types) through ID card, the number of application institutions in non-bank institutions - licensed online small loan institutions in recent 6 months, the number of applications in non-bank institutions in recent 6 months, and the number of applications in non-bank institutions in recent 6 months. ...

Table 9: Data Source 1 Field Analysis Table (2018)

Data Source 2 (Local Big Data): local external big data; local e-Business big data; farmer big data;

Compared with data source 1, these data sources obtained by BTB are localized data. These data have distinctive characteristics of local customers in Baotou, but external Internet banks cannot or are difficult to obtain them. This part of the data source is the key basis of this study.

No.	Data Source 2	Field description
1	Local External Big Data	There are 50 fields, including: Personal communication information, suspected card maintenance (abnormal use number) identification, average bill amount in the past three months, mobile phone number attribution city label, mobile phone number total shutdown frequency tag (yuan) in the past six months.... The base number of provident fund deposits, the monthly deposit amount of provident fund, etc. The base and amount of individual income tax payment. Personal car ownership. Personal marital status. Personal household registration status.
2	Local e-Business Big Data	Sales amount, number of customer orders, customer unit price, product number, transaction time... A total of 14 fields.
3	Farmers' Big Data	There are 156 fields including basic information, village group, main family members, house property, land and agricultural machinery….

Table 10: Data Source 2 Field Analysis

2.4. Modeling with External Public Big Data and Local Data

The key points of this study are as follows:

1. External public big data (data source 1) is used to model and build Public model to achieve the same performance as competitors.

2. On the basis of external public big data modeling, the local data source (data source 2) is introduced. After the combination of the two data sources, the joint modeling is carried out to build the Local model, and the Knowledge Graph and other tools are used to solve the difficulties and problems of feature selection in the process of joint modeling. The goal is to make the optimized results of joint modeling significantly better than the former.

3. The cost-benefit analysis of joint modeling is carried out to determine the economic feasibility of joint modeling.

The selected modeling methods are as follows:

The modeling method selected in this study is XGBoost, which is mainly based on the following reasons:

1. At the same time, it can also process large-scale data efficiently.

2. It is easy to engineering and easy to realize in practical application.

3. The algorithm has good effect and strong ability to solve practical problems.

4. High tolerance for missing and missing data.

2.5. Our Findings

<u>In the research of this paper, the first aspect</u>: through two data sources, the existing data sets (30774 BTB 2018 sample data) are compared by using XGBoost, and the modeling effect is very significant:

Item	Data Source 1 (Public Model)	Data Source 1 + Data Source 2 (Local Model)
Under the Same Passing Rate (Pass Rate = 80%)	Bad Debt = 1.9%	Bad Debt = 1.1%
When the Bad Debts Are Consistent (Bad Debt <= 1%)	Passing Rate=30%	Passing Rate=75%
Lift Analysis (Decile=10)	3.62	5.56

Table 11: Comparative Analysis of Modeling Effectiveness of Two Different Data Sources

Under the same pass rate, the bad debt rate of the Local Model is lower than that of the Public Model; under the same bad debt rate, the pass rate of the Local model is higher than that of the Public Model. If the bad debt rate is controlled within 1%, the pass rate of the Local Model can reach 75%, while that of the Public Model can only reach 30%; if the pass rate is controlled at 80%, the bad debt rate of the Local Model is 1.1%, and that of the Public Model is 1.9%.

From the lift analysis, the gain of local model is better than that of the Public Model.

<u>In the second aspect of this paper</u>: the new local model is used to establish the Citizen loan products, which were operated in one year in 2019, and the actual income is very significant:

Before the system went online (2018) and after the system went online (2019), comparisons are as follow:

Customer benefit analysis: (1) the acquisition time of a small loan was reduced from 3 days to 3 hours; (2) the number of transactions: the number of transactions increased from 73,788 in 2018 to 197,233 in 2019, with wider customer coverage; (3) interest rates: the average annual interest rate of customers decreased from 10.96% to 8.72%;

Analysis of different types of customers: (1) it can cover 70% of the local population, only local people who have no bad records can borrow more than 8,000 yuan; (2) among them, it can cover 800,000 customers who have not borrowed money before; (3) the average credit line of the original 800,000 customers having personal credit references can be adjusted from 3,000 yuan to 20,000 yuan;

Bank profit analysis: (1) the loan balance reached 1.535 billion in 2019; (2) credit line: the average credit line of customers reaches 43,728 yuan, and the risk is further dispersed, which is significantly better than that in the past (96,683 yuan), close to the Internet Bank (5,000 yuan) and better than other local banks (72,187 yuan); (3) the current non-performing loan and expected non-performing loans are less than 1.2%; and (4) the estimated net profit can reach 4.4%. (5) cost: the cost decreased rapidly, which decreased by 60.04%;

In this paper, the third aspect: in the specific research innovation, this paper mainly makes two important innovations:

1. On the basis of data source 1 and data source 2, XGBoost modeling tool is applied to model, and good results are achieved.
2. Compared with data source 1, the local data source has a large number of problems such as data error and data missing. In this paper, we use the Knowledge Graph to clean the local information of data source 2 and enhance the domain knowledge by adjusting the weight. After cleaning and adjusting, the data quality has been significantly improved, and the effect of the model has been further optimized.

In the application of Knowledge Graph, Knowledge Graph standardizes the analysis framework and prediction process of different data sources.

There are three supplements and improvements to the model:

(1) Any node in the Knowledge Graph can be predicted and analyzed. For example, to study the integrity of a family, we only need to select all the nodes that are related to a family's relatives as the initial variable set;

(2) The structural information in the Knowledge Graph can be brought into the quantitative model. For example, using principal component analysis (PCA) to reduce the dimension, we can consider summarizing the changes of items under each inclusion relationship into a principal component grade;

(3) The traditional quantitative model can be superimposed on the perspective of the field and industry knowledge. For example, if the price of agricultural products is good, the willingness and ability of customers to repay will be high.

Using the Knowledge Graph, for example, a significant finding and benefit is that the IV value of "net family income of immediate family members" is calculated to be 0.342. In other words, the dimension of "family net income of immediate family members" has high modeling value.

3. The Personal Credit Analysis of BTB Before Building Big Data Model

In 2016, 2017 and 2018, the personal credit approval of BTB did not rely on big data, but mainly relied on manual model for risk judgment. The specific data of credit line, credit extension increment, loan average, average interest rate, loan number, development cost, processing speed, non-performing loan rates, etc. in the past 3 years are as follows:

Year	Loan Balance (10000 yuan)	New Loan (10000 yuan)	Loan per Person (10000 yuan)	Average Annual Interest Rate of Loan	Number of Loan Persons	NPL Rate
2016	677,651.04	14192.2	13.019	11.52%	52086.	3.31%
2017	684,257.61	6606.57	11.54	11.32%	59656	2.98%
2018	704,018.62	19761.01	9.60	10.96%	73788	3.20%

Table 12: Personal Credit Data of BTB in 2016, 2017 and 2018

Year	Developing Customer Cost Per Person		Processing Speed (Weekday)			
			Developing Customers		Handle Procedures	
	Credit	Mortgage	Credit	Mortgage	Credit	Mortgage
2016	450	1590	3	8.5	2	10
2017	450	1620	4.5	9	2.5	10
2018	470	1650	5	10	3	10

Table 13: Personal Credit Cost and Processing Effectiveness Data Sheet of BTB in 2016, 2017 and 2018

As of the end of 2018, there were a total of 73788 outstanding personal loans of BTB, with an average of RMB 96 thousand, an average annual interest rate of 10.96% and a non-performing loan rate of 3.20%. These customers are distributed in 16 retail products of the bank, with the main structure of mortgage loans accounting for 36% and credit loans accounting for 64%.

Observing the data of 3 years from 2016 to 2018, we can see that:

1. During the 3 years from 2016 to 2018, the loan balance increased by 264 million yuan in total, and the average increase in the 3 years was 135 million, which was relatively sluggish;

2. During the 3 years from 2016 to 2018, the average of credit line decreased year by year, the number of loans increased year by year, the cost of a single business development increased year by year, the growth rate of the overall business development cost exceeded the growth rate of the scale of the overall loan balance, and the cost remained high. However, the average annual interest rate of loans decreased year by year due to market factors, and the profit margin decreased year by year;

3. During the 3 years from 2016 to 2018, the business processing speed was in a long processing cycle, customer satisfaction decreased, the credit competition in the external market was relatively fierce, and the loss of high-quality customers was serious.

4. During the 3 years from 2016 to 2018, the non-performing loan rate of retail loans remained high, the development cost increased, the loan interest rate decreased, the non-performing loan rate was at a high level, which further squeezed the profit space, and the personal credit business was in an unsustainable development state.

Therefore, since 2018, BTB has introduced various big data sets, and selected 30774 customers with complete data, clear structure and high data value from 73788 personal credit customers in 2018 as examples, established a new risk control model, launched new products of public loans, and adopted a new model for loan extension and management.

4. Analysis

4.1. Public Data vs. Local Data: Data Preparing

Through 3 types of interface methods, BTB constructed its own data set, and the relationship between the interface and the data source is as follows:

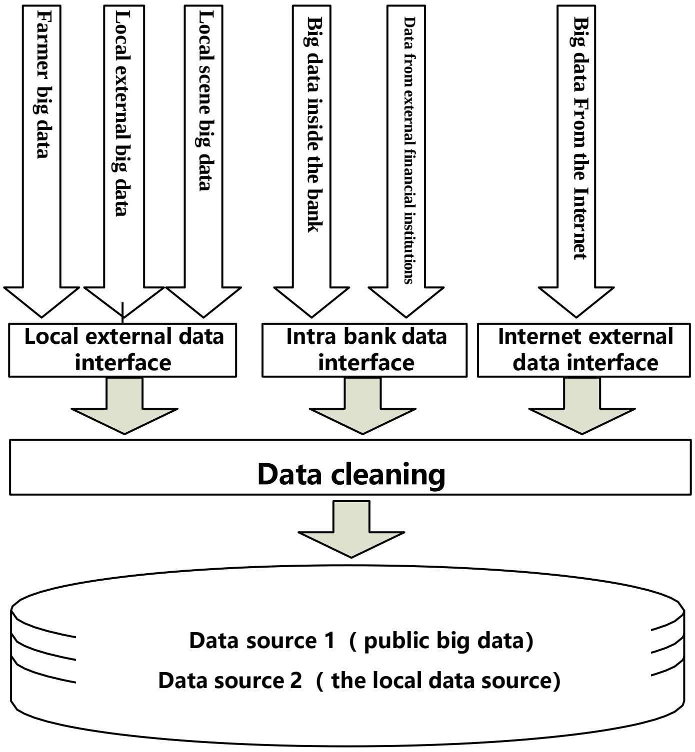

Figure 1: Data Preparation Process

In the modeling work, exploratory data analysis and data cleaning of data samples account for 70% of the workload, so the main work of risk assessment model is to build and determine the labels of each field in the data set. In the process of determining the label indicators, extreme value and non-normal distribution labels are removed through variable basic information analysis, variable distribution analysis, variable frequency analysis, etc.; data is reexamined and verified through data cleaning, with the purpose of removing duplicate information, rectifying existing errors, and ensuring data consistency. The two most important steps are as follows to build an effective label system and prepare for modeling.

4.1.1. Consistency Checking

According to the reasonable value range and mutual relationship of each variable, check whether the data meets the requirements, and find out the data beyond the normal range, which is logically unreasonable or inconsistent. Simple consistency check such as gender, age, etc., for example under the same ID card number, according to the logic of the ID card number, verify and correct the gender and age fields in the data source; for some data in the case of different source values, for example, use the voting mechanism to set up a consistency check, such as the customer's family population, the value has (2 persons, 3 persons, 3 persons) 3 situations, according to the Vote and select 3 persons.

4.1.2. Handling Invalid Values and Missing Values

Due to the survey, coding and input errors, there may be some invalid values and missing values in the data, which need to be properly handled. Common processing methods include estimation, whole column deletion, variable deletion and pair deletion. For example, when processing the POS records, the consumption amount is necessary. For the data that cannot be obtained in this field, the whole column is deleted. For the data with a single amount of more than 100 thousand yuan, it is replaced with a special code (usually 9, 99, 999, etc.) to represent the invalid value and the missing value. Meanwhile, all variables and samples in the data set are retained. However, only samples with complete answers are used in the specific calculation, so the effective sample size will be different for different analyses due to different variables involved. This is a conservative approach that maximizes the information available in the data set.

4.2. Public Data and a First Optimal Credit Scoring Model

4.2.1. Defining the Target of Samples

4.2.1.1. Pre-Loan Approval Model

For the pre-loan approval process of Citizen loan products, it is necessary to evaluate the credit qualification of the applicant and predict the probability of overdue or bad debts in the future to decide whether to grant loans to the applicant. The pre-loan approval model is the main basis for pre-loan approval.

The pre-loan approval model is a 2-category model, which takes whether the applicant is overdue in the future as the target variable and the information submitted by the applicant and other information collected by the bank are as the forecast variable.

4.2.1.2. Defining the Target Variable Y in Pre-Loan Approval Model

By analyzing the rolling rate of the stock samples, the target variable is defined as:
1) For good customers, the maximum historical overdue days are less than 30 days, i.e. Y=0;
2) The largest overdue days in history greater than or equal to 30 days are bad customers, i.e. Y=1.

4.2.2. Data Samples

In order to make the pre-loan approval model have better prediction ability, the selection and preparation of data samples are particularly important. Sufficient and representative data samples are the premise of model establishment. The data sample selects 30774 customers of individual loan business who granted loans in 2018, including 1002 bad customers, accounting for 3.26%, and 29742 good customers, accounting for 96.74%.

4.2.3. Feature Engineering

According to the different data attributes and derivative methods of features, the features are roughly divided into 2 types. One is business features based on business logic, and the other is derived business features generated by using methods such as mathematical transformation, algorithm derivatives, features crossover and combination.

4.2.3.1. Business Features

In Chapter 1.2.1 of this paper, the preliminary research on business features are reviewed. The business features studied in this paper are derived from the data in the actual business scenarios, through which a large number of features reflecting the business characteristics can be constructed. Common business features are divided into 3 categories: basic attribute features, features based on various scene data and associated information features.

The basic attribute features are mainly the description of the inherent nature and characteristics of the research object, mainly involving identity information, education information, work information and other application form information. Through the analysis of such record information, the features that can be used for quantitative description or classification can be obtained.

For the features of various scene data, detailed data about the research object obtained from various scenes can be used for classification modeling.

The association information is mainly to establish the relationship between people through social data. With the help of Knowledge Graph, the association path depth, relationship type, relationship weight, relationship density, association node attribute and other indicators of the group or node are calculated and extracted, and the complex relationship network is visualized.

4.2.3.2. Feature Derivative

The mathematical transformation, algorithm derivative, feature cross and combination of business features can derive features with new meanings, which is more conducive to model calculation and thus improves the prediction ability of the model. According to the changes in the number of features before and after the derivative, the derivative methods are divided into 1-to-1 feature derivative, 1-to-N feature derivative and N-to-N feature derivative.

（1） 1-to-1 feature derivative

Function transformation of single variable: commonly used transformation functions include absolute value transformation, square transformation, logarithm transformation, index transformation and reciprocal transformation.

Box Division: mainly used for discretization of continuous variables and combination of multi-class value discrete variables. The features after discretization have strong robustness to abnormal data and are not vulnerable to the impact of extreme values, and can avoid the impact of meaningless fluctuations in the features on the model, which will make the model more stable. The box division method mainly includes the equidistance division and the equifrequency division, among which the equidistance Division: divides the value range of variables into k equal parts, each of which is a box; the equifrequency Division: divides the number of observed values of variables into k parts, so that each part contains approximately the same number of cases.

WOE conversion: WOE conversion is a supervised coding method in which the concentration attribute of the forecast category is used as the coding value. Generally speaking, it is a kind of mapping of the default probability when the feature is in a certain value range.

（2） 1-to-N feature derivative

1-to-N derivative method refers to the processing of a single feature to output a number of new features. The main methods include 2 types of One-Hot coding and mean value coding, both of which are used to process classified variables.

One-Hot coding: it is mainly applied to unordered classification variables, and cannot be directly used because the classifiers usually default such data to continuous and orderly variables for processing. Using one hot coding, the category can be 'binarized' and then used as a feature of model training.

Mean value coding: mean value coding is used to process high base amount of class features. When there are too many instance values of class features, One-Hot coding is easy to cause dimensional disaster, which reduces the effect of the model. Under the framework of Bayes, the mean value coding uses the target variables to be predicted to determine the most suitable coding method for this qualitative feature. The most important feature of this method is that it uses the known data to estimate the prior probability and the posterior probability based on the empirical Bayes method, and calculates the final characteristic code value by weighted average of the prior probability and the posterior probability.

（3） N-to-N Feature Derivative

N-to-N derivative method refers to the processing of multiple features to output multiple new features. The main methods include multinomial transformation and decision tree algorithm derivative features.

Transformation of multinomial: it is mainly to combine the multinomial features of the existing features to form a new feature matrix, such as, to X= (x1, x2) for 2 order transformation, the output result is: (x1, x2, x1^2, x1*x2, x2^2), which is often used in the linear model to achieve non-linear effect.

Derivative features of the decision tree algorithm: in a series of decision tree algorithms, each sample will fall on a leaf node, and the leaf node will be used as a new feature in the training model. The tree model itself does not generate features, but it can generate feature combinations by using the characteristics of its algorithm. To a certain extent, the algorithm makes up for the time-consuming and laborious defect of artificial combination features.

4.2.3.3. Feature Selection

For the above business features, when assessing whether a feature is included in the model variables, the following factors should be considered comprehensively and ranked by priority:

（1） Logical and explanatory;
（2） Strong prediction ability;
（3） Low correlation with other variables;
（4） Stable and accessible;
（5） Compliance without legal or ethical restrictions;
（6） Relevant to the applicant and not a strategy of the financial institution;
（7） The information loss will be large after removal.

For 30774 modeling samples, the local data has a total of 5006 fields. How to select the appropriate features for the model is very key.

The feature selection methods adopted in this paper are as follows:

（1）The variable with too-high loss rate is directly eliminated, which is based on the threshold of 70%;

（2） All variables with values close to constant in numerical variables are eliminated;
（3） Variables that cannot be explained by business logic are directly eliminated;
（4） The direct elimination of the unique value greater than 20 (excluding 20) in the classification variable;
（5） Variables with IV value less than 0.02 are directly eliminated;
（6） If the correlation between the 2 variables is greater than 0.7, the variable with smaller IV value is removed.

It is necessary to explain the IV value here. The IV value measures the amount of information of a variable, and its value reflect the impact of the variable on the target variable. In general, the prediction ability of variables with IV value less than 0.02 is very poor, and it may affect the stability of the model. Therefore, independent variables with IV value less than 0.02 will be directly eliminated.

After the screening of the above process, the remaining feature fields are 284, among which the more representative features are:

（1） Age

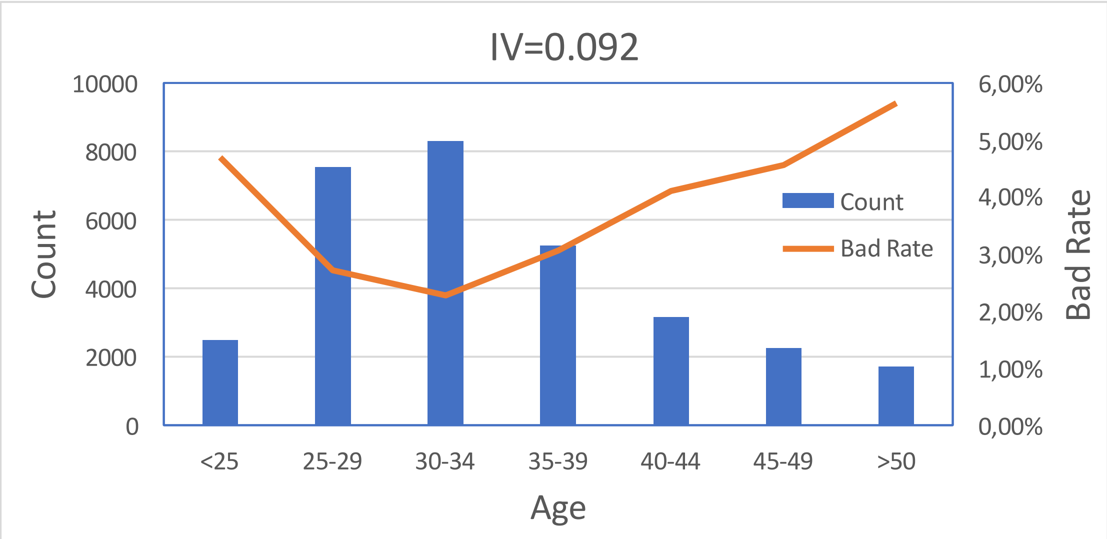

Figure 2: IV Value of Age

The age feature is derived from the customer application form. For the analysis of its IV value, the age is first discretized and divided into 7 groups as shown in the figure above, and the number of samples in each box and the bad debt rate are calculated. As can be seen from the figure, the main age range of Citizen loan customers is 30-34 years old, and the bad debt rate of these customers is the lowest, which is 2.28%. As a whole, with the increase of age, the bad debt rate shows a ‘V’ shape of first reduction and then increase. After calculation, the IV value of age is 0.092.

（2） Number of applications in the last 4 months

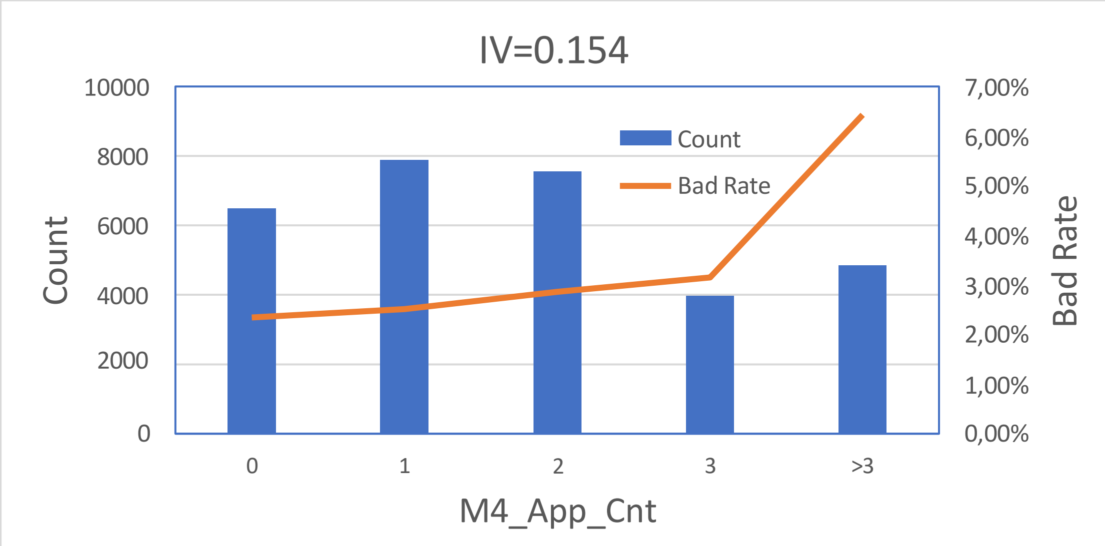

Figure 3: The Number of Applications in the Past Four Months IV Value

‘Number of applications in the past 4 months’ is a feature of Internet big data category, which describes the number of times that applicants apply for other loans in the past 4 months. As can be seen from the figure, with the increase of ‘number of applications in recent 4 months’, the bad debt rate gradually increased, especially if the ‘number of applications in recent 4 months’ of the applicant is more than 3, its bad debt rate is much higher than that of other applicants. The IV value of ‘number of applications in recent 4 months’ is 0.154.

（3） Monthly average balance of deposits

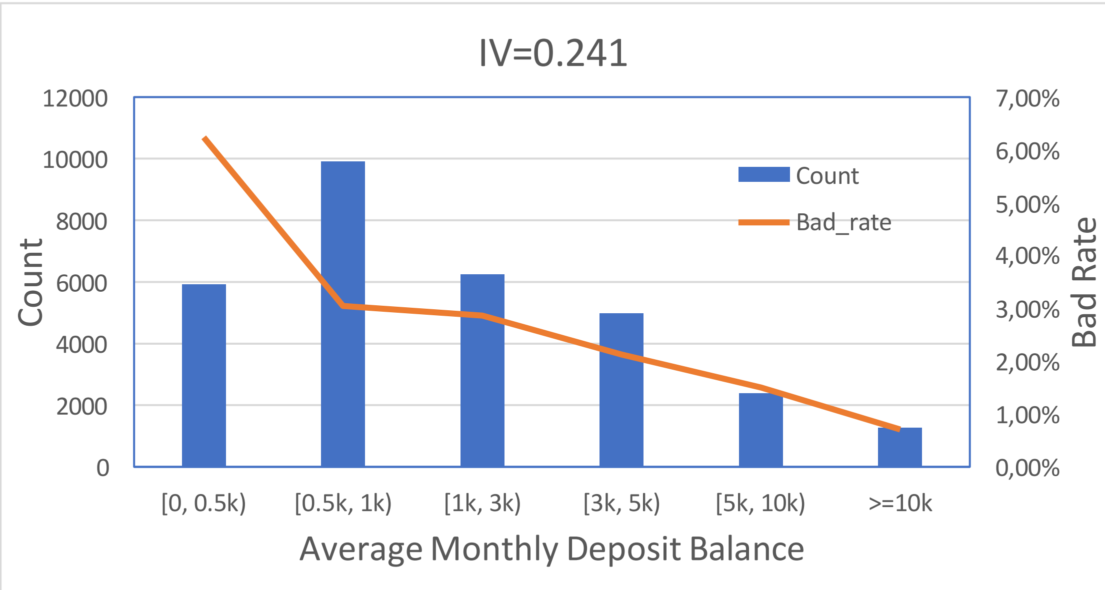

Figure 4: Average Monthly Deposit Balance IV Value

‘Monthly average balance of deposits’ is derived from the internal data in the BTB. First, we do the discretization and box separation, as shown in the figure below, and divide it into 6 groups, and calculate the number of samples and bad debt rate in each box respectively. As can be seen from the figure, customers with ‘monthly average balance of deposits’ less than 0.5K have the highest bad debt rate. With the increase of ‘monthly average balance of deposits’, the bad debt rate gradually decreases. After calculation, the IV value of "monthly average balance of deposits" is 0.241.

4.2.4. Establishing and Evaluating the Public Model

4.2.4.1. XGBoost Model

XGBoost (eXtreme Gradient Boosting) algorithm is an efficient version of Gradient Boosting algorithms. It has been widely praised by the industry due to its excellent effect and efficiency in practice. In 2014, Dr. Tianqi Chen proposed the XGBoost algorithm. XGBoost can be considered as a further optimization based on the GBDT algorithm. First of all, the XGBoost algorithm introduces a regular term into the loss function of the basic learning device to control the reduction of over fitting in the training process; secondly, the XGBoost algorithm not only uses the first derivative to calculate the false residual, but also calculates the 2 derivatives to approximate the rapid pruning to build a new basic learning device; in addition, the XGBoost algorithm also does a lot of engineering optimization, such as supporting parallel computing and improving the computing efficiency, processing sparse training data and so on. The XGBoost algorithm originated from the Boosting integrated learning method, and integrated the advantages of the Bagging integrated learning method into the evolution process. The customized loss function of Boosting framework improves the ability of the algorithm to solve general problems. At the same time, more controllable parameters can be introduced to optimize the problem scenarios. Finally, the detailed optimization of the engineering implementation can ensure the stability of the algorithm results, and at the same time, it can efficiently process large-scale data. It can be expanded to support different programming languages.

These factors together make it one of the mainstream machine learning algorithms.

4.2.4.2. Establishing the Public Model

Before the model is established, the sample data set needs to be randomly divided into 2 parts: 70% of the sample is used as the training set to train the model; 30% of the sample is used as the test set to measure and evaluate the trained model. In particular, in the sample segmentation, we should ensure that the bad debt rate in the 2 data is basically consistent.

The following is the percentage and details of good customers and bad customers in the training set test set:

Sample Category	Customer Category	Amount	Percentage
Training Set Sample Size: 21520 Cases	Bad Customer	709	3.29%
	Good Customer	20,811	96.71%
Sample Number of Test Set: 9224 Cases	Bad Customer	293	3.18%
	Good Customer	8,931	96.82%

Table 14: Proportion and Details of Good and Bad Customers in Training Set and Test Set

In the process of model training, the method of 5-fold cross validation is adopted for the super parameter of XGBoost algorithms such as Num_Round, Max_Depth, etc. conducted grid search and parameter adjustment, and selected a group of parameters with the largest ROC as the final parameters of the model to complete the model training.

Variable	Data source	Relative Significance
Average Monthly Deposit Balance	Internal Data	5.6%
Accumulated Overdue Times In the Last 6 Months	Data of External Financial Institutions (Credit Reference in PBC)	5.3%
Number Of Applications In the Last 4 Months	Data of External Financial Institutions (Credit Reference in PBC)	4.5%
High Debt Service Pressure Index	Internet Big Data	3.9%
Age	Application Form	3.7%
Account Status Is Overdue	Data of External Financial Institutions (Credit Reference in PBC)	3.1%
Many Times of Rejection In Non-Bank Institutes	(Data of External Financial Institutions (Credit Reference in PBC)	2.8%
Number of Night Applications In the Past 3 Months	Internet Big Data	2.6%
Financial Balance	Internal Data	2.5%
Customer Rating	Internal Data	2.3%

Table 15: Top 10 Key Variables in the Model, as well as Data Sources and Relative Importance of Variables

The table gives the top 10 key variables in the model, as well as data sources and relative importance of variables.

4.2.4.3. Evaluating the Public Model

Model evaluation is used to evaluate the quality of the model. Generally, after completing the model training on the training set, the model is used to predict the samples in the test set, and evaluate the quality of the model by calculating the accuracy of the model prediction. For the 2-classification model, the commonly used evaluation indicators are KS and AUC.

On the test set, the probabilities predicted by the model are arranged from small to large, and the different (FPR, TPR) values are calculated with these probabilities as the threshold values, and then the ROC curve is drawn with FPR as the horizontal coordinate and TPR as the vertical coordinate. KS is the maximum value of the difference between FPR and TPR in the 2-value group (TPR, FPR); AUC is the area under the ROC curve, which can quantitatively reflect the model performance measured based on the ROC curve. The larger the KS and AUC, the more likely the classifiers are to rank the real positive samples in the first place, and the better the classification performance.

After calculation, the Public Model on the test set, KS=0.3637, AUC=0.7317.

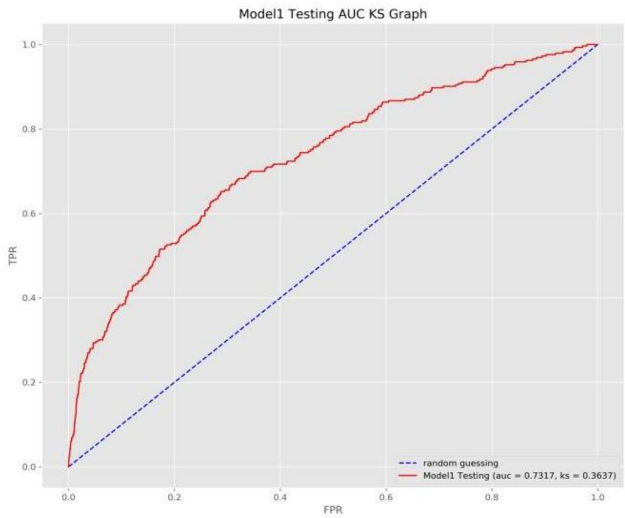

Figure 5: Public Model on Test Set, KS, AUC Figure

For the samples in the test set, they are divided into 10 groups according to the model prediction probability from small to large, and each group contains 10% of the samples. Then the number of users in each group, the number of good users and the number of bad users are counted, and the bad debt rate of each group is calculated, as shown in the table below. From the table, it can be seen that from the first group to the tenth group, the bad debt rate has a monotonic increasing trend. And the first group of customers with the best quality selected by the model has a bad debt rate of 0.5%, which is far lower than the overall bad debt rate of 3.2% of the test set; while the tenth group with the worst quality selected by the model has a bad debt rate of 11.5%, which is far higher than the overall bad debt rate of 3.2%, indicating that the model has a better ability to distinguish good from bad users.

In addition, the table also calculates the cumulative indicators, including the cumulative number of users, the cumulative number of good users, the cumulative number of bad users, the cumulative bad debt rate, the cumulative proportion of good users and the cumulative proportion of bad users, which can reflect the results after the use of the model. If the passing rate of the model is required to be 80%, the user risks predicted by the model need to be ranked from low to high. 80% of the users with lower risks are approved and the remaining 20% are rejected. As can be seen from the table, the bad debt rate of the pass through part is 1.9%, which is 1.3% lower than the overall bad debt rate of 3.2% without using the model. Among the 20% rejected users, 1-81.1%=18.9% of the good users were killed by mistake and the more gain that 1-47.4%=52.6% of the bad users were rejected was got.

Decile	Population	Current	Overdue	Bad_ Rate	Cum_ Pop	Cum_ Current	Cum_ Overdue	Cum_ Bad_ Rate	Cum_ Current_ Rate	Cum_ Overdue_ Rate
1	923	918	5	0.5%	923	918	5	0.5%	10.3%	1.7%
2	922	912	10	1.1%	1845	1830	15	0.8%	20.5%	5.1%
3	922	910	12	1.3%	2767	2740	27	1.0%	30.7%	9.2%
4	923	911	12	1.3%	3690	3651	39	1.1%	40.9%	13.3%
5	922	906	16	1.7%	4612	4557	55	1.2%	51.0%	18.8%
6	922	902	20	2.2%	5534	5459	75	1.4%	61.1%	25.6%
7	923	896	27	2.9%	6457	6355	102	1.6%	71.2%	34.8%
8	922	885	37	4.0%	7379	7240	139	1.9%	81.1%	47.4%
9	922	874	48	5.2%	8301	8114	187	2.3%	90.9%	63.8%
10	923	817	106	11.5%	9224	8931	293	3.2%	100.0%	100.0%

Table 16:User Distinguishing Ability Analysis Table of Public Model on Test Set

4.3. Local Data and a Second More Comprehensive Model

4.3.1. Local Data

4.3.1.1. Local Features

Local data includes local and external big data, local scene big data and farmer big data. Among them, the local external big data is the detailed telecommunication record of the users and the data of the closeness level between the users obtained by BTB and the local telecommunication department through cooperation and authorization; the local scene big data mainly comes from the local payment platforms of Baotou (e-commerce POS flow), which is the data of the offline consumption expenditure of the users; the farmer big data mainly comes from the survey of local farmers by BTB. Data obtained, including income, expenditure, assets, liabilities, social relations, etc. of farmers.

For 30774 modeling samples, the local data has a total of 34179 features. After data cleaning, indicator processing, feature deriving and feature screening, there are 147 effective features.

The following table shows some examples of these features:

	Feature name	Data Sources	Relative Significance
1	Family Income	Field Investigation	4.90%
2	Average Consumption Amount In the Past 6 Months	Field Investigation	3.60%
……			
11	Housing Area	Field Investigation	3.32%
……			
15	Average Value of Transaction of e-Commerce Consumption	Local e-Commerce POS System Data	3.11%
……			
23	Whether or Not Support College Students	Field Investigation	2.84%
……			
54	Number of Livestock	Field Investigation	2.44%
……			
73	Present Value of Homestead	Village Committee Records	1.88%
……			
88	Average Amount of Communication Bill In the Past 3 Months	Carrier Data	1.72%
……			
95	Present Value of Agricultural Machinery	Field Investigation	1.32%
……			
111	Family Relationship	Back to Back Comment	0.98%
……			
123	Neighborhood Relationship	Back to Back Comment	0.72%
……			
147	Insured Amount	Insurance Company Data	0.38%

Table 17: Examples of 147 Local Features

‘Family income’ is a feature of farmer big data, which describes the total income of the applicant's family in the past 1 year. As can be seen from the figure, with the increase of ‘family income’, the bad debt rate gradually decreases, especially for applicants with ‘family income’ greater than 100 thousand, the bad debt rate is far lower than other applicants. The IV value of ‘family income’ is 0.317.

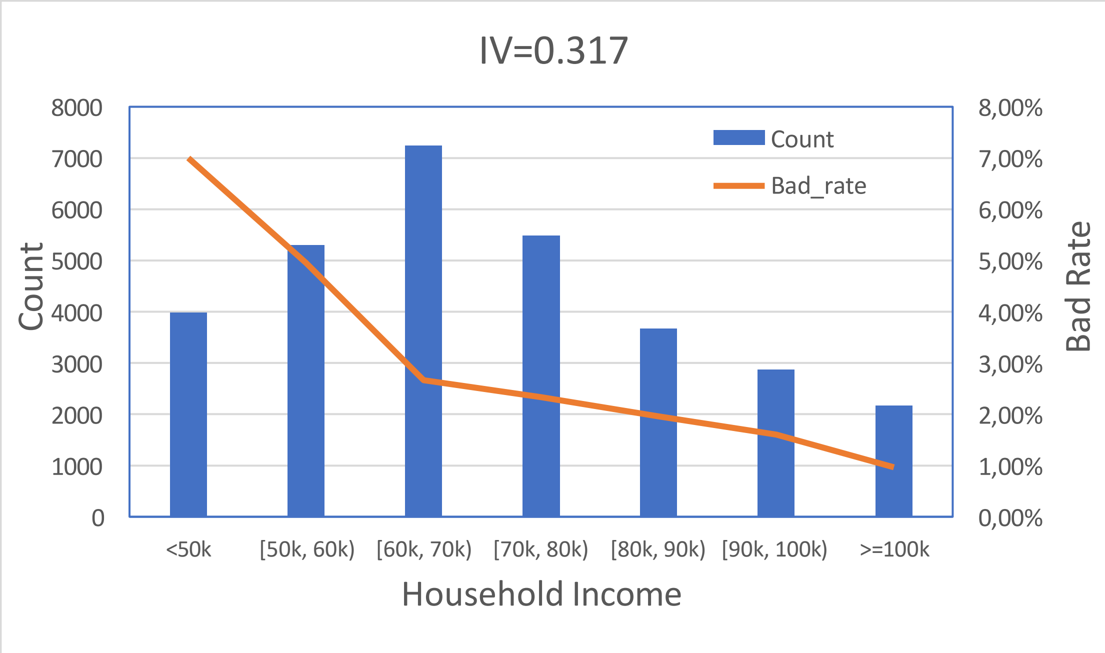

Figure 6: Family Income IV Value

It should be noted that there are some features that have no value, as shown in the following table:

	Feature Name	Data Sources	IV Value
1	Total Household Population	Field Investigation	0.019
2	Difficulty of Selling Houses	Field Investigation	0.017
3	Position of Village Committee	Village Committee Records	0.016
4	Type of Agricultural Machinery	Field Investigation	0.014
5	Time of Mobile Internet Access	Carrier Data	0.013
6	Number of Products Held by Households	Statistical Data	0.009
7	Frequency of Mobile Phone Down in Recent 6 Months	Carrier Data	0.006
8	Premium Amount	Insurance Company Data	0.005
9	Difficulty of Selling Agricultural Machinery	Field Investigation	0.003
10	Present Value of Vehicle	Field Investigation	0.003

Table 18: 10 Worthless Features

4.3.1.2. Associating Features in Knowledge Graph

Knowledge Graph is essentially a kind of knowledge base called semantic network, which is a knowledge base with a directed graph structure. The nodes of the graph represent entities or concepts, and the edges of the graph represent various kinds of semantic relationships between entities and concepts, such as the similar relationship between 2 entities.

At present, with the continuous development of the application of intelligent information services, Knowledge Graph has been widely used in the fields of intelligent search, personalized recommendation and financial big data. Knowledge Graph technology provides a way to extract structured knowledge from massive text and images, which is the key to big data analysis. The construction of Knowledge Graph mainly includes 3 parts: knowledge acquisition, data fusion and knowledge calculation and application.

（1）Knowledge acquisition: there are a large number of structured and unstructured data in the local data. In the process of unstructured data, the first step is to extract the text from the user's unstructured data, in which information is extracted mainly through Natural Language Processing (NLP) technology, and specific types of information are extracted from the unstructured or semi-structured text to effectively identify and disambiguate the unstructured information.

（2）Knowledge fusion: when the knowledge is obtained from each data source, it is necessary to provide a unified term to integrate the knowledge obtained from each data source into a huge knowledge base.

（3）Knowledge Computing and application: Knowledge Computing is mainly to obtain more implied knowledge according to the information provided by the graph; link prediction can predict the implied relationship between entities; meanwhile, different algorithms of community computing are used to calculate the communities existing in the knowledge graph on the knowledge network to provide the path of knowledge association; noises and defects in the data are detected through the inconsistency detection technology.

Due to the large amount of social relationship information of farmers in the big data of local farmers, in order to excavate effective features and improve the effect of the pre-loan approval model, a Knowledge Graph of local farmers has been established, as shown in the following figure.

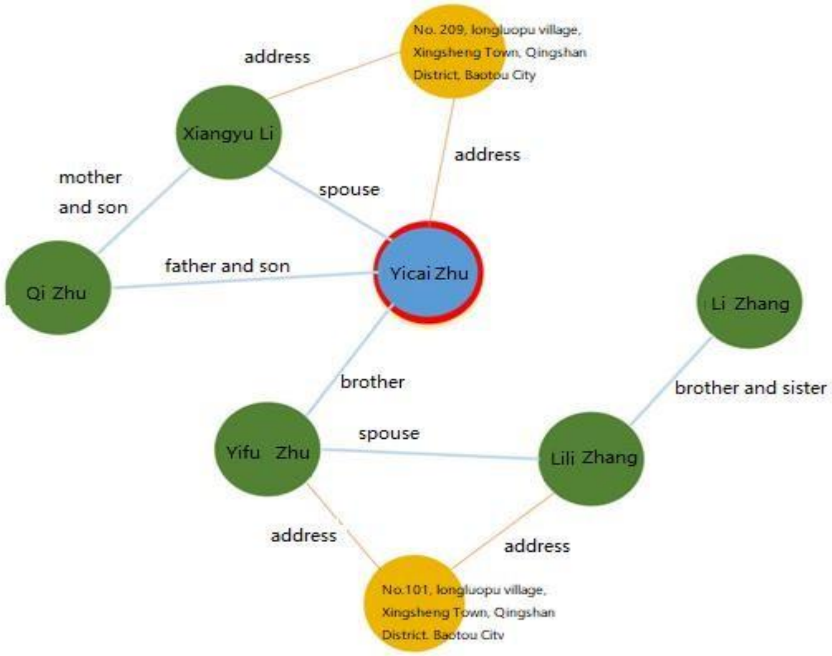

Figure 7: Knowledge Graph of Local Farmers

In the Knowledge Graph, this paper mainly takes individuals and addresses as nodes, and the relationship between people as edges. In the individual nodes, there are also attributes such as age, marital status, assets, income, liabilities, etc.

Using Knowledge Graph, the first step is to verify the information, such as whether the husband and wife live in the same address, whether the two brothers have the same father or mother, to improve the authenticity of the information collected. Secondly, some features are extracted from the Knowledge Graph and used to establish the Local model, such as the total assets of the first-related parties, the total liabilities of the first-related parties, the total borrowing times of second-degree-related parties and other related features.

The following table shows some examples of these features:

	Feature name	**Data Sources**	**Relative Significance**
1	Net Family Income of Immediate Family Members	Field Investigation + Knowledge Graph	5.40%
2	Total Borrowing Times of Second-Degree Related Parties	Credit Data + Knowledge Graph	1.90%
……			
6	Total Assets of the First Related Parties	Field Investigation + Knowledge Graph	0.93%
7	Average Value of Transaction of e-Commerce Consumption of First Related Parties	Local e-Commerce POS System Data + Knowledge Graph	0.71%
……			
11	The Average Communication Bill of the Second Related Parties in the Past 3 Months	Communication Data + Knowledge Graph	0.68%
……			
13	Total Liabilities of the First Related Parties	Credit Data + Knowledge Graph	0.55%
14	Second Related Parties' Neighborhood Relationship	Back to Back Comment + Knowledge Graph	0.42%

Table 19: 14 Effective Features After Knowledge Graph Extraction

'Net income of immediate family members' is an effective feature extracted from the Knowledge Graph. First, we do the discretization and box separation, as shown in the figure below, and divide it into 6 groups, and calculate the number of samples and bad debt rate in each box respectively. As can be seen from the figure, the users with 'net income of immediate family' less than 10k have the highest bad debt rate. With the increase of 'net income of immediate family', the bad debt rate gradually decreases. After calculation, the IV value of "net income of immediate family" is 0.342.

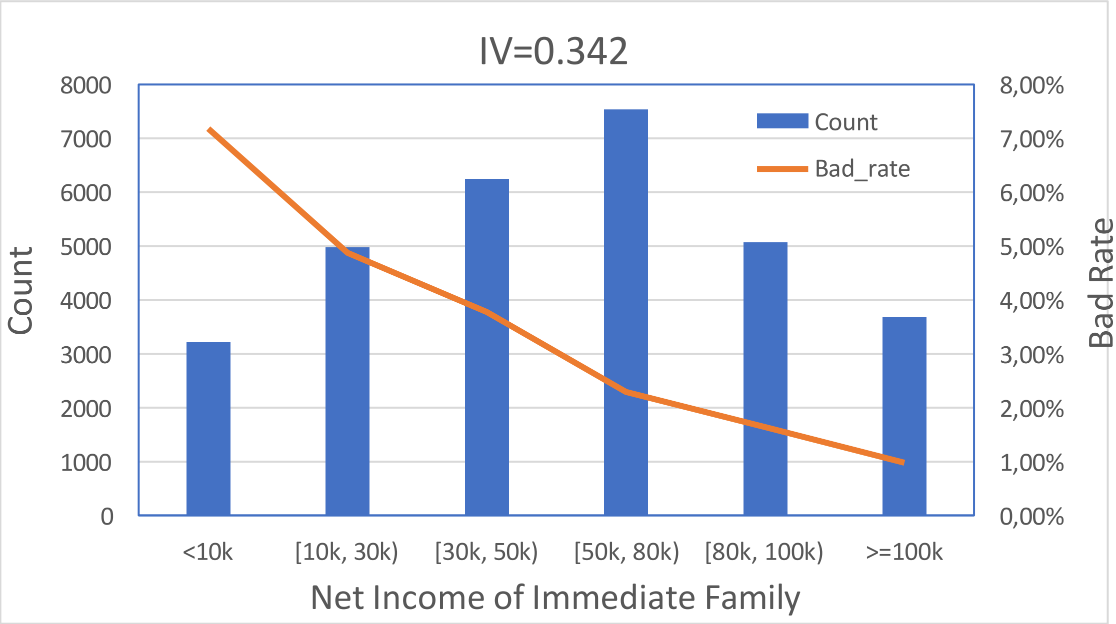

Figure 8: Family net income IV of immediate family members

4.3.2. Establishing and Evaluating the Local model

4.3.2.1. Establishing the Local Model

The Local model also adopts the XGBoost algorithm. The training set is consistent with the training samples of the Public model, with 21,520 cases. However, the model adopts features, including 284 features of Public data screening, 147 features of local data, and 14 features derived from the Knowledge Graph, with a total of 445 features.

In the process of model training, the method of 5-fold cross validation is adopted for the super parameter of XGBoost algorithms such as Num_Round, Max_Depth, etc. conducted grid search and parameter adjustment, and selected a group of parameters with the largest ROC as the final parameters of the model to complete the model training.

Variables	Data Source	Relative Significance
Net Family Income of Immediate Family Members	Knowledge Graph (Local Data)	5.4%
Family Income	Local Data	4.9%
Monthly Average Balance of Deposits	Public Data	4.7%
Accumulated Overdue Times in the Past 6 Months	Public Data	4.1%
Average Consumption Amount in the Past 6 Months	Local Data	3.6%
Number of Applications in the Past 4 Months	Public Data	2.9%
High Debt Service Pressure Index	Public Data	2.5%
Age	Application Form	2.2%
Overdue Account Status	Public Data	2.1%
Total Borrowing Times of Second-Degree Related Parties	Knowledge Graph (Local Data)	1.9%

Table 20: Top 10 Key Variables in the Local Model

The table above lists the top 10 key variables in the Local Model, as well as the data source and relative importance of the variables. Among the 10 key features, 4 are from local data and the rest are from Public data, which is also a key feature of the Public model.

4.3.2.2. Evaluating the Local Model

The Local Model was evaluated in the same test set, and its ROC curve is shown in the figure below, with KS=0.5310 and AUC=0.840 for the model. Compared with the KS and AUC indicators of Public model, the effect of the Local model is much better than that of the Public model.

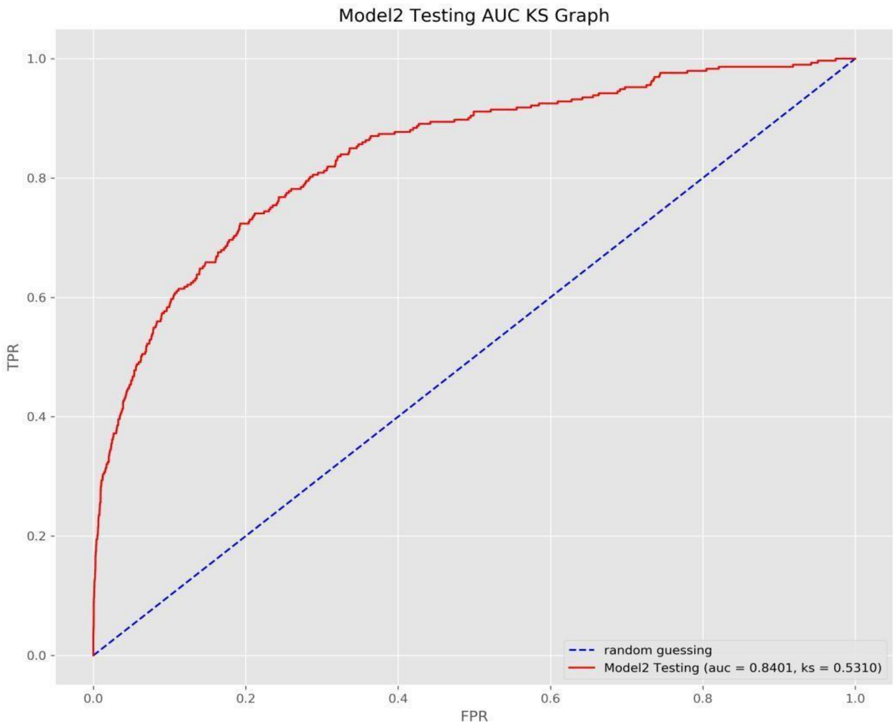

Figure 9: Local Model on Test Set, KS, AUC Figure

When the 14 variables derived from the knowledge graph are not considered, the Local-14 model is established and evaluated in the same test set, whose ROC curve of the model is shown in the figure below, with KS = 0.4917 and AUC = 0.8027. Compared with the KS and AUC indexes of public model and local model, the effect of Local-14 model is between that of local model and public model, which is still better than that of the Public Model.

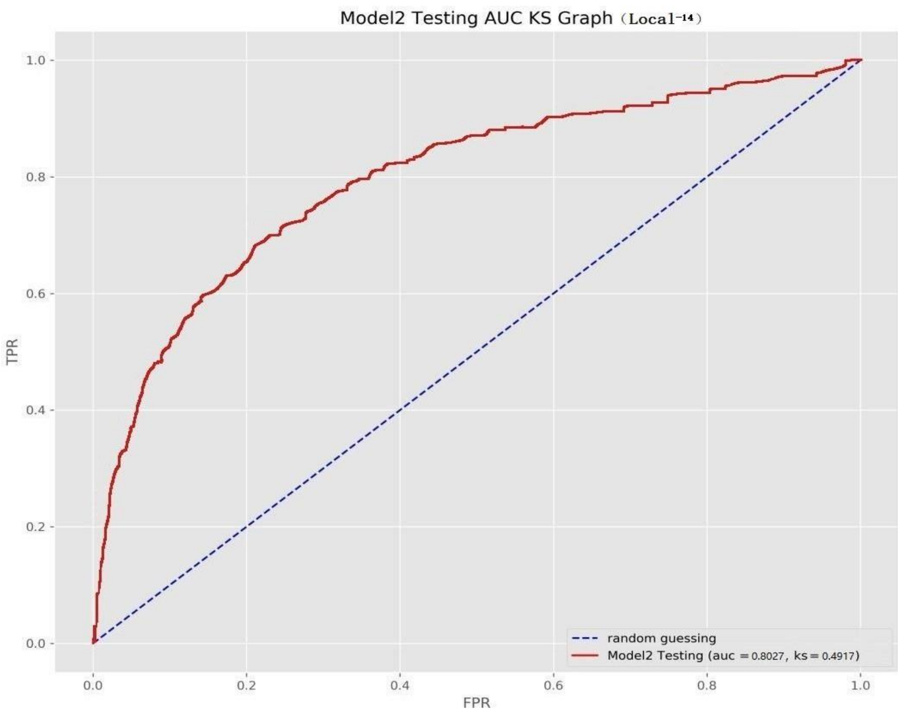

Figure 10: Local-14 Model on Test Set, KS, AUC Diagram

	Public Model	Local-14 Model	Local Model
KS	0.3637	0.4917 Compared with the public model, it increased by 35.19%	0.5510 Compared with the Public Model, it increased by 51.5%. Compared with Local-14 Model, it increased by 12.06%.
AUC	0.7317	0.8027 Compared with the public model, it increased by 10%	0.8401 Compared with the Public Model, it increased by 14.8%. Compared with Local-14 Model, it increased by 4.7%.

Table 21:Public, Local-14, Local Model KS, AUC Analysis Table

Therefore, the Local-14 model is no longer considered separately in practice and the following analysis.

For the samples in the test set, according to the probability predicted by the Local model from small to large, they are divided into 10 groups with equal frequency, each group contains 10% of the samples. Then, the number of users, the number of good users, the number of bad users, the bad debt ratio, the cumulative number of users, the cumulative number of good users, the cumulative number of bad users, the cumulative bad debt ratio, the cumulative proportion of good users and the cumulative proportion of bad users in each group are calculated respectively, as shown in the table below. From the table, it can be seen that from the first group to the tenth group, the bad debt rate has a monotonic increasing trend. And the first group of customers with the highest quality screened by the Local model only has a bad debt rate of 0.1%, while the tenth group with the worst screened by the Local model has a bad debt rate of 17.7%, which indicates that the Local model has a better ability to distinguish good from bad users, and its ability to distinguish good from bad is better than that of the Public model.

If the passing rate of the model is required to be 80%, approval is required for the 80% with lower risk predicted by the model, and the remaining 20% is rejected. As can be seen from the table, the bad debt rate of the pass through part is 1.1%, which is 2.1% lower than the overall bad debt rate of 3.2% without using the model. Among the 20% rejected users, 1-81.7%=18.3% of the good users were killed by mistake and the more gain that1-28.7%=71.3% of the bad users were rejected was got.

Decile	Population	Current	Overdue	Bad_ Rate	Cum_ Pop	Cum_ Current	Cum_ Overdue	Cum_ Bad_ Rate	Cum_ Current_ Rate	Cum_ Overdue_ Rate
1	923	922	1	0.1%	923	922	1	0.1%	10.3%	0.3%
2	922	919	3	0.3%	1845	1841	4	0.2%	20.6%	1.4%
3	922	916	6	0.7%	2767	2757	10	0.4%	30.9%	3.4%
4	923	915	8	0.9%	3690	3672	18	0.5%	41.1%	6.1%
5	922	912	10	1.1%	4612	4584	28	0.6%	51.3%	9.6%
6	922	910	12	1.3%	5534	5494	40	0.7%	61.5%	13.7%
7	923	904	19	2.1%	6457	6398	59	0.9%	71.6%	20.1%
8	922	897	25	2.7%	7379	7295	84	1.1%	81.7%	28.7%
9	922	876	46	5.0%	8301	8171	130	1.6%	91.5%	44.4%
10	923	760	163	17.7%	9224	8931	293	3.2%	100.0%	100.0%

Table 22: User Distinguishing Ability Analysis Table of Local Model on Test Set

4.4. The Incremental Value of Local Information

4.4.1. Pass Rate and Bad Debt Rate Analysis

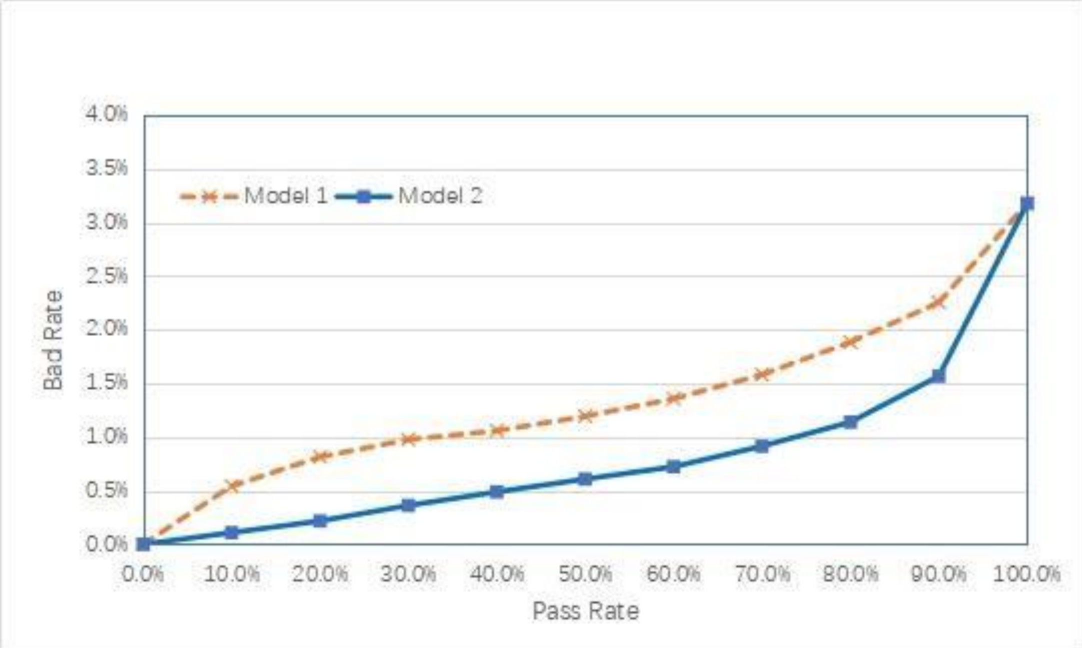

Figure 11: Pass Rate and Bad Debt Rate Analysis Chart

The above figure compares the effect of Public model and Local model with passing rates as abscissa and bad debt rate as ordinates. As can be seen from the figure, the Local Model is located at the bottom of the Public Model, that is, under the same pass rate, the bad debt rate of the Local model is lower than that of the Public Model; under the same bad debt rate, the pass rate of the Local Model is higher than that of the Public Model. If the bad debt rate is controlled within 1%, the pass rate of the Local Model can reach 75%, while that of the Public Model can only reach 30%; if the pass rate is controlled at 80%, the bad debt rate of Local model is 1.1%, and that of Public model is 1.9%.

4.4.2. Lift Analysis

Lift is a measure to evaluate whether a model is effective or not. It measures the multiple of a model's predictive ability over random selection. Take 1 as the boundary, and lift greater than 1 means that the model captures more bad users than random selection. A lift equal to 1 indicates that the model's performance is independent of random selection, and less than 1 means that the model captures fewer bad users than random selection.

Usually, according to the prediction value of the model, the test samples are sorted from low to high, equally divided into 10 equal parts, and then the proportion of each group of bad samples in the overall bad samples is calculated.

In the figure below, the analysis is carried out according to three sorting methods: 1) random sorting; 2) sorting according to the prediction value of the Public Model; 3) sorting according to the prediction value of the Local Model.

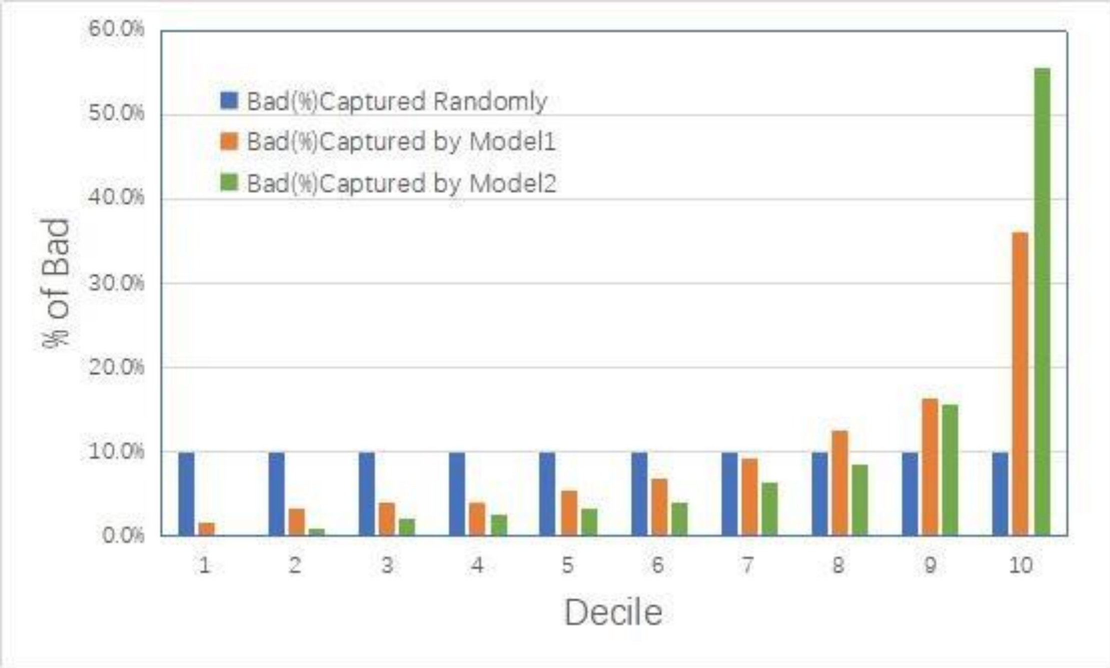

Figure 12: Lift Analysis Chart

According to the cumulative calculation step by step from bad to good, take its ratio with the cumulative calculation of random groups to get the lift degree, that is, how many times of the score card's ability to capture bad customers are randomly selected. The lift of the Public Model and local model is shown in the following table, and the cumulative lifting diagram is shown in the figure below.

Decile	Lift_Model1	Lift_Model2
1	1.00	1.00
2	1.09	1.11
3	1.19	1.23
4	1.30	1.38
5	1.44	1.56
6	1.62	1.81
7	1.86	2.16
8	2.17	2.66
9	2.63	3.57
10	3.62	5.56

Table 23:Public Model and Local Model Lift Table

The cumulative lifting graph can intuitively compare the gain degree of discrimination ability brought by different models. As can be seen from the figure, the gain of the Local Model is better than that of the Public Model.

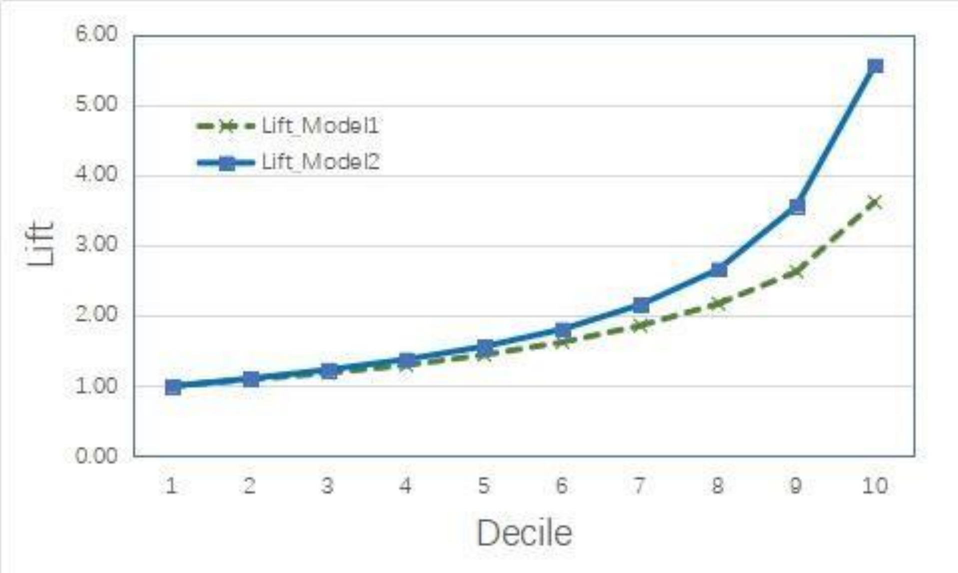

Figure 13: Comparative Analysis of Cumulative Improvement Between the Public Model and the Local Model

4.5. Other Competitive Advantages

Another important application area of the model is in the selection and establishment of marketing whitelist.

The telecom corporations in Baotou have a large number of data (more than 2 million customers) and have more than 2000 tags to identify these customers. After negotiation, BTB can deploy a new marketing model based on the Local Model at telecom corporations, model 2 million customers and 2000 tags, filter target customers (with low risk and interested in products) in advance, establish marketing whitelist, and then send marketing information through SMS to obtain customers.

There are two limitations in the application of the model: 1) the original data cannot be obtained from the corporations, only the tags can be obtained; 2) the PBC credit reference cannot be obtained without the customer's authorization. Therefore, after the original the Local Model has been adjusted to adapt to these two factors, a new pre-credit model is generated and deployed to corporations to generate marketing whitelist.

The advantage of this marketing method is that it can reach the target customers accurately and efficiently, push the loan marketing information to the people who really need it, reduce the disturbance to invalid customers (high risk and not interested in products), and can effectively improve the success rate of marketing.

Items	Results
Number of Original Customers	2,371,456
Number of Customers Filtered by the Marketing Model	128,475
SMS Amount (20 SMSs per Customer)	2,569,500
The Number of Customers Touched Who Begin to Apply	8,264
The Application Success Rate	94.23%

Table 24:Marketing Model Operation Analysis Table (2018)

The detail process of establishing the marketing model is shown in the figure below:

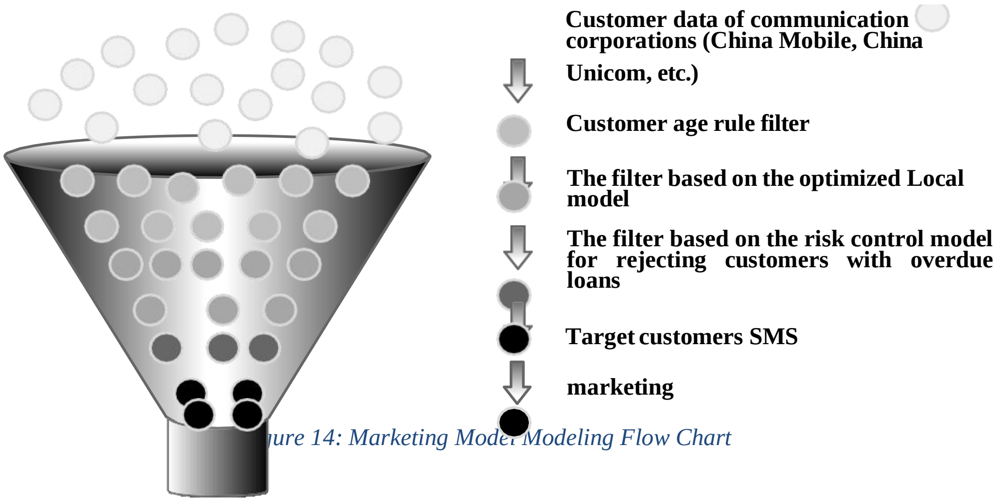

Figure 14: Marketing Model Modeling Flow Chart

4.6. The Application and Operation Information of Local Model and Financial Analysis in 2019

4.6.1. The Basic Operation Information of New Product in 2019

In 2019, based on the local model, Baotou rural commercial bank established the Citizen loan products, the specific forms of which are as follows:

Product Elements	Description
Product Name	Citizen loan
Customer	Citizens, farmers, small- and micro- business owners, etc.
Credit Line	2000~1000000
Annual Interest Rate	6% - 18%; average interest rate 8.72%
Loan Term (Month)	12，24，36
Repayment Method	Equal installment payment, Before interest after principal payment
Guarantee Method	Credit, Mortgage
Average Credit Line	43,728

Table 25: Elements of Citizen Loan Products Based on Local Model (2019)

Based on such products, the loan scale and efficiency of the whole year in 2019 have been greatly improved, and the cost has been well controlled. In 2019, the monthly new loans will increase month by month, and the increasing trend will increase month by month. The total new loans in 2019 will be 1.535 billion.

Item	1	2	3	4	5	6	7	8	9	10	11	12
New Loan	3855	4910	6360	8146	10491	12946	15506	17083	20171	23354	26809	30597
Accumulative Amount	3855	7710	12620	18980	27126	37617	50563	66069	83152	103323	126677	153486

Table 26: Accumulative Amount of Citizen Loan in 2019 (From January to December)

4.6.2. Comparative Analysis of Bad Debt in 2019 and Bad Debt in 2018

After the introduction of the Local model in 2019, after the loan scale gradually increased, the non-performing loan rate showed a downward trend in 2019 due to the effect of external filtering and risk control strategies of the model. Moreover, due to the uniform risk control standards of various sub-branches, the non-performing loan distribution of sub-branches in each quarter was relatively even. Through the monitoring of the model, it is found that the risk can be adjusted quickly and uniformly, and the adverse situation of each sub-branch is effectively controlled.

Quater	A Sub-Branch	B Sub-Branch	C Sub-Branch	D Sub-Branch	E Sub-Branch	F Sub-Branch
2019.1	1.25%	0.83%	1.52%	1.28%	0.92%	1.68%
2019.2	1.28%	0.75%	1.54%	1.25%	0.94%	1.72%
2019.3	1.24%	0.78%	1.42%	1.13%	0.88%	1.66%
2019.4	1.23%	0.72%	1.45%	1.20%	0.84%	1.65%

Table 27: More Balanced Bad Debt Performance of Sub-Branches

Sub-Branch	2018	2019
A	2.85%	1.23%
B	1.37%	0.72%
C	2.54%	1.45%
D	1.55%	1.20%
E	1.77%	0.84%
F	3.55%	1.65%

Table 28: The Overall Decline of Bad Debt in Sub-Branches

Non-performing loans of BTB as a whole decreased (62.8% in 2019 compared with that in 2018)

	2018	2019
BTB	3.20%	1.19%

Table 29: Overall Decline of Non-Performing Loans in BTB (Down 62.8% in 2019 Compared With That in 2018)

4.6.3. Financial Performance of New Products

Taking a loan of 100,000 yuan as an example, the cost of credit and guaranteed loans decreased from 470 yuan to 319.42 yuan, and the single cost decreased by 150.58 yuan; the mortgage loan decreased from 1,650 yuan to 425.42 yuan, and the single cost decreased by 1,224.58 yuan. It is particularly worth mentioning here that due to the introduction of online evaluation, the evaluation cost has been reduced from 1,000 yuan of the original manual evaluation to 26 yuan of the evaluation data only.

The new loan of 1.535 billion yuan in 2019 will save 676.7 + 147.93 = 8.2463 million yuan through the new mode, and the loan cost will be reduced by 60.04%.

Cost Comparison of Two Credit and Guaranteed Loan Models (100,000 Yuan as an Example)			
Traditional Model (Yuan)		**After Introducing Big Data and Local Model (Yuan)**	
Manual Query	150	Big data query (data source 1 + data source 2) - specifically to obtain a valid customer	14.20
Notarization	300	Notarization	300
Information Fee	20	SMS service	0.15 yuan per time
		CFCA	2.7 yuan per time
		OCR identification of bank card	0.06 yuan per time
		Face recognition	0.45 yuan per time
		OCR identification of ID card	0.06 yuan per time
		Checking consistence between face and ID card	1.8 yuan per time
Total	470	Total	319.42

Table 30: Cost Comparison of Two Credit and Guaranteed Loan Models (RMB 100,000 as an Example)

Cost Comparison of Two Mortgage Models (100,000 Yuan as an Example)

Traditional Model (Yuan)		After Introducing Big Data and Local Model (Yuan)	
Manual Query	250	Big data query (data source 1 + data source 2) - specifically to obtain a valid customer	14.20
Notarization	300	Notarization	300
Assessment Fee	1000	Assessment fee	26
Charge for Handling Mortgage	80	Charge for handling mortgage	80
Information	20	SMS service	0.15 yuan per time
		CFCA	2.7 yuan per time
		OCR identification of ID card	0.06 yuan per time
		Face recognition	0.4 yuan per time
		OCR identification of bank card	0.06 yuan per time
		Checking consistence between face and ID card	1.8 yuan per time
Total	1650 元	Total	425.42

Table 31: Cost Comparison of Two Mortgage Models (RMB 100,000 as an Example)

	New Loans (100 Million)	Number of New Loans	Cost per Transaction in 2018 (Yuan)	Total Cost in 2018 (Yuan)	Cost per Transaction in 2019 (Yuan)	Total Cost in 2019 (Yuan)	Compared with 2018, The Single Transaction in 2019 Decreased (Yuan)	How Much Cost Will Be Saved in 2019 Compared with That in 2018 (Yuan)	The Cost In 2019 Is Lower Than That In 2018
Credit and Guaranteed Loan	9.824	9824	470	4617280	319.42	3137982.08	150.58	1479297.92	32.04%
Mortgage	5.526	5526	1650	9117900	425.42	2350870.92	1224.58	6767029.08	74.22%
Total	15.35	15350	2120	13735180		5488853		8246327	60.04%

Table 32: Analysis of Cost Reduction of New Loans in 2019

4.6.4. Performance and Comparison of Customer Timeliness of Citizen Loan

Loan Mode	Traditional Model		Online Mode (New Products of Citizen Loan)	
	Developing Customers Via Local Manager	Time for Handle Procedures	Developing Customers Online	Time for Handle Procedures
Credit and guaranteed loan	5 Working Days	3 Working Days	Anytime	3 hours
Mortgage	10 Working Days	10 Working Days	Anytime	1 Working Days

Table 33: Performance and Comparison of Customer Timeliness of Citizen Loan

4.6.5. Analysis of Old Customers

After the introduction of the new model in 2019, we analyzed and filtered the existing old customers. Through the analysis of the existing old customers, the average credit line of a total of 32452 old customers was increased. At the same time, due to the introduction of new analysis methods, the investigation and operation costs were greatly reduced. The average credit line increase rate of old customers was 9800 yuan/transaction, with a total amount of 318 million yuan. The investigation and operation costs were reduced by 115.41 + 21.22 = 1.3663 million yuan. At the same time, the new loan of 318 million yuan decreased the non-performing loan rate from 3.20% in 2018 to 1.1%, decreased by 2.1%, and saved losses of 6.679 million yuan. A total of 8.0453 million yuan was saved in the investigation and operation cost and risk control.

Referring to the experience of fully mining old customers in 2019, BTB will continue to increase the average credit line and reduce the fees of old customers in the first quarter of 2020. In the first quarter of 2020, 9,245 old customers will be excavated, and 87.82 million yuan of new loans will be added. The cost of investigation will be saved by 370,000 yuan, the non-performing loan rate will be reduced to 0.13%, and the non-performing loss will be saved by 851,900 yuan. The effect is remarkable.

4.6.6. Further Research

The Local model and financial analysis of citizen loans show that the current research is successful. The next research will focus on the following three aspects:

1. Further improve the local big data and introduce more data sources. At present, some important local big data sets, such as the real estate ownership of ordinary citizens, still need to be connected to improve the local data collection.

2. Further improve the ability of the XGBoost model. The previous model is mainly based on the bank's existing 30774 loan customers before 2019. With the increase of the number of customers in 2019 and the clarity of customer performance, the model needs to be upgraded iteratively on the basis of larger data.

3. Further improve the ability of Knowledge Graph. In Baotou, local people account for more than 83%. All kinds of people have a variety of inextricable relationships. The Knowledge Graph needs to further explore these relationships to improve the identification ability of the model.

4.6.7. Summary

In this paper, the new local model was used to establish the Citizen loan, which was operated in one year in 2019, and the actual income was very significant.

Before the system went online (2018) and after the system went online (2019), comparisons are as follow:

Customer Benefit Analysis:
- (1) the acquisition time of a small loan was reduced from 3 days to 3 hours;
- (2) the number of transactions: the number of transactions increased from 73,788 in 2018 to 197,233 in 2019, with wider customer coverage;
- (3) interest rate: the average annual interest rate of customers decreased from 10.96% to 8.72%;

Analysis of Different Types of Customers:
- (1) it can cover 70% of the local population, only local people who have no bad records can borrow more than 8,000 yuan;
- (2) among them, it can cover 800,000 customers who have not borrowed money before;
- (3) the average credit line of the original 800,000 customers having personal credit references can be adjusted from 3,000 yuan to 20,000 yuan;

Bank Profit Analysis:
- (1) the loan balance reached 1.535 billion in 2019;
- (2) credit line: the average credit line of customers reaches 43,728 yuan, and the risk is further dispersed, which is significantly better than that in the past (96,683 yuan), close to the Internet Bank (5,000 yuan) and better than other local banks (72,187 yuan);
- (3) the current non-performing loan and expected non-performing loan are less than 1.2%; and
- (4) the estimated net profit can reach 4.4%.
- (5) cost: the cost decreased rapidly, which decreased by 60.04%;

From these data, we can see that both the customer benefit and the bank benefit have been greatly improved.

5. Conclusion

Through the application of an online ‘Citizen loan’ personal credit product of BTB, this paper deeply studies and analyzes all kinds of data fusion and modeling that can be obtained by small and medium-sized banks, and establishes the risk control model of localized Citizen loan personal credit product by using XGBoost and Knowledge Graph tools. In this practice process, this paper deeply compares and analyzes the advantages and disadvantages of the models based on public data sets and local data sets (superimposing the former), and draws an exciting conclusion that the optimization local model based on local data sets has better performance. This ‘Citizen loan’ risk control model based on local data set was actually deployed and applied in BTB in 2019, and it has greatly improved both customer benefits and Bank benefits (for example, the fastest time for customers to obtain loans is 3 hours; the average interest rate of customers is reduced from 10.96% to 8.72% annually; the non-performing loan rate in 2019 decreases by 6% compared with that in 2018 2.8%; the cost of the new 1.535 billion yuan loan decreased by 60.04% compared with that in 2018. This result fully shows that the research results of this paper can help small and medium-sized banks to find a feasible way to realize the successful transformation of retail strategy through financial technology empowerment, which has universal application value.

The value of this paper is that it can help more than 2,000 local small- and medium-sized banks in China to compete with online internet giants and offline homogeneous banks in terms of retail loans and obtain competitive advantages.

In the future, with the further collection and expansion of local big data and the further optimization and improvement of local mode, the model and new products of Citizen loans will play a greater value in BTB.

References

Arjunwadkar, Parag Y. 2018. *FinTech: The Technology Driving Disruption in the Financial Services Industry* [M]. New York: Auerbach Publications.

Baptista, José A. G., Ramalho, Joaquim J. S., & da Silva, Jacinto V. 2006. Understanding the Microenterprise Sector to Design a Tailor-made Microfinance Policy for Cape Verde [J]. *Portuguese Economic Journal*, 5(3): 225-241.

Batiz-Lazo, Bernardo, Maixé–Altés, Carles J., & Thomes, Paul. 2011. *Technological Innovation in Retail Finance: International Historical Perspectives* [M]. New York: Routledge.

Bian, Weilin 卞维林. 2017. *Zhuanxing zhisheng: Yinhang lingshou yewu xiaoneg tisheng zhuanxing celve* 转型致胜：银行零售业务效能提升转型策略 (Transformation to Win: Transformation Strategies to Improve Bank Retail Business Efficiency) [M]. Nanjing: Jiangsu People's Publishing.

Cai, Liyan 蔡丽艳, Feng, Xianbin 冯宪彬, & Ding, Rui 丁蕊. 2011. *Jiyu juece shu de nonghu xiao e daikuan xinyonf moxinf yanju* 基于决策树的农户小额贷款信用评估模型研究 (Research on the credit evaluation model of farmers' small loans based on decision trees) [J] *Anhui nonye kexue* 安徽农业科学 (*Anhui Agricultural Sciences*), 39(2): 1215-1217.

Cao, Lei 曹磊, & Qian, Haili.钱海利. 2016. *Fintech jinrong keji gemin.* FinTech 金融科技革命 (*FinTech Financial Technology Revolution*) [M]. Taipei: Business Weekly Culture Press.

Chen, Chunbao 陈春宝, Xu, Xiaogang 徐筱刚, & Tian Jianzhong 田建中. 2017. *SAS jinrong shuju wajue yu jian mo: Xitong fangfa yu anli jiexi* SAS 金融数据挖掘与建模: 系统方法与案例解析 (SAS Financial Data Mining and Modeling: Systematic Approach and Case Analysis) [M]. Beijing: Machinery Industry Publishing.

Chen, Lianwei.陈良维. 2008. *Juce shu suanfa zai nonghu xiao e daikuan Zhong de yingyong yanjiu* 决策树算法在农户小额贷款中的应用研究 (Research on the Application of Decision Tree Algortithm in Farmers' Small Loan) [J]. *Jisuanji gongcheng zu yingyong* 计算机工程与应用 (*Computer Engineering and Application*), 44(31): 242-248.

Chen, Tianqi, & Guestrin, Carlos. 2016. XGBoost: A Scalable Tree Boosting System [J]. *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 16: 785-794.

Chi, Guotai 迟国泰, & Wang, Wei 王卫. 2009. *Jiyu kexue fazhan de zonghe pingjia lilun, fangfa yu yingyong* 基于科学发展的综合评价理论、方法与应用 (*Comprehensive Evaluation Theory, Method and Application Based on Scientific Development*) [M]. Beijing: Science Press, ch. 9: 27-79.

Chi, Guotai. 迟国泰. 2010. *Zhonggup ypizheng chuxu yinhang nonghu xiao e daikuan xinyong fengxian juece pingjia xitong yanjiu jie xiang baogao* 中国邮政储蓄银行农户小额贷款信用风险决策评价系统研究结项报告 (*Final Report on Research on the Credit Risk Decision-Making and Evaluation System for Small-Amount Loans to Farmers of the Postal Savings Bank of China*) [R]. Dalian: Dalian University of Technology.

Copestake, James. 2007. Mainstreaming Microfinance Social Performance Management or Mission Drift [J]. *World Development*, 35(10): 1721-1738.

Croxford, Hugh, Abramson, Frank, & Jablonowski, Alex. 2007. *The Art of Better Retail Banking: Supportable Predictions on the Future of Retail Banking* (Translated into Chinese by Zhao Ruian). Beijing: Economy Science Publishing House.

Cui, Jian 崔健. 2005. *Shangye yinhang geren xinyong fengxian pingjia* 商业银行个人信用风险评价 (*Commercial Bank Personal Risk Credit Assessment*) [D]. Tianjin: Tianjin University.

Cui, Junyang 崔军扬. 2011. *Jilin sheng nongcun xinyong she nonghu daikuan weiyue de yingxiang yinsu fenxi* 吉林省农村信用社农户贷款违约的影响因素分析 (*Analysis of Factors Affecting Farmers' Loan Defaults in Rural Credit Cooperatives in Jilin Province*) [D]. Guilin: Guilin University.

Dai, Yu 戴宇. 2018. *Nonghu daikuan xinyong fengxian pingjia de yanjiu jiyu lin'an qu A yinhang de shili*. 农户贷款信用风险评价的研究基于临安区 A 银行的实例 (*Research on Credit Risk Evaluation of Farmers' Loans Based on the Case of Bank A in Lin'an District*) [D]. Zhejiang Agriculture and Forestry University: 3-7.

Ding, Zhenhui 丁振辉. 2014. *Da shuju beijing xia de xiao wei qiye xinyong pingji yanjiu* 大数据背景下的小微企业信用评级研究 (Research on Credit Rating of Small and Micro Enterprises in the Context of Big Data) [J]. *Zhen xin* 征信 (*Credit Report*), 11：1674-747.

DiVanna, Joseph A. 2005. *The Future of retail Banking: Dleivering Value to Global Customers*. (Translated into Chinese by Qin Donghai and Zheng Ying) [M]. Beijing: Economy Science Publishing House.

Dutta, Soumitra, & Shekhar, Shashi. 1988. Bond Rating: A Non-Conservative Application of Neural Networks [J]. *IEEE International Conference on Neural Networks*: 443-450.

Fan, Wenzhong范文仲. 2014. *Hulianwang jinrong lilun, shijuan yu jianguan* 互联网金融理论、实践与监管 (*Internet Finance Theory, Practice and Supervision)* [M]. Beijing: Economy Science Publishing House.

Goldfinch, Peter. 2018. *A Global Guide to FinTech and Future Payment Trends* [M]. New York: Routledge.

Han, Jiawei, & Kamber, Micheline. 2007. Data Mining: Concepts and Techniques (Translated into Chinese by Men Xiaofeng) [M]. Beijing: Machinery Industry Publishing.

Harper, Richard, Randall, David, & Rouncefield, Mark. 2012. *Organisational Change and Retail Finance: An Ethnographic Perspective* [M]. New York: Routledge.

Hartarska, Valentina, & Nadolnyak, Denis. 2007. Does Rating Help Microfinance Institutions Raise Funds? Cross-country Evidence [J]. *International Review of Economics and Finance*, (5):1-14.

He, Pingping 何平平. & Che, Yueyun.车云月. 2017. *Da shuju jinrong yu zheng xin*大数据金融与征信 (Big Data Finance and Credit Reporting) [M]. Tsinghua University Press.

Huang, Zijian 黄子健, & Wang, Long 王龑. 2015. *Da shuju, hulianwang jinrong yu xinyong ziben, pojie xiao wei qiye rongzi bei lun* 大数据、互联网金融与信用资本，破解小微企业融资悖论 (Big Data, Internet Finance and Credit Capital solve the Financing Paradox of Small and Micro Enterprises) [J]. *Jinrong jingjixue yanjiuyuan* 金融经济学研究院 (*Institute of Financial Economics*), 30(1).

Jha, Saurav, & Bawa, Kamaljit S. 2007. The Economic and Environmental Outcomes of Microfinance Projects: An Indian Case Study [J]. *Environment, Development and Sustainability*, 9(3): 229-239.

Lei, Fengyu. 雷丰羽. 2018. *Zhishi tupu zai jinrong xindai lingyu de yingyong siandai shangye* 知识图谱在金融信贷领域的应用 (Application of Knowledge Graph in the Field of Financial Credit) [J]. *Xiandai Shangye* 现代商业 (*Modern Business*): 89-90.

Li, Baoxu 李保旭, Han, Jiyang 韩继炀, & Feng, Zhi 冯智. 2018. *Hulianwan jirong chuangxin yu fengxian guanli* 互联网金融创新与风险管理 (*Internet Financial Innovation and Risk Management*) [J]. Beijing: Machinery Industry Press.

Li, Yang 李杨, Sun, Guofeng 孙国峰, Zhu, Yedong 朱烨东, & Wu, Xuchuan 伍旭川. 2018. *Zhongguo jinrong keji lanpish: Zhongguo jinrong keji fazhan baogao* 中国金融科技蓝皮书：中国金融科技发展报告(China FinTech Development Report)[R]. Beijing: Social Sciences Academic Press.

Li, Qingping 李庆萍. 2017. *Lingshou zhi dao* 零售之道 *The Way of Retailing* [M]. Beijing: CITIC Press Corporation.

Li, Zhengbo 李正波, Gao, Jie 高杰, & Cui, Weijie 崔卫杰. 2006. *Nongcun xinyong she nonghu daikuan de xinyong fengxian*农村信用社农户贷款的信用风险评价研究 (Research on Credit Risk Evaluation of Farmer Loans from Rural Credit Cooperatives) [J]·*Beijing Dianzi keji xueyuan xuebao* 北京电子科技学院学报 (*Journal of Beijing Institute of Electronic Science and Technology*), 14 (1):69-74.

Lv, Jing 吕京娣, & Lv, Dehong 吕德宏. 2011. *Qian fada diqu xiao e xindai huan kuan lv yingxiang qiu* 欠发达地区小额信贷还款率影响因素实证分析 (Empirical Analysis of Factors Affecting Microcredit Repayment Rates in Underdeveloped Areas) [J]. *Guangdong nongye ke* 广东农业科 (*Guangdong Agricultural Science*): 245-247.

Mahjabeen, Rubana. 2008. Microfinancing in Bangladesh: Impact on Households, Consumption and Welfare [J]. *Journal of Policy Modeling*, 30(6): 1083-1092.

Mayer-Schönberger, Viktor, & Cukier, Kenneth. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think* [M]. Boston: Houghton Mifflin Harcourt.

Mayer-Schönberger, Viktor, & Cukier, Kenneth. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. (Translated into Chinese by Zhou Tao) [M]. Hangzhou: Zhejiang People's Publishing House.

Nicoletti, Bernardo. 2018. *The Future of FinTech: Integrating Finance and Technology in Financial Services*. (Translated into Chinese by Cheng Hua) [M]. Beijing: Posts & Telecom Press.

Tan, Pang-Ning, Steinbach, Michael, & Kumar, Vipin. 2014. *Introduction to Data Mining*. London: Pearson.

Sun, Guofeng 孙国峰. 2019. *Jinrong keji shidai de defang jinrong jianguan* 金融科技时代的地方金融监管 (*Local Financial Supervision in the FinTech Era*) [M]. Beijing: Economy Science Publishing House.

Wang Chunfeng 王春峰, Zhao Xin 赵欣, & Han Dong 韩冬. 2005. *Jinyu gaijin yi qun suanfa de shangye yinhang xinyong fengxian pinggu fangfa* 基于改进蚁群算法的商业银行信用风险评估方法 (Credit Risk Assessment Method for Commercial Banks Based on Improved Ant Colony Algorithm) [J]. *Tianjin daxue xuebao* 天津大学学报 (*Journal of Tianjin University*), 7(2): 81-85.

Wei, Qiang 魏强, & Wang, Diguang 王迪光. 2016. *P2P wang dai pingtai jiekuan ren xinyong fengxian pingjia zhibiao tixi de yanjiu* P2P 网贷平台借款人信用风险评价指标体系的研究 (Research on the Credit Risk Evaluation Index System of Borrowers on P2P Online Lending Platforms) [J]. *Shang* 商 (*Business*) 14: 157.

Wei, Yanling 韦艳玲. 2009. *Jiyu mohu ju lei de nonghu xinyong xinxi fenxi* 基于模糊聚类的农户信用信息分析 (Analysis of Farmer Household Credit Information Based on Fuzzy Clustering) [J]. *Guanxi minzu daxue xuebao (ziran kexue ban)* 广西民族大学学报(自然科学版) (*Journal of Guangxi University for Nationalities (Natural Science Edition))*, 15(1): 78-80.

Wen, Tao 温涛, Ran, Guanghe 冉光和, Wang, Huangyu 王煌宇, & Xiong, Deping 熊德平. 2004. *Nonghu xinyong pinggu xitong de sheji yu yunyong yanjiu* 农户信用评估系统的设计与运用研究 (Research on the Design and Application of Farmer Credit Evaluation System) [J]. *Yunchou yu guanli* 运筹与管理 (*Operations and Management*), 13(4): 82-87.

Xia, Meng 夏萌, Zhao, Banghong 赵邦宏, & Wang, Junqin 王俊芹. 2015. *Jiyu xiangguan fenxi fa de nonghu daikuan xinyong yingxiang yinsu fenxi* 基于相关分析法的农户贷款信用影响因素分析 (Analysis of Factors Affecting Farmer Loan Credit Based on Correlation Analysis Method) [J]. *Keji tongbao* 科技通报 (*Science and Technology Bulletin*), 31(10): 266-268.

Xie, Xunli.谢绚丽. 2018. *Keji fu neng: Zhongguo shuzi jinrong de shangye Shijian* 科技赋能：中国数字金融的商业实践 (*Technology Empowerment: Business Practice of China's Digital Finance*) [M]. Beijing: Remin University Publishing

Xu, Jiana.徐佳娜. 2004. Guanyu AHP・ANN moxing de shangye yinhang xinyong fengxian pinggu 基于 AHP・ANN 模型的商业银行信用风险评估 (Credit Risk Assessment of Commercial Banks Based on AHP・ANN Model) [J]. *Journal of Harbin University of Science and Technology*, 9(3): 96-99.

Yang, Tao 杨涛, Ben, Shenglin 贲圣林, Yang, Dong 杨东, & Song, Ke 宋科. 2018. *Zhongguo jinrong keji yunxing baogao* 中国金融科技运行报告 2018 (*China FinTech Operation Report*) [R]. Beijing: Social Sciences Academic Press.

Yao, Zhiyong. 姚志勇. 2010. *SA biancheng yu shuju wajue shangye anli* SAS 编程与数据挖掘商业案例 (*SAS Programming and Data Mining Business Cases*) [M]. Beijing: Machinery Industry Publishing.

Yu, Fenghui 余丰慧. 2018. *Jinrong keji : Da shuju, qu kuai lian he rengong zhineng de ying yong yu weilai.* 金融科技：大数据、区块链和人工智能的应用与未来 (*Financial Technology : Application and Future of Big Data, Blockchain and Artificial Intelligence* [M]. Hangzhou: Zhejiang University Publishing.

Yu, Xiaoyang 于小洋. 2018. *Qiye faren zhishi tupu de goujian ji yingyong yanjiu* 企业法人知识图谱的构建及应用研究 (Research on the Construction and Application of Corporate Legal Person Knowledge Graph) [M]. Qindao University.

Zhang, Runchi 张润驰, Du, Yabin 杜亚斌, Jing, Wei 荆伟, et al. 2017. *Nonghu de xiao e daikuan weiyue yinsu yingxiang yanjiu* 农户的小额贷款违约因素影响研究 (Research on the Factors Affecting Farmers' Small Loan Default) [J].*Xibei nonglin keji daxue xuebao 8shehui kexue ban)* 西北农林科技大学学报(社会科学版) (*Journal of Northwest Agricultu re & Forestry University (Social Science Edition)),*17(3): 67-75.

Zhang, Tao 张涛. 2017. *Juyu juece shu moxing de nonghu chaquan diya daikuan fenlei de shizheng yanjiu* 基于决策树模型的农户产权抵押贷款分类的实证研究 (Empirical Study on the Classification of Farmers' Property Rights Mortgage Loans Based on Decision Tree Model) [J]. *Shaanxi Nongye kexue* 陕西农业科学 (*Shaanxi Agricultural Sciences*), 63(4): 91-94.

Zheng, Lanxiang 郑兰祥, & Wan, Xue 万雪. 2014. Jiyu Logit fa de woguo nongcun xiao e daikuan gongsi xinyong fengxian pingfen moxing goujian yanjiu 基于Logit法的我国农村小额贷款公司信用风险评分模型构建研究 (Research on the Construction of Credit Risk Scoring Model for Rural Small Loan Companies in My Country Based on Logit Method) [J]. *Anhui nongye daxue xuebao* 安徽农业大学学报 (Journal of Anhui Agricultural University), 23(4): 49-54.

Resume of the Author

My name is Yunxiang Chen, and I was born in Bayannao'er City of Inner Mongolia Autonomous Region in September 1970. I am a member of Communist Party of China. I've gotten a master's degree and qualifications of senior economist and intermediate auditor of China.

Education:

2016/6 *to now*: The third class of global finance GFD Jinbo, Wudaokou business school, Tsinghua University.

2014/6-2020/6：School of economics and management, National University of Singapore, EMBA course.

2008/4-2010/7: School of economics and management, Inner Mongolia University, MBA course;

1988/9-1990/7: Majoring in agriculture and animal husbandry finance in Inner Mongolia Agricultural Bank school;

Work Experience:

2015/10 *to now*: Chairman of Baotou Rural Finance Research Institute

2014/2 *to now:* Secretary of the Party committee and chairman of the Baotou Rural Commercial Bank Co., Ltd.

2012/7-2014/1: Secretary of the Party committee and chairman of the board of directors of Baotou Suburban Rural Credit Union Co., Ltd;

2011/6-2012/6: Secretary of the Party committee and chairman of the board of directors of the rural credit cooperative association in the suburbs of Baotou city;

2006/5-2011/5: Deputy Secretary of the Party committee and chief supervisor of Baotou Suburban Rural Credit Cooperatives (transferred from October 2005 to January 2007 in the audit and supervision department of rural credit cooperatives in the autonomous region);

2001/4-2006/4: The chief supervisor of Rural Credit Union in Linhe District, Bayannaoer City;

1998/3-2001/3: The chief supervisor of Rural Credit Union in Dengkou county;

1996/4-1998/2: The director of Shuguang credit union of Linhe City;

1995/2-1996/3: The director of Bayi credit cooperative of Linhe Rural Credit Association;

1992/10-1995/1: The director of Xiaozhao credit cooperative of Linhe Rural Credit Association;

1990/9-1992/9: The credit officer, accountant in charge and deputy director successively;