# Risk Assessment of Internet Credit Based on Big Data Analysis

WANG HAORU[1], Yi Zhixuan[2], WEI YUJIA[3*a], Tianpeng Yao[4], ZhaoShuoheng[5], Xuzhiqiang[6]

[1]Northeastern University, Shenyang, China
[2]Southwestern University of Finance and Economics, Chengdu, China
[3]North China University of Technology, Beijing, China
[4]University College London (postgraduate), London, England
[5]Tianjin Zhonghuan Information College of Tianjin University of Technology, Tianjin, China
[6]Nantong Institute of Technology, Nantong, China

**Abstract—**In recent years, network technology has continued to develop, and Internet finance has rapidly developed into a new business area. Internet credit is one of the important ways for banks to conduct business, and the scale of online credit has continued to expand. Due to the existence of various unpredictable factors, frequent emergencies, and online financial fraud, the overall market risk in the field of online credit has increased, and the rate of non-performing loans has continued to increase. Online financial fraud cases show that online credit risk has become one of the most prominent risks in the operation of commercial banks, which has a direct impact on the stability and development of commercial banks. We can build a bank database system based on big data, introduce professional big data analysis technical personnel, and constantly improve the big data sharing analysis platform, so that commercial banks can use system data more fully and effectively, and facilitate relevant business personnel to use big data technology for analysis and calculation. Big data is constantly produced, which provides basic materials for online credit risk assessment. Big data analysis technology is gradually mature, and it has the necessary conditions for online credit risk assessment. Based on the theories and technologies related to big data analysis, this paper comprehensively evaluates the online credit risk in the form of example data analysis, thereby effectively reducing the online credit risk coefficient.

## 1 Introduction

Big data analysis methods have become a new topic that needs to be studied in various fields. "Big data technology" is slowly integrated into the work and life of human beings. With the role of computers and network platforms, the use of large amounts of data for analysis is becoming increasingly common. As a pioneer in the field of big data, the Internet has become increasingly integrated with the financial industry. Therefore, the application of big data in the financial industry is also more extensive, and financial services such as personal credit also requires the further support of big data analysis. In the design of the bank's bad credit early warning system under big data, it is necessary to centrally process the inaccurate data in the mass data and establish corresponding data relationship algorithms to optimize the bank's bad credit risk problem. Commercial banks are financial intermediaries and need to bear corresponding operating and management risks, especially when they are serving the real economy. There are many types of risks, including credit market operations and liquidity risks. In recent years, China has begun to establish information technology reforms. The use of big data, mobile technology, and the Internet of Things has led to commercial transformation of financial institutions, especially in the context of the new economic normal. Commercial banks must pay attention to modern science and technology Application to strengthen risk prevention and control. At present, many banks have begun to apply the IT system framework, and continue to explore. They have established a credit risk early-warning mechanism in combination with big data technology and carried out a series of practices, and achieved good results.

## 2 Overview of big data and big data analysis technology

### 2.1 Big data overview

Big data is also called massive data. Its essence is the collection of data. It is mainly big data generated during the development of Internet information technology. The use of computer technology can capture management or

[a*]Corresponding author's e-mail: WYJ18301191900@163.com

process data in a corresponding time. According to the International Data Corporation (IDC), the amount of data generated by humans is increasing exponentially, doubling approximately every two years.

The basic characteristics of big data include:

Mass is huge. Big data includes three main forms: structured, semi-structured, and unstructured. The Internet era has brought PB-level semi-structured and unstructured data. Compared to structured data, unstructured data that is inconvenient to be represented by a two-dimensional logical table in a database is not only huge in size but also growing rapidly.

High value and low density. The value of scarcity, uncertainty and diversity of big data, valuable information may be fleeting, which poses a serious challenge to predictive analysis. In the era of big data, the idea of receiving and processing data needs to be changed.

### 2.2 Big data analysis overview

The raw data of big data analysis is extracted and integrated from heterogeneous data sources. Effectively expressing, interpreting and learning these data becomes the core problem of big data analysis. The main goals of data analysis include inferring or interpreting data and determining how to use the data, data legality verification, decision support, error cause identification, and predictive analysis. The big data processing flow is shown in Figure 1 below.

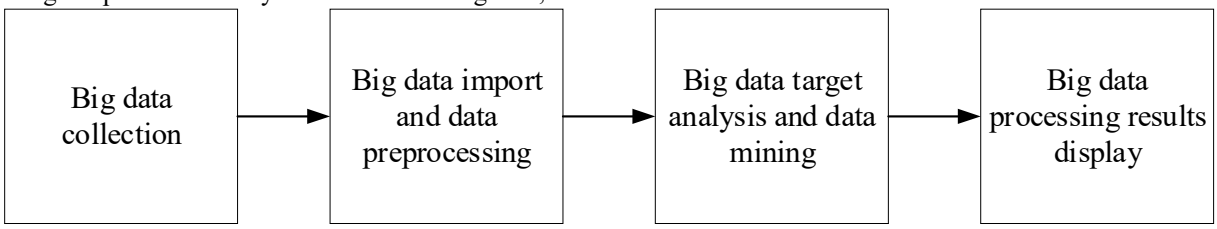


**Figure 1.** The basic flow of big data processing

Big data analysis technology involves the following aspects.

Big data information visualization analysis technology. By constructing a data information processing model, the original data is transformed into data in a visual form. The information visualization technologies of mainstream big data applications include text data visualization technology, network data visualization technology, spatio-temporal data visualization technology, and multidimensional data visualization technology.

Data mining technology. In the process of big data analysis, data mining is the core work. After a long collection of raw data, a data source for data mining and analysis is formed. These data generally have no regularity, which includes subject data and non-subject data. The mass of data gathered together needs to classify the original data according to the need scenario and extract the required data content from it. The basic process of data mining is shown in Figure 2 below.

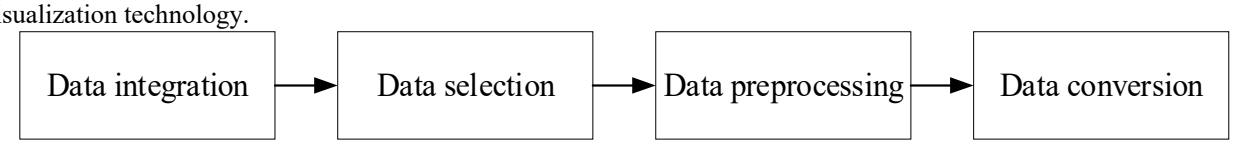


**Figure 2.** Basic process of data mining

Among the classification and prediction methods, classification is based on the basic characteristics of a known training data set, constructing a classifier, and classifying the data set through supervised learning. According to the classification results, a data prediction model is constructed through association analysis.

Predictive analytics technology. Based on the existing big data, modeling, data mining, and statistical analysis techniques can be used to complete the forecasting work.

## 3 Data preparation for online credit risk assessment

### 3.1 The main risk factors of online credit

*1) Market risk:* According to the current situation, in terms of personal online credit, market risk is a factor that hinders online credit and repayment in the economic environment. Exchange rates, interest rates, and leverage often cause market risks and cause loan losses. Personal loan risks are mainly the risks and policies to respond to changes in interest rates. Customers who apply for online credit will have an impact on profits and earnings. Changes in national policies will affect credit business. At the same time, the measures of the People's Bank of China will also cause floating interest rates on loans, which can cause market risks to varying degrees.

*2) credit risk:* Credit risk is divided into two types: default risk and credit spread risk, which correspond to the following: the borrower does not repay or loses the ability to repay due to subjective reasons; The downgrade of customers' credit rating leads to the risk of increasing loan spreads. Internet credit risk is one of the common risks at present, and it has become an important factor leading to bad debts of banks. Effective measures need to

be taken for risk assessment and prevention. Before the implementation of online credit, an effective evaluation model is used to predict the customer's ability to repay, and finally the online credit risk is evaluated.

*3) Operational risk:* The operational risk of online credit mainly refers to the fact that before the lending, the relevant business personnel did not conduct a comprehensive investigation and evaluation of the basic situation of the loan subject. For online credit, banks generally have auditing and operating regulations. In the actual application process, due to the large number of online credit customers, the variety of credit applications, and the complexity of credit subjects, the investigation and verification work was not thorough enough; In the process of online credit review and management, due to the lack of professional operation processes and the professional quality of employees, the lack of real-time monitoring of online credit; At present, the risk assessment methods for online credit of many banks are relatively backward, and the risk assessment model needs to be further optimized.

### 3.2 Data sources for online credit risk assessment

In the era of big data, the channels for collecting data on participants in online credit are more extensive. We can collect these data information from multiple aspects, and use it as the raw data for the later input of the online credit risk assessment model. One part is the existing data information of the traditional credit reference model, which mainly includes: application information submitted by customers to commercial banks, customer historical transaction data accumulated internally by the bank, and data provided by external agencies such as the People's Bank of China Credit Reference. These data dimensions are relatively narrow, the information value density is highly concentrated, and the pertinence is strong. The other part is the era of big data. The data collected in multiple dimensions after the information collection channels are expanded, mainly includes daily transaction information such as clothing, food, housing, and transportation on social network platforms, and business, tax, court, and provident funds for government service platforms, hydropower and other information.

### 3.3 Data Design for Cyber Credit Risk Assessment

*1) Sample data selection:* In order to verify the classification performance of personal credit assessment models, they often need to be trained and tested through historical credit data. Considering the current privacy protection of customer credit information by financial institutions in China, it is difficult to obtain personal credit data of commercial banks for research. Therefore, this article uses crawler technology to capture P2P personal credit data on Renrendai website. Using data mining technology in big data processing methods, key information is extracted from the massive data obtained as a sample of customer credit data for online credit risk assessment.

Considering that there are too many indicators for the Renren loan sample, the following 15 indicators were retained after screening: age, marriage, education, income, real estate, mortgage, car production, car loan, company industry, company size, working hours, annual interest rate , Term, total borrowings, number of overdue. The basic information of the online credit risk assessment data set is shown in Table I.

**TABLE I.** THE BASIC INFORMATION OF THE ONLINE CREDIT RISK ASSESSMENT DATA SET

| Total number of samples | Risk-free customers | Risk client | Number of attributes | Category attributes |
|---|---|---|---|---|
| 2783 | 1740 | 1043 | 15 | 1 |

*2) Indicator data analysis:* Credit indicators include information such as loan information, credit card information, and historical credit information, so as to understand its credit risk and debt pressure, as well as its historical credit. Both samples involve more credit indicators. Credit amount, credit term, credit use, credit score, and number of overdue indicators can better reflect the borrower's credit history and credit status, which can be used by commercial banks for loan approval for reference. Generally speaking, borrowers with historical overdue times and long credit terms have weak repayment ability and high risk of default.

Economic indicators include information such as the borrower's position, seniority, and income, and are important indicators of loan applicants' ability to repay. The income indicator can more directly reflect the repayment ability of the loan applicant. A higher income can not only guarantee its daily basic living expenses, but also provide repayment ability.

## 4 Data processing and online credit risk assessment

### 4.1 Sample data preprocessing

Data preprocessing can be divided into four steps: processing of missing values, assignment of sample data, data normalization, and grouping of sample data. The Renrendai credit data set contains personal information, credit information, economic information, authentication information, and other aspects, which are relatively complicated, and there are cases of missing data. When sorting out the Renrendai sample, it was found that among the 10 samples, 97 indicators, such as marriage, education, and work status, had missing data, which would adversely affect the empirical results. Therefore, the 97 samples with missing data were deleted, leaving a total of 2686 samples. The sample data pre-processing process is shown in Figure 3.

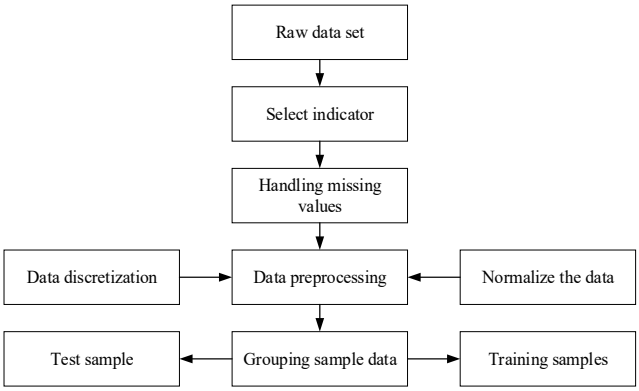


**Figure 3.** Sample data preprocessing process

### 4.2 Construction of Network Credit Risk Assessment Model

So far, there has not been a universally accepted method for assessing credit risk on the Internet. After years of reform and development, in terms of online credit risk assessment, quantitative management models have become mainstream. Currently, in the era of big data, after obtaining multi-dimensional data, the calculation of loan risk is used to evaluate online credit risk. Based on the actual situation in China, this paper believes that the use of Logistic regression model to measure customer credit risk is more suitable for online credit risk assessment.

*1) Variable setting:* To construct a network credit risk assessment model using the Logistic model method, the corresponding dependent and independent variables need to be set. For the convenience of research, this paper sets the dependent variable as whether the online credit customer has a default loan. If there is a default bandwidth, it is a default customer (assigned to 1), if there is no default loan, it is a normal customer (assigned to 0). The default probability of the online credit customer calculated by the model is less than 0.5, then the customer is judged as a normal customer, otherwise, it is judged as a default customer. The independent variables are 15 credit information indicators that reflect the basic status of online credit customers.

*2) Principal component analysis:* Due to the high correlation and high dimensionality of the information and data of online credit customers, when applying regression analysis to the prediction of default risk of online credit customers, it will affect the analysis process and results, leading to the loss of most of the original data and estimates Problems such as multicollinearity appear in the equations. Therefore, regression analysis requires that the explanatory variables of the model cannot have a linear relationship. Although collinearity does not change the estimated value of the coefficient itself, increasing the standard deviation will reduce the reliability of the coefficient, which will reduce the stability of the prediction results of the model. In order to further improve the model, this paper first uses principal component analysis to eliminate multicollinearity between explanatory variables by dimensionality reduction. Before the principal component analysis was performed, the principal components were extracted using the variance factor common factor, and the extraction results are shown in Table II.

**TABLE II.** VARIANCE DECOMPOSITION COMMON FACTOR EXTRACTION ANALYSIS TABLE

| Initial Eigen Values | Number of principal components | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Eigenvalues | 7.5 | 3.45 | 1.81 | 1.19 | 0.6 |
| Contribution rate | 0.45 | 0.17 | 0.09 | 0.07 | 0.03 |
| Cumulative contribution rate | 0.46 | 0.65 | 0.74 | 0.86 | 0.9 |

As can be seen from Table II, the eigenvalues of the first 4 principal components are greater than (or equal to) 1 and the cumulative contribution rate exceeds 80%, indicating that it is more appropriate to extract the principal components.

*3) Logistic regression analysis:* According to the risk assessment criteria for default customers on the Internet, among the selected 2686 sample data, the number of default customers is 1028, of which the control group sample is 1686, the default customer is 727, and the forecast group sample is The number of defaulted customers is 301.There are 4 main components entering the Logistic model, which are $F^1$, $F^2$, $F^3$, $F^4$, as shown in Table III.

**TABLE III.** LOGISTIC REGRESSION MODEL PARAMETERS BASED ON BIG DATA ANALYSIS

| Var | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|
| $F^1$ | 5.67 | 1.13 | 66.10 | 1 | 0 | 0 |
| $F^2$ | 2.13 | 0.16 | 32.59 | 1 | 0 | 0.35 |
| $F^3$ | 0.55 | 0.25 | 6.39 | 1 | 0.01 | 1.79 |
| $F^4$ | 1.56 | 0.33 | 27.68 | 1 | 0 | 0.20 |
| Constant | -1.54 | 0.25 | 20.99 | 1 | 0 | 0.36 |

Build model as:

$$F= 5.67F^1+2.13F^2+0.55F^3+1.56F^4-1.54 \quad (1)$$

The principal component model is:

$$P = \frac{\exp(F)}{1+\exp(F)} = \frac{1}{1+\exp(-F)} = \frac{1}{1+\exp(1.54-5.67F^1-2.13F^2-0.55F^3-1.56F^4)} \quad (2)$$

*4) Cyber Credit Risk Forecast Results:* The regression model constructed by the control group has an average accuracy rate of 82.5% for the credit risk

prediction of the control group, of which the prediction accuracy of normal customers is 83.84%, and the prediction result of default customers is 81.16%; The average prediction accuracy rate for the prediction group is 83.64%, of which the prediction accuracy rate of normal customers is 86.55%, and the prediction accuracy rate of default customers is 80.73%. The results of network credit risk prediction are shown in Table IV

**TABLE IV.** FORECAST RESULTS OF ONLINE CREDIT RISK

| Group | Judgment value 0 | Judgment value 1 | Normal percentage | Default percentage |
|---|---|---|---|---|
| Control group | 804 | 590 | 83.84% | 81.16% |
| Forecast Group | 605 | 243 | 86.55% | 80.73% |

## 5 Conclusion

In the era of big data, all walks of life are facing unprecedented data volumes and data analysis needs. Data analysis will be fully integrated into the process of enterprise product development and service. The value of data analysis needs to be explored at a deeper level, and the results of data analysis will really aid management decisions. From the source of big data, the channels for data acquisition are diversified; now the development of computers and information technology, and the continuous maturity of big data analysis technologies such as data mining, can enable us to obtain more research data more quickly, accurately, and deeply. Under the current circumstances, online credit risk assessment has become an important research work. After data mining of big data, we extract targeted data information in a targeted manner, build a mature data analysis model, and evaluate online credit risk, which is effective avoid online credit risk.

## References

1. Lusher S J，Mcguire R, Van schaik RC，et al. Data − driven Me-dicinal Chemistry in the Era of Big Data[J]. Drug Discovery Today, 2014, 19(7): 859 - 868.
2. Lee Y, Madnick S，Wang RE. A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data[J]. M IS Quarterly Executive, 2014, 13(1): 1-13.
3. Lu Yanfeng, Yu Pengfei. Research on the Big Data Strategy of Commercial Banks under the Background of Internet Finance-Application of Internet Finance in the Transformation and Upgrade of Commercial Banks [J]. Economy and Management, 2015 (5)
4. Huang Yandan. Research on Credit Risk Management Strategies of Commercial Banks in the Era of Big Data [J] .Research on Industrial Innovation. 2018 (11): 75-76 + 89.
5. Liu Lei.Analysis of Credit Risk Control of Construction Bank's Small and Micro Enterprises Based on Big Data Credit [J]. Economic and Trade Practice, 2017 (19): 94.
6. Chen Huijuan, Jia Yungang.Research on Credit Risk Early Warning System in the Big Data Era [J] .Software, 2018 (01): 39-44.
7. Qiu Guiling. Personal credit risk control of commercial banks under the background of big data [J] .Economist, 2018 (07): 142-143.
8. Huang Leidan.Establishment and Empirical Research on Credit Risk Evaluation Models of Commercial Banks Based on Big Data Algorithms [J] .Contemporary Economy, 2018 (22): 59-61.
9. Lu Yifeng, Wang Tingting.Strategic Research on Digital Credit Risk Control Management Based on Digital Banking [J]. Finance Theory and Practice, 2020 (01): 21-26.