Research on the Construction of Intelligent Risk Control Service Platform for Financial Institutions under Big Data Technology

Fan Zhang[1,a*], Yuhao Liao[2,b], Ye Ding[1,c]

[1]School of Economics and Trade, ShangHai Urban Construction College, 201415, Shanghai, China
[2]Business School, University of ShangHai for Science and Technology, 200093, Shanghai, China

[a*]zfansky@163.com, [b]liaoyuhao2003@163.com, [c]dingye@succ.edu.cn

Abstract. Due to the rapid change of financial technology, Internet finance has encountered a series of risk problems while continuously optimizing financial services, and the original risk control measures of financial institutions can no longer meet the current application needs. In this regard, based on the current risk control system of Internet financial institutions, this paper will propose a set of construction scheme of intelligent risk control service platform, aiming at taking advantage of the advantages of big data technology in risk control and optimization, thus helping financial institutions to assess risks conveniently and efficiently, monitor abnormal behaviors and realize sustainable risk management. The platform takes Hadoop cluster as data management and processing server, and combines Javaweb technology to form a comprehensive application service platform integrating online application, model operation, visual analysis and other functions. Practice has proved that the platform can use Logistic and XGBoost models to evaluate credit and behavior, effectively identify potential risk factors, and provide necessary data support for risk management of Internet financial institutions.

Keywords: Big data; Internet financial risks; Machine learning algorithms; Hadoop; Computer software applications

1 Introduction

At present, the global wave of scientific and technological innovation and the wave of industrial change converge, and the digital economy has become a "new engine" to lead the global economic growth. The rapid development of digital economy has promoted the deep application of digital technology in the financial field, and spawned a series of financial technologies. [1] Among them, Internet finance, as one of the important directions of financial technology development, can break through the barriers of traditional financial service mode by virtue of its convenient and efficient application advantages, provide users with richer financial products and services, and

C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_52

also help the financial structure achieve the purpose of reducing costs and increasing efficiency. [2] At the same time, Internet financial institutions are also facing risks such as network security, credit review, liquidity, market environment and laws and policies. The rule model under the traditional risk control measures can no longer meet the current application requirements, and it is urgent to rebuild the risk control system of Internet financial institutions with the practical advantages of big data technology and machine learning model. [3] In view of this, this paper proposes a construction scheme of an intelligent risk control service platform. The platform takes user behavior and user credit as research objects, and improves the accuracy of evaluation by setting a brand-new risk control model, so as to effectively identify potential risk factors and provide necessary data support for risk management of Internet financial institutions.

2 Platform construction

Firstly, according to the platform application requirements, Hadoop is built according to the highly available framework distributed cluster. It contains six functional nodes, two for master and four for worker, and each node is equipped with a separate server. The server consists of 3.2GHz 4-core CPU, 16G memory and 1TB hard disk. In the hardware environment, the bottom operating system is Linus CentOS 6.8, the basic development environment is Java, and the JDK version is jdk-1.8. Hadoop architecture is installed and deployed in the JVM virtual machine environment, and according to the data processing flow, components such as Flume, Sqoop, Hive, HDFS and MapReduce are installed and deployed in turn. [4] Secondly, the construction of business application layer and presentation layer of the platform mostly depends on Javaweb technology. Under the Javaweb technology system, the presentation layer is the user interactive interface, which is led by JSP technology and supplemented by HTML, CSS and JavaScript to complete the page development and deployment. The business logic layer takes the Spring framework as the core, and cooperates with Apache Tomcat 9.0 Web server to complete the definition and declaration of various functions of the system, encapsulates the algorithms and calculation processes of all functional applications needed by the system, and interacts with the data access layer and the presentation layer. [5] In addition, the development process will also need the support of IntelliJ IDEA 2019, Maven 3.5.0, MySql 8.0 and other software tools. Finally, after the installation and configuration of the above software systems one by one, the final application will be released to the server, and the corresponding network and IP address will be configured to support users to log in and use.

3 Functional implementation

3.1 Data source and data processing

In this paper, an overseas peer-to-peer lending platform is taken as the application target, and its actual user lending data is selected as the data source. The data span is

from February 2021 to February 2022, with a total of 565,741 rows and 76 columns of data field attributes. As shown in Table 1, it is part of the original data information.

Table 1. Some raw data information

Data field attribute classification	Field attribute number	Field description
Basic information	A1	Age
	A2	Sex
	A3	Job position
Application information	B1	Application time
	B2	Application quota
Log record information	C1	Registration time
	C2	Active within 7 days
Communication information	D1	Mobile phone number
Historical information	E1	Default record
...	...	...

The original data exists in the source data layer, and the platform needs to use data acquisition tools to extract the data. The data processing layer of the platform is built on the Hadoop framework, and data such as application information, transaction information and logging information will be stored in HBase, while a large number of offline batch data such as basic information and communication information will be stored in Hive. When users initiate risk assessment online, the platform will automatically extract the original data into HDFS through Kafka, Flume, Sqoop and other tools, and then after Hive cleaning, processing and calculation, all kinds of data will be stored in HBase and Hive respectively. [6] The process of cleaning, processing and calculating the original data in Hive is called data preprocessing. Common processing methods include deletion of missing values, detection of abnormal values, etc.According to the requirements of subsequent model operation, the field attributes with missing rate exceeding 60% are deleted.

3.2 Feature screening

Each model and algorithm in the platform has a certain machine learning upper limit, so it is necessary to strike a balance between running efficiency and calculation accuracy, that is, to filter the necessary features of the data to determine the field attributes of the final input model and algorithm. The feature screening algorithm is preset in the platform, and WOE coding will be used to calculate the IV value of field attributes to complete the final selection of field features. Formula 1 shows the calculation formula of WOE coding, and formula 2 shows the calculation formula of IV value, where mxi and mni are the sample proportions in the ith grouping. The calculation results of some field attributes are shown in Table 2. The IV value represents the contribution of the field attribute to the final result. When the IV value is less than 0.02, it can be judged that this field attribute has almost no effect on the operation of the model. When the IV value is between 0.5 and 0.7, it has a great contribution and is

suitable for input model mining. When the IV value is over 0.7, it is necessary to further verify the actual utility of this field feature. [7] After feature screening, 12 field features were finally determined.

$$WOE_i = \ln\left(\frac{m_{xi}}{m_{ni}}\right) \tag{1}$$

$$IV = \sum_N^i \left| m_X - m_N \right| \times WOE_i \tag{2}$$

Table 2. Some field properties calculation results

Field property number	WOE	IV
A1	-0.0471	0.0002
A5	0.1055	0.0049
C4	-0.1936	0.0031
D5	-1.3714	0.1875
...	...	...

3.3 Model setting

After feature screening, the platform supports users to select evaluation items. Common items include pre-loan credit evaluation, application evaluation, user behavior evaluation and collection evaluation. Taking application evaluation as an example, users set the historical data of the past 12 months to predict the default probability of users after applying for loans according to actual requirements, and divided them into three types according to their actual performance, namely, good customers, bad customers and fraudulent customers, and the criteria for judging the three types are shown in Table 3. [8]

Table 3. Model prediction target

No.	Type	Standard
01	Good customers	There is no default within 4 months, or there is slight default.
02	Bad customers	There is overdue arrears beyond 4 months.
03	Fraudulent customers	There are overdue arrears within 4 months.

At the same time, the platform also supports users to choose different evaluation models for operation, such as Logistic, XGBoost, GBDT and LightGBM. Different models have different operation effects, and the platform provides confusion matrix, ROC curve and AUC index to evaluate the model and help users to choose. The confusion matrix is shown in Table 4, and the definitions of precision and recall are derived. The ROC curve and AUC are generated according to the confusion matrix, in which the abscissa of ROC curve is FPR=FP/(TN+FP) and the ordinate is TPR=TP/(TN+FN), as shown in Figure 1. AUC is derived from ROC curve, that is, the area under the curve. The closer the AUC value is to 1, the better the classifier

performance is. Conversely, the closer the AUC value is to 0, the worse the performance of the classifier. [9]

Table 4. Confusion Matrix

Actual situation	Forecast results		Precision (P)	Recall (R)
	Positive	Negative		
Positive	True Positive	False Negative	$P = \frac{TP}{TP + FP}$	$R = \frac{TP}{TP + FN}$
Negative	False Positive	True Negative		

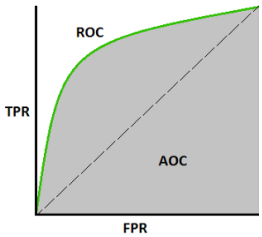

Fig. 1. ROC curve

3.4 Operation results

According to the model evaluation standard, when the user selects Logistic, the platform will fit the model according to the backward step-by-step method. Formula 3 shows the Logistic regression model, where e is the natural logarithm, θ is the parameter or regression coefficient, and x represents the field feature vector. [10]

$$W = \frac{1}{1+e^{-\theta x}}, \quad \log\left(\frac{W}{1-W}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_m x_m \tag{3}$$

The calculation result of Logistic model is the default probability of users, and a single probability can not meet the risk control of institutions, so it is necessary to introduce sample proportion odds to calculate the application score of users. Formula 4 shows the calculation formula of user's application score, where A and B are constants and S stands for score. The final application score calculation results and coping strategies are shown in Table 5.

$$S = A + B * \ln(odds) = A + B * \ln(\frac{W}{1-W}) \tag{4}$$

Table 5. Final application score calculation results and coping strategies

Score	Odds	Probability of default	Coping strategy
500	1:32	0.9696	Refuse the application
540	1:8	0.8888	
600	1:1	0.5000	
660	8:1	0.1111	Manual audit

700	32:1	0.0303	
720	64:1	0.0153	Pass the audit
800	1024:1	0.0009	

4 Conclusion

In order to promote the change of risk control system of Internet financial institutions, this paper builds an intelligent risk control service platform based on big data technology and machine learning algorithm. The platform can start with data collection, storage, analysis and processing, and set up Logistic, XGBoost and other algorithm models to improve the accuracy of evaluation, so as to effectively identify potential risk factors and provide necessary data support for risk management of Internet financial institutions. In the follow-up research, the platform will further enrich the system's support for other algorithm models and contribute to the realization of intelligent risk management.

References

1. Du Han. Research on Financial Risk Analysis and Supervision Countermeasures under the Background of Financial Science and Technology[J]. China Journal of Commerce.2023.08.
2. Wang Yu. Strategies for Internet Finance to Promote Economic Transformation and Upgrading[J]. Investment and Entrepreneurship.2023.08.
3. Cai Linxue, Wang Yongchang. Thoughts on the Prevention of Internet Financial Risks in the Information Age[J].Shanghai Business.2023.06.
4. Fakhri Nifdalizada, Maliha Nifdalizada. Creating A High-Performance Hadoop Cluster Consist Of Weak Computers Within The Organization[J].ETM Equipment Technologies Materials .2023.10
5. Shi Feng. Development and Application of JavaWeb Based on MVC Pattern[J]. Electronic Technique.202105.
6. Marwa Khadji Samira Kholji. Sustainable MapReduce: Optimizing Security and Efficiency in Hadoop Clusters with Lightweight Cryptography-based Key Management[J].E3S Web of Conferences.2023.08
7. Zheng Bei. Research on Online Credit Risk Control of Small and Micro Enterprises in Bank B[D]. Southwestern University of Finance and Economics.2021.09.
8. Feng Jiayin et al. Research on Financial Credit Risk Assessment Method Based on Deep Learning[J]. Investment and Cooperation.2023.08.
9. Liu Xinxing. Evaluation and Analysis of Credit Scoring Model Based on Multi-source Heterogeneous Credit Data[D]. Shanxi University.2021.06.
10. Zhao Jiaojiao. Research on Financial Risk Assessment Method of Supply Chain Based on Logistic Model[J]. China Collective Economy.2022.07.