Identification Model for Gambling and Fraud in Bank Personal Settlement Accounts Based on XGBoost Algorithm

Xin Li[1, a], Yan Ke[2, b], Jianbin Yu[3, c], Yao Zhang[4, d*]

[1]Technology Department, Nanchang branch, Bank of Beijing, Nanchang, Jiangxi, China
[2]Software College, Xinjiang University, Urumqi, Xinjiang, China
[3]Technology Department, Nanchang branch, Bank of Beijing, Nanchang, Jiangxi, China
[4]Technology Department, Nanchang branch, Bank of Beijing, Nanchang, Jiangxi, China

[a]lixin6@bankofbeijing.com.cn
[b]keyan@stu.xju.edu.cn
[c]yujianbinnc@bankofbeijing.com.cn
[d*]zhangyaonc@bankofbeijing.com.cn

Abstract. Based on the basic characteristics of sample accounts involved in gambling and fraud, this paper constructs a risk identification model for gambling and fraud in bank personal settlement accounts through machine learning algorithms, uses the model to accurately identify suspected fraudulent accounts, and analyzes risks through key features of gambling and fraud. The rules of account behavior provide decision support for the establishment of anti-fraud models. The model proposed in this paper can enhance the recognition accuracy and predictability based on the collision rules, and effectively avoid the problem of blocking normal accounts by mistake.

Keywords: data mining; anti-gambling and anti-fraud; XGBoost

1 Introduction

The application of digital technology has accelerated the comprehensive online and digital transformation of commercial banking business, while the advance of the business model and the sinking of the target customer base have made the fraud scenarios faced by banks also diversified. In particular, personal settlement accounts of banks, as carriers and mediums of capital flows, are easily used by various types of gangs of telecommunication network frauds, gambling organizations and other unscrupulous elements, posing a great threat to financial security [1]. It is necessary for financial institutions to use structured processes and intelligent systems to identify suspicious transactions such as gambling and frauds [2].

Many methods have been proposed in academia to identify bank accounts involved in gambling and fraud. [3] proposes a new semi-supervised text mining method, namely

H. Wu et al. (Eds.): CSIEDE 2022, ACSR 103, pp. 556–563, 2022.
https://doi.org/10.2991/978-94-6463-108-1_63

the plain collision algorithm, to mine bank risk factors. [4] uses integrated algorithms such as LightGBM for classification to identify gambling accounts. [5] utilizes logistic regression algorithms to detect suspicious transactions in bank accounts. Machine learning can be used to focus on the detection of online fraud and suspicious transactions, and thus can be applied in the banking industry to detect and prevent risk [6]. Although these methods have solved the problem of gambling and fraudulent account identification to a certain extent, they have also provided a seminal idea for subsequent scholarly research. However, the previous studies focused on improving the positive sample checking rate and did not consider the huge impact of mistakenly blocked numbers on bank customers.

To address this situation, we proposed identification model for gambling and fraud in bank personal settlement accounts (hereinafter to be referred as IMGF) for detecting the riskiness of gambling and fraud in bank personal settlement accounts, considering the industry characteristics such as high dimensionality, massive size and noise information complexity of the Bank of Beijing customer dataset [7]. The model innovatively combines collision rules proposed by the bank's business experts with machine learning algorithms. It improves the accuracy and identification prospect of gambling-related and fraudulent accounts, and effectively avoids the problem of false blocking caused by traditional collision rules to find gambling-related and fraudulent accounts.

2 Model Construction and Experiments

2.1 Overview of Model Construction

The construction process of the risk identification model for gambling and fraud in the bank's personal settlement account consists of three parts, namely the original feature extraction link, the feature validity analysis and processing link, and the machine learning model building link.

The first step is to analyze the relevant information of customers in the bank based on the experience of anti-gambling and anti-fraud business in the banking industry, and select a group of characteristic attribute values that have the function of judging the risk of gambling-related fraud, and obtain the "Original Feature Table".

The second step is to perform data exploration and necessary model training experiments on the extracted "Original Feature Table", exclude the feature attributes that are obviously not suitable for modeling, retain the feature attributes suitable for machine learning, and undergo feature engineering. Process the remaining feature attribute values to obtain a "feature table" suitable for input into the machine learning algorithm.

The third step is to input the "feature table" into the constructed machine learning model, and through the model validity analysis, continuously optimize the algorithm construction method of the overall model, obtain the model that can best learn the feature information contained in the data set, and then adjust the parameters to optimize Until convergence, the final model and its optimal parameter settings are obtained; according to the model and parameter settings in this paper, gambling-related fraud risk accounts can be captured from bank data in real time, and gambling-related fraud risks can be eliminated in time.

To sum up, this paper processes the relevant data of bank personal settlement accounts through the three-step process of the bank's personal settlement account gambling-related fraud risk identification model, and mines the gambling-related fraud risk account from it, which provides a different way for banks to combat gambling and fraud. The new risk screening system of the rules has improved the accuracy of gambling and fraud risk exclusion.

2.2 Feature Collection and Data Processing

For the identification of gambling-related fraudulent accounts, this chapter starts from the data available to banks, extracts features from it to analyze the gambling-related fraudulent behaviors of accounts, and builds a machine learning model for timely investigation and early warning of such behaviors.

After exploring the original data in the bank combined with business experience, this paper selects some important information from the three major account information types of personal information, behavior information and transaction information to construct a preliminary feature wide table.

There are 17 original features in the wide table, which are customer number, customer type, customer creation time, gender, age, marital status, nationality, expiration sign, CIF last update time, active days in the past 5 months, and accumulated transactions in the past 5 months. The number of transactions, the number of credit transactions in the past 5 months, the number of mobile phone changes in the past 5 months, the 150-day average daily balance, the ratio of the loan amount, the transaction balance ratio, and the last updated CIF serial number.

By eliminating meaningless fields, high degree of missing information, low content of effective information and low importance, the final 10 fields are obtained: customer creation time, gender, age, nationality, active days in the past 5 months, and nearly 5 Monthly cumulative transactions, recent 5-month credit transactions, 150-day average daily balance, loan-to-loan ratio, and transaction balance ratio.

When dealing with missing values, for categorical variables, missing values are filled with mode, integer variables, and missing values are filled with median, floating-point variables, The missing values are filled with the mean. When dealing with outliers, the boxplot method is used for detection, and the detected outliers are assigned as NaN and filled with the median. The final data format is suitable for various machine learning algorithms.

2.3 Prediction Algorithm Selection

The modeling initially uses four corresponding machine learning algorithms commonly used in academia to predict black samples, namely XGBoost [8-10] (extreme gradient boosting machine), LightGBM [11] (light gradient boosting machine), GBDT [12] (gradient boosted decision tree) and LR [13] (logistic regression).

This paper randomly selects 65% of the samples from the wide-table data according to the proportion of black samples as modeling samples for model training, and the remaining 35% of samples as validation samples for model testing. The obtained

prediction results are shown in the Figure 1 and 2. Finally, XGBoost is selected as the machine learning algorithm of this model. Its comprehensive performance of black sample recall and precision is significantly better than other commonly used algorithms.

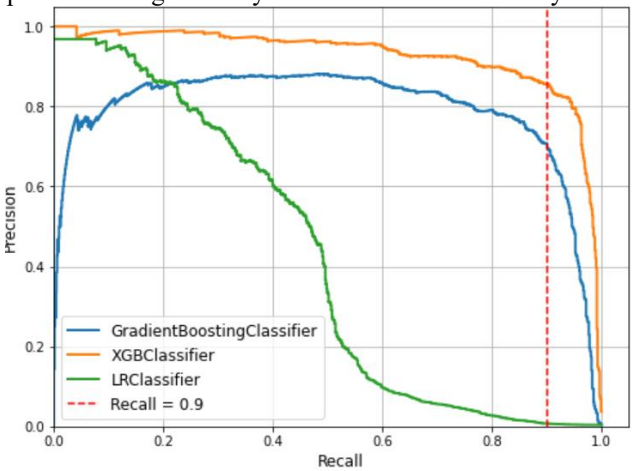

Fig. 1. PR curve

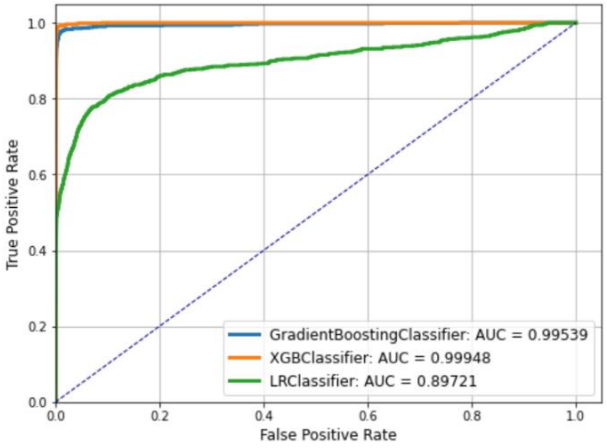

Fig. 2. ROC curve

3 Results And Discussion

In the previous chapter, we conducted an in-depth exploration of the data to find out the features suitable for modeling, and experimentally selected the most suitable model algorithm XGBoost for the final feature table. In this section, we further optimize XGBoost and analyze the validity of the recognition results of the final model.

3.1 Model Optimization

For the XGBoost algorithm, this paper uses four optimization methods to optimize the algorithm's ability to capture feature information and adjust the model over-fitting phenomenon [14].

- The model is trained by under-sampling the positive samples and adjusting the proportion of the normal samples to alleviate the imbalance between the positive and negative samples.
- The sample weight is adjusted and the scale parameter of XGBoost algorithm is set to 300 to alleviate the imbalance between positive and negative samples.
- The parameters such as learning_rate, n_estimators and max_depth were searched by parameter grid to optimize value. Table 1 shows the experimental results of the optimal hyperparameters.

Table 1. The Primary Hyperparameter Value

Parameters	*learning_rate*	*n_estimators*	*max_depth*
Value	0.05	1500	6

3.2 Model Validity Analysis

In the verification stage, more than 660,000 verification samples divided by the above are used. Among them, more than 650,000 samples are normal samples, and 2,176 samples are black samples. The ratio of normal samples and black samples is about 303:1. The positive and negative distributions of the validation samples and the modeling samples are consistent. Save the established model, predict the validation samples, and evaluate the model through PR curve, precision, recall and other indicators.

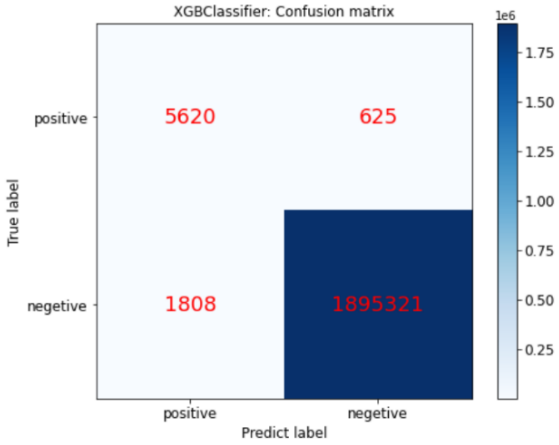

Fig. 3. Confusion Matrix

The confusion matrix shows the classification of all validation samples. The number 5620 in the upper left corner represents the real risk account and is predicted to be a risk account, the number 625 in the upper right corner represents the real risk account but is predicted to be a normal account, and the number 1808 in the lower left corner represents the real account It is a normal account but is predicted to be a risk account. The number 1895321 in the lower right corner represents a real normal account and is predicted to be a normal account.

As the confusion matrix shown in Figure 3 and the classification report in Table 2, The precision accuracy of IMGF for negative samples is 1.00 while the recall and F1-Score is 1.00 as well. This shows that our model can effectively avoid the normal account is mistakenly sealed by the system of bank.

The precision accuracy, recall accuracy, F1-Score of IMGF for positive samples is 0.76, 0.90 and 0.82. This shows that our model can detect most of the fraudulent accounts in the data set with extremely unbalanced positive and negative samples, and will have a better performance when combined with collision rules.

Table 2. Classification Report

Content	*Precision*	*Recall*	*F1-Score*	*Support*
Negetive	1.00	1.00	1.00	1895321
Positive	0.76	0.90	0.82	6245
Accuracy	-	-	1.00	1903374
Macro avg	0.88	0.95	0.91	1903374
Weighted avg	1.00	1.00	1.00	1903374

PR curve [15] is the recall rate, and the ordinate is the precision rate. The higher the precision rate and recall rate of a model, the more accurate the model and the better the effect. As shown in Figure 4 and 5, XGBoost has achieved good results in the identification of gambling and fraudulent accounts.

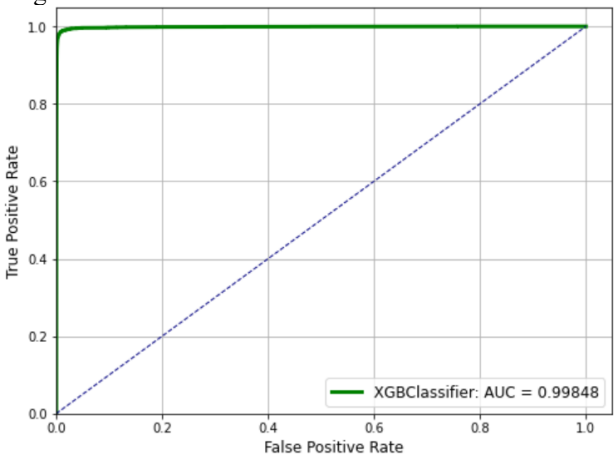

Fig. 4. Final model ROC curve

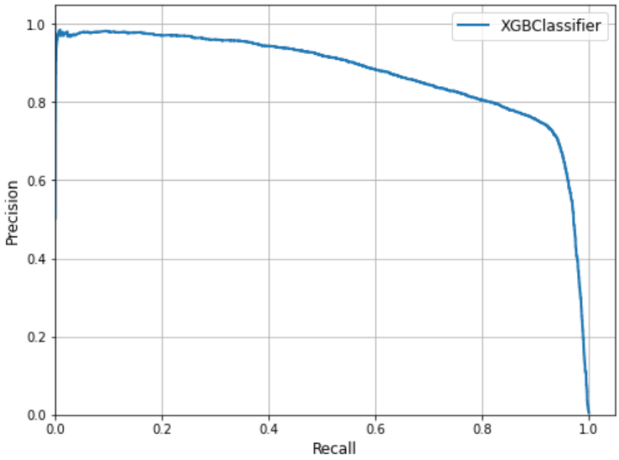

Fig. 5. Final model PR curve

4 Conclusions

This paper innovatively combines the collision rules set by banking experts with machine learning prediction algorithm and proposes identification model for gambling and fraud in bank personal settlement accounts. IMGF can effectively identify and predict the bank's personal settlement accounts involved in gambling or fraud, and at the same time, it can avoid the normal account being mistakenly closed. Provide safer and unaffected services to bank customers.

In later work, we can use more digital methods, legitimate and legitimate bank internal data reasonable application, mining its potential value. For example, you can apply more comprehensive and detailed information on this project to build a multimodal model of anti-gambling and anti-fraud.

Acknowledgment

This work was supported by Bank of Beijing Nanchang Branch with research funding and desensitized data.

References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. Kannan, Somasundaram, and K. Somasundaram. "Autoregressive-based outlier algorithm to detect money laundering activities." Journal of Money Laundering Control (2017).

3. Wei, Lu, et al. "Discovering bank risk factors from financial statements based on a new semi-supervised text mining algorithm." Accounting & Finance 59.3 (2019): 1519-1552.
4. Catherine, Denny and M. R. Shihab, "Bank Account Classification for Gambling Transactions," 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2021, pp. 302-308, doi: 10.1109/EIConCIT50028.2021.9431874.
5. Khrestina, Marina Pavlovna, et al. "Development of algorithms for searching, analyzing and detecting fraudulent activities in the financial sphere." (2017).
6. Leo, Martin, Suneel Sharma, and Koilakuntla Maddulety. "Machine learning in banking risk management: A literature review." Risks 7.1 (2019): 29.
7. Diniz, E.H., Luvizan, S.S., Hino, M.C., Ferreira, P.C., "Unveiling the big data adoption in banks: Strategizing the implementation of a new technology." Digital Technology and Organizational Change. Springer, Cham, 2018. 149-162.
8. L. Sun, "Application and Improvement of Xgboost Algorithm Based on Multiple Parameter Optimization Strategy," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020, pp.1822-1825.
9. Shimin, L. E. I., Ke, X. U., Huang, Y., & Xinye, S. H. A. "An Xgboost based system for financial fraud detection." E3S Web of Conferences. Vol. 214. EDP Sciences, 2020.
10. Song, Zijian. "A data mining based fraud detection hybrid algorithm in E-bank." 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE, 2020.
11. Khemais, Zaghdoudi, Djebali Nesrine, and Mezni Mohamed. "Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression." International Journal of Economics and Finance 8.4 (2016): 39.
12. LI, Hao, and Yan ZHU. "Xgboost algorithm optimization based on gradient distribution harmonized strategy." Journal of Computer Applications 40.6 (2020): 1633.
13. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
14. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).
15. Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. 2006.