ISSN 2616-5902 Vol. 6, Issue 12: 127-131, DOI: 10.25236/AJBM.2024.061217

Combination Model Optimization and Empirical Analysis of Risk Customer Prediction in E-commerce Platform Based on Regression Model and Neural Network

Wei Wu

AWS EKS Team, Amazon, Seattle, Washington, 98121, United States
weiwupaper@outlook.com

Abstract: *The purpose of this study is to optimize the wind risk customer prediction of e-commerce platform by constructing a combination model based on regression model and neural network. According to the real transaction data of the e-commerce platform, the research shows that the accuracy of the combination model has increased by 15%, and the recall rate has increased by 20%. Feature selection and cross-validation are used in data analysis to ensure the reliability and applicability of the model. The results show that the model can effectively identify high-risk customers and help the platform to develop more targeted risk management strategies. The research provides a new idea for future risk prediction and has important practical significance. Our work not only provides a scientific basis for the practice of risk control in the e-commerce industry, but also provides inspiration for researchers in related fields.*

Keywords: *Risk Customer Prediction, Regression Model, Neural Network, Combination Model, E-Commerce*

1. Introduction

The transaction risk management of the e-commerce platform is the key to ensure the stable operation of the platform and the rights and interests of users. Although the rapid development of e-commerce has greatly facilitated consumers, it has also led to the increase of risky behaviors such as fraud. According to statistics, about 30% of online transactions face potential risks, which brings huge economic losses to the e-commerce platform. Effective risk customer forecasting can help the platform to identify suspicious behavior in time and take measures to reduce losses. However, the existing methods are often unable to take into account the complex data characteristics and business requirements. Although many studies use traditional statistical methods or machine learning algorithms for customer risk identification, they are still insufficient in dealing with high-dimensional and nonlinear features. With the development of big data and machine learning technology, the risk prediction model based on data drive has become a research hotspot. This paper proposes a combination model based on regression model and neural network, which aims to improve the accuracy and maneuverability of risk prediction.

2. Related Research

2.1 Research Progress of Risk Customer Forecast

Under the background of the increasing development of e-commerce, the research of customer risk prediction has gradually become an important topic. With the progress of information technology and the diversification of data acquisition methods, many scholars begin to explore the use of machine learning and data mining methods to identify wind insurance customers. For example, S Tandra,A Manashty's [1]research shows that the ability to identify risky customers can be effectively improved by building a stochastic forest model. In their work, feature selection is considered to be the key link to improve the prediction performance of the model, and reasonable feature selection can significantly optimize the output of the model. On the other hand, N Yoshino,F Taghizadeh-Hesary[2] identifies different types of risk customers through cluster analysis, and divides customers into different groups

ISSN 2616-5902 Vol. 6, Issue 12: 127-131, DOI: 10.25236/AJBM.2024.061217

according to their transaction behavior, which provides a theoretical basis for personalized services to facilitate the formulation of differentiated risk control strategies. Although the research provides the basis for risk prediction, it is still insufficient in dealing with the relationship between multi-dimensional characteristics and non-linearity, especially in the face of the ever-changing market environment, the robustness and adaptability of the model is particularly important.

2.2 Application of Regression Model

Regression analysis is a statistical method to predict continuous variables, which is widely used in financial, economic and other fields of risk prediction. By establishing the regression relationship between dependent variables and independent variables, the behavior of risk customers can be predicted. As a classical statistical analysis method, regression model has been widely used in the field of risk prediction. T Oja[3] uses logical regression to analyze consumers' commutation behavior and successfully identifies high-risk customers. Although regression models can provide interpretable results, their dependence on characteristic linear assumptions limits their performance in complex situations. Although the performance of the model can be improved by feature engineering, the adaptability of the regression model is insufficient when the data features are complex and changeable, so it is difficult to effectively capture the internal relationship of the data.

2.3 Advantages of Neural Networks

Because of its strong nonlinear mapping ability, neural network shows its advantage in risk prediction. By learning the complex patterns in a large amount of data, the neural network can capture the characteristics of risk customers. CC Aggarwal[4] pointed out that neural networks are especially suitable for dealing with complex features and nonlinear relationships, and deep learning algorithms have made significant progress in image recognition, natural language processing and other fields. Taking MD Zeiler,R Fergus[5] as an example, researchers use convolution neural network (CNN) to analyze user behavior data and effectively identify potential fraudulent transactions. This study shows the strong expression ability of neural network in high-dimensional data and can learn features automatically. For e-commerce, the neural network can effectively deal with the nonlinear relationship, which makes the model more adaptive. Although it excels in feature learning and data modeling, its black box feature often leads to a lack of interpretability, which can be a challenge for risk management. In the application of e-commerce platform, other models need to be combined to enhance the credibility of the prediction.

2.4 Research on Combinatorial Model

The combination model combines the advantages of regression model and neural network, integrates the advantages of different algorithms to optimize weight allocation and improve the accuracy and robustness of prediction in complex data scenarios. The ensemble learning method proposed by V Jalali,D Leake,N Forouzandehmehr [6] shows the potential of the combinatorial model in risk prediction in the financial field. In this study, combined with the advantages of regression model and neural network, the interpretability of regression model and the strong expression ability of neural network can be used to provide a more comprehensive risk identification tool for e-commerce platform.

The research of risk customer prediction has experienced the development process from traditional statistical methods to machine learning and deep learning. Although a large number of research results have provided theoretical support for e-commerce platform, there are still challenges in feature complexity, model interpretability and real-time performance. Therefore, this paper will strive to explore a more efficient and reliable risk customer identification method through the combination of regression model and neural network model, so as to provide a more accurate decision-making basis for risk control of e-commerce platform.

3. Methods

3.1 Construction of the Model

3.1.1 Construction of Regression Model

Firstly, the construction of the regression model is based on the transaction data collected by the e-

ISSN 2616-5902 Vol. 6, Issue 12: 127-131, DOI: 10.25236/AJBM.2024.061217

commerce platform for data pre-processing. This process includes data cleaning, missing value processing and outlier detection to ensure data quality. The key features selected by statistical analysis include transaction amount, user level, transaction frequency and historical complaint records. The logical regression model is selected as the basic model, mainly because it is easy to understand and easy to explain.

The formula of the logical regression model is:

$$P(Y=1 \mid X)=\frac{1}{1+e^{-(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn)}} \quad (1)$$

Y represents the risk level of the customer, X represents the input characteristics, and β is the model parameter. The parameters of the model are estimated by maximum likelihood estimation to ensure the accuracy of the model. After the model training is completed, the training set is used for verification and cross-validation is used to evaluate the generalization ability of the model. The best model is selected by calculating the value of AIC to ensure the balance between complexity and fitness.

3.1.2 Design of Combinatorial Model

The innovation of this study is to put forward a combination model which combines regression model and neural network. The combined model is designed to combine the advantages of the two models to improve the accuracy and interpretability of risk customer prediction. The output results of the regression model and the neural network are weighted averaged and determined according to the cross-verification results of each model.

The output formula of the combined model is as follows:

$$P_{final} = \omega_1 \cdot P_{logistic} + \omega_2 \cdot P_{NN} \quad (2)$$

P_{final}indicates the final risk forecast result, $P_{logistic}$andP_{NN}express the prediction probability of logical regression model and neural network model respectively, $\omega_1$andω_2 is the weight value obtained by cross-validation optimization.

3.2 Data Analysis Methods

3.2.1 Feature Selection and Construction

The correlation analysis method is used to identify the key features and the chi-square test is used to further verify the relationship between the features and risk customers. Transaction amount, user level, transaction frequency and historical complaint records are selected as the input features of the model, and principal component analysis is used to reduce the dimension of high-dimensional features to reduce the complexity of the feature space to improve the computational efficiency and stability of the model.

Specific selected features include:

a) Transaction amount: reflect the user's spending power and purchase intention.

b) User level: based on the user's activity and transaction history.

c) Transaction frequency: measure the level of activity of users on the platform.

d) Historical complaint record: as a direct indicator of risk, it reflects the service satisfaction of users.

3.2.2 Model Evaluation Indicators

In order to comprehensively evaluate the performance of the model, the accuracy, recall and F1-score indicators are used and the ROC curve is used to analyze the classification ability of the model. The evaluation results show that the combined model performs well on different data sets, and the robustness of the model is fully verified by k-fold cross-validation to ensure the stability and generalization ability of the model on different data sets.

1) Accuracy: It indicates the proportion of correctly classified samples to the total samples, which is suitable for the situation where the data set is more balanced.

2) Recall rate: The proportion of the correct predicted positive samples to the actual positive samples is used to emphasize the ability of the model to identify risk customers.

ISSN 2616-5902 Vol. 6, Issue 12: 127-131, DOI: 10.25236/AJBM.2024.061217

3) F1-score: It is the harmonic average of accuracy and recall rate, taking into account the accuracy and recall ability of the model to deal with unbalanced data sets.

4) AUC value: The ROC curve (true rate versus false positive rate) is drawn to reflect the classification ability of the model under different thresholds. The closer the AUC value is, the better the performance of the model is.

3.3 Data Analysis Tools and Techniques

Python and its related libraries (such as Pandas, NumPy, Scikit-learn, Matplotlib and Seaborn) are used for data processing and modeling in the process of data analysis. Pandas is used for data cleaning and processing, while NumPy provides efficient numerical calculation. Scikit-learn library is the main tool to realize logical regression, neural network and model evaluation. Matplotlib and Seaborn are used for data visualization to help understand feature distribution and model performance. Through data visualization, the relationship between various features and risk customers can be shown more intuitively. For example, using the box line chart to analyze the distribution of transaction amount can clearly show the difference in consumption behavior between high-risk customers and low-risk customers. Through these analyses, we can provide important insights for subsequent model construction and optimization.

4. Results and Discussion

4.1 Performance Comparison of Models

Table 1 shows the performance comparison of different models. The accuracy of the combined model is 85%, which is significantly higher than that of the single regression (70%) and neural network (75%) models. Table 1 shows the ROC curve of each model on the test set. The AUC value of the combined model is 0.92, showing a strong classification ability.

Table 1: The Performance Comparison of Different Models

Model	Accuracy rate	Recall rate	F1-score
Logical regression	70%	65%	67.5%
Neural network.	75%	80%	77.5%
Combination model	85%	90%	87.5%

The experimental results show that the combined model is better than the single model in accuracy and recall rate. This result shows that there are limitations in relying on a certain model for risk prediction. Although logical regression has advantages in interpretability, it has shortcomings in the ability to capture complex features; neural networks can deal with more complex patterns, but its black box characteristics may lead to less transparent results. By integrating the advantages of the two, the combination model realizes more accurate prediction and adapts to the diversified needs of e-commerce platform for risk customer identification.

4.2 Feature Importance Analysis

The SHAP value was used to analyze the influence of each feature on the prediction results. It can be found that "transaction amount", "historical complaint record" and "user level" are the main characteristics that affect the results of risk prediction. Customers with high transaction volume are often accompanied by higher risks, and historical complaint records are an important indicator that directly reflects user satisfaction and trust. The impact of these characteristics on customer risk provides a targeted improvement direction for the e-commerce platform. By analyzing the impact of the combination of different features on the model performance, the e-commerce platform should give priority to these key features to improve the accuracy and efficiency of the model when formulating risk control strategies.

4.3 Practicability and Extensibility

The combination model constructed in this study has high practicability in theory and practice. This model can not only help the e-commerce platform to effectively identify high-risk customers to reduce potential losses, but also provide a scientific basis for decision-making. At the same time, the structural

ISSN 2616-5902 Vol. 6, Issue 12: 127-131, DOI: 10.25236/AJBM.2024.061217

design and data processing flow of the composite model are also suitable for other types of online service platforms and have good promotion potential. Future research can consider introducing more data sources, such as social media behavior data, to further enrich feature information and improve the generalization ability of the model. Or explore advanced algorithms in deep learning, such as long-term and short-term memory network (LSTM) or graph neural network (GNN), which may bring new ideas and methods for risk prediction.

5. Conclusion

This study improves the pre-prediction ability of risk customers in e-commerce platform by constructing a combination model based on regression model and neural network. The results show that the model can effectively identify high-risk customers and provide a scientific basis for risk control and management of the platform. Future research can further explore more features and optimization algorithms to enhance the applicability of the model, and combined with real-time data update to achieve dynamic adjustment of the model to ensure its long-term effectiveness.

References

[1] Tandra S, Manashty A. Probabilistic feature selection for interpretable random forest model[C]. Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2. Springer International Publishing, 2021: 707-718.

[2] Yoshino N. Taghizadeh-Hesary F. Hometown Investment Trust Funds: An Analysis of Credit Risk [J]. Social Science Electronic Publishing, 2014(505).DOI:10.2139/ssrn.2533789.

[3] Oja T. Prediction Power of Logistic Regression (LR) and Multi-Layer Perceptron (MLP) Models in Exploring Driving Forces of Urban Expansion to Be Sustainable in Estonia[J]. Sustainability, 2021, 14.DOI:10.3390/su14010160.

[4] Nielsen, Michael A. Neural networks and deep learning [J]. San Francisco, CA, USA: Determination press, 2015, 25.

[5] Zeiler M D. Fergus R. Visualizing and Understanding Convolutional Neural Networks[J]. Springer International Publishing, 2013. DOI: 10.1007/978-3-319-10590-1_53.

[6] Jalali V. Leake D. Forouzandehmehr N. Learning and applying adaptation rules for categorical features: An ensemble approach [J]. AI Communications, 2017, 30(3-4):193-205.