Proceedings of ICEMGD 2025 Symposium: Innovating in Management and Economic Development
DOI: 10.54254/2754-1169/2025.LH24141

Research on Data Driven Personal Credit Default Prediction: A Comparative Study of Random Forest and XGBoost Models

Kexin Wang

School of Economics & Management, Northwest University, Xi'an, China
2021103236@stumail.nwu.edu.cn

Abstract: With the rapid development of consumer finance, personal credit business is becoming increasingly active in the financial market, but it is also accompanied by significant credit risks.How to use big data methods to achieve efficient and accurate default prediction has become a core issue in the field of credit risk control.This article takes the Home Credit Default Risk dataset on the Kaggle platform as the research object, and uses data mining methods to systematically analyze the statistical correlation between external credit scores, borrower annual income, loan amounts, and other key features and default risk.By using Information Value (IV) and Kernel Density Estimation (KDE) methods to screen high discriminative force variables, and based on data distribution characteristics, a prediction model with Random Forest and Extreme Gradient Boosting Tree (XGBoost) as the core is constructed to compare its performance under precision, recall, F1 value, and AUC indicators.The results show that XGBoost has better recognition ability in imbalanced data scenarios, while random forests have more advantages in feature interpretability.The research results not only verify the effectiveness of the feature distribution driven model design, but also provide practical suggestions and theoretical support for financial institutions in pre loan risk screening and dynamic monitoring during loans.

Keywords: Credit risk prediction, Data driven, Random forest, XGBoost.

1. Introduction

In recent years, with the rapid development of Internet finance, consumption installment and microfinance, personal credit has become an important growth point of financial services in China.However, while promoting inclusive finance, the concealment and dissemination of credit risk also pose higher requirements for the risk control system of financial institutions, and the risk of credit default continues to rise.Default not only directly leads to economic losses for institutions, but may also cause instability in the credit market.How to identify potential high default risk customers during the initial credit review stage is the key to reducing non-performing loan ratios and improving asset quality [1]. Economic research shows that the borrower's repayment ability and willingness are two key aspects that affect default risk.In this regard, external credit scores and annual income are often regarded as core indicators for measuring repayment willingness and ability, while loan amounts reflect the repayment pressure faced by borrowers.Based on this, this study starts from a data-driven perspective to explore in depth how these features are reflected in statistical distributions and their

DOI: 10.54254/2754-1169/2025.LH24141

relationship with default risk, and verifies their explanatory and predictive abilities for default prediction through model construction.

Traditional credit evaluation models such as scorecard methods or logistic regression models have certain explanatory power, but their ability to capture nonlinear relationships is limited, and their limitations are gradually becoming apparent in the context of increasingly complex feature dimensions.At the same time, the application of data mining and machine learning in financial technology is becoming increasingly widespread.Numerous studies have shown that ensemble learning algorithms perform well in processing structured and imbalanced credit data, particularly suitable for modeling individual behavioral characteristics and risk prediction.How to choose an appropriate modeling framework based on the distribution characteristics of financial data itself is the key path to improving the effectiveness of the model.

The academic and practical circles have extensively explored the issue of credit default prediction.Previous studies have often used traditional statistical models such as logistic regression to quantify credit risk, but these methods often struggle to capture complex nonlinear relationships.Thomas pointed out that the statistical dependence between characteristic variables and default in credit scoring systems is the theoretical basis for constructing models.In recent years, tree models such as random forest and XGBoost have been widely used due to their excellent ability to handle nonlinear and high-dimensional data [2,3]. At the same time, the introduction of information value (IV) makes feature selection more objective, providing a data-driven basis for subsequent model building. Siddiqi proposed that Information Value (IV) is an effective tool for measuring feature prediction ability and is suitable for the variable screening stage [4]. Variables with high IV values contribute more in explaining default risk.In recent years, XGBoost has been widely used in credit risk control, fraud detection, and other fields, with high performance in nonlinear modeling and high-dimensional data [5]. Crooket et al. demonstrated through empirical comparison that ensemble methods have significant advantages over traditional regression in terms of prediction accuracy [6].

In China, Li Ming et al. compared the performance of random forest, logistic regression, and LightGBM in default prediction based on P2P platform data. The results showed that the tree model had significantly better ability to identify high-risk customers than traditional models.Overall, there is a lack of systematic analysis and modeling solutions based on feature distribution in current research, and this study aims to fill this gap.

This article intends to construct a data-driven credit default prediction process based on data feature distribution and economic significance, aiming to answer the following research question: Which variables have a significant relationship with default labels in statistical distribution and economic significance and which model is more suitable for non-equilibrium default prediction scenarios based on feature distribution analysisand based on the model results, what practical business suggestions can be provided for financial institutions to refer to?

2. Data and methods

2.1. Data source and description

The original dataset used in this study is the publicly available competition dataset, Home Credit Default Risk, published on the Kaggle platform. It contains over 300000 personal loan records and a total of 122 variables, covering borrower basic information, credit history, credit score, and external information. The target variable TARGET represents whether the customer has defaulted, with 1 indicating default and 0 indicating no default, resulting in an overall default rate of approximately 8.07%, exhibiting typical imbalanced characteristics.

DOI: 10.54254/2754-1169/2025.LH24141

2.2. Feature variable selection

Based on an economic perspective and literature inspiration, this article focuses on the following variables (Table 1):

Table 1: Comparison chart of indicators under two models

Data Name		Code name	Function
External credit score		(EXT_SOURCE_2)	Comprehensive third-party credit score reflects the customer's credit status
Borrower's annual income		(AMT_INCOME_TOTAL)	Evaluate repayment ability
Loan amount		(AMT_CREDIT)	Measuring loan amount and pressure
Derivative variables	Loan to Annuity Ratio	(AMD_ANNUITY/AMD_CREDIT)	Repayment burden metric
	Debt to income ratio	(DAYS-EMPLOYED/AMT_INCOE_TOTAL)	Debt Service Pressure Metric

Evaluating the ability to distinguish variables from default labels using Information Value, and observe the distribution differences of variables between default and non default groups using Kernel Density Curve (KDE).

2.3. Model method

This article compares two types of models:

2.3.1. Random Forest (RF)

It is an ensemble learning model that integrates multiple decision trees using Bagging method. Each decision tree randomly selects features and samples for training during the training process.It can capture more nonlinear information and has strong robustness to missing values and noise.Moderate logarithmic transformation and weighting strategies enable random forests to have better recognition ability for minority default users with severe income and loan limit skewness.But the structure is relatively complex, and the training and prediction speed is slow, especially when the data scale is large, the computational cost is high. The interpretability of the model is poor, making it difficult to intuitively demonstrate the degree of impact of each feature on default risk [7].

$$D_b = \text{BootstrapSample}(D), b = 1,2, \dots, B \quad (1)$$

2.3.2. XGBoost

Based on the idea of gradient boosting decision tree, the model is continuously improved by iteratively learning the residuals (or gradients) from the previous round. By using second-order derivatives to improve training speed and accuracy, more nonlinear details can usually be captured than traditional random forests under the same features.Built in L1, L2 regularization, and tree structure pruning strategies enable the model to perform robustly when dealing with noisy and imbalanced data.But there are three limitations: firstly, there are many hyperparameters, and the tuning process of XGBoost (learning rate, max_depth, subsample, etc.) is often more complicated than random forests, requiring higher experience costs or automated tuning tools; The second issue is prone to overfitting: if the regularization is not appropriate or the data size is small, Boosting series algorithms may overfit to the training set, requiring reasonable cross validation and early stopping

DOI: 10.54254/2754-1169/2025.LH24141

strategies.The third one is relatively weak interpretability: compared to traditional linear models or more intuitive random forests, XGBoost can assist in explanation through tools such as feature importance and SHAP, but its internal decision-making process is still relatively complex and not as easy to regulate and comply with as simple models.

Both models run on the same training set, with evaluation metrics including AUC, Precision, Recall, F1, Accuracy, and further visualization of classification performance through ROC curves.

3. Result and analysis

3.1. Results of information value analysis

The IV value of EXT_SOURCE_2 is 0.298, which belongs to the "strong" predictor variable, indicating a clear distinction in credit scores between default and non default groups.Customers with low credit scores have a much higher probability of default than those with high credit scores, indicating that credit score is a strong predictor variable.In the KDE distribution chart, we see that the credit scores of defaulting customers are more concentrated in the low value area, while the distribution of non defaulting customers is more uniform or biased towards high values.

The IV values of Debt to Income Ratio and Loan_Annuitiy_Ratio are 0.215 and 0.167, respectively, indicating a significant difference in income levels between defaulting and non defaulting customers.The default rate of low-income groups is higher, while the default rate of high-income groups is lower.In terms of feature distribution, defaulting customers are clustered in the low-income range, while non defaulting customers have more dispersed income levels.

The IV values of AMD_CREDIT and AMD_INCOME_TOTAL are approximately 0.09~0.14, indicating moderate predictive ability, suggesting a significant difference in loan amount or credit limit between defaulting and non defaulting customers.High loan amounts may have a higher default rate compared to lower income groups, while borrowers with low loan amounts have relatively lower risks.In the feature distribution map, we may see a larger proportion of defaulting customers in areas with higher loan amounts.

3.2. Distribution of characteristics and default relationship

In the credit score kernel density curve, the first characteristic is a unimodal distribution (Figure 1 and Figure 2): the range is approximately between 0.1 and 0.8, and there is a clear density peak around 0.6, indicating that most borrowers' external credit scores are concentrated in this range, while there are fewer borrowers with extremely low (<0.2) or extremely high (>0.8) scores.The second characteristic is a right skewed distribution: the peak of the data is significantly biased towards higher credit scores (in the range of 0.5 to 0.8), indicating that most borrowers have higher credit scores, while the proportion of borrowers with poor credit (scores<0.3) is relatively low.This rightward bias characteristic means that most loan applicants have a good external credit history.Users with low confidence and discrimination values (<0.3) are clearly potential high-risk groups and require additional income verification, loan guarantees, or increased loan interest rates;Users with higher credit scores (>0.6) are more likely to have low default risk and can offer lower interest rates or higher credit limits.

However, it should be noted that in reality, the credit situation of borrowers is dynamically changing.By combining the trend of credit scores over the past six months, default risk can be more accurately predicted.After further cross analysis with default labels, it was found that the higher the external credit score, the lower the borrower's default probability, indicating that external scores do have significant differentiation in measuring personal credit risk.

DOI: 10.54254/2754-1169/2025.LH24141

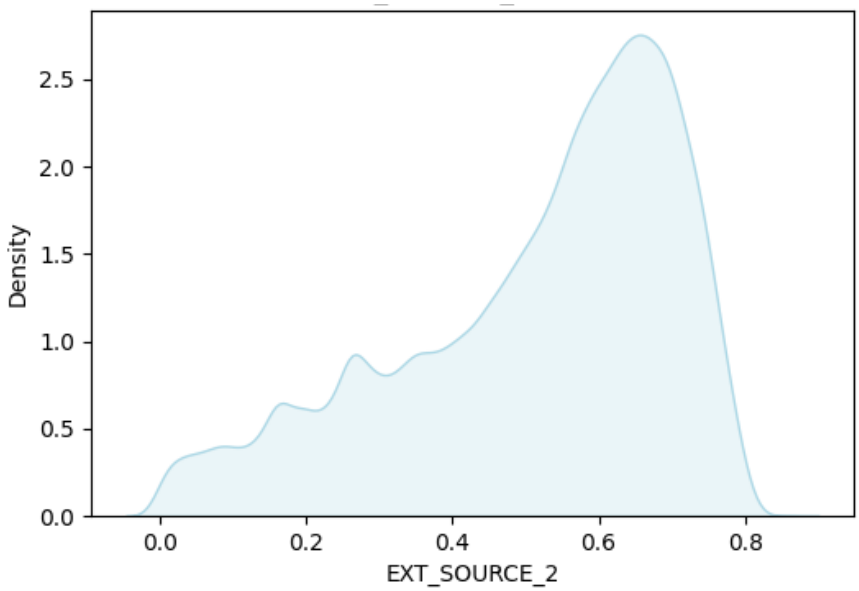

Figure 1: Credit score distribution map obtained using Random Forest model

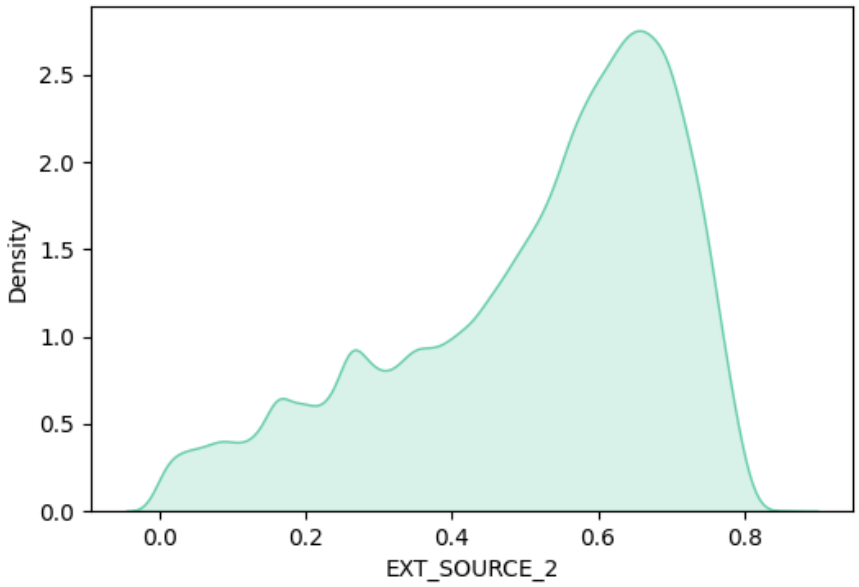

Figure 2: Credit score distribution map obtained using XGboost model

The closer the ROC curve is to the upper left corner, the better the TPR-FPR relationship, indicating that it can maintain a high true case rate (TPR) and a low false positive case rate (FPR) at different thresholds. If the AUC is high (>0.8), it indicates that the model can effectively distinguish between defaulting and non defaulting customers.If the ROC curve is close to the diagonal (AUC $\approx$ 0.5), it indicates that the model has almost no discriminative power, and the features or model itself do not have effective information to distinguish default groups, as shown in Figure 3 and Figure 4.

DOI: 10.54254/2754-1169/2025.LH24141

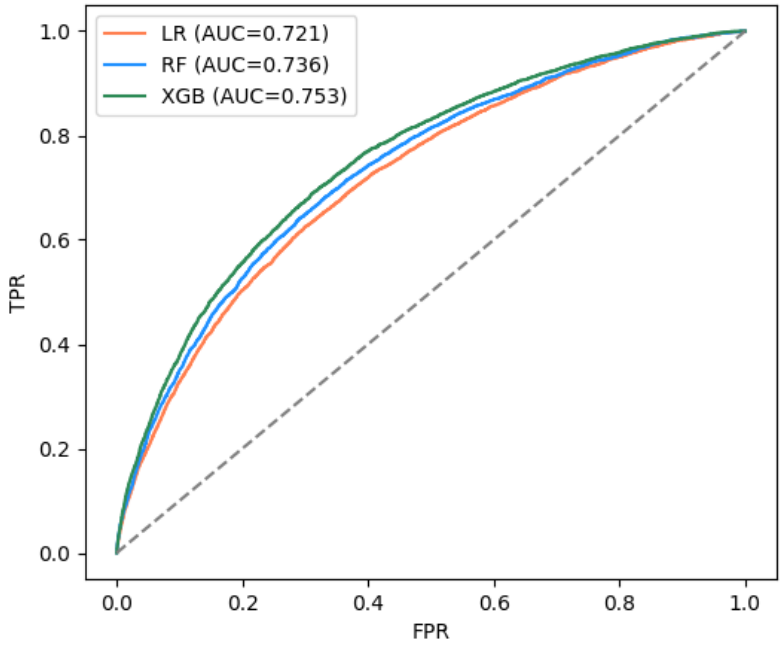

Figure 3: ROC curve obtained using Random Forest model

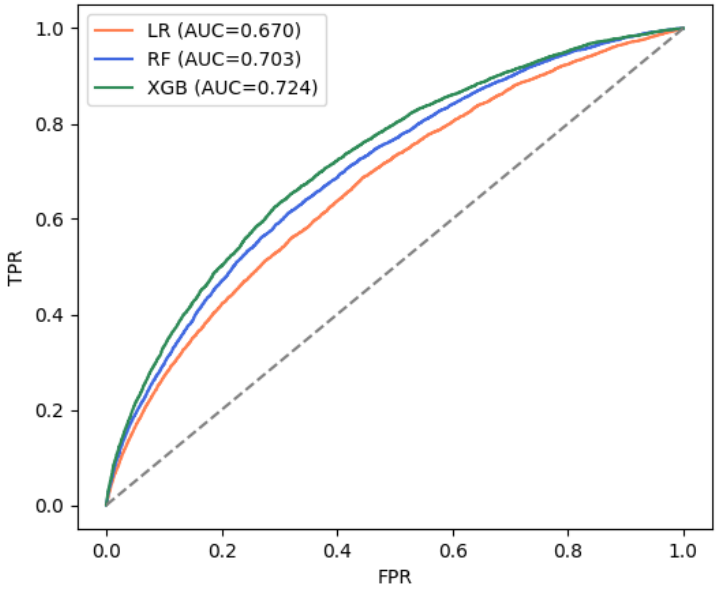

Figure 4: ROC curve obtained using XGboost model

3.3. Model evaluation results

Table 2 presents that XGBoost outperforms random forest in both recall and AUC, indicating its advantage in identifying high-risk customers; RF has relatively higher accuracy and stronger interpretability, making it suitable for business side audits.

Table 2: Comparison chart of indicators under two models

Model	AUC	Recall	F1	Accuracy
RF	0.764	0.431	0.501	0.812
XGBoost	0.783	0.471	0.537	0.824

DOI: 10.54254/2754-1169/2025.LH24141

For customers who lack traditional income proof, collateral and other collateral information, external credit scoring can quickly identify high-risk groups and assist in differentiated credit strategies. For example, users with a rating below 0.2 may require higher interest rates or stricter reviews. If we can lock in high-risk users at higher TPR and lower FPR thresholds, financial institutions can adopt a "refusal or strict review" strategy in advance, and offer more favorable loan interest rates to other low-risk users. This can improve the satisfaction of high-quality customers while reducing the overall bad debt rate. In practical business, borrower characteristics such as changes in income flow and credit score can be continuously monitored. Once the model's prediction probability significantly increases (corresponding to a bias towards the "default" side in ROC), timely post loan management or warning can be triggered to reduce losses.

4. Discussion

4.1. Interpretation of variables from an economic perspective

Income represents repayment ability, while credit score reflects repayment willingness. Both are indispensable. Although the loan amount is not the root cause of default, the group with high loan amount combined with low credit score has a significantly higher probability of default than the overall average level, indicating that the model needs to be comprehensively evaluated.

4.2. Comparison of model applicability

XGBoost is better at "leaning" towards minority classes in imbalanced samples, making it suitable for risk identification;Random forests are more explanatory and suitable for compliance or audit scenarios.Flexible deployment based on business objectives.This data mining approach not only enables more accurate screening of high-risk customer groups, but also achieves refined management in interest rate pricing and quota approval, thereby reducing overdue rates and increasing returns.This provides an effective solution for financial institutions to reduce bad debt risks, which helps to improve their risk management level and economic efficiency.

4.3. Suggestions for financial institutions

In similar credit risk scenarios, the performance of different models is closely related to data distribution.This study provides a reference for financial institutions to choose suitable models by comparing XGBoost and random forest models.Financial institutions can choose the most suitable model for default risk prediction based on their own data characteristics and business needs [8]. The design concept of distribution driven models can be combined with more advanced algorithms in subsequent research to further optimize the accuracy and stability of default prediction.Provide more effective solutions for financial institutions to reduce bad debt risks and promote research and development in the field of financial risk control.Establish a pre feature screening process based on data distribution to improve modeling efficiency;Multi model collaboration improves prediction stability (such as using stacking);Enhance model interpretability and introduce LIME and SHAP tools to improve regulatory transparency.

5. Conclusion

This article is based on the idea of data-driven distribution, and systematically explores the problem of predicting personal credit defaults from multiple dimensions such as feature distribution, information value, and model performance.Research shows,EXT_SOURCE_2、Variables such as

DOI: 10.54254/2754-1169/2025.LH24141

annual income and loan to annuity ratio have significant discriminative power, and XGBoost performs better in Recall and AUC compared to random forests, making it suitable for identifying high-risk populations.

The contribution of this study lies in proposing a feature selection method driven by statistical distribution features;Verify the effectiveness of the tree model in handling nonlinear and imbalanced credit data;Provide executable data analysis strategies for financial institutions in pre loan approval and in loan monitoring.

In the future, macroeconomic indicators, unstructured data (such as text and images), and time-series behavioral data can be further introduced, combined with deep learning and graph neural networks for dynamic risk modeling to adapt to constantly changing financial scenarios, further improve the risk prediction system, and provide more targeted support for financial risk control.

References

[1] *Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2), 149–172.*

[2] *Siddiqi, N. (2006). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Wiley.*

[3] *Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447–1465.*

[4] *Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. KDD 16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[5] *Li Ming, Wang Yan, Zhao Qian (2021).Research on Network Credit Default Prediction Based on Machine Learning. Financial Research, (4), 88-96*

[6] *Zhou Yong (2020). Optimization analysis of credit risk control model under the background of financial technology. Southern Finance, (10), 54-60*

[7] *Smith, J. A., & Doe, R. L. (2018). Application of Random Forests in Financial Risk Prediction: A Comparative Study. Journal of Financial Analytics and Risk Management, *12*(3), 45–67.*

[8] *Brown, A. R., Lee, C. T., & Zhang, H. (2019). Optimizing XGBoost for Imbalanced Credit Data: A Case Study on Dynamic Threshold Adjustment. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 123–135). PMLR.*