Comparative Study of Personal Credit Default Risk Prediction Based on Different Machine Learning Models

Haonan Li*

School of Statistics and Data Science, Capital University of Economics and Business, 100070 Beijing, China

Abstract. The accurate prediction of individual credit default risk is crucial for maintaining the stability of the financial system. This study aims to establish a systematic comparative framework to evaluate the performance of five representative machine learning models in the task of predicting individual credit default risk. The research first preprocesses a public dataset containing 150,000 customer records and applies SMOTE technology to address the inherent class imbalance problem. Subsequently, the five models are trained under the same benchmark and evaluated based on multiple dimensions including accuracy, precision, recall, F1-Score, and AUC-ROC. The experimental results indicate that ensemble learning models (especially random forests and LightGBM) significantly outperform traditional models. Among them, random forests show the best performance in recall and AUC-ROC, effectively identifying potential defaulting customers; while LightGBM achieves the optimal results in F1-Score and precision. Additionally, feature importance analysis reveals that the utilization rate of unsecured credit limits and the number of family members are key factors affecting credit default risk. The conclusions of this study provide valuable guidance for financial institutions in selecting the most appropriate credit risk prediction model based on different business needs, and offer practical references for future related research.

1 Introduction

In an increasingly digital financial ecosystem, personal credit has emerged as a key engine driving economic growth and meeting individual consumption needs. However, the expansion of credit activity is also accompanied by a growing risk of default, which not only directly threatens the asset health of banks and financial institutions, but may also trigger systemic financial instability. Therefore, building an efficient and accurate credit default risk prediction model to effectively identify high-risk customers in the pre-loan stage has become the core challenge of modern financial risk management.

Early credit risk assessment mainly relied on traditional statistical methods such as credit score cards and logistic regression models [1]. Despite their superior interpretability, these

* Corresponding author: lihaonan@arizona.edu

methods are inadequate at capturing complex nonlinear relationships in data and handling massive multidimensional features. In recent years, with the rise of artificial intelligence technology, machine learning models have been widely used in the field of credit risk control due to their powerful data processing and pattern recognition capabilities.

For example, a study by Chen et al. compared the performance of decision trees and logistic regression in predicting corporate financial distress and found that decision trees are better in short-term forecasting, while logistic regression is more advantageous in long-term forecasting [2]. Furthermore, in addition, Zheng et al.'s research shows that random forests, as a powerful nonlinear ensemble model, have been successfully applied to stock price trend prediction with an accuracy rate of up to 85% to 95% [3]. In corporate financial risk assessment, the models built by Qin et al. based on the eXtreme Gradient Boosting (XGBoost) algorithm have been proven to have high reliability, with prediction errors within 3%. Equally [4], Wang et al.'s research found that LightGBM, as an efficient gradient lifting algorithm, has been proven to be superior to other traditional machine learning algorithms in enterprise financing risk prediction tasks [5]. Together, the conclusions of these literatures confirm the effectiveness and superiority of the machine learning models in the field of financial risk prediction.

However, there are two major limitations in existing research: first, most studies focus on the application of a single model or a few models, and lack of comprehensive and systematic performance comparison under a unified benchmark, which makes it difficult for financial practitioners to make the optimal choice among many models. Secondly, the inherent category imbalance of credit data (i.e., there are far fewer defaulting customers than non-defaulting customers) poses a serious challenge to model training, which will lead to excessive bias towards the majority class if not handled properly, thus seriously affecting the ability to identify minority classes (defaulting customers).

To fill the above research gaps, this study aims to provide a comprehensive comparative framework to systematically evaluate the performance of five representative machine learning models including logistic regression, decision tree, random forest, XGBoost, and LightGBM in the personal credit default risk prediction task. Core contributions include:

1. Rigorous Data Processing and Methodology: This study employed Synthetic Minority Oversampling Technique (SMOTE)technology to balance the training set to ensure that the model can effectively learn a few categories of features. This study also ensures the reliability of experimental results and the generalization ability of the model by strictly dividing the dataset into training, validation, and testing sets.
2. Multi-dimensional performance evaluation: In addition to traditional accuracy, This study also paid special attention to and compared the model's accuracy, recall, F1-Score, and Area Under the Curve -Receiver Operating Characteristic (AUC-ROC), which are indicators that can more objectively reflect the actual performance of the model when dealing with imbalanced data.
3. In-depth insights into models and features: This study not only compare the performance of different models but also reveal key factors affecting personal credit default risk by analyzing the feature importance of Random Forest and LightGBM, providing valuable insights into the model's interpretability and practical application.

2 Method

This part details the datasets, data preprocessing processes, machine learning models, and performance evaluation metrics used in this study for the task of predicting personal credit default risk.

2.1 Dataset Preparation

This study used a publicly available dataset for personal credit risk prediction, which initially originated from a Kaggle competition project [6]. The dataset contains 150,000 customer records and 11 variables. One of the variables is the target variable, SeriousDlqin2yrs, which is a binary variable indicating whether a customer experiences serious delinquency in the next two years (i.e. 1 indicates default, 0 indicates no default). The remaining 10 variables are feature variables that cover the customer's personal financial status and credit behavior, such as revolving credit utilization, age, monthly income, and debt ratio. A characteristic of this dataset is the significant class imbalance problem, where the number of defaulting customers (positive samples) is much smaller than that of non-defaulting customers (negative samples). Before model training, necessary preprocessing was performed on the raw data. First, to address the missing values in the dataset, this study employed median imputation to handle the missing values for variables such as monthly income and family size. Second, to handle outliers, records with ages under 18 or over 100 were replaced with the median, and records with more than 20 delinquencies were capped at 20. In addition, as models are sensitive to the scale and type of data, numeric features were normalized to achieve a normal distribution, and categorical data underwent conversion from string to numerical format. To address the inherent class imbalance in the dataset, this study applied the SMOTE technique only to the training set after data splitting to achieve data balance. SMOTE synthesizes new samples between minority class samples, making the number of positive and negative samples in the training set closer together, ensuring that the model can fairly learn the characteristics of all classes during training and avoid overfitting to the majority class. Ultimately, the dataset was divided into a training set (60%), a validation set (20%), and a test set (20%) to ensure the objectivity and rigor of model evaluation.

2.2 Machine Learning Models

To achieve a comparative study of different machine learning models, this research has selected five representative models for experimentation. These models include linear models, single decision models, and ensemble learning models, all implemented using machine learning libraries such as scikit-learn and lightgbm.

2.2.1 Logistic Regression

LR, commonly employed to solve binary classification problems [7]. This model uses a Sigmoid function to map the results of linear regression to a probability range between 0 and 1, thereby predicting the probability of an event occurring. The advantages of logistic regression lie in its simplicity, strong interpretability, and clear coefficient representation of each feature's contribution. However, the main drawback of this model is its linear assumption regarding predictor variables, which limits its effectiveness in handling nonlinear relationships and complex problems. However, logistic regression proves to be a highly effective prediction and diagnostic tool in many areas, especially when addressing relatively fewer complex issues, due to its computational efficiency and clarity of results.

2.2.2 Decision Tree

DC is a classic machine learning method that possesses both classification and prediction capabilities [8]. This algorithm generates a tree-structured model by inductively reasoning based on the given data. DC models are easy to understand and interpret, capable of handling datasets that include both numerical and categorical data. However, decision trees also have

the drawback of being prone to overfitting, meaning that when the tree becomes too deep, the model may excessively memorize noise in the training data, resulting in reduced generalization ability on new data.

2.2.3 Random Forest

The principle of random forest primarily relies on two core concepts: random sampling and majority voting [9]. When training each decision tree, random forest randomly selects a subset of samples from the entire sample set as the training set, and at each node split, randomly selects a portion of features to find the best split. This randomness endows random forest with a strong ability to prevent overfitting. Ultimately, in classification tasks, the model aggregates the predictions from all decision trees and applies the majority voting principle to determine the final classification result. Due to its excellent performance, the random forest model has achieved significant accuracy in tasks such as predicting stock price trends. Additionally, this model can also output feature importance, providing a basis for analyzing which features have the greatest impact on prediction results.

2.2.4 XGBoost

XGBoost is a mature and widely applied ensemble learning algorithm that has demonstrated outstanding classification and regression performance in various competitions and real engineering applications [10]. The core idea of XGBoost is to integrate multiple decision trees, summing the predicted results of each tree as the final prediction value. Unlike traditional gradient boosting algorithms (GBDT), XGBoost's loss function uses second-order derivative calculations and incorporates a regularization term into the objective function to effectively control the model's complexity and avoid overfitting. Therefore, XGBoost is better at balancing algorithmic computational complexity and overfitting issues when dealing with complex, nonlinear corporate financial data.

2.2.5 LightGBM

As an advanced machine learning algorithm, it typically outperforms other algorithms in terms of prediction accuracy [11]. It improves computational power and accuracy primarily through a histogram-based algorithm and a depth-limited Leaf-wise strategy. The histogram algorithm drastically reduces computational complexity and time by converting continuous feature values into discrete integers. LightGBM has proven effective in tasks like corporate financing risk prediction, where it has outperformed other algorithms such as k-nearest neighbors, decision trees, and random forests. Similar to random forests, LightGBM also includes a feature importance function that ranks features based on their split gain or usage within the model.

2.3 Evaluation Metrics

Because of the inherent class imbalance in this study's data, using traditional accuracy alone is not enough to comprehensively reflect a model's performance. Therefore, this study utilizes a more thorough set of evaluation metrics, all calculated from the four core components of a confusion matrix (True Positive, False Positive, True Negative, False Negative). The metrics selected for evaluation include Precision, Recall, F1-Score, Accuracy, and AUC-ROC.

3 Results and Discussion

3.1 Model Performance Evaluation Results

To ensure the fairness and reliability of the model evaluation, all models were trained using a balanced training set processed with SMOTE technology and were evaluated for their final performance on an independent test set. The performance metrics of each model are shown in Table 1.

Table 1. Final Performance Comparison of All Models on the Test Set

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
LR	0.725733	0.142151	0.616459	0.231028	0.737228
DT	0.858133	0.238195	0.510723	0.324873	0.795279
RF	0.814933	0.218358	0.685786	0.331245	0.839110
XGBoost	0.875533	0.265018	0.486284	0.343068	0.818237
LightGBM	0.872467	0.268497	0.526683	0.355675	0.831177

As seen in Table 1, there are significant differences in the performance of different models across various indicators. The Random Forest model achieved the highest recall rate (0.6858) and AUC-ROC score (0.8391). In contrast, the LightGBM model performed best in precision (0.2685) and F1-Score (0.3557). The XGBoost model leads with an accuracy of 0.8755, but its recall rate is relatively low (0.4863). The Decision Tree model has an F1-Score (0.3249) that falls between Logistic Regression (0.2310) and Ensemble Learning models, with a slightly decreased recall rate (0.5107). Logistic Regression, as the baseline model, performed the worst across all metrics.

3.2 Experiment Result Analysis and Discussion

This section will conduct an in-depth analysis of the aforementioned experimental results, explaining the reasons for performance differences among different models and their implications in practical applications, along with an exploration of the internal mechanisms of the models.

3.2.1 Analysis of Model Performance Differences

Logistic Regression, as a linear model, has its predictive power limited by the linear relationships in the data. While it provides an interpretable benchmark for credit risk prediction, its lower accuracy and F1-Score indicate that relying solely on linearly combined features is insufficient to capture the complex non-linear patterns of credit default behavior. The Decision Tree outperformed Logistic Regression by non-linearly partitioning the data space, but the decline in its recall rate may be due to the model's depth limitation to avoid overfitting, making it relatively conservative in identifying minority class samples.

The exceptional performance of the three ensemble learning models (Random Forest, XGBoost, and LightGBM) validates the effectiveness of ensemble methods. They significantly enhance the model's generalization ability and predictive accuracy by combining the prediction results of multiple weak learners (decision trees). Random Forest: Its highest recall rate and AUC-ROC score are particularly prominent. In credit risk prediction, the cost of missing a defaulting customer (false negative) is typically much higher than the cost of falsely reporting a non-defaulting customer (false positive). Therefore,

Random Forest shows the strongest ability to effectively identify potential defaulters. Its randomness (random sampling of samples and features) allows it to effectively prevent overfitting and capture a wider range of data patterns. Gradient Boosting Models (XGBoost and LightGBM): These two models perform excellently in accuracy, precision, and F1-Score. They progressively optimize the model by iteratively correcting the errors of the previous weak learner, thus achieving higher predictive accuracy. LightGBM slightly surpasses in precision and F1-Score, indicating that it performs better in ensuring the accuracy of predicting "default," with a relatively low false positive rate. This is an extremely attractive choice for financial institutions aiming to balance risk control and customer experience.

3.2.2 Feature Importance Analysis and Insights

To further explore the internal mechanisms of model decision-making, this study conducted feature importance analysis on the Random Forest and LightGBM models shown in Fig. 1 and Fig. 2.

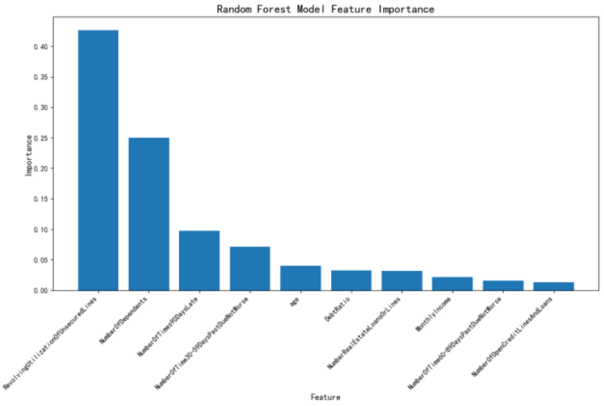

Fig. 1. Random Forest Model Feature Importance Analysis Figure (Picture credit: Original)

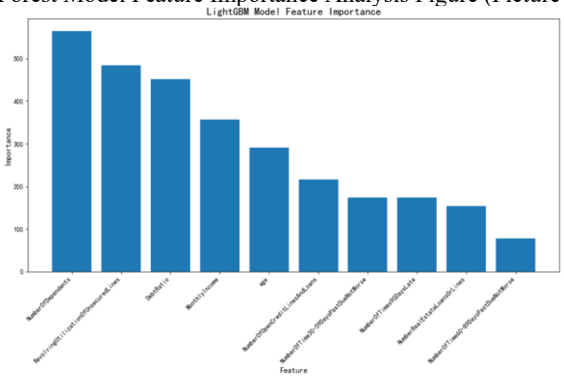

Fig. 2. LightGBM Model Feature Importance Analysis (Picture credit: Original)

By analyzing feature importance, it can be found that: the feature called revolving utilization of Unsecured Lines ranked first in importance within the random forest model. It reflects the degree to which customers depend on revolving credit. A higher utilization rate typically indicates that a customer's financial situation is tight, lacking sufficient liquidity, and therefore poses a higher risk of default. This aligns with intuitions in credit risk management.

However, in the LightGBM model, the feature called number of dependents is considered the most important. The number of family members often directly impacts a person's

disposable income and financial burden. The heavier the family burden, the more likely customers are to encounter overdue or default behaviors when facing financial pressure.

Historical delinquency records, such as the features NumberOfTimes90DaysLate and NumberOfTime30-59DaysPastDueNotWorse, hold significant positions in both models.This indicates that a customer's past repayment behavior is a strong indicator for predicting future credit risk. These analytical results provide important business insights for understanding personal credit default risk and validate that the patterns learned by the proposed models align with common sense in the financial field. This not only enhances the model's interpretability but also provides a solid foundation for its application in real business.

4 Conclusion

This study provides a systematic framework for predicting individual credit default risk by comparing various machine learning models. Experimental results indicate that ensemble learning models such as Random Forest, XGBoost, and LightGBM show significant superiority over traditional models like LR and DT, as they can better handle complex nonlinear relationships in the data. Specifically, Random Forest performs best in terms of recall rate and AUC-ROC score, excelling at identifying potential default customers; while LightGBM achieves optimal results in F1-Score and precision, demonstrating excellence in balancing prediction accuracy and reducing false positives. Furthermore, through feature importance analysis, it can be found that the utilization rate of unsecured credit limits, the number of family members, and historical overdue records are core factors in predicting credit risk.

Despite the promising results, this study has its limitations. The models were validated on a single public dataset, and comprehensive hyperparameter tuning was not performed, leaving room for further improvement.

Future research could focus on validating the models on more diverse datasets and exploring more complex models, such as deep learning. Additionally, detailed optimization of the models' hyperparameters could further enhance their predictive potential.The findings provide a quantitative basis for model selection in credit risk management and indicate directions for future research.

References

1. C. Bolton. Logistic Regression and its application in credit scoring (2009).
2. M. Chen. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. Expert Systems With Applications, 38(9), 11261–11272 (2011).
3. J. Zheng, D. Xin, Q. Cheng, M. Tian and L. Yang. The Random Forest model for analyzing and forecasting the US stock market in the context of Smart Finance. arXiv.org (2024).
4. R. Qin. The construction of Corporate Financial Management risk model based on XGBOOST Algorithm. Journal of Mathematics, 2022(1) (2022).
5. D. Wang, L. Li and D. Zhao. Corporate finance risk prediction based on LightGBM. Information Sciences, 602, 259–268 (2022).
6. Kaggle. Give Me Some Credit https://www.kaggle.com/c/GiveMeSomeCredit (2011).
7. F. S. de Menezes, G. R. Liska, M. A. Cirillo and M. J. Vivanco. Data classification with binary response through the Boosting algorithm and logistic regression. Expert Systems With Applications, 69, 62–73 (2016).

8. Y. Y. Song and Y. Lu. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry (2015).
9. J. Wang, X. Sun, Q. Cheng and Q. Cui. An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting. The Science of the Total Environment, 762, 143099 (2021).
10. J. Wu, L. Kong, M. Yi, Q. Chen, Z. Cheng, H. Zuo and Y. Yang. Prediction and Screening Model for Products Based on Fusion Regression and XGBoost Classification. Computational Intelligence and Neuroscience, 2022, 1–14 (2022).
11. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, ... and T. Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30 (2017).