Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Comparative Analysis Research on Machine Learning Models in Credit Risk Assessment

Bohan Zhang *

School of economic and management, Beijing Jiaotong University, Beijing, China

* Corresponding Author Email: 22711117@bjtu.edu.cn

Abstract. Credit risk assessment is crucial for the risk management and control of financial institutions, but it faces challenges such as sample imbalance, complex characteristics and the lack of model interpretability. In this study, two public datasets, "Give Me Some Credit" and "Loan Default", were used. The Synthetic Minority Over-Sampling Technique (SMOTE) was employed to balance the sample distribution and conduct feature engineering. Construct new features such as the income-debt ratio (Income_Debt_Ratio) to reduce variable redundancy. Meanwhile, by comparing the model's different performance among logistic regression, Random Forest (RF), the study improves the training efficiency. The experiment results depict that the integrated models (XGBoost, LightGBM) perform better on both datasets, with an average accuracy rate of 94% and an AUC value of 0.98 compared with the traditional models. Furthermore, SHapley Additive exPlanations (SHAP) values were used to develop the interpretability analysis. This study provides credit institutions with a high-precision and interpretable model construction scheme, and verifies the generalization ability of the model through cross-datasets, laying a theoretical and practical foundation for future credit risk control and the construction of an integrated system.

Keywords: Credit Risk Assessment; SMOTE; Machine learning; SHAP Interpretability; Cross-Dataset Validation.

1. Introduction

Credit risk assessment is the key for financial institutions to reduce defaults. In the financial market, one of the core tasks of risk management for financial institutions is to use quantitative methods and empirical theories to analyze the credit actions of customers, assess their default probability, and initially confirm and reduce potential losses [1]. Traditional linear models (such as logistic regression) lack the ability to capture nonlinear relationships in multi-dimensional data. Therefore, with the rapid development of fintech, continuously evolving and improving machine learning and deep learning methods are gradually replacing traditional modeling methods [2]. In addition, the complexity of credit data (such as the unbalanced distribution of positive and negative samples, redundancy of characteristic variables, and noise interference) poses severe challenges to the generalization and robustness of the model [3]. For example, the proportion of default samples in actual scenarios is often less than 10%. This fact will cause the model to tend to predict the majority class, seriously weakening the risk identification ability [4]. In recent years, the academic community has proposed a variety of solutions for credit risk assessment. For the problem of uneven distribution of positive and negative samples, Chawla et al. proposed to use of the Synthetic Minority Over-sampling Technique to handle the problem of severe data imbalance through synthetic sample data [5]. In the fields of machine learning and model optimization, Chen et al. studied and developed the XGBoost algorithm [6]. This method based on the gradient boosting framework is an optimized decision tree ensemble model, which has a better ability to capture nonlinear features than traditional logistic regression, and also solves the problem of excessively high time complexity of decision tree models. Regarding the explainability analysis, Lundberg et al. introduced the SHapley Additive exPlanations values and used the game theory method to explain the modeling results, revealing the specific impact of characteristic variables on the default risk in the model [7]. However, the above-mentioned research has solved some problems, but still has limitations. For example, it has not verified the

generalization ability of the model across datasets, and there is a lack of systematic comparison of multiple models, cross-comparison and in-depth analysis of characteristic variables [8].

Therefore, this study attempts to establish a relatively comprehensive credit risk assessment framework, comprehensively considering three key elements: Firstly, based on the method adopted by Hlongwane et al. [9], a multi-model comparison system is established; Secondly, referring to the consideration of the model's migration ability by Dudovskiy et al. [10], the generalization ability verification of the model across datasets is introduced; Finally, combined with the principles of transparency and interpretability emphasized by Bücker et al. [11], an interpretability analysis of the model was added. The research will be carried out in the following way. From data preprocessing and feature engineering at the data level, to model visualization display and model comparison at the model level, and then to the interpretation level, the nonlinear impact of key variables (such as debt ratio and credit limit utilization rate) is quantified based on SHAP values.

2. Data sets and Methods

2.1. Give ME Some Credit Dataset

Initially, the data set used in the study was sourced from Kaggle, an open data platform. The data set includes 201, 198 records, which are divided into 119,898 training sets and 81,300 test sets. Each data sample contains 12 fields, including one data number (ID), 10 feature variables, and one target variable. Overall, the dataset contains the required elements for credit risk analysis and machine learning, which can be used as a baseline dataset for data analysis and model training. The related data fields are described in Table 1.

Table 1. Variable description

Variable name	Description
Serious Dlqin2yrs	Whether the lender will experience serious delinquency in the next two years (1= yes, 0= no)
Revolving Utilization of Unsecured Lines	Unsecured revolving credit Line utilization rate (total credit card and personal credit lines ÷ total credit limits)
age	Age of the recipient of credit
Number of Time30-59Days Past Due Not Worse	The number of times the borrower has been slightly past due in the last 30-59 days
Debt Ratio	Debt ratio (monthly debt payments ÷ monthly gross income)
Monthly Income	Monthly income of borrower
Number of Open Credit Lines And Loans	Number of outstanding loans and lines of credit
Number of Times 90 Days Late	The number of overdue dates of 90 days or more
Number Real Estate Loans or Lines	Number of mortgages and real estate loans (including home equity lines of credit)
Number of Time60-89Days Past Due Not Worse	The number of times the borrower has been slightly past due in the last 60-89 days
Number of Dependents	Number of people supported by the household (excluding the borrower)

As shown in Figure 1, for partial data in the data set, this study uses t-SNE method to reduce the dimensionality of the original data sample, displays the distribution of the data by visualization method, and uses logistics regression to build a decision interface to preliminarily determine the model method and training ideas of the data.

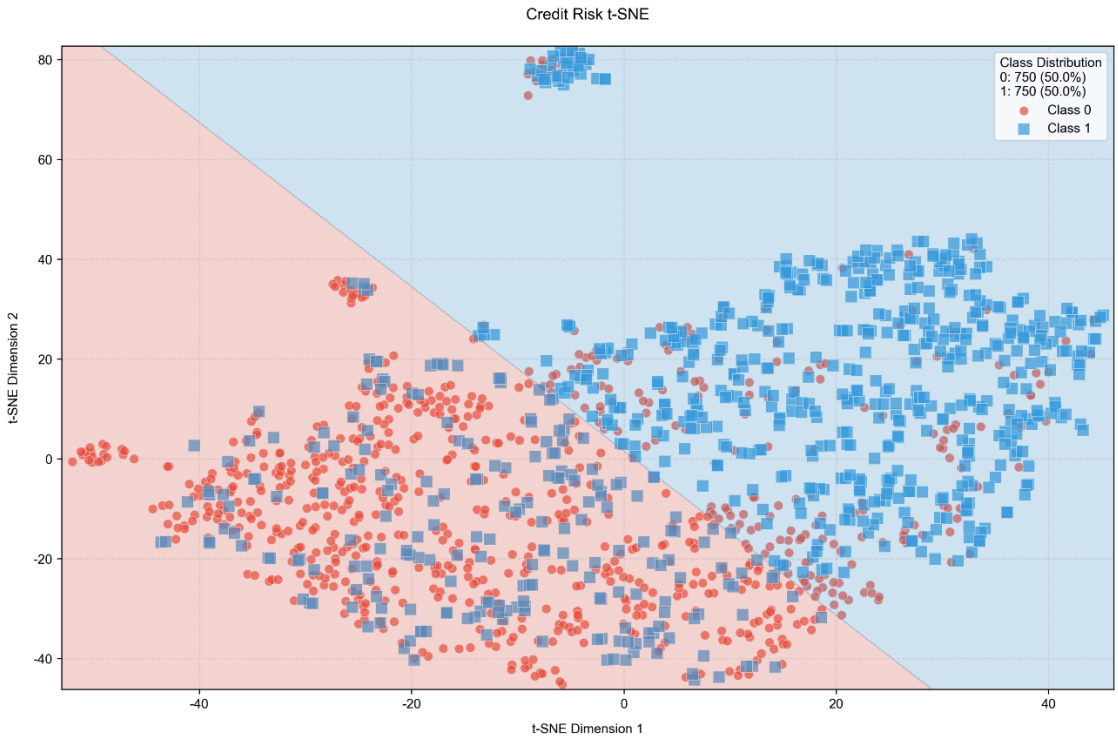

Figure 1. t-SNE data distribution map (Picture credit: Original)

At the same time, by drawing the thermal map of the variable correlation matrix, the effectiveness of variables can be preliminarily observed, and variables with high correlation can be removed, which will help the subsequent variable screening and variable construction in feature engineering. The corresponding content is shown in Figure 2.

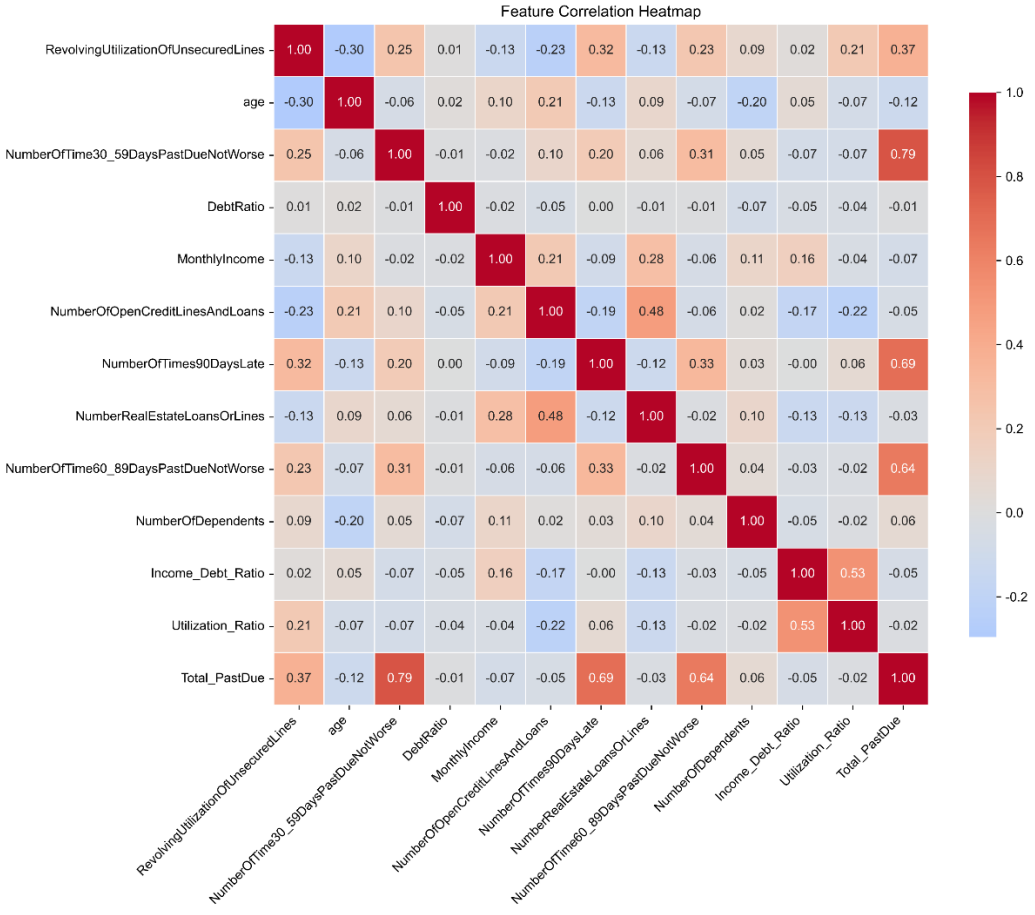

Figure 2. Variable correlation coefficient matrix (Picture credit: Original)

2.2. Loan Default Dataset

To further investigate the generalization of the model and the change in importance of the feature variables, this study simultaneously used another dataset on Kaggle. The data set is comprised of credit data that can be used for credit risk assessment, data modeling, and machine learning. The total sample data is about 150,000 rows, which contains 16 feature variables and 1 target variable (Default: 1= default,0 = no default).

The second dataset has added more detailed information about lenders, mainly including three aspects: behavioral characteristics, social background and credit intentions. For instance, in terms of credit behavior, variables such as “CreditScore”, “InterestRate”, and “DTIRatio” have been added to more directly quantify credit quality. In terms of credit intent, adding “LoanPurpose”, “LoanAmount”, and “LoanTerm” can help consider the impact of high-risk uses. At the same time, in terms of social stability, the quantification of several dimensions such as “Education”, “Employment Type”, “Months Employed”, and “Marital Status” has been highlighted.

2.3. Data Preprocessing

After the initial data reading and data checking, the following problems were found in dataset 1. First, the distribution of positive and negative samples is unbalanced, with non-defaulting borrowers accounting for the majority, resulting in a ratio column of only 1:15 between positive and negative samples. Second, the data contains some missing values, mainly Monthly Income and Number of Dependents. Third, for some features, there are extreme values or outliers in the data. At the same time, the correlation between the characteristic variables should be considered.

To solve the abovementioned problems, the following methods are mainly used in data processing:

First, use the median fill method. For the main missing fields in numeric features, Monthly Income (missing about 5.6%) and Number of Dependents (missing about 2.1%), using the data median to fill in, to avoid the model's insufficient capture of data details due to missing data, and ultimately achieve no missing data. Second, deal with outliers. For outliers and extreme values in Number of Time30-59DaysPastDueNotWorse, Number of Time60-89DaysPastDueNotWorse, Number of Times90 DaysLate, when processing data, Filter using a Z-score to remove values that do not meet the criteria. Thirdly, the class imbalance treatment is carried out. In terms of sample imbalance, SMOTE oversampling method was used. The positive/negative sample ratio of 1:15 is converted to 1:1. Fourth, construct feature engineering. In order to avoid correlation between variables and enhance model performance and training efficiency, the research creates new features after connecting domain related knowledge: Income_Debt_Ratio: monthly income/ (debt ratio + 1e-5). This characteristic variable reflects the borrower's solvency. Utilization_Ratio: credit limit utilization. The revolving credit utilization ratio (debt ratio + 1e-5) is used to measure the efficiency of the use of the credit line. Total_PastDue: indicates the total number of past due dates, including 30-59 days, 60-89 days and 90 or more. This variable reflects a comprehensive assessment of credit history. Finally, the data are standardized. For the overall data, Z-score standardization was used in the study to eliminate dimensional differences in data, thus effectively improving the convergence speed and generalization of the model.

2.4. Model

Logistic regression model, by using generalized linear model, features variables through the Sigmoid function to project data into the objective function space, to establish the corresponding probability mapping relationship. The Sigmoid function is calculated as follows.

$$P(Y = 1|X) = \sigma(W^T X + B) = \frac{1}{1+e^{-(W^T X+B)}} \quad (1)$$

Where W is the weight vector that determines the contribution of each component in X to the target variable, B is the offset of the data, and $\sigma(X)$ projects linear outcome into the interval [0,1]

Random forest enhances the interpretability and generalization of the model by integrating multi-class decision trees. This model

Firstly, the Bootstrap Aggregating (Bagging) process was used to extract N samples from the returned data set and generate B subsets. For each subset, a corresponding training decision tree is built, and m randomly selected features are considered respectively (m≤M, where M is the total number of features).

At the same time, the splitting criterion of a single decision tree is established, and Gini impurity is used as a reference index. Gini impurity reduction at the split point s for feature ij:

$$\Delta Gini(j,s) = Gini(D) - \frac{|D_L|}{|D|} Gini(D_L) - \frac{|D_R|}{|D|} Gini(D_R) \tag{2}$$

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{3}$$

As an integrated model, the gradient lift tree iteratively optimizes the loss function through an addition model based on decision trees.

$$\mathcal{L}^{(t)} = \sum_{\{i=1\}}^{N} L\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

By approximating the above expression with Taylor's second-order expansion, the following result can be obtained:

$$\mathcal{L}^{(t)} \approx \sum_{\{i=1\}}^{N} [L\left(y_i, \hat{y}_i^{t-1}\right) + g_i f_{t(x_i)} + \frac{1}{2} h_i {f_t}^2(x_i) + \Omega(f_t) \tag{5}$$

Where $\boldsymbol{g_i}$ and $\boldsymbol{h_i}$ are the first and second partial derivatives of the above function, respectively:

$$g_i = \partial_{\hat{y}_i^{t-1}} L\left(y_i, \hat{y}_i^{t-1}\right) \tag{6}$$

$$h_i = \partial^2_{\hat{y}_i^{t-1}} L\left(y_i, \hat{y}_i^{t-1}\right) \tag{7}$$

Then, the optimal weight of each leaf node can be calculated:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{8}$$

Table 2 shows a comparison between XGBoost and Light GBM on decision tree generation strategies and split point Settings.

Table 2. Comparison between XGBoost and LightGBM

Character	XGBoost	LightGBM
Decision tree growth strategy	Exact Greedy	Histogram-based
Split location	Sort the eigenvalues first, and then go through all possible segmentation points	For eigenvalue classification, only bucket boundaries are evaluated

Multilayer Perceptron (MLP) is a basic and classical neural network architecture consisting of multiple neurons and multiple layers that pass inputs to fully connected transformation layers in turn and are gradually optimized through backpropagation to achieve the approximation of complex functions. Gradient of the loss function to the weight $W^{(l)}$ is as follows:

$$\frac{\partial \mathcal{L}}{\partial W^{(l)}} = \delta^{(l)} a^{(l-1)T} \tag{9}$$

Where $\delta^{(l)}$ can be calculated by the chain rule:

$$\delta^{(l)} = \left(W^{(l+1)T}\delta^{(l+1)}\right) \odot \sigma'^{(L)}\left(z^{(L)}\right) \tag{10}$$

For the practical sample data during training, we need to use Dropout regularization processing to randomly remove neurons with probability p, and modify the forward propagation to:

$$a^{(l)} = \sigma^{(l)}\left(z^{(L)} \odot m^{(l)}\right) \tag{11}$$

Where $m^{(l)} \sim Bernoulli(p)$, as the mask vector, is used to scale the weight during testing.

3. Experiment

In the model training process of this study, the methods of grid search and Bayesian optimization were also used for cross-validation, and the hyperparameters corresponding to each model were optimized to improve the efficiency and effect of the model. Meanwhile, during the training process, an early stop mechanism is set up to monitor the AUC situation of the validation set. If there is no improvement for five consecutive rounds during the training process, the training will be terminated, and the optimal weights will be saved to prevent the model from overfitting. In addition, a series of evaluation metrics and visualization analysis were used in the study. Based on this, the performance of the four main models in two datasets was compared intensively.

3.1. Model Comparison

The following Tables 3 and 4 are the evaluation metrics obtained for different models of Dataset 1 and Dataset 2.

Table 3. Model evaluation for Give Me Some Credit Dataset

Model	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.77	0.76	0.76	0.76
XGBoost	0.97	0.94	0.95	0.94
LightGBM	0.97	0.95	0.95	0.95
Random Forest	0.93	0.95	0.94	0.94
MLP	0.77	0.77	0.77	0.78

In Table 3, the model results illustrate that the overall recognition ability of the logistic regression model and MLP is relatively weak. Among them, in data visualization, the data distribution presented through T-distributed stochastic neighbor embedding (t-SNE） can already illustrate that the data distribution is not a strict linear distribution, so the effect of the logistic regression model is poor. As for MLP, due to the large amount of data processed, after weighing the time cost of training and the complexity of the model, it shows a relatively weak recognition ability, but at the same time, it also demonstrates better model stability. For model types based on decision trees, nonlinear data features

can be captured very precisely. Therefore, the comprehensive performance of XGBoost, LightGBM and random Forest is all satisfying.

To further explore whether the models that perform well in Dataset 1 can maintain good generalization ability and stability when dealing with different datasets facing a larger number of samples and more types of feature variables, the study introduced Dataset 2 and used XGBoost, LightGBM, random forest and MLP models. And the same type of evaluation indicators was recorded, as shown in Table 4.

Table 4. Model Evaluation for Loan Default Dataset

Model	Precision	Recall	F1-score	Accuracy
XGBoost	0.91	0.90	0.90	0.90
LightGBM	0.91	0.90	0.90	0.90
Random Forest	0.91	0.91	0.91	0.91
MLP	0.87	0.86	0.86	0.86

It can be found from Table 4 that regardless of which dataset it is for, the main models XGBoost, LightGBM, and Random Forest all have good model performances, and each index reaches above 0.9, indicating that the models have good universality and generalization ability.

3.2. Visualization Analysis

3.2.1. Model Convergence Analysis.

In the experiment, by setting the training strategy of the model, using methods such as the early stop mechanism, dynamic adjustment of the learning rate, model regularization and Dropout, validation-driven hyperparameter optimization, incremental training, dynamic recording of the model training log, and data preprocessing, the convergence of the model was ensured. Figures 3, 4 and 5 record the training curves and convergence situations of different models for dataset 1.

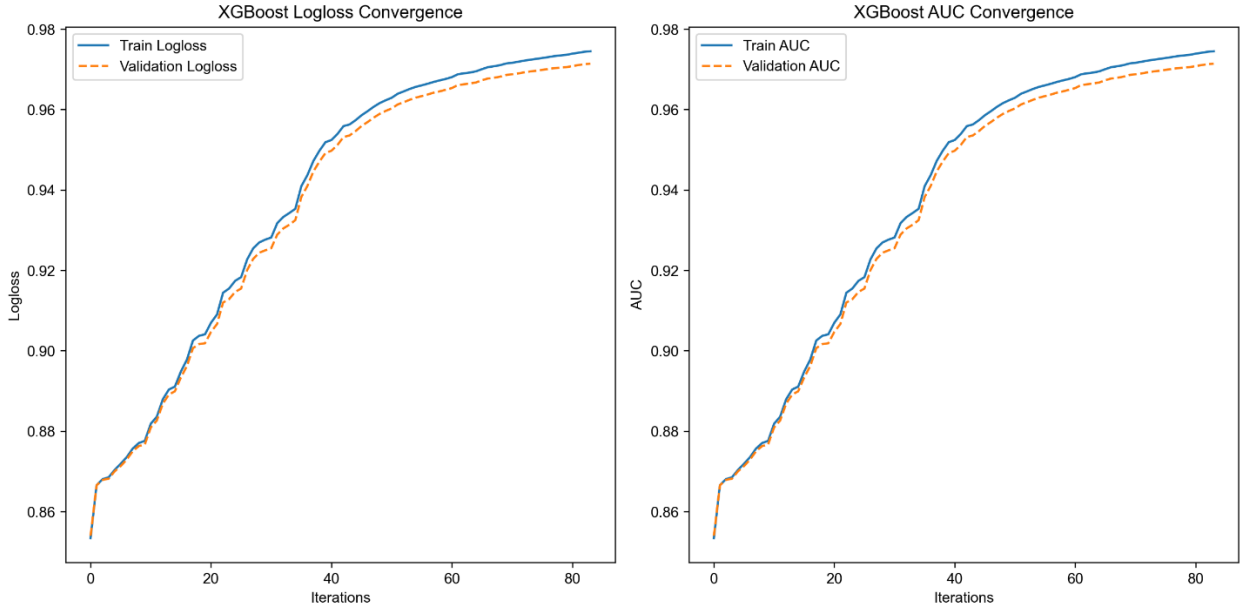

Figure 3. XGBoost training curves (Picture credit: Original)

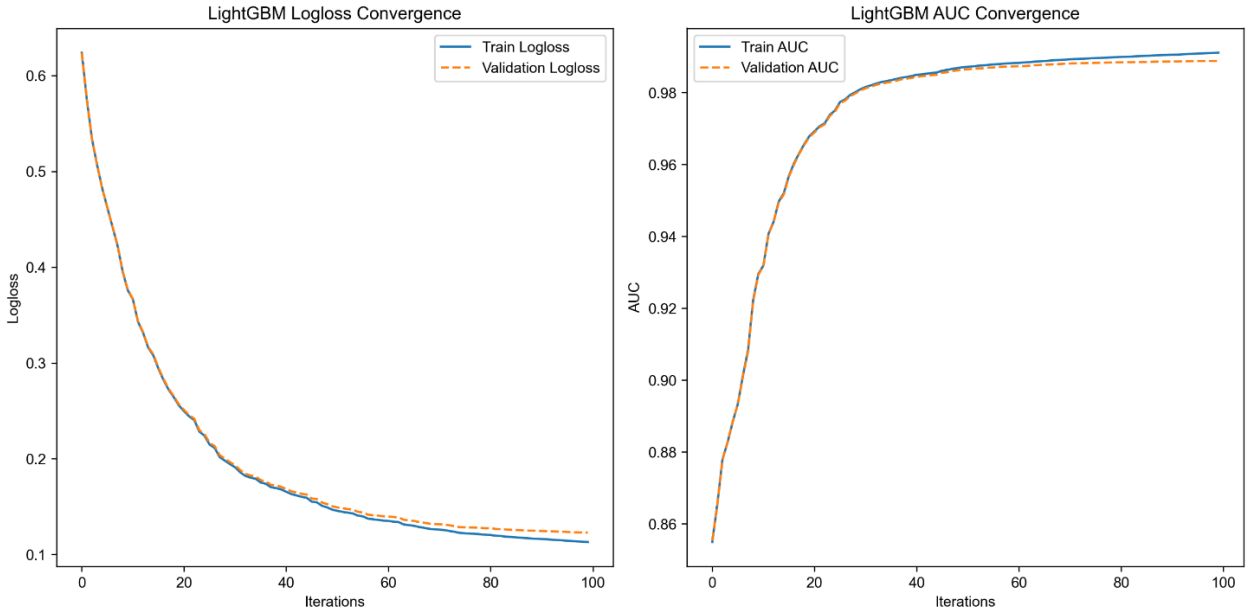

Figure 4. LightGBM training curves (Picture credit: Original)

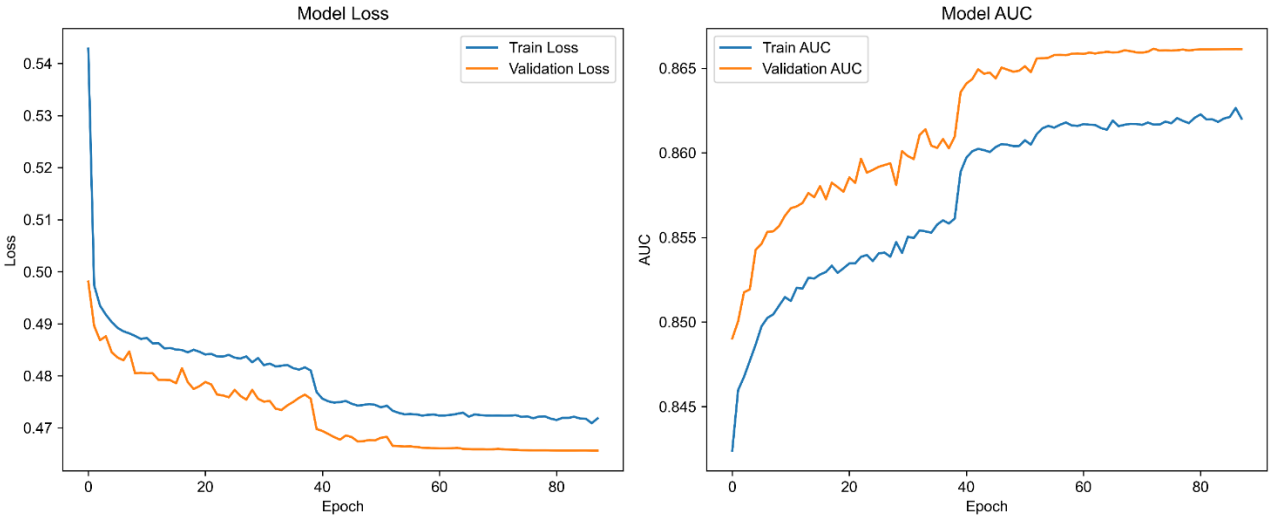

Figure 5. MLP training curves (Picture credit: Original)

The study also plotted the ROC and PC curves of different models (Figure 6) to compare the classification performance of different models. The horizontal axis of the ROC curve is FPR, which stands for False Positive Rate, indicating the proportion of negative samples that are wrongly predicted as positive samples. The vertical axis of the corresponding ROC curve is TPR, which stands for True Positive Rate, representing the proportion of positive samples that are correctly predicted. The ROC curve reflects the comprehensive evaluation ability of the model under different decision-making interfaces. Meanwhile, the area AUC below the curve can quantify this indicator well. The horizontal and vertical axes of the PR curve represent Recall and Precision respectively. The area below the curve reflects the model's recognition ability on positive samples, making it suitable for scenarios such as disease detection and risk assessment, where the recognition accuracy of positive samples is of greater concern. It can be seen from the ROC and PR curves that each model has a good recognition ability for the data.

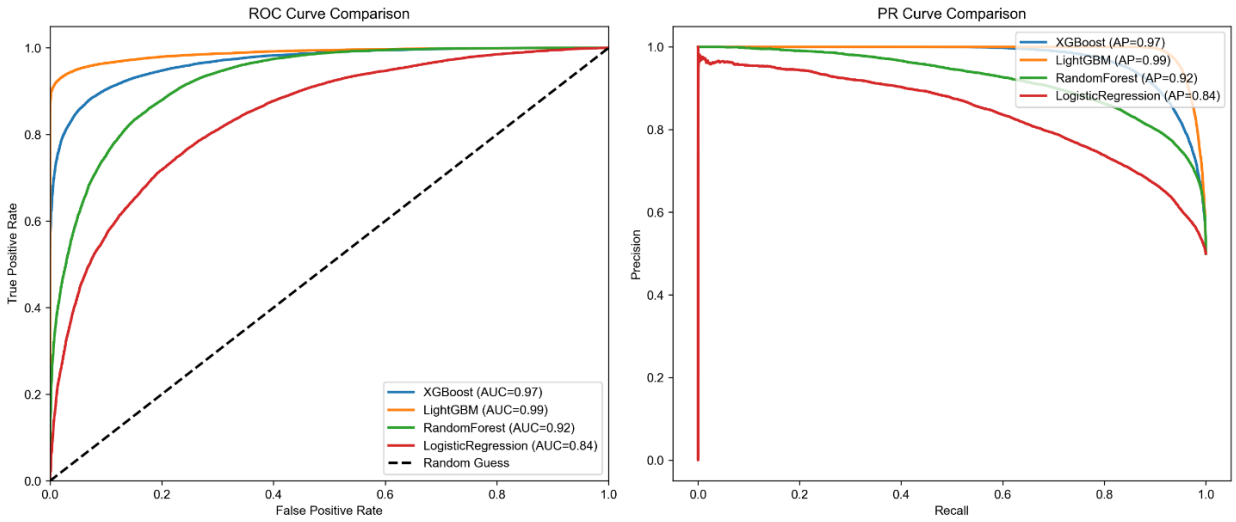

Figure 6. The ROC and PR curves of the model (Picture credit: Original)

3.2.2. Overfitting Detection.

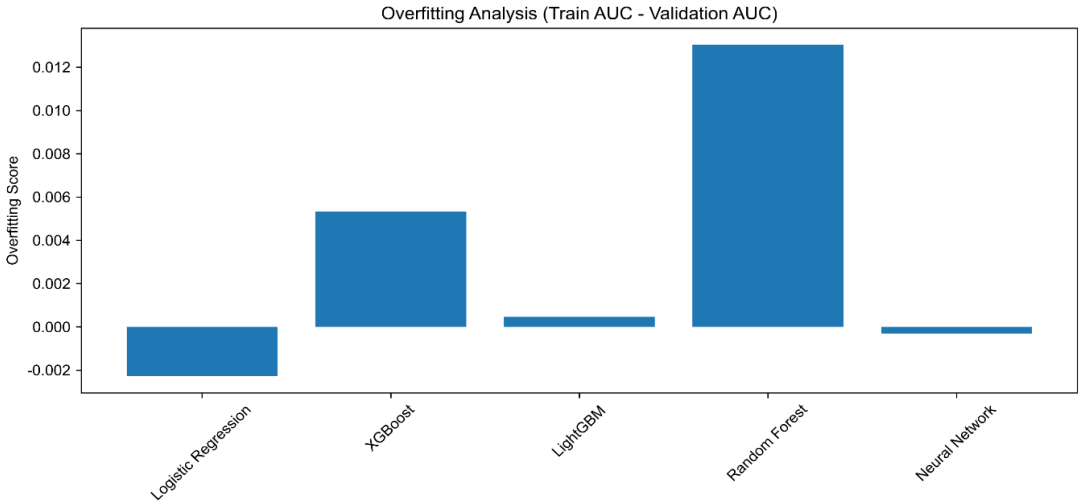

Figure 7. Overfitting analysis (Picture credit: Original)

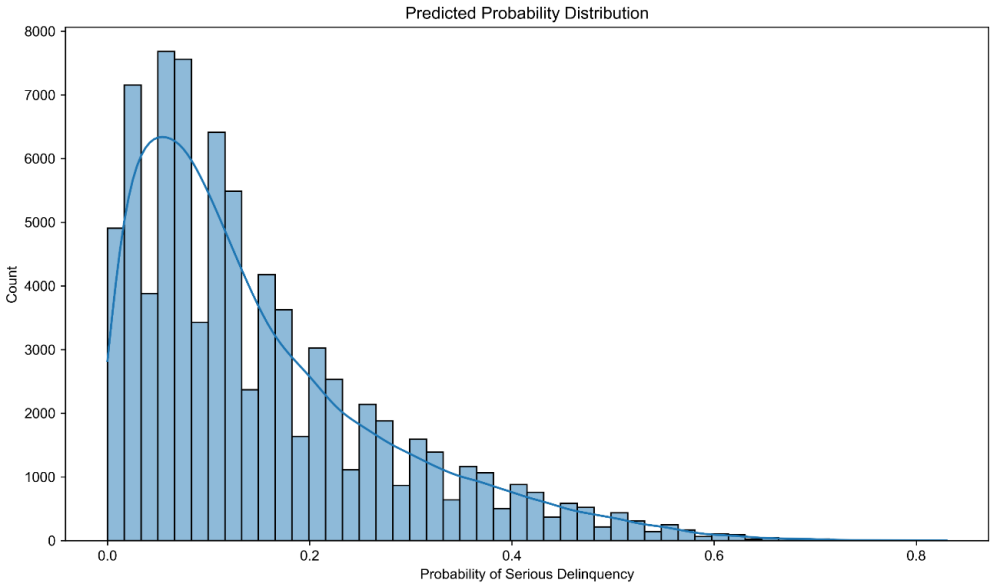

Figure 8. Histogram of the probability distribution of the predicted outcome (Picture credit: Original)

In the experiment, the analysis of the overfitting situation of each model was analyzed (Figure 7). It can be found from it that except for the value of Random Forest being slightly greater than 0.012 and possibly having a slight overfitting situation, the models of other models fit well and there is no obvious overfitting situation. Furthermore, the value of Logistic Regression being -0.0023 indicates that the model is not overfitted and is even slightly conservative. In the experiment, for the Random Forest, the corresponding OOB Score was also output additionally. The output result was 0.85, indicating that the model fits well and has a strong generalization ability. Figure 8 shows the probability distribution of the output results of Random Forest, the model with the best comprehensive evaluation in the training set of datasets 1.

3.3. Feature Importance Analysis

For the model results of the two datasets, the models with better results were selected respectively. The interpretability analysis of their training results was continued, and it was explored whether the feature variables had the same contribution to different datasets, providing data support for subsequent research. In the SHAP graph, the horizontal axis represents the SHAP value corresponding to each feature variable, reflecting the degree of influence of that variable on the model's predicted output. The larger the absolute value, the greater the influence. Each point in the figure represents the SHAP value of a sample on that feature. The color indicates the magnitude of the feature value itself, with red representing high values and blue representing low values.

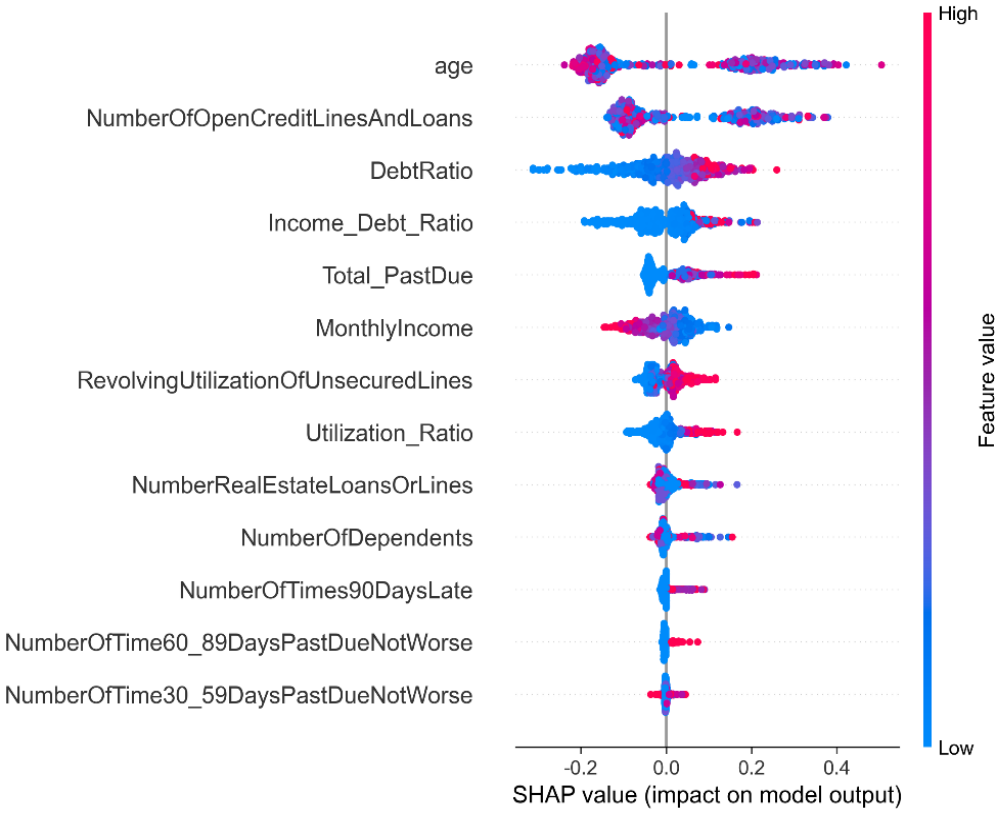

Figure 9. SHAP analysis of the best model for Give Me Some Credit Dataset (Picture credit: Original)

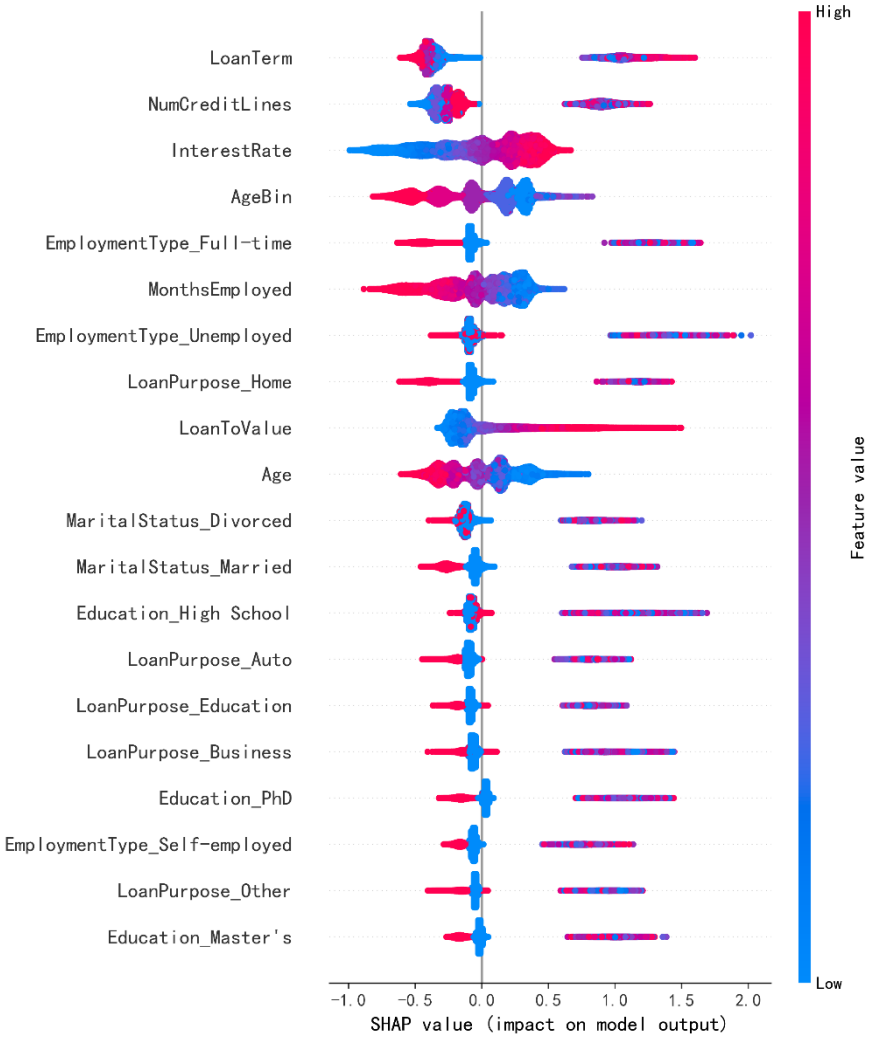

Figure 10. SHAP analysis of XGBoost model for Loan Default Dataset (Picture credit: Original)

The interpretability analysis results of the features in Dataset 1 are shown in Figure 9, and the interpretability analysis of the top several feature variables in dataset 2 is shown in Figure 10. The performance of different model variables basically shows consistency.

By comparing the performance of the characteristic variables in these two datasets, the following related influencing factors are emphasized: Including age, credit limit, debt ratio, income level, loan-to-value ratio, as well as demographic information and financial behavior variables related to the borrower. Subsequently, it is necessary to further strengthen feature engineering to better capture the complex changes of variables in different situations, conduct more in-depth analysis of some variables, and enhance the overall interpretability and practicability of the model.

4. Conclusion

This study focuses on the issue of credit risk assessment and explores the performance differences and interpretability of various machine learning and deep learning models. The research results show that the integrated models (XGBoost, LightGBM and Random Forest) have the best comprehensive performance on the two types of datasets, with accuracy rates exceeding 90% and AUC values close to 0.98, which is significantly better than logistic regression and MLP. The MLP model, on the other hand, shows better stability, with the smallest difference in AUC between the training set and the test set. Add the displayed shapes in addition, the experimental analysis, in view of the risk influence factors are analyzed as the core of the debt ratio (DebtRatio), credit line utilization (RevolvingUtilizationOfUnsecuredLines) and age (age), gives the interpretability of the specific analysis, it was compared and studied whether there were significant differences in the roles of

characteristic variables in different datasets. Meanwhile, the consistency of the influence direction of core variables with the experience in the financial field was investigated, verifying the credibility and practicability of the model.

However, this study still has certain limitations. Firstly, the source of the data is merely limited to public datasets and does not cover the acquisition and research of real-time dynamic credit behavior data. Secondly, in the model construction, since the training of MLP takes a long time, the predictive potential of its deep architecture has not been fully explored. Finally, although feature engineering alleviates some correlations, in order to better explain the contribution of feature variables to the target variables and the interaction effects between variables, more refined modeling strategies are still needed.

The practical significance of this paper lies in providing financial institutions with multi-model comparison and interpretable comprehensive evaluation schemes, and at the same time providing data support for the formulation of risk control strategies through feature importance analysis. Against the backdrop of the increasing importance of financial risk management and control, this study has laid a practical and technical foundation for the subsequent construction of an intelligent credit decision-making system.

References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19 (10): 53 - 58.

[2] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C., Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research, Eur. J. Oper. Res., vol. 247, no. 1, pp. 124 – 136, 2015.

[3] Chen, H., Yang, C., Du, M., & Zhang, Y., Research on Credit Risk Prediction Under Unbalanced Dataset Based on Ensemble Learning, Math. Probl. Eng., vol. 2023, Article ID 2927393, 18 pages, 2023.

[4] He, H., & Garcia, E. A., Learning from Imbalanced Data, IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263 – 1284, 2009.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., SMOTE: Synthetic Minority Over-Sampling Technique, J. Artif. Intell. Res., vol. 16, pp. 321 – 357, 2002.

[6] Chen, T., & Guestrin, C., XGBoost: A Scalable Tree Boosting System, in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., San Francisco, CA, USA, 2016, pp. 785 – 794.

[7] Lundberg, S. M., & Lee, S. I., A Unified Approach to Interpreting Model Predictions, in Adv. Neural Inf. Process. Syst., vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[8] Bumin, M., & Ozcalici, M., Predicting the Direction of Financial Dollarization Movement with Genetic Algorithm and Machine Learning Algorithms: The Case of Turkey, Expert Syst. Appl., vol. 213, p. 119301, 2023.

[9] Hlongwane, R., Ramabao, K., & Mongwe, W., A Novel Framework for Enhancing Transparency in Credit Scoring: Leveraging Shapley Values for Interpretable Credit Scorecards, PLoS One, vol. 19, no. 8, p. e0308718, 2024.

[10] Didkovskyi, O., Jean, N., Pera, G. L., et al., Cross-Domain Behavioral Credit Modeling: Transferability from Private to Central Data, arXiv preprint, arXiv: 2401.09778, 2024.

[11] Bücker, M., Szepannek, G., Gosiewska, A., et al., Transparency, Auditability, and Explainability of Machine Learning Models in Credit Scoring, J. Oper. Res. Soc., vol. 73, no. 1, pp. 70 – 90, 2022.