Domain-Informed Default Risk Prediction: Interpretable Features and Model Performance Analysis

Xiyun Yan
South China Normal University, Guangdong, China
20233637004@m.scnu.edu.cn

Abstract. The rapid development of peer-to-peer lending platforms has provided a more efficient financing channel for the consumer credit market. However, limited borrower information and the regulatory environment increase the difficulty of predicting default risks. In the problem of default prediction, there are usually challenges such as sample imbalance and how to select appropriate models. In the field of credit default prediction, existing research mostly focuses on model selection and model optimization, while less attention is paid to improving model performance by integrating financial domain knowledge to construct interpretable features. Therefore, this paper proposes three new features - DTI_Interest (debt-to-income ratio for interest), stability score, and loan density, which respectively describe repayment pressure, financial stability, and borrowing intensity. Experiments based on five different models show that although the new features slightly improve the prediction performance, this improvement is not statistically significant. This study provides new ideas for P2P risk feature engineering and highlights the potential value of domain knowledge in credit assessment.

Keywords: default risk prediction, interpretable features, machine learning

1 Introduction

The expansion of the P2P lending model has transformed the consumer credit market, offering faster and more comprehensive financing options. However, these platforms usually operate under conditions where borrower information is limited and regulation is lax, which poses new challenges for predicting default risks. The main challenges include how to select appropriate models to accurately predict default risks and how to improve model performance in the case of imbalanced samples.

Addressing these challenges, extensive research has focused on two primary aspects: advanced modeling techniques and data imbalance treatment. In parallel, modern gradient boosting implementations, including eXtreme Gradient Boosting (XGBoost), Gradient Boosting Decision Tree (GBDT), and Light Gradient Boosting Machine (LightGBM), have demonstrated superior predictive accuracy when processing structured financial data [1, 2]. To address the prevalent challenge of class imbalance, researchers have developed multiple solution strategies. Advanced oversampling methodologies have emerged, such as K-means Synthetic Minority Over-sampling Technique (SMOTE), SMOTETomek, and FCM_Weight_SMOTE, which employ

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_26

various mechanisms to rebalance dataset distributions [3]. Additionally, innovative ensemble approaches have been proposed, including weighted soft voting heterogeneous ensembles and Bayesian model averaging techniques, which aim to improve detection performance for minority class instances [4].

Despite these advancements, a critical gap remains in the application of domain knowledge through feature engineering. Current literature mainly emphasizes model selection, ensemble strategies, and imbalance treatment. However, most studies operate primarily on original feature sets, with limited attention to constructing interpretable features informed by financial expertise. Insufficient investigation exists regarding the impact of domain-specific composite features on predictive power in lending default risk.

To address this gap, three novel interpretable features—DTI_Interest, Stability Score, and Loan Density—are proposed to capture core borrower risk dimensions: repayment burden (incorporating interest cost), financial stability, and borrowing intensity. Subsequently, this study compares the model performance differences between the basic dataset and the dataset with these three new features added and further analyzes whether the new features can significantly enhance the predictive ability of the model.

2 Method

2.1 Framework

Fig. 1 displays the workflow for loan default prediction clearly. The pipeline begins with raw data ingestion and cleaning, followed by the creation of domain-informed features. The SMOTETomek resampling technique is then applied to address class imbalance. Finally, five classifiers are trained and evaluated using AUC, Precision, Recall, and F1-score as primary metrics.

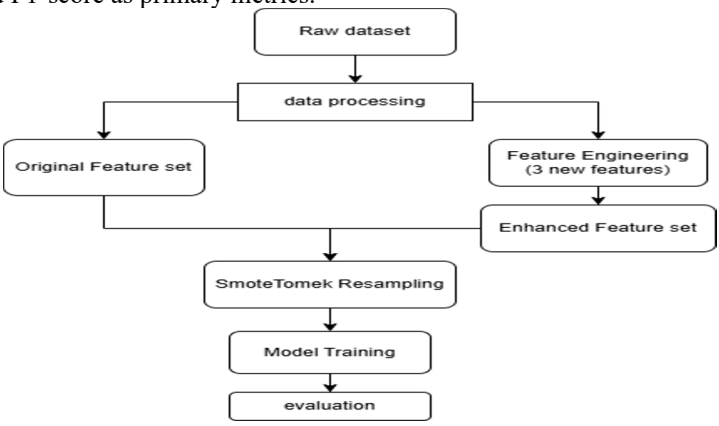

Fig. 1. The structure of the experiment (Picture credit: Original).

2.2 Dataset and Data Processing

The loan default prediction dataset comes from Kaggle, which contains 255,347 loan records and 18 variables related to borrower demographics, financial status, loan details, and repayment outcomes [5]. The dataset was used to construct a binary classification model aimed at predicting loan defaults.

This investigation focuses on binary classification of loan repayment outcomes, with the target variable "default status" encoded as 1 (default) versus 0 (successful repayment). The feature space comprises both continuous and discrete variables: numerical attributes include demographic and financial metrics (age, income, credit score, etc.), while categorical variables encompass socio-economic indicators (education level, employment type, marital status) and loan characteristics (purpose, co-signer status). Preliminary EDA revealed substantial class imbalance, with defaults representing merely 11.9% of observations. Feature engineering involved type-specific transformations: numerical features underwent standardization via StandardScaler, whereas nominal categorical variables were processed through one-hot encoding

2.3 Feature Engineering

To explore whether the new features can enhance the forecasting power of the model, this research innovatively constructed three risk characteristic indicators with clear financial explanatory power. These features depict the default risk characteristics of borrowers from different perspectives.

The first feature is DTI_Interest, which is obtained by multiplying the borrower's debt-to-income ratio by the interest rate stipulated in the loan contract. This composite characteristic not only takes into account the borrower's relative debt level but also incorporates the factor of capital cost, enabling a more accurate reflection of the borrower's actual debt repayment pressure.

The second feature is the Stability Score (Stability Rating), which is a comprehensive indicator that assesses the borrower's social and economic stability. This scoring system consists of three key dimensions: marital status (Married), mortgage status (Has Mortgage), and career stability (Months Employed > 36). Each dimension uses a binary scoring method, with 1 point awarded for meeting the conditions and 0 points for not meeting them. Specifically, being married earns 1 point, reflecting the possibility of family responsibility and additional income sources; having a mortgage earns 1 point, indicating a higher asset base and default risk; being continuously employed in the current position for more than 36 months earns 1 point, representing career stability. The scores of these three dimensions are added together to obtain the total score, ranging from 0 to 3 points.

The third feature is Loan Density, which is calculated by dividing the loan amount by the loan term (in months). This design ingeniously avoids the problem of indicator distortion that may occur with extremely short-term loans. This feature quantifies the intensity of repayment pressure that the borrower needs to bear within a certain period of time.

2.4 SmoteTomek

Additionally, in the dataset of this study, the class imbalance problem is quite severe, with only 11.9% of the observed samples being labeled as default samples. To address this issue, this paper adopts the SMOTETomek hybrid resampling technique [6]. This technique combines the minority class over-sampling method with the Tomek link removal method, aiming to clean up the ambiguous class boundaries. SMOTETomek combines the SMOTE with Tomek link removal. The SMOTE component generates synthetic minority samples via linear interpolation:

$$x_{\text{new}} = x_i + \lambda(x_j - x_i), \lambda \sim \mathcal{U}(0,1) \tag{1}$$

This enriches the minority class distribution by populating sparse regions in feature space. The Tomek link removal phase identifies borderline instances (xi, xj) from different classes where each is the other's nearest neighbor:

$$\text{NN}(x_i) = x_j \text{ and } \text{NN}(x_j) = x_i \tag{2}$$

The majority of class samples in such pairs are removed to increase inter-class separation and reduce overlap. This hybrid technique thus combines synthetic oversampling with boundary cleaning, resulting in more balanced and separable training data.

2.5 Model

To verify the predictive performance of the enhanced feature set, this paper selected five common classification models for evaluation, including logistic regression (LR), random forest (RF), XGBoost, LightGBM, and the soft voting ensemble method.

Logistic regression is a classic supervised learning algorithm and is widely applied in various data classification problems. Its core mechanism is to establish a classification model through a training dataset (including known output labels) and use the Sigmoid function to map the linear regression results to the [0, 1] interval, outputting probability values to determine the category affiliation (such as in binary classification problems). Logistic regression has a higher classification accuracy rate. For example, in Narayanan's research, the average accuracy rate of logistic regression reached 95.79% (or 97.92% for the new logistic regression), significantly higher than that of linear regression at 92.94% (or 94.88%) [7].

As an ensemble method, Random Forest generates numerous decision trees through bootstrap aggregation and random feature selection. Classification outputs derive from majority voting among trees, whereas regression results are tree averages [8]. This approach improves prediction stability while reducing overfitting risks.

Both LightGBM and XGBoost are efficient machine learning algorithms based on GBDT, widely used in classification, regression, and ranking tasks. They build multiple weak learners (decision trees) through iterative optimization and eventually combine them into a strong learner. However, there are still some differences between the two.

The main advantages of XGBoost include regularization to prevent overfitting, parallel processing, support for custom optimization objectives, dealing with missing data, controlling model complexity through pruning strategies, evaluating performance using cross-validation methods, and having the feasibility of continuous training

According to the reference information, the Soft Voting Ensemble Classifier is an ensemble learning method that enhances overall performance by combining the prediction results of multiple base classifiers [9]. In this study, this method calculates the weighted average of the probability values output by 4 base models (XGBoost, Random Forest, Logistic Regression, LightGBM).

2.6 Evaluation Metrics

This study conducted the evaluation of the model's performance from multiple perspectives, with the selected evaluation metrics. The assessment incorporated four key indicators that collectively provide a robust measurement of classification effectiveness. The area under the receiver operating characteristic curve (AUC) served as the primary measure of the model's discriminative capability between classes. Precision reflects the ratio of actual positive samples among those predicted as positive by the model; recall assesses the model's effectiveness in capturing all true positive instances; and the F1 score, being the harmonic average of precision and recall, provides a comprehensive measure of the model's balanced performance in classification scenarios.

3 Result

3.1 Comparison of Model Performance

Five predictive models were systematically compared using original and enhanced feature sets. Performance metrics (AUC, precision, recall, F1) for both configurations are presented in Table 1.

Table 1. Model performance comparison on original and enhanced feature set

Feature Set	Model	AUC	Precision	Recall	F1 Score
Original	LR	0.7491	0.2089	0.6652	0.3179
Original	RF	0.7037	0.4737	0.0396	0.0732
Original	XGBoost	0.6916	0.3810	0.1057	0.1655
Original	LightGBM	0.7111	0.5128	0.0881	0.1504
Original	Soft Voting	0.7388	0.4925	0.1454	0.2245
Enhanced	LR	0.7475	0.2097	0.6696	0.3193
Enhanced	RF	0.7114	0.3448	0.0441	0.0781
Enhanced	XGBoost	0.6852	0.4444	0.1057	0.1708
Enhanced	LightGBM	0.7039	0.5208	0.1101	0.1818
Enhanced	Soft Voting	0.7390	0.5538	0.1586	0.2466

In terms of overall predictive ability (AUC), the logistic regression (LR) model stands out, demonstrating the strongest discriminatory ability between default and non-

default samples. Its AUC value on the original feature set is 0.7491, and on the enhanced feature set, it is 0.7475, both being the highest among all models. On the enhanced feature set, the AUC value of the ensemble model (soft voting) increases to 0.7390, indicating that multi-model integration can reduce the bias of individual models. Tree models (such as XGBoost and LightGBM) perform moderately, with their AUC values slightly decreasing after adding new features, suggesting possible overfitting or redundant features.

From the perspective of accuracy and recall rate, under the original feature set, LightGBM has the highest accuracy of 0.5128, which increases to 0.5208 after feature enhancement, but its recall rate is only 0.1101, indicating severe underreporting. Logistic regression has a significantly higher recall rate than other models, being 0.6652 on the original feature set and 0.6696 after enhancement, capable of identifying approximately 67% of true default samples, although its accuracy is relatively low at around 0.209, with many false positives. After feature enhancement, the accuracy of the random forest model drops from 0.4737 to 0.3448, and its recall rate does not improve significantly, suggesting that the feature engineering effect is limited.

Finally, all models exhibited relatively low F1-scores (range: 0.0732-0.3193), a phenomenon closely associated with the class imbalance characteristics of the dataset (default samples accounting for only 11.9%). Notably, logistic regression (LR) demonstrated significant advantages across both feature sets (original feature set: 0.3179; enhanced feature set: 0.3193), outperforming other models by approximately 2-4 times. This superiority primarily stems from LR's outstanding performance in recall metrics (0.6652-0.6696), despite its relatively lower precision (approximately 0.209)

3.2 T-Test

To evaluate the influence of the additional features on predictive performance, statistical comparisons were performed using paired t-tests (Table 2).

Table 2. T-test results

metrics	Mean difference	std	t-value	p-value
AUC	-0.0015	0.0060	-0.54	0.615
Precision	+0.0009	0.0755	0.03	0.980
Recall	+0.0088	0.0088	2.24	0.089
F1 Score	+0.0126	0.0126	2.24	0.089

Through paired sample t-test analysis, it was found that the improvement effect of feature engineering on AUC, precision, recall rate, and F1 score did not show statistical significance (all p-values were higher than 0.05). Nevertheless, the model performance still showed a certain positive trend: for instance, the average recall rate improved by 0.0088; the average F1 score improved by 0.0126. In conclusion, although feature engineering has a certain promoting effect on model performance, the improvement magnitude has not yet reached a significant level.

4 Discussion

This research explores machine learning applications in predicting loan defaults with particular emphasis on how feature engineering affects model performance. The study conducted a thorough evaluation of five widely used algorithms, including logistic regression, random forest, XGBoost, LightGBM, and a soft voting ensemble method. The assessment employed four key metrics measuring model effectiveness, which were AUC, accuracy, recall rate, and F1 score. Results indicated that while the additional features DTI_Interest, StabilityScore, and LoanDensity contributed to slight performance improvements, these enhancements did not prove statistically significant.

The findings present both confirmations and contradictions when compared to previous studies. Supporting existing literature, LightGBM demonstrated better performance than XGBoost in terms of AUC and accuracy measurements, reinforcing the advantages of its data processing methodology as referenced in prior work [10]. However, contrary to established research suggesting random forests typically outperform logistic regression in default risk assessment, this study found logistic regression delivered superior overall results [11]. Specifically, it achieved recall rates between 0.6652 and 0.6696, making it particularly effective at identifying actual default cases. These differences may be attributed to variations in dataset composition, preprocessing approaches, or feature selection techniques.

Several limitations should be noted regarding this study. The models' generalizability may be constrained by the sample size and feature range used in the research. Although feature engineering produced measurable improvements, their statistical insignificance indicates a potential need for more advanced feature selection or extraction methods. Future research could benefit from conducting a detailed feature importance analysis to identify and retain only the most predictive variables, potentially improving model efficiency. Expanding datasets to include more balanced samples might enhance classifier performance. Additionally, investigating neural networks or hybrid models could lead to better prediction accuracy by capturing complex nonlinear relationships present in loan default data.

5 Conclusion

This study proposed three novel financial indicators—DTI_Interest, StabilityScore, and LoanDensity—to enrich the existing feature space. A comprehensive evaluation was conducted across five machine learning algorithms using both the baseline and enhanced feature sets. The analysis revealed minimal difference in discriminative power between the two sets (AUC 0.7491 vs 0.7475), though the enriched features showed marginally better performance in identifying actual defaults (recall 0.6652-0.6696). Among the tested models, tree-based approaches maintained higher classification accuracy at the expense of recall, while the soft voting ensemble demonstrated consistent but relatively weaker results when benchmarked against logistic regression.

The experimental results indicated modest performance gains from the additional features, with F1-score improving by 0.0126 and recall increasing by 0.0088. However, statistical testing through paired t-tests determined these improvements lacked significance. These findings suggest potential research directions including the acquisition of larger datasets and investigation of more sophisticated neural network architectures to potentially achieve more substantial performance enhancements.

References

1. Zhu, L., Qiu, D., Ergu, D., Ying, C., Liu, K.: A study on predicting loan default based on the random forest algorithm. Procedia Comput. Sci. **162**, 503–513 (2019)
2. Akinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O., Ogunleye, B.: Ensemble-based machine learning algorithm for loan default risk prediction. Mathematics **12**(21), 3423 (2024)
3. Cao, W., Zhang, J.: Prediction of farmer loan default risk based on imbalanced data processing and weighted soft voting heterogeneous ensemble. Comput. Appl. Softw., **42**(08),71-79. (in Chinese)
4. Weng, F., Zhu, M., Buckle, M., Hajek, P., Abedin, M.Z.: Class imbalance Bayesian model averaging for consumer loan default prediction: The role of soft credit information. Res. Int. Bus. Finance **74**, 102722 (2025)
5. Kaggle: Loan Default Prediction Dataset. https://www.kaggle.com/datasets/nikhil1e9/loan-default, last accessed 2025/07/24
6. Hairani, H., Anggrawan, A., Priyanto, D.: Improvement performance of the random forest method on unbalanced diabetes data classification using SMOTE-Tomek Link. JOIV: Int. J. Inform. Vis. **7**(1), 258–264 (2023)
7. Bhavani, R., Balamanigandan, R., Priscilla, A.A.: Analyzing the performance of novel logistic regression over linear regression algorithms for predicting fake job with improved accuracy. In: 2024 5th Int. Conf. Electron. Sustain. Commun. Syst. (ICESC), pp. 1728–1732. IEEE, (2024)
8. Aria, M., Cuccurullo, C., Gnasso, A.: A comparison among interpretative proposals for Random Forests. Mach. Learn. Appl. **6**, 100094 (2021)
9. Behera, D.K., Dash, S., Behera, A.K., Dash, C.S.K.: Extreme gradient boosting and soft voting ensemble classifier for diabetes prediction. In: 2021 19th OITS Int. Conf. Inform. Technol. (OCIT), pp. 191–195. IEEE, (2021)
10. Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., Niu, X.: Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning. Electron. Commer. Res. Appl. **31**, 24–39 (2018)
11. Coşer, A., Maer-Matei, M.M., Albu, C.: Predictive models for loan default risk assessment. Econ. Comput. Econ. Cybern. Stud. Res. **53**(2) (2019)