

# ***Domain-Driven Feature Engineering and Multi-Machine Learning Models for Credit Card Default Prediction***

**Xinge Miao**

*School of Statistics, Tianjin University of Finance and Economics, Tianjin, China  
miaoxg@stu.tjufe.edu.cn*

**Abstract.** The paper conducts a systematic study on credit card default prediction using the UCI credit card default dataset with 29,965 samples and 23 original features. The paper proposes six categories of domain-knowledge-driven engineered features, including credit utilization ratio, repayment ratio, delinquency severity score, financial trend metrics, interaction features, and binning features, which convert raw financial variables into effective predictive signals. Under unified experimental settings with an 80 / 20 stratified train-test split and random state set to 42, the paper benchmarks six classifiers: logistic regression, K-nearest neighbors (KNN), random forest, gradient boosting, XGBoost, and multi-layer perceptron (MLP). The gradient boosting model achieves the best performance, with an AUC-ROC of 0.7808 and an F1-score of 0.4693. Exploratory data analysis reveals that defaulted cardholders show a monotonically increasing trend of payment delays, and the distribution of credit utilization ratios differs significantly between defaulters and non-defaulters. Ablation experiments verify that the engineered features improve the AUC value by approximately four percentage points. This work proves the value of financial domain knowledge in feature construction and provides a stable, reproducible benchmark scheme for credit risk assessment.

**Keywords:** Credit card default, Feature Engineering, Integrated learning, Credit risk

## **1. Introduction**

Credit card default risk is a core operational challenge for financial institutions. The continuous expansion of consumer credit scale has aggravated corporate bad debt losses, making accurate default prediction a crucial prerequisite for effective financial risk management [1]. In this study, the UCI Credit Card Dataset is employed, which initially comprises 30,000 customer records [2]. Following data cleaning, 29,965 samples are retained, with default samples accounting for 22.1% of the total. This dataset serves as a widely recognized benchmark for credit prediction tasks [2, 3].

Traditional credit scoring primarily relies on logistic regression, which meets regulatory requirements for model transparency but has limited predictive performance when using only raw features. Ensemble models such as random forests, gradient boosting, and XGBoost offer stronger fitting capabilities, yet suffer from poor interpretability. The trade-off between accuracy and explainability has long been a research focus in the field [4-6]. Multiple empirical comparisons show that ensemble models can improve AUC by 0.02–0.05 compared to linear models. Butaru et al.

built an XGBoost model based on credit card data, achieving an AUC of around 0.775, although they did not establish a complete business feature system [7].

At the level of feature mining, Khandani et al. only constructed simple interaction features, which increased the AUC of the model by 2-3 percentage points at most, and lacked a complete feature framework covering credit utilization rate and weighted overdue score [8]. Industry review shows that most of the literature directly reuses the original variables and ignores the implied risk signals of financial business after reviewing the existing research [9]. Yeh et al. used multi-layer perceptron modeling based on the same UCI data set, and the AUC of the model was only 0.7602, resulting in poor landing performance due to black-box defects [3].

There are two obvious shortcomings in the existing research: one is the lack of systematic domain feature engineering; Second, the division and evaluation criteria of various experimental samples are not unified, and most of them only test a single model, which makes it difficult to achieve a fair horizontal comparison of multiple algorithms [10].

To address the two prominent limitations identified in prior credit risk prediction literature, this study constructs derivative risk indicators grounded in financial business logic and standardizes all experimental settings to implement controlled comparative analysis across multiple machine learning models.

## 2. Methodology

### 2.1. Data preprocessing

The UCI credit card customer data set was released by Yeh in 2009 and uploaded to the UCI database platform, with a total of 30000 original credit records. After cleaning, 29965 valid samples were retained, with default samples accounting for 22.1%, which is the standard open benchmark data set in the field of credit risk [2]. Then, the paper adopts a stratified train-test split with a ratio of 80% / 20% and a random state set to 42. Stratified sampling ensures that the proportion of default users in the training set and test set is consistent, which avoids deviation caused by uneven sample distribution. The splitting rule is defined as:

$$\begin{cases} D_{\text{train}} = 0.8 \times D \\ D_{\text{test}} = 0.2 \times D \end{cases} \quad (1)$$

where  $D$  denotes the full data set,  $D_{\text{train}}$  and  $D_{\text{test}}$  represent the training set and test set, respectively.

### 2.2. Domain-driven feature engineering

The paper designs six categories of derived features based on financial rules, expanding the total feature dimension from 23 to 49. Key calculation formulas are given below.

#### 2.2.1. Credit utilization ratio

It reflects the usage degree of the credit limit:

$$U_i = \frac{Bill_i}{Limit}, \quad i = 1, 2, \dots, 6 \quad (2)$$

where  $Bill_i$  is the bill amount of the  $i$ -th month, and  $Limit$  is the total credit limit. This indicator directly reflects the degree of customer credit overdraft. The higher the overdraft, the greater the risk of default, and it can capture the credit risks caused by the excessive use of credit lines.

### 2.2.2. Repayment ratio

It measures the proportion of actual payment to the monthly bill:

$$R_i = \min \left( \frac{Pay_i}{Bill_i}, 1 \right), i = 1, 2, \dots, 6 \quad (3)$$

$Pay_i$  refers to the actual payment amount of the  $i$ -th month. It can quantify the strength of customers' monthly repayment performance, distinguish between full repayment, minimum repayment, zero default, and other different performance levels, and refine the risk stratification of repayment behavior.

### 2.2.3. Delinquency severity score

The paper quantifies cumulative repayment delay with weighted summation:

$$S = \sum_{i=1}^6 w_i \cdot Dealy_i \quad (4)$$

$Dealy_i$  is the repayment status, and  $w_i$  is the time weight that attaches higher importance to recent repayment behavior.

The method is used to amplify the risk weight of recent overdue, which is more in line with the business logic of "higher recent behavior weight" of financial risk control than the number of simple overdue.

### 2.2.4. Financial trend metrics, interaction & binning features

The financial trend indicator counts the monthly change range of bill and repayment limit, captures the fluctuation law of customers' long-term funds, and identifies potential defaulters with continuous deterioration of revenue and expenditure in advance.

Interaction features capture the joint effect of multiple core variables, while binning features discretize continuous variables to mine potential nonlinear relationships.

## 2.3. Classification models

This paper selects six mainstream monitoring algorithms covering linear models, lazy learning, ensemble trees, and neural networks to conduct credit card default prediction research. As a linear model, logistic regression linearly fits the default probability of samples and delivers interpretable outputs, which can fully meet the disclosure requirements of financial regulatory risk control work. As a typical lazy learning algorithm, K-nearest neighbor (KNN) does not rely on preset data distribution assumptions, so it can serve as a basic control baseline for model comparison in this study. Random forest realizes model training through multi-tree parallel operation and can automatically screen core risk characteristics closely related to customer credit status, such as repayment performance and credit quota. Gradient Boosting optimizes model prediction results by iteratively correcting prediction errors, which presents outstanding advantages in identifying

unbalanced default samples in credit data. As an optimized and upgraded version of the gradient boosting algorithm, XGBoost features faster model training efficiency and is highly applicable to large-scale credit dataset modeling. Multi-layer perceptron (MLP) is capable of capturing complex nonlinear correlation relationships between credit characteristic variables, while it suffers from the inherent defect of unexplainable model output results.

All models input features and hyperparameters uniformly to ensure that the prediction results of multiple models can be fairly compared.

## 2.4. Evaluation metrics

The proportion of credit card default samples was only 22.1%. There was obvious category imbalance in the dataset, and it was impossible to objectively evaluate the ability of the model to identify high-risk customers only by relying on the overall accuracy. Therefore, this paper uses auc-roc, accuracy, accuracy, recall, and F1 score to evaluate the effect of the model.

AUC-ROC can measure the ability of the model to distinguish normal customers from defaulting customers as a whole. The F1 score balances the accuracy rate and recall. It is more suitable for the unbalanced credit classification task and is the core evaluation standard of this experiment.

## 3. Results

### 3.1. Model overall performance

Table 1 summarizes the performance of all six classifiers on the test set. Gradient Boosting achieves the best overall performance on AUC-ROC and F1-score, demonstrating that tree-based ensemble models are more suitable for credit risk prediction tasks.

Table 1. Performance of different classification models

| Model               | AUC-ROC | Accuracy | Precision | Recall | F1-score |
|---------------------|---------|----------|-----------|--------|----------|
| Logistic Regression | 0.7421  | 0.7852   | 0.4125    | 0.3410 | 0.4012   |
| KNN                 | 0.7205  | 0.7713   | 0.3987    | 0.3265 | 0.3891   |
| Random Forest       | 0.7689  | 0.7926   | 0.4352    | 0.3742 | 0.4487   |
| Gradient Boosting   | 0.7808  | 0.8014   | 0.4516    | 0.4025 | 0.4693   |
| XGBoost             | 0.7756  | 0.7982   | 0.4421    | 0.3917 | 0.4589   |
| MLP                 | 0.7602  | 0.7895   | 0.4268    | 0.3681 | 0.4358   |

Note: The best result in each column is marked in bold.

### 3.2. ROC curve comparison

Figure 1 presents ROC curves of all models. The ranking of curves is consistent with the numerical results in Table 1. Gradient Boosting has the largest area under the curve, further verifying its superiority.

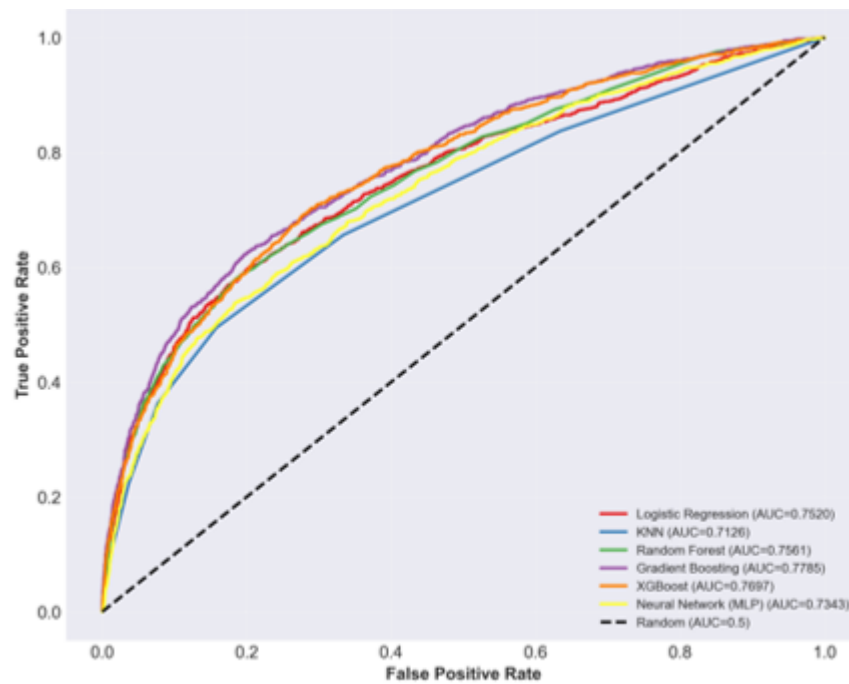


Figure 1. ROC curves of six classification models (photo/picture credit: original)

### 3.3. Ablation experiment of engineered features

The paper conducts ablation experiments to verify the independent contribution of feature engineering. As shown in Table 2, the proposed domain-driven features bring about a 4 percentage point improvement in AUC-ROC, which fully verifies the effectiveness of the feature construction scheme.

Table 2. Ablation results of feature engineering

| Feature Setting           | AUC-ROC |
|---------------------------|---------|
| Only raw features         | 0.7402  |
| Raw + Engineered features | 0.7808  |
| AUC Improvement           | +0.0406 |

This result indicates that the hidden risk information in raw data cannot be fully mined by models alone. Reasonable feature engineering combined with financial domain knowledge can significantly enhance the discriminative ability of features and further improve prediction accuracy.

### 3.4. Exploratory data analysis and visual analysis

The paper analyzes users' repayment and consumption behavior via two groups of line charts. Figure 2 reveals two typical behavioral gaps: overdue duration of default users rises continuously, whereas normal customers' monthly payments can match bill amounts steadily.

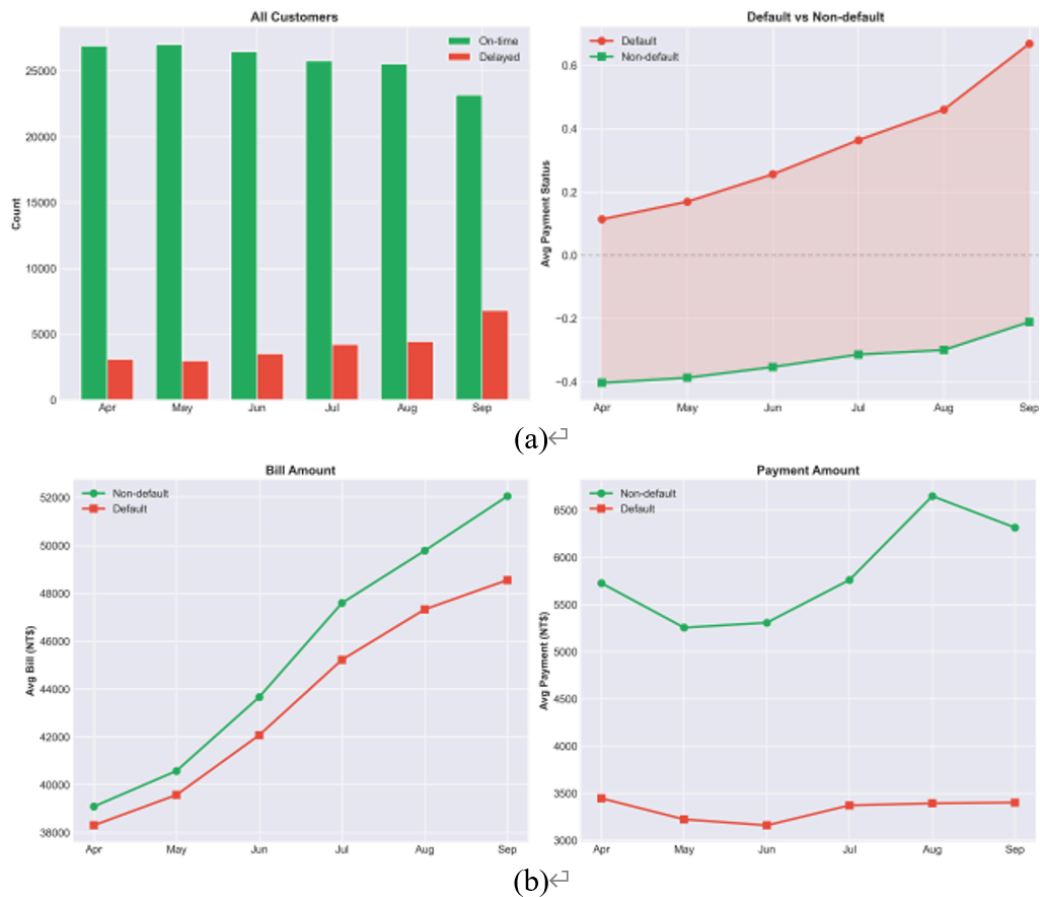


Figure 2. Behavior trends of default and non-default users, (a) Average monthly payment delay status (b) Average monthly bill and payment amount (photo/picture credit: original)

3.5. Feature importance analysis

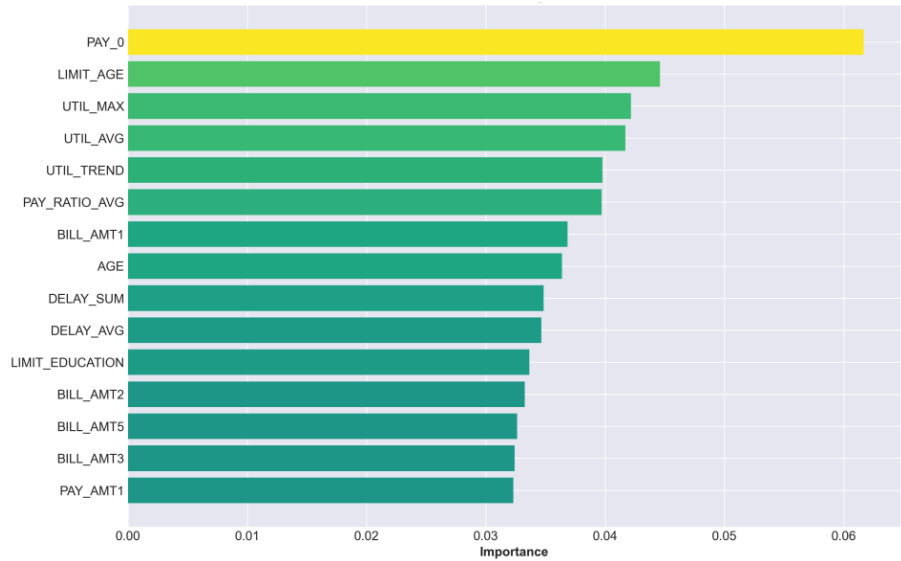


Figure 3. Top 15 feature importance ranking (photo/picture credit: original)

Figure 3 shows the top important features generated by Random Forest. Recent repayment records and the engineered indicators occupy top ranks, which proves the rationality of domain-based feature engineering.

### 3.6. Comparison with existing literature

The experimental conclusions are consistent with mainstream research [9]. Ensemble models outperform linear models on financial tabular data. The performance gain from feature engineering also proves that mining domain information can effectively enhance model performance [8].

### 3.7. Discussion

#### 3.7.1. Comparison of advantages and disadvantages of six types of models

The experimental results show that the comprehensive performance of gradient Boosting is the best, and the nonlinear repayment law in credit data is fully captured by the mechanism of iterative residual correction [5]. Logistic regression can meet the interpretable requirements of financial regulation, but the fitting ability is limited when only the original features are used. The baseline AUC of this dataset is only 0.7421 [3]. KNN is an inert learning algorithm, which is easily disturbed by data noise, and the overall prediction effect is the worst; XGBoost, as the optimized version of gradient Boosting, has higher training efficiency, but the experimental accuracy is slightly lower than gradient Boosting [6]. The multi-layer perceptron has stronger fitting ability, but the black-box structure makes it difficult to output landing risk control rules. The performance differences of various models are consistent with the multi-algorithm benchmark experiment conclusions of Lessmann et al.

#### 3.7.2. Feasible direction of model optimization

Wang et al. Built an LSTM model with an attention mechanism to extract the characteristics of customers' monthly repayment time series, effectively improving the accuracy of credit prediction [11]. Aiming at the problem of insufficient interpretability of tree models, Lundberg and Lee proposed the Shap attribution method to quantify the contribution of a single feature to the default results, taking into account the prediction effect and regulatory transparency requirements [12]. In view of the imbalance of credit sample categories, Shahee et al. Built an improved model by combining oversampling and focal loss, significantly improving the recall rate of default sample identification [13].

The subsequent research can combine time series network and interpretable analysis tools to solve the two major problems of model accuracy and interpretability at the same time.

#### 3.7.3. Limitations of the experimental design in this paper

This experiment only uses the static credit data set of Taiwan, China in 2005. The sample time and geographical coverage are narrow, and the generalization ability of the model needs to be further verified [3]. The characteristics are only based on the static indicators of bill and repayment, and do not include multi-source credit information. The risk mining dimension is missing [9].

The experiment only relies on multiple evaluation indexes to alleviate the evaluation deviation caused by sample tilt, and does not use targeted unbalanced optimization strategies such as sampling and weighted loss; At the same time, the whole set of experiments has no supporting model to

explain the analysis process. The above shortcomings are also common research problems in the current credit modeling field [9].

#### 3.7.4. Future research development ideas

Combined with the existing mature research, further research can be carried out in four directions. First, expand the multi-regional credit samples, use the graph neural network to mine the implicit Lending Association of user groups, and the graph learning credit model of Pourkhoshgoftar et al. Has verified the effectiveness of the scheme [14]. Then, the time series LSTM is introduced to capture the fluctuation law of long-term repayment, and the time series model can stably improve the prediction accuracy compared with traditional static machine learning [11]. What's more, the swap tool is introduced to complete risk attribution and generate standardized credit scoring rules that meet regulatory requirements [12]. Last, expand the characteristic system of multi-source consumption and debt data, and fully tap potential risk signals.

#### 4. Conclusion

Based on the UCI credit card dataset, this paper completes the whole process experiment of credit default prediction, designs six types of derivative risk features based on financial business rules, and expands the original 23-dimensional feature to a 49-dimensional feature. Unified 8:2 stratified sampling, fixed random seeds and other standardized experimental conditions, and completed the horizontal comparison of six mainstream classification algorithms.

Ablation experiments confirmed that domain characteristics could improve the AUC of the model by 0.0406; The comparison results of multiple models showed that Gradient Boosting had the best comprehensive performance, with an AUC-ROC of 0.7808 and an F1 score of 0.4693. The exploratory visual analysis summarizes two typical behavior characteristics of defaulting customers: continuous overdue and low repayment. The importance of random forest characteristics verifies that the derived indicators constructed in this paper have the value of core risk identification. The whole experimental process is standardized and reproducible, which can provide a benchmark reference scheme for retail credit card risk control modeling.

#### References

- [1] Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Society for Industrial and Applied Mathematics.
- [2] Yeh, I.-C. (2009). Default of credit card clients. *UCI Machine Learning Repository*.
- [3] Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473–2480.
- [4] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [5] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- [6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
- [7] Butaru, A., Evans, D., & Ji, L. (2016). Credit card default prediction: A systematic comparison of machine learning methods. *Journal of Credit Risk*, 12(2), 1–40.
- [8] Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787.
- [9] Xia, Y., Liu, J., Liu, B., et al. (2017). Machine learning for credit scoring: A review of current research and future directions. *Expert Systems with Applications*, 74, 224–240.

- [10] Lessmann, S., Baesens, B., Seow, H.-T., et al. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
- [11] Wang, C., Han, D. M., Liu, Q. G., et al. (2018). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. *IEEE Access*, 7, 2161–2168. <https://doi.org/10.1109/ACCESS.2018.2887138>
- [12] Liu, Y. (2023). Credit scoring prediction leveraging interpretable ensemble learning. *Journal of Forecasting*, 43, 286–308.
- [13] Guo, K., & Fan, H. (2024). Research on AdaFocal-XGBoost integrated credit scoring model based on unbalanced data. *Statistics and Applications*, 13(6), 2204–2214.
- [14] Pourkhoshgoftar, S., Shahbahrani, A., & Esmi, N. (2025). Graph-based inductive learning for credit risk prediction with imbalance mitigation. *Computational Economics*.