International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 5, Sep-Oct 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Research on Financial Credit Risk Assessment Model Based on WOE Encoding and Ensemble Learning

Chang Xinyuan

Beijing Wuzi University, Beijing, China

ABSTRACT

Aiming at the limitations of traditional logistic regression in handling high-dimensional nonlinear data and class imbalance in financial credit risk assessment, this paper uses the GiveMeSomeCredit dataset. Missing monthly income values were filled using random forest, skewed features were corrected by log-quantile transformation, expanding features from 4 dimensions to 27; nonlinear features were transformed using WOE encoding, and an ensemble model was constructed combining logistic regression, random forest, and gradient boosting machine, while class imbalance of 1:14 was handled by under sampling and cost-sensitive learning. Experiments show that the GBM model is optimal, improving 9.4% over baseline logistic regression and saving 390,000 yuan annually in bad debt costs; WOE-encoded logistic regression maintains full interpretability and meets regulatory requirements. The study provides support for risk control decision-making in financial institutions.

KEYWORDS: *Credit risk assessment; WOE coding; ensemble learning; class imbalance; feature engineering.*

How to cite this paper: Chang Xinyuan "Research on Financial Credit Risk Assessment Model Based on WOE Encoding and Ensemble Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-5, October 2025, pp.851-853, URL: www.ijtsrd.com/papers/ijtsrd97600.pdf

IJTSRD97600

1. INTRODUCTION

Credit risk assessment is central to the sound operation of financial institutions. Traditional logistic regression offers strong interpretability but limited performance, while complex ensemble models provide high accuracy but struggle to meet regulatory transparency requirements, and existing studies often overlook the quantification of business value.

1.1. Research Background

After the COVID-19 pandemic, the global financial non-performing loan (NPL) rate increased by 1.2 percentage points (according to BIS data), and the balance of non-performing loans in Chinese commercial banks rose; Basel III requires KS>0.3, GDPR prohibits 'purely automated decisions,' and regulation is tightening. At the same time, behavioral data brought by digital transformation (such as payment frequency and device location) enhance risk identification potential but also increase the complexity of feature processing. Logistic regression, due to its parameter interpretability and high training efficiency, remains the core model for FICO credit scores and Lending Club, with a strong industry foundation.

1.2. Core Contributions

Build an end-to-end automated modeling process, reducing manual intervention by 60%; balance performance and interpretability, with WOE-encoded logistic regression and GBM adapted to different scenarios; quantify business value, with the optimal model saving 390,000 yuan annually in bad debt costs.

2. Data Preparation and Preprocessing

2.1. Overview of the Dataset

The Kaggle public dataset GiveMeSomeCredit contains 150,000 records, 11 original features, and 1 binary target (SeriousDlqin2yrs, where 1 indicates serious delinquency). It is split 70/30 for training/testing, with a default rate of 6.68% in both sets.

2.2. Missing Value Handling

For monthly income(with a missing rate of 13.8%), Random Forest Regression was used for imputation (n_estimators=100, max_depth=10). Compared with mean imputation, this method increased the subsequent Logistic Regression AUC by 2.9 percentage points.

For the number of family dependents(with a missing rate of 2.67%), MICE (Multiple Imputation by Chained Equations) was adopted for imputation, which maintained the stability of the data distribution.

2.3. Skewed Distribution Correction

For features with skewness > 2 (e.g., unsecured credit line utilization rate), the method of "log1p + quantile transformation (n_quantiles=100)" was applied. After correction, the skewness was < 0.3, and the data passed the Kolmogorov-Smirnov (K-S) normality test, thus eliminating dimensional differences.

3. Feature Engineering and Variable Expansion

Features were expanded from 4 dimensions to 27, with the core contents as follows:

Debt Burden: Total debt amount (DebtRatio × MonthlyIncome); actual debt-to-income ratio (clipped to the range [0,10]).

Family Pressure: Per capita income (MonthlyIncome / (number of family dependents + 1)); family burden coefficient.

Interaction and Behavior: 3 interaction features (e.g., high income × number of delinquencies); 3 behavioral features derived from consolidating delinquency records (including total number of delinquencies and delinquency severity index). This alleviated multicollinearity, reducing the Variance Inflation Factor (VIF) from over 40 to 1.2.

Distribution Transformation: Log1p transformation was performed on variables such as income and debt ratio to make their distributions closer to the normal distribution.

4. Model Building and Optimization

4.1. WOE Encoding and Logistic Regression

4.1.1. Principles of WOE Encoding and Binning Methods

WOE encoding is implemented through "equal-frequency binning and chi-square merging (p>0.05)", with the formula as follows:

$$WOEi = ln\left(\frac{\frac{Good\_i}{Good\_Total}}{\frac{Bad\_i}{Bad\_Total}}\right)$$

4.1.2. IV Screening and Model Performance

IV Screening and Modeling Keep 15 core features with IV>0.02 (top 3: Unsecured Credit Utilization IV=1.06, Total Delinquency Count IV=1.05, Delinquency Severity IV=0.87). After WOE encoding, the logistic regression test set AUC=0.8521, and parameters are traceable.

4.2. Ensemble Models and Hyperparameter Optimization

Core optimal parameters: Logistic Regression: C=0.1, L2 regularization; Random Forest: n_estimators=50, max_depth=10; GBM: n_estimators=100, max_depth=4, learning_rate=0.1;Soft Voting Ensemble: weights [3,2,1] (Logistic Regression 3, Random Forest 2, GBM 1).

4.3. Handling Class Imbalance

A combined strategy of undersampling (training set 14,000 samples, 1:1 balance) and cost-sensitive learning (class_weight=balanced) increased the recall rate of bad customers from 38% to 71.7%, turning an annual loss of 120,000 into savings of 390,000.5 Experimental Results and Analysis

5. Experimental Results and Analysis

TABLE I.

Model	AUC	Recall Rate (%)	KS Value
Logistic Regression (WOE Encoding)	0.8521	62.5	0.48
Random Forest	0.8620	69.7	0.51
GBM (Gradient Boosting Machine)	0.8646	71.7	0.53
Soft Voting Ensemble	0.8624	69.4	0.50

GBM achieved the optimal performance, the Soft Voting Ensemble had the lowest overfitting, and the WOE-based Logistic Regression exhibited the best interpretability.

5.1. Feature Importance and Business Insights

The top 5 features in the GBM model are as follows:

Total Number of Delinquencies (0.28, positive correlation)

Delinquency Severity (0.22, positive correlation)

Unsecured Credit Line Utilization Rate (0.18, risk surges sharply when exceeding 80%)

High Income × Delinquency (0.12, positive correlation)

Actual Debt-to-Income Ratio (0.08, risk increases when exceeding 1.5)

This confirms that historical behavior and debt burden are core risk signals.

5.2. Quantification of Business Value

For GBM's optimal threshold (0.59): The proportion of users approved automatically reached 78%, manual review costs were reduced by 44%, and annual bad debt savings amounted to 390,000 yuan.

For threshold 0.3: Although the recall rate reached 89%, the review cost was high, resulting in only 120,000 yuan in annual bad debt savings.

6. Conclusions and Recommendations

6.1. Research Conclusions

Data preprocessing increased the model's AUC by 2.9 percentage points; Core features from feature

engineering contributed 45% to the AUC and eliminated multicollinearity; GBM improved the AUC by 9.4% compared with the baseline model; The class imbalance combination strategy was utilized to achieve the transition from cost to profit.

6.2. Practical Recommendations

Data Level: Improve behavior and asset data, and establish a "missing-anomaly-skewness" monitoring mechanism.

Model Level: Adopt a three-stage architecture: pre-screening (WOE-based Logistic Regression) - fine-screening (GBM) - post-loan (online learning), and generate risk reports using SHAP.

Business Level: Optimize thresholds by scenario, and establish a dual "technology-business" monitoring system.

References

[1] White, M. A. (1996) Environmental Finance: Value and Risk in an Age of Ecology. Business Strategy and the Environ ment, 5, 198-206.

[2] Feng, Y. (2022) Bank Green Credit Risk Assessment and Management by Mobile Computing and Machine Learning Neural Network under the Efficient Wireless Communication. Wireless Communications and Mobile Computing, 2022, Article 3444317.