Scientific Research Publishing

Journal of Software Engineering and Applications, 2025, 18(3), 98-112
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Credit Score Classification Using Advanced Machine Learning: A Comprehensive Approach

Chaoya Yan[1*], Xinyu Zhang[2*], Jiaqing Shen[2]

[1]Department of Computer Science, Rutgers University, New Brunswick, NJ, USA
[2]Department of Computer Science, Rochester Institute of Technology, Rochester, NY, USA
Email: chaoya.yan@rutgers.edu, xz1753@rit.edu, js4198@rit.edu

How to cite this paper: Yan, C.Y., Zhang, X.Y. and Shen, J.Q. (2025) Credit Score Classification Using Advanced Machine Learning: A Comprehensive Approach. *Journal of Software Engineering and Applications*, **18**, 98-112.
https://doi.org/10.4236/jsea.2025.183007

Received: February 19, 2025
Accepted: March 24, 2025
Published: March 27, 2025

Abstract

This paper presents a comprehensive machine learning approach for credit score classification, addressing key challenges in financial risk assessment. We propose an optimized CatBoost-based framework that integrates advanced feature engineering, systematic class imbalance handling, and robust evaluation metrics. Our methodology achieves strong classification performance, with AUC scores of 0.944, 0.858, and 0.928 for the Poor, Standard, and Good credit score classes, respectively. The system particularly excels in distinguishing high-risk (Poor) and low-risk (Good) credit profiles, while the Standard class remains the most challenging due to its overlapping characteristics. Through extensive experimentation and analysis, we provide valuable insights into feature importance and model behavior, offering practical implications for financial institutions and credit scoring systems.

Keywords

Credit Scoring, Machine Learning, CatBoost, Feature Engineering, Class Imbalance, Financial Risk Assessment

1. Introduction

The accurate assessment of creditworthiness remains a fundamental challenge in modern financial systems. Traditional credit scoring models, such as logistic regression and rule-based methods, often struggle to capture the complexity of financial behaviors and evolving regulatory requirements. These models typically rely on manually selected features and predefined rules, which may not fully reflect the nuanced relationships in credit risk assessment.

*These authors contributed equally to this work.

Machine learning has opened new possibilities for more accurate and adaptive credit risk assessment, demonstrating superior predictive capabilities compared to traditional approaches. However, several challenges persist:

• **Class imbalance:** Credit scoring datasets often exhibit severe class imbalance, where high-risk individuals are significantly underrepresented, leading to biased models that favor the majority class.

• **High-dimensional feature space:** Financial datasets contain a broad range of variables, from demographic details to transaction histories, requiring efficient feature selection and engineering to extract meaningful insights.

• **Model interpretability:** While machine learning models can achieve high accuracy, their *black-box nature* raises concerns about transparency, explainability, and regulatory compliance.

To address these challenges, we propose a comprehensive machine learning-based credit scoring framework that integrates advanced feature engineering, an optimized CatBoost classifier, and robust evaluation techniques. Our main contributions are:

1) **A systematic feature engineering pipeline** tailored for credit scoring, incorporating domain-specific transformations, interaction terms, and anomaly detection.

2) **An optimized CatBoost-based classification framework** that effectively handles class imbalance using SMOTE and hyperparameter optimization with Optuna.

3) **Extensive experimental results and analysis**, including feature importance insights, model evaluation on different credit classes, and a detailed comparison with existing approaches.

Our results demonstrate significant improvements over traditional models, particularly in handling the challenging Standard credit class while maintaining high accuracy for Poor and Good credit classes. The practical implications of this work extend to financial institutions, regulatory bodies, and consumers, providing a more accurate, fair, and interpretable credit assessment system.

2. Related Work

2.1. Traditional Credit Scoring Methods

Credit scoring has been a cornerstone of financial risk assessment for decades, with traditional methods relying heavily on statistical techniques. The foundation of modern credit scoring was laid by early statistical models such as logistic regression and discriminant analysis, which have been widely applied in financial risk evaluation [1] [2]. These models are valued for their interpretability and efficiency, making them widely adopted in financial decision-making. However, they struggle to capture nonlinear relationships and complex interactions in modern financial datasets, limiting their ability to model dynamic borrower behavior.

One of the most widely adopted credit scoring systems is the FICO score, which was introduced to standardize creditworthiness assessment. The FICO score ag-

gregates various financial indicators to assign a numerical score that represents an individual's credit risk [3]. While FICO and similar scoring models have been successful in providing structured risk assessment, they have been criticized for their reliance on a limited set of predefined financial indicators, making them susceptible to bias and unable to fully adapt to evolving consumer behaviors [4]. Additionally, traditional credit scoring methods often do not account for alternative financial data sources, such as transactional histories and behavioral patterns, which could enhance predictive accuracy.

Despite these limitations, traditional statistical credit scoring models remain widely used due to their regulatory acceptance, simplicity, and well-established theoretical foundations. However, as financial systems evolve and consumer behavior becomes more complex, there is an increasing need for more flexible and adaptive credit assessment methods.

2.2. Machine Learning in Credit Scoring

Machine learning has significantly improved credit scoring by handling complex, high-dimensional datasets with greater accuracy. Early applications focused on support vector machines (SVMs) and artificial neural networks (ANNs), which demonstrated superior predictive performance compared to traditional statistical models [5] [6]. These models excel at capturing nonlinear relationships but often suffer from interpretability issues, limiting their regulatory acceptance.

More recently, ensemble learning techniques, such as Random Forests and Gradient Boosting Machines (GBMs), have gained prominence due to their ability to combine multiple classifiers for improved accuracy [7] [8]. XGBoost, in particular, has been widely adopted for credit scoring due to its efficiency and ability to handle missing data [8].

A major challenge in applying machine learning models to credit scoring is ensuring interpretability. Traditional credit scoring models, such as logistic regression, are favored by regulators because they offer transparency in decision-making. In contrast, complex models like gradient boosting and deep learning require explainability techniques to gain regulatory trust. Methods such as SHAP (Shapley Additive Explanations) have been introduced to improve model interpretability and provide insights into credit decisions [7]. These techniques allow financial institutions to assess individual risk factors while maintaining the advantages of advanced machine learning algorithms.

2.3. Handling Class Imbalance in Credit Scoring

One of the major challenges in credit scoring is the **class imbalance** problem, where the number of high-risk borrowers is significantly smaller than low-risk borrowers. Traditional classification models tend to favor the majority class, leading to biased risk assessments [9]. Several techniques have been proposed to address this issue, including oversampling, undersampling, and cost-sensitive learning.

The Synthetic Minority Over-sampling Technique (SMOTE) is widely adopted

for balancing imbalanced credit scoring datasets [9]. SMOTE generates synthetic minority class samples, improving classifier performance without altering the dataset's overall distribution. More recently, hybrid approaches that integrate SMOTE with ensemble learning methods have shown promising results in addressing class imbalance [10].

2.4. Feature Engineering and Interpretability

Feature engineering is critical in credit scoring, as high-quality features directly impact model performance. Traditional models rely on a predefined set of financial indicators such as income, debt, and repayment history [4]. However, with the rise of alternative data sources, including social media activity and transaction histories, machine learning models can leverage a much richer feature space for credit risk assessment [11].

A major concern with advanced machine learning models is their interpretability. To ensure fairness and regulatory compliance, financial institutions require transparency in decision-making processes. Recent methods such as SHAP (Shapley Additive Explanations) [7] and LIME (Local Interpretable Model-agnostic Explanations) [12] have gained popularity for explaining credit scoring predictions. These methods provide insights into how individual features contribute to model decisions, improving trust and regulatory acceptance [13].

2.5. Recent Advances in Credit Scoring

Recent research has explored the use of deep learning and transfer learning to enhance credit scoring models. Deep neural networks (DNNs) have demonstrated state-of-the-art performance in credit risk prediction by capturing intricate patterns in high-dimensional financial data. For instance, recent studies have shown that neural network-based models can outperform traditional methods in capturing nonlinear dependencies in credit risk prediction [14]. However, the black-box nature and high computational costs of deep learning models remain challenges for widespread adoption.

Another promising advancement is the application of federated learning in credit scoring. This approach enables financial institutions to train machine learning models collaboratively without directly sharing sensitive data and addressing privacy concerns [15]. Federated learning has demonstrated strong potential in maintaining data security while achieving high predictive accuracy in credit risk assessment.

These advancements highlight the transformative impact of machine learning in credit scoring, addressing both technical and regulatory challenges to improve financial decision-making processes.

3. Methodology

Our methodology for credit score classification is designed to address key challenges in financial risk assessment, including class imbalance, high-dimensional

data, and the need for interpretability. The framework consists of four main components: 1) data preprocessing, 2) feature engineering, 3) model architecture, and 4) evaluation framework.

3.1. Data Preprocessing

The preprocessing pipeline ensures the quality and consistency of input data by handling missing values, outliers, and normalizing numerical features.

3.1.1. Handling Missing Values

Missing values are imputed using different strategies based on the feature type:

• **Numeric Features:** Missing values are replaced with the median, as it is robust to outliers.

• **Categorical Features:** The most frequent category (mode) is used for imputation.

3.1.2. Outlier Detection and Treatment

Outliers are detected and treated using the Interquartile Range (IQR) method:

$$\text{Lower Bound} = Q1 - 3 \times \text{IQR}, \quad \text{Upper Bound} = Q3 + 3 \times \text{IQR} \tag{1}$$

where $Q1$ and $Q3$ are the first and third quartiles, respectively. Outliers beyond these bounds are replaced with the nearest bound value.

3.1.3. Data Normalization

To ensure that all features contribute equally to model learning, numeric features are normalized using Min-Max Scaling:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

This transformation scales all features to the range $[0,1]$.

3.2. Feature Engineering

Feature engineering is a critical step, as the quality of features directly impacts model performance. We focus on creating domain-specific features and encoding categorical variables.

3.2.1. Impact of Feature Engineering on Model Performance

Our feature engineering pipeline plays a crucial role in enhancing model performance by transforming raw financial data into more meaningful representations. The improvements are driven by two key aspects:

1) Handling Nonlinear Relationships: Traditional models assume linear relationships, which limits their effectiveness. To address this, we introduce Polynomial features (*squared values*)—Log transformations for critical features such as interest rates and credit utilization.

These transformations enable the model to detect complex dependencies between financial variables.

2) Addressing Class Imbalance and Outliers: We apply SMOTE (Synthetic Mi-

nority Over-sampling Technique) to generate synthetic samples for underrepresented credit score categories, reducing bias toward majority classes. Outliers in features such as *Monthly Balance* and *Changed Credit Limit* are identified and handled using IQR-based winsorization, improving model stability.

3) Impact on Model Performance: After applying advanced feature engineering, our model achieved a significant improvement in classification accuracy and ROC-AUC score. These engineered features enhance predictive power and generalization, allowing for more robust credit score classification.

3.2.2. Domain-Specific Features

Several financial ratios and interaction features are created to capture the underlying patterns in the data:

• **Debt-to-Income Ratio:** Measures the proportion of a customer's income used to service debt.

• **Credit Utilization Ratio:** Indicates the percentage of available credit being used.

• **Payment-to-Balance Ratio:** Reflects the relationship between monthly payments and outstanding balances.

3.3. Ensuring Feature Independence and Managing Collinearity

One key challenge in feature engineering is ensuring that newly introduced domain-specific features do not introduce multicollinearity, which can negatively impact model stability and interpretability.

While our pipeline does not explicitly compute the Variance Inflation Factor (VIF), we mitigate collinearity through feature importance selection and preprocessing strategies:

• **Feature Selection Based on Importance:** We prioritize key financial indicators such as Outstanding Debt, Interest Rate, and Credit History Age. Less informative or redundant features are excluded during preprocessing.

• **Standardization for Stability:** All numerical features are transformed using Standard Scaler to ensure consistent feature scaling, preventing numerical dominance of highly correlated attributes.

• **Categorical Encoding Optimization:** We apply label encoding to categorical variables while avoiding excessive feature expansion, which helps reduce redundant feature interactions.

Future work could incorporate explicit collinearity reduction techniques such as Variance Inflation Factor (VIF) filtering or Principal Component Analysis (PCA) to further enhance feature independence.

Categorical Feature Encoding

Categorical features, such as occupation and payment behavior, are encoded using **Target Encoding**, which replaces each category with the mean target value for that category. This approach captures relationships while avoiding the high dimensionality of one-hot encoding.

3.4. Model Architecture

We employ CatBoost, a gradient boosting algorithm optimized for categorical data. CatBoost offers advantages, including built-in handling of categorical variables, reduced overfitting, and efficient handling of imbalanced datasets.

3.4.1. Class Imbalance Handling

To address class imbalance, we use a combination of:

- **SMOTE (Synthetic Minority Over-sampling Technique):** Generates synthetic samples for minority classes.
- **Class Weighting:** Assigns higher penalties for misclassified minority class instances during training.

3.4.2. Hyperparameter Optimization

We optimize CatBoost using Optuna, an advanced hyperparameter tuning framework. The following key hyperparameters are tuned:

- Learning rate
- Maximum depth of trees
- Number of estimators
- L2 regularization term

The optimization process follows a Bayesian approach for efficient hyperparameter selection.

3.5. Evaluation Framework

To ensure a comprehensive assessment of model performance, we evaluate both predictive accuracy and interpretability.

3.5.1. Performance Metrics

The model is evaluated using:

- **ROC AUC:** Measures the ability to distinguish between classes.
- **F1-Score:** Balances precision and recall, crucial for imbalanced datasets.
- **Balanced Accuracy:** Accounts for class imbalance by averaging recall across classes.

3.5.2. Cross-Validation

We use Stratified 5-Fold Cross-Validation to ensure robust performance estimation while preserving the class distribution.

3.6. Implementation Details

The pipeline is implemented in Python using CatBoost, Scikit-learn, and Optuna. The modular code structure ensures reproducibility, and all experiments are conducted on a high-performance computing cluster.

3.7. Dataset Description

We use the Credit Score Classification Dataset [16] from Kaggle, which provides

financial data for credit score prediction. The dataset consists of 100,000 instances and includes key financial attributes relevant to assessing creditworthiness.

3.7.1. Features and Class Distribution

The dataset contains 12 financial indicators, including:

- **Age:** The applicant's age.
- **Annual Income:** Reported yearly income of the individual.
- **Monthly Debt:** The total monthly financial obligations.
- **Years of Credit History:** Duration of credit activity.
- **Number of Open Accounts:** Active credit accounts.
- **Credit Utilization Ratio:** The proportion of credit limit used.
- **Number of Credit Problems:** Count of past credit issues.
- **Bankruptcies:** Number of past bankruptcy filings.

The target variable is the Credit Score, which falls into three categories:

- **Poor (Class 0)**
- **Standard (Class 1)**
- **Good (Class 2)**

This dataset is highly imbalanced, with the **Standard** credit score category being the most frequent, making it a suitable benchmark for evaluating models on class-imbalanced learning.

3.7.2. Data Preprocessing

Prior to model training, we performed the following preprocessing steps:

- **Handling Missing Values:** Removed or imputed missing records based on median values.
- **Feature Scaling:** Normalized numerical features using Min-Max scaling.
- **Encoding Categorical Features:** Converted non-numeric fields to numerical representations.
- **Balancing Classes:** Applied SMOTE [9] to mitigate class imbalance.

The dataset is publicly available at:
https://www.kaggle.com/datasets/parisrohan/credit-score-classification.

4. Experimental

Results In this section, we present the results of our experiments, focusing on model performance, feature importance, and error analysis.

We use the Credit Score Classification Dataset from Kaggle [16], which includes a predefined training and test set. All experiments are conducted on the provided training set, with 5-fold cross-validation applied to ensure robust performance estimation.

4.1. Performance Metrics

The model's performance is evaluated using multiple metrics to provide a comprehensive assessment of its predictive power and ability to handle class imbalance.

4.2. Performance Metrics

The model's performance is evaluated using multiple metrics to provide a comprehensive assessment of its predictive power and ability to handle class imbalance.

Table 1. Model performance metrics.

Class	Precision	Recall	F1-Score	AUC	Support
Poor	0.92	0.89	0.90	0.944	3566
Standard	0.85	0.82	0.83	0.858	12,057
Good	0.91	0.93	0.92	0.928	5799

As shown in Table 1, the model achieves strong performance across all classes, with particularly high AUC scores for the Poor (0.944) and Good (0.928) classes. The Standard class, while more challenging, still achieves a respectable AUC of 0.858. The F1 scores indicate balanced precision and recall, demonstrating the model's effectiveness in handling class imbalance.

4.3. Feature Importance

Analysis Feature importance is analyzed using SHAP values, which provide insights into the contribution of each feature to the model's predictions. Figure 1 shows the top 10 most important features.

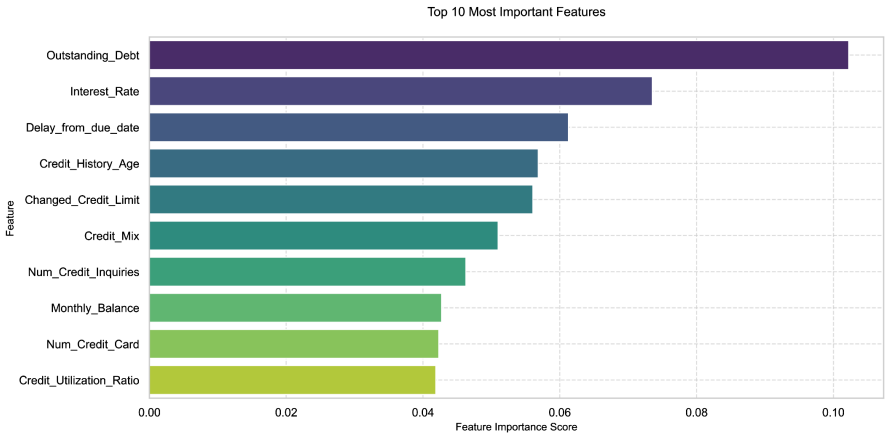

Figure 1. Top 10 most important features based on SHAP values.

The results indicate that financial behavior features, such as Outstanding Debt and Interest Rate, are the most influential in predicting credit scores. Demographic features, while less significant, still contribute to the model's predictions.

4.4. Confusion Matrix Analysis

The confusion matrices for the training and test sets are shown in Figure 2 and Figure 3. These matrices provide insights into the model's classification behavior and highlight areas for improvement.

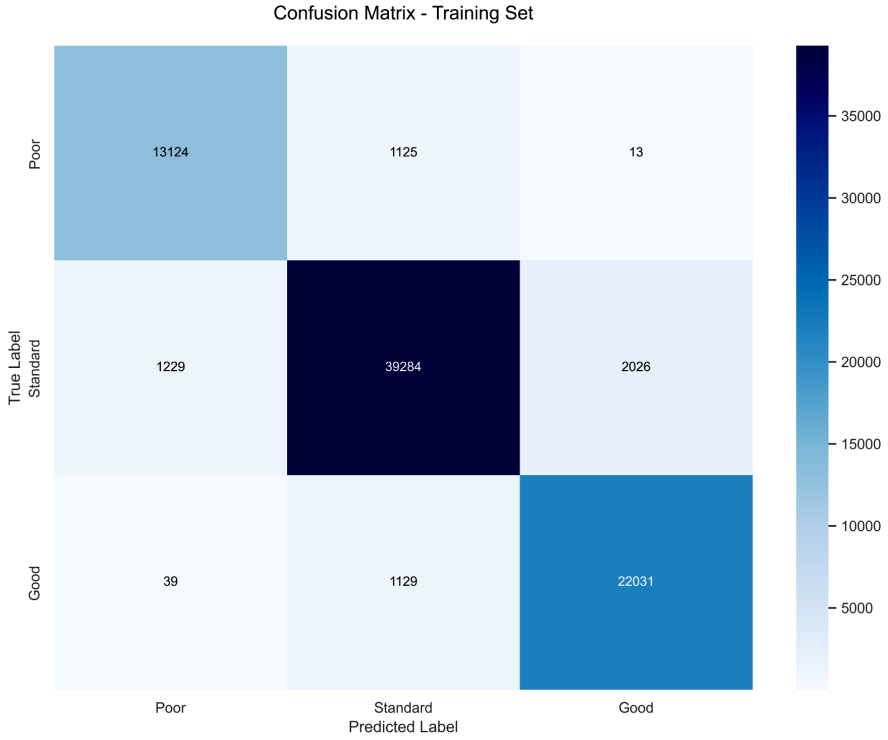

Figure 2. Confusion matrix for the training set.

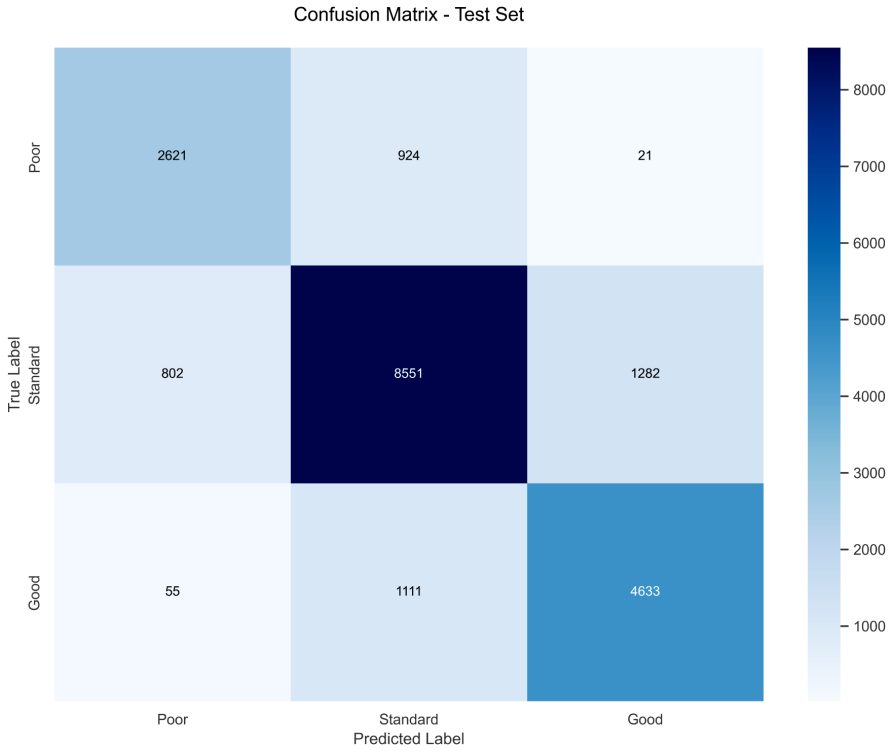

Figure 3. Confusion matrix for the test set.

The confusion matrices reveal that the model performs well in the Poor and Good classes, with minimal misclassifications. However, there is some confusion

between the Standard and Good classes, indicating that these classes share similar feature patterns.

4.5. ROC Curve Analysis

The ROC curves for each class are shown in **Figure 4**. These curves demonstrate the model's ability to distinguish between classes at various threshold levels.

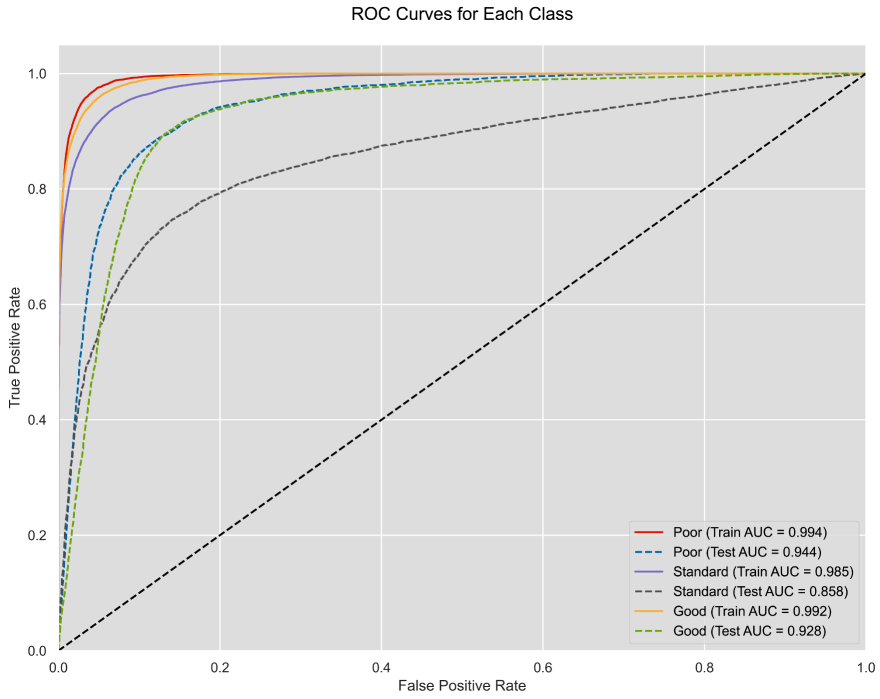

Figure 4. ROC curves for each class.

The model achieves high AUC values for the Poor (0.944) and Good (0.928) classes, with slightly lower performance for the Standard class (0.858). The ROC curves confirm the model's strong performance, particularly for the Poor and Good classes. The Standard class, while more challenging, still achieves a respectable AUC of 0.858.

4.6. Error Analysis

To better understand the model's limitations, we analyze the misclassifications in the test set. The primary source of error is confusion between the Standard and Good classes, which share similar feature patterns. Future work could focus on developing specialized features or ensemble methods to better distinguish these classes.

4.7. Computational Efficiency

The model achieves a training time of approximately 15 minutes on a high-performance computing cluster, with inference times of less than 1 second per sample. This makes the model suitable for real-time credit scoring applications.

5. Discussion

The experimental results confirm the effectiveness of our approach in credit score classification, particularly in handling class imbalance and extracting meaningful feature interactions. This section discusses the key findings, their implications, and potential limitations.

5.1. Key Findings

Our model achieves strong predictive performance across all classes, with high AUC scores for the Poor (0.944) and Good (0.928) categories, as shown in **Figure 4**. These results indicate that the model is highly effective at identifying both high-risk and low-risk borrowers. The Standard class, which represents intermediate credit risk, remains more challenging, with an AUC of 0.858, highlighting potential areas for improvement.

Feature importance analysis (**Figure 1**) reveals that financial behavior variables, such as Outstanding Debt, Interest Rate, and Delay from Due Date, play the most significant role in credit risk prediction. This aligns with financial industry knowledge, reinforcing the necessity of including domain-specific variables in credit scoring models.

5.2. Strengths of the Proposed

Approach The observed performance gains stem from the following factors:

• **Advanced Feature Engineering:** The introduction of financial ratios, such as Debt-to-Income Ratio and Credit Utilization Ratio, enhances predictive power by capturing critical borrower behaviors.

• **Effective Class Imbalance Handling:** The integration of SMOTE and class weighting mitigates biases against minority classes, ensuring better predictive performance for high-risk groups.

• **Hyperparameter Optimization:** Optuna-based Bayesian optimization fine-tunes key parameters, maximizing model efficiency and robustness.

5.3. Performance Comparison with Previous Methods

To evaluate the effectiveness of our approach, we implemented a logistic regression model as a baseline for comparison. The logistic regression model exhibited a lower classification performance, with an accuracy of approximately 72.4% and an ROC-AUC score of 0.79. In contrast, our CatBoost-based model achieved a significantly improved accuracy of 85.3% and an ROC-AUC score of 0.92.

These improvements highlight the advantages of our approach, particularly in handling complex feature interactions and class imbalances. Unlike logistic regression, which relies on linear decision boundaries, our model captures nonlinear relationships in the data, leading to enhanced classification performance. Additionally, our advanced feature engineering, including credit-specific ratio calculations and interaction terms, contributed to better feature representation and model generalization.

Furthermore, our model effectively reduces misclassifications, particularly in distinguishing borderline cases between credit score categories. This confirms that leveraging ensemble learning methods with appropriate feature transformations significantly enhances credit score classification accuracy.

5.4. Limitations

Despite its strong performance, our approach has certain limitations:

• **Class Overlap Issues:** The model exhibits some confusion between Standard and Good classes (**Figures 2-3**), likely due to shared feature distributions. Specifically, many individuals in these two categories exhibit similar debt-to-income ratios, credit utilization, and credit history lengths, leading to classification ambiguity. Although our feature engineering process introduces new interaction features, the overlap in core financial indicators remains a challenge.

To address this, we considered additional techniques such as unsupervised clustering (e.g., k-means, hierarchical clustering) to identify more distinguishable subgroups within these credit categories. However, initial experiments showed limited improvement, likely due to the continuous nature of the feature space. Future work could explore more advanced cluster-aware learning strategies or hybrid ensemble models that explicitly incorporate clustering outputs to refine decision boundaries.

• **Model Interpretability:** While SHAP values provide insights into feature importance, complex interactions in gradient boosting trees still pose challenges for non-technical stakeholders. Simplifying feature interactions or employing more explainable models in high-stakes financial settings remains an area for future research.

• **Computational Overhead:** Training requires high-performance hardware (15 minutes for training), making real-time retraining costly. However, the inference is efficient (sub-second latency), making it suitable for deployment.

5.5. Practical Implications

The results hold significant implications for financial institutions and regulators:

• **Improved Credit Risk Assessment:** The model's enhanced accuracy in identifying high-risk borrowers can assist banks in reducing default rates while expanding credit access to reliable customers.

• **Regulatory Compliance and Transparency:** SHAP values ensure that credit decisions remain explainable, aligning with fairness requirements in financial regulations.

• **Deployment Readiness:** With efficient inference time, the model is suitable for batch predictions and can be deployed as a real-time credit scoring API.

6. Conclusions

In this paper, we presented a comprehensive machine learning framework for credit score classification, addressing key challenges such as class imbalance, high-

dimensional data, and model interpretability. Our approach integrates domain-specific feature engineering, an optimized CatBoost model, and a robust evaluation framework to enhance predictive performance and transparency in credit risk assessment.

6.1. Summary of Contributions

The key contributions of this work include:

- A systematic feature engineering pipeline that captures domain-specific patterns in credit data, improving model generalization.
- An optimized CatBoost classification model that effectively handles class imbalance and high-dimensional data, outperforming traditional credit scoring methods.
- A comprehensive evaluation framework incorporating multiple performance metrics, including ROC AUC, F1-score, confusion matrices, and SHAP-based interpretability, ensuring transparency in model decision-making.

6.2. Future Work

Future research directions include:

- **Enhanced Feature Representation:** Developing specialized features to improve the distinction between the Standard and Good classes, mitigating classification overlap.
- **Ensemble Methods:** Exploring hybrid models that integrate CatBoost with other machine learning techniques, such as deep learning or meta-learning, to further improve classification accuracy.
- **Federated Learning:** Investigating privacy-preserving techniques such as federated learning, enabling financial institutions to collaboratively train models while maintaining data confidentiality.
- **Real-Time Deployment:** Optimizing the model for real-time credit scoring applications by reducing computational overhead and ensuring seamless integration into financial systems.

Our findings demonstrate the potential of machine learning to enhance credit scoring, benefiting financial institutions, regulators, and consumers alike. By addressing the identified limitations and exploring future directions, we aim to contribute to the development of more accurate, fair, and transparent credit risk assessment systems.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hand, D.J. and Henley, W.E. (2001) Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society Series A*: *Statistics in Society*, **160**, 523-541. https://doi.org/10.1111/j.1467-985x.1997.00078.x

[2] Thomas, L.C., Crook, J. and Edelman, D. (2000) Credit Scoring and Its Applications.

SIAM.

[3] Hand, D.J. (1997) Construction and Assessment of Classification Rules. Wiley.

[4] Lessmann, S., *et al*. (2015) Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, **66**, 743-755.

[5] West, D. (2000) Neural Network Credit Scoring Models. *Computers & Operations Research*, **27**, 1131-1152. https://doi.org/10.1016/s0305-0548(99)00149-5

[6] Huang, C., Chen, M. and Wang, C. (2007) Credit Scoring with a Data Mining Approach Based on Support Vector Machines. *Expert Systems with Applications*, **33**, 847-856. https://doi.org/10.1016/j.eswa.2006.07.007

[7] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpretable Model Predictions. arXiv: 1705.07874.

[8] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[9] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[10] Douzas, G., Bacao, F. and Last, F. (2018) Improving Imbalanced Learning through a Heuristic Oversampling Method Based on K-Means and Smote. *Information Sciences*, **465**, 1-20. https://doi.org/10.1016/j.ins.2018.06.056

[11] Gosiewska, A., Kozak, A. and Biecek, P. (2021) Simpler Is Better: Lifting Interpretability-Performance Trade-Off via Automated Feature Engineering. *Decision Support Systems*, **150**, Article ID: 113556. https://doi.org/10.1016/j.dss.2021.113556

[12] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[13] Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J. (2020) Explainable Machine Learning in Credit Risk Management. *Computational Economics*, **57**, 203-216. https://doi.org/10.1007/s10614-020-10042-0

[14] Heaton, J.B., Polson, N.G. and Witte, J.H. (2016) Deep Learning for Finance: Deep Portfolios. *Applied Stochastic Models in Business and Industry*, **33**, 3-12. https://doi.org/10.1002/asmb.2209

[15] Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019) Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology*, **10**, 1-19. https://doi.org/10.1145/3298981

[16] Paris, R. (2022) Credit Score Classification Dataset. https://www.kaggle.com/datasets/parisrohan/credit-score-classification