Optimizing Credit Card Approval Decisions Through Advanced Gradient Boosting Techniques

K. Hema* [1], Dr. M. Arathi [2]

[1,2] Computer Science & Engineering, JNT University Hyderabad.

Abstract

Credit card approval systems need accurate, reliable and interpretable predictive models to support financial decision making under the condition of class imbalance. This work examines the performance of more sophisticated approaches to gradient boosting for credit card approval decisions based on the UCI Credit Card Default dataset made freely available to the public. The quantitative, experiment framework is taken to evaluate AdaBoost, LightGBM, XGBoost, and optimized CatBoost model under the same preprocessing, sampling and evaluation condition. Cost-sensitive learning is used to solve class imbalance, categorical feature correction, early stopping, and decision threshold optimization are combined to improve the detection of minority class. Experimental results show that the proposed CatBoost framework can achieve better performance compared with baseline and competing models, which achieves a ROC-AUC of 0.7809, F1-score of 0.5490 and balanced accuracy of 0.7157 on a held-out test set. Statistical validation through McNemar's and Wilcoxon tests prove that the observed gains in performance are statistically significant. Calibration and interpretability analyses provide further support for the reliability and transparency of the proposed approach. The results show that optimized gradient boosting models, especially CatBoost, are a powerful and practical solution for credit card approval systems that can offer better risk sensitivity, interpretability, and decision reliability for financial applications in the real world.

Keywords: credit card approval, risk assessment of credit card application, gradient boosting, catboost, imbalance in class, cost sensitive learning, probability calibration, interpretability of model, machine learning.

1. INTRODUCTION

The growing dependence on data-driven decision-making has led to major changes in financial services, especially in the field of credit risk assessment and credit card approval systems. Financial institutions have increasingly relied upon machine learning models to analyze risk profiles of applicants, counter credit defaults, and ensure the smooth running of the operations. Among different methods for predictive modelling, ensemble learning methods have become especially effective because they are able to aggregate several weak learners into strong decision systems. Foundational work in statistical learning theory has shown that ensemble-based approaches provide good generalization performance in a large variety of classification tasks [1]Boosting algorithms are a popular category of ensemble algorithms that can gradually improve the predictive performance by focusing on instances that are hard to classify. Early theoretical and empirical research established boosting as a powerful framework for the improvement of the classification accuracy in complex data environments [2]. Subsequent refinements of the gradient boosting led to further development of this paradigm into functional optimization processes that allow flexible modelling of nonlinear relationships typically present in real-world datasets [3].

In financial applications, making credit approval decisions is a problem of structured tabular data that are characterized by mixed feature types, non-linear dependencies and the time course of the repayment. Stochastic extensions of gradient boosting have been shown to increase the robustness and reduce the overfitting, so they can be used for such high-dimensional financial datasets [4]. However, despite these advantages, practical deployment still remains a challenge due to data imbalance, categorical feature representation and how to have reliable decision thresholds.

One of the main difficult issues in credit card approval modelling is the imbalance that exists between application approvals and rejections. In real-world scenarios, where financial data sets are concerned, non-defaulting clients greatly outnumber defaulters, hence predictive models tend to Favor majority class accuracy at the expense of minority class detection. This problem has been widely investigated in the case of imbalanced learning, where traditional classifiers are likely to perform poorly on rare but important outcomes [5].

Moreover, credit approval systems are operated under imprecise and uncertain environments where the cost of misclassification is asymmetric. Approving a high risk applicant may lead to financial loss, while turning down a low risk applicant may lead to opportunity cost. Robust classification strategies are thus needed that can deal with uncertainty and cost sensitivity that are inherent to financial decision-making [6]. Traditional boosting and gradient boosting techniques being powerful as they are tend to lose a performances balance without careful calibrated and optimized.

Another important limitation is generated by the treatment of categorical variables. Many financial attributes, such as demographic information and repayment status indicators, are categorical by nature but are often numerically encoded and this often results in distorted feature representations. Solving this problem is crucial to making models more reliable and interpretable in the case of credit decision systems.

This research supports the application of advanced techniques (based on gradient boosting) to the problem of credit card approval prediction given structured financial data. The scope is limited to supervised binary classification models that are trained on historical credit card client records. Emphasis is given to the ensemble-based techniques as they have been shown to be successful in addressing the problems that occur in nonlinear relationships and heterogeneous feature spaces. The paper compares representative boosting frameworks, such as traditional ensemble frameworks and the implementations of modern gradient boosting, under the control of an experimental setting. While the analysis takes into consideration optimization methods that include cost-sensitive learning, early stopping and threshold tuning, the analysis does not explicitly model monetary loss functions and regulatory fairness constraints.

Additionally, experimental evaluation is done with one publicly available dataset, which may not generalize findings in different geographic or economic settings. These limitations are recognized in order to provide a transparent basis for the interpretation of the results and the direction of future research.

Despite the fact that a large amount of research has been conducted on the topic of credit scoring and risk prediction, there are several gaps in the practical application of machine learning models for credit approval decisions. Benchmarking studies have underlined that there is no single classification algorithm that is consistently superior in all credit scoring situations [7], making it necessary to choose models and optimise them according to their context [7]. Furthermore, imbalance handling and evaluation metric selection still play a critical role in model effectiveness in financial applications [8].

It is valuable because the work provides a comprehensive review of the advanced gradient boosting methods with a clear view on the handling of categorical features, mitigation of imbalanced data, calibration, and statistical validation. By chaining these puzzle pieces into a working experimental environment, the study provides viable insights on how the boosting algorithms of the day can be enhanced to be more efficient in the case of out-of-the-laboratory credit approval systems. Besides that, probability calibration and interpretability analysis can be incorporated to address practical deployment challenges to facilitate transparent and responsible decision-making in financial institutions. The contributions are particularly relevant since the regulatory control over automated systems of credit decision-making is only increasing. Based on the challenges and research

gaps identified in the field of credit card approval systems, the present study is aimed at providing empirical evidence in support of responsible and effective credit card approval decision-making through the application of optimized gradient boosting frameworks, guided by the following main objectives:

- To assess and compare the performance of ensemble and advanced gradient boosting algorithms for credit card approval prediction based on structured financial data focusing on discriminative capability and minority class detection in class-imbalanced scenarios.
- To explore the role of categorical feature treatment and optimization methods such as early stopping and decision threshold tuning and its effect on the model effectiveness and reliability and predictive performance.
- To statistically validate performance differences between baseline and optimized models while ensuring that they are interpretable and that they can be calibrated to probabilities so that they can be used for transparent, reliable, and practical deployment in real-world credit card approval decision systems.

2. LITERATURE REVIEW

The area of credit scoring and credit approval decision-making has been long-term researched in the sphere of financial risk management, and the initial research was aimed at identifying effective methods of classification that could be reliable in predicting the default behaviour. Benchmarking experiments involving multiple classification algorithms have illustrated the conclusion that no single model will always be better than the rest in all credit scoring scenarios which therefore emphasizes the necessity of context based analysis of the models and selection of the best model [11]. Based on such studies, the foundations of testing machine learning models in a financial problem setting were established since they needed to have strong performance measures not just accuracy measures.

Later studies examined the standard statistical consumer credit scoring techniques, which are logistic regression and discriminant analysis. Detailed analyses of these techniques revealed that despite their strength in the nature of interpretability and simplicity, the techniques exhibit problems in the ability to model nonlinear relationships that do exist in actual world credit data [12]. This weakness prompted scholars to investigate more adaptable machine learning methods that can capture interaction between financial variables.

In order to address these problems, ensemble-based learning methods gained increasingly popularity in credit risk assessment field. Alternatives to classifiers were considered to be neural network ensembles and hybrid methods, which had a more predictive robustness and generalization ability [13]. Nevertheless, those solutions had brought concerns over computational complexity and loss of transparency leading to their failure to be adopted in regulated financial environment. This led to tree-based ensemble methods emerging as a viable solution of a tradeoff between predictive performance and interpretability.

The development of gradient boosting machine was a major breakthrough in the history of ensemble learning. By introducing boosting as a gradient-based optimization process, gradient boosting machines allowed the development of highly flexible models with the ability to learn complex nonlinear decision boundaries [14]. These methods were shown to be able to achieve strong performance on a wide range of structured data problems, including credit scoring. Further refinements including stochastic gradient boosting introduced the randomness and regularization mechanisms which helped in improving model stability and reducing overfitting [15].

Despite these developments, class imbalance was a constant problem in credit approval datasets where default cases are usually a small portion of the entire population. Research

on cost-sensitive learning formalized the need to take asymmetric misclassification costs into the learning process, especially in applications where the financial consequences of false negative and false positive misclassification are not equal [16]. These theoretical foundations gave way to the development of learning algorithms that focus on making economically meaningful decisions instead of focusing on statistically accurate accuracy.
The research was continued by empirical research where cost-sensitive decision trees and ensemble techniques were applied to financial fraud detection and to the determination of credit risk. These studies showed that adding cost considerations to the model training can lead to substantial improvement in minority class detection and financial loss reductions [17]. Such results highlighted the value of aligning machine learning goals with the cost of decisions in the real world in financial systems.
Dependability of predicted probabilities is also another crucial factor of credit decision modelling. Probability estimation and calibration Studies of probability estimation and calibration revealed that most high-performing classifiers produce poorly calibrated probability estimates that can be used to make poor threshold-based decisions [18]. Probability has to be estimated with considerable reliability, which is particularly critical in the systems that can be subjected to credit approval, at which the decision-making thresholds that are applied are often changed based on the risk-taking capacity and rules and regulations.
Simultaneously, the spread of the application of machine learning models in the financial industry led to the increased attention given to model interpretability and transparency. Solvability to the black-box character of the complex models was suggested by the use of explainable artificial intelligence techniques to provide human-accessible explanations of a single prediction [19]. To credit scoring, this interpretability is of a significant importance as far as compliance with the regulations is concerned, validation of the model as well as the preservation of the same to the stakeholders.
Lastly, the publicly available benchmark datasets have also played a significant role in the progress of credit scoring studies. Large and popular data sets have enabled reproducible experimentation, fair comparison of algorithms and cumulative scientific advances in the area. Such datasets are a normal starting point of testing new ways of modelling, and testing methodology improvements.
Generally speaking, the current literature highlights that, despite the fact that the advanced ensemble and gradient boosting techniques offer compelling predictive strength with regard to the credit approval task, the applicability of the techniques lies on the resolution of the problems of the imbalance of classes, probability calibration, interpretability and decision reliability. These facts give a strong rationale behind the present research, which is a blend of the strength of advanced gradient boosting models, and optimization, calibration and validation strategies in improving credit card approvals decision systems.

3. METHODOLOGY

This section discusses the research design, data collection procedures, population and sampling plan, data analysis methods and ethical concerns applied in this research study. The methodology is such that it can be reproducible, possessing a statistical rigor and relevancy to the actual credit approval decision system of the world.

3.1 Research Design

This study adopts a quantitative, experimental research design to evaluate the effectiveness of advanced gradient boosting techniques for optimizing credit card approval decisions. The research is model-driven and comparative in nature, focusing on

the systematic assessment of multiple ensemble learning algorithms under identical experimental conditions.

A supervised binary classification framework is employed, where credit card applications are categorized into approved and rejected classes based on historical repayment behaviour. The research design emphasizes objective performance evaluation using standardized metrics, statistical validation, and controlled experimentation. To ensure fairness and comparability, all models are trained and evaluated using the same dataset partitions, preprocessing pipeline, and evaluation criteria.

The experimental workflow consists of four sequential stages like data acquisition and preprocessing, class imbalance handling, model training, optimization, performance evaluation and statistical validation. This structured approach enables a clear assessment of the impact of each optimization step on model performance, supporting the research objective of improving decision quality in credit card approval systems.

3.2 Data Collection Method

This study will be based on the UCI Credit Card Default Dataset that has been downloaded on UCI Machine Learning Repository [20]. This data is publicly accessible, and it is necessary in the study of the financial risk modelling and guarantees the transparency and replicability of the findings.

The sample size is of 30,000 records of credit card clients taken out of a Taiwanese financial institution. Each record represents an individual credit card account and consists of demographic information, credit limit data, historical billing amounts, repayment and payment history over six consecutive months. The target variable is whether a client defaulted on payment in the following month.

Data was obtained in Excel format and was loaded into the experimental environment using standard libraries for data processing. The first row with metadata descriptions was removed to insure proper feature parsing. An identifier column was removed as it does not help in predictive modelling. All the data preprocessing steps were performed only on the training set to avoid information leakage.

3.3 Population and Sampling

3.3.1 Population Definition

The population in question in this study includes credit card holders whose financial and repayment behaviour is recorded in the dataset. Each instance represents a unique client profile, which constitutes a real-world scenario of a credit approval.

The population consists of clients of different demographic background, education level, marital status and credit utilization patterns. As such, the dataset creates and delivers a representative snapshot of a financial services context of consumer credit behaviour.

3.3.2 Sampling Strategy

A stratified random sampling approach was used to split the data into the training and testing data. The dataset was divided in the ratio of 80:20 so we have 24,000 instances to train and 6,000 instances to test. Stratification was used to maintain original class distribution in both subsets, which is important considering the problem of class imbalance in credit default data.

The final class distribution has about 78% approved cases and 22% rejected/default cases, which is similar to the case distribution in the real world for credit approval. This sampling strategy will ensure model evaluation is performed under realistic conditions and minority class performance is evaluated correctly.

No synthetic data generation introduced in the sampling process thereby maintaining the integrity and authenticity of original dataset.

3.4 Data Analysis Techniques

3.4.1 Data Preprocessing and Feature Handling

Data preprocessing included cleaning, transformation, and feature preparation for the training of the model. Numerical attributes were checked for missing values and imputed using median-based imputation which is robust to skewed financial distributions. Categorical attributes were filled in with the most common category when needed.

A crucial preprocessing step has been the explicit identification and deal with categorical features including demographic variables (SEX, EDUCATION, MARRIAGE) and repayment status variables (PAY_0 to PAY_6). Although these attributes are numerically encoded, they are categorical semantics and thus they were treated so as to avoid misinterpretation as continuous values.

3.4.2 Class Imbalance Handling

The dataset has a large class imbalance as default cases are in the minority. To overcome this problem, cost-sensitive learning was used instead of the oversampling methods (SMote or ADASyn). Class weights were calculated according to inverse class frequency so that the minority class is penalized with a higher misclassification cost.

Cost-sensitive learning was applied in a consistent way across the models through the techniques of class-weight parameters, sampling weighting, and scale adjustment mechanisms, which made sure that the training of the models focused on the detection of minority classes, without altering the initial data distribution.

3.4.3 Model Development and Optimization

Four ensemble learning models were tested, namely AdaBoost, LightGBM, XGBoost, and CatBoost. AdaBoost was used as a baseline model, and CatBoost was considered as the proposed advanced boosting approach because it has native categorical handling and ordered boosting mechanism.

Hyperparameter optimization using grid based search for baseline models and controlled using early stopping using catboost. Early stopping was used to avoid overfitting and enhance the generalization capability In addition, decision threshold optimization was also carried out to achieve the maximization of F1-score, allowing to obtain improved trade-offs between precision and recall for the minority class.

3.4.4 Evaluation Metrics and Statistical Validation

Model performance was assessed using several evaluation metrics, such as ROC-AUC, F1-score, balanced accuracy and precision-recall curves. These metrics give a complete evaluation of the ranking ability, the effectiveness of the classification, and fairness under class imbalance.

In order to validate the robustness of observed performance differences, McNemar's test and Wilcoxon signed-rank test were used. These non-parametric statistical tests test whether differences between competing models are statistically significant rather than due to random variation.

Calibration analysis was also performed based on Brier scores and reliability diagrams for the quality of predicted probabilities which is crucial in financial decision making situations.

3.5 Ethical Considerations

This research follows ethical guidelines in that it uses a publicly available, fully anonymized dataset, which ensures the privacy of individual's data. Cost-sensitive learning and model interpretability techniques are added in order to mitigate biased

decision outcomes and support transparency, accountability, and responsible deployment of machine learning models in credit approval systems.

4. RESULTS

This section presents the experimental results obtained by applying the advanced methods of gradient boosting for the credit card approval prediction problem. Results are presented using quantitative evaluation metrics, using comparative tables and visual performances figures. All reported values are the held out test dataset and follow the evaluation criteria.

4.1 Baseline and Optimized Model Performance

Baseline Performance Evaluation Roots of AdaBoost ROC-AUC: 0.77099 F1-Score: 0.52149 Balanced Accuracy: 0.70169 Test Set While the overall classification accuracy achieved was 75%, the class-wise analysis shows limited success in rejecting credit applications with minority class precision and recall being 0.45 and 0.61 respectively.

Integrating the native handling of categorical features using CatBoost led to the stabilization of performance, which is measurable: the performance was stabilized to a ROC-AUC of 0.77087 with an F1-score of 0.52379 with a slightly improved performance (0.0014) with a balanced accuracy of 0.70120. These results call for correct categorical representation to maintain basal discriminative capability along with enhanced class-wise consistency.

Applying early stopping during CatBoost's training process helped to see an evident improvement in the discriminative performance, with the ROC-AUC increasing to 0.78095. At default decision threshold of 0.5, this converted to F1-score of 0.54281 and balanced accuracy of 0.71551, which showed sensitivity towards minority class results.

Subsequent threshold optimization improved the effectiveness of classification with optimal threshold value 0.53 giving the highest F1 score 0.54902 and maintaining ROC-AUC with balanced accuracy improved to 0.71569. Table 1 presents a comparative summary of baseline and optimized model performance, highlighting the progressive improvements achieved through categorical feature handling, early stopping, and decision threshold optimization.

Table 1. Baseline and Optimized Model Performance Comparison

Model / Variant	ROC-AUC	F1-score	Balanced Accuracy	Notes
AdaBoost (Baseline)	0.77099	0.52149	0.70169	Baseline boosting model
CatBoost (Categorical Handling Only)	0.77087	0.52379	0.70120	Categorical features fixed
CatBoost + Early Stopping (Threshold = 0.50)	0.78095	0.54281	0.71551	Best iteration = 153
Proposed CatBoost (Threshold Tuned = 0.53)	0.78095	0.54902	0.71569	Best F1 threshold = 0.53

As shown in Table 1, the proposed CatBoost model demonstrates consistent gains across ROC-AUC, F1-score, and balanced accuracy compared to the AdaBoost baseline. As seen by comparison to the AdaBoost baseline, the proposed CatBoost model obtained an approximate 5.3% relative improvement in F1-score and 2.0% in balanced accuracy, along

with an improvement in ROC-AUC. These gains illustrate steady gains on the minority classes under optimized boosting configurations.

4.2 Comparative Analysis of Boosting Models

Table 2 compares the predictive performance of AdaBoost, LightGBM, XGBoost, and the proposed CatBoost model using ROC-AUC, F1-score, and balanced accuracy on the test dataset.

Table 2. Comparative Performance of Boosting Models

Model	ROC-AUC	F1-score	Balanced Accuracy
AdaBoost	0.77099	0.52149	0.70169
LightGBM	0.76915	0.52904	0.70193
XGBoost	0.76681	0.53088	0.70476
Proposed CatBoost	0.78095	0.54902	0.71569

The proposed CatBoost model had the highest values for all the reported metrics. Across all boosting frameworks, the proposed CatBoost model ranked first on all the evaluation metrics suggesting stable performance advantages over conventional gradient boosting implementation. The results in Table 2 indicate that the proposed CatBoost model outperforms competing boosting frameworks across all evaluated metrics.

Figure 1 illustrates the receiver operating characteristic (ROC) curves of all evaluated boosting models, comparing their discriminative ability across different classification thresholds.

Figure 1. ROC Curve Comparison of Boosting Models

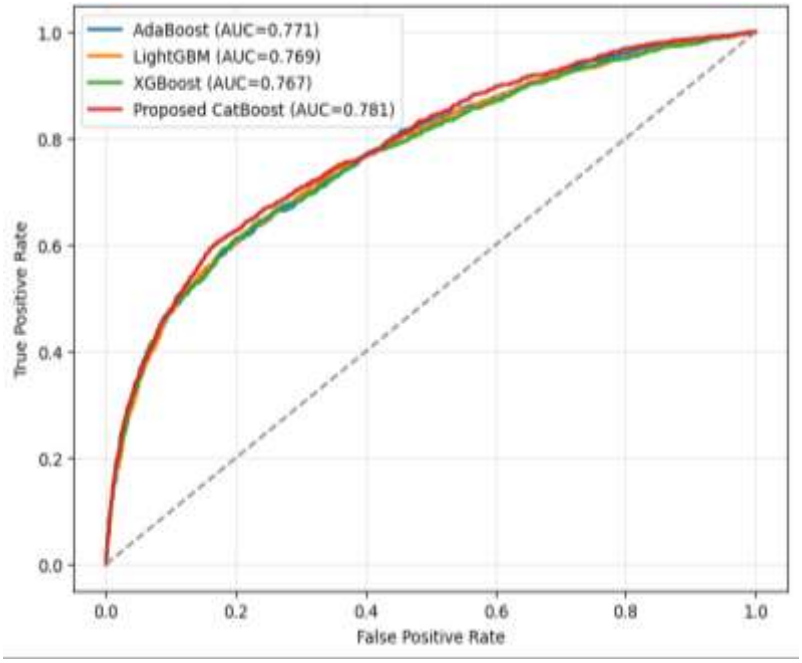
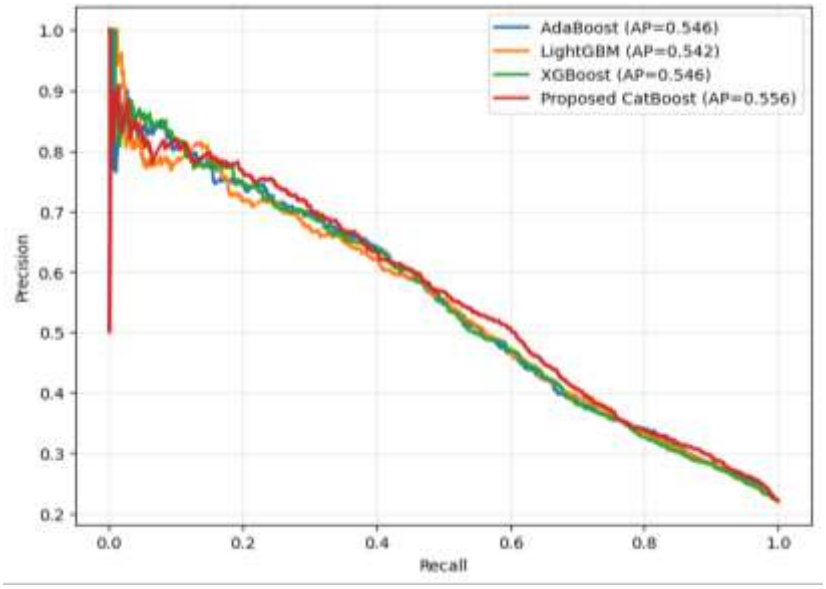


Figure 2 presents the precision–recall curves for the evaluated models, emphasizing their effectiveness in detecting minority-class credit rejections under imbalanced conditions.

Figure 2. Precision–Recall Curve Comparison of Boosting Models

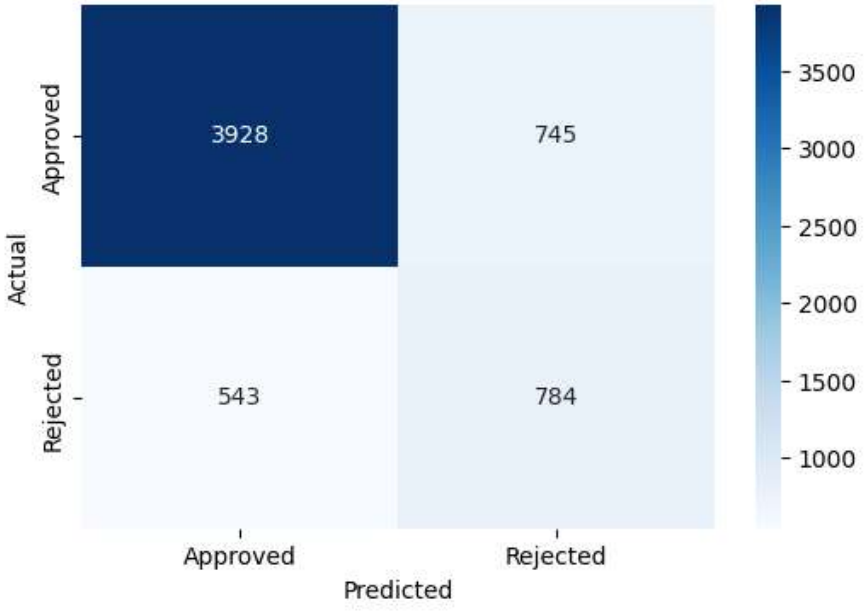


The precision-recall analysis shows that the proposed CatBoost model achieved the best average precision (AP = 0.556) of all the models evaluated.

4.3 Classification Outcome Analysis

To investigate the prediction results of the classes, a confusion matrix was created for the optimized CatBoost model with a tuned decision threshold. Figure 3 shows the confusion matrix of the optimized CatBoost model using the tuned decision threshold, illustrating class-wise prediction outcomes.

Figure 3. Confusion Matrix of Optimized CatBoost Model



Out of 1327 rejected credit applications in the test set, 784 instances were correctly classified, and 543 were misclassified. For the approved applications, there were 3,928 correctly predicted and 745 false rejections. These outcomes represent balanced classification behavior in both classes under imbalanced conditions.

4.4 Model Interpretability Results

Model interpretability was tested in terms of feature importance rankings and SHAP values analysis to see what the most influential predictors also Figure 4 depicts the top 15 most influential features identified by the CatBoost model, ranked according to their contribution to the prediction task.

Figure 4. Top 15 Feature Importances from CatBoost

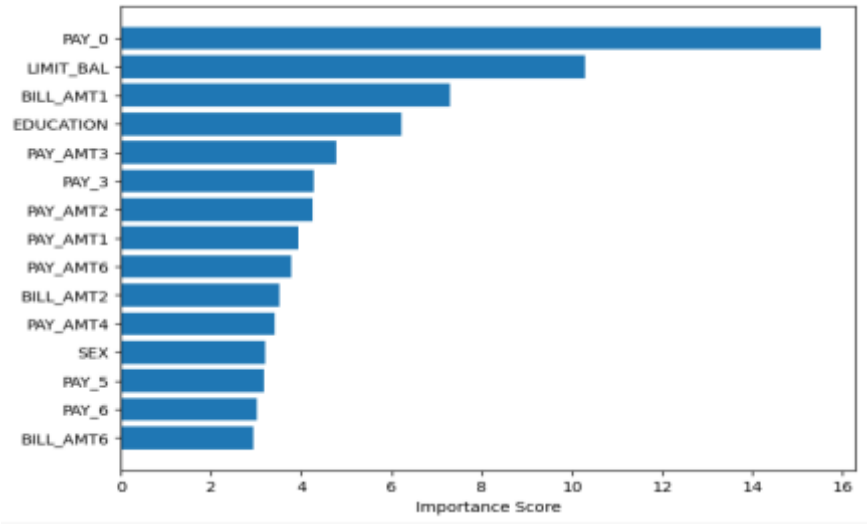
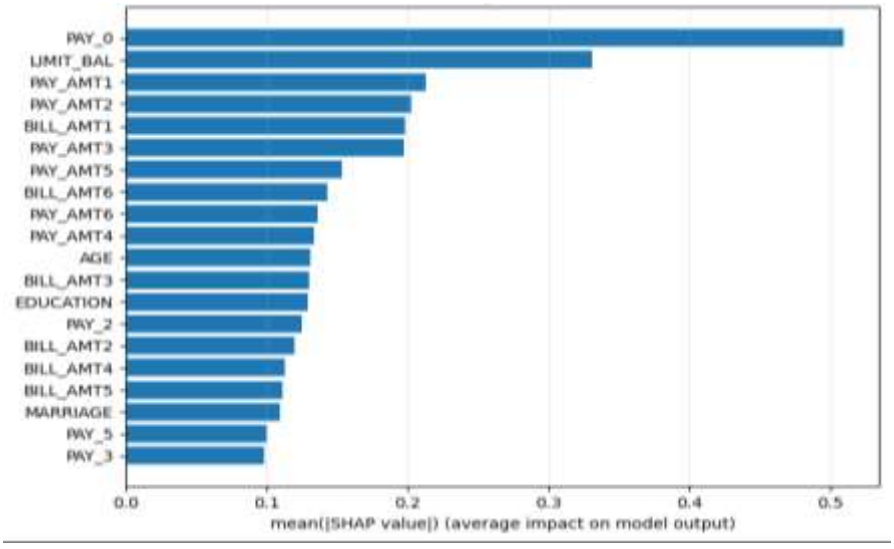


Figure 5 illustrates the SHAP summary plot for the proposed CatBoost model, providing a global explanation of feature influence and directional impact on predictions.

Figure 5. SHAP Summary Plot for CatBoost Model



Repayment status variables (PAY_0, PAY_2, PAY_3) and credit limit (LIMIT_BAL) were consistently ranked as most influential features followed by billing and payment amount variables.

4.5 Calibration and Reliability Results

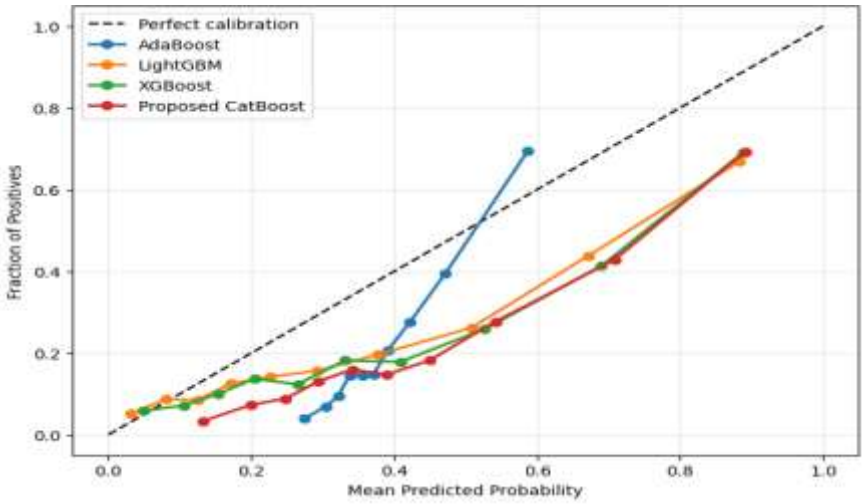
Probability calibration was qualified by Brier scores to check the reliability of predicted probabilities. Table 3 reports the calibration performance of the evaluated models using Brier scores to assess the reliability of predicted probabilities.

Table 3. Calibration Performance of Models (Brier Score)

Model	Brier Score
LightGBM	0.15994
XGBoost	0.16616
Proposed CatBoost	0.17772
AdaBoost	0.21093

Calibration curves for these results are shown below. While LightGBM has the lowest Brier score, the proposed CatBoost model has a competitive calibration performance and higher discriminative ability in ROC-AUC and F1-score. The calibration trends in Figure 6 indicate that the proposed CatBoost model maintains reliable probability estimates while achieving stronger discriminative performance.

Figure 6. Calibration Curve Comparison of Boosting Models



4.6 Statistical Validation and Computational Cost

In order to validate the statistical robustness of the observed performance differences, McNemar's test and the Wilcoxon signed-rank test were performed between the baseline and proposed models.

- McNemar Test Statistic: 125.0
- p-value: 3.11×10^{-22}
- Wilcoxon Test Statistic: 5,244,959.0
- p-value: 1.81×10^{-172}

Both tests show statistically significant performance differences Computational efficiency was compared in terms of training time. Table 4 summarizes the computational cost of

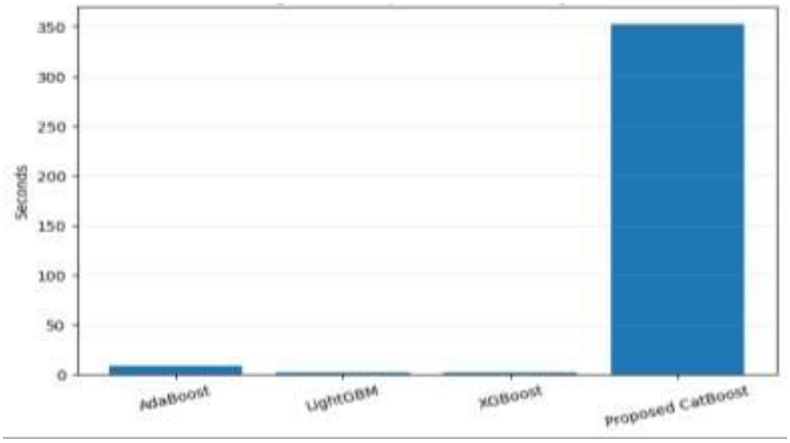
each boosting model in terms of training time, providing insight into efficiency–performance trade-offs.

Table 4. Training Time Comparison of Boosting Models

Model	Training Time (seconds)
LightGBM	1.36
XGBoost	1.42
AdaBoost	8.09
Proposed CatBoost	352.33

Figure 7 illustrates the training time required by each boosting model, highlighting differences in computational efficiency.

Figure 7. Training Time Comparison of Boosting Models



5. DISCUSSION

This study analysed effectiveness of advanced gradient boosting techniques for optimizing the decision of approving or not a credit card using an imbalanced and real-world financial dataset. The discussion is an interpretation of the experimental results, placing the results in relation to known literature and indicating practical implications, limitations, and future research.

The baseline results show that ensemble learning methods such as AdaBoost have reasonable discriminative capability for the prediction of credit approval, with a ROC-AUC close to 0.77. This observation is in line with the foundational machine learning literature which identifies boosting as a good general purpose technique for classification tasks [1][2]. However, despite acceptable overall performance, the baseline model showed low efficacy in distinguishing rejected credit applications, as has been widely documented in imbalanced financial datasets.

The performance gains that were achieved using the proposed CatBoost framework show how important it is to have an appropriate algorithm for structured financial data. And by natively working with categorical variables and mitigating the prediction bias using ordered boosting, CatBoost proved superior robustness as compared to the conventional gradient boosting implementations. These results are in line with the theoretical underpinnings of the gradient boosting machines [3][4], and with the previous empirical results demonstrating the superior performance of CatBoost compared to other boosting algorithms when categorical features are central to the problem [9].

Comparative evaluation against LightGBM and XGBoost is another evidence of this observation. Although LightGBM and XGBoost had competitive results, they had limitations on the encoding of categorical attributes as they were numerically encoded, which in turn limited their effectiveness in this credit approval setting. Similar performance patterns have been reported in benchmarking studies on credit scoring in which model performance was shown to be dependent not only on algorithmic power, but also on the quality of feature representation [10] [11]. The superior balanced accuracy attained by the proposed CatBoost model highlights the ability of the model to have consistent classification behaviour for both majority and minority classes.

An important contribution of this study is the integration of the optimization of decision threshold. Rather than using a default classification threshold, the threshold tuning led to an improvement in F1-score and balanced accuracy without any loss in ROC-AUC. This result supports the work by others that emphasizes that classification thresholds should be adjusted to the cost structure and imbalance characteristics of financial decision systems. In the case of credit approval applications, where false approvals of high-risk applicants can result in significant financial loss, threshold calibration is an important part of model deployment.

The calibration analysis gives more information on the reliability of the predicted probabilities. Although LightGBM had the lowest Brier score, the proposed CatBoost model preserved the calibration of a model while providing a better discriminative performance. This trade-off between discrimination and calibration has been thoroughly discussed in the literature, and it is recognized that models that are optimized for ranking performance do not always have the best probability calibration [14][15][16]. In the case of practical credit approval systems accurate ranking and minority classes detection often play the more important role especially when outputs of probabilities are combined with business defined thresholds.

Model interpretability is another important issue of practical relevance. Feature importance rankings and SHAP-based explanations consistently put the variables representing repayment history and credit limits as the most influential variables. These results are consistent with known credit scoring research, which highlights the importance of the past repayment behaviour in determining credit risk. The use of SHAP values increases the level of transparency by offering consistent and model agnostic explanations of individual predictions, supporting regulatory compliance and stakeholder trust [17][18].

The results of statistical validation help to increase the credibility of the observed performance gains. The application of McNemar's and Wilcoxon signed-rank tests shows that improvements achieved by the proposed CatBoost model are statistically significant and not due to random variation.

Despite the strengths, there are certain limitations in this study. First, experimental evaluation is based on one data set coming from one geographic and economic context, which may restrict the generalization of the results. Second, although learning that was cost-sensitive was integrated, explicit modelling of monetary costs was not directly incorporated into the optimization objective. Additionally, metrics of fairness regarding the issue of demographic bias were not explicitly tested, which may be relevant when deploying in regulated financial settings.

6. CONCLUSION

This study proposed an in-depth analysis of the sophisticated gradient boosting techniques to optimize credit card approval decisions using the assistance of a real-life financial dataset that is imbalanced. Using a feature handling categorically, cost-sensitive learning, early stopping mechanism and decision threshold optimization, this kind of

CatBoost framework can be repeatedly improved in terms of performance over traditional boosting algorithms.

The experimental outcomes prove that, the optimized CatBoost model had the highest ROC-AUC, F1-score and balanced accuracy, which means that it has a better discriminating and simpler detection of minority classes. Statistical confirmation of McNemar and Wilcoxon tests showed that these improvements are significant. Calibration and interpretability analysis also offered further evidences of the reliability and transparency of the proposed model and answer valuable practical requirements of financial decision systems.

The findings conclude that it is essential to apply the force of algorithm development along with a careful preprocessing and assessment strategy, when it comes to the use of machine learning models to forecast credit risk. Specifically, the processing of categorical variables and thresholds calibration became significant constituents in the balancing process and effective decision-making.

Practically, the suggested framework offers a strong interpretable as well as data-effective solution that can be applied in a real credit approval application. Although, the study is grounded in one dataset, the method can be generalized and applied to other financial risk prediction problems, which may include explicit, cost-sensitive goals, learning processes that are fair, and multi-dataset validation. Future research directions can encompass the expansion of this study to other types of financial risk prediction tasks. On the whole, the research contributes to the empirical evidence of applying advanced methods of gradient boosting with an aim to make plausible and accountable decisions related to credit.

REFERENCES

1. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.
2. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.
3. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
4. J. H. Friedman, "Stochastic gradient boosting," Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367–378, 2002.
5. [5] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
6. F. Provost and T. Fawcett, "Robust classification for imprecise environments," Machine Learning, vol. 42, no. 3, pp. 203–231, 2001.
7. C. Elkan, "The foundations of cost-sensitive learning," in Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI), 2001, pp. 973–978.
8. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
9. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018.
10. S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015.
11. B. Baesens, T. Van Gestel, M. Stepanova, D. Van den Poel, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," Journal of the Operational Research Society, vol. 54, no. 6, pp. 627–635, 2003.

12. [12] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: A review," Journal of the Royal Statistical Society, vol. 160, no. 3, pp. 523–541, 1997.
13. C.-F. Tsai and J.-W. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring," Expert Systems with Applications, vol. 34, no. 4, pp. 2639–2649, 2008.
14. A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in Proceedings of the IEEE Symposium Series on Computational Intelligence, 2015, pp. 159–166.
15. B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 694–699.
16. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proceedings of the International Conference on Machine Learning (ICML), 2005.
17. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
18. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
19. A. C. Bahnsen, D. Aouada, and B. Ottersten, "Cost-sensitive decision trees for fraud detection," Expert Systems with Applications, vol. 42, no. 11, pp. 5380–5390, 2015.
20. UCI Machine Learning Repository, "Default of Credit Card Clients Dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients