A Cost-Sensitive Method for Credit Card Default Prediction

Yiqi Zhou[a]
Department of Mathematics, The Ohio State University, 281 W. Lane Avenue, Columbus, Ohio 43210, U.S.A.

Keywords: Credit Card Default, Cost-Sensitive, Tuning Threshold, Gradient Boosting, Financial Risk.

Abstract: With an increasing number of credit cards being issued, financial institutions must predict credit card default to minimize bad debts and remain profitable. Traditional machine learning approaches typically emphasize maximizing overall accuracy but neglect the higher cost of missing actual defaulters (false negatives, FNs). This research introduces a cost-sensitive framework that involves robust data cleaning, multi-model comparisons, and threshold tuning. Based on a University of California, Irvine (UCI) dataset containing 30,000 records and 25 features, the procedure merges anomalous categories, removes duplicates, and standardizes skewed numeric columns. Benchmarking three popular algorithms—Logistic Regression, Random Forest, and eXtreme Gradient Boosting (XGBoost)—reveals that XGBoost attains the highest Area Under the Curve (AUC) and recall for credit defaulters. Building on these findings, the XGBoost decision threshold is further adjusted by assigning heavier penalties to FNs than to false positives (FPs). Experimental results indicate that without such an intervention, potential bad debt nears 877,900 (or 11.58 percent of the total), whereas the cost-sensitive approach reduces it to approximately 432,400 (or 5.64 percent), highlighting the limits of raw accuracy metrics. This paper concludes with a discussion of interpretability, advanced hyperparameter tuning, and deployment considerations in real-world finance, reflecting data up to October 2023.

1 INTRODUCTION

Credit card lending remains a major revenue source for banks, yet it can also involve default risks that erode profits and jeopardize long-term stability. Expanding consumer credit access around the globe has made the effective identification of potential defaulters increasingly vital. Machine learning has the potential to reinforce and automate credit risk decisions, sometimes outperforming conventional models in both speed and predictive power. Brown and Mues (2012) indicate that ensemble methods often handle imbalanced credit scoring more effectively than older techniques, while Yeh and Lien (2009) show that data mining approaches frequently surpass standard statistical models in identifying defaulters. These findings illustrate the promise of advanced algorithms—such as gradient boosting or cost-sensitive classification—for leveraging large-scale finance data and achieving more accurate, efficient predictions than standard solutions. However, existing methods tend to focus on metrics like accuracy or AUC without investigating how misclassifications differently influence institutional outcomes.

Cost asymmetry, if ignored, may result in suboptimal performance. A missed defaulter (FN) can entail a far higher cost than a false alarm (FP), because lenders risk losing the entire unpaid balance plus fees, whereas a wrongly flagged customer experiences only minor inconvenience or limited credit access (Bahnsen, Aouada, & Ottersten, 2015; He & Garcia, 2009). Certain studies address class imbalance through oversampling or weighting, but often overlook actual monetary losses. Meanwhile, data quality issues—such as “filler” categories, duplicates, and skewed numeric fields (credit limit, bill amounts)—can distort modeling results. Merging anomalous labels, removing repeated records, and normalizing columns helps mitigate these challenges (Li, Zhao, & Wu, 2019). Another key concern is interpretability: although complex ensembles like gradient boosting yield strong predictive performance, they often appear opaque to loan officers or regulators who seek transparency in decision-making.

[a] https://orcid.org/0009-0000-1834-8521

Zhou, Y.
A Cost-Sensitive Method for Credit Card Default Prediction.
DOI: 10.5220/0013701200004670

In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 516-520
ISBN: 978-989-758-765-8

A cost-sensitive framework is introduced to manage these problems, combining enhanced data preprocessing on the UCI dataset (merging rare labels, removing duplicates, and scaling skewed features) with threshold-based penalization of FNs in XGBoost. Small threshold adjustments can substantially reduce potential losses, thereby bridging standard classification metrics and the real-world economic landscape. This perspective raises credit modeling beyond routine data cleansing, aligning machine learning decisions more closely with the financial realities of default risk.

2 METHODS

2.1 Dataset

The "Default of Credit Card Clients" dataset made publicly available by Yeh and Lien (2009) in the UCI Machine Learning Repository (Dua & Graff, 2019) is employed. It contains 30,000 records and 25 features, with each record representing a unique cardholder's demographic and financial attributes. The target label specifies whether a cardholder is in default (1) or not (0). According to Yeh and Lien (2009), around 22% of cardholders default in the original data, whereas 78% do not. If further filtering or cleaning is performed, the final distribution should be reconfirmed, but in its default version the data generally splits at approximately 22% default vs. 78% non-default. This imbalance can motivate naive models to prioritize the majority class if maximizing accuracy alone is the objective.

2.1.1 Data Cleaning

Combining certain anomalous values in EDUCATION (0,5,6) and MARRIAGE=0 into an "other" category (such as EDUCATION=4, MARRIAGE=3) prevents fragmentation arising from extremely rare labels. Any duplicate entries are eliminated to maintain unique records, given that duplication can distort metrics if repeated samples are uncommonly easy or difficult to classify. Numeric columns such as LIMIT_BAL, monthly bill amounts, and payment amounts can display significant skewness, so standard scaling ($(x - \mu)/\sigma$) is applied to reduce outlier influence. Although alternative transformations such as logs are feasible, standardization aligns well with the pipeline setup.

2.1.2 Exploratory Analysis

A check of the distribution of default vs. non-default, the range of numeric columns, and correlations across features (for example, monthly billing correlated with repayment) is carried out to identify whether any derived features could be impactful. Despite these examinations, the original columns remain in place for the main modeling stage.

2.2 Models

Logistic Regression is a linear model that calculates log-odds for the default class based on weighted feature inputs. While it offers convenient interpretability via its coefficients, it may overlook significant nonlinearities in financial data unless supplementary transformations are introduced. Random Forest, in contrast, merges multiple decision trees constructed via bootstrap aggregating, which often delivers robust performance on tabular credit data by modeling more complex relationships than purely linear methods. Some interpretability is retained through feature importance measures, yet the model structure is inherently more opaque than Logistic Regression. XGBoost applies a gradient boosting strategy (Chen & Guestrin, 2016), incrementally refining shallow trees through advanced regularization and efficient handling of missing data. Research by Brown and Mues (2012) suggests that boosting algorithms frequently outdo simpler ensembles in terms of AUC and recall when dealing with credit scoring challenges, indicating that XGBoost is likely to be either the top performer or a strong benchmark for real-world default detection. All models undergo identical transformations, including standardized numeric columns, one-hot encoded categorical variables (with merged "other" categories), and an 80–20 split into train and test subsets.

2.3 Experimental Setup

2.3.1 Train–Test Split

A random partition is carried out, with 80% of the available records (24,000) dedicated to training and 20% (6,000) allocated to testing, while preserving the proportion of defaulters in each subset. This approach replicates the default and non-default distribution encountered in the original dataset (Hand & Henley, 1997).

2.3.2 Performance Metrics

Before each model is trained, the standard transformations mentioned above are applied, and default hyperparameters are used unless noted otherwise. Evaluation includes accuracy, AUC, recall (for defaulters), precision, and F1. Accuracy can be deceptive under imbalanced conditions, while AUC reflects how effectively the model ranks defaulters over non-defaulters. High recall for defaulters is vital in credit settings, since missed defaulters (false negatives) impose considerable costs. Precision measures how many predicted defaulters indeed default, whereas the F1 score balances recall and precision in a single index. Earlier findings (Brown & Mues, 2012) imply that XGBoost tends to surpass simpler baselines in terms of AUC and recall for credit defaulters, thereby providing a stable basis for further threshold calibration.

3 RESULT

Logistic Regression, Random Forest, and XGBoost are trained and the metrics are collected in Table 1:

Table 1: Comparison of model performance indicators

Model	Accuracy	AUC	Recall(1)	Precision(1)	F1(1)
Logistic Regression	0.81	0.7062	0.25	0.69	0.37
Random Forest	0.81	0.7550	0.36	0.64	0.46
XGBoost	0.81	0.7588	0.36	0.61	0.45

All three perform at an accuracy of ~81%, suggesting moderate separability of the dataset. But XGBoost gives the highest AUC (0.7588) and the same recall(1) (=0.36) as Random Forest. On the contrary, the recall of Logistic Regression(1) is only 0.25, indicating that it misses more defaulters. For institutions that give priority to capturing defaulters, XGBoost is the best candidate before threshold tuning.

In standard classification, a threshold of 0.5 is commonly employed: predicted probabilities greater than or equal to 0.5 result in a “default” label. However, as Bahnsen et al. (2015) suggest, the cost of a missed defaulter (FN) in credit risk substantially exceeds that of a false positive (FP), and failing to account for these differences frequently leads to suboptimal default detection. Consequently, the approach here defines a cost function that imposes a heavier penalty on FNs.

4 DISCUSSION

4.1 Comparison with Other Work

Many studies have highlighted advanced boosting or ensemble methods for credit scoring but have centered primarily on optimizing metrics such as AUC or recall (Li et al., 2019). By systematically scanning thresholds under an explicit cost function and translating different error types into direct monetary impacts, the present approach offers a clearer understanding of how false negatives and false positives each influence potential losses. This perspective extends beyond basic imbalance handling (He & Garcia, 2009) by identifying the precise cost ratio that best aligns with real-world risk. Consequently, the measure proposed here provides a tangible metric of potential bad-debt reduction, distinguishing it from methods that merely improve top-line classification rates without clarifying financial consequences.

4.2 Real-World Implementation

If the framework is to be deployed in an actual banking environment, it must be established who would revise the threshold and how frequently. One possibility is to perform monthly updates to the ratio of Cost_FN vs. Cost_FP, using default outcomes or macroeconomic indicators observed within that period. Such updates would then feed into the production XGBoost scoring model, causing acceptance rates for borderline applicants to fluctuate depending on current risk conditions. Regulators might question the equity or transparency of a frequently shifting threshold, thus suggesting the need for interpretability measures or disclaimers that clarify the strategic selection of cost-based thresholds.

4.3 Managing Macroeconomic Ambiguity

Credit default rates can rise and fall alongside broader economic trends. When unemployment increases, more borrowers might fail to repay their balances, rendering a static threshold increasingly risky. Adopting a dynamic threshold—one that lowers automatically when default rates climb—may mitigate some portion of potential damage. If combined with advanced forecasting tools, this tactic could guide lenders to tighten acceptance policies before an expected recession. Still, implementing real-time or frequent updates to the cost function may become logistically complex, requiring additional processes for data ingestion and threshold recalibration.

4.4 Potential Extensions

One extension might involve incorporating hybrid models that blend XGBoost with specialized anomaly detection techniques, particularly for unusual behavior in credit usage. Another possibility involves deeper exploration of feature contributions, moving beyond threshold scans to examine which attributes drive cost outcomes most strongly. In some cases, a time-series or sequence-based approach might be pursued if payment patterns over multiple months can be formulated as a sequence, thus applying RNN or transformer architectures. Finally, methods such as SHAP or LIME can clarify which features consistently trigger a default label, thereby satisfying stakeholders who need transparency (Zhang & Li, 2020).

4.5 Limitations Revisited

Although this cost-sensitive approach unifies data preprocessing, multi-model comparison, and threshold optimization in XGBoost, several constraints remain. First, the 1000:100 ratio is heuristic and may not reflect actual institutional practices. Different banks could adopt distinct scales based on average credit limits or interest structures, indicating that a tiered or dynamic cost matrix might better capture genuine lending risks. Second, only a basic version of XGBoost was tested here, and advanced hyperparameter tuning or specialized cost-sensitive objectives could further refine outcomes. Third, the study hinges on one UCI dataset, suggesting a need for external validation on proprietary data to verify real-world performance. Finally, threshold scanning occurs offline, whereas continuous risk environments require pipelines for frequent data ingestion, model retraining, and threshold readjustment. Nonetheless, emphasizing penalties for missed defaulters reduces estimated bad debt from roughly 877,900 to 432,400, underscoring how accuracy (or AUC) alone may conceal the severe costs of high-risk borrowers. Future directions include multitier cost ratios guided by credit lines, deeper tuning of XGBoost, and integration of interpretability tools such as SHAP or LIME for regulatory approval. Online or dynamic thresholds could also adapt to evolving economic signals, ensuring that modeling keeps pace with real-world lending operations (Zhou, 2012).

5 CONCLUSION

This study proposed a cost-sensitive framework for credit card default prediction, which integrates comprehensive data preprocessing, multi-model comparison, and threshold optimization in the XGBoost. Highlighting the imbalanced cost of missing defaulters brought potential bad debt down from around 877,900 to 432,400, pointing to the fact that some metrics like accuracy or AUC can miss the point in high-stakes finance contexts.

Discussion of these results emphasized how a decision threshold adjustment, instead of just maximizing standard measures, can better suit classification to real-world lending goals. This insight advances practice through cost-based threshold tuning and overcoming missing sentiment, which will improve risk management and profits.

Still, some limitations do exist. Decisions made using the single-cost ratio present in this paper may not generalize to all lenders, nor is it based on any numerous datasets. Refining this approach with multi-tiered cost frameworks or more sophisticated hyperparameter searches would enhance its robustness and adaptability. Moreover, the interpretability issue is not solved in complex ensemble approaches such as gradient boosting, highlighting the need for explainable tools that can help fulfill the requirements of transparency in financial systems.

Therefore, in practice, this cost-sensitive strategy can be implemented as part of the day-to-day decision processes of banks and credit issuers that can update their thresholds as market conditions change. Next steps may be for models to become adaptive, updating cost ratios in real time to further improve the risk capture while maintaining a good balance with overall portfolio performance.

REFERENCES

Bahnsen, A. C., Aouada, D., & Ottersten, B. 2015. Example-dependent cost-sensitive logistic regression for credit scoring. *Expert Systems with Applications*, 42(6), 2476–2486.

Brown, I., & Mues, C. 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446–3453.

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*, 785–794.

Hand, D. J., & Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A*, 160(3), 523–541.

He, H., & Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

Li, C., Zhao, K., & Wu, Y. 2019. Deep learning for credit risk assessment: A review of challenges and solutions. *Expert Systems with Applications*, 122, 206–220.

Thomas, L. C. 2009. Consumer credit models: Pricing, profit, and portfolios. *Oxford University Press*.

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.

Zhang, L., & Li, J. 2020. Cost-sensitive boosting trees for financial risk assessment. IEEE Transactions on Knowledge and Data Engineering, 32(5), 909–919.

Zhou, Z.-H. 2012. Ensemble methods: Foundations and algorithms. *Boca Raton, FL: CRC Press*.