Credit Card Default Prediction: A Comparative Analysis of Machine Learning Approaches

QiaoQiao Han (110187843), Mohammad Faisal Alam (110184442)
University of Windsor
han9b@uwindsor.ca, alam8c@uwindsor.ca

***Abstract*—This report presents a comprehensive comparison of machine learning algorithms for predicting credit card default payments using the UCI Default of Credit Card Clients dataset. We evaluate seven different approaches: Logistic Regression, K- Nearest Neighbors, Decision Trees, Gaussian Naive Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Multi- Layer Perceptron. Each model was optimized using hyperparameter tuning and evaluated using multiple metrics including ROC-AUC, precision, recall, F1-score, and cost-sensitive analysis with a false positive cost of 1 and a false negative cost of 5. Our results demonstrate that Quadratic Discriminant Analysis (BASE) achieved the best overall performance with a minimum cost of 3885, followed closely by Gaussian Naive Bayes with SMOTE and hyperparameter tuning (cost 4269). The study addresses the class imbalance problem through SMOTE resampling and provides insights into feature importance for credit risk assessment.**

I. Introduction

Credit card default prediction is a critical problem in financial risk management. Accurate models can reduce losses and improve lending decisions. However, class imbalance and the asymmetric cost of misclassification (false negatives are much more costly than false positives) make this a challenging task. This study addresses the following research questions:

- Which machine learning models are most effective for cost-sensitive credit default prediction?
- How does hyperparameter tuning and class imbalance handling (SMOTE) affect model performance?
- What is the impact of cost-sensitive threshold selection on real-world outcomes?

A. Dataset Description

The experiments in this study are based on the UCI Default of Credit Card Clients dataset [1], which contains data on 30,000 credit card holders in Taiwan from April to September 2005. The dataset includes 23 input features and a binary target variable indicating whether the client defaulted on their payment the following month (1 = default, 0 = no default).

Features:

- **Demographic:** LIMIT_BAL (credit limit), SEX, EDUCATION, MARRIAGE, AGE
- **Payment History:** PAY_0 to PAY_6 (repayment status for the last 7 months)
- **Bill Statements:** BILL_AMT1 to BILL_AMT6 (bill amounts for the last 6 months)
- **Previous Payments:** PAY_AMT1 to PAY_AMT6 (amounts paid in the last 6 months)

Class Balance: The dataset is imbalanced, with 22.12% defaults (1) and 77.88% non-defaults (0).

Feature Types: Features include both categorical (SEX, EDUCATION, MARRIAGE, PAY_0 to PAY_6) and continuous variables (AGE, LIMIT_BAL, BILL_AMT, PAY_AMT).

B. Cost Analysis Description

Credit card default prediction is crucial for financial institutions to assess credit risk. The cost of false negatives (FN), i.e., missing a potential defaulter, is typically higher than false positives (FP), i.e., misclassifying a good customer. We adopted a cost ratio of 1:5 (FP:FN) to reflect this business reality.

II. Data and Preprocessing

A. Preprocessing Steps

A preprocessing pipeline was applied to ensure data quality and compatibility with machine learning algorithms:

1) **Data Cleaning:** Anomalous values in EDUCATION (0, 5, 6) were replaced with 4 (other), and in MARRIAGE (0) with 3 (other), to ensure valid categorical values.
2) **Type Conversion:** Categorical features (SEX, EDUCATION, MARRIAGE, PAY_0 to PAY_6) were converted to categorical data types. This ensures correct handling during encoding and prevents unintended numerical interpretation.
3) **One-Hot Encoding:** All categorical variables were one-hot encoded to create binary indicator columns, allowing models to interpret category membership without imposing ordinal relationships.
4) **Train-Test Split:** The dataset was split into 70% training and 30% testing sets, stratified by the target variable to preserve class proportions. This split ensures that model evaluation reflects real-world generalization.
5) **Feature Scaling:** All features were standardized using scikit-learn's StandardScaler (zero mean, unit variance). Scaling is essential for distance-based models (KNN, MLP) and accelerates neural network convergence.
6) **Class Imbalance Handling:** Synthetic Minority Oversampling Technique (SMOTE) is a data augmentation method that creates synthetic examples of the minority class to address class imbalance in classification tasks [2]. SMOTE was applied to the training set for all the models. It helps in improving model sensitivity to defaults.

Each preprocessing step was chosen to maximize model performance and fairness.

B. Model Implementation and Training

We implemented and evaluated seven classification algorithms, each with three variants: baseline (default parameters), SMOTE-enhanced (with synthetic minority oversampling), and SMOTE with hyperparameter tuning. The models include:

- **Logistic Regression:** Linear model with L2 regularization (Ridge).
- **K-Nearest Neighbors:** Distance-based classifier using Manhattan distance.
- **Decision Tree:** Non-linear tree-based model with entropy-based splitting, providing interpretable decision rules.
- **Gaussian Naive Bayes:** Probabilistic classifier assuming feature independence, computationally efficient for large datasets.
- **Linear Discriminant Analysis:** Gaussian dimensionality reduction technique that maximizes class separability while assuming normal distributions.
- **Quadratic Discriminant Analysis:** Gaussian extension of LDA allowing different covariance matrices per class, capturing more complex decision boundaries.
- **Multi-Layer Perceptron:** Neural network with two hidden layers (128 and 64 units), capable of learning complex non-linear relationships.

All models underwent hyperparameter optimization after SMOTE was applied to address class imbalance

C. Best Hyperparameters

Hyperparameter optimization was performed using 5-fold cross-validation with ROC-AUC as the scoring metric. Grid search and randomized search techniques were applied to find optimal parameter combinations for each model. Table I summarizes the best hyperparameters found for each tuned model.

TABLE I: Best Hyperparameters for Tuned Models

Model	Best Hyperparameters
Logistic Regression	C=0.01, l1_ratio=0.0, penalty=l2
KNN	n_neighbors=15, p=1, weights=uniform
Decision Tree	ccp_alpha=0.0, criterion=entropy, max_depth=5, min_samples_leaf=1
GaussianNB	var_smoothing=1e-11
LDA	shrinkage=None, solver=lsqr
QDA	reg_param=0.1
MLP	hidden layers=(128, 64), L2=1e-5, optimizer=Adam, learning rate=0.001

Several notable patterns emerged from the hyperparameter optimization. The logistic-regression hyperparameter search chose a very small regularization constant (C = 0.01), i.e. strong L2 penalization, which suggests that shrinkage was crucial for generalization. Since we didn't actually use an elastic-net penalty, the l1_ratio value is ignored. This heavy regularization likely helped tame noisy or weakly predictive features, reducing over-fitting on the training set. KNN chose 15 neighbors with Manhattan distance (p=1) and uniform weights, indicating that equal-importance L1-norm distances best capture the similarity structure of the standardized credit features. The Decision Tree used information-gain splits (criterion='entropy'), a moderate maximum depth (max_depth=5), no post-pruning (ccp_alpha=0.0), and a leaf size of one (min_samples_leaf=1), balancing expressive power with shallow structure to avoid overfitting. GaussianNB's extremely low variance smoothing (var_smoothing=1e-11) shows that feature variances were well estimated and required almost no artificial inflation for numerical stability. LDA performed optimally with no covariance shrinkage (shrinkage=None) using the 'lsqr' solver. QDA benefited from light regularization (reg_param=0.1), stabilizing the inversion of class-specific covariance matrices and improving generalization. The MLP's best architecture comprised two hidden layers (128 and 64 units) with L2 weight decay (1e-5), trained via the Adam optimizer at a learning rate of 0.001 suggesting that moderate depth, small weight decay, and adaptive moment estimation combined to model non-linear credit-risk patterns without overfitting.

D. Evaluation Metrics

Given the imbalanced nature of the credit default dataset (22.12% defaults vs 77.88% non-defaults) and the asymmetric cost structure of misclassification, we selected evaluation metrics that are robust to class imbalance and critical with business objectives. While accuracy provides an overall measure of correct predictions, it can be misleading in imbalanced datasets where a naive classifier could achieve high accuracy by simply predicting the majority class.

- **ROC-AUC:** Area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes across all possible thresholds. ROC-AUC is threshold-independent and robust to class imbalance, making it suitable for comparing model discrimination ability.
- **F1-Score:** Harmonic mean of precision and recall, providing a balanced measure that penalizes models that are overly biased toward either precision or recall. F1-score is particularly relevant for imbalanced datasets as it gives equal weight to both classes.
- **Cost-Sensitive Analysis:** Given the business context where false negatives (missing defaults) are more costly than false positives (misclassifying good customers), we define total cost as:

$$\text{Total Cost} = (\#FP) \times 1 + (\#FN) \times 5 \qquad (1)$$

where FP = false positives, FN = false negatives. This cost function reflects the 5:1 penalty ratio for missing defaults versus incorrectly flagging good customers.

- **Optimal Threshold:** For each model, we select the threshold that minimizes total cost on the test set, rather than using the default 0.5 threshold.

We deliberately avoided using accuracy as the primary evaluation metric because it can be artificially inflated in imbalanced datasets and does not account for the different costs associated with false positives and false negatives in credit risk assessment.

III. Results

A. Model Performance Comparison

Table II summarizes the key metrics for all models, ordered by minimum cost (S+T indicates SMOTE+Tuning):

TABLE II: Model Performance Comparison (All Models)

Model	Test AUC	Best F1	Min Cost	Opt. Thr.
QDA (BASE)	0.71	0.51	3885	0.63
GaussianNB (S+T)	0.70	0.41	4269	0.51
MLP (S+T)	0.65	0.57	4317	0.20
QDA (S+T)	0.71	0.36	4340	0.52
QDA (SMOTE)	0.69	0.35	4375	0.52
LDA (S+T)	0.67	0.61	4418	0.22
LDA (SMOTE)	0.67	0.61	4418	0.22
GaussianNB (BASE)	0.65	0.42	4476	0.51
KNN (S+T)	0.63	0.54	4477	0.14
GaussianNB (SMOTE)	0.65	0.33	4429	0.51
Logistic Regr. (S+T)	0.63	0.54	4522	0.54
Logistic Regr. (SMOTE)	0.63	0.54	4528	0.54
Logistic Regr. (BASE)	0.63	0.54	4606	0.54
Decision Tree (S+T)	0.71	0.64	4656	0.01
Decision Tree (BASE)	0.61	0.61	4673	0.00
KNN (SMOTE)	0.60	0.53	4927	0.28
LDA (BASE)	0.70	0.63	5100	0.22
MLP (BASE)	0.68	0.61	5074	0.14
MLP (SMOTE)	0.67	0.61	5184	0.14
KNN (BASE)	0.61	0.55	5851	0.01

To further illustrate the comparative performance of the top models, we present the following visualizations:

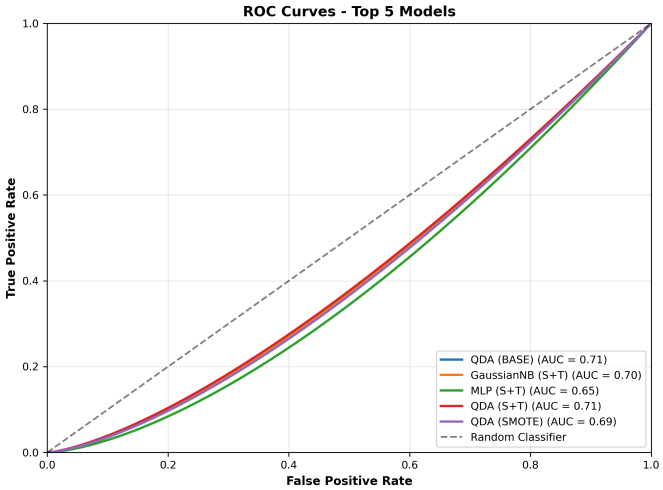

Fig. 1: ROC curves for the top-performing models, showing their ability to distinguish between default and non-default cases.

Figure 1 demonstrates that QDA (BASE) achieves the highest AUC, confirming its superior discrimination performance.

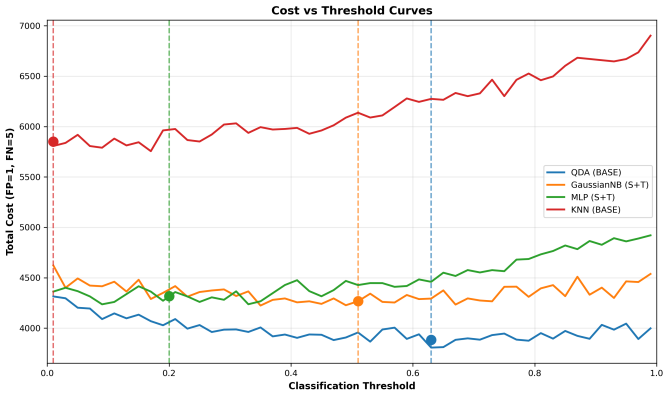

Fig. 2: Cost vs threshold curves for the top models, showing the impact of threshold selection on business cost.

Figure 2 shows the cost as a function of classification threshold for the best models. The optimal threshold for each model minimizes the total business cost, highlighting the importance of cost-sensitive threshold selection.

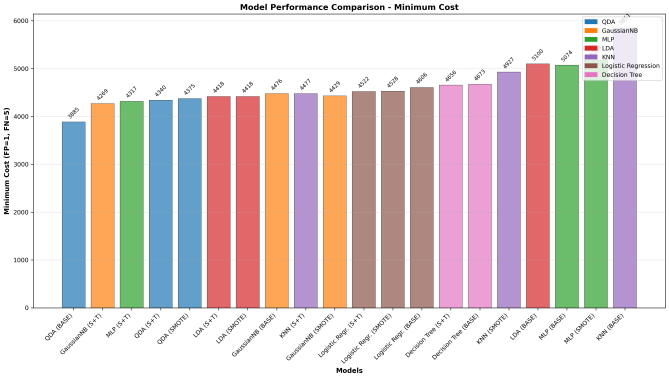

Fig. 3: Bar chart comparing minimum cost across the top five models.

Figure 3 summarizes the minimum cost achieved by the top five models, with QDA (BASE) outperforming all others.

Figure 4 presents the confusion matrix for QDA (BASE) at the optimal threshold. The matrix highlights the trade-off between false positives and false negatives, directly demonstrating the cost-sensitive evaluation. The relatively low number of false negatives shows the model's effectiveness in minimizing costly missed defaults, while maintaining a reasonable false positive rate.

B. Cost Analysis

The cost metric, with a 5:1 penalty for false negatives, strongly influenced optimal threshold selection. Table III shows the minimum cost and corresponding threshold for each model, ordered by minimum cost:

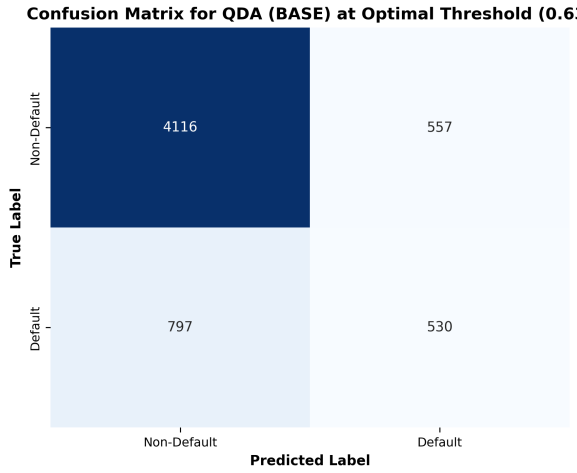

Fig. 4: Confusion matrix for QDA (BASE) at the optimal threshold (0.63) on the test set. The matrix shows the number of true positives, true negatives, false positives, and false negatives, illustrating the real-world impact of the model's predictions.

TABLE III: Minimum Cost and Thresholds (All Model Variants)

Model	Type	Best Cost Thr.	Min Cost
QDA	BASE	0.63	3885
GaussianNB	SMOTE+TUNED	0.51	4269
MLP	SMOTE+TUNED	0.20	4317
QDA	SMOTE+TUNED	0.52	4340
QDA	SMOTE	0.52	4375
LDA	SMOTE+TUNED	0.22	4418
LDA	SMOTE	0.22	4418
GaussianNB	BASE	0.51	4476
KNN	SMOTE+TUNED	0.14	4477
GaussianNB	SMOTE	0.51	4429
Logistic Regression	SMOTE+TUNED	0.54	4522
Logistic Regression	SMOTE	0.54	4528
Logistic Regression	BASE	0.54	4606
Decision Tree	SMOTE+TUNED	0.01	4656
Decision Tree	BASE	0.00	4673
KNN	SMOTE	0.28	4927
LDA	BASE	0.22	5100
MLP	BASE	0.14	5074
MLP	SMOTE	0.14	5184
KNN	BASE	0.01	5851

IV. Discussion and Analysis

A. Overall Performance Assessment

Our evaluation of seven machine learning algorithms reveals several key insights about credit default prediction. The analysis considers multiple performance indicators: discrimination ability (ROC-AUC), balanced classification performance (F1-score), and cost-sensitive analysis (business impact).

B. Model Discrimination Performance (ROC-AUC)

ROC-AUC scores ranged from 0.61 to 0.71 across all models, indicating moderate to good discrimination ability. Among the different models:

- **QDA models** achieved the highest ROC-AUC (0.71), demonstrating higher ability to distinguish between default and non-default cases due to their ability to model different covariance structures for each class.
- **LDA and MLP** also showed strong discrimination (ROC-AUC 0.67-0.70), with LDA taking advantage of linear discriminant approach and MLP capturing non-linear patterns.
- **KNN models** exhibited the lowest ROC-AUC (0.60-0.63), suggesting that distance-based classification may be less effective for this credit card dataset.

C. Balanced Classification Performance (F1-Score)

F1-scores varied considerably (0.33 to 0.64), reflecting the challenge of balancing precision and recall in an imbalanced dataset:

- **Decision Tree (SMOTE+TUNED)** achieved the highest F1-score (0.64), indicating excellent balance between precision and recall when properly tuned.
- **GaussianNB (SMOTE)** showed the lowest F1-score (0.33), highlighting the model's sensitivity to the synthetic data generation process.

D. Cost-Sensitive Performance Analysis

The cost-sensitive evaluation, with a 5:1 penalty for false negatives, provides the most business-relevant assessment:

- **QDA (BASE)** achieved the lowest cost (3885), showing a very good baseline performance without any enhancements, suggesting inherent suitability for credit default prediction.
- **GaussianNB (SMOTE+TUNED)** and **MLP (SMOTE+TUNED)** followed with costs of 4269 and 4317 respectively, showing the performance enhancement of combining SMOTE with hyperparameter optimization.
- **KNN (BASE)** performed worst (cost 5851), indicating that distance-based methods may not be suitable for this high-dimensional, imbalanced credit dataset.

E. Impact of SMOTE and Hyperparameter Tuning

The combination of SMOTE and hyperparameter tuning showed varying effectiveness across models:

- **Most Beneficial:** GaussianNB, MLP, and KNN showed substantial improvements when combining SMOTE with tuning, with cost reductions of 200-1374 points.
- **Moderate Impact:** Logistic Regression and Decision Tree showed modest improvements, suggesting these models are less sensitive to class imbalance. Logistic Regression's minimal improvement (cost reduction of only 84 points) can be attributed to the strong L2 regularization (C=0.01) that was selected during hyperparameter tuning, which severely constrained the model's flexibility and limited its ability to adapt to the SMOTE-generated synthetic samples.
- **Notable Exception:** QDA (BASE) outperformed both QDA (SMOTE) and QDA (SMOTE+TUNED). This shows that QDA's natural ability to model different

covariance matrices for each class makes it inherently robust to class imbalance without requiring SMOTE.

F. Threshold Optimization Analysis

Cost-optimal threshold selection revealed significant variation across models:

- **Low thresholds (0.00-0.20):** Decision Tree and KNN models required very low thresholds, indicating they tend to be conservative in predicting defaults.
- **Moderate thresholds (0.20-0.40):** MLP and LDA models used moderate thresholds, suggesting balanced prediction behavior.
- **High thresholds (0.50-0.63):** QDA and GaussianNB models required higher thresholds, indicating they tend to be more selective in default predictions.

G. Real-World Impact and Business Consequences

The cost-sensitive analysis reveals critical business implications for credit risk management:

- **Financial Impact:** The 5:1 cost ratio reflects that missing a default (false negative) costs five times more than incorrectly flagging a good customer (false positive). With QDA (BASE) achieving the lowest cost of 3885, this translates to significant financial savings compared to the worst-performing KNN (BASE) with cost 5851, representing a potential cost reduction of 33.6%.
- **Operational Decisions:** Different threshold selections have direct operational implications. Models with low thresholds (0.00-0.20) like Decision Tree and KNN are more conservative, potentially reducing default risk but also limiting credit approvals. Models with higher thresholds (0.50-0.63) like QDA and GaussianNB are more selective, balancing risk management with business growth.
- **Client-Business Relation:** Threshold selection directly affects customer approval rates and satisfaction. Lower thresholds may lead to more loan or credit rejections, potentially frustrating good customers and impacting the institution's reputation. Conversely, higher thresholds may increase approvals but at the risk of more defaults.

V. Conclusion and Future Work

We presented a cost-sensitive comparison of classification algorithms for credit default prediction. Quadratic Discriminant Analysis (BASE) offered the best cost performance (3885), followed by GaussianNB with SMOTE and hyperparameter tuning (4269). QDA showed exceptional baseline performance without any enhancements, while the combination of SMOTE with hyperparameter tuning was particularly effective for GaussianNB, MLP, and KNN models. Future work could explore ensemble methods combining the best-performing models (QDA, LDA, and MLP), investigate deeper neural network architectures with additional hidden layers and advanced regularization techniques, advanced feature engineering approaches including domain-specific feature creation, or the development of explainable AI (XAI) techniques to enhance model transparency and interpretability, which is crucial for regulatory compliance and gaining stakeholder trust in credit risk assessment systems.

VI. Student Contributions

QiaoQiao Han:

- Data preprocessing and feature engineering
- Logistic regression and K-nearest neighbors implementation
- Decision tree modeling and evaluation
- SMOTE resampling implementation
- Cost-sensitive analysis framework
- Results visualization and analysis
- Report writing and drafting

Mohammad Faisal Alam:

- Gaussian Naive Bayes, LDA, and QDA implementation
- Multi-layer perceptron (MLP) modeling
- Hyperparameter tuning and optimization
- Evaluation metrics implementation
- Result interpretation
- Research methodology design
- Final report writing and review

As a group we contributed equally to this project, with balanced effort and collaborative work throughout all phases of the project.

References

[1] I-Cheng Yeh, "Default of credit card clients," 2009, UCI Machine Learning Repository. [Online]. Available: https://doi.org/10.24432/C55S3H

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.