# Credit Risk Prediction Using Supervised Machine Learning

**Author: Bozhao Wang**

**Department of Economics**

**University of Calgary**

**ECON 611.97 – Machine Learning in Economics**

**Instructor: Dr. Arvind Magesan**

**April 21, 2025**

## 1. Abstract

Today in digital finance, banks and financial institutions often use credit risk analysis models to predict the probability of default and classify potential defaulters. Its goal is to minimize losses and build stronger relationships with long-term clients. This project applies supervised machine learning algorithms to predict credit risk using a publicly available credit card default dataset from Taiwan. The motivation for the project narrows down to the increasing need for data-driven decision-making in financial institutions. To better assess institutions with high exposure to credit risk, we explored the classification of default vs. non-default clients through various models including Logistic Regression, Decision Trees, Random Forests, XGBoost, K-Nearest Neighbors, and Neural Networks.  Given the dataset's imbalance, we applied SMOTE (Synthetic Minority Over-sampling Technique) to improve minority class detection. Among all models, Random Forest + SMOTE and Logistic Regression + SMOTE emerged as strong performers. While Random Forest + SMOTE achieved the highest F1-score (0.53) and a strong AUC (0.77), Logistic Regression + SMOTE achieved the highest recall (0.62) for the default class.

In order to optimize model's performances, we performed hyperparameter tuning on the two best performing models, identified the most significant financial behaviors that contribute to default risk, such as past payment delays and recent delays. These insights efficiently support financial decisions. At the end, the final model has potential for real-world deployment, offering a transparent and reliable scoring mechanism to assist lenders in managing risk and improving credit policy.

## 2. Introduction

At the macro level, credit risk describes how large-scale debt problems such as high inflation or sudden increases in unemployment can put pressure on banks, impacting the ability of businesses to repay their debts. At the micro level, credit risk is about whether an individual borrower such as a credit card user or loan applicant might fail to repay their loan. In this report, the primary focus will be the analysis at the micro economics perspective.

In this context, accurately predicting default risk serves as an important goal for banks and financial institutions. With a large volumes of financial transactions occur daily, this developed machine learning model can benefit both lenders and borrowers. For lenders, it helps reduce financial losses, manage risk exposure, and allocate resources more efficiently. For borrowers, accurate risk assessment leads to fairer access to credit and more responsible lending practices.

Additionally, implementing high-accurately prevent misclassifying clients. Misclassification occurs when high-risk borrowers are incorrectly classified as low-risk. It make institutions suffer financial loss due to default. Vise versa, if low-risk borrower is wrongly denied credit, the institution loses potential revenue and damages its relationship with clients. At scale, these issues can influence lending policies, impact market confidence, and even contribute to broader economic problems.

To address these challenges, the project have these objectives:

- Predicting the likelihood of loan default using supervised ML models.
- Identifying key factors associated with loan default using feature importance.
- Comparing model performance across several classifiers, including Logistic Regression, Decision Tree, Random Forest , XGBoost, K-Nearest Neighbours, and Neural Network, to evaluate their predictive power and trade-offs in accuracy, recall, and AUC.
- Improving model performance by addressing class imbalance and optimizing hyperparameters.

These objectives provide a completed pipeline for developing, evaluating, and refining credit risk models. It offers a data-driven framework that financial institutions can use to make better lending decisions.

## 3. Literature Review

Galindo and Tamayo(2000)[1] conducted a comprehensive comparative analysis of various classification algorithms using a mortgage loan dataset from a large financial institution. The study is motivated by the past financial crises in the 1980s and 90s, such as the U.S. savings and loan crisis[2], where over 1000 savings and loan institutions went bankrupt due to risky lending practices. To address such problems, their study argues that using advanced machine learning method can lead to stronger financial systems.

The paper compared four main methods for credit risk classification:

- Probit Regression: Traditional Statistical Method
- CART(Classification and Regression Trees): Decision tree algorithm
- Neural Networks: Non-linear modelling
- K-Nearest Neighbors(KNN): distance-based classification

[1] 1. Jorge Galindo and Pablo Tamayo, (PDF) credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications, February 2000, https://www.researchgate.net/publication/5144412_Credit_Risk_Assessment_Using_Statistical_and_Machine_Learning_Basic_Methodology_and_Risk_Modeling_Applications.

[2] 1. Kenneth J. Robinson, “Savings and Loan Crisis,” Federal Reserve History, November 22, 2013, https://www.federalreservehistory.org/essays/savings-and-loan-crisis.

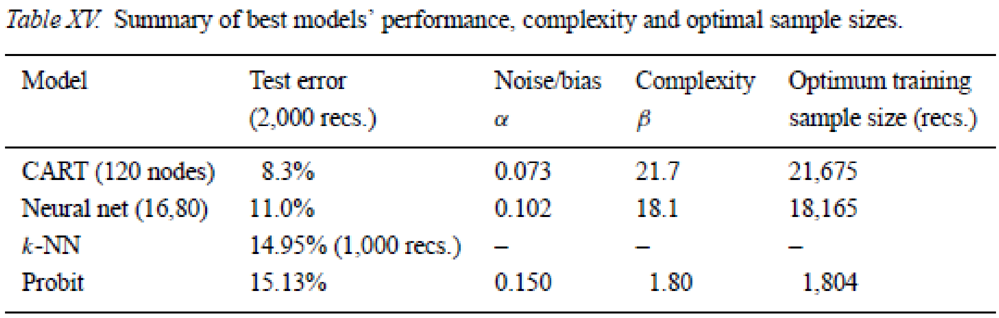


*Table XV.* Summary of best models' performance, complexity and optimal sample sizes.

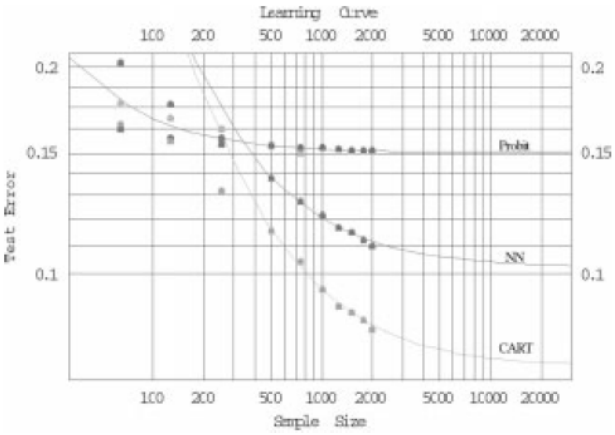
| Model | Test error<br>(2,000 recs.) | Noise/bias<br>$\alpha$ | Complexity<br>$\beta$ | Optimum training<br>sample size (recs.) |
|---|---|---|---|---|
| CART (120 nodes) | 8.3% | 0.073 | 21.7 | 21,675 |
| Neural net (16,80) | 11.0% | 0.102 | 18.1 | 18,165 |
| *k*-NN | 14.95% (1,000 recs.) | – | – | – |
| Probit | 15.13% | 0.150 | 1.80 | 1,804 |

*Figure 1. CART model performance in default prediction (Galindo and Tamayo 2000, 28)*[3]

Among the main methods, Classification and Regression Trees (CART) outperformed with an 8.3% error rate. Neural networks shows an average error of 11%, and k-Nearest Neighbour yielded a 14.95% error rate. Traditional statistical models such as the probit model performed poorly with 15.13% error rate. From the result, the study shows that logistic regression and probit models lack the flexibility to capture non-linear patterns in borrower behaviour. However, machine learning methods such as Random Forest and Neural Networks demonstrate greater adaptability in modelling complex relationships within high-dimensional data. Additionally, the paper introduces a methodology based on error curve analysis, which provides insights into noise sensitivity and model bias. These findings are particularly relevant for individuals aiming to build scalable, automated credit-scoring systems. The authors also emphasize the practical application of such models in institutional-level risk management and early warning systems.

[3] Jorge Galindo and Pablo Tamayo, *Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications*, Computational Economics 15 (2000): 28, https://doi.org/10.1023/A:1008699112516.

*Figure 2. Comparison of test error across four algorithms (Probit, CART, Neural Nets, and k-NN). Source: Screenshot from Galindo and Tam (2000, 29)*[4]



Overall, this work directly supports my current research direction, which I aim to evaluate and compare modern machine learning classifiers such as Random Forest and XGBoost against logistic regression. However, this paper only focused on binary classification accuracy and does not consider metrics such as AUC-ROC curves or precision-recall trade-offs. In this research, I will address these gaps by integrating contemporary models with modern evaluation metrics and techniques.

Montevechi et al. (2024)[5] present a more recent review that looks at how machine learning has explained credit risk modelling. The author examined 67 studies published between 2000 and 2022, went through each stage of the credit scoring model pipeline from data preprocessing to classification algorithms, hyperparameter tuning and post-model evaluation. One of the key insights from the paper is the importance of handling class imbalance. The review recommends techniques like Synthetic Minority Oversampling Technique (SMOTE),

[4] Jorge Galindo and Pablo Tamayo, *Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications*, Computational Economics 15 (2000): 29, https://doi.org/10.1023/A:1008699112516.

[5] 1. André  Aoun Montevechi et al., “Advancing Credit Risk Modelling with Machine Learning: A Comprehensive Review of the State-of-the-Art,” Engineering Applications of Artificial Intelligence, August 3, 2024, https://www.sciencedirect.com/science/article/pii/S0952197624012405?via%3Dihub.

which was adopted in this project to improve model recall and balance precision in the minority class.

Several findings from the review inform the methodological decisions in this research. Logistic regression model remains the most used statistical baseline due to its simplicity and regulatory acceptability. However, ensemble models such as Random Forest(RF) and XGBoost outperform both statistical models and individual ML classifiers in predictive accuracy. The review also highlights that many research studies relied too heavily on accuracy. It recommends that comparing accuracy with metrics like AUC, recall, and F1-score will provide a more comprehensive evaluation of model performance. Furthermore, the review reports that over 39% of the analyzed studies used SMOTE to address class imbalance. As shown in Figure 3, oversampling techniques were used in 59.26% of cases, compared to 40.74% for undersampling approaches.

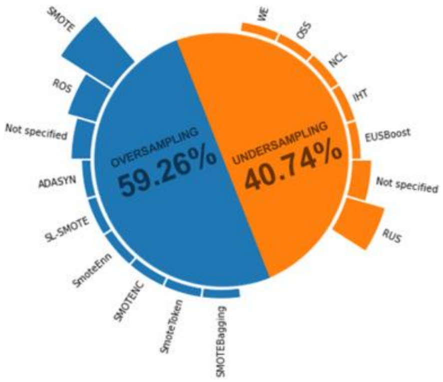


*Figure 3.[6] Proportion of oversampling and undersampling methods used to address class imbalance in credit risk modelling. Adapted from Montevechi et al. (2024, 17).*

[6] André Aoun Montevechi et al., *"Advancing Credit Risk Modelling with Machine Learning: A Comprehensive Review of the State-of-the-Art," Engineering Applications of Artificial Intelligence* 137 (2024): 12, Figure 9, https://doi.org/10.1016/j.engappai.2024.109082.

Overall, the findings of Montevechi et al. (2024) support the methodological decisions in this study: Starting with a logistic regression baseline, incorporating advanced ML classifiers, applying SMOTE for class imbalance, and eventually using regularization and model evaluation metrics like AUC and F1-score to guide performance comparisons.

In addition to traditional and ensemble learning methods, deep learning models have also been explored for credit risk prediction. The study by Moula and Pandey (2018)[7] compared various machine learning classifiers with neural networks. The authors found that while decision trees and Random Forest performed well in accuracy, deep learning methods were more effective in capturing non-linear relationships and hidden patterns in complex borrower data. These insights support the inclusion of neural networks in my project as a complementary approach.

In conclusion, the reviewed literature provides a comprehensive foundation for this study's methodological design. Galindo and Tamayo (2000) highlight how machine learning models can outperform traditional statistical methods in credit risk prediction. Montevechi et al. (2024) expand on this by highlighting the importance of handling class imbalance and using more modern evaluation metrics like AUC and F1-score. Finally, Moula and Pandey (2018) show that deep learning can capture complex patterns in borrower data. These literature studies help structure this project's strategy flow: Starting with logistic regression as a baseline model, expanding on tree-based models and deep learning models, finally applying the SMOTE technique and hyperparameter tuning to address class imbalance issues and optimize model performance.

[7] 1. Peter  Martey Addo, Dominique Guegan, and Bertrand Hassani, "Credit Risk Analysis Using Machine and Deep Learning Models," MDPI, April 16, 2018, https://www.mdpi.com/2227-9091/6/2/38?amp=1.

# 4. Data Description & Methodologies

## 4.1. Data Description and Preprocessing

This study uses the “Default of Credit Card Clients”[8] dataset from Taiwan. It is a public dataset for credit risk analysis. It contains records of 30,000 credit card holders, each with 23 attributes reflecting demographic characteristics, billing information, payment history, and a binary target variable indicating default in the next month. The column “Default Payment Next Month” is used as the binary outcome, where 0 indicates a client who repaid on time and 1 represents a client who defaulted.

**Original Features**:

- Demographics: “SEX”, “EDUCATION”, “MARRIAGE”, “AGE”
- Billing amount: “BILL_AMT1” … “BILL_AMT6” (Monthly Bill Statements)
- Payments: “PAY_AMT1”... “PAY_AMT6” (Actual Payment Amounts)
- Payment History: “PAY_0” … “PAY_6” (Repayment Status Each Month)
- Credit Limit: “LIMIT_BAL” (Initial Credit Limit)

**Preprocessing Steps**:

1. Data Cleaning: No missing values existed in the original dataset. 795 “Nan” values exist after generating additional features. Rows with “Nan” values were dropped.
2. Categorical Encoding: Applied “One-Hot Encoding” to convert categorical variables “SEX”, “EDUCATION”, “MARRIAGE” into numerical format.
3. Train-Test Split: The dataset was split into 80% training and 20% testing, maintaining class distribution.

[8] Yeh, I-Cheng. 2009. Default of Credit Card Clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H.

4. Feature Scaling: Continuous numerical features such as “LIMIT_BAL”, “PAY_AMT1”... “PAY_AMT6” were scaled using “StandardScaler” to ensure features had mean of zero and standard deviation of one. It benefits distance-based and regulatized models.

This preprocessing pipeline ensured that the dataset was clean, numerically stable, and ready for feature engineering and machine learning modeling.

### 4.2. Feature Engineering

To enhance the model’s ability to detect default risk, several new features were engineered based on financial domain knowledge and temporal trends observed in billing and repayment behavior. These derived variables provide a more nuanced understanding of a client’s payment capacity, utilization behavior, and risk exposure.

Engineered Features:

- AVG_UTIL_RATIO = $\frac{1}{6}\sum_{i=1}^{6} \frac{BILL\_AMT}{LIMIT\_BAL}$. Captures the average proportion of credit limit used.
- AVG_REPAY_RATIO = $\frac{1}{6}\sum_{i=1}^{6} \frac{PAY\_AMT}{BILL\_AMT}$. Measures how much of the bill was repaid on average.
- PAY_TO_LIMIT = $\frac{1}{6}\sum_{i=1}^{6} \frac{PAY\_AMT}{LIMIT\_BAL}$. Indicates general repayment capacity based on credit availability.
- NUM_DELAYS = $\sum_{i=0}^{6} 1(PAY_i > 0)$. Counts the number of months where the client delayed payment.

- LONGEST_DELAY = MAX($PAY_0 ... PAY_6$). Represents the most severe delay encountered by the client.
- NUM_ZERO_PAYMENT_MONTHS = Count(PAY_AMT1–6 = 0). Flags customers who made no payment in multiple months.
- CREDIT_LIMIT_RECENT_DELAY = LIMIT_BAL × PAY0. Quantifies current risk exposure based on the latest repayment delay.
- AGE_TIMES_PAYMENT = AGE × PAY_AMT1. Combines age and recent payment size as a proxy for repayment ability.

### 4.3. Modeling and Evaluation Strategy

To assess the predictive power of various classification techniques, we implemented and compared a set of supervised machine learning algorithms.

#### 4.3.1. Models Implemented

Logistic Regression (Linear baseline Model)

- The baseline classifier Extended with L1 (Lasso) and L2 (Ridge) regularization to address multicollinearity and control overfitting.

Decision Tree Classifier (Non-linear, Interpretable)

- A tree-based algorithm that learns hierarchical splitting rules.

Random Forest (Ensemble, Bagging)

- An ensemble of decision trees trained on bootstrapped subsets with random feature selection.

XGBoost (Ensemble, Boosting)

- A gradient-boosted tree model optimized for speed and performance.

K-Nearest Neighbors (KNN)

- A non-parametric method that classifies based on the majority class of the nearest neighbors.

Neural Network (Deep Learning)

- Captures complex non-linear patterns but lacks transparency.

#### 4.3.2. Handling Class Imbalance

The dataset had approximately 22% default cases which lead to biased models that favor the majority class. To address this, we used SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic default cases in the training set.

#### 4.3.3. Hyperparameter Tuning

Model hyperparameters were tuned using GridSearchCV with 5-fold cross-validation, optimizing for the AUC (Area Under the ROC Curve) score.

#### 4.3.4. Evaluation Metrics

Performance was assessed using the following metrics:

- Accuracy: Overall correctness.

- Recall: Focus on correctly identifying defaulters.

- Precision: Proportion of predicted defaulters who were actually defaulters.

- F1 Score: the combination of precision and recall.

- AUC-ROC: Area under the curve, representing model’s ability to rank defaulters above non-defaulters.

- Confusion Matrix: Breakdown of true positives, false positives, true negatives, and false negatives.

In financial applications, recall and AUC are often prioritized to avoid missing risky borrowers.

### 4.4 Mathematical Formulation of Key Models

In this section, we will go over the mathematical formulation of those key models mentioned earlier.

#### 4.4.1 Logistic Regression

Logistic Regression estimates the probability that a given input belongs to the positive class.

$$P(y = 1|X) = \frac{1}{(1+e^{-(\beta 0+\beta 1X1+...+\beta nXn)})}$$

The model is trained by minimizing the binary cross-entropy loss:

$$\text{Loss Function: } L(\beta) = -\sum_{i=1}^{n}[y_i log(\widehat{y_i}) + (1 - y_i)log(1 - \widehat{y_i})]$$

Lasso(L1) & Ridge(L2) regulations are added to prevent overfitting.

### 4.4.2 Decision Tree

A decision tree splits data by selecting features and thresholds that maximize information gain at each node:

Gini Impurity:

$$G(t) = 1 - \sum_{k=1}^{k} P_k^2$$

### 4.4.3 Random Forest and XGBoost

Random Forest is an ensemble of decision trees trained on random subsets with majority voting.

XGBoost builds trees sequentially, optimizing a regularized objective function:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{k=1}^{k} \Omega(f_k)$$

“ $l$ ” is the loss function.

$\Omega(f_k)$ penalizes model complexity.

### 4.4.4 K-Nearest Neighbours (KNN)

KNN classifies a test point x* based on the majority vote of its k nearest neighbours using a distance metric:

$$d\,(x_i, x\ ^*) = \sqrt{\sum_{j=1}^{p} (x_{i,j} - x_j^*)^2}$$

### 4.4.5 Neural Network

In this report, we use a simple feedforward neural network with one hidden layer.

Input Layer → Hidden Layer → Output Layer

# 5. Results and Evaluation

## 5.1. Performance Metrics

To evaluate model performance in this study, we are using four modern metrics:

- Accuracy: measures how many cases are correctly predicted out of all predictions.
- Recall: The model's ability to identify actual defaulters
- F1-Score: The combination of Precision and Recall. Measures the model's ability to correctly predict the recall rate while maintaining a low false positive rate.
- AUC(Area Under the ROC Curve): Measures the model's ability to distinguish between classes.

In this setting, we are predicting the default rate for financial risk assessment; recall and AUC are thus emphasized as the most relevant metrics given the high cost of false negatives in credit scoring. We will prioritize these two when comparing model performance.

## 5.2. Individual Model Results

In this section, we will compare the effectiveness of the SMOTE technique across selected models. Showing the impact of addressing class imbalance on model performance.

### 5.2.1. Logistic Regression

Logistic regression was used as a baseline model to evaluate credit default risk. Before addressing class imbalance, the model achieved a relatively high overall accuracy of 81.8%, but recall for the minority class was only 32%, indicating poor sensitivity to actual default cases. This suggests the model was heavily biased toward the non-default class, signaling SMOTE for improvements.

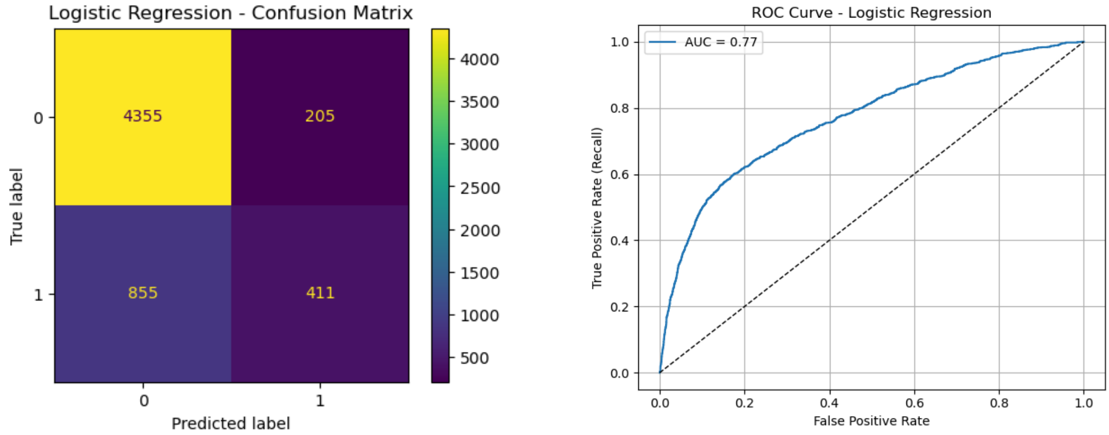


*Figure 4. Confusion Matrix & ROC Curve for Logistic Regression*

After applying SMOTE to address class imbalance, the performance of the logistic regression model improved, with overall accuracy dropping slightly from 81.8% to 73.7%. The recall for defaulters increased significantly from 32% to 62%, indicating the model became much more sensitive to identifying actual defaulters. The F1-score for the minority class also rose from 0.44 to 0.51, reflecting a better balance between precision and recall. Finally, the AUC score improved to 0.74, demonstrating better overall classification capability compared to the pre-SMOTE model. The model becomes more effective at flagging high-risk borrowers.

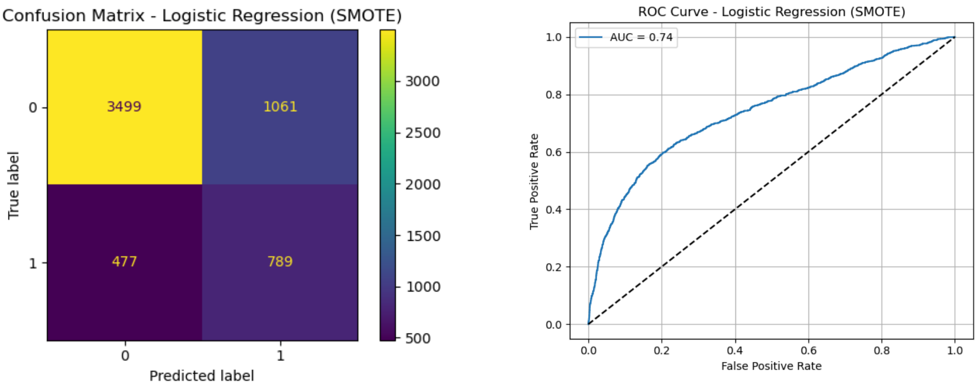


*Figure 5. Confusion Matrix & ROC Curve for Logistic Regression - After SMOTE*

### 5.2.2. Decision Tree

After applying SMOTE, the Recall for decision tree model increases from 41% → 50%. The F1-score remained stable at 41%. The overall accuracy dropped from 73.1% to 68.9%, which is expected when improving minority class detection in imbalanced datasets. The AUC score improved slightly from 0.61 to 0.62, suggesting only marginal gains in discriminatory power. These results reinforce the need for further tuning to optimize performance across classes.

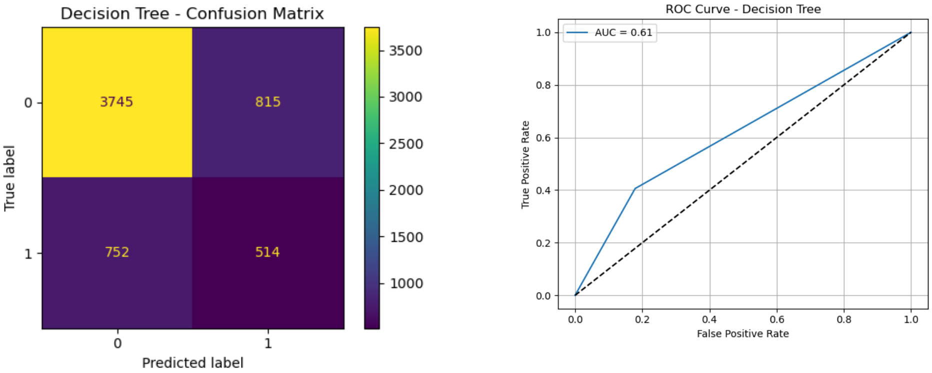


*Figure 6. Confusion Matrix & ROC Curve for Decision Tree*

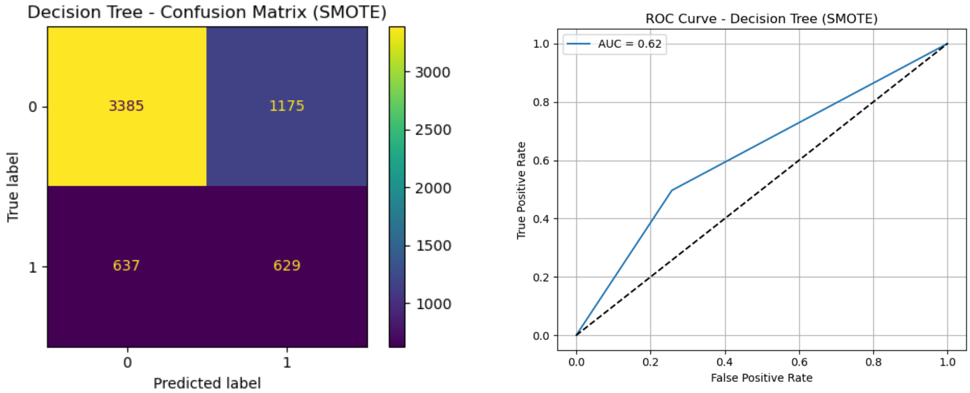


*Figure 7. Confusion Matrix & ROC Curve for Decision Tree - After SMOTE*

### 5.2.3. Random Forest

The Random Forest model delivered the highest overall accuracy among all models at 82.42%, with an AUC score of 0.773. After applying SMOTE, the recall for defaulters rose from 38% to 51%, while the F1-score also increased from 49% to 53%. Although the overall accuracy decreased slightly from 82.4% to 80.3%, and the AUC dropped slightly from 0.77 to 0.766, the model became significantly more responsive to default cases.

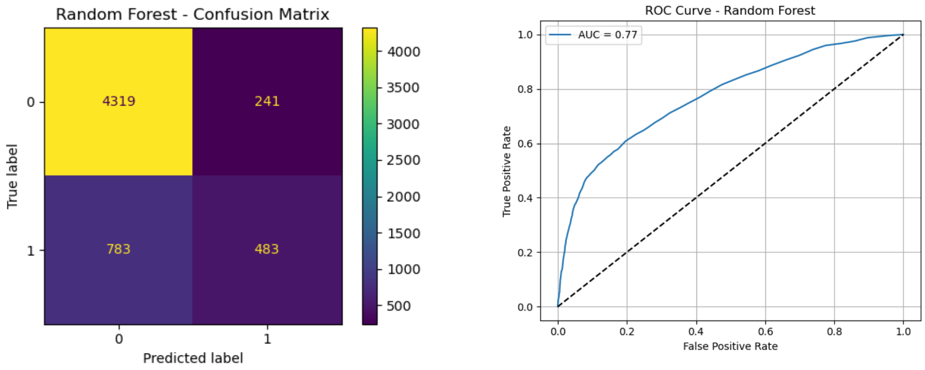


*Figure 8. Confusion Matrix & ROC Curve for Random Forest*

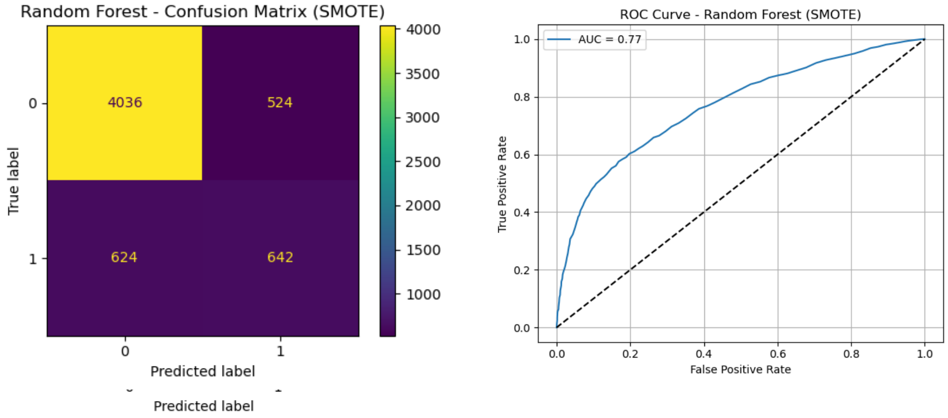


*Figure 9. Confusion Matrix & ROC Curve for Random Forest - After SMOTE*

### 5.2.4. XGBoost

After applying SMOTE, the recall for XGBoost increased from 39% to 48%, and the F1-score is 49%. The model achieved a slightly lower accuracy of 77.97% and a marginally reduced AUC score of 0.743. Overall, the XGBoost + SMOTE combination gives a better balance between class sensitivity and general performance.

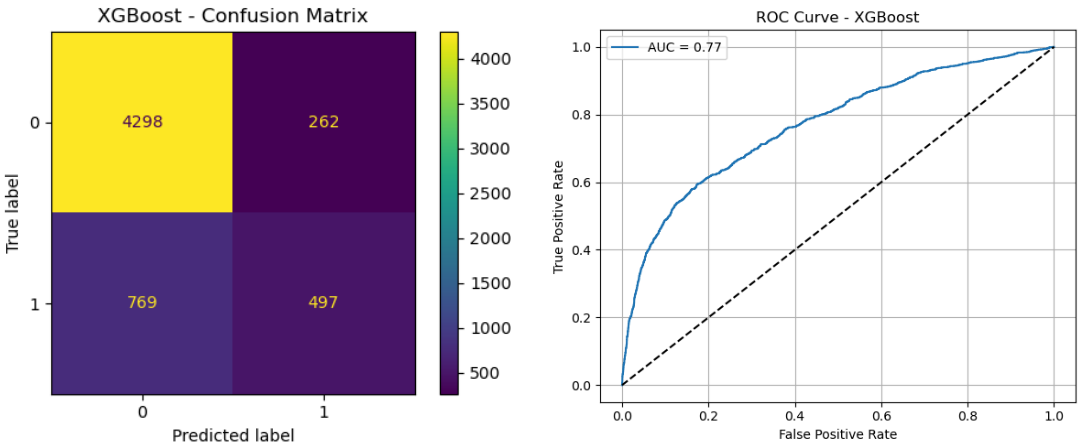


*Figure 10. Confusion Matrix & ROC Curve for XGBoost*

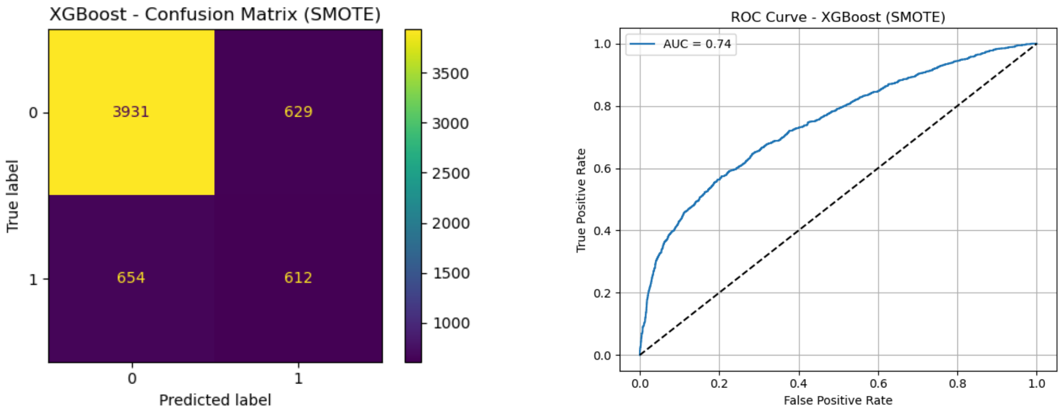


*Figure 11. Confusion Matrix & ROC Curve for XGBoost - After SMOTE*

### 5.2.5. K-Nearest-Neighbour (KNN)

After applying SMOTE, the minority class recall improved from 36% to 57%. However, the trade-off is that overall accuracy dropped to 71.19%, and the AUC score slightly decreased to 0.700. The F1-score for default prediction only increased from 45% to 46%. Despite the drop in accuracy, this model demonstrates better balance between sensitivity and general performance than its non-SMOTE statistics.

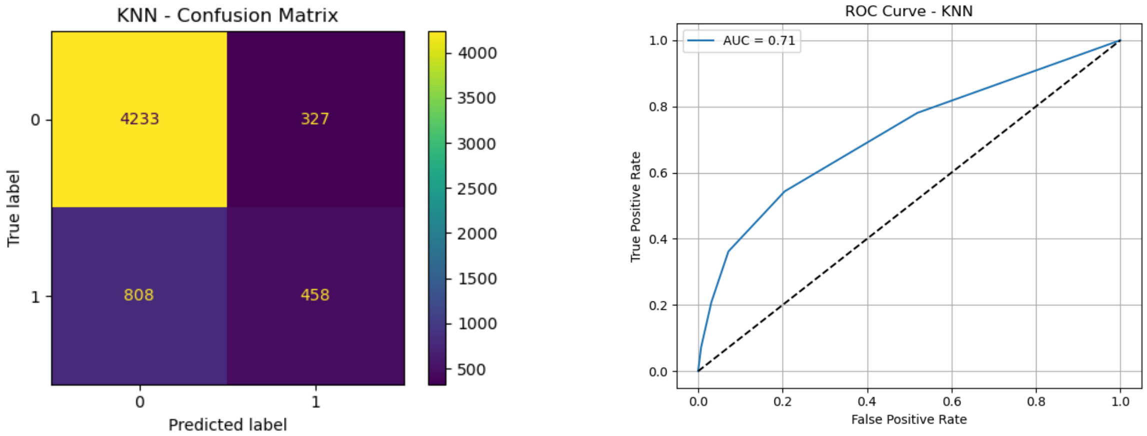


*Figure 12. Confusion Matrix & ROC Curve for KNN*

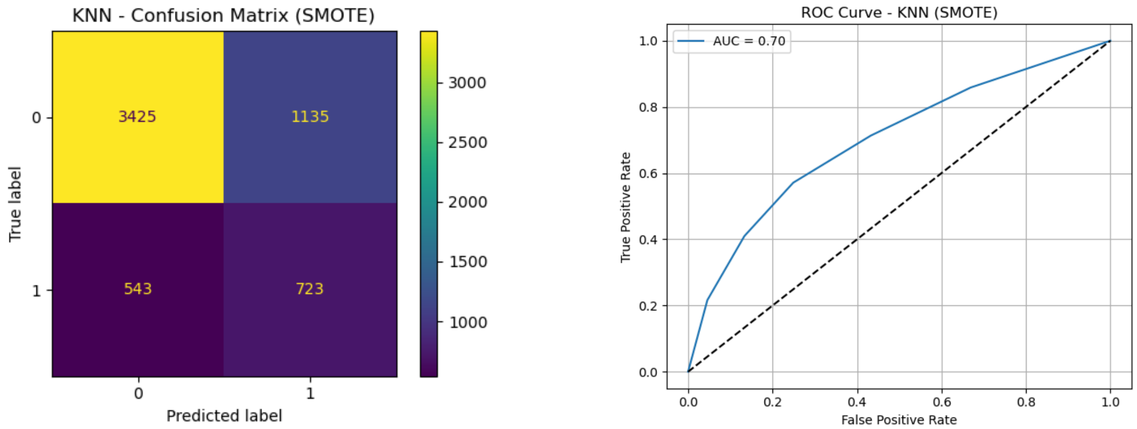


*Figure 13. Confusion Matrix & ROC Curve for KNN - After SMOTE*

### 5.2.6. Neural Network

After applying SMOTE, the neural network’s recall for defaulters improved from 40% to 50%, and the F1-score increased to 50%, indicating a better balance between identifying defaulters and minimizing false positives. The accuracy dropped slightly to 78.145, and the AUC dropped to 0.73; the trade-off improved overall model performance.

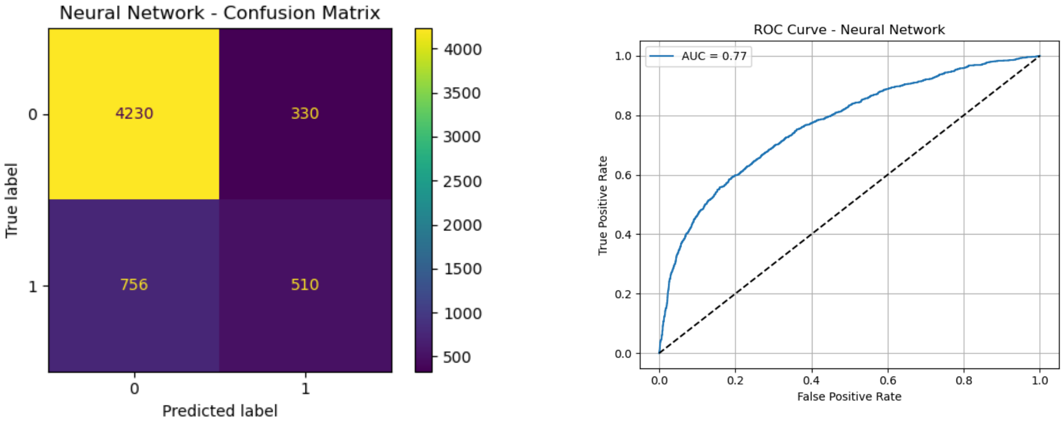


*Figure 14. Confusion Matrix & ROC Curve for NN*

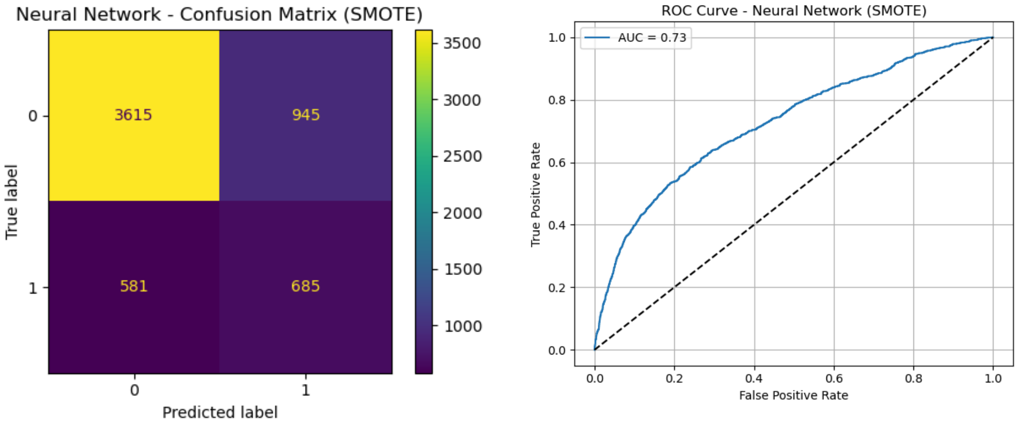
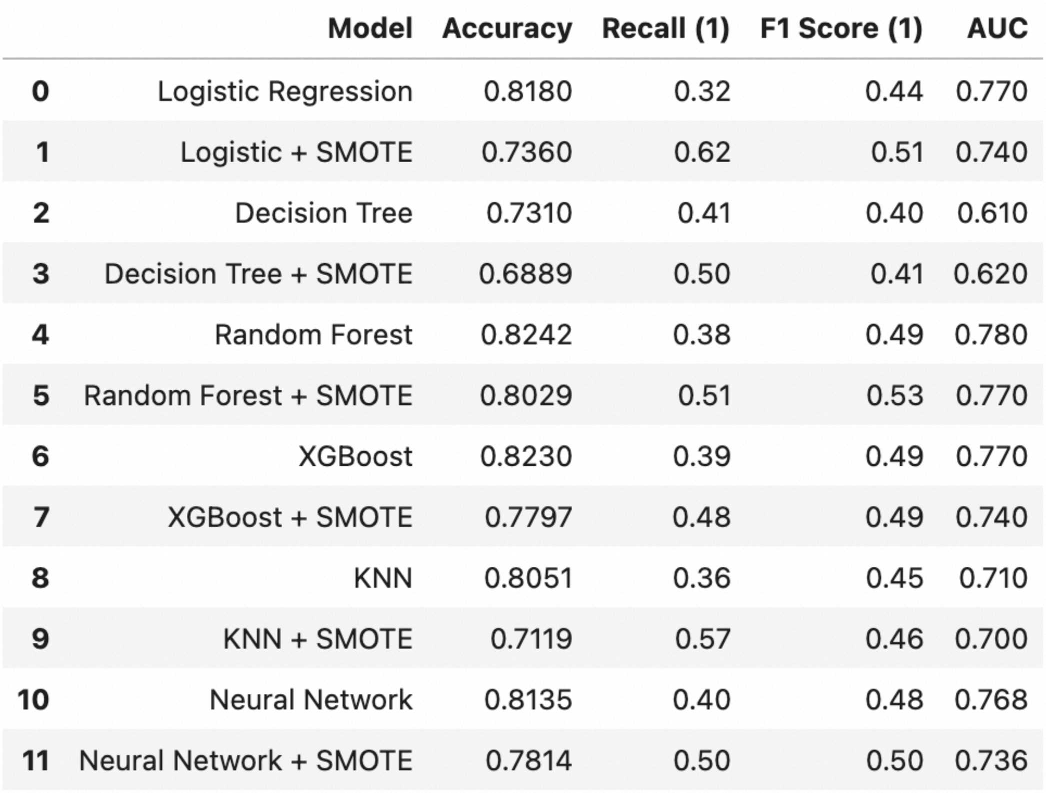


*Figure 15. Confusion Matrix & ROC Curve for NN - After SMOTE*

Overall, among all models, Logistic Regression and Random Forest with SMOTE showed the greatest recall gains with minimal AUC trade-off, while XGBoost offered the most balanced AUC and F1-score. These findings guide the final model selection.

## 5.3. Model Comparison Table



|  | Model | Accuracy | Recall (1) | F1 Score (1) | AUC |
|---|---|---|---|---|---|
| 0 | Logistic Regression | 0.8180 | 0.32 | 0.44 | 0.770 |
| 1 | Logistic + SMOTE | 0.7360 | 0.62 | 0.51 | 0.740 |
| 2 | Decision Tree | 0.7310 | 0.41 | 0.40 | 0.610 |
| 3 | Decision Tree + SMOTE | 0.6889 | 0.50 | 0.41 | 0.620 |
| 4 | Random Forest | 0.8242 | 0.38 | 0.49 | 0.780 |
| 5 | Random Forest + SMOTE | 0.8029 | 0.51 | 0.53 | 0.770 |
| 6 | XGBoost | 0.8230 | 0.39 | 0.49 | 0.770 |
| 7 | XGBoost + SMOTE | 0.7797 | 0.48 | 0.49 | 0.740 |
| 8 | KNN | 0.8051 | 0.36 | 0.45 | 0.710 |
| 9 | KNN + SMOTE | 0.7119 | 0.57 | 0.46 | 0.700 |
| 10 | Neural Network | 0.8135 | 0.40 | 0.48 | 0.768 |
| 11 | Neural Network + SMOTE | 0.7814 | 0.50 | 0.50 | 0.736 |

*Figure 16. Model Comparison Summary*

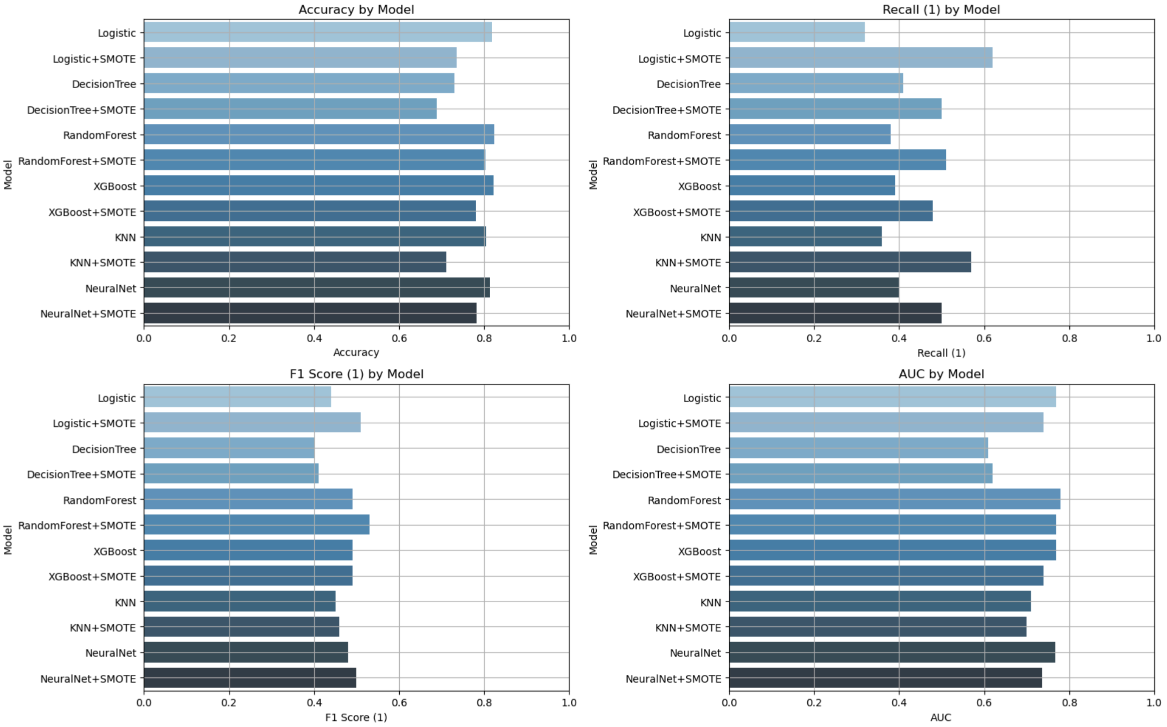


*Figure 17. Model Comparison Bar Chart*

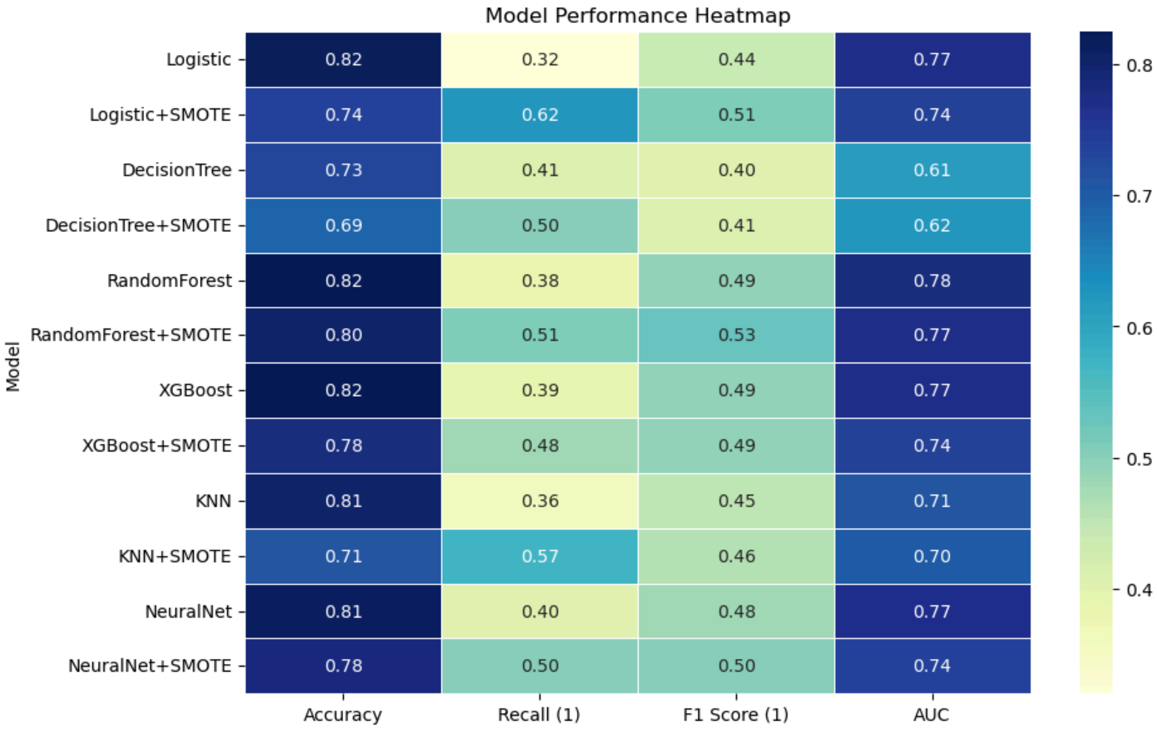


*Figure 18. Model Comparison HeatMaps*

To compare the effectiveness of different classification models, both with and without SMOTE, four key evaluation metrics were analyzed: Accuracy, Recall (Default class), F1-Score (Default class), and AUC. Figure 16 presents the comparison using bar plots and a heatmap for better visualization.

Overall, Random Forest + SMOTE and Logistic Regression + SMOTE emerged as strong performers. While Random Forest + SMOTE achieved the highest F1-score (0.53) and a strong AUC (0.77), Logistic Regression + SMOTE achieved the highest recall (0.62) for the default class, which is critical in credit risk detection.

In contrast, models without SMOTE generally demonstrated higher overall accuracy but significantly lower recall for the minority class. For instance, baseline Logistic Regression achieved 81.8% accuracy but only 32% recall, indicating a strong bias toward the majority class. This pattern was consistent across other non-resampled models.

Among non-linear models, XGBoost + SMOTE provided a good balance across all metrics (Recall: 48%, F1: 49%, AUC: 0.74), while KNN + SMOTE showed the highest recall improvement (from 36% to 57%), despite a notable drop in accuracy. The Neural Network + SMOTE also saw balanced improvements in both recall and F1-score, reinforcing the benefit of resampling.

These results confirm that applying SMOTE consistently improves the sensitivity of models to the minority class, with only marginal trade-offs in accuracy and AUC. The choice of the best model depends on whether overall classification accuracy or minority class recall is prioritized in deployment.

## 6. Feature Importance and Interpretability

### 6.1. Overview

To gain insights into model behaviour and ensure transparency in decision-making, I analyzed feature importance rankings from three top-performing models: Logistic Regression + SMOTE, Random Forest + SMOTE, and XGBoost + SMOTE. From the output, the three models consistently highlighted 6 features, which are: *LONGEST_DELAY, PAY_0, LIMIT_BAL, PAY_AMT2, CREDIT_LIMIT_RECENT_DELAY and NUM_DELAYS*. Among these 6 features, “LONGEST_DELAY”, “PAY_0” and “NUM_DELAYS” ranked the top 3 most important features across models. From this, we can analyze that all three models pointed to similar behavioural indicators, such as late payments and recent delays show that certain patterns are consistently and reliably signs of financial risk. This kind of consistency across models makes the results more trustworthy and easier to explain to both decision-makers and regulators. “LONGEST_DELAY” reflects the worst delay on record, “PAY_0” captures the most recent payment status, and “NUM_DELAYS” aggregates the borrower’s history of late payments. By prioritizing these behavioural indicators, financial institutions can better understand risk patterns and implement credit control strategies more efficiently.

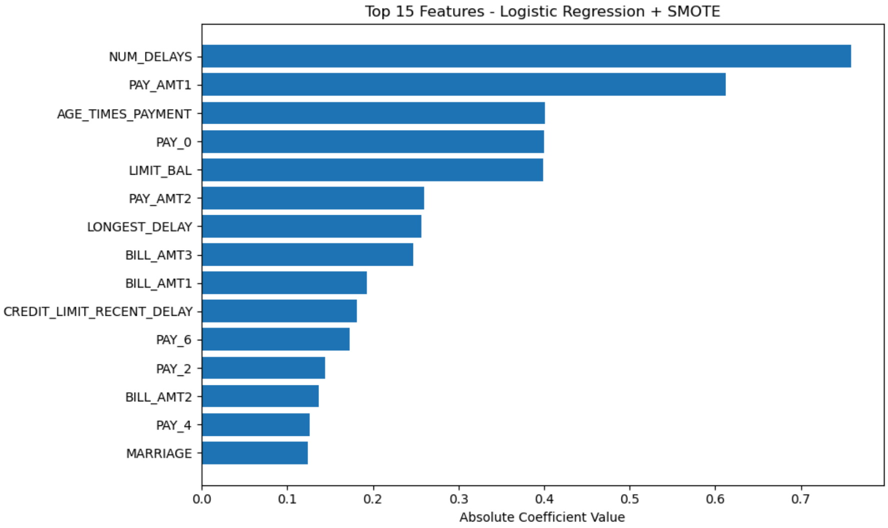


*Figure 19. Feature Importance - Logistic Regression + SMOTE*

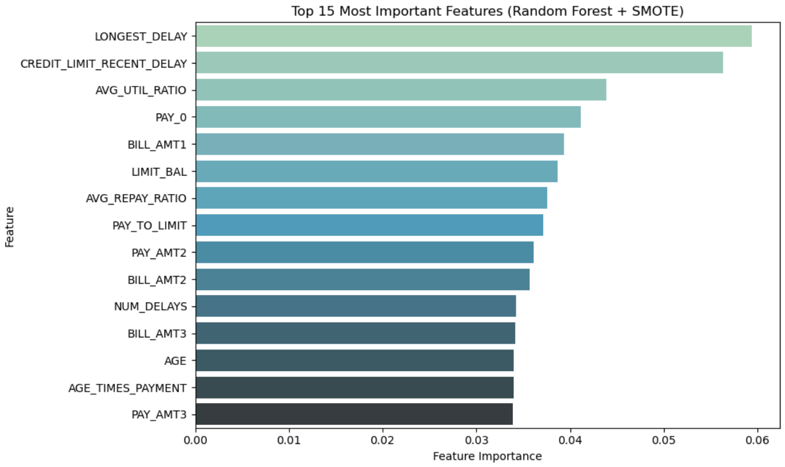


*Figure 20. Feature Importance - Random Forest + SMOTE*

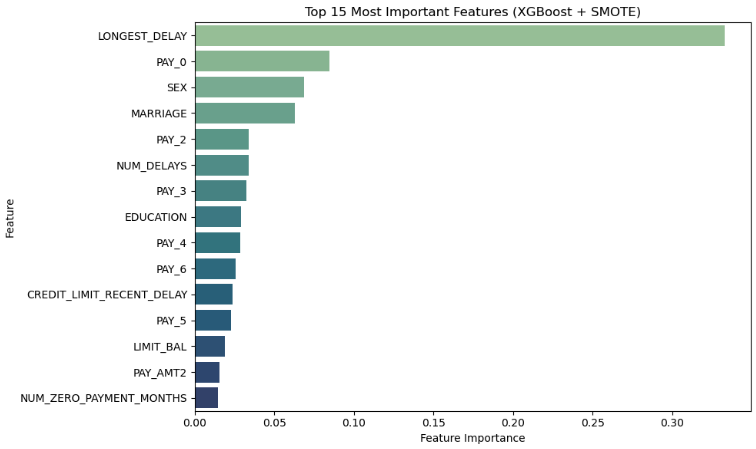


*Figure 21. Feature Importance - XGBoost + SMOTE*

# 7. Hyperparameter Tuning

## 7.1. Overview of Tuning Approach

To further enhance the model’s performance and address the trade-off between accuracy and recall, I conducted hyperparameter tuning using the “GridSearchCV” method within a pipeline structure. In this case, GridSearchCV performed a search over predefined hyperparameter grids using 5-fold cross-validation, which allows for robust model selection based on generalization performance. The scoring metric was set to recall, aiming to prioritize the correct identification of default cases. In this section, I performed tuning on two models: Random Forest and Logistic regression.

### 7.2. Tuned Random Forest

For Random Forest, I performed hyperparameter tuning using a pipeline that incorporates SMOTE, feature scaling and model training. A grid search with 5-fold cross-validation was applied to the training set, then it was balanced using SMOTE. From the output, the best parameter has “Max_depth” = 10, “Min_samples_split = 5”, “n_estimators = 100, and “Min_samples_leaf = 1”. From that, I got the tuned model’s performance with an accuracy of 80%, a recall rate of 58% and a F1-score of 55% for the default class. It reflects improved sensitivity in identifying defaulters compared to the untuned model. The result demonstrates that combining hyperparameter tuning with SMOTE oversampling techniques can significantly improve the model’s ability to detect defaulters.

### 7.3. Tuned Logistic Regression

For Logistic Regression, I performed the same approach with 5-fold cross-validation, optimizing for the AUC score. The best combination of hyperparameters was: “C = 0.01”, “Penalty = L2”, “solver = lbfgs”. This tuned model achieved an overall accuracy of 74%, improved the recall for the default class to 62% compared to the baseline model. The confusion matrix shows that more default cases were correctly identified after tuning. These results confirm that tuned logistic regression combined with SMOTE remains a strong model for credit risk classifications.

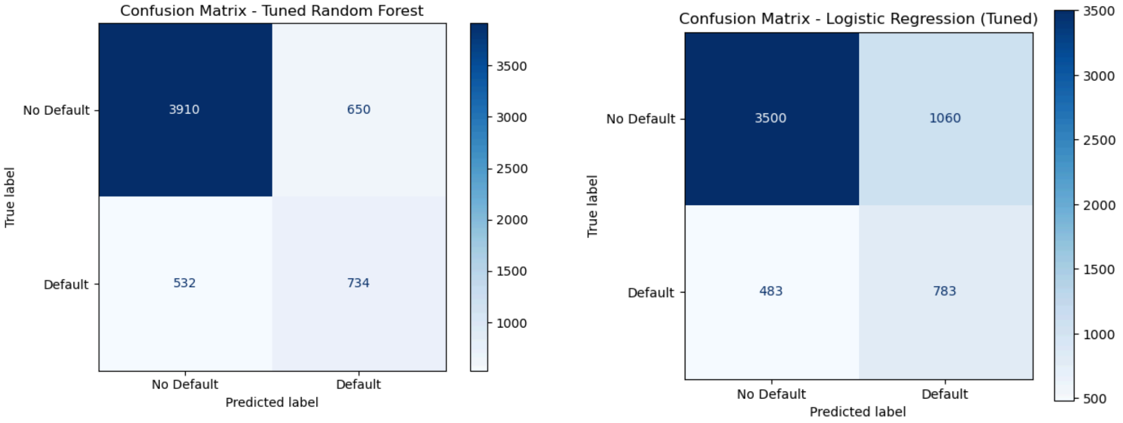


*Figure 22. Feature Importance - Random Forest + SMOTE*

## 8. Discussion and Business Implications

This paper demonstrates the practical value of using machine learning techniques with Oversampling techniques in credit risk modeling. Starting from a logistic regression baseline, we observed noticeable improvements in the recall rate when applying SMOTE, particularly in models such as Random Forest and XGBoost. Furthermore, when applying hyperparameter tuning, it further enhanced model's performance. The tuned Random Forest model achieved the best balance between accuracy and recall. The logistic regression demonstrated competitive recall and interpretability while sacrificed slightly in the accuracy. These findings all align well with Montevechi et al. (2024) in the literature review section where he also highlights the role of SMOTE techniques in improving the identification of defaulters.

From the business perspective, these findings suggest that financial institutions should not rely solely on traditional accuracy metrics when building credit risk model, instead, it should favour recall score and F1-score in order to ensure better detection on the potential defaulters, which will directly reduce financial loss for insitutions. By deploying the model to the real application systems, it can help lenders spot high-risk borrowers instantly by flagging those clients for review. From that, the model manages the loan portfolio more effective. Altogether, this approach leads to stronger risk control. Overall, the combination of advanced modeling techniques, class balancing strategies, and hyperparameter tuning approach in this paper provides a more reliable foundation for credit risk assessment. Financial institutions can leverage these insights to improve their lending policy and business strategies.

## 9. Conclusion

In conclusion, this study aims to improve credit risk prediction by comparing multiple machine learning models and examining the effect of class imbalance correction using SMOTE. By analyzing a publicly available credit card default dataset, we successfully examined six supervised machine learning models using key metrics including accuracy, recall, F1-score, and AUC. Among these, ensemble models like Random Forest and XGBoost achieved the highest overall accuracy and AUC, while Logistic Regression combined with SMOTE delivered the best recall and F1-score for the minority class. Further analysis on feature importance gives us the top three most influential predictors of credit default: “LONGEST_DELAY,” “PAY_0,” and “NUM_DELAYS”. Moreover, the findings highlight the importance of not only selecting robust algorithms but also addressing data imbalance and interpretability in financial applications. While this study offers meaningful insights, future work could expand on incorporating alternative resampling strategies and hybrid models to further enhance predictive power and stability.

## Bibliography

- Galindo, Jorge, and Pablo Tamayo. (PDF) credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications, February 2000. https://www.researchgate.net/publication/5144412_Credit_Risk_Assessment_Using_Statistical_and_Machine_Learning_Basic_Methodology_and_Risk_Modeling_Applications.
- Robinson, Kenneth J. "Savings and Loan Crisis." Federal Reserve History, November 22, 2013. https://www.federalreservehistory.org/essays/savings-and-loan-crisis.
- Table XV. Jorge Galindo and Pablo Tamayo, *Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications*, December 1997, table, Computational Economics (2000), 15, reproduced from page 28, https://doi.org/10.1023/A:1008699112516.
- Montevechi, André  Aoun, Rafael  de Carvalho Miranda, André  Luiz Medeiros, and José  Arnaldo Barra Montevechi. "Advancing Credit Risk Modelling with Machine Learning: A Comprehensive Review of the State-of-the-Art." Engineering Applications of Artificial Intelligence, August 3, 2024. https://www.sciencedirect.com/science/article/pii/S0952197624012405?via%3Dihub.
- Addo, Peter  Martey, Dominique Guegan, and Bertrand Hassani. "Credit Risk Analysis Using Machine and Deep Learning Models." MDPI, April 16, 2018. https://www.mdpi.com/2227-9091/6/2/38?amp=1.
- Yeh, I-Cheng. 2009. Default of Credit Card Clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H.