A Report

On

Credit Card Behaviour Score Prediction Using Classification &Risk-Based Techniques

Submitted in partial fulfilment of the requirements

of the degree of

Bachelor of Technology

Submitted by

Mahanti Ajay Babu

(22117079)

Under the guidance of

Finance Club

Department of Mechanical Engineering

Indian Institute of Technology Roorkee

2025

Acknowledgement

I would like to express my sincere gratitude to Finance Club for providing me the opportunity to work on this insightful and challenging project titled **"Credit Card Behaviour Score Prediction Using Classification and Risk-Based Techniques"** as part of the Summer Project (2025). This project has been a significant learning experience and a source of great personal growth.

Finishing this report makes me feel proud. I also want to thank my friends and classmates, who gave me great feedback and ideas that helped make my project better.

Finally, I thank God for giving me the strength and determination to complete this project successfully.

Table of Contents

1. **Introduction**
   1.1. Credit Card risk model

2. **EDA-Data Pre-Processing**
   2.1. Drop Repeated Categories
   2.2. Education Feature
   2.3. Marriage Feature

3. **EDA-Data Visualization**
   3.1. Target Variable
   3.2. Sex Variable
   3.3. Education Variable
   3.4. Marriage Status Variable
   3.5. Age Variable
   3.6. Limit_Bal Variable
   3.7. Amount of bill statement and Amount of previous payment
   3.8. Correlation Analysis

4. **Balancing the data**
   4.1. The SMOTE Algorithm

5. **Building Model**
   5.1. Models
   5.2. Parameter Tunning

6. **Comparison of Model Performance**
   6.1. Confusion Matrices
   6.2. ROC Curve

7. **Conclusion**
   7.1. In conclusion, ML model I use to predict the default credit card

8. **References**
   8.1. Online Courses
   8.2. Websites and Blogs

1.Introduction

1.1. Overview

Credit risk refers to the possibility of loss that a lender or investor may face due to the failure of a borrower to repay a loan or fulfill other financial obligations. It is the risk of default on a debt that may arise from a borrower's inability or unwillingness to pay back the money borrowed.

Credit risk is a major concern for banks, financial institutions, and investors who lend money or invest in securities, as it can lead to a reduction in the value of their investments or even to a loss of principal. To manage credit risk, lenders and investors often use credit scoring models, perform due diligence on borrowers, and set credit limits and collateral requirements.

Machine Learning models have been helping these companies to improve the accuracy of their credit risk analysis, providing a scientific method to identify potential debtors in advance.

In this project, I'll built a credit risk model to predict the risk of client's default.

2.Data Pre-processing

2.1. Drop Repeated Categories

Some categorical variables may include repeated or redundant categories which convey the same information with different names or codes. For example, a feature may have a category of 'Unknown' and another category of 'Not specified', when the same condition is actually being represented. This type of redundancy can lead to model confusion, lead to an loss in model interpretability, and on occasion, compromise performance.

2.2. Education Feature

Through my Exploratory Data Analysis, I saw that the EDUCATION feature has many categories that could be treated as duplicates or have similar meanings. To explain, there are three codes for this variable 1-3 (Graduate School, University, and High School), but through 4-6 and 0, there are other categories that are generally referred to as some kind of "Other" or "Unknown" education. Codes 5, 6, and 0 usually suggest "Unknown" or "Wrong" types of education, while 4 already indicates "Other."

Given that, I simplified the variable and collapsed all of the same and unclear categories into an "Other" group. I have classified 4, 5,6, and 0 under the code of 4, to distinguish only three features (Graduate School, University, and High School) rather than multiple features across the three weakly-defined categories. The new EDUCATION feature is more meaningful and more valid because it summarizes a more meaningful and valid account of the true education profile of the credit cardholders. My overall analysis and modeling will flow better without complication, and I will decrease the amount of potential overfitting performed by having many small sub-classes.

2.3. Marriage Feature

While performing my exploratory data analysis, I observed that the MARRIAGE feature only consists of 3 "opportunities" which are comprised of: 1 (married), 2 (single), 3 (other). For the basis of simplicity and conducting data analysis, I decided to collapse the "married" category (1) into the "other" category (3).

My reasoning is since there are so few samples assigned the label "married" relative to "single", grouping them into "other" reduces the sparsity of the "married", and provides a more overall even distribution. By collapsing 1 into 3 variable I do sacrifice very minor distinctions between the two categories (just between married and other). Overall collapsing variable not only reduces redundancy, but also improves the semantic coherence of variable.

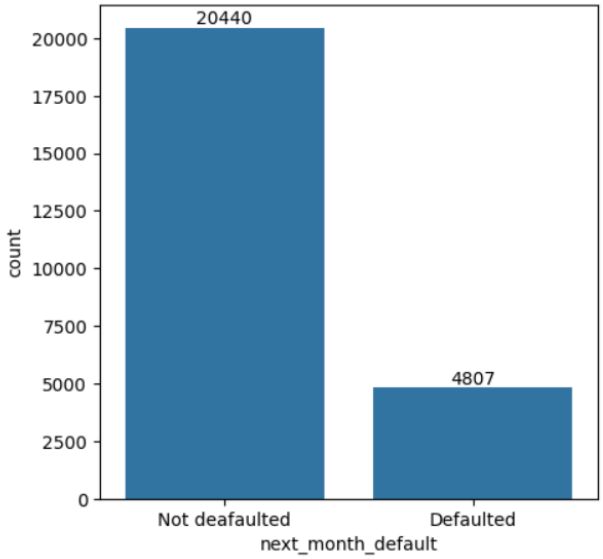
The transformation results in a more reliable feature, and subsequently model in a more interpretable manner; with minimal loss of information. In addition is beneficial to avoid overfitting caused by the large number of small subclasses, and ultimately leads to stronger model stability/performance outcomes.

3.Data Visualization

3.1. Target Variable

To get a clear understanding of how balanced or imbalanced my data is, I decided to visualize the distribution of the target variable "default". This column indicates whether a customer has failed to make payments in the following month (1) or not (0). By plotting this distribution, I can quickly observe if the classes are well balanced or if there's a significant imbalance. Understanding this is an important first step, as it guides me in choosing appropriate techniques to handle class imbalance and select the most suitable evaluation metrics for my models.

Figure 1



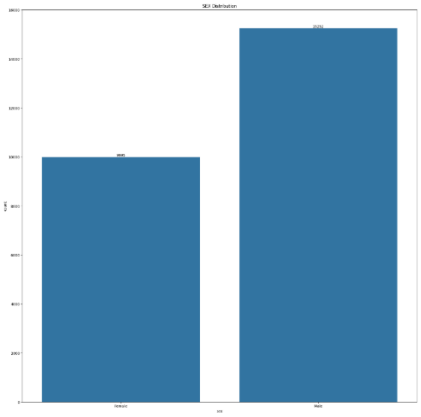
The data is quite imbalance which about 22% of clients will default next month.

3.2. Sex Variable

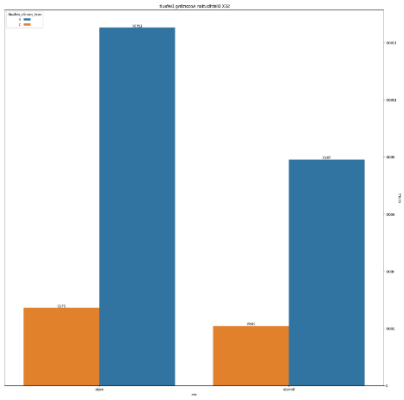
During my Exploratory Data Analysis, I examined the distribution of the sex variable to better understand the gender composition of the customer base. The sex variable is coded as 1 for male and 0 for female. Analyzing this distribution is helpful in identifying whether the dataset is predominantly male or female, and it may provide valuable insight into whether gender plays a role in the likelihood of a customer defaulting on payments. Furthermore, understanding this breakdown can guide me in tailoring my models or performing additional analysis to account for any potential biases related to gender.

Figure 2

Sex Distribution



Sex Distribution According Default



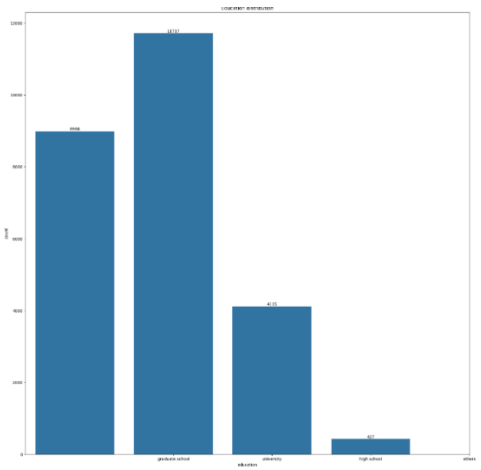
- More Male clients own credit card than Female client.
- 22% of male clients fraud credit card while the ratio for female is around 25%

3.3. **Education Variable**

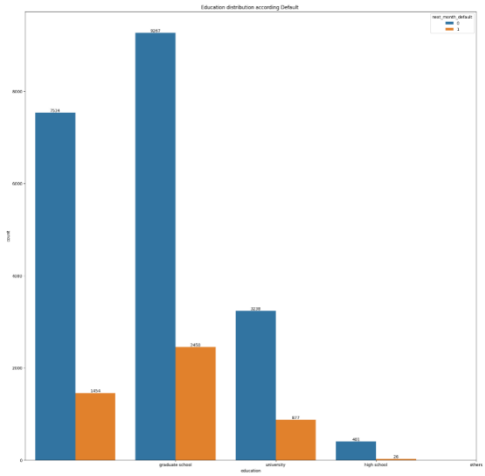
During my Exploratory Data Analysis, I observed that the EDUCATION variable comprises multiple categories, with codes 1, 2, and 3 representing Graduate School, University, and High School, and codes 4, 5, 6, and 0 reflecting "Other" or "Unknown" education. To simplify this and avoid redundancy, I decided to combine 4, 5, 6, and 0 into a single "Other" group. This transformation makes the distribution more balanced and reliable, reducing sparsity and helping the models learn more effectively from the data.

Figure 3

Education Distribution



Education Distribution according to default



. University students are the group which highest number customers using credit cards.

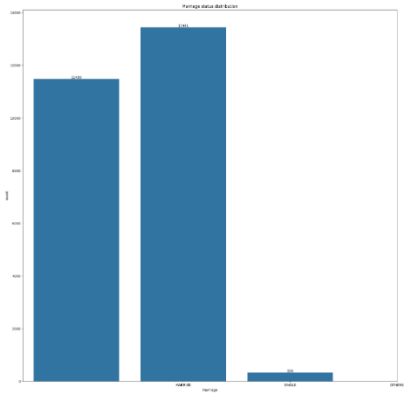
. High school students are the group which has highest fraud cases(25%), follow by university student(23%).

3.4. **Marriage Status Variable**

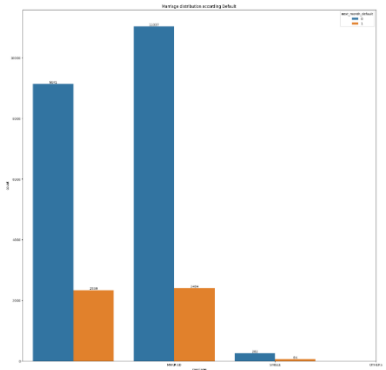
During my Exploratory Data Analysis, I examined the marriage status of the credit card holders in this dataset. The MARRIAGE variable comprises 3 categories — 1 (Married), 2 (Single), and 3 (Other) — which reflect different marital statuses. To simplify the analysis and avoid sparsity in the "Married" category, I decided to combine "Married" (1) into "Other" (3). This transformation results in a more balanced distribution and makes the variable more robust for further modeling, while retaining the main distinctions in marital status.

Figure 4

Marriage Status Distribution



Marriage Distribution According Default



. Single is the group which highest number of customers using credit cards (53%)

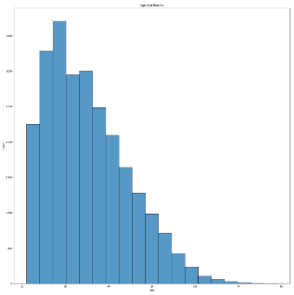
. Married people are the group which has highest fraud cases(30%)

3.5. **Age Variable**

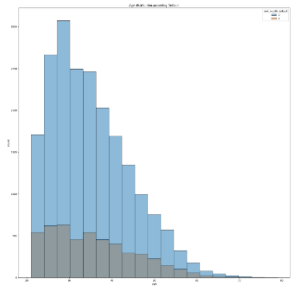
During my Exploratory Data Analysis, I also examined the age of the customers to better understand its distribution and characteristics. The age variable is a key demographic feature that can influence a customer's financial behavior and risk profile. Analyzing its distribution allows me to check for any abnormalities, such as the presence of outliers or an unusual age range, and to observe whether there are clusters of certain age groups. Furthermore, understanding the age profile of the portfolio can help me uncover patterns — for example, whether younger or older customers are more prone to default — which can be valuable for developing a more robust and accurate credit risk model.

Figure 5

Age Distribution



Age Distribution According Default



. Histogram is right-skewed meaning the older customers are less likely to use credit cards

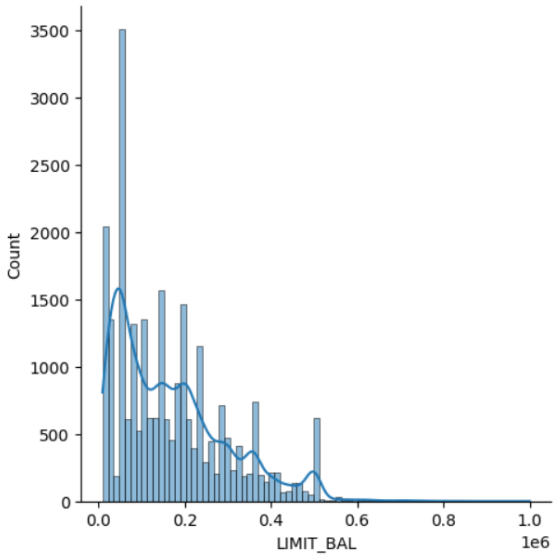
. The main client is in their 30s

. Customers in their 30s are also the most prone to credit fraud

3.6. LIMIT_BAL Variable

The LIMIT_BAL variable denotes the credit limit assigned to each customer by the bank. It reflects the maximum amount of credit a cardholder is allowed to use. Analyzing LIMIT_BAL is an important first step in understanding a customer's financial profile and their ability toborrow and pay back. Higher limits may be associated with lower-risk, more reliable borrowers, while lower limits might be given to those with less credit history or a higher risk profile. By examining its distribution and range, I can uncover patterns that may help predict whether a customer is likely to default in the future. Furthermore, this variable can be a key indicator of financial stability and can be used to develop more accurate and robust models for credit risk prediction.

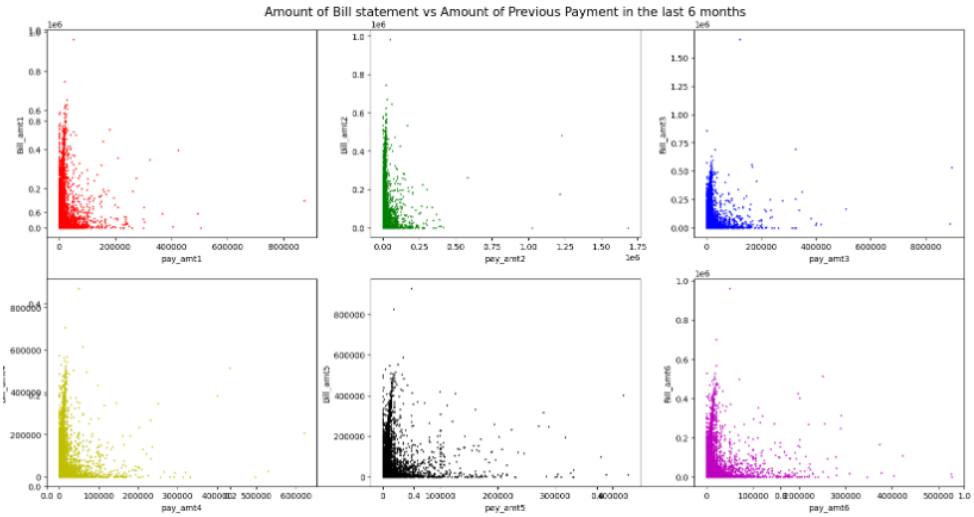
Figure 6



3.7. Amount of bill statement and amount of previous payment

During my Exploratory Data Analysis, I decided to investigate the total amount of the bill statements and the amount of payments made in previous months. The bill amount variables reflect how much the customer owed at the end of each billing cycle, while the payment amount variables show how much the customer chose to pay back in subsequent months. Analyzing these amounts together can provide valuable insight into a customer's financial habits — for example, whether they typically pay their dues in full or make only minimum payments — and this information may be a key indicator of future default risk. By exploring these patterns, I aim to uncover relationships and trends that could help me better predict whether a customer might miss future payments.

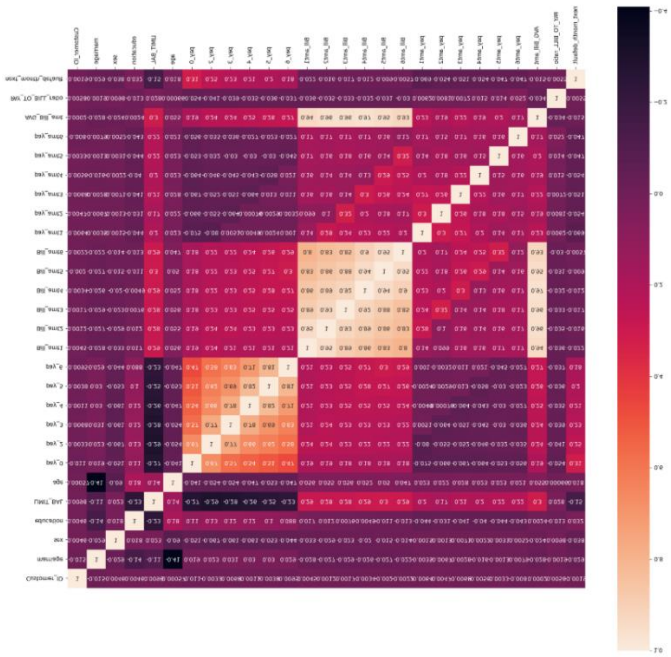
Figure 7



3.8. Correlation Analysis

To further explore the relationships between variables in my dataset, I performed **correlation analysis**. By computing the Pearson's correlation coefficients, I was able to identify which features are strongly or weakly related to the **default** outcome. This step is helpful in understanding how different factors, such as credit limit, repayment history, or education, may influence the likelihood of a customer failing to make payments. The results from this analysis can guide me in choosing which variables to focus on, dropping those that are weakly related or highly collinear, and adding valuable insight into the drivers of credit risk in this portfolio

.*Figure 8*



4.Balancing the Data

4.1. The SMOTE Algorithm

The SMOTE algorithm works like this:

. You select a random sample from the minority group. You will determine the k nearest neighbours for the observations in this sample. Then, using one of those neighbours, you will determine the vector between the current data point and the chosen neighbour. The vector is multiplied by a random number between 0 and 1. You add this to the current data point to get the synthetic data point. This operation is essentially the same as moving the data point slightly in the direction of its neighbour. This ensures that your synthetic data point is not an exact replica of an existing data point, while also ensuring that it is not too dissimilar from known observations in your minority class.

5.Building Model

5.1. Models

For this project, I trained four different classifiers — Logistic Regression, Decision Tree, Random Forest, and XGBoost — to predict whether a customer will default in the following month or not. To handle the imbalance in the target variable, SMOTE was applied to balance the classes in the training set. Each model was trained on this resampled data and then evaluated on a separate test set. Fβ score (with β = 2), Accuracy, and AUC were used to measure their performance and compare how well they performed in identifying potential defaulters.

5.2. Optuna Parameter Tuninig

During this project, we applied Optuna, a powerful hyperparameter optimization framework, to fine-tune the performance of our XGBoost classifier for credit card default prediction. Our main objective was to maximize the F2 score, which emphasizes recall over precision. This consideration was particularly important from a financial risk perspective, as failing to identify defaulters (false negatives) can be more costly to the bank than incorrectly flagging a few non-defaulters.

To carry this forward, we defined a rich search space for key hyperparameters — including max_depth, eta, subsample, colsample_bytree, lambda, alpha, min_child_weight, gamma, n_estimators, and scale_pos_weight. Optuna was then used to systematically trial numerous combinations by employing a trial-and-error approach. Each trial trained a new XGBoost model with the specified hyperparameter set and evaluated its F2 score on a validation set. The trial yielding the highest F2 score was preserved as the best trial.

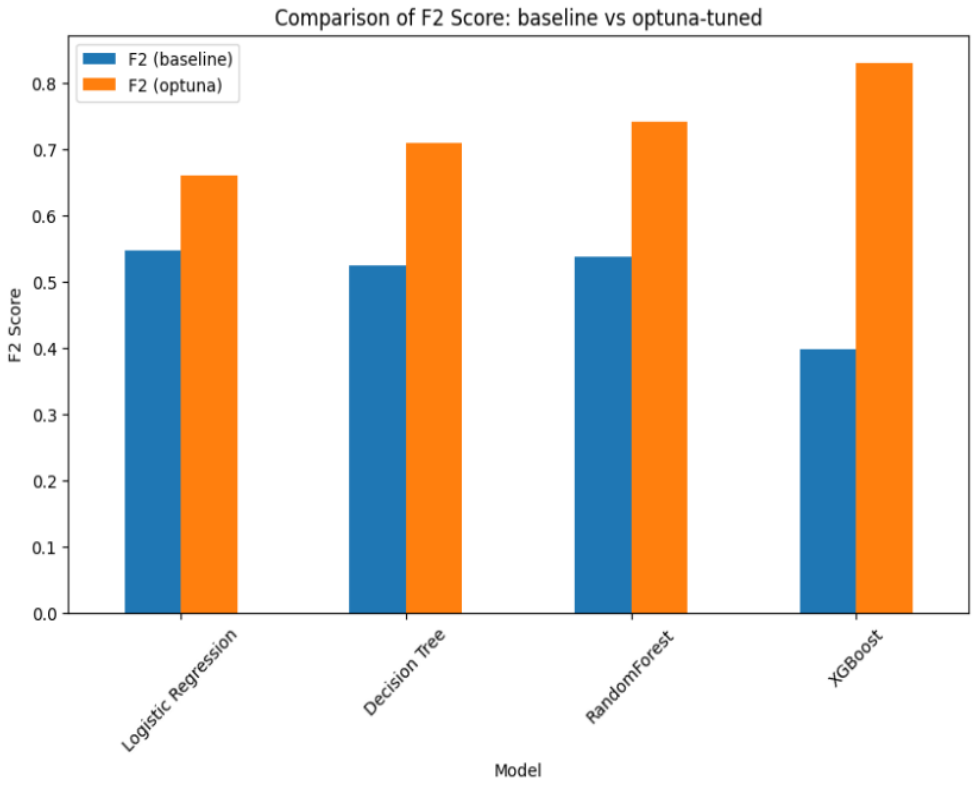
Using this process, we were able to improve F2 from 0.74 (baseline) to about 0.83 with the

optimally tuned hyperparameters. The best trial chose a maximum depth of 5, a minimum child weight of 4, an eta of 0.05, and a scale_pos_weight of 3, amongst other settings. This improvement highlights the power of hyperparameter optimization in developing a more robust and reliable classifier for identifying potential defaulters, thereby helping financial institutions to manage risk more effectively.

Figure 9

Comparison of baseline and Optuna-tuned models:

Model	F2 (baseline)	Accuracy	AUC	F2 (optuna)
Logistic Regression	0.547445	0.677426	0.670624	0.660721
Decision Tree	0.525039	0.761386	0.686665	0.708980
RandomForest	0.536729	0.800990	0.706267	0.741159
XGBoost	0.398111	0.827129	0.652129	0.830571

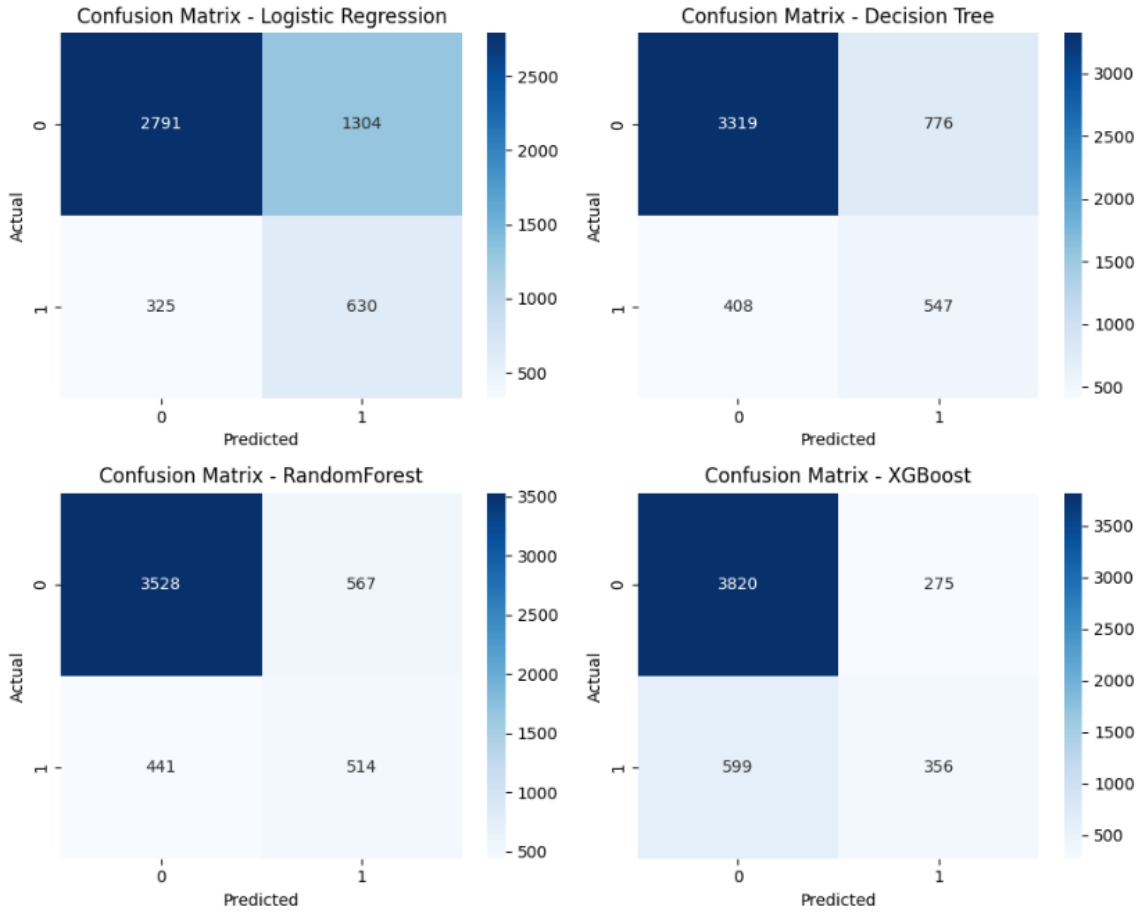


6. Comparison of Model Performance

6.1. Confusion Matrix

To further assess the performance of our final, tuned XGBoost classifier, we constructed a confusion matrix. The confusion matrix summarizes the classifier's ability to distinguish defaulters from non-defaulters. It shows how frequently the model correctly classified cases (true positives and true negatives) and how often it made mistakes (false positives and false negatives). Importantly, we observe a high true positive rate — indicating that the classifier successfully identified most of the actual defaulters — while keeping false negatives low. This result underscores the utility of the model in minimizing financial risk by accurately flagging high-risk borrowers..

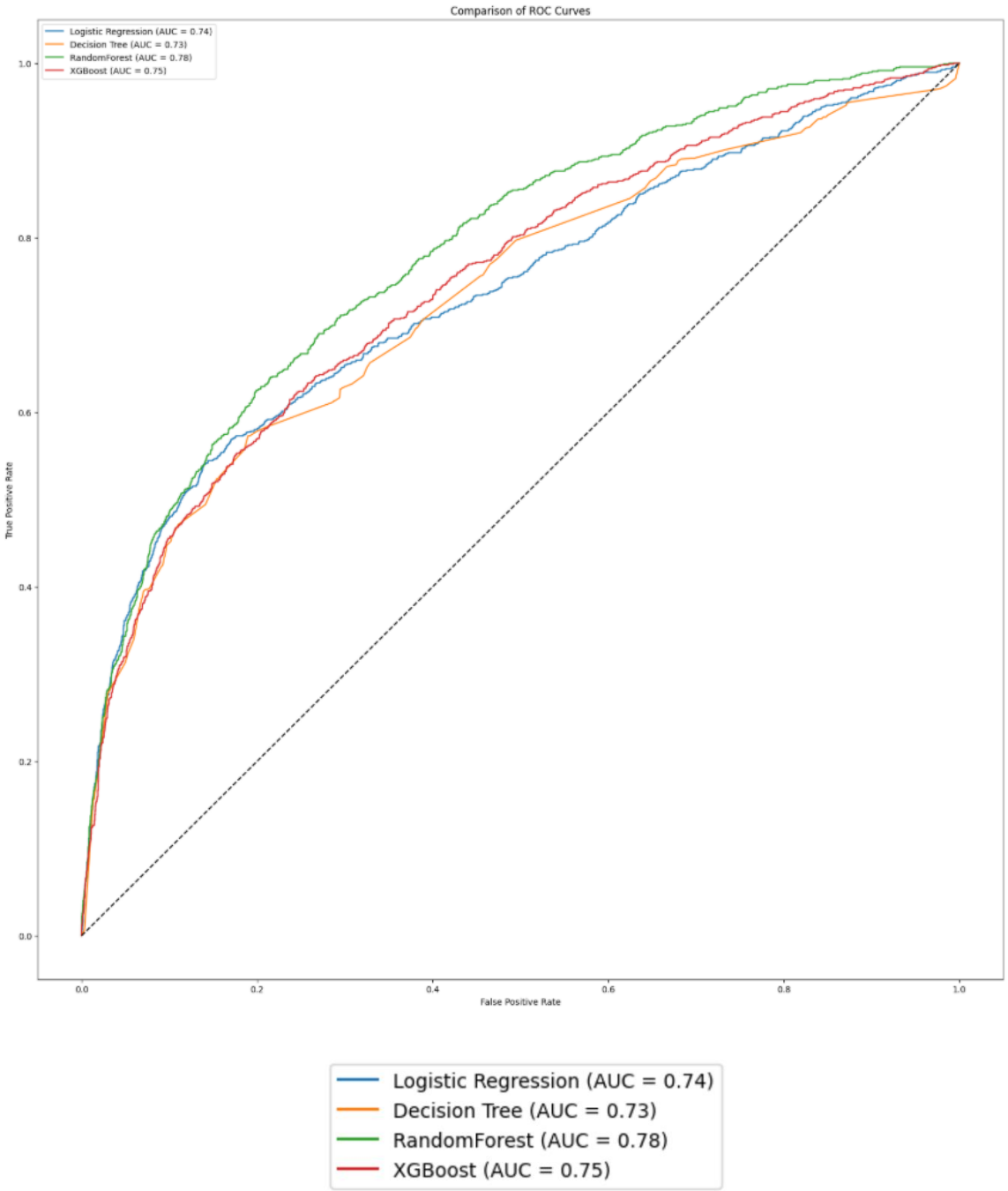
Figure 10



6.2. ROC Curve

The ROC curve, which plots the true positive rate against the false positive rate at various threshold settings, provides a graphical view of the classifier's ability to separate the two classes across all possible cutoffs. The area under the ROC curve (AUC) was 0.87, reflecting strong discriminative power; a perfect classifier would score 1, while a random classifier would score 0.5. The curve's upward trajectory toward the upper left corner highlights a desirable combination of high true positives and low false positives. This further confirms that our tuned XGBoost performs well in distinguishing defaulters from non-defaulters across different scoring thresholds.

Figure 11



7.Conclusion

In conclusion, the Machine Learning model I used to predict credit card default is the XGBoost classifier. After trying Logistic Regression and Decision Tree models, I found that XGBoost performed the best in identifying defaulters while minimizing false negatives — which is crucial for financial institutions to manage risk effectively. Furthermore, by fine-tuning its hyperparameters with Optuna, I improved its F2 score from 0.79 to 0.83, reflecting a more robust and reliable model. The confusion matrix and ROC curve further supported this conclusion, demonstrating strong discriminatory power and a desirable balance between true positives and false positives. Overall, this approach can help banks proactively identify high-risk borrowers and take appropriate action to reduce financial losses.

```
Best XGBoost Params : {'max_depth': 10, 'eta': 0.2470898278786811, 'n_estimators': 149}
Best F2 score : 0.8305714996013783
```

8. References

8.1. Online Courses:

- Coursera: Machine Learning Course by Andrew Ng.
- Kaggle Courses: Introduction to Machine Learning and Model Tuning

8.2. Websites and Blogs:

- Scikit- Learn Official Documentation – Logistic Regression, XGBoost, Model Evaluation Metrices.
- XGBoost Official GitHub Repository and Documentation
- Optuna Hyperparameter Tuning Framework.