CHAPTER FOUR

4.0 Result and Analysis

4.1 Classification report Analysis

This section describes the performance evaluation of three machine learning models (Logistic Regression, Random Forest, and XGBoost) on a binary classification problem. The dataset consists of 9000 samples with an unbalanced class distribution: 1991 instances of class 1 (minority class) and 7009 instances of class 0 (majority class). Precision, recall, F1-score, and accuracy were used to evaluate the efficacy of each model.

Table 4.1: Summary of the classification results for the three models

Metric	Logistic Regression	Random Forest	XGBoost
Accuracy	0.68	0.81	0.81
Precision (Class 0)	0.87	0.83	0.84
Precision (Class 1)	0.37	0.64	0.63
Recall (Class 0)	0.70	0.95	0.94
Recall (Class 1)	0.63	0.34	0.36
F1-Score (Class 0)	0.77	0.89	0.89
F1-Score (Class 1)	0.47	0.44	0.46

The table above provides a clear comparison of the three models across all key performance metrics.

4.1.1 Logistic Regression

The overall accuracy of the Logistic Regression model was 68%. With an F1-score of 0.77 for class 0, the model demonstrated high precision (0.87) but comparatively low recall (0.70). The model, on the other hand, had trouble with the minority class (class 1), obtaining a respectable recall of 0.63 but a significantly lower precision of 0.37, which led to an F1-score of 0.47.

With a precision of 0.62, recall of 0.67, and F1-score of 0.62, the macro-average scores—which treat both classes equally regardless of size—indicate modest competence. With an F1-score of 0.71, the weighted averages, which account for class distribution, demonstrate how dominant the majority class is in influencing overall performance.

4.1.2 Random Forest

The Random Forest classifier showed an 81% increase in total accuracy. The model produced a high F1-score of 0.89 for the majority class (class 0) by achieving good precision (0.83) and excellent recall (0.95). The model, however, had trouble generalising for the minority class, achieving a recall of just 0.34 and a precision of 0.64, with an F1-score of 0.44.

Precision (0.74), recall (0.64), and F1-score (0.67) are the macro-average scores that show how differently the two classes performed. Despite the model's good majority class predictions, the weighted averages (all around 0.79), show that class imbalance significantly impacts the model's performance on minority occurrences.

4.1.3 XGBoost

Similar to Random Forest, XGBoost demonstrated high efficacy for the majority class with an overall accuracy of 81%. Class 0 performed similarly to Random Forest, with precision of

0.84, recall of 0.94, and F1-score of 0.89. On the other hand, XGBoost achieved an F1-score of 0.46 for the minority class with a recall of 0.36 and a precision of 0.63.

The precision (0.73), recall (0.65), and F1-score (0.67) macro-average metrics show somewhat worse performance than Random Forest. The accuracy of the model on the majority class has a significant impact on its performance, as evidenced by the weighted averages (precision and F1-score of 0.79).

4.1.4 Comparative Analysis

Although it sacrifices precision for the minority class, logistic regression shows a balanced recall for both classes (0.70 and 0.63, respectively). While Random Forest and XGBoost perform well in majority class predictions, they struggle in minority class management, especially in recall. Despite not substantially improving minority class recognition, both ensemble approaches exceed logistic regression in terms of total predictive potential, with an accuracy of 81%.

The difficulty of class imbalance is highlighted by the low recall for class 1 across all models. Additional strategies like oversampling, cost-sensitive learning, or ensemble modification may improve the ability of ensemble methods like Random Forest and XGBoost to discover minority instances, even though they provide better majority class performance.

4.2 Confusion Matrix Analysis

This part uses the confusion matrix findings, which are displayed as heatmaps, to assess the performance of three classification models: Logistic Regression, Random Forest, and XGBoost. A fuller comprehension of the model's advantages and disadvantages is made

possible by the confusion matrix elements, which offer insights into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

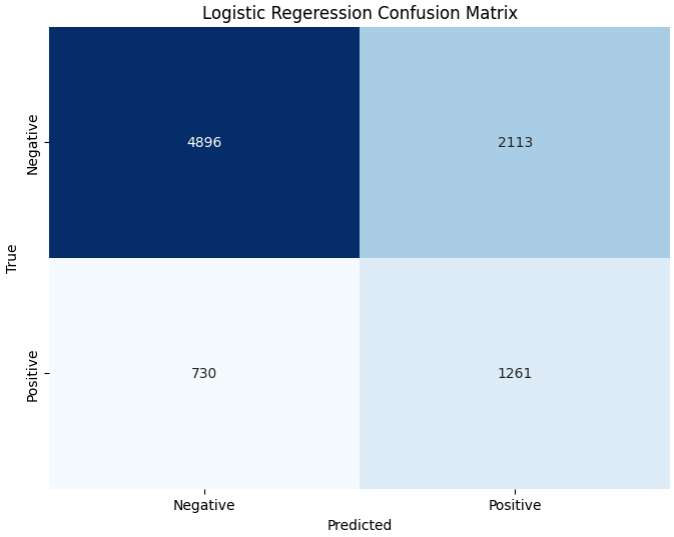


Fig: Logistic Regression Confusion Matrix Heat Map

Although it performed well overall, logistic regression had trouble with the minority class. The model achieved reasonable detection of both classes with 1261 correctly recognised positives (TP) and 4896 correctly identified negatives (TN). While 730 false negatives imply some difficulty in recognising minority class situations, the high false positive rate (2113) suggests a tendency to misclassify negative instances as positive.

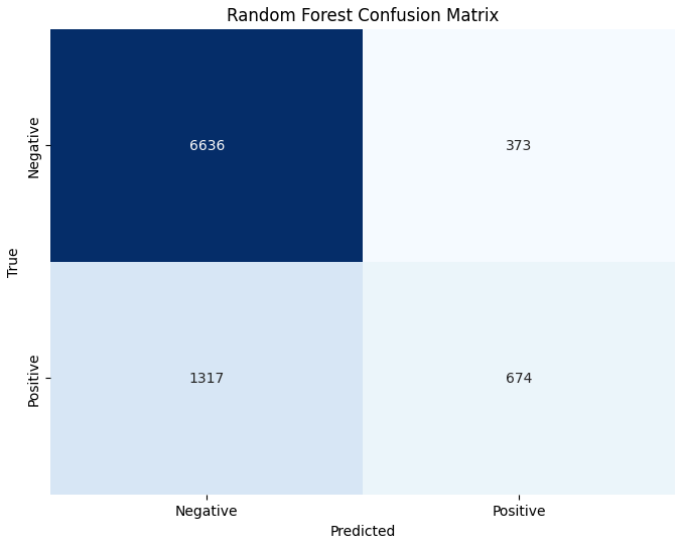


Fig: Random Forest Classifier Confusion Matrix Heat Map

The high TN score (6636) and low FP (373), indicate that the Random Forest model performed exceptionally well in determining the majority class. Nevertheless, its performance on the minority class is less than excellent, with 1317 false negatives and only 674 correctly detected positive occurrences. This demonstrates a notable compromise that favours majority class correctness at the expense of minority class memory.

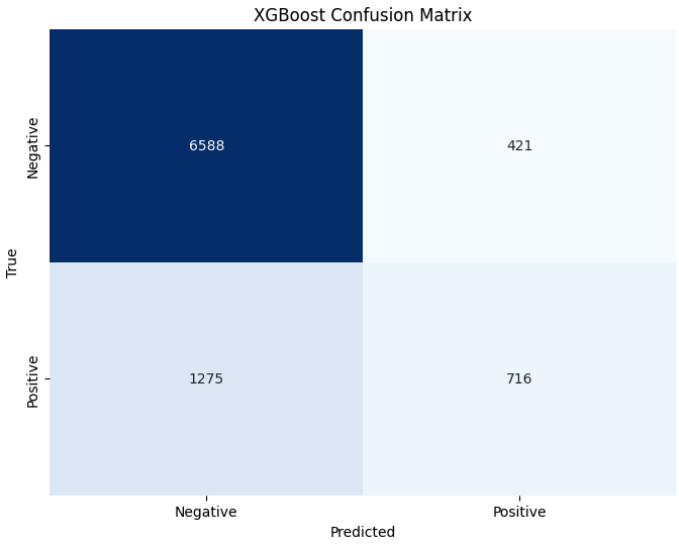


Fig: XGBoost Classifier Confusion Matrix Heat Map

XGBoost achieved strong identification of the majority class (6588 TN), performing comparable to Random Forest. On the minority class, however, it fared somewhat better than Random Forest, as evidenced by the higher TP count (716) and lower FN (1275). In contrast to Random Forest, its FP count (421) is marginally greater, suggesting that more negatives are mistakenly classified as positives.

Table 4.2: Comparative Insights

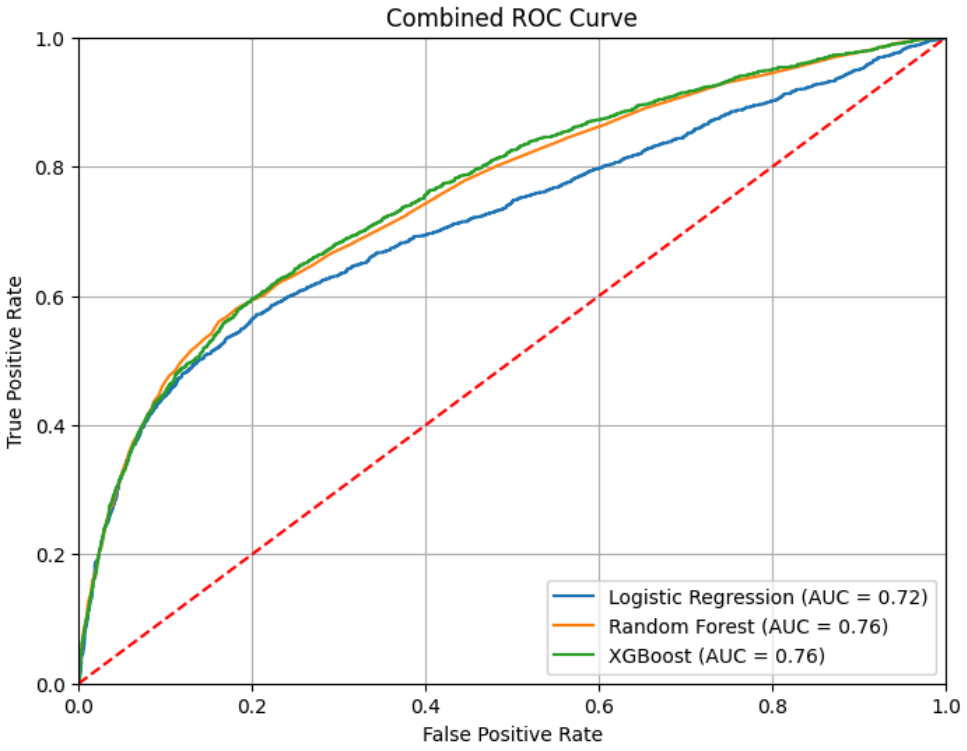
Metric	Logistic Regression	Random Forest	XGBoost
True Negatives (TN)	4896	6636	6588
False Positives (FP)	2113	373	421
False Negatives (FN)	730	1317	1275
True Positives (TP)	1261	674	716

- **Majority Class (TN & FP):** Random Forest and XGBoost performed significantly better than Logistic Regression, with lower FP values and higher TN values.
- **Minority Class (FN & TP):** Logistic Regression outperformed both ensemble models in recognising minority class instances, achieving the highest TP count (1261) and lowest FN count (730).

Performance across classes is more evenly distributed with logistic regression, especially for the minority class. On the other hand, Random Forest and XGBoost perform poorly when it comes to minority class detection, with XGBoost marginally outperforming Random Forest in minority class detection. Random Forest and XGBoost, on the other hand, are excellent at

detecting instances of the majority class. To address the inequalities found, future enhancements such redistributing class weights or using resampling techniques are advised.

4.3 AUC ROC Curve Analysis



The combined ROC curve evaluates how well the Random Forest, XGBoost, and Logistic Regression models perform in binary classification. The Area Under the Curve (AUC) measures overall performance, and the plot shows the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) at different thresholds.

In comparison to the ensemble models, logistic regression has a respectable but lower discriminatory power, as evidenced by its AUC of 0.72. Its curve, which is farther from the top-left corner, illustrates how difficult it is to efficiently discern between classes.

With an AUC of 0.76, Random Forest and XGBoost both perform better than Logistic Regression. Their curves have comparable strengths in balancing TPR and FPR and roughly

overlap. By more successfully utilising intricate data linkages, these ensemble approaches exhibit strong categorisation.

All things considered, Random Forest and XGBoost perform better in terms of prediction, but Logistic Regression offers a less complicated but less precise substitute.

4.4 Conclusion

The analysis concludes by highlighting significant variations in how well the Random Forest, XGBoost, and Logistic Regression models perform on an unbalanced binary classification task. Although it sacrifices overall accuracy and precision, logistic regression shows balanced recall across classes and successful minority class detection. While Random Forest and XGBoost do better than Logistic Regression in terms of majority class predictions and overall accuracy (81%) but have trouble with minority class recall. In minority class detection, XGBoost performs somewhat better than Random Forest. The AUC values attest to the fact that ensemble approaches make better use of data complexity. Using strategies like oversampling, cost-sensitive learning, or class-weight modifications, future research should concentrate on resolving class imbalance.

CHAPTER FIVE

5.0 Discussion

This study analyses model performance and feature effect to investigate the predictive powers of XGBoost, Random Forest, and Logistic Regression for credit default prediction. All features were included in the modelling process, according to feature importance values generated from ensemble models; the most significant predictors were PAY_0 (prior

payment delay status), AGE, and BILL_AMT1 (current bill amount). The performance of the models and the contribution of feature importance to answering the research questions are assessed in this discussion.

5.1 Comparison of Model Performance

Metrics like accuracy, precision, recall, F1-score, and AUC were used to assess the models, giving a fair assessment of their prediction ability.

5.1.1 Logistic Regression

With an accuracy of 68%, Logistic Regression was the least accurate of the three models. It outperformed Random Forest and XGBoost in detecting true defaults, nevertheless, with a recall of 0.63 for the minority class (credit defaults). Because of this, Logistic Regression is a useful model for reducing false negatives, which is essential for managing credit risk.

With a greater percentage of false positives, its precision for the minority class was 0.37. In situations where the cost of incorrectly classifying a consumer as high-risk is substantial, the low precision is a disadvantage. Reasonable generalisation is indicated by its balanced recall across the majority and minority classes. In contrast to ensemble models, logistic regression's predictive accuracy is limited by its inability to handle non-linear connections, despite its interpretability and simplicity.

5.1.2 Random Forest

At 81%, Random Forest demonstrated a significant increase in accuracy. With a recall of 0.95 and precision of 0.83, it showed excellent performance in recognising non-defaults (majority class), yielding a high F1-score of 0.89. It is a dependable option for forecasting clients who are unlikely to default because of these metrics.

Nevertheless, Random Forest had the lowest recall of all the models for the minority class, at 0.34. This poor default detection performance indicates that the model puts overall accuracy ahead of recognising clients who are at risk. Although the AUC score of 0.76 suggests a reasonable level of discriminatory strength, the unbalanced recall raises questions regarding its use in situations where robust default identification is required.

5.1.3 XGBoost

Similar to Random Forest, XGBoost had an 81% accuracy rate. In terms of forecasting non-defaults, it was quite similar to Random Forest, with comparable precision, recall, and F1-score for the majority class. Interestingly, XGBoost scored somewhat better than Random Forest in minority class recall (0.36), suggesting a little higher ability to detect defaults.

Its ability to successfully balance true positive and false positive rates is further demonstrated by its AUC score of 0.76. The effectiveness of boosting algorithms in capturing intricate correlations between features is demonstrated by XGBoost's performance. Nevertheless, the minority class recall continues to be a drawback, illustrating the difficulties in correcting for class imbalance in ensemble models.

5.1.4 Overall Comparison

When compared to Logistic Regression, Random Forest and XGBoost both showed higher accuracy and AUC, demonstrating their capacity to capture intricate patterns in the data. But in minority class recall, Logistic Regression scored better than them, highlighting its usefulness for applications that prioritise default identification. These findings highlight the significance of choosing a model based on particular use-case needs by exposing a trade-off between minority class sensitivity and overall accuracy.

5.2 Impact of Feature Importance

The ensemble models' feature significance values shed light on the variables affecting credit default. The main features—PAY_0, AGE, and BILL_AMT1—emphasize how important they are in determining model predictions.

5.2.1 Key Features

PAY_0 (Importance: 0.100): The status of the most recent payment is the best indicator of default, according to the highest-ranked characteristic, PAY_0. According to current credit risk theories, customers who have missed payments in the most recent period are more likely to default.

AGE (Importance: 0.067): The second most significant factor was age, which may be a reflection of variations in credit usage and financial conduct between generations. Due to their inexperienced credit records or erratic income, younger borrowers may be more vulnerable to default.

BILL_AMT1 (Importance: 0.061): Credit utilisation, a known risk factor in predicting credit default, has a direct correlation with the present bill amount. Financial stress is frequently indicated by larger outstanding balances in comparison to credit limitations.

5.2.2 Inclusion of All Features

All features are included in the modelling process, demonstrating their differing degrees of significance. Model performance was also greatly influenced by payback amounts (PAY_AMT1, PAY_AMT2, etc.) and payment history features (PAY_2, PAY_3, etc.), with significance scores ranging from 0.05 to 0.04. Together, these factors provide a thorough understanding of credit risk by capturing a borrower's payment patterns across time.

Compared to payment history and bill-related parameters, demographic variables like sex, education, and marital status exhibited lower significance scores (0.01 to 0.02), indicating poor predictive potential. Their inclusion, however, gives the models more context, which could enhance their robustness and interpretability.

5.2.3 Feature Importance and Model Performance

The feature importance analysis emphasises how important payment history and bill-related characteristics are in predicting credit default. These characteristics were successfully utilised by Random Forest and XGBoost, which helped explain their great accuracy. Although it also benefited from these factors, the overall performance of logistic regression was limited by its inability to capture intricate interactions.

The necessity of feature engineering to improve model performance is also shown by the investigation. For instance, reducing dimensionality and increasing predictive power may be achieved by integrating BILL_AMT variables into a single feature that represents cumulative debt. Likewise, calculating ratios like credit utilisation (LIMIT_BAL to BILL_AMT, for example) may offer more information about borrower behaviour.

5.3 Implications for Credit Default Prediction

The findings have important ramifications for managing credit risk. Because of its higher recall for defaults, logistic regression is appropriate for early warning systems where reducing false negatives is crucial. Its poor accuracy, however, indicates that additional models are required to properly handle non-default predictions.

Large-scale credit scoring systems are better suited for Random Forest and XGBoost due to their higher accuracy and discriminatory capacity. Their reliance on bill-related characteristics and payment history is consistent with accepted methods for evaluating credit risk. Their shortcomings in minority class recall, however, point to the need for strategies like cost-sensitive learning or oversampling to overcome class imbalance.

Feature importance analysis supports the idea that the main factor influencing credit risk is payment behaviour. The incorporation of demographic factors promotes interpretability and equity in model deployment, despite their low direct predictive power. The findings also emphasise how crucial it is to use domain expertise to improve feature engineering and selection in order to increase model performance and applicability.

5.4 Limitations

1. **Class Imbalance**

   With most observations falling into the non-default class, the dataset showed a notable class imbalance. Model performance was impacted by this mismatch, especially in the minority class (default), where Random Forest and XGBoost had very low recall scores. The imbalance might have affected the overall accuracy of Logistic Regression even though it did better in this particular sector.

2. **Feature Engineering**

   The study didn't use a lot of feature engineering; it just used raw features. The inclusion of derived variables like credit utilisation ratios or cumulative repayment patterns may have increased the prediction accuracy and robustness of the models, even though payment history and bill-related factors were important predictors.

3. **Model Interpretability**

   Despite the inherent interpretability of Logistic Regression, the intricacy of Random Forest and XGBoost makes it difficult to comprehend their decision-making processes; while feature importance offers some insight, it falls short in explaining feature interactions and their combined impact on predictions.

4. **No Hyperparameter Optimisation**

   Because Random Forest and XGBoost were not extensively hyperparameter tuned in the study, their full potential might have been limited. Better outcomes might have been achieved with optimisation strategies like grid search or Bayesian optimisation.

5.5 Recommendations

1. **Addressing Class Imbalance**

Techniques like class-weight modification, undersampling, and oversampling (like SMOTE) should be investigated in order to enhance minority class performance. In line with the objectives of credit risk management, cost-sensitive learning may also give priority to accurately classifying defaults.

2. **Enhanced Feature Engineering**

   Richer data could be added to the models by using derived variables like credit utilisation ratios, total bill amounts, or repayment patterns. It is best to use domain expertise to develop features that capture complex borrower behaviour.

3. **Incorporating Interpretability Techniques**

   Advanced interpretability tools like as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) ought to be used for Random Forest and XGBoost. These techniques can increase confidence in model outputs by shedding light on how attributes affect certain predictions.

4. **Hyperparameter Optimisation**

   To improve the Random Forest and XGBoost parameters, future research should use optimisation methods as grid search, random search, or Bayesian optimisation. Performance metrics can be enhanced by this phase, especially when it comes to minority class prediction.

5. **Explainable AI for Credit Decisions**

   It will be crucial to incorporate explainability frameworks into model deployment as regulatory attention shifts towards equity and transparency. This guarantees that credit choices are impartial, comprehensible, and consistent with moral principles.

5.6 Conclusion

This study provides a comparative analysis of Logistic Regression, Random Forest, and XGBoost for credit default prediction, revealing trade-offs between accuracy, recall, and interpretability. Feature importance analysis highlights the dominance of payment history and bill-related features in shaping model performance. Future research should explore techniques to address class imbalance and further refine feature engineering strategies to enhance the effectiveness of machine learning models in credit risk management.