Predicting Credit Card Defaults Using Supervised Machine Learning

By Kieran Dhooper

Executive Summary

Financial institutions face significant risks when issuing credit cards, as some customers may default on their payments. By leveraging supervised machine learning, these risks can be mitigated through predictive modelling based on customer characteristics and credit history.

A 22% class imbalance in the target variable made it difficult to predict defaulters, so SMOTE was applied to balance the dataset.

The project identified XGBoost as the best-performing algorithm, with an optimal combination of F1 and recall metrics of 0.53 and 0.56, respectively. The feature with the most significant impact on the model was repayment status in September (PAY_1).

The chosen model identified 575 more defaulters than the Dummy Classifier and demonstrated strong generalisation with a Receiver Operating Characteristic (ROC) score of 0.78.

Although the model performed well, I recommend further exploration of feature engineering and dedicating time to tuning a Support Vector Machine (SVM) model, as an untuned SMOTE-based SVM model performed nearly as well as the best model.

Data Understanding

Source: Default of Credit Card Clients Dataset (Yeh, 2009)

Data Overview:

- Credit card clients in Taiwan between April – September 2005.
- Dataset size is 30,000 rows and 25 columns.
    - Columns include information on default payments, demographic factors, credit data, payment history, and bill statements.
- Target feature is the default status next month (‘defaulter’).

Data Cleaning/Pre-processing

Data Cleaning:

- Merged 'Education' and 'Marriage' integer categories into a single 'Other' category to prevent confusion during model training.
- No null values or out-of-range categories.
- Removed ID column to prevent unintentional relationships.

Data Pre-processing:

- Applied SMOTE to address class imbalance (22% defaulters).
- Stratified train-test split to maintain class distribution, with 'defaulter' as the target feature.
- Standard Scaler used for normalising continuous, non-categorical features.
- No encoder transformations necessary as categorical features were in integer format.

Model Selection:

- Best model chosen by a combination of F1-score and recall, with an optimal precision. This is to reduce the financial loss of not predicting defaulters but not sacrifice the potential customers by neglecting precision.
- Accuracy is not a reliable metric on an imbalanced dataset, so it’s not widely considered.

EDA Insights

Graph 1 (Data Imbalance):

- A 78% class imbalance skews the dataset toward non-defaulters, indicating that credit score evaluations fail to catch the 22% of actual defaulters.

Graph 2 (Feature Correlation):

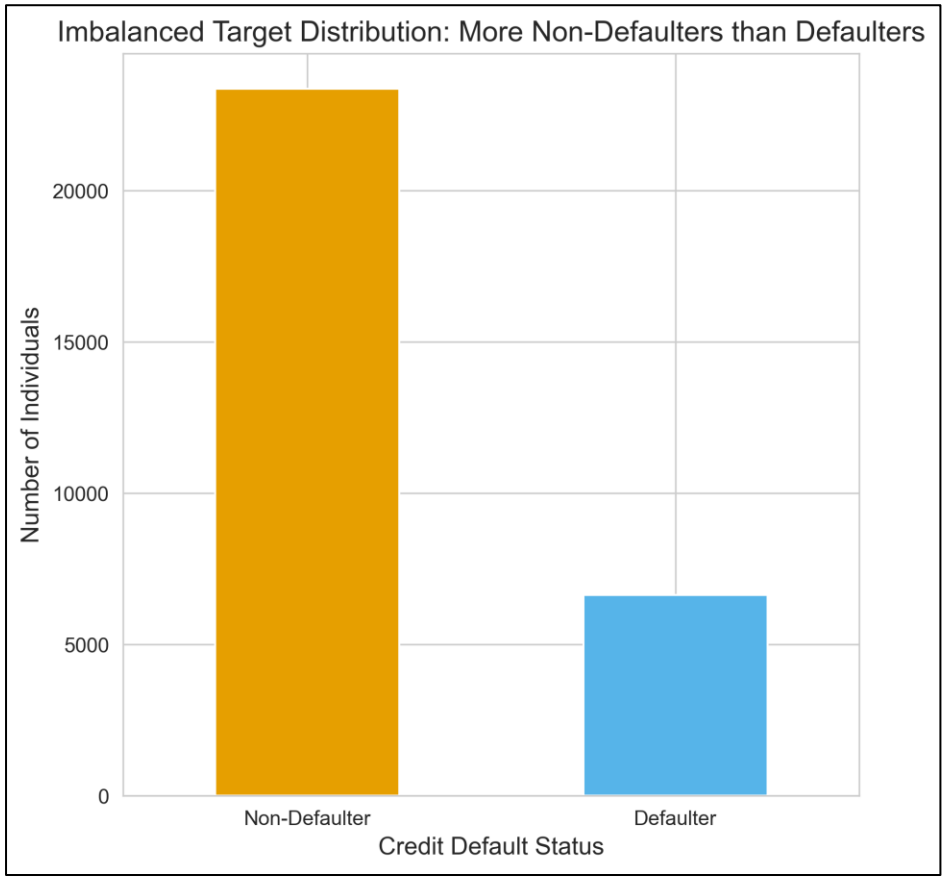
- Repayment status (PAY_1) in September has a 0.32 correlation with default status, making it the strongest predictor of future defaults.
- The April bill amount has a -0.01 correlation with default, showing that past balances do not predict future defaults.
- Repayment status correlation increases by 68% from April to September, confirming that recent payment behaviour is more predictive than past payments.
- Demographic features (e.g., age, gender, marital status) show low correlation with default, adding little predictive value.

Graph 3 (Balance Limit Box Plot):

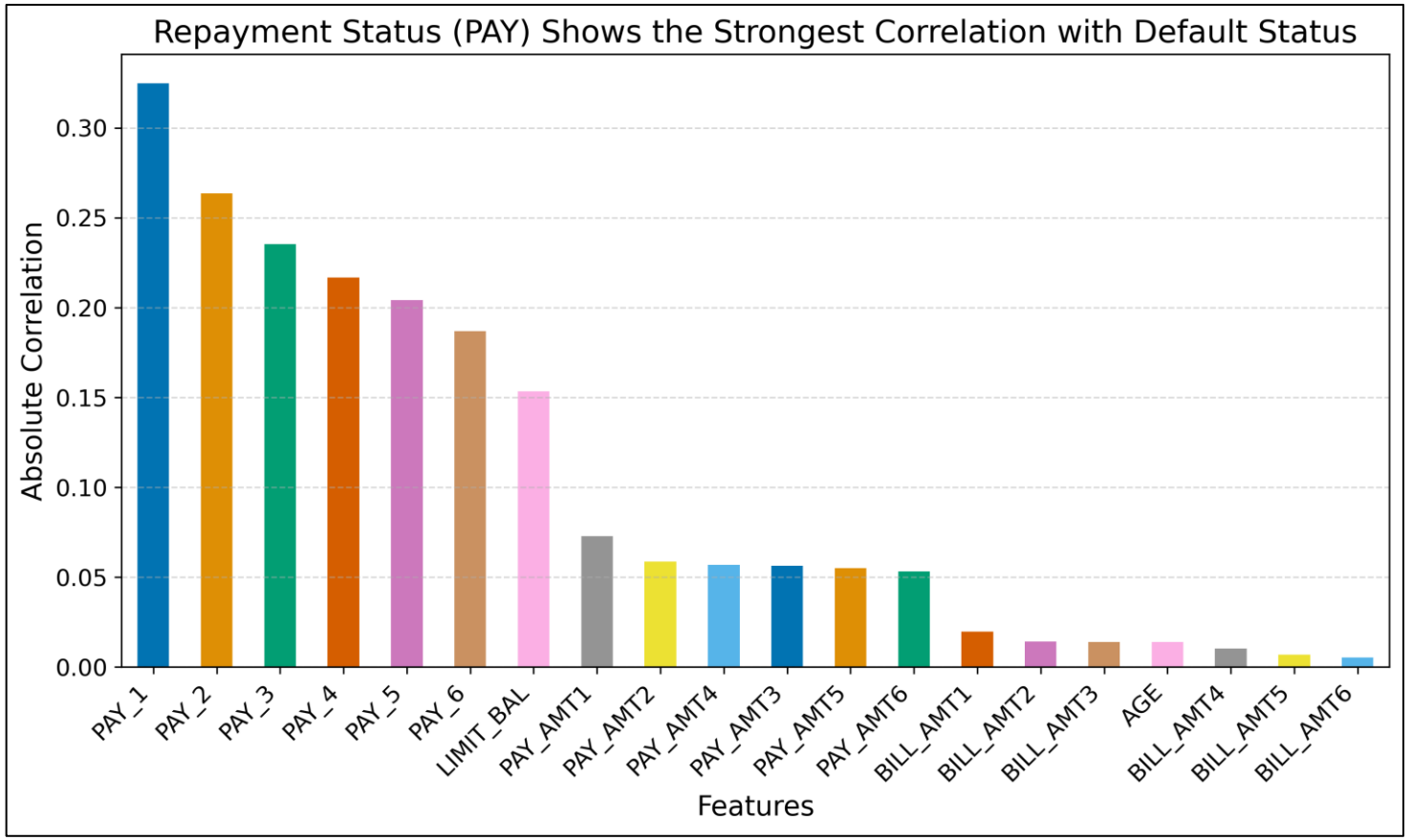
- Non-defaulters have a median credit limit 60,000 TWD higher than defaulters. High credit limits may influence predictions, but repayment status remains the stronger predictor.

EDA Insights

Graph 1



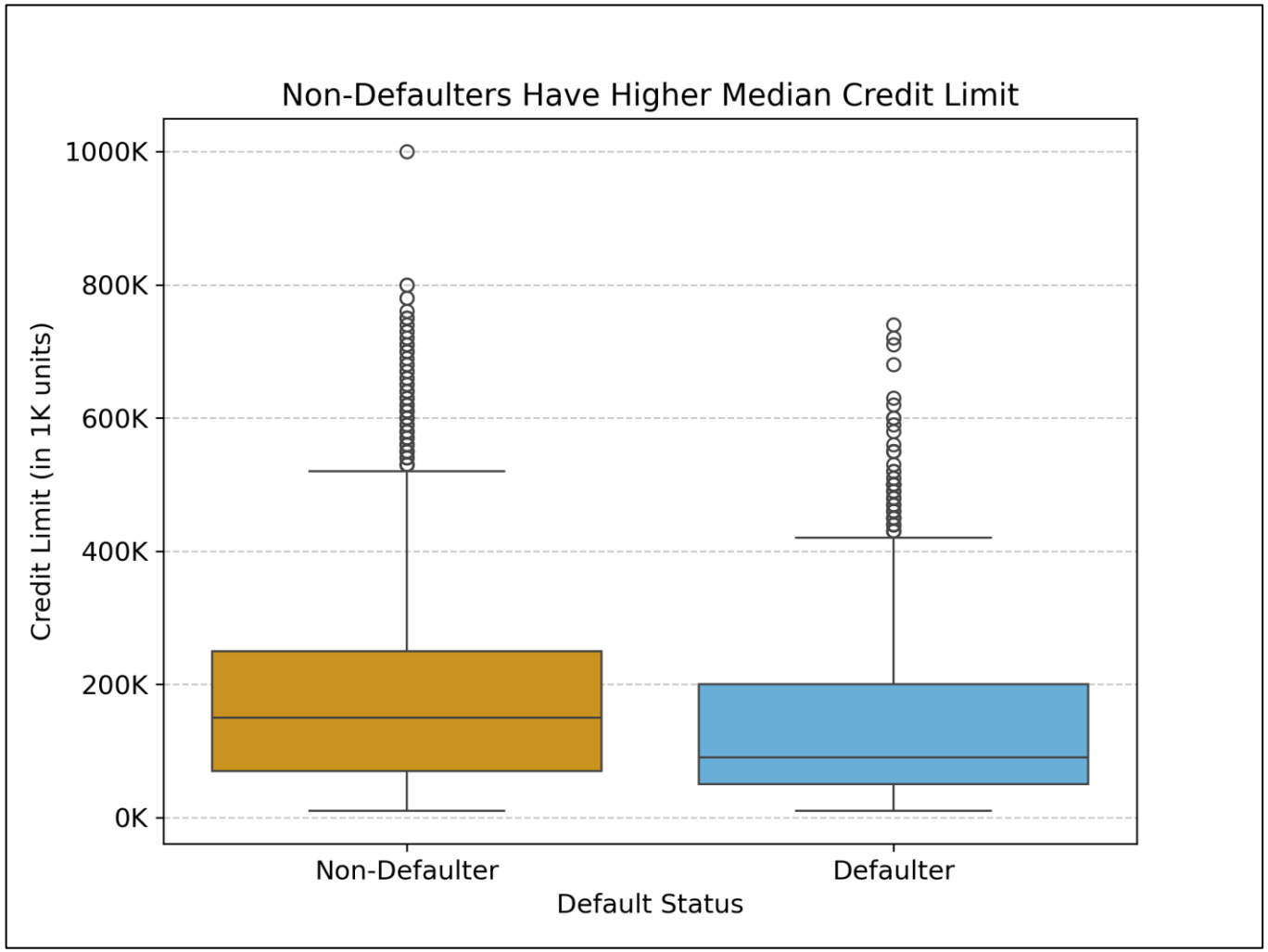
Graph 2



* Feature definitions in Appendix section on slide 13.

EDA Insights

Graph 3



Modelling Results

Graph 4 (Model Metrics):

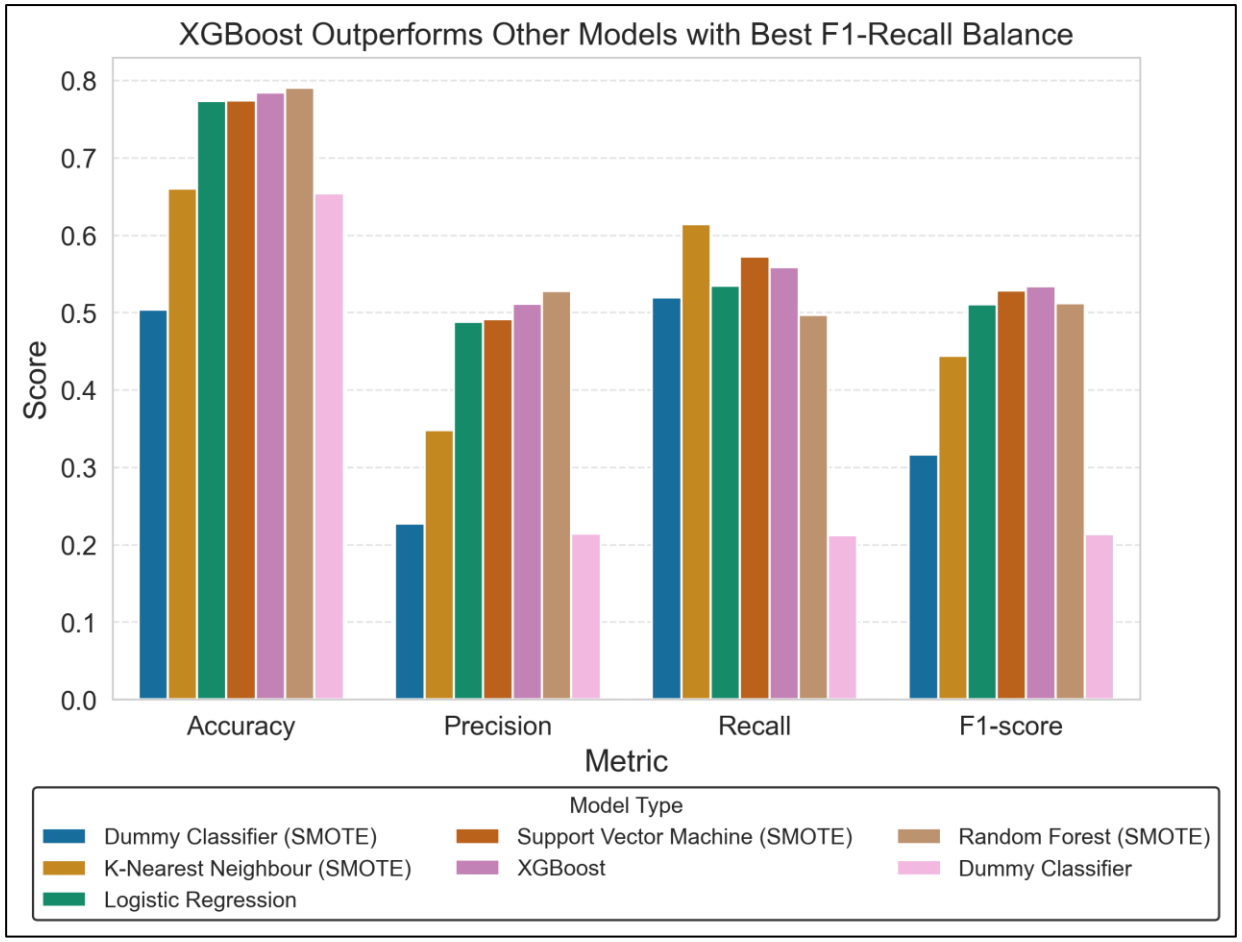
- XGBoost and Logistic Regression performed best without SMOTE, with a focus on optimizing the F1-score and prioritizing recall to minimize financial losses from false positives.
- XGBoost achieved the best balance of F1 (0.53) and recall (0.56), outperforming the Dummy Classifier by 0.22 in F1.
- KNN demonstrated a high recall of 0.61 but low precision (0.34), risking potential customer and revenue loss due to false positives.

Graph 5 (SHAP Analysis of XGBoost):

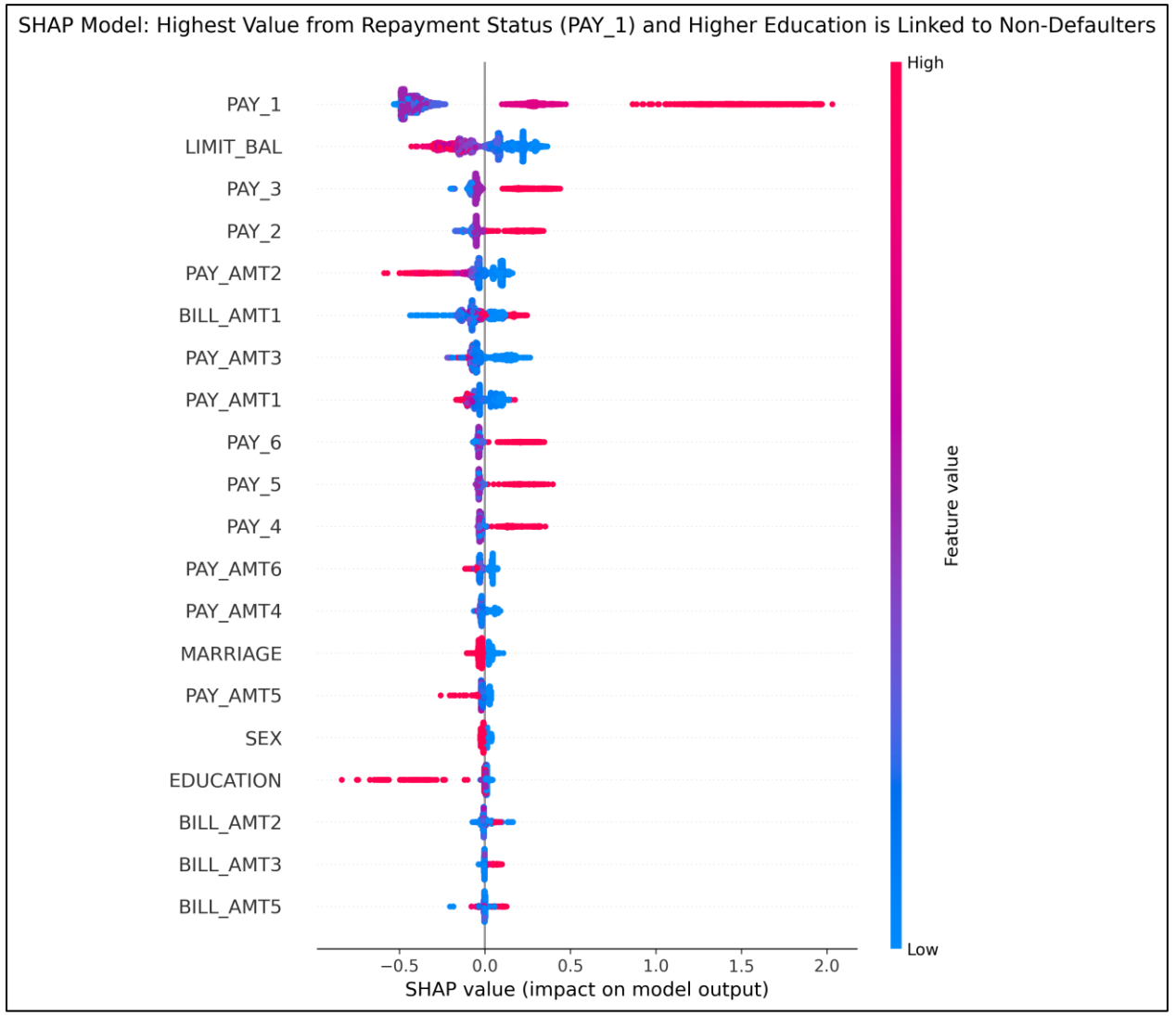
- Repayment status (PAY_1) had the highest SHAP value in the XGBoost model, making it the most influential predictor. Education was the only demographic feature with a significant impact, where higher education correlated with a lower likelihood of default.
- Previous payment amounts from August had a stronger predictive impact for non-default than more recent payments from September.

Modelling Results

Graph 4



Graph 5



Modelling Results

Graph 6 (Dummy and XGBoost Confusion Matrix):

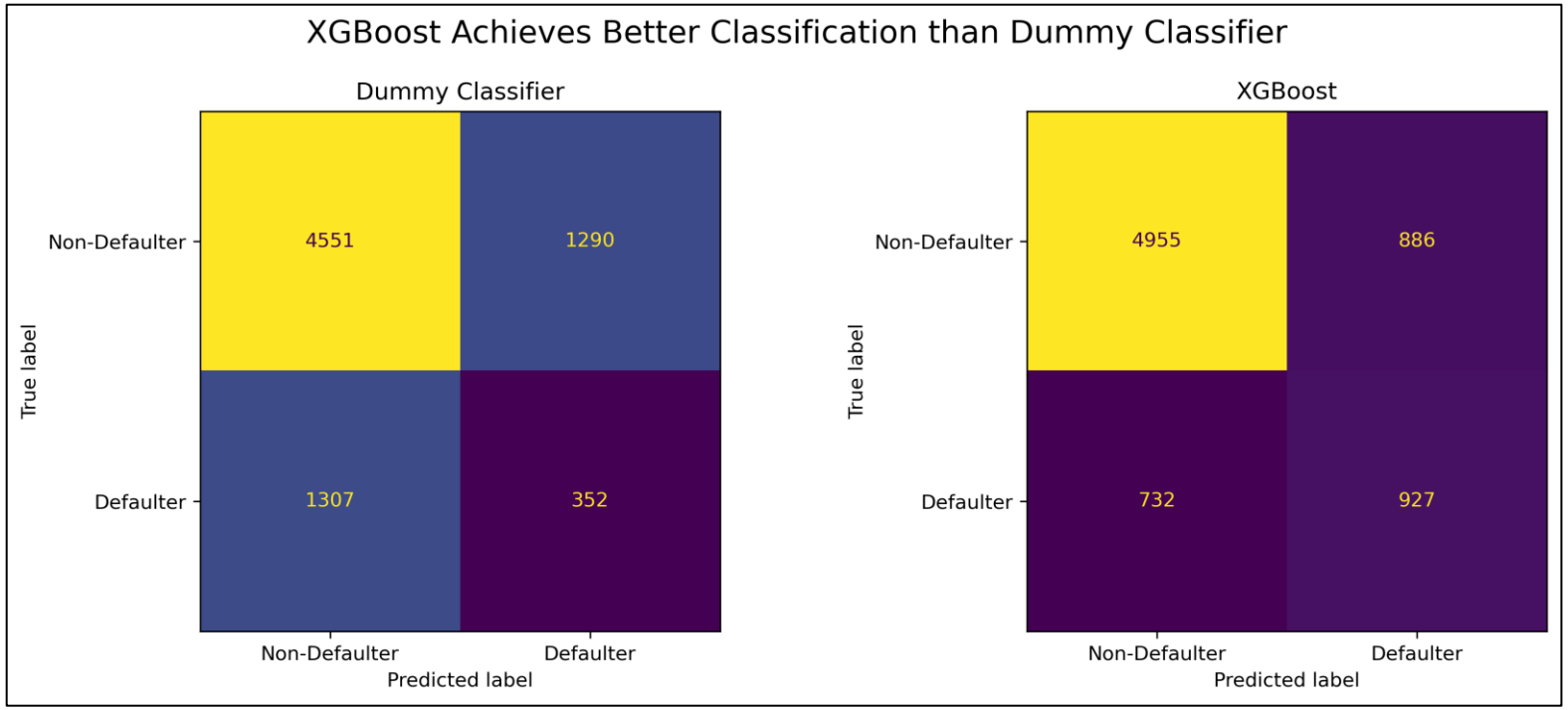
- Correctly classified 575 additional defaulters, reducing financial loss by targeting high-risk customers effectively.
- Reduced misclassification of 404 non-defaulters, improving customer experience and minimising unnecessary loss of customers and potential reputation damage.

Graph 7 (Precision-Recall Curve):

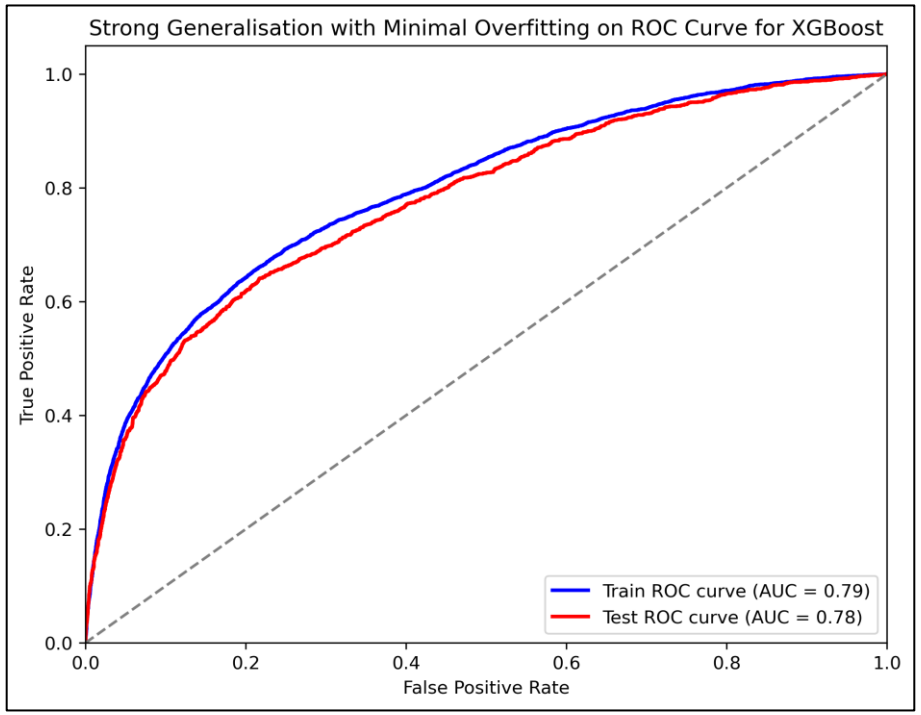
- XGBoost shows a minimal 0.01 difference between the test and train datasets, indicating low overfitting and strong generalization.
- Achieved an ROC-AUC score of 0.78, demonstrating the model's ability to accurately discriminate between defaulters and non-defaulters.

Modelling Results

Graph 6



Graph 7



Next Steps

- Focus more on the Support Vector Machine (SVM) model, as an untuned SMOTE-based SVM performed nearly as well as the tuned XGBoost model.
- Investigate other resampling techniques, such as under-sampling and over-sampling, to address class imbalance.
- Shift focus on feature engineering for model improvement, rather than predominantly concentrating on hyperparameter tuning.

Appendix

Table 1: Data Dictionary.

Field	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_1	Repayment status in September 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2	Repayment status in August 2005 (scale same as above)
PAY_3	Repayment status in July 2005 (scale same as above)
PAY_4	Repayment status in June 2005 (scale same as above)
PAY_5	Repayment status in May 2005 (scale same as above)
PAY_6	Repayment status in April 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July 2005 (NT dollar)

Appendix

Table 1: Continued.

BILL_AMT4	Amount of bill statement in June 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April 2005 (NT dollar)
defaulter	Default payment (1=yes, 0=no)