PROJECET: PREDICTIVE MODELING FOR CREDIT CARD DEFAULT RISK

Credit risk is a fundamental concept in the financial world, referring to the potential for loss a lender (banks, credit card companies, etc.) faces when extending credit (loans, credit cards, etc.) to a borrower. This means there's a chance the borrower might not be able to repay the loan according to the agreed terms, leading to financial losses for the lender. Predicting credit risk can empowers financial institutions to make well-informed lending decisions manage portfolios more effectively, and ultimately reduce the impact of credit-related losses.

This project focuses on developing a model to predict potential credit card defaults within the next few months. The dataset was obtained from Kaggle, containing information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card of 30000 clients in Taiwan from April 2005 to September 2005. The research employed a binary variable, default payment (Yes=1 and No=0) as the response variable and other 23 variable related to credit limits, demographic details, repayment status, bill amounts, and payment history as explanatory variables. The primary objective of the study is to predict which customer might default in payment in upcoming months.

Exploratory Data Analytics (EDA)

The EDA (Exploratory Data Analytics) reveals key insight into the factors influencing credit card usage and default risk. While the dataset lacked missing values, it exhibited a significant class imbalance, with 22.1% of clients defaulting next month. Despite females comprising the majority of (60.2%) of credit card users, males showed a higher default rate (20.8% v/s 24.2%). Similarly single clients, although more in number (53.2%), married clients are comparatively more prone to default. Clients with higher education levels, especially university graduates, are more inclined to use credit cards, with a higher chance of defaulting. The age distribution is right skewed, indicating that older customers (50 above) are less likely to use credit cards. Clients in their 30s are primary credit card users, and those in their 20s and 30s are more likely to default. Analysing credit limits revealed most clients possessed limits below 200,000, and smaller limits were associated with higher default risk. Most clients consistently make their payments on time. However, the highest number of defaults occurred with payment delays of 2 months or more.

Class Imbalance Handling with SMOTE:

To address the significant class imbalance in the dataset with only 21% clients defaulting, Synthetic Minority Over-sampling Technique (SMOTE) was applied. It is a powerful technique used to address class imbalance in machine learning datasets. It stands out by oversampling the minority class by creating synthetic instances rather than replicating existing ones, thereby mitigating the risk of model overfitting.

Model Selection and Evaluation Metrics:

Primary modelling approaches considered were Logistic Regression, Random Forest, XGBoost, and AdaBoost. These models were evaluated on the original dataset and the SMOTE-enhanced dataset. Subsequently, hyperparameter tuning was performed using GridSearchCV to optimize the performance of different machine learning models.

Accuracy can be misleading in imbalanced datasets as it prioritizes the majority class, potentially making poor performances on the minority class. Therefore, we focused metrics like AUC-ROC, Recall and F1 score which provide more informative insight into a model's ability to handle imbalanced data. AUC-ROC (Area Under the Receiver Operating Characteristic Curve) assesses the model's ability to distinguish between positive and negative classes across different probability thresholds. A higher AUC-ROC score indicates better discrimination between those who will default and those who won't. F1 score on the other hand, balances precision and recall, providing a good comparison between false positives and false negatives in imbalanced settings. Recall, which measures the model's ability to identify true defaulters, is particularly crucial in credit risk prediction. Additionally, metrics such as precision and confusion matrices were also considered to gain a comprehensive understanding of the model performance.

Impact of SMOTE on Model Performance:

As expected, SMOTE oversampling significantly improved recall scores for all models, with increase ranging from 18.1% to 21.2%. This demonstrates its effectiveness in addressing class imbalance and enhancing the identification of defaulters. Precision generally decreased with SMOTE, but the trade-off was acceptable considering the significant gain in recall. The impact of SMOTE on ROC AUC varied across models. While some modes exhibited slight decrease XGBoost showed highest ROC AUC value (0.768) suggesting its superior ability to discriminate between defaulters and non-defaulters.

Model Comparison and Selection:

Among the evaluated models XGBoost (resampled using SMOTE) emerged as the preferred choice due to its strong recall (0.552), high ROC AUC (0.768), great F1 score (0.52) and moderate precision (0.492). It effectively identifies defaulters while maintaining a reasonable balance over false positives. Resampled Random Forest also demonstrated strong performance, particularly in recall (0.580), indicating its ability to identify most defaulters, and maintains a good balance of F1-score (0.531), ROC AUC (0.771), and accuracy (0.775). Both XGBoost(smote) and Random Forest(smote) offer significantly improved performance compared to other models and address the class imbalance challenge effectively, making them highly suitable for credit card default prediction.