SUMMARY

Name : Raja Chowdhury

Gmail:- **rajachowdhury2468@gmail.com**

Github Link:- **https://github.com/RajaChowdhury/Credit-Card-Default-Prediction---Capstone-Project.git**

Problem Statement:

This project is aimed at predicting the case of customer default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S chart to evaluate which customers will default on their credit card payments.

Our Approach:

Our dataset contains 30000 observations of 25 variables from a bank.

- There are no null values present in the dataset as most of the dataset consists of categorical variables.
- We tried to remove most of the outlier from our dataset.
- We treated multicollinearity and treated the variables which were showing high multicollinearity.
- After that we did one hot encoding to convert categorical features to numeric.
- We used SMOTE technique to handle data imbalance.
- In Machine Learning we used algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, Random Forest and XG boost and fetched respective scores from all of these.
- We took out variable importance of every variable of our dataset.
- We did hyperparameter tuning for all the algorithms.

From all the features the most important features are dues, payment_value, limit_value and age.

Conclusion:

Key Findings from EDA

- From Subplot 1 we can conclude that number of defaulters are very less than non-defaulters
- From Subplot 2 we can conclude that people of age 79 and above are less like to be defaulters
- From Subplot 3 we can say that female tends to pay on time than male.
- From Subplot 4 conclusion is married people are most defaulters than singles and others.

In classification problems, an imbalanced dataset is also crucial to enhance the performance of the model, so different resampling techniques were also used to balance the dataset. We first investigated the datasets by using exploratory data analysis techniques, including data normalization. We started with the random forest model, then compared the results with traditional machine learning-based models. The prediction accuracy rate of the random forest model is higher than the traditional machine learning-based models. The random forest method given the best accuracy of 86.7% while utilizing the K-means SMOTE resampling method on Taiwan client's credit dataset.

In our project, we used many algorithms like Logistic Regression, Support Vector Classifier, Decision Tree Classifier, XGBoost Classifier and Random Forest Classifier. Below is the best algorithm for the metrics.

Random forest classifier-

- The precision on test data : **86%**
- The recall on test data : **86%**
- The accuracy on test data : **86%**
- The f1 score on test data : **86%**
- The roc_auc on test data : **86%**