Capstone Project Submission

Team Member's Name, Email and Contribution:
NAME- RAJAKUMARAN S **EMAIL –** rajakumaransrs@gmail.com **CONTRIBUTION -** Data Analysis, Data Visualization, Feature Engineering, Fitting Models, Model Explainability and Report Writing.
GitHub Repo link.
https://github.com/Rajakumaran-S/Credit_Card_Default_Prediction_Capstone-Project_III
Short summary of your Capstone project

Problem Statement:

The contents of the data came from a country called Taiwan. The purpose of this project is to conduct quantitative analysis on credit card default risk by applying 6 classification machine learning models. Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting. Heavily relying on credit scores could cause banks to miss valuable customers who are new immigrants with repaying power but little to no credit history. This analysis is a machine learning application on default risk itself and the predictor features do not include credit score or credit history. Due to the regulatory constraints that banks are facing.

The problem statement was to build a machine learning model that could predict the customer who default in upcoming months.

From the EDA we observed that,

- 78% of non-default customer and 22% of default customer. We have an imbalance dataset of the target variable.
- Female: Non-Default - 76%, Default 24%

  Male: Non-Default - 78%, Default 22%

  Females have lower default risk than males in this dataset.
- 20 to 45 years customer are on average for defaulters

  Age above 60 years are almost defaulters
- Customer which had education at university level has more user as well as defaulters
- Married customer count is greater of all. Married and single defaulter customers does not have much difference but, married customers takes lead for defaulters
- Checking correlation between columns.

Feature Engineering

- Apply Standard scaler to the independent features.
- Apply SMOTE of Oversampling of Target variable
- Then Train Test Splitting of the dataset

Fitting Model

- Logistic Regression
- Support Vector Classifier
- K-Nearest Neighbors Classifier

- Random forest Classifier
- XG Boosting Classifier

  With Hyperparameter tuning.

Checking Model performance with precision, recall value and KS Chart

Conclusion:

With every classification model, there is a general trade-off between precision and recall. A model's recall can be adjusted to arbitrarily high at the cost of lower precision. In these 6 models, if the firm expects high recall, then XGB classifier model is the best candidate. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model.

We understand creditors need to make decisions efficiently and, in the meantime, to abide by regulations, the machine learning models in this analysis can be served as an aid to credit card companies, loan lenders, and banks make informed decisions on creditworthiness based on accessible customer data. We suggest the model outputs probabilities rather than predictions, so that we can achieve higher accuracy and allow more control for human managers in decision making.