Capstone Project Summary

Team Member's Name, Email and Contribution:

1. Tito Varghese

   Email - tito.varghese1992@gmail.com

   - Data Preprocessing
   - Feature Engineering
   - Model Formulation- KNN
   - Model Formulation- XGBoost
   - Model Formulation- Naïve Bayes
   - Evaluation Metrics
   - Hyper Parameter Tuning

2. Lakshmi Keerthana

   Email - keerthana826@gmail.com

   - Data Cleaning
   - Exploratory Data Analysis
   - Model Formulation-Logistic Regression
   - Model Formulation-Support Vector Machine
   - Model Formulation- Random Forest
   - Evaluation Metrics
   - Hyper Parameter Tuning

Please paste the GitHub Repo link.

GitHub Profile Link: - https://github.com/7692TITO

GitHub Repository Link: - https://github.com/7692TITO/Credit-Card-Default-Prediction

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Credit Card Default Prediction
Summary

A payment default occurs when you fail to pay the Minimum Amount Due on the credit card for a few consecutive months. We shall predict the credit card defaulting using cardholder characteristics. The dataset contains details of credit card holders of a bank in Taiwan for the period April to September, 2005. The features available include some basic customer demographics, credit line and history of payment/default for six months, bill amounts and payment amounts for that period and a target variable indicating default.

We will first start by preparing the dataset for our machine learning models. After loading the dataset, we then get to Data Cleaning, we perform the tasks such as handling missing values, handling duplicate values, handling outliers and handling Irrelevant records.

Then we started with the EDA to understand the data better where by visualizing using seaborn and matplotlib, we have checked the educational qualification, marital status, age, gender and their relationship with the credit default. Through bar plots we have checked repayment status and previous amount paid.

We then get to the process of Feature Engineering where we have derived new features like total bill amount and total paid amount. We have also implemented binning and later did the process of feature encoding where we used hot encoding techniques to train our model more effectively. We have also handled imbalanced target variable using smote.

Before training the model, we prepare the dataset by splitting it into train and test data, we have used Standard Scalar to perform the scaling. We build a function to train the model using multiple supervised learning algorithms and classify the different models on the basis of their test - train accuracy. The models that we have used were Logistic Regression, Gaussian Naive Bayes, Support Vector Machines, K-Nearest Neighbour, XGBoost and Random Forest**.**

After hyper parameter tuning the XGBoost Model comes out to be the best model in terms of its AUC_ROC score (0.875) and Recall score (0.82) and we can predict with 87.45% accuracy, whether a customer is likely to default next month. The Second best model was the Support Vector Machine with an AUC_ROC score of 0.875 and a Recall score of 0.805 and we can predict with 87.5% accuracy, whether a customer is likely to default next month.

Except Naive Bayes model, all the models have got really good ROC_AUC scores with a probability of 0.85 on an average.The Random Forest and KNN models were really overfitting with default parameters and we handle the over fit in both these model by fine tuning the model. We see that being Female, More educated, Single and between 30-40 years old means a customer is more likely to make payments on time.