Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email, and Contribution:

Team Member's Role:-

- **Sanjaya Kumar Khadanga**

  eMail- skhadanga38@gmail.com
  - Data Understanding
  - Feature Analysis
  - Feature Engineering
  - Exploratory Data Analysis
  - Implementing Logistic Regression
  - Implementing Random Forest
  - Hyperparameter Tuning
  - Evaluting Models

- **Bibhuti Bhusan Sahu**

  eMail- sahubibhuti45@gmail.com
  - Data Understanding
  - Feature Analysis
  - Data Visualization
  - Multivariate Analysis
  - SMOTE
  - SVC
  - ROC AUC Curve
  - Research Analytics
    - Technical documentation

- **Balaram panigrahy**

  eMail- balarampanigrahy42@gmail.com
  - Data Understanding
  - Data Visualization
  - Multivariate Analysis
  - One Hot Encoding
  - XGBoost
  - Research Analytics
    - Technical documentation

Please paste the GitHub Repo link.

Github Link:- https://github.com/sanjaykhadanga/credit-card-default-prediction

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

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 4 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 this study, we discovered a few interesting insights which may or may not hold for other datasets. We learned the most important predictors of default are not human characteristics, but the most recent 2 months' payment status and customers' credit limit.The conventional thinking of younger people tend to have higher default risk is proven to be only partially true in this dataset. Also, surprisingly, customers being inactive for months doesn't mean they have no default risk.

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 4 models, if the firm expects high recall, then random forest and xgboost classifier models are 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.