Project on Expected Credit loss (PD,LGD & EAD) & CCAR

Comprehensive Capital Analysis and Review (CCAR) is an annual exercise by the Federal Reserve to check whether bank has sufficient capital to continue operation in economic and financial stress. In this process the Federal Reserve evaluates institutions' capital adequacy, internal capital adequacy assessment processes, and their individual plans to make capital distributions. DFAST is a complementary exercise to CCAR which is a forward-looking component conducted by the Federal Reserve and financial companies supervised by the Federal Reserve to help assess whether institutions have sufficient capital to absorb losses and support operations during adverse economic conditions. This model is stress tested on 3 scenario which is baseline, severe & adverse.

The unprecedented times we are in right now makes this project ever so relevant because of the need of maintaining banks in good health when the economy is going through a turmoil. The aim of the project is to make a model which can predict the if a customer will be able to make repayments of the amount he would be borrowing from the bank. The aim of doing this is to determine the risk involved in lending money to a customer. Probability of Default/Loss Given Default analysis is a method used by generally larger institutions to calculate expected loss. Having a model provide precise estimates of risk/default when lending money to a customer will not only minimize the losses which bank incurs due to defaults, but it would enable the bank to cater to more customers since they would not need to maintain excess liquid funds to counter the credit risks. The expected loss calculation is Expected Loss = EAD x PD x LGD.

This project was data intensive and hence pre-processing the data was one of the most crucial and time-consuming steps. The exploratory data analysis performed using libraries like pandas, numpy, seaborn and matplotlib enabled me to handle the missing values in the dataset. It took domain knowledge while deciding about either interpolating the missing value or dropping the rows with missing values. Generally, I restored to more conservative approach of interpolating the missing data to not miss out on any significant information while choosing to drop the missing values. Statistical analysis of dataset gave me insights about the overall distribution of the dataset by giving me insights about the mean, median, standard deviation.

The dataset was not perfect which called for the need of feature engineering. Proper feature engineering could render the dataset helpful in making right predictions. For categorical variables, dummy variable encoding technique was used. For continuous variables, I used polynomial and log features as well as binning to convert some of the continuous variables into discrete. I have also made some interaction terms. There were two target variables which the model would be aiming to predict. They were – Probability of Default (a categorical target variable) and Recovery Rate (in case of default).

For modeling I used scikitlearn to split my datasets into training and validation and for modeling I used Logistic Regression for categorical target and linear regression for continuous target.
For categorical target, the confusion matrix was used to evaluate the performance of the model. It was further used to calculate the sensitivity, specificity.
For continuous target variable, accuracy measures like Root Mean Squared Error and R-squared was used.
While predicting the categorical target variable (Probability of Default), the model had an overall accuracy of 65% which could be improved further by performing hyperparameter tuning and feature engineering.

The business impact of the model could further be improved if the model performs better at predicting defaulter as compared to predicting non-defaulters. This was a challenge with this dataset since the model was not trained enough (because of lack of defaulter data) to predict defaulters. This issue could be solved by performing stratification and under sampling in python.

Overall, the model was able to predict the probability of default with high accuracy which in turn enabled to focus on the predictions of recovery rate. This model and the project could be implemented in IBs to calculate Loss Given Default, Exposure at Default and Probability of Default which could be leveraged to get details about expected loss. Thus, maintaining banks in good health, good business at lesser risk.

For reference please my work on Github account for referring my project on https://github.com/Rajsawant25

- Expected credit Loss in Python (PD, LGD, EAD)
- CCAR on Fannie mae Loan data set on R/SAS
- Option Pricing model on R