Capstone Proposal

Business Understanding

Lending is a fundamental income-generating activity for banks, and the default rate on loans is a critical factor in determining a bank's profitability and liquidity. Effective credit risk assessment is essential to minimize defaults and grant loans only to creditworthy applicants.

One of the main challenges that banks confront is determining the risk level of loan applicants. Banks require a way to assess the creditworthiness of loan applicants for this task. As a result, internal credit risk rating systems are becoming increasingly important in this situation.

This study is designed to assist banks in predicting the risk level of a loan applicant based on the information applicant provides in the loan application when applying for a loan. The bank can then decide whether or not to provide the loan based on the results. The model is built using the payback information of previous loan debtors.

If an applicant is classified as high risk, the bank may either decline the loan application or impose a risk premium to compensate for the risk.

Compared with the traditional subjective judgmental methods, credit scoring has several benefits. Mainly it offers greater efficiency and savings in terms of labor and time in the loan approval process. Therefore experienced risk managers can allocate more of their time for important issues. On the other hand it reduces subjectivity, which makes the lenders apply the same standards to all applicants, eliminating human biases or prejudices in the process.

Data Understanding

The source for this dataset is kaggle
https://www.kaggle.com/datasets/yasserh/loan-default-dataset/data

The dataset is with 148,000+ records & consists of multiple deterministic factors like borrower's income, gender, loan purpose, interest rate, LTV values etc. There are lot of missing values as well in different fields

Data Preparation

There are 30 categorical features and numerical features in the dataset. The first step will be splitting the dataset to test and training sets. As there are a lot of missing values it will be required to apply strategies for missing values as the next step. It is hoping to step a pipeline to handle missing values. Also would be required to calculate a few more ratios and loan monthly premium using available information.

Modeling

The target variable of the data set is borrower default or not. Therefore will be addressing a classification model. In order to identify, the most efficient model will be starting with the dummy model and will progress with logistic regression, decision trees, ensemble models and at last neural networks.

Evaluation

As model evaluation methodology hoping to use log loss, classification reports, area under ROC curve on training and test datasets. Also confusion matrix as a visual approach.

As MVP hoping to develop an efficient model using logistic and ensemble models. As the stretch goal it is expected to develop a neural network model for the prediction.