Capstone Project Proposal

LendingClub Loan Default Prediction

Business Understanding

Banks have a huge volume of applicants applying for loans. Some of the applicants do not have record with the National Credit Bureau. Or some might have very light credit score. It doesn't mean that lending money to them is highly risky so we should reject all of these applicants. The goal is to develop a model that could step by step explain the results for all processed cases marking which input parameter had what impact on the likelihood of the case falling into one of the binary categories (loans paid-off and charged). The model is to help loan issuers to maximize the profit and reduce NPL.

Data Understanding

The dataset was downloaded from https://www.kaggle.com/datasets/wordsforthewise/lending-club . The dataset contains 2007 -2018 applicants' information, including loan status, interest rate, annual income, loan amount, grade, FICO score, etc.

Data Preparation

In this dataset, there are 2.3M observations and 151 features. It's time-consuming to run the models by using such large dataset. In order to run the model more efficient, I will clean the null values in the dataset and drop the features that won't help modeling. After that, I will do data exploratory.

Modeling

I will use Logistics Regression Model, Decision Tree Model, Random Forest Model and XGBoost Model to train the data and test the result. And I will choose a baseline model to compare the performance of the models.

Evaluation

I will report both accuracy and recall score on training and test data. And I will use GridSearchCV to increase model performance through parameter tuning on the model with highest recall score.

Deployment

I expect to deploy the model to AWS SageMaker to train the models.