ORIE 5741 Project Proposal

Yijun ZHU(yz2745), Xinyue LIU(xl533), Ruifan CHEN(rc654)

Background

Financial institutions as banks play a crucial role in the market economy by financing corporates and individuals to form the liquid market. Thus, it is of great significance for financial institutions to make decisions wisely about which individuals or businesses to offer credits to avoid bad debt, which may lead to the bankruptcy. Under this scenario, our project is to build a model to make predictions and give suggestions on whether credit should be offered or not to a client. We aim to help financial institutions in New York state to analyze their applicants, make good decisions on approval of clients' loan application, trying to decrease the default rate.

Research Question

How can financial institutions do better analysis, and make better decisions on whether to offer credit to a client or not?

Applications of Dataset

The dataset we plan to use is the Home Mortgage Disclosure Act dataset, and we would narrow down our analysis on data of New York state in the year 2020. This dataset contains basic financial and geographical information about clients like gross annual income, credit score type, race, ethnicity and so on. It also provides the information about the loan customer applied like loan term, interest rate, and institutions' decision on the clients' loan application. Based on the information above, we plan to apply this dataset to analyze and reach the following objective:

- Explore the dataset by visualization to give a general view and find some interesting directions for analysis.
- Understand the features that are highly related to the decision making of a client's loan application.
- Build a classification model allowing financial firms to make wiser financial decisions on whether to offer credit to a client or not.

Proposed Approaches (why we can succeed)

Firstly, in data preprocessing, we will deal with the null values, apply unit-length scaling to reduce the colinearity impact, along with other data cleaning process. Then we would also analyze the current dummy variables, trying to reset them in a better way.

Secondly, we will do some data exploration to find relationship between certain variables, in which we may find directions for further data analysis and conduct data transformation for next modelling process.

We then split the whole dataset into training and test sets for feature engineering, selecting related features and conducting data transformation if needed. Then our plan is to use n-fold cross validation to score the performance of different models including logistic regression and random forest. Finally, we will compare the test error and propose the best model, helping financial firms to make better decision on credit approval.

After obtaining model based on original dataset, we will try to improve the model by joining other datasets to increase model accuracy. For instance, add a dataset about each county's house price or income class. The initial reason is that living expenses and house price can vary among different counties, which means with the same income, clients from different counties may have various solvency. Thus, it might be helpful to include these variables. Additionally, we would also consider other possible improvements to enhance our prediction accuracy.