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. Financial institutions should make decisions wisely about which individuals or businesses to offer credits to avoid bad debt, leading to the bankruptcy. Under this scenario, Our project is to build a model to make predictions on the probability of default to give suggestions on whether a load should be denied. We aim to help financial institutions in New York state to choose their applicants correctly and try to avoid most default situations.

Applications of Dataset

The dataset we plan to use is the Home Mortgage Disclosure Act dataset, choosing data of New York state in the year of 2020. It contains basic information about clients and the institutions' decision on the clients' loan application. We plan to apply this dataset to reach the following objective:

- Present the dataset by visualization to give a general view.
- Explore and Understand the features that are related to the clients' bankruptcy.
- Build a classification model allowing financial firms to identify clients who are "risky", and make wise financial decisions.

Proposed Approaches

After data preprocessing that we are going to organize the current dummy variables, standardize, clean the null values and split it into train and test datasets, we would like to observe the features and make selections to fit different models for our predictions. After comparing the testing error, we propose the current best model.

According to the performance of the current models, we are planning to include other datasets in the feature engineering process to increase the accuracy of models. For instance, we would add a dataset about each county's house price or income class. The reason for consideration is that living expenses and house price may vary among different counties, which means with the same income, clients from different counties can have diverse repaying abilities, indicating various solvency. Thus, it would be helpful to the whole prediction model.