5741 Project Proposal

Project Introduction:

Balancing risk and profit is an eternal topic for firms, especially for the financial firms that live on loan interests from the credible customers. Correspondingly, it is necessary to have a comprehensive and fair evaluation of the applicants on whether they should get credits. Hereby, we propose to establish a series of machine learning models to accurately score the credits of each applicant based on their historical data such as demographic information, previous financial records, current financial status etc.

Goal of the project:

Accurately estimate the risk of default of applicants based on their features and assist companies understand profitability and make solid decisions.

Dataset Description:

The Home Mortgage Disclosure Act (HMDA) by FFIEC provides sufficient data fields about home mortgage loan application records, including the features and the label of whether the application got denied or passed. The data set we adopt is New York State HMDA data for 2020, which contains demographic information, applicant credit, application status, and mortgage & loan information.

- **Demographic Information:** the demographic information in this data set could provide the data about age group, sex and race about the applicants. Such information would be useful to determine the capability of a certain applicant to pay back the loan.
- **Applicant Credit**: this dataset included the information of the credit score for each applicant as well the evaluation from the system. The credits of applicants are the most important factor for banks to consider whether to permit their applications.
- **Mortgage/Loan Information**: this category contains the detailed information of each mortgage/loan including the loan amount, the loan type and the loan period, etc. These data would give us an idea whether the loan has a proper usage or a reasonable amount.
- **Application Status**: information like pre approval, acceptance and denial.

Solution Strategy:

Through the data set, we would process, model and make evaluations via a series of techniques such as data cleaning, feature engineering, and modeling. A precise, fair, and quickly-responded model is estimated to be proposed after model selection and validation. The model should be able to quantify the risk more thoroughly and estimate profits more efficiently. This model would allow the company to make accurate and quick decisions on whether applicants would be offered credits, and whether it is profitable for the company to take the risk.

With the help of this data set, our model should be able to learn basic metrics to determine an eligible applicant for credit. This data set not only has a great size of data, but also incorporates essential information for determination of eligible applicants. The data size allows for solid training for modelling, while the quality of data guarantees the fair results.