Yiqing Xin, Yiwen Sun, Yihao Luo
ORIE 5741-Project Proposal-Data Analysis
3/19/2023

Who Can Get a Loan?

The objective of this project is to analyze loan application trends across various financial institutions in New York State in 2021. While giving credit to individuals, financial companies face credit risk, the possibility of a loss resulting from borrowers' failure to repay loans or meet contractual obligations. It can result in an interruption of cash flow and increase costs for collection. So financial institutions require applicants' information to decide whether to lend a loan. We will be utilizing data from the FFIEC Home Mortgage Disclosure Act (HMDA), an official United States government website that provides information about loan applications and decisions. By examining the dataset, we aim to identify the applicant characteristics favored by lenders and have a higher probability of application acceptance and to gain insights into the qualities lenders seek in New York State.

To achieve this goal, we will use the HMDA dataset of New York State in 2021, which contains 787,436 observations. Currently, the dataset includes 99 variables which are information about applicants that the government records. Among these variables, 12 are essential and will be analyzed in depth. Among these variables, action taken is an outcome variable representing loan application acceptance or rejection, with 1 indicating acceptance, while values of 2 and 3 indicating non-acceptance. We will also examine the outcome variable under different subgroups, such as loan types, applicants' ages, race, and sex, to identify any potential relationships or biases among these variables. To be specific, we aim to investigate potential lending discrimination, such as preferential treatment for young individuals over older ones, male applicants over female applicants, and white individuals over people of color.

Before analyzing data and building models, we will first examine the dataset to ensure data quality. For example, we will check missing values in the dataset and implement appropriate ways to handle them, such as removing rows with missing entries, imputing with mean or median, etc. In fact, this is a classification problem as the outcome variable is binary (whether a financial firm will give credit or not). We plan to use logistic regression to decide whether an applicant should be approved for credit. More specifically, for the outcome variable "action taken," we will transform it to value 0 if it is equal to 2 or 3, which means the loan application is not approved. If it is equal to 1, it will be kept at its original value since it means that the loan application is approved. The list of features we plan to use includes loan type, loan purpose, lien status, age, ethnicity, sex, etc.

As a result, our model will determine the characteristics of loan applicants that are likely to be accepted or rejected by financial institutions in New York State. By identifying the features favored by the top 5 financial institutions in the state, our model can help potential loan applicants select the institution with the highest probability of accepting their application, increasing their chances of loan approval.