Midterm Report:The Home Mortgage Disclosure Act (HMDA) Data Exploration

Chengyuan Yang cy362@cornell.edu|Zihao Fu zf223@cornell.edu|Zhengyi Sui zs359@cornell.edu

1. PROBLEM STATEMENT

Balancing risk and profit is an eternal topic for firms, especially for the financial firms that live on loan interests from credible customers. Every year, there are millions of applicants for the home mortgage loan. However, financial institutions haven't found a way or method to differentiate between positive and negative applicants, thus failing to balance their risk and profit. Our goal is to develop a convenient method by establishing a series of machine learning models to score the credits of each applicant based on their historical data such as demographic information, previous financial records, and current financial status. Therefore, this model can provide some reference for these financial institutions whether certain applicants should be qualified.

2. DATASET DESCRIPTION

To fulfill our goal to develop an effective prediction model on credible home mortgage applicants, the HMDA dataset provided by FFIEC is used by our group. FFIEC began to collect HMDA data around 2017 and financial institutions can use this platform to upload their loan/application data, review edits, certify the accuracy and completeness of the data, and submit data for the filing year.

Our group specifically selected the New York HMDA data in 2020 to train and develop our prediction model. In general, the original dataset we have incorporates 99 features with more than 700,000 samples. Among those features, data can be separated into three groups: real-valued data, boolean data, and categorical data.

3. PREPROCESSING AND DATA CLEANING

At the first glance at the dataset, there are many "NaN" values in certain columns and distributed among real valued data and categorical data. For columns with more than 50% data to be "NaN", the entire column is deleted since these features will have rare effects on the accuracy of our model. For those columns with few "NaN" values, certain approaches are applied based on histogram. For numerical data, "NaN" values are replaced either with mean or median based on the distribution in histogram plots. For categorical data, "NaN" values are replaced with the majority vote.

After having processed all the missing values for each feature, another check for outliers was operated. Boxplots and Violin plots are applied to help better visualize outliers in the sample data. In order to preserve most characteristics, detected outliers were replaced with the value of 95 percent quantile. And according to the magnitude of the feature, logarithm is implemented correspondingly.

For our model, the response variable is selected as “action_taken”. This feature has variables 1,2, and 3 to indicate early acceptance, acceptance, and denial. We combine the first two variables to be acceptance, indicated by 0, and the other condition to be 1 as denial. Meanwhile, the feature series “denial_reason” are deleted as they are exactly the hidden information about “action_taken”.

For all categorical variables in the dataset, one-hot encoding has been applied for the preparation of logistic regression.

Some features are deleted as they have no effect on the model. For example, “lei” is deleted as it is a financial institution’s legal entity identifier and demonstrates who is the data provider.

In total, 27 features are deleted for having too many missing values or giving no information.

3. HISTOGRAMS AND DESCRIPTIVE STATISTICS

a. histogram

Action taken is the response variable in our model. A histogram is generated to understand the distribution of the data. From the following histogram figure, we can see that imbalance response variables for action taken. We can use sampling methods to balance the data in the future.

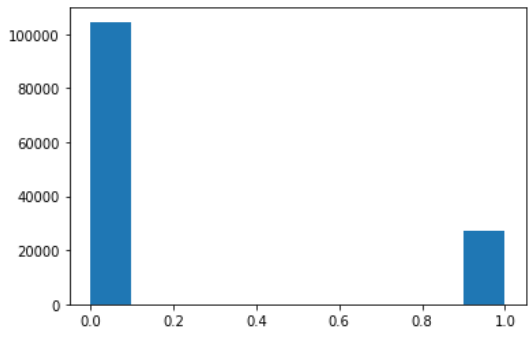

Figure1: histogram for feature “action_taken”

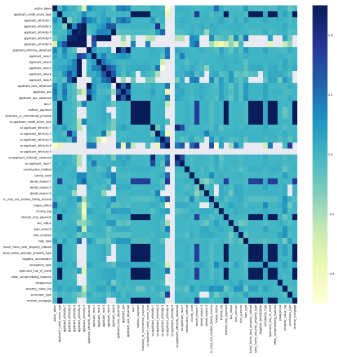

Figure2: Heatmap

b. Heatmap

A correlation heatmap displays the relationship among the selected features. It's clear that some features are highly correlated, which gives a hint that feature engineering might be needed for a more accurate result in future models.

4. PRELIMINARY ANALYSIS

Our group has generated a logistic regression model for the data set after the data cleansing. While doing this model, in case of crashing, we resample the data to 25% fraction of the original data set.

The overall accuracy is already 100% which should be impossible in the real world. This means that some features in the current data set are highly correlated with our response feature. So, our next step will be identifying those features and delete them from the model.

5. PLAN TO TEST MODEL EFFECTIVENESS

- Overall accuracy: in order to promise the company to get enough profit, we need to maintain a high overall accuracy to precisely distinguish the people who have the quality to apply for the mortgage and those who do not.
- Recall score: another important thing for the company is that they do not want to give mortgages to applicants with low credit which would lead to bad debt. So, a low false negative rate would be important, which means we need a high recall score.

6. PLAN TO AVOID OVERFITTING

In order to avoid overfitting, several methods are proposed to fix. One idea would be decreasing the amount of features by removing the less related ones or PCA. It's also worth a try to implement penalty terms as Lasso and Ridge regression will do. We may still try other models like random forest, and we might restrict the width or depth of trees to avoid overfitting.

7. PLAN TO DEVELOP THE PROJECT OVER THE REST OF THE SEMESTER

- Continue on selecting features and pick the most correlated ones to fit the model
- Upsample the minority class of the label to avoid the issue of data imbalanced
- Apply some more models besides logistic regression to fit our data such as random forest.