Fair or Not: Evidence from HMDA Data (Midterm Report)

Anji Zhao (az529), Chenghao Li (cl2567), Yiwei Zhang (yz2454)

***Abstract*— An analysis that explores which types of candidates are most likely to get approved for housing mortgage loan application. The result can not only be used by applicants who want to maximize their chances to get approval, but also be used by regulatory departments to ensure the fairness in terms of granting loans.**

I. INTRODUCTION

Housing mortgage loan granting procedure relies on applicants' background information, and it is important for the financial institutions to predict whether an applicant should be approved in order to minimize their default risk. However, one model may not fit all people in such diversified country. Different races, ethnicities, genders are supposed to have equal opportunities to get approved, therefore, it is also crucial to ensure the model takes only financial conditions into account.

In this project, HMDA data from District of Columbia is used to explore the fairness of loan granting procedure. D.C. has the highly diversified population, which provides less biased dataset for this project. Dataset has all loan application records from 2007 to 2012. Note that 2007 is during the global financial crisis, therefore, data in 2007 may provide different insight, which will be explored in this project.

II. EXPLORATORY DATA ANALYSIS

A. Data Characteristics

The team selected and downloaded the HMDA dataset for District of Columbia from 2007 to 2012. The dataset has 36 feature columns and 254976 records of loan application. For every loan application record, outcome is nominal with value 1 to 5, each representing being approved, rejected, etc. The data also contains nominal values for other features including loan agency names, loan types, property types, and more importantly for this study, applicants' race and ethnicity. Other than categorical features, two features with continuous values are loan amount and applicant income.

The data is complete by itself without any missing values, there are, however, certain features having values of ”No information provided” or ”Not applicable”, these values will be considered missing values in this study. In addition, some features are involved with co-applicants' information, while most of the applications are filed alone, which also makes the dataset messy. We picked 17 most relevant features to conduct this study.

Below is the features being used in this project:

- Categorical: Year, Agency, Loan Type, Property Type, Loan Purpose, Owner Occupancy, Loan Amount, Preapproval Status, Applicant Ethnicity, Co-applicant Ethnicity, Applicant Race, Co-applicant Race, Applicant Sex, Co-applicant Sex, Purchaser Type
- Continuous: Applicant income, Loan Amount
- Output: Action taken

B. Data Visualization

Correlation matrix serves as a great preview tool to see the relationships between features. We used heatmap to visualize the matrix:

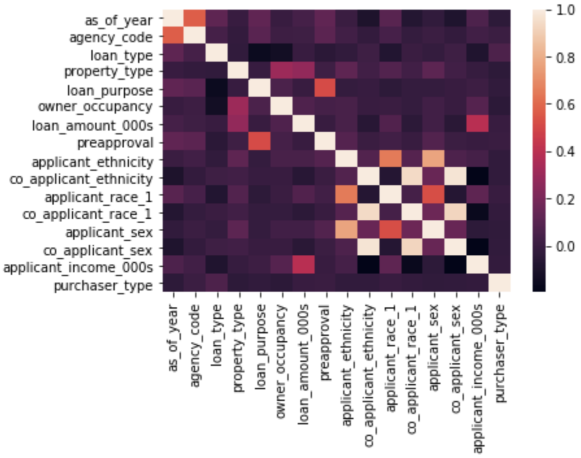

Fig. 1. Correlation matrix visualization between all features

We observe that except (co-)applicant race and ethnicity have relatively strong correlation, correlations are low among almost all other features, which means co-linearity is not significant between those features.

Distributions on the features or output space (y) are important as well to have a basic understanding of the data. For continuous features like applicant income and loan amount, there are some outliers, after excluding the outliers, their distributions are shown in Figure 2 and Figure 3.

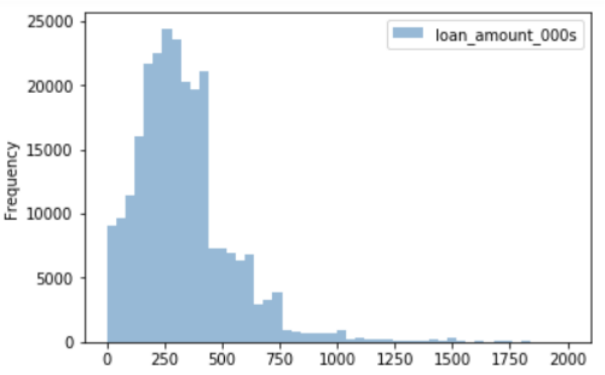

Fig. 2. Histogram of loan amount being applied

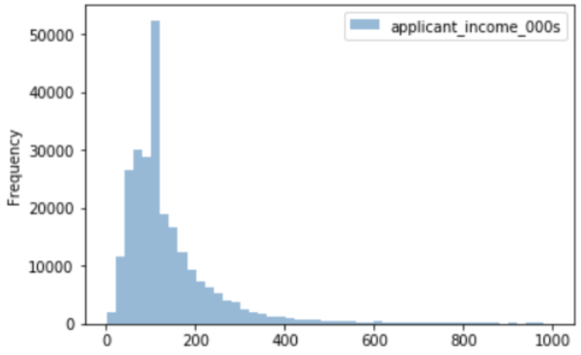

Fig. 3. Histogram of applicants' income level

We can observe that most loan amount being applied are around 250,000 dollars, and majority of applicants have annual income around 100,000 dollars, and the distribution for income is right-skewed.

For output space, to have a first check of the fairness of the loan granting process, it is ideal to group records by sex or race, and plot the bar charts separately.

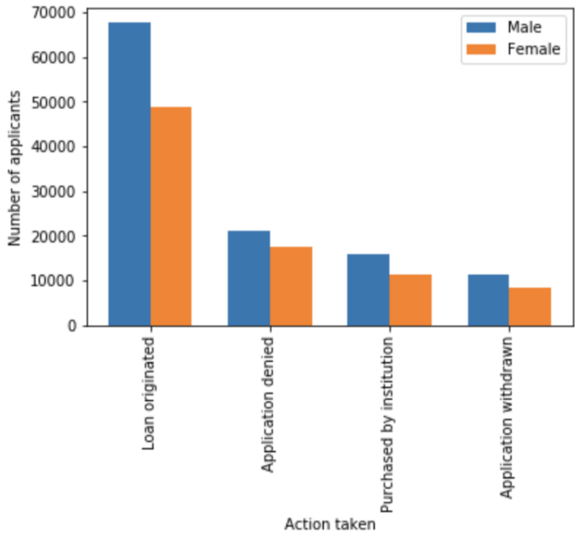

Fig. 4. Comparison between male and female

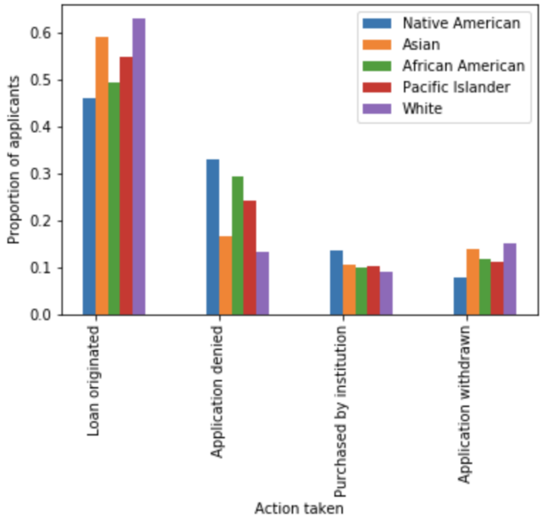

Fig. 5. Comparison among races

Based on Figure 4, we cannot see much difference in pattern for male applicants and female applicants, their distributions are similar. For distributions of difference races, however, we observe that African American, Pacific Islander and Native American have higher denial rates compared to White and Asian.

C. Feature Engineering

We did several feature engineering work. First of all, we explored the features and found the variable applicant_income_000s contains 28608 missing values. Since this variable indicates the applicant's income level and from the visualization part we know it contains extreme outliers. As a result, we chose to fill the missing values with the median. Secondly, we created all the dummies for the categorical values. Thirdly, we handled outliers in the data set. We know that the two real value variables, applicant income and loan amount contains extreme outlier and these are probably incorrect information. So we drop the entries which have applicant income or loan amount that is three standard deviations above the mean. Fourthly, since some of the classes has very little samples and their definitions are very similar to some other classes, so we merged the classes. We added the observations of ”Preapproval request approved but not accepted” and ”Application approved but not accepted” into the ”Loan originated” class. Also, we added the observations of ”File closed for incompleteness” and ”Preapproval request denied by financial institution” into the ”Application denied by financial institution” class. Lastly, we used SMOTE methods to oversample the minority classes in order to balance the data set.

After doing all these feature engineering work, we split the data into training set and testing set (0.7:0.3). Then we normalize the data using MinMaxScaler.

III. Model Implementation

A. Multi-class Logistic Regression

Logistic Regression is one of the fundamental classification methods. It is often used for binary classifications since Sigmoid function gives probabilities of the two classes between 0 and 1. For multi-class predictions, if we are using a linear solver, this model works in the a one-vs-all way: it trains multiple logistic regression classifiers, one for each of the K classes in the training dataset. If we use a multinomial loss function, then we should use solver like ”sag” or ”lbfgs”. We used ”sag” solver in this case and tested the regularization size from 0.1 to 3.

Logistic Regression is designed to be a very efficient algorithm so it computes very fast. In this model, we used grid search to tune the parameter ”C”, which is the inverse of regularization strength. We used 5-fold cross validation to make our results more reliable.

The best parameter we found using logistic regression is C = 1. The accuracy score achieved is 60% and the confusion matrix is shown below.

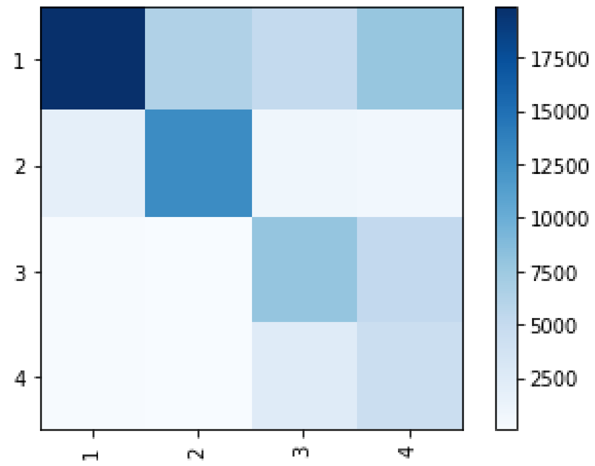

Fig. 6. Confusion Matrix of Logistic Regression

IV. Next Step

Next steps include fitting other classification models like decision tree, random forest, SVM and comparing the performance of each model. Since we are solving a multi-class classification problem, it is inherently reasonable to apply decision tree model into our selected features. Moreover, to reach a possible lower error, we can fit a random forest model. As for linear classifier, SVM could be utilized to make a convincing prediction. Then we can find which model has better performance for the HMDA dataset.

To discuss the relation between race and application approval, we can focus on the information gain or Gini index for tree models. Because the earlier branch node means the bigger information gain or lower Gini index, which means the corresponding feature is more convincing to make classification. For further discussion, we are going to find more accurate statistical characterizations.

As for logistic regression, After fitting the model, it is likely that researchers will want to examine the contribution of individual predictors. To do so, they will examine the regression coefficients. In linear regression, the regression coefficients represent the change in the criterion for each unit change in the predictor. In logistic regression, however, the regression coefficients represent the change in the logit for each unit change in the predictor. Given that the logit is not intuitive, researchers are likely to focus on a predictor’s effect on the exponential function of the regression coefficient – the odds ratio. In linear regression, the significance of a regression coefficient is assessed by computing a t test. In logistic regression, there are several different tests designed to assess the significance of an individual predictor, most notably the likelihood ratio test and the Wald statistic. We can apply those tests to modify the contribution of predictor race.