ORIE 5741 Project Midterm Report
Credit Approval Prediction

Ruifan Chen
rc654

Xinyue Liu
xl533

Yijun Zhu
yz2745

Abstract

It is of great significance for financial institutions to make wise decisions about which individuals or businesses to offer credit to avoid bad debt. The main dataset is home mortgage disclosure action provided by HMDA Platform combined with the dataset of median housing price in counties in NY state. This project applied big data techniques to build machine learning models to make better predictions on whether to offer credit or not to a client. We develop logistic regression, random forest models, together with the L1 regularization approach and PCA techniques to improve the model accuracy. In the end, the project will discuss corresponding limitations and improvements for our models.

1 Exploratory Data Analysis

1.1 Data Characteristics

The main dataset for our project is from the HMDA platform, including our label, home mortgage disclosure action, with 98 explanatory variables from 2018 to 2020 in New York State, which has 1825608 entries in total. Among all the 99 variables, there are 86 categorical variables with 25 ordinal data and 61 nominal data, 13 numerical variables including 2 discrete variables, and 11 continuous variables. These variables focus on the three aspects: 1. financial and demographic background of applicants such as gross annual income, credit score type, ethnicity. 2. demographic information of the county or tract of applicants including tract-population, median-housing-units. 3. Information about the loan such as loan term, loan amount, and interest rate. The second dataset we apply is the median house price data of each county in New York state from 2018 to 2020 from the official website of New York state(https://www.ny.gov/). We merge the two datasets to get a combined one with 100 variables in total.

In addition, we plan to consider the influence of the Covid-19 Pandemic on loan approval cases, so we add one nominal variable: "Covid". It is set to be 1 if the application is in 2020, otherwise 0. In this case, we obtain our final dataset, composed of the label "action-taken" and in total 100 explanatory variables from 2018 to 2020 in New York State.

Considering the aim of the project, the current dataset is messy and large that many variables contain high volume of null values. Thus, in the first step, we would discuss the quality of the dataset and clean data for our further data analysis.

1.2 Data Cleaning

1.2.1 Missing Values

According to figure 1.1, there are more than 30 variables which have a proportion of over 50 percent of NA values. For most variables, "NA" values do not have actual meaning because applicants leave them blank when filling in the application form. Moreover, for many of these variables, the information is repetitively inferred in other variables.

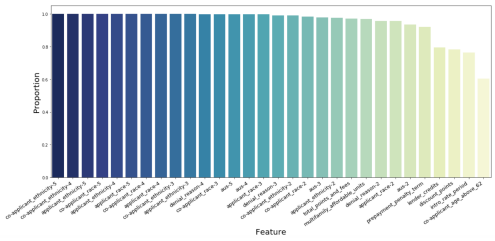

Figure 1.1: Missing values proportion for top 30 variables

To deal with the null values, variables over 50 percent NA values (31 in total) are dropped. For variables containing less than 5 percent of NA values, we delete all the corresponding entries with NA values in corresponding variable column. Lastly, for those variables with 5 to 50 percent NA values, we fill in the NA values using our estimate by random forest models separately.

1.2.2 Outlier Detection

Outliers in the numerical variables may exert a negative influence on our data visualization and analysis. Thus, we apply the Interquartile Range Approach and regard values not in the range of [Lower quartile - 1.5 $\times$ Interquartile range, Upper quartile + 1.5 $\times$ Interquartile range] as outliers, and then delete their corresponding entries.

1.2.3 Data Corruption

To facilitate our further analysis and increase the interpretability of our models, we relabel our target variable "action-taken". Originally, this nominal variable has in total 7 values including Application in progress, Application denied, Application withdrawn, Application approved but not accepted, and so on. In this case, we combine "Application approved but not accepted" and "Purchased loan" as "Approve", combine "Application denied" and "Preapproval request denied" as "Deny", and drop other values'

entries. As a result, we obtain a binary label for our target variable.

1.3 Data Visualization

1.3.1 Correlation Analysis

After we have done the basic data cleaning process, we are able to conduct data visualization to explore relationship among variables, and understand the significance of different explanatory variables in credit approval prediction.
First of all, we split variables into categorical variables and numerical variables. For numerical ones, we conduct a correlation heatmap to explore the relationship between each variable and the label. As we can see from Figure 1.2, the label "Action-taken" is relatively highly correlated with features "loan-term" and "median-housing-price".

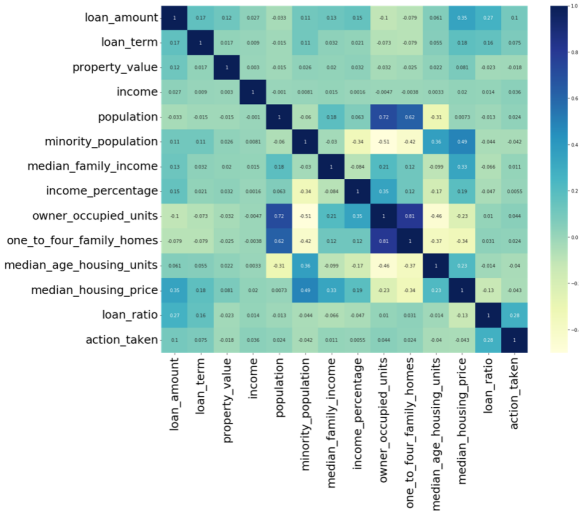

Figure 1.2: Numerical Variable Correlation Analysis

1.3.2 Important Numerical Variable Visualization

According to our findings in Figure 1.2, we consider the two most correlated numerical variables: "loan-term" and "median-housing-price". We create several groups based on value, and plot group bar charts to see their influences on the credit approval rate.
As we can see from Figure 1.3 and Figure 1.4, there is an obvious positive relationship between "loan-term" and "action-taken", and there exists a negative relationship between "median-housing-price" and "action-taken".

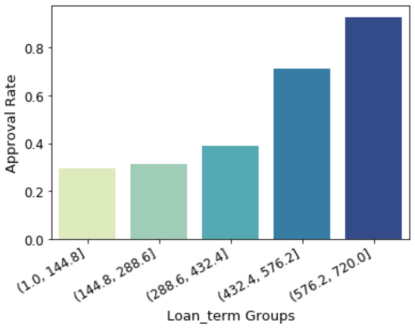

Figure 1.3: Loan-term Variable Group Analysis

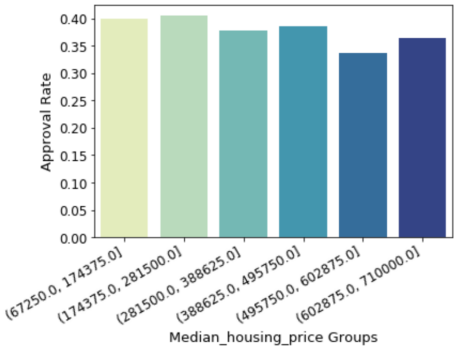

Figure 1.4: Median-housing-price Variable Group Analysis

1.3.3 Important Categorical Variable Visualization

Intuitively, we choose two important categorical variables to draw bar charts below to see the influences of their different levels on the credit approval rate.
According to Figure 1.5 and Figure 1.6, we could find that as the value varies, "credit score" and "sex" have a great impact on our target variable, and we would conduct further analysis on their relationship in the following sections.

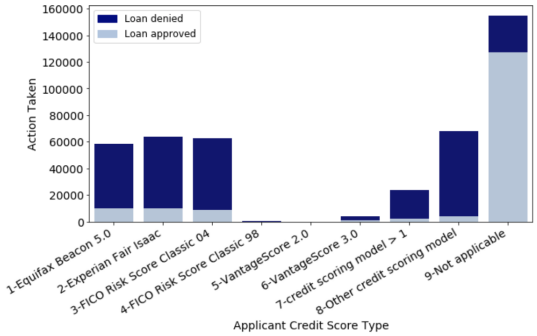

Figure 1.5: Credit-score Variable Analysis

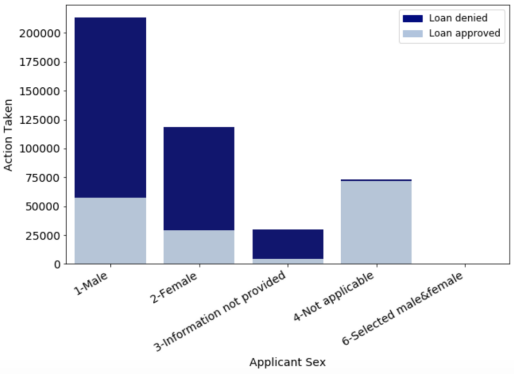

Figure 1.6: Sex Variable Analysis

2 Feature Engineering

2.1 Categorical Variable Encoding

In order to prepare for feature selection and modelling process, we need to complete feature engineering for different kinds of variables. We apply one-hot encoding techniques

for all the nominal variables. As for ordinal variables, we employ Boolean encoding methods, which will generate n-1 dummy variables to represent the information of n-level ordinal variable.

2.2 Numerical Variable Normalization

For numerical variables, we apply the normalization approach to scale all the variables from different range to the similar one with mean = 0 and standard deviation = 1. In this case, we could reduce the negative impact of multicollinearity problem, enhance the interpretability of our models, and make it easy to compare different features to do feature selection.

3 Model Selection

3.1 Feature Selection

3.1.1 Pearson Correlation Approach

In this approach, we consider numerical variables and category variables separately.
For category variables, after we apply the one-hot encoding and boolean encoding methods to nominal as well as ordinal variables, we compute the correlation coefficient between all the dummy variables and our target variable "action-taken".After this, we select the top 20 category dummy variables in terms of the absolute value of their correlation with target variable.
As for numerical variables, similarly, we compute their correlation with target variable and select top 5 numerical variables.

3.1.2 Logistic regression with L1 regularization

Employing logistic regression method with L1 regularization, we plan to make use of the sparsity property of its coefficient due to L1 regularization, to complete the feature selection. We include all the numerical variables as well as dummies from categorical variable encoding in to our logistic regression model, and select corresponding numerical variables and dummy variables whose coefficient in this model is significantly nonzero. As a result, we select 62 predictors containing 1 numerical variable "income" and 61 dummy variables.

3.2 Model Selection

3.2.1 Random Forest Approach

Firstly, we apply random forest approach for our dataset under two feature selection methods. For this modelling, we include 100 trees in the forest, and set up its maximum depth at 50.
As for the first feature selection method, correlation approach, we employ the random the forest model and apply the 10-fold cross-validation to obtain its model accuracy. The average model accuracy is very high to reach a value of 0.9133.
As for the second feature selection method, logistic lasso regression approach, we employ the forest model with similar parameter, and apply the 10-fold cross-validation again. The results turn out that the average model accuracy is 0.8897, which is relatively lower than first feature selection method.

3.2.2 Logistic Regression

Secondly, we use logistic regression model under two feature selection methods separately. For this modelling approach, we include an L1 regularization penalty term.
Similar as random forest, we apply the 10-fold cross-validation to obtain its model accuracy. The average model accuracy is 0.9068 for the correlation approach for feature selection, which is high but a bit lower than random forest model. As for the second feature selection approach, the results turn out that the average model accuracy is 0.9154, which is the highest one among all the combination of models and feature selection methods.
Table 3.1 below summarizes the performance of two models under different feature selection approaches.

Feature Selection Methods	Average Model Accuracy	
	Random Forest	Logistic Regression
Pearson Correlation	0.9133	0.9068
Logistic Regression with L1	0.8897	0.9154

Table 3.1: Model Accuracy Comparison

4 Further Analysis and Improvement

4.1 Model Analysis

According to results in Table 3.1, we find that under two distinctive feature selection scenarios, random forest and logistic regression models may have different performance. Thus, we may conduct deeper analysis on this issue, and try to figure out the rationale behind it, which may be beneficial for the final model selection.

4.2 Improvement

4.2.1 Variable Rebalance

For many variables, the data points are not distributed evenly into different classes so we combine low-occurrence levels to rebalance these variables to obss potential improvements.

4.2.2 Interaction Terms

Based on our feature selection, we plan to consider the two-dimension interaction terms of all the features to increase the model interpretability and prediction accuracy. Moreover, we will consider further feature selection methods after including all the interaction terms, trying to further explore relationship between label and explanatory variables and construct a better prediction model through cross-validation.

4.2.3 Overfitting Issue Analysis

After including the interaction terms, there may be a great many predictors in our model, which may lead to overfitting problem. Therefore, we currently apply PCA approach to reduce the dimension of our features and plan to explore more methods to improve our model prediction accuracy.