The Home Mortgage Disclosure Act (HMDA) Data Exploration

ORIE 5741 Project Final Report

Chengyuan Yang (cy362); Zhengyi Sui (zs359); Zihao Fu (zf223)

Cornell University

12/05/2021

Abstract

Identifying risk and recognizing profit are important to loan companies. An accurate and robust model plays a key role in the inspection process. This report establishes a series of prediction models to assess applications' credit score and support making approval decisions. Gradient boosting decision tree(GBDT) and random forest(RF) turn out to outperform other models in recall, accuracy, precision and F1-score. Correspondingly, GBDT is regarded as the relatively better predictor for loan approval decision-making.

Background

Identifying risk and making profit are eternal topics for enterprise, especially for the financial firms that live on loan interests from credible customers. When an applicant applies for a credit card or a loan, the bank should decide on whether to approve the application or not. This is crucial to banks because they are businesses that live on such loans and the revenue from the interests. For banks, manual inspections and machine filtering are two ways to assess an application in order to make the decision. Correspondingly, accurately identifying potential default risk and rejecting the application, and recognizing stable profit chances and approving the application, are important to the overall revenue and profits to the banks.

Introduction

A robust and scalable model to accurately assess each applicant's credit score is necessary for decision-making. This project established a series of models to score each application's loan score based on the loan details as well as applicant's demographic information and financial records. More concretely, with the home mortgage disclosure act dataset, which data is big and messy, data preprocessing methods to deal with the missing values and outliers are adopted to clean the data. Furthermore, logistic regression, principal component analysis, bagging and boosting models are fitted and compared based on their performances. It turns out that gradient boosting decision tree(GBDT) and random forest(RF) both outperformed logistic regression significantly. Ideally, the model could enable quick determination of whether the application should be approved or not.

Data Description

To fulfill our goal to develop an effective prediction model on credible home mortgage applicants, the home mortgage disclosure act(HMDA) dataset provided by the Federal Financial Institutions Examination Council(FFIEC) is used by our group. FFIEC began to collect HMDA data around 2017 and financial institutions are required to 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.

The project adopted the HMDA data of New York state in 2020 to train and evaluate the prediction models. In general, the original dataset incorporates 99 features involving loan details and applicant's demographic information, historical financial records and current financial status, with numerical data, boolean data and categorical data. The dataset contains more than 700,000 samples, which is sufficient for the model training.

Data Processing and Cleaning

The dataset is big and messy in that missing values and outliers are obvious after exploring the distribution and summary statistics. Missing values are important to the modeling as the rare and imbalance information it contains could impact the prediction significantly. Fixing outliers are also crucial as such far-away data points could drag the model to extreme bias and calibrating outliers enables the model to be more stable.

Missing Values

By counting the missing values of each column, some features are rarely missed, while some features have over half missing values. For features with over 50% missing values, the features are removed from the dataset as the information these features provide is rare and imbalanced. For those with fewer missing values, certain methods are applied based on the feature distribution. For numerical data, missing values are replaced with mean or median based on the skewness of the feature. For categorical data, missing values are replaced with the majority vote.

Outliers

In order to check the outliers within the dataset, visualized and quantitative methods are used. Box plots are applied to visualize the outliers in sample data. Summary statistics including min, lower quartile, median, mean, upper quartile, 95 percentile, 99 percentile and max are computed for each feature. For those features with extreme skewness possibly due to magnitude, logarithm transformation is implemented to get a normal-like distribution. For other features, outliers are replaced with 95% quantile to mitigate the effect of outliers and preserve most characteristics.

Encoding

For the model, the response variable is “action_taken”. This feature has values 1,2, and 3 that indicate early acceptance, acceptance, and denial. The first two values are combined to be “approved”, indicated by 0; and the other condition to be 1 as “denied”. Meanwhile, the feature series “denial_reason” are deleted as they are exactly the hidden information about “action_taken”. For all other 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.

Data Balancing

From *figure1* below, it is clear to see that there is a huge difference between two classes of the response variable (action_taken). There are about 400,000 records for accepted applications but only 100,000 records for denied applications. Such an imbalance pattern would heavily affect the accuracy of the model, the model will tend to have high accuracy with low recall, since there are much more negative cases than positive cases. So our group used the oversampling to random duplicate the records of minority classes.

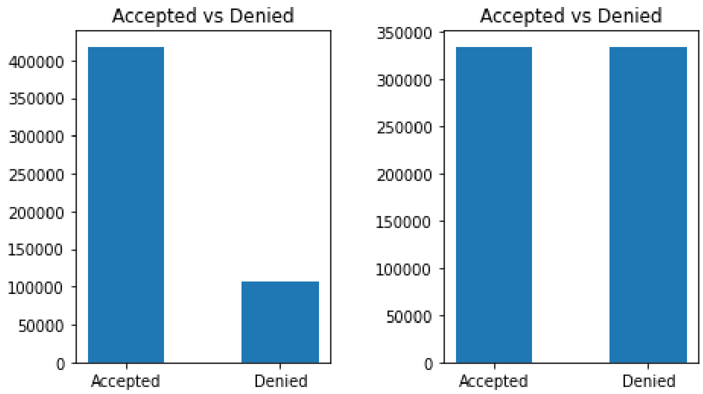


Figure 1

Modeling & Results

Performance metrics: in this project, two metrics are mainly used to determine whether the model is solid for the classification and prediction problem, which are accuracy and recall score.

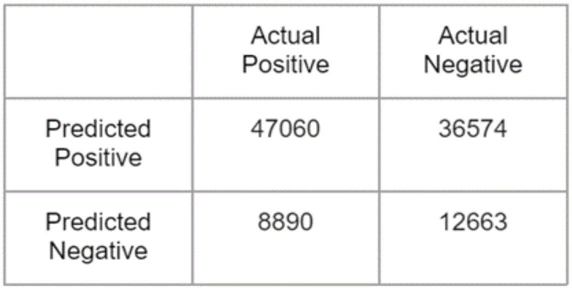
- **Accuracy**: Accuracy of the model indicates the overall accuracy of the model when predicting whether the application should be accepted. Hence, the accuracy would be important for us to determine whether the model could correctly predict each case. A high accuracy would show that the model has a high quality.
- **Recall score**: the recall score shows the percentage of negative cases the model could correctly identify. So, a high recall score for the model would indicate that the model could precisely find the application that should be denied. This would be important for this project, since the overall goal is to help the company to avoid the loan application that might not be paid back.

Data modeling: four models are established for the choice of predictor, including logistic regression, principal component logistic regression, gradient boosting tree and random forest.

- **Logistic regression**:

  Logistic regression is the simplest model and could give us a baseline performance for this problem. Basically, the logistic regression would generate a function which could calculate the probability of the case to be positive. Since 0.5 is chosen as the threshold, if the probability is higher than 50%, then it can be categorized as positive. If it is lower than 50%, it can be categorized as negative.

As the *figure 2* shows, accuracy and the recall score are generated. For this model the accuracy is only about 56.7% and the recall score about 58.7%. Such performance means that this model could only correctly classify less than 60% cases. This is not ideal for targeted problems.

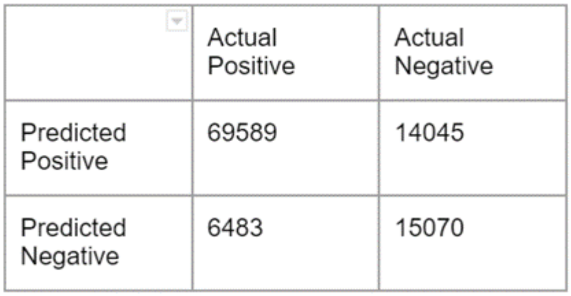


	Actual Positive	Actual Negative
Predicted Positive	47060	36574
Predicted Negative	8890	12663

figure 2

- **PCA and logistic regression:**

Principal component analysis(PCA) could help us reduce the dimensions of the data set while containing the information of the data set. In addition, it could improve the performance of this model. With the help of PCA, the dimension of the PCA model was reduced from 437 to 267. Then a logistic model is developed on the result data set to see whether there is improvement.



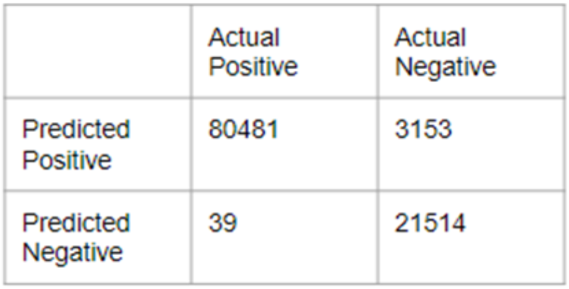
	Actual Positive	Actual Negative
Predicted Positive	69589	14045
Predicted Negative	6483	15070

figure 3

From the confusion matrix of *figure 3*, the results show that the accuracy has increased from 56.7% to 80.5% and the recall score has increased from 58.7% to 69.9%. Though this is a large increase in the performance, it is still not good enough for the ideal classification model.

- **Gradient Boosting Decision Tree**

Since the performance of the logistic model is not ideal, a decision tree model is developed. However, the single decision tree usually would generate a high variance result. So eventually gradient boosting trees and random forests are integrated. The boosting tree generally fits a tree on the dataset and then fits another tree on the residual of the previous tree. At last, it will sum up all the trees to compute a result.



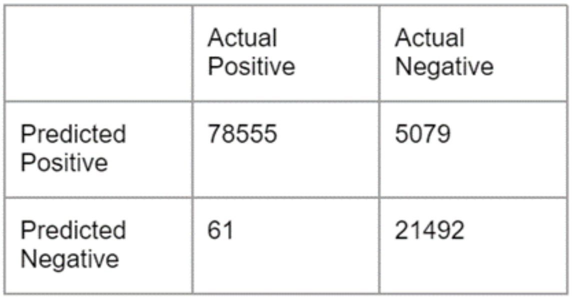
	Actual Positive	Actual Negative
Predicted Positive	80481	3153
Predicted Negative	39	21514

figure 4

From the *figure 4*, it can be concluded that the accuracy reached about 96.9% and the recall is about 99.8%. This is a huge improvement from the logistic model. Such performance has already met the requirement of our target problem. It could correctly identify 99.8% cases that should not be approved which is really close to 100%.

- **Random forest:**

Though the boosting tree already generates great performance, another random forest model on the data set is developed to see whether there is more improvement. Random forest is a special case of bagging trees, in which each tree is fitted in a subset of the whole data set and then averages all the trees as the result of the whole model.



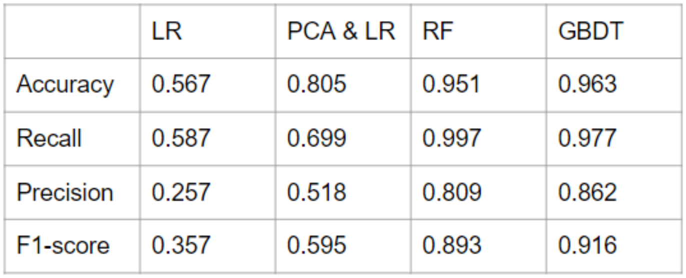
	Actual Positive	Actual Negative
Predicted Positive	78555	5079
Predicted Negative	61	21492

figure 5

From the *figure5*, accuracy is about 95.7% and the recall score is about 95%. It is clear that this model does not have much improvement than the boosting tree, though it also generates pretty good performance.

- **Summary table:**

Figure 6 displays the overall performance of all models and this table is used as reference to select the model with best performance.



	LR	PCA & LR	RF	GBDT
Accuracy	0.567	0.805	0.951	0.963
Recall	0.587	0.699	0.997	0.977
Precision	0.257	0.518	0.809	0.862
F1-score	0.357	0.595	0.893	0.916

figure 6

Discussion

Selected Model & Feature Importance

Among all models applied to select important features and make predictions, the gradient boosting decision tree model performs the best, as it has the greatest accuracy score and recall value. The high accuracy score means that the model predicted value fits the real value very well. High recall score indicates that there is a higher chance if the applicants should be rejected, the model will predict that the applicants would be rejected.

Figure 7 displays 10 most important features for the selected model. Among them, whether it is a high-cost mortgage, interest rate, whether the purchase type is applicable, total loan costs, and whether the mortgage value is applicable are the most important features.

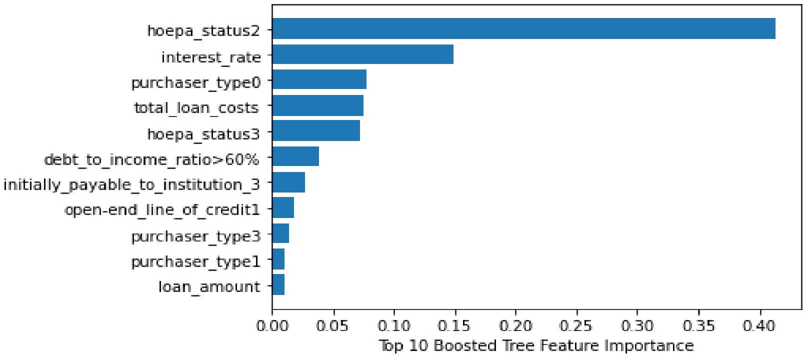


figure 6

Limitation of Data Science

In this project, complex machine learning models have been applied. Even though the performance of the model is eligible, there are some limitations that need consideration.

The first and most important is there would be some human factors issues with the data and the way how FFIEC collects the data. They collect the data by letting companies upload their credit approval record. It might have detrimental effects if some of the financial institutions lied when they uploaded their information. Whether there is fraudulent data will affect the ultimate credibility of the model.

Fairness Issue

Through the entire model development process, the requirements of protected attributes have been reached by not selecting gender, ethnicity, race, and some other similar features into models. However, there are some other fairness issues.

One limitation is the individual fairness problem. It is very difficult to define the similarity of targets, mortgage applicants. And it is one of the limitations of data science about failure to define similarity. Therefore, the model is developed carefully and tries to minimize the effect of fairness issues.

Direction for future improvement

During the data exploration process, one problem is data bias. The label of the current dataset is whether approved or not. However, the real label for bad loans should be the historical default-or-not data. Such differences cause data bias. Besides, it appears that the model developed will be a little optimistic, because the actual default data only contains users that were approved, who were considered as the "good" applicants.

The bias discussed above will have a bad effect on the trained model. Therefore, it is worth discussion and development for the next stage to fix the data bias problem in the future. Therefore, solutions to fix the data bias problem, such as the reject inference methods, are worth discussion and development for the next stage.

Conclusion

Gradient Boosting Decision Tree model has been selected as the ultimate model for the credit approval classification. It has relateively best recall score and is considered as a reliable model for clients to use to balance between profit and risk. With gradient boosting decision tree model, an applicant with a not high cost mortgage, low loan interest rate, with type of the entity applicable, low loan cost, and low mortgage value will have a higher chance to be approved.

The next step of the project is trying to avoid the negative effect brought by data bias. This will avoid the prediction model being too optimistic.