**I．Introduction**

It is normal in the US that people submit a mortgage application to get financial support when they buy houses. Basically the application process is about submitting documents to prove the applicant is a credible person who deserves a money loan. I would like to prediction the result of respondent's mortgage applications, which is the variable `action\_taken`(a categrocial variable) in the dataset. The label variable `action\_taken` includes three kinds of application results: approved, denied and withdrawn. I believe it is meaningful if we have a model capable of predicting mortgage applications' result since people could estimate the result based on some property characteristics, loan information, demographics data, and they could make more preparations in advance. The application process is complicated and it is not 100% promised that people could get application approved. Therefore knowing the application status ahead of time could be beneficial for applicants to have other plans. Also, it is meaningful for to have a model helpful for economists to research which factors associate with an individuals’ credit level and to forecast the proportion of applicants who get the applications approved.

**II. Data Wrangling and Cleaning**

The name of the dataset I used is called ‘hmda\_2017\_il\_data\_all\_record’, which is derived from the website of consumer financial protection bureau. The dataset includes all the mortgage application records in Illinois, 2017. It is a single dataset so I did not make any merge. I directly dropped the observations having any missing values since I have enough records to train the model. I did some recode for some vary iables. For example, there is a dummy variable called `preapproval`, initially the non-preapproval level is coded as 2 but I recoded it 0. I also converted the type of my label variable `action\_taken`, initially it have three integers 0,1, 2 but I coded them as strings: ‘approved’, ‘denied’ and ‘withdrawn’.

**III. EDA**

Figure 1: The figure of label variable `action\_taken`:

图表, 条形图

描述已自动生成

Given my label variable is a categorical variable, I showed a histogram to how the variable distribute over 3 categories: approved, denied, and withdrawn. Based on the above figure, we could see the distribution is basically balanced and there is no need to make transformation for the label variable.

Figure 2:

图表, 条形图

描述已自动生成

The above graph is a barplot of applicant income vs applicant result over sex. It showed that the applicant whose applications are approved have higher income than those whose applications are denied. Thus application\_income should be included as x variable into my model. Also, female(applicant\_sex = 0) has lower income than male under each application result so we may need to include sex and income:sex interaction term into my model.

Figure 3:

图示

低可信度描述已自动生成

The above graph is a violin plot showed the distribution of loan amount over each application result category. It showed that the distribution of loan amount is very right-skewed so we may need to do some log transformation for this variable.

**IV. Inference**

**表格

描述已自动生成** **表格

描述已自动生成**

The final alpha chosen is 0.01, and there are 4 features removed after regularization.

The interpretation for action = 0:

one standard deviation increase in *loan\_amount* is associated with an increase in the probability of `Denied` relative to `Approved` by 0.05,

one standard deviation increase in *Black* is associated with an increase in the probability of `Denied` relative to `Approved` by 0.11,

one standard deviation increase in *Black* is associated with an decrease in the probability of `Denied` relative to `Approved` by 0.05,

The interpretation for action = 1:

one standard deviation increase in *preapproval* is associated with an decrease in the probability of `Withdrwan` relative to `Approved` by 0.084,

one standard deviation increase in *applicant\_income* is associated with an decrease in the probability of `Withdrawn` relative to `Approved` by 0.0064,

one standard deviation increase in *Family\_income* is associated with an decrease in the probability of `Withdrawn` relative to `Approved` by 0.027.

**V. Prediction**

The first model is random forest model with n\_estimators = 500, random\_state = 432 and max\_feature = ‘sqrt’.

The second model is a stacking model with ‘saga’ solver and 5 folders, and it has three estimators: The first one is a Naïve-Bayes model; the second one is a random forest model n\_estimators = 500,random\_state = 432, max\_features = 'sqrt'; the third model is adaboosting model with random\_state = 490,n\_estimators =50,learning\_rate = 0.5, and the max\_depth = 1.

**VI. Comparison**

The confusion matrix for random forest model: The confusion matrix for stacking model:

**图表

描述已自动生成** 图表

描述已自动生成

The accuracy of multinomial logistics regression model is 56.81%, the accuracy of random forest model is 56.29%, and the accuracy of stacking model is 57.27%.

The stacking model performed best out of three models since it incorporated the advantages of the three stacked models. It is easy to interpret the multinomial logistics regression model but both random forest and stacking model are not interpretable.

**VII. Conclusion**

My project is focused on predicting the result of mortgage application, which includes three categories(approved, denied, withdrawn). The data used for model building are 19460 records of mortgage applicants with their demographics data, loan information, property characteristics. I evaluated three models: the multinomial regression model, the random forest model and the stacking model. The best performing model is the stacking model. For the project’s further improvement, I may try to build a variety of stacking model based on different sets of models to find the stacking model with highest accuracy.