Credit Approve Using Machine Learning

Introduction

One-of the largest business sectors for any financial institution is mortgage. It is very important to determine which applicant to approve and which to reject. This decision often affects the company's revenue performance and sometimes even determines the company's life and death.

However, it would be difficult for every business to investigate every aspect of the credit record of an individual customer in today's society. So, our goal is to use a few features to create a model for the decision-making process. In our vision, the model should help to filter out weak and fraudulent applicants and therefore increase the investment returns for the company and reduce the occurrence of bad debts.

Dataset

We now have our data from FFIEC named 21st Mortgage Corporation in California. Since it is unrealistic to evaluate every feature, we filtered out many variables that were considered to have less impact on loan approval. First, we removed the variables which are meaningless to our predictor. For example, the variable 'lei' is useless since it is just a series number. Second, we replaced missing values to 0 so that it will not affect our future model evaluation.

Variables

After cleaning our dataset, here are some important features that we may want to explore first.

Applicant_age: It is the age of the applicant or borrower. The values are the group range and we want to know how each group could affect the predictor.

Derived_sex: It is single aggregated sex categorization derived from applicant/borrower and co-applicant/co-borrower sex fields.

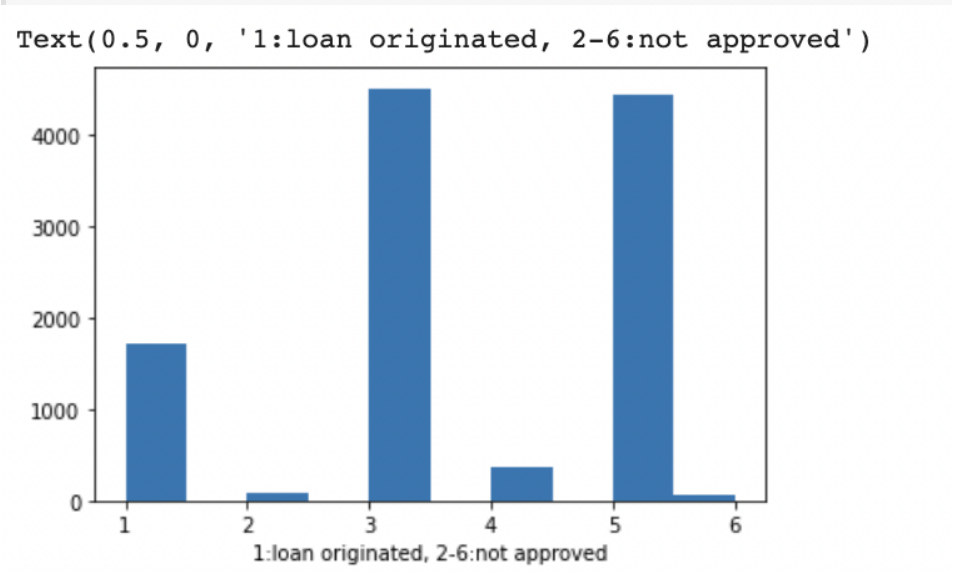
Action_taken: The action taken on the covered loan or application. It is the categorical variable including values from 1 to 6 which represent different loan status for each person.

loan_type：The type of covered loan or application. It is a categorical variable including values 1 to 4. We will transform it into one-hot vectors.

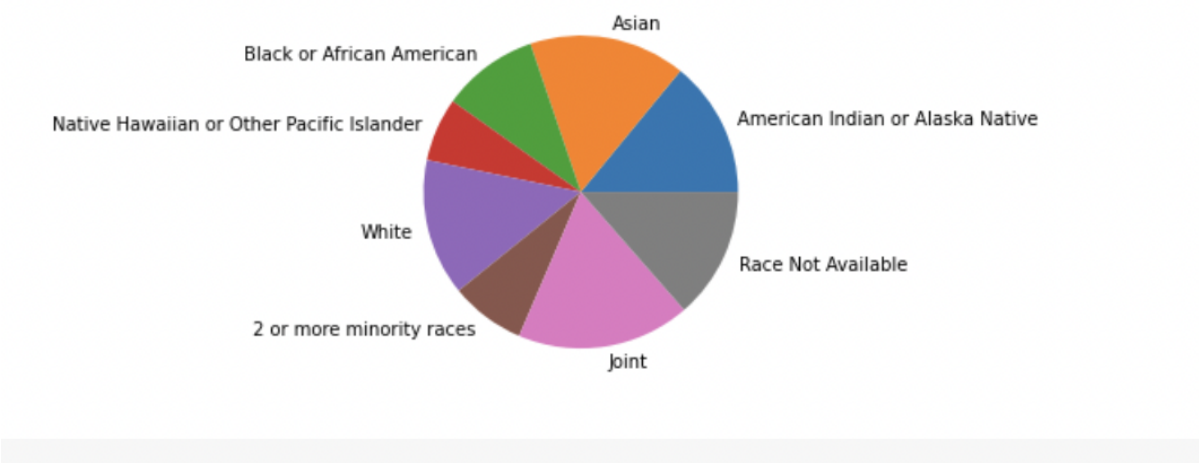
reverse_mortgage: Whether the covered loan or application is for a reverse mortgage. It is the boolean variable which 1 represents 'Yes' and 2 represents 'No'.

Data visualization

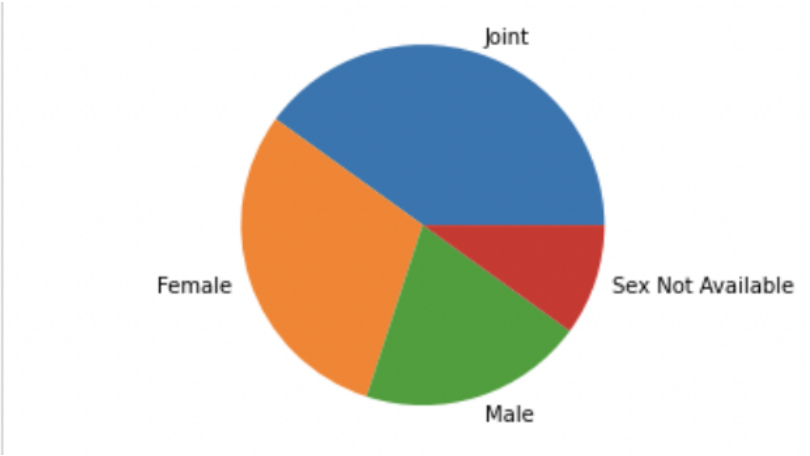
Since our purpose is to model whether a given credit application should be approved, we choose **action_taken** as our predictor, which is a categorical variable with indicators 1 to 5 where 1 means ‘loan originated’, 2 means ‘application approved but not accepted’, 3 means ‘application denied’, 4 means ‘application withdrawn by applicant’, 5 means ‘file closed for incompleteness’. As we can see from the histogram, the number of credit approved is around 1800 cases. For all the disapproved cases(2 - 5), the main causations distribute in 3 - ‘application denied’ and 5 - ‘file closed for incompleteness’.



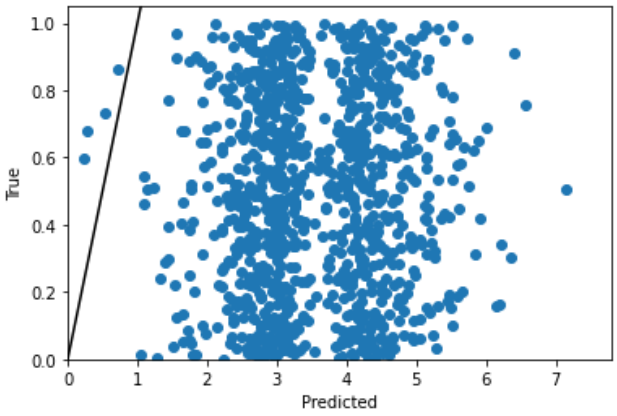
For the features, we did a preliminary exploratory analysis on the effect of sex and race on credit approval rate. As we can see from the first pie chart, the highest approval rates are among Asian, White, American Indian / Alaska Native and Joint. Black / African American, Native Hawaiian / Other Pacific Islander have relatively low approval rate.



From the second pie chart, we can see that Joint and female have higher approval rate while male and sex not available have much lower approval rate. These pie charts show that race and sex are two very influential features o include in our model.



Below is the plotting by boolean variables. From the plotting, we find that boolean variables may not be very informative to the predictor. We still need further investigation in the process of the project.



Next Step

Since our predictor is the categorical variable, one-hot would be an important method to our project. We will transform the data into one-hot vectors as our next step. Also, we need to get more data to have more features to explore. We will add more to our input space to see whether we should keep them as significant ones or reject them to prevent overfitting.