- Are minorities discriminated against in lending?
  Judging from the result of analysis, yes.
- What is your hypothesis before running experiments?
  Minorities are not discriminated against lending.
- Which minorities did you chose to analyze?
  I considered White Non-Hispanic-or-Latino applicants as majorities and other races and ethnicities as minorities and analyzed the minorities as a whole.
- What was the analytical process that you chose to use?
  For preprocessing, I cleaned data and did some transformations on some variables. Then I selected variables has a large impact on the result of application using Random Forest. Then I control the important variables to be approximately the same between minority group and majority group and compared approval rates of both group.
- What metrics/variables did you use in this data to prove or disprove? Why?
  Tract-related variables, applicant income, loan amount. Because they are considered most important variables by Random Forest model.
- What metrics/variables did you Not use in this data to prove or disprove? Why?
  Variables with too many null values and variables which do not make much sense in modeling.
- Are these variables explanatory or correlated or both?
  Some are explanatory and some are correlated.
- What factors did you control for? Did you use other data sets? Why? Why not?
  Tracts, loan amount and applicant income.
  No. Because it is hard to find relevant datasets to be aligned with so many records and given dataset are sufficient in predicting application result.
- Is there a Geographic effect?
  Yes. The discrimination in Texas is more severe that in California
- Are the effects you are describing getting worst or better over time? Was your hypothesis supported by your analysis? Why?
  Discrimination against minority is more severe in 2013 than in 2010.
  My hypothesis was not supported by my analysis because controlling important variables to be approximately the same, there is still a significant difference between application approval rates of minority group and non-minority group.
- Should you use HMDA data to improve mortgage prepayment models? How would you do it?
  No because the data does not include applicants' behaviors.

A detailed report

Data selection

The question is whether minorities are discriminated against lending. When it comes to racial and ethnical discrimination issues, people tend to associate them with political stands, such as being Democrat or Republican. So, I decided to choose the data of California and Texas as the dataset for analysis as they are each a Blue state and a Red state, which helps get a more comprehensive picture of the issue. Also, these two states are with comparable economy status – both of them are among the states with largest GDP in United States, which helps control the variable of economy status. Furthermore, the volumes of data from these two states are among the largest, which is expected to help yield more accurate analysis.

In order to see if there is a change in situation with time, I selected data of 2010 and 2013.

According the question, I am looking at relation between applicant's race and ethnicity and how likely his or her application will be approved. So, the data only need to include applications with definite results (approved or denied) and applicants who can be identified as minority or not. Based on these criteria, I only selected records whose action is among Loan originated, Loan purchased by the institution, Application approved but not accepted, and Loan Application denied by financial institution, where the first three actions are considered approval and the last action is considered denial. Records whose Sex, Race and Ethnicity variables are Not applicable were also excluded because being Not applicable means the applicant is not a natural person, which is not relevant to this question.

General Overview

First, I got an overview of minority's application success rate by directly looking at the ratios of denied cases to all cases of minorities and non-minorities. To get this ratio from provided data, data recoding is needed. That is, based on current data, labeling whether the case of an applicant is a minority and whether the case is approved.

In the data, the column of "applicant_race_name_1" indicates the applicant's race and the column of "applicant_ethnicity_name" indicates the applicant's ethnicity. From these two columns, it can be decided whether an applicant is a minority or not. Here applicants whose race is White and ethnicity is "Not Hispanic or Latino" are defined as non-minorities and applicants with other combinations of race and ethnicity as minorities. Note there are missing values in both columns. For applicants whose race is not White, no matter what is their ethnicity, they are minorities. However, for applicants whose race is White and ethnicity information is missing, they are labeled as unknown. Also, for applicants whose ethnicity is "Hispanic or Latino", they as minorities no matter what their races are. After labeling, a new variable indicating whether the applicant is minority or not is created.

For the result of a case, records are labeled as Approved or Denied according to the criteria mentioned above and a new variable called Minority to indicating the result is created

With the data labeled, I then computed the ratios of different groups. Where

Ratio of Deny = Number of Denied / (Number of Approved + Number of Denied)

Minority or Not	Result	Number	Ratio of Deny
Yes	Approved	1358351	20.85%
	Denied	357887	
No	Approved	2344948	15.32%
	Denied	424177	

It can be observed that cases with minority applicants are more likely to be denied than those with non-minority applicants.

By selecting data, I calculated data from California and Texas separately.

Here is a summary from data of California:

Minority or Not	Result	Number	Ratio of Deny
Yes	Approved	935520	18.40%
	Denied	210903	
No	Approved	1428651	15.23%
	Denied	256664	

Then let's take a look at the data of Texas:

Minority or Not	Result	Number	Ratio of Deny
Yes	Approved	422831	25.80%
	Denied	146984	
No	Approved	916297	15.46%
	Denied	167513	

By comparing the data from Texas with that from California, it is clear that the success rate for minorities to obtain loans differentiate between two states while for non-minorities, there is no such significant geographic effect. In both states, minorities are less likely to get approved. As people may have expected, in Texas, minorities are less likely to get approved than in California.

However, it is still uncertain that minorities are discriminated against lending and such discrimination is more severe in Texas as it is a Red State. There can be some factors common for minorities but not so common for non-minorities that keep them from getting approved. To

determine if such factors exists, I need to find out the most important variables that affects a case's success and compare minorities and non-minorities from perspective of these variables.

Dealing with missing values

As mentioned above, when recoding data, for some records, whether the applicant is a minority and whether the case was approved is unknown. Records of which such information are not provided should be excluded.

After excluding the records above, the total number of records is 4485363. In the dataset, there are some columns that contains over or nearly 4 million null values, such as rate, denial reason and so on. For such columns that contains too many null values, they should be excluded from analysis because they don't provide much useful information.

Also, there are over 16000 null values in each of location– related variables: 'tract_to_msamd_income', 'population',' minorituy_population', 'number_of_owner_occupied_units', 'number_of_1_to_4_family_units', and 'number_of_owner_occupied_units'. When looking at the distribution of these null values, I found that the number of records containing at least one null values in one of these columns, which is fewer than 0.5% of number of all records. So, I decided to exclude these records from analysis.

Data processing and feature engineering for modeling

To improve the performance of model, variables that do not make much sense in modeling need to be excluded, such as 'respondent_id' and 'sequence_number'. Also, since the column of 'msamd_name' and 'census_tract_number' contains so many values, and other location-related variables, such as population and median income, are basically dependent on the name of MSA/MD and tract number, so the MSA/MD names and tract census numbers can be excluded. 'county_name' is similar to 'msamd_name' so it should also be excluded.

Only one of 'state_name' and state_abbr' need to be kept. Here I kept the abbreviate as it takes less space. This applies the same to the agency.

'purchaser_type_name' also should be excluded because its value indicates whether an application was approved or not to some extent. For example, if an application was rejected, the 'purchaser_type_name' is 'Loan was not originated or was not sold in calendar year covered by register'.

As mentioned above, to prove that minorities are not discriminated against lending, I need to find the factors that play important roles in getting approved and differentiate between minorities and non-minorities. Based on this assumption, if such factors exist, they must be dependent on or highly correlated with race and ethnicity variables. Provided that values dependent to each other can undermine the performance of models, race and ethnicity are excluded from modeling.

Since I am looking at minority-related issues, the ratio of minority population makes more sense than the absolute number of minorities in an area. So I created a new variable called 'minority_ratio', which is the ratio of minority population to the total population of the tract. Regarding the economic status of the tract where the property is located, instead of 'tract_to_msamd_income', the percentage of the median family income for the tract compared to the median family income for the MSA/MD, the absolute value of median income of the tract makes more sense. So I created another new value called 'tract_median_income', which is calculated by 'tract_to_msamd_income' * 'hud_median_family_income'. Then 'tract_to_msamd_income' and 'hud_median_family_income' are excluded to avoid dependency among variables.

Modeling

Random forest is used to find the most important features for predicting whether an application will be approved and correlation matrix is used to find out a feature has a positive or negative impact on the result of the application.

According to the result of random forest model, it can be observed that the most important features affecting application's result are applicant's income and the amount of loan, followed by some tract-related variables, including 'tract_median_income', 'minority_ratio', 'number_of_owner_occupied_units', 'number_of_1_to_4_family_units' and 'hud_median_family_income'. It can be concluded that the economic status and demographic status of where the property is located are also affecting the success rate of a loan application. Following amount of loan are income and tract-related variables are lien status, lien of the mortgage and loan purposes. These variables, along variables indicating the race and ethnicity of the applicant, are included for further analysis.

Further analysis and controlling

By looking at correlation matrix generated from included variables, I can see which variables' impact on result and their relations with Minority variable. I found that the higher the median income of the tract where property is located is, the more likely the application will be approved, while properties of minorities' applications tend to have lower median income. This goes same with number of owner-occupied units of the tract, number of 1-to-4 family units of the tract, loan amount and the applicant's income. If the application's property is located in tract with lower ratio of minority population, it tends to be approved, while properties of minorities' applications tend to locate in tracts with higher ratio of minority population.

To find out whether minorities are discriminated against lending, I decided to control the above variables to be approximately the same between minorities and non-minorities and calculate the respective approval rates of minority and non-minority applicants.

To control the tract-related variables, I simply selected records from the same tract. To control the loan amount and income to be the same, I selected a subset of all records where based on

T-test, the means of income and loan amount of minority group and non-minority group are approximately equal, respectively. Given that for most tracts, number of minority applicants are much smaller than the number of non-minority applicants, the goal basically to select a subset of non-minority applicants whose average income and loan amount are approximately equal to those of minority applicants in the tract. I did this by sorting the applicants by income and loan amount and perform T-test on both mean income and mean loan amount (since the distribution of income and loan amount is skewed, I actually performed T-test on logarithm of them). If the null hypothesis of equal means cannot be rejected, it is considered that they are approximately equal. If based on T-test, the null hypothesis of equal means is rejected, check if the mean income of selected non-minority group is larger or smaller than that of minority group. If it is larger, exclude the record with largest income from selected non-minority group, and if it is smaller, exclude the record with smallest income from selected non-minority group. Repeat these steps until the two groups are considered approximately equal on both income and loan amount. Then calculate and compare the approval rate of minority applicants and non-minority applicants to see if they are significantly different. To ensure accuracy, the remaining records of each group should be more than 30. Tracts with no more than 30 records on each group after multiple times of excluding will not be considered in analysis. With most important variables controlled to be almost the same, if there is still significant difference between approval rate of minority applicants and non-minority applicants, it can be considered minorities are discriminated against lending in this tract.

I repeated the above process on the 4000 tracts with largest number or records, which include 2459578 records. It shows that even with tract, loan amount and applicant income controlled approximately the same, in 77.75% of the 4000 tracts, minority applicants are discriminated against lending. In Texas, minorities are discriminated in 87.13% of 2000 largest tracts. In California, 74.3%. In 2010, the number is 72% and in 2013 it goes up to 78%.

Conclusion

In 2010 and 2013, minorities are discriminated against lending. Such discrimination is worse in Texas than in California. The situation became worse from 2010 to 2013.