Mortgages in the US

Final Report

Dina El-Kholy - 201601463
Amr Mohamed - 201500358
Abdelmoez Elsaadany - 201500438

I. Problem description

A mortgage is a kind of loan in which a bank or any financial institution lends an applicant an amount of money that he/she will repay to the institution later with an added interest. The presence of different races and religions in the US led to **mortgage discremination** that started in the 1930s. It is believed that this discremenation is still present till today. **Home Mortgage Disclosure Act (HMDA)** is a US federal law that enforced banks to collect and share their mortgages' data to the public in order to ensure they are not discriminating or misusing their money. In this project, we analyzed the HMDA public dataset [1] (collected in 2014) to investigate the fairness of the mortgages approval criteria.

II. Assumptions & Analysis

While working on the dataset, we made a few assumptions:

1. There were eight actions that a financial institution can do. We assumed that the actions, "*loan originated*" and "*Loan purchased by the institution*" correspond to a loan approval. We assumed that the actions, "-*Preapproval request denied by financial institution*", "*Preapproval request approved but not accepted*" and "*Application denied by financial institution*", correspond to a loan denial. We also neglected the two actions, "*Application withdrawn by applicant*" and "*File closed for incompleteness*".
2. We neglected the co-applicant information as they did not add new information. So, the analysis focused on the main applicant.

Most of the statistics were made following the following methodology:

1. The approved applicants' data and the rejected applicants' data are saved to two different dataframes.
2. We then choose an attribute and calculate its percentage of approval/denial. Hence, we know how it affects the selection process.

III. Final pipeline of the solution

1. Data Preprocessing & Cleaning

- We downloaded the data in [1] and used the first 9M rows only resulting in a dataset of size 5GB.
- Some columns were mostly empty; for example, the co-applicants' information (when there is more than one co-applicant). These columns were removed. These columns were null as the financial institution is not obliged to add them.
- Some rows contained null values; hence, they were also removed.
- The cleaned data contains 6,769,112 rows and its size is nearly 4GB.

2. Data Analysis

- We used Apache Spark in calculating statistics about the data. We investigated the relation between the percentage of approval and the gender, race, and ethnicity of the applicant.
- We also checked the relation between the percentage of approval and the location of the property, the loan type, the loan purpose,...etc.

3. Machine learning model

- We used Spark ML to predict the approval or denial of loans. We fed all the available information in the data to the model and checked the result of three models, decision trees, logistic regression and naive bayes models.

4. Recommendation

- We implemented something similar to a recommendation system. An applicant will input his information (similar to that found in the data); our system will feed that information to our ML model. If the model predicts an approval, we will notify the applicant to proceed. If the model predicts a denial, we will suggest some modifications that the applicant can change in order for his application to be approved.

IV. Trials made, but not included

- We tried decreasing the number of features fed to the model. However, this did not improve the AUC.

- We tried using one hot encoding on the categorical data; however, the model took too long (~30 mins); so, we stopped it.
- We tried using a parameter grid to tune the model's parameters; however, the previous issue happened again and we stopped the model.
- More statistical relations can be found in the project's notebook.

V. Results and evaluation

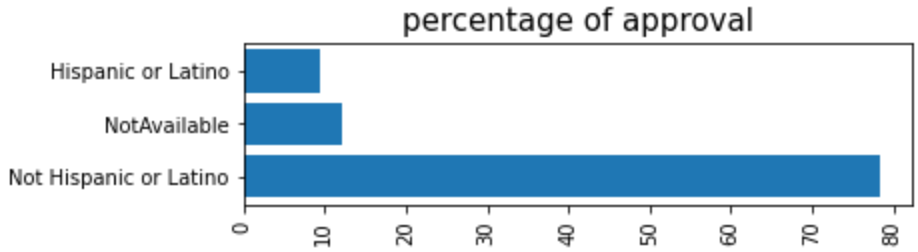
a) Machine learning Model Results:

We tried three different machine learning models namely; decision trees, logistic regression and naive bayes models. The decision trees model achieved the highest AUC among them with an AUC of 0.8506.

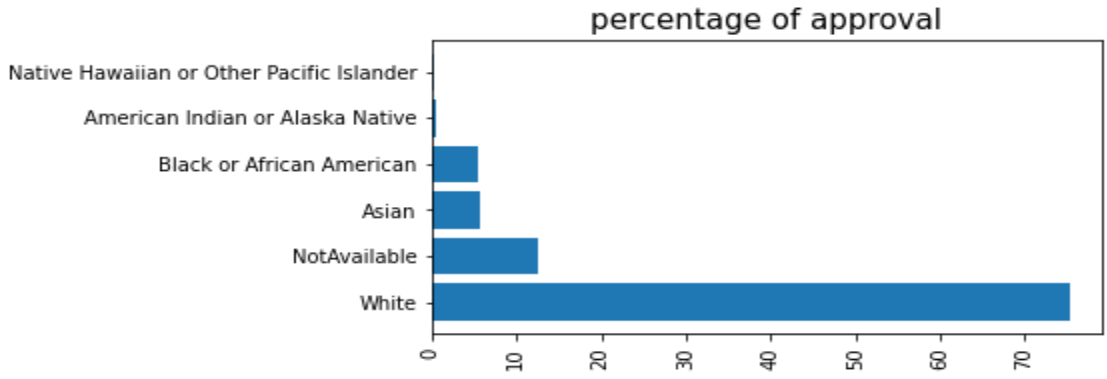
b) Data Analysis Results:

1. Ethnicity, race and gender:

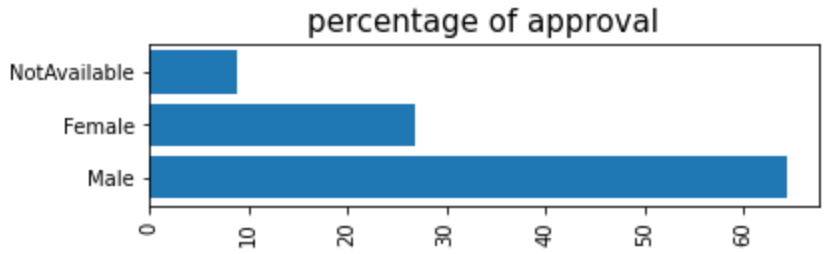
- *Most of the approved applicants were not hispanic or latino.*



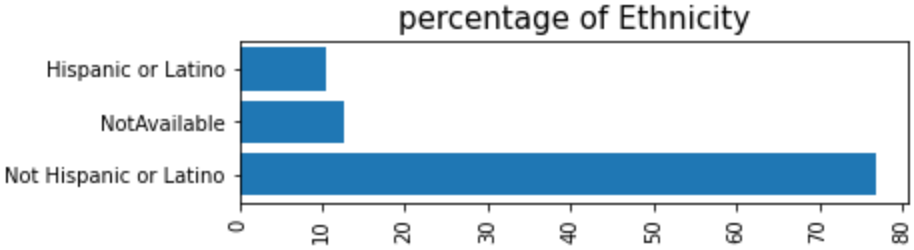
- *Most of the approved applicants were White.*



- *Most of the approved applicants were males.*



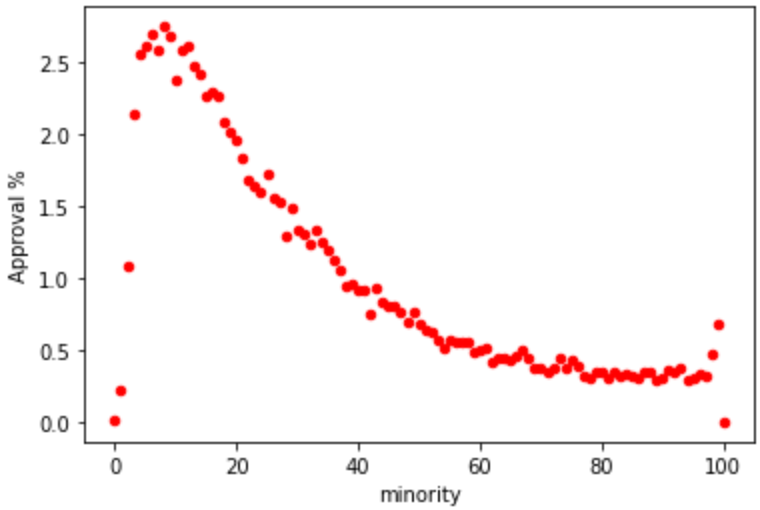
This refers to *gender, race and ethnicity discremenation*. However, when we checked the total number of applicants representing each category, we found that most of the applicants were white, not hispanic or latino and males. So, one can not validate that claim. For example, the total applicants' ethnicity is shown below.



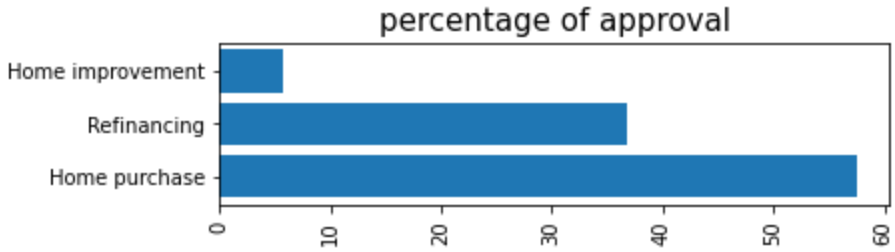
- *The females' average income is 30% less than the males' average income.*

2. The location of the property:

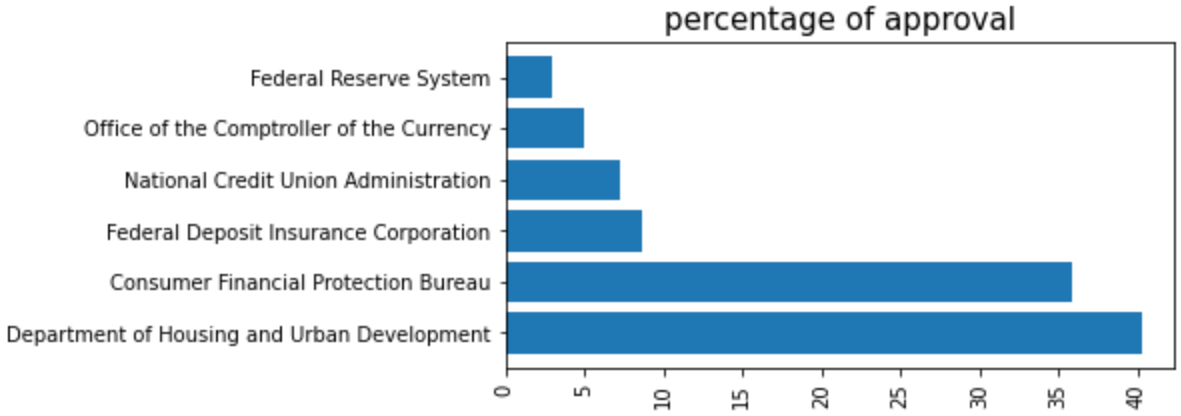
- *We found that California state has the maximum approval percentage.*
- *The percentage of approval in each county or msamd is acceptable (no bias) as the highest percentage for a county/msamd is nearly 2.5%. Also, when we checked the percentage of approval for the whole location (msamd, state, county), we found that the maximum percentage is nearly 2%.*
- *As the minority percentage in a location increases, the percentage of approval decreases.*



3. **loan purpose:**
   Most popular loan purpose is Home Purchase.

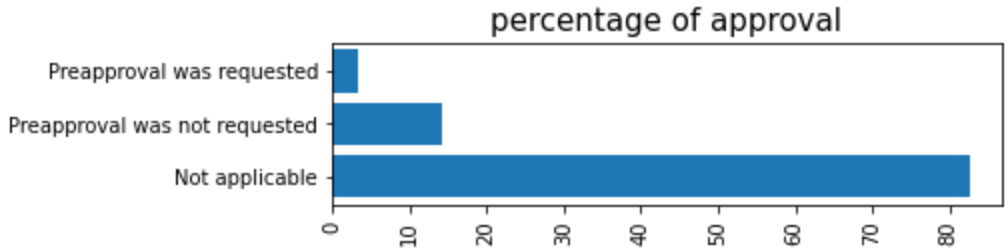


4. **Agencies:**
   The *Department of Housing and Urban Development (HUD)* agency has the highest approval percentage. It also has the highest rejection percentage. So, it looks like this is the most famous agency.

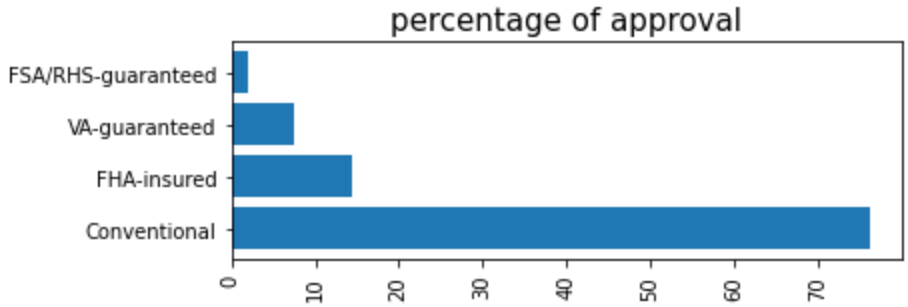


5. **Pre Approval:**
   Most applicants do not apply for preapprovals. Having a "Not applicable" value means that some information was

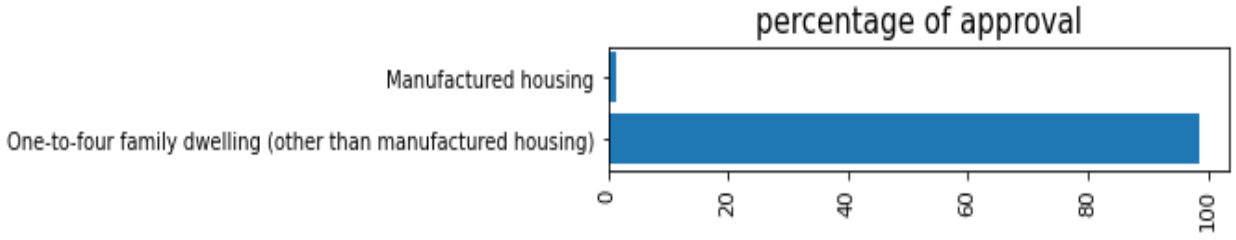
missing or the applicant requested an application, but did not complete it. It looks like applicants are not comfortable with showing their financial status before applying.



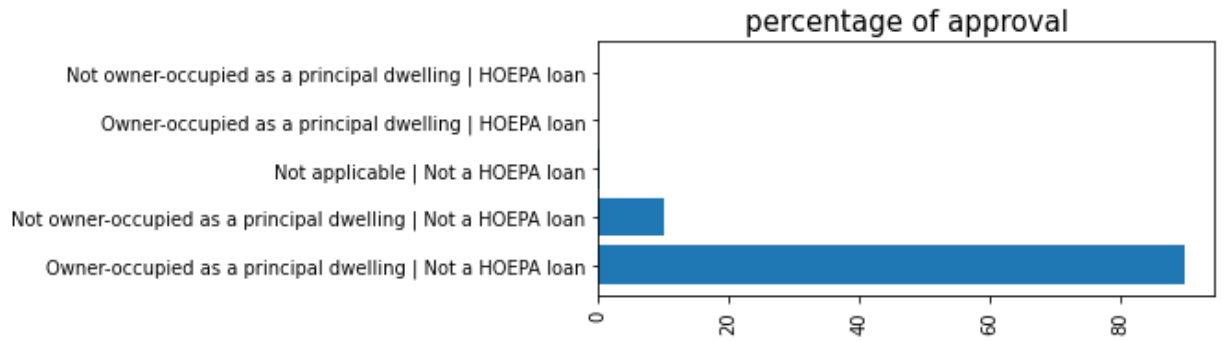
6. **Loan Type:**
   Most popular loan type is a conventional loan.



7. **Property Type:**
   The most accepted property type is *One to four-family property with an approval percentage of 98.74%*. This is because the lender usually looks for a borrower who will pay fast. Someone who buys a *One to four-family property* is usually an investor with much money.



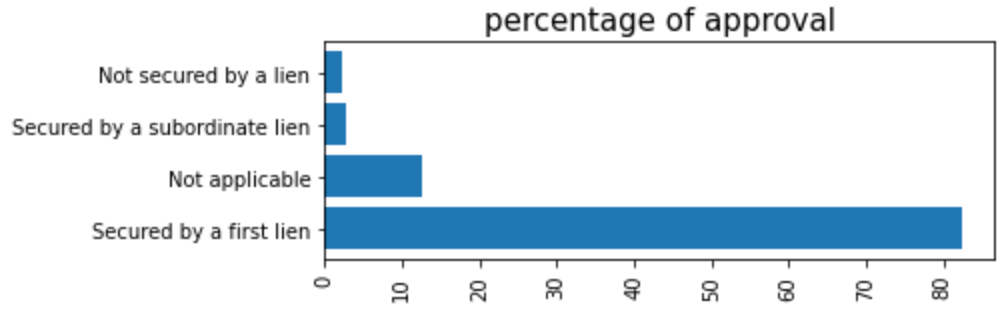
8. **Owner Occupancy Type:**
   The most approved occupancy type is *Owner-occupied as a principal dwelling* with an approval percentage of 89.84%.
   Checking the relation between HOEPA status and the owner occupancy:



9. **HOEPA Status:**
   Most of the accepted loans were not HOEPA loans (99.98%); however, all HOEPA loans were approved.
10. **Loan Amount and Income:**
    It looks like larger loans with larger income are more approved.

Loan State	Average Loan Amount	Average Income
Approval	222.42677302340542	110.94555421714901
Denial	182.3154467535072	91.77987976554688

11. **Lien Status**
    A loan secured by a first lien has a higher percentage of approval. A loan secured by a first lien means that the lender is either the only lender to one's loan or he/she is the first lender. The first or only lender has an advantage over the loans secured by a second lien in that the borrower can not sell or use the property for any similar purpose unless the debt to the lender is paid. Moreover, the first lender has a priority.



VI. Future work

- The data retrieved was biased as most of applicants were of the same gender, race or ethnicity. We may want to retrieve more data to investigate.
- We only investigated the loans that took place in 2014. We may investigate more years (2015,2016, and 2017) to check if these statistics changed with time.
- Investigate more relations about the race, ethnicity, and minorities in general. For example, we can check their average income.
- Relation between the income and the requested loan amount.

VII. References

[1] "Download Historic HMDA Data | Consumer Financial Protection Bureau", *Consumer Financial Protection Bureau*. [Online]. Available: https://www.consumerfinance.gov/data-research/hmda/historic-data/?geo=nationwide&records=all-records&field_descriptions=labels. [Accessed: 19- Jan- 2021].

VIII. Role of each member

Abdelmoez Elsaadany	Visualizations
Amr Mohamed	ML model & Recommendation
Dina Adel	Statistics