Post Ma Loan: Project Abstract

rpemmire, faleksic, elakhoti, rdulepet

Hypothesis Testing Goals

We wanted to examine a large loan dataset to explore whether there may be particular biases in the data. In particular, we wanted to see if there may be differences between socioeconomic groups in the likelihood of loan denials.

Data

We found data as part of the Home Mortgage Disclosure Act dataset, which is a collection of financial institutions reporting on their mortgage data. (Download Historic HMDA Data | Consumer Financial Protection Bureau) The entire HMDA dataset contains data from 2007-2017 for every state. However, we used data from the years 2013-2017 in Rhode Island to make our dataset manageable. The data included attributes such as loan amount, county, loan type, applicant ethnicity, race, sex, and more. This dataset had around 224,000 entries, but after filtering any incomplete entries and defining our subset of data, we ended up with a dataset with around 100,000 entries. Specifically, we filtered out entries with incomplete attributes and only focused on those that had no co-applicants and were for 1 to 4 family dwellings, to simplify our testing and modeling processes.

Hypothesis Testing Findings- This section reflects our hypothesis test findings regarding race, sex, and income for the individuals in our dataset. We have 5 claims we investigated.

1. Minorities have a different likelihood of having their loan denied than white individuals.
   Analysis: We completed a two sample independent proportions z test to examine this claim. We found a rate of denials of 26% and 17% for minorities and white individuals respectively. The p-value from this test was .000, so our findings were statistically significant.
2. Minorities have lower income levels than white individuals.
   Analysis: We completed a one-sided two sample t-test to see if there was a significant difference in the sample means. We found a p-value of .000 for the difference between the mean incomes between the two groups, supporting our claim.
3. Females have a different likelihood of having their loan denied than males.
   Analysis: We completed a two sample independent proportions z test to examine this claim. We found a rate of denials of 18.1% and 18.4% for female and male individuals respectively. The p-value from this test was .18, so our findings were not statistically significant.
4. On average, females have lower income levels than males.
   Analysis: We completed a one-sided two sample t-test for sample means. We found that average income for females was $67,000, while it was $91,000 for males. The p-value for the difference between the two was .000, which was statistically significant.
5. Income levels influence the likelihood of having a loan approved or denied.
   Analysis: We completed a logistic regression on loan approval vs. income level. We found that the coefficient between income level and denial was negative. The z-score of

the coefficient was -11.32 with a p-value of 0. This result is statistically significant, therefore we can support the claim that higher income lowers the likelihood of denial.

Hypothesis Testing Takeaway: After analyzing our claims, it is easy to see that race is correlated with income and denial. However, we found that gender was correlated with income but not with loan denial. This is because, by looking at Figure 1., we see that for every race except white individuals, females have higher denial rates than males. However, white individuals make up 93197/101190 entries in the dataset, leading to a non-significant difference in denials between races. Therefore, there are nuanced but apparent biases in the data.

Machine Learning Models- Goals and Data

We wanted to additionally see if we could build classifiers to predict loan denials in our dataset. We used only certain variables for our classifier: race, sex, income, loan amount, median hud income, county, census tract, tract_to_msamd_income, population, minority population. This is because these are variables we understood and thought were relevant to the classification problem. Additionally, any categorical features were one-hotted into binary variables. Given that our dataset was so large, we used samples of our data to analyze model performance in some cases.

Models Set-Up

We tried Decision Trees, K Nearest Neighbors, and Logistic Regression as potential classifiers. A Support Vector Classifier did not run in time given the size of our data set. We shuffled the data then randomly selected training and testing datasets with a 80-20 split. Once we settled on K Nearest Neighbors as our classifier of choice, we used cross-validation to tune our choice of K.

Results and Analysis

Claim 1: No model significantly outperforms the dummy classifier given our choice of variables.
Support: Figure 2 shows us that no model has testing accuracy much higher than the dummy classifier. This may be because there are variables outside the ones we chose which are instrumental to decisions in loan approvals.

Claim 2: K-Nearest Neighbors is the most promising classifier given our analysis.
Support: Figure 2 shows us that no model except the KNN outperforms the dummy classifier which always predicts that a loan is approved. Although the increase in performance is pretty small, examining confusion matrices shows us that KNN predicts a small amount of loan denials and tends to be correct when predicting this small number denials. Our hypothesis, which should be explored given more time, is that there are small clusters within the data that are more commonly denied loans. K-Nearest-Neighbors is able to find this cluster, while other classifiers which try to find more global decision boundaries have a tougher time.

Figures and Tables

Figure 1: Percentage Denial by Race and Sex

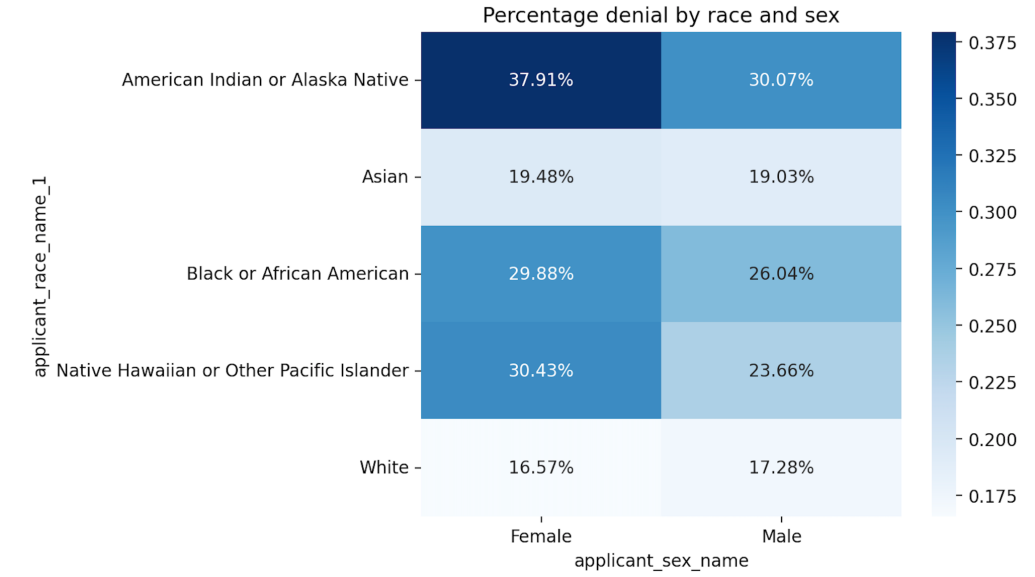

Figure 2: Model Performances on a sample of 10,000 data points

Model	Testing Accuracy
Dummy Classifier	.8185
Logistic Regression	.8185
Decision Trees	.733
K-Nearest Neighbor (K=36) <- found by cross-validation	.8219

Socio-historical context and Impact Report

Socio-Historical Context

To get a mortgage, people used to go to a bank and have a face-to-face meeting with someone who would review their financial situation and make a determination on their ability to repay the loan. Nowadays, financial institutions have large datasets and machine learning algorithms to assist them in that decision. However, if biases have existed in the past data, algorithms will proliferate these biases without any modification. Racial biases in the real estate industry have existed in the past with the era of "redlining", a practice of declining loans for people who live in specific neighborhoods. Stakeholders in the use of algorithms for mortgage lending decisions include both lenders and borrowers. Financial institutions cannot afford to discriminate against certain minority groups because people looking for loans will go to competing lenders. Similarly, borrowers depend on fair algorithms to receive equal loans. Evidence has shown that in the recent past, borrowers have paid higher interest rates due to bias in algorithms, costing minority groups financially.[1] Additionally, just as we have seen in our findings, it is found in research that denials are higher among minority borrowers than white individuals.[2] However, research shows that there are ways to make algorithms more fair without sacrificing accuracy.

Researchers at the University of Texas at Austin found that making fair algorithms with high accuracy in mortgage lending predictions is not feasible with the HMDA dataset.[3] This is because there is not enough information in this dataset in order to really capture the intricacies of the biases in the data. For example, the researchers argue that simply differences in denial rates may not be enough to claim bias, as lenders may be justifiably more likely to lend to those with higher income. This indicates to our group that the HMDA dataset may have been a good starting point for understanding issues in the use of data science for human decision making in the mortgage lending industry. However, more work needs to be done and more data has to be uncovered to make algorithms which are fairer but are still good at lending to those who will repay their loans.

Ethical Considerations

In our hypothesis tests, we found that there are differences in denials between minorities and white individuals. However, there are underlying biases in the data in that there are income and employment differences as well between the two. This is important given that income and employment are big determinants of whether people can repay their loans. These biases cannot

[1] https://www.cbsnews.com/news/mortgage-discrimination-black-and-latino-paying-millions-more-in-interest-study-shows/

[2] https://personal.utdallas.edu/~liebowit/mortgage/mortgages.pdf

[3] https://personal.utdallas.edu/~liebowit/mortgage/mortgages.pdf

be completely ignored, as lenders must take into account income and employment. However, these biases should be mitigated to the point that borrowers with similar income but different races should not receive different predictions from algorithms.[4] There are two ways to achieve this. First, algorithms can completely ignore sensitive variables. However, an alternative more active way to encourage fairness in the algorithm is to measure metrics such as adverse impact on minority groups.

The collection of this data was unbiased in that all financial institutions are required to report on this data in a way that preserves privacy and treats all loan applicants equally. However, the choices we made in how to use this data were not completely free of bias. We came into this project with previous knowledge that biases existed in mortgage lending, and we wanted to examine the extent of these biases. Therefore, we did not look at data without sensitive attributes such as race and sex. A misinterpretation of this is that we were trying to make a predictor to take advantage of the biases in the data. This was not our goal, and while researching for this portion of the project, we learned that there are ways to make models more fair in order to account for biases in the data.

Works Cited- 3 sources needed

1. https://news.mit.edu/2022/machine-learning-model-discrimination-lending-0330
2. https://www.cbsnews.com/news/mortgage-discrimination-black-and-latino-paying-millions-more-in-interest-study-shows/
3. https://personal.utdallas.edu/~liebowit/mortgage/mortgages.pdf

[4] https://news.mit.edu/2022/machine-learning-model-discrimination-lending-0330