Raghav Pemmireddy | Esha Lakhotia | Filip Aleksic | Riya Dulepet

Post Ma Loan!

An Analysis of Loan Applications in Rhode Island and Attributes Leading to Acceptance vs Denial

Background

Whether it be organizing large purchases, covering an emergency expense, consolidating debt, or any other fiscal needs, loans provide otherwise impossible opportunities for businesses, homeowners, students, etc and are essential to the growth of overall money supply in an economy. However, the loan application approval process isn't always easy or straightforward, with many factors like race, socioeconomic status, gender, geographic location, etc playing a role in the decision.

We were interested to see which **specific factors** of an application most affect the rate of approval vs denial of loans within the state of Rhode Island. Specifically, we analyzed how gender, race, gender and race, income level, & county and income level all impact **denial** and **approval** rates.

Dataset and Methodology

Dataset: Home Mortgage Disclosure Act dataset, a collection of financial institutions reporting on their mortgage data. 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.

Sample: We generated our sample by joining the tables which had loans data for Rhode Island for years 2013, 2014, 2015, 2016, and 2017.

Attributes: loan amount, county, loan type, applicant ethnicity, race, sex, and more

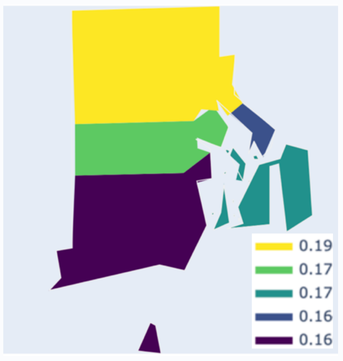
Hypothesis

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.

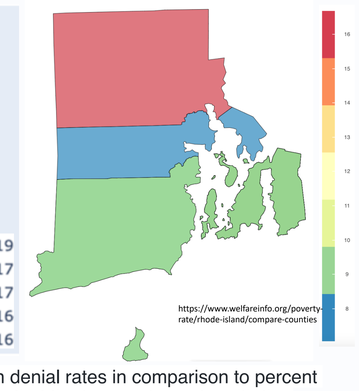
3 variables of interest: sex, race, and income. We wanted to investigate if potential differences in mortgage denials are dependent on sensitive variables or the more traditional ones you would expect such as income.

County Level Analysis

Loan Denial Rates by County in Rhode Island



Percent Below Poverty Line by County in Rhode Island



Looking at the county analysis of loan denial rates in comparison to percent below the poverty line, it makes sense that the top most county, with the highest denial rate (19 percent), also has the highest percentage of people below the poverty line. The rest of the percentages of denial rates were pretty similar, so conclusions are difficult to make surrounding the rest in correlation to the poverty line graphic.

Machine Learning Models

We built classifiers to predict loan denials in our dataset in samples of randomly selected training and testing sets in a 80-20 split. Additionally, any categorical features were one-hotted into binary variables
We tried **Decision Trees, K Nearest Neighbors, and Logistic Regression** as potential classifiers. Once we settled on KNN as our classifier of choice, we used cross-validation to tune our choice of K.

- Figure 2: 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.
- Figure 2: 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, KNN predicts a small amount of loan denials and tends to be correct when predicting this small number denials.

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

- **PCA visualization** shows that the data was not obviously linearly separable, and in fact, around 80% of the data was made up of non-denials. The data fell into little clusters, some of which may have had more denials than others, but most clusters seemed to still have a majority of non-denials. We chose PCA because it effectively shows us the basic structure of the data and the labels of all of the data points.

Income Level Analysis

Alternative Hypothesis #1: Minorities have lower income levels than whites.

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.

Intuitions: Intuitively, the severe gaps in mean of minority vs white incomes make sense. We can reason that there are many reasons for this disparity, most of them being entrenched discrimination and unequal opportunities.

Income Level Breakdown across Racial Groups $	
American Indian or Alaska Native	62,000
Black or African American	58,000
Native Hawaiian or Other Pacific Islander	60,000
Asian	82,000
White	83,000

Alternative Hypothesis #2: 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.

Intuitions: We know women earn less than men,averaging only 80% of their income. However, it is shocking to see the gap being so high (24,000). Additionally, with the above test, we concluded that men and women have similar rates of approval, which seems to be inconsistent with the findings of this test (or meaning that income holds less of a weight than other factors).

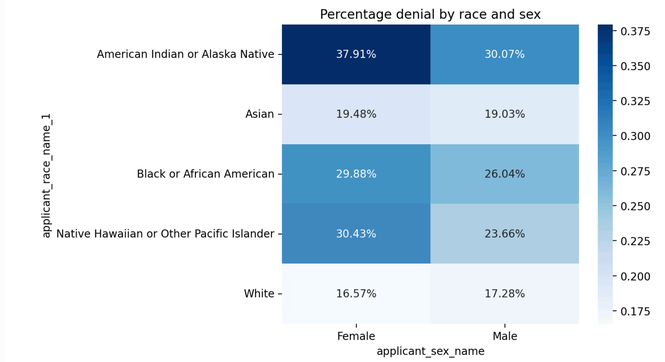
Alternative Hypothesis #3: Income levels influence the likelihood of having a loan approved or denied.

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.

Intuitions: It makes sense that the higher the income, the lower the denial and the lower the income, the higher the denial. After our intuitions of the above two tests, it makes sense that the relationship between income levels and loan denials is slight.

Analysis of Denials over Race and Sex

Figure 1: Percentage Denial by Race and Sex



Alternative Hypothesis #4: Minorities have a different likelihood of having their loan denied than whites.

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.

Intuitions: Intuitively, it makes sense that minorities have a higher chance of having their loan denied. Due to inherent prejudices and racism built into the system, many factors like income differences, discrimination, etc could have led to this disparity.

Alternative Hypothesis #5: 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.

Intuitions: Intuitively we were surprised to see the rate of denial being higher for men than women, since women on average have lower incomes in comparison. However, since the rates of denial are pretty similar, we can see how these findings make some sort of sense.

Conclusions

Our results somewhat correspond with our initial belief that in a domain such as mortgage loans, there may be other data missing that is more important for classifying denials in mortgage loans. Our main takeaways:

- **Race** is correlated with income levels and likelihood of denials
- **Gender** is correlated with income levels but not likelihood of denials

Future Improvements and Limitations

Limitations

- Our analysis was limited to loan applications that involved properties of **one-to-four** family dwellings as well as loans that did not have a co-applicant. This could have potentially **biased** our models and analyses.
- Some of the data lacked interpretability and documentation. Variables were unusable to us because we could not figure out exactly what they meant. Given the dataset had so many features, choosing "important" variables was a challenging task.
- Lack of **domain knowledge**.

Future Work

- Compare loan statistics with **another state** to accurately address how bias is affecting our data collection and analysis.
- Examine **deep learning** approaches or other datasets that contain more personalized variables. We originally chose the HMDA dataset because of its volume, but it might have been worth looking into smaller, more comprehensive datasets.
- Researchers at UT Austin have found that making fair algorithms with high accuracy in mortgage lending predictions **may not be feasible** with the HMDA dataset due to the fact that there is not enough information in the dataset to capture the intricacies of the biases. 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 fairer, accurate algorithms.