Post Ma Loan - Analysis Deliverable

Preprocessing

Our first challenge was to hone in on a specific part of the dataset. We limited our analysis to loans that concern properties of one-to-four family dwellings as well as loans that did not have a co-applicant. We realize how this could bias our data findings and models but decided this to keep the data as cohesive as possible.

For hypothesis testing, we had to clean/restructure our data: we removed missing data and fields that were filled in with inconclusive values such as “information not provided.” For classification, we kept the missing data because we thought that it could be valuable in determining the likelihood that someone receives a loan (i.e. people with less information or incomplete applications may be less likely to receive a loan.) We additionally found that there were many possible labels in the action_taken column, but we reduced these labels to whether or not a financial institution denied the loan application. This turned the problem into a binary classification problem.

Hypothesis Testing

**Side note: we used alpha = 0.05 as our threshold for identifying statistically significant outcomes*

Our hypothesis tests primarily concern 3 variables of interest: sex, race, and income. We believed going into the project that these would be interesting variables to look at given the potential for bias in this type of dataset. We wanted to see whether differences in mortgage denials were dependent on sensitive variables or the more traditional ones you would expect such as income.

Alternative Hypothesis #1: Minorities have a different likelihood of having their loan denied than whites.

- **Test:** 2 sample independent proportions z test.
- **Why Test:** We used this test to determine if there is a significant difference between loan denial rate between minorities and whites. Since rates are proportions, we needed to use the z test to come to a conclusion.
- **Statistical Results:** The p-value was approximately 0, so our findings are significant. The rate of denials for minorities was approximately 26% while the rate of denials for whites was approximately 17%.
- **Conclusion:** Because of the p-value, we accept our hypothesis, meaning the statistical results support our assumption. From our test, we can conclude there is a significant difference in loan denial between white people and minorities.
- **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 #2: Minorities have lower income levels than whites.

- **Test:** One-sided two sample t-test.

- **Why Test:** We used this test to examine whether there is a significant difference in mean of two samples (minorities and whites incomes), which are independent.
- **Statistical Results:** Since the p-value was approximately 0, our findings are significant. This was the income level breakdown of averages 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
- **Conclusion:** The statistical results support our hypothesis, since it shows that there is a clear difference in income levels between white people and minorities.
- **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.

Alternative Hypothesis #3: Females have a different likelihood of having their loan denied than males.

- **Test:** 2 sample independent proportions z test.
- **Why Test:** We used this test to determine if there is a significant difference in loan denial rates between men and women. Since rates are proportions, and the form indicated gender as independent proportions, we needed to use the independent proportion z test.
- **Statistical Results:** Since the p-value was approximately 0.18, our findings are not significant**.** Our results showed the rate of denial for women is approximately 18.1%, while the rate of denial for men is approximately 18.4%.
- **Conclusion:** Since the statistical results showed there is not a statistically significant difference in denial rates between men and women, our hypothesis is not supported.
- **Intuitions:** Intuitively we were surprised to see the rate of denial being higher for men than women, since women on average have lower incomes (shown in the test below), etc in comparison. However, since the rates of denial are pretty similar, we can see how these findings make some sort of sense.

Alternative Hypothesis #4: On average, females have lower income levels than males.

- **Test:** One-sided two sample t-test.
- **Why Test:** We used this test because we were interested in whether there is a difference in mean of two samples (female and male income).
- **Statistical Results:** Since the p-value was close to 0, our findings are significant. The average income for women was approximately $67,000, while the average income for men was approximately $91,000.
- **Conclusion:** Since the statistical results showed there is a difference in income levels between females and males, our hypothesis is supported.
- **Intuitions**: Intuitively, we can see why our results showed women to have a lower average income than men. Outside this dataset, we know women earn less than men, still averaging only 80% of their income (though varying on occupation). 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 as shown in the below test).

Hypothesis #5: Income levels influence the likelihood of having a loan approved or denied.

- **Test:** Logistic Regression test.
- **Why Test:** We considered applying linear regression, but given that the dependent variable is not continuous, logistic regression analysis allows us to describe the relationship between an independent variable that is continuous and a dichotomous dependent variable.
- **Statistical Results:** The correlation coefficient between income levels and loan denial is -0.0012, with a p-value of .0000 so our findings are significant.
- **Conclusion:** The statistical results support our hypothesis since it shows that there is a relationship between income levels and loan denial–slightly negative.
- **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.

False discovery rate is very low due to all tests having significance below alpha/5 = 0.01. Because of this, we are confident in not identifying significant results from random luck.

Overall, the results mainly correspond to our initial beliefs of the data. The only assumption that did not hold true was women getting denied at a higher rate than men (men were denied slightly more in fact).

As for the other assumptions, since they were general and we had a big dataset, it wasn't surprising that it would represent a population. Our analysis was appropriate because we used hypothesis testing that made sense in terms of whether we were looking at proportions or whether it was one or two samples.

Visualizations for Data Exploration

For one of our visualizations, we depicted the percentage of denial by race and sex. We decided to depict this based on denial and not acceptance to evoke a stronger reaction to the heatmap. We included the applicant's race on the y axis (American Indian or Native Alaskan, Asian, Black or African American, Native Hawaiian or Pacific Islander) and applicant's sex on the x axis (female, male). Each sub-square was colored according to the percentage legend on the right side, with lower percentages in a lighter blue and higher in darker blue. We could have depicted this as a slightly confusing or messy pie chart or bar graph (or two separate ones), but we chose a heatmap, since it efficiently and effectively relayed the disparities not only by race and by gender but by each race and gender pairing. For example, we can readily see American Indian females had a higher rate of denial in comparison to American Indian males (37 percent vs 30 percent). Additionally, the colors make it easy to visualize the differences, making the data relationship evident before anyone even looks at the actual percentages. The title, labeled axis, and color legend to the right make the heatmap easy to understand and devoid of any further requirements of explanation or context along with it.

The inspiration behind this visualization is that in our hypothesis tests, we saw that denials were correlated with race and income but not with sex. Therefore, it was worth looking at the

breakdown of denials over race and sex. Interestingly, we roughly see that for minorities, females are denied more than males, while the opposite is true for white individuals. Given that white individuals make up a large part of the data set, this could have led to a lack of a significant difference in denials between males and females.

For our other visualization, we depicted denial rates by county in the state of Rhode Island. It intuitively made sense to use a choropleth map since we were depicting different rates within a geographic region. Visually, the map was easy to interpret due to the distinct boundaries and colors of each county. It was interesting to see it depicted as a map to compare the differences and similarities of rates of distant vs close-by counties, especially since Rhode Island is such a small state and neighboring counties are pretty close by to each other. One would assume nearby counties would have similar rates, but that wasn't always the case (even though the yellow and dark blue counties are right next to each other, they had the second greatest difference in loan rejections). One could do further analysis examining why and which factors between counties affected these loan rates. Using any other chart or graph (like pie or bar) would not have been as visually appealing, easy to interpret, or revealing of spatial data. The title and color legend make the choropleth easy to understand and devoid of any further requirements of explanation or context along with it. However, if we wanted to, we could add the names of the counties over each section, or even break down by city to dive deeper.

K-Nearest Neighbors

Data

The data was adequate for our analysis, with not many missing values for the variables of interest. For race/sex, around 5-10% of entries had missing or unprovided values. We still considered these values for the sake of considering the group of applicants which chose not to submit this sensitive information. For classification preprocessing, we kept the same dataset we used for the hypothesis testing (limited our analysis to loans that concern properties of one-to-four family dwellings as well as loans that did not have a co-applicant) but included the missing data since we thought that it could be valuable in determining the likelihood that someone receives a loan (i.e. people with less information or incomplete applications may be less likely to receive a loan.) We had to perform one-hot encoding on certain columns of our data set. This issue was not because of problematic data, but because of our assumption of our models. We used the features race, sex, income, loan amount, median hud income, county, census tract, tract_to_msamd_income, population, minority population with the denied/not denied classifier. As our data was very large (over 100,000 entries with around 250 variables after one-hot encoding), we ran our models on a sample of the data. We found that running them on the entire data set took more time but did not have a significant impact on classifiers' performance, so we used a sample to aid with time management. Using distributed computing or cloud services may have helped remedy the lack of time or computing resources to run classifiers on the entire dataset.

Some of the data values were tough to understand, and the data was not very well documented. Therefore, some variables were unusable to us because we could not figure out exactly what they meant or what it meant to not have that specific value filled in. Given that the dataset had so many features as well, we ran into a problem with choosing variables. Having domain knowledge may have helped us with this, but we tried our best to pick variables that we felt were relevant to the classification problem.

Visualization With PCA

Looking at a visualization from PCA, it could be seen 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. Therefore, we had to pick a classifier which could classify non-separable data as well as classify a relatively rare event of loan denial. Given that a dummy classifier which always predicted non-denial could have an accuracy of around 80 percent, we had to have a higher threshold for model accuracy as well.

We chose PCA because it effectively shows us the basic structure of the data and the labels of all of the data points. It can be tough to view data in 3 dimensions but the interactivity of matplotlib helped clarify how the data is structured. The PCA only takes into account the same variables used for our classification.

Classification Models

We tried using classifiers like a Decision Trees, Support Vector Classifier, and Logistic Regression. Each of these classifiers had either lower or the same accuracy as a dummy classifier. We found that the eventual predictions from the SVC and Logistic Regression were to always predict non-denial. However, we found that K-Nearest-Neighbors to be a slightly better-performing classifier. After running KNN on samples of around 10,000 data points, we saw an improvement of about .5-1% in accuracy for certain values of k. We hypothesize that KNN was able to find a small cluster within the data which were easier to classify. Looking at the confusion matrices for KNN, most predictions were still non-denials, but the few predictions that were denials tended to be correct more often than not, increasing accuracy by a small amount. Our hypothesis for using KNN was that financial institutions would reject similar applicants, so the similarity between data points would be important to classifying denials vs. non-denials.

Our measure of success was calculating the average cross validation accuracy, since there wasn't a distinct indication of failure vs success. Below you can see a plot for average cross validation accuracy vs. k. The results of the graph make intuitive sense, with the testing accuracy being lower when K = 1, rapidly improving between 1 and 6, and stabilizing at around the accuracy of a dummy classifier as k increases. This is an example of the bias variance tradeoff, and if we were to choose an optimal k, it would have to balance these two aspects. Given that our KNN model accuracy was only an improvement of .5-1% on the dummy model accuracy, it shows that KNN might be a good classifier for this data set, but more work might need to be done to choose the right variables and preprocess the data.

Visualization of KNN k vs. accuracy

We chose this graph because it was easy and clear to interpret. We could have presented it in a bar graph or pie chart, but we found it more helpful to be able to see the accuracy with respect to the hyperparameter k. From the graph, we can see an increase in accuracy as k increases, but a leveling off around the same accuracy as the dummy classifier. Depending on the sample, certain k's will achieve higher training accuracy than the dummy classifier, but this isn't always consistent/might be due to overfitting. The graph is stand alone and doesn't require text to explain apart from the title and the labels of the axis. It does, however, require context of the classifier we chose, the data, and the classification problem at hand.

Conclusion

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. Given more time, we could examine deep learning approaches or other data sets which contain more personalized variables and analysis. Additionally, we could have looked at datasets from other states. We originally chose the HMDA dataset because of its size, but it might have been worth looking into smaller datasets with more comprehensive data.