THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

Unintentional Single-Family Loan Price and Approval Discrimination on Race through Financial Deserts

DERIC LIANG
SPRING 2021

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Statistics and Economics
with honors in Statistics

Reviewed and approved* by the following:

Qi Li
Assistant Professor of Economics
Thesis Supervisor

Jia Li
Professor of Statistics and Computer Science
Honors Adviser

* Electronic approvals are on file.

Abstract

In this study, we explore the factors which contribute to the continued discrimination against racial and ethnic minorities, particularly Black and Hispanic communities, in mortgage loan pricing and approval. We draw on the work of Bartlett et. al. and hypothesize that despite laws against discrimination based on race and ethnicity, financial deserts (areas with low market competition) justify high loan prices, inevitably affecting disadvantaged groups such as racial minorities. We particularly suspect that financial deserts with a larger minority population would see exacerbated effects of low competition on their loan pricing, compared to areas with a smaller minority population. We also explore the effects of financial deserts on loan approval, to support previous findings. Data are sourced from the Home Mortgage Disclosure Act database, Freddie Mac database, and Census data. The data are used to construct models predicting loan interest rates, rate spread, and rejection rates on a zip code level, utilizing Ordinary Least Squares regression, regularization, and regression tree techniques, providing robustness to our findings through drawing on different modeling approaches.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would first and foremost like to thank Dr. Qi Li for her role in guiding me towards research in the mortgage market. I began the thesis process with little idea of what topics I was interested in or wanted to explore, except for the fact that I wanted to bring statistical applications into areas in economics. Dr. Li guided me towards literature and data which would become essential to the research I perform in this thesis. Dr. Li was also incredibly patient with me as I found numerous dead ends with my methodology and as I made several significant adjustments to my research questions and the data I included in my analysis.

Secondly, I would like to thank all the faculty under whom I have taken Statistics and Economics courses over my four years of study. Each class I have taken, quantitative or qualitative in nature, has allowed me to develop my abilities in processing, analyzing, and interpreting data which has been invaluable to me in the process of writing.

Next, I would like to thank my Honors Advisors over the course of my time in the Schreyer Honors College. Dr. Hunter was helpful in the process of selecting the courses which would be most advantageous to my development as a statistician, and Dr. Jia Li has been helpful in ensuring I received research credit for the work I did for this thesis.

Finally, I would like to thank my family and friends, who have been supportive throughout the thesis process, the pandemic, and the entire duration of my undergraduate studies. I particularly would like to thank my mother, an incredible statistician with whom I had many conversations about my research whenever I faced difficulties in my data processing or analysis. I do not know what I would do without all of you.

Chapter 1: Introduction

1.1 Motivation

Minority groups, especially the Black and Hispanic communities, experience many disadvantages in the United States economy. This fact has become widely accepted in recent years by researchers, politicians, and the public. As researchers, we have the power and obligation to utilize data in furthering our understanding of the specific avenues in which minorities experience these disadvantages so that legislators can implement policy that increases equity in the U.S. economy. In this study, we specifically examine racial discrimination in the mortgage market and how it arises through location.

In order to demonstrate the importance of location in promoting racial discrimination in the mortgage market, we explore literature on the concept of redlining. Aaronson et. al., researchers from the Federal Reserve Bank of Chicago, find that the practice of redlining, or discriminating against borrowers due to the racial composition of their neighborhood, resulted in lower home values, home ownership rates, and credit scores in these areas (2017). This presents obvious disadvantages in the sense that racial minorities face difficulties acquiring wealth compared to non-minorities and will have less options in terms of borrowing money.

However, recent research has shown a decrease in discrimination in the mortgage market through the emergence of Financial Technology (FinTech) lenders. These are lenders making approval decisions based on algorithms, as opposed to the traditional face-to-face lenders. Researchers from the National Bureau of Economic Research (NBER) have shown that FinTech lenders discriminate 40% less than face-to face lenders. Additionally, their study demonstrates that FinTech lenders discriminate on loan pricing, but not on loan approval (Bartlett et. al., 2019).

The researchers declined to explore the factors behind this discrimination, stating, “How discrimination happens is an important question. We leave a full exploration of this topic to a separate research project.” They speculate that the FinTech algorithms detect individuals living in financial services deserts, and therefore less prone to shop for competitive loan prices. We will discuss mortgage market competitiveness more in the next section, where we outline our research objectives.

1.2 Research Objective

The first objective is to gain an overview of the mortgage market for minority groups, which will be done through exploratory data analysis. What are the loan rejection rates for each racial group? What are the average interest rates and rate spreads for each racial/ethnic group? Do univariate relationships exist between market competitiveness and loan pricing/approval variables?

The second objective is to build robust models to determine whether racial discrimination can be explained through financial services deserts. This is possible using the Herfindahl-Hirschman Index (HHI), which calculates market competitiveness and can be found for each location. We account for legitimate factors (credit-risk variables, variables correlating with race through hidden relationships) and illegitimate factors (race) to isolate the effect of competitiveness. The models also explore the effect of an interaction term between race and competitiveness to explore if a lack of competition exacerbates discrimination. We will be looking at interest rate, rate spread, and rejection rate to analyze price discrimination and approval discrimination.

Framing our objectives into research questions, we attempt to answer the following: How does lender discrimination arise? Specifically, does it arise through detection of an individual's presence in a financial service desert? If so, do the deserts with higher minority percentages see larger increases in their average interest rate? These questions all are worthwhile to consider due to previous findings on redlining and may help us as a society take tangible steps to increase equity in the overall economic landscape.

Chapter 2: Methodology

2.1 Data Sources

The data come from the Home Mortgage Disclosure Act (HMDA), Freddie Mac, and Census data. The HMDA "requires many financial institutions to maintain, report, and publicly disclose loan-level information about mortgages." This data provides valuable information on the actions of lenders alongside factors which may influence these actions, whether legitimate or illegitimate ("Mortgage Data (HMDA)"). The Consumer Financial Protection Bureau explicitly states that this data source can be used to expose discriminatory lending patterns.

From our literature review, the HMDA data contains variables on interest rate and rate spread for individual loans, as well as other useful information outlined in Table 2.1. However, it omits two crucial variables to our analysis: credit score and data on the percent of workers in the public sector. Credit score comes from Freddie Mac, a loan purchaser whose mission is, "to provide liquidity, stability, and affordability to the U.S. housing market in all economic conditions extends to all communities from coast to coast" ("About Freddie Mac," 2021). Freddie Mac provides loan-level data on loan performance and includes the credit score associated with the borrower on each loan. Credit score is a crucial credit-risk factor to include, as it accounts for debts/payment history and is a large consideration for lenders in making lending decisions ("How are fico scores calculated?" 2019).

We include a variable indicating the percent of workers in the public sector, calculated from Census data, as research from NBER has shown that employment stability is one of the most important considerations in assessing credit standing according to 126 banks (Chapman, 1940),

and that working in the private sector as a male correlates with lower employment stability (Hollister, 2011).

2.2 Data Aggregation

The HMDA data and the Freddie Mac data are impossible to merge on a one-to-one loan level due to the lack of information; the common variables in both data sets will not guarantee unique matches. The HMDA data has county data, while the Freddie Mac data has zip code data. In addition, the Census data gives observations for each county, not a loan level. Therefore, to bring the information from all the data sets together, we aggregate variables by location and merge by location. The challenges in merging by location arises from the mismatch of location level and the coding of zip code in the Freddie Mac data omitting the last 2 digits, such that it takes the format "XXX00" (here forth referred to as the "abbreviated zip code"). This problem is addressed through the following process:

1. In the HMDA data, calculate the percentage of males, the percentage of minorities (defined for our purposes as Black or Hispanic), rejection rate, median interest rate, rate spread, income, and loan-to-value ratio by county. Additionally, calculate the HHI for each county (the HHI calculation is explained in Section 2.3.1).
2. In the Freddie Mac data, average credit score by zip code.
3. In the Census data, extract the data on the percentage of people working in the public sector in each county.
4. Find zip-county conversion data with "full zip code" data (which does not convert the last 2 digits into zeroes), as well as county code data. We combine this by county code with data giving the county name. We generate an abbreviated zip code variable using the full zip code variable to match the "XXX00" zip format of the Freddie Mac data.

5. Merge zip-county data with Freddie Mac data by abbreviated zip code. Merge this with HMDA data by county code. Merge this with Census data by county name. The resulting data will have fully coded county codes and zip codes associated with all the variables.

A limitation to this approach is that information is obfuscated through aggregation in general, but more specifically through applying aggregations from a general level to observations in a more specific level; however, this approach is unavoidable given the data, and we approach our result with the assumption that zip codes beginning with the same three first digits are similar in character. We must also consider the fact that the county codes associated with the Freddie Mac abbreviated zip codes are approximated in this approach, such that counties may not have been assigned to loans with 100% accuracy. Finally, by aggregating by zip and county codes, we assume homogeneity within these locations. The quality of these assumptions is unclear but are necessary to build a data set with all the desired variables for modeling.

The full data description is shown in Table 2.1 below. All data sources come from 2018, as this is the most recent year in which data is available from all sources.

Variable	Source	Use	Abbreviation
Zip Code	Zip-County	Denotes observation	zip
Interest Rate	HMDA	Response variable	med_ir
Rejection Rate	HMDA	Response variable	rej_rate
Rate Spread	HMDA	Response variable	med_rs
Income	HMDA	Credit-risk variable	med_inc
Loan-to-Value Ratio	HMDA	Credit-risk variable	med_ltv
Percent Male	HMDA	Demographic variable	perc_male
Percent Minority	HMDA	Demographic variable	perc_min
HHI	HMDA	Measure of competitiveness	hhi
Credit Score	Freddie Mac	Credit-risk variable	avg_credit
Percent Public Sector	Census	Demographic variable	perc_pub

Table 2.1: Data description, with sources and variable use in the model

2.3 Variable Definitions

Most of the variables in the data set are self-explanatory; however, we believe the reader would benefit from definitions of the HHI, rate spread, and loan-to-value ratio variables.

2.3.1 Market Competitiveness – The Herfindahl-Hirschman Index

The Horizontal Mergers Guidelines (2010), published by the U.S. Department of Justice (DOJ) and the Federal Trade Commission (FTC), defines the HHI as the following formula:

$$HHI = 10000 \sum_{i=1}^{N} s_i^2$$

In the formula, s_i is the market share of party i, and N is the number of parties in the market. The DOJ and the FTC typically use the HHI to calculate the competitive effects of mergers; since it quantifies competitiveness in a market, it works for our models to measure the extent to which a location may be a financial desert. The HHI gives greater weight to lenders with larger market shares, pronouncing the effect of an individual lender's concentration and thus making it a popular measure for competition.

The HHI can range from 0 to 10000. The DOJ and FTC classify HHIs of less than 1500 as unconcentrated, between 1500 and 2500 as moderately concentrated, and above 2500 as highly concentrated.

2.3.2 Rate Spread and Loan-to-Value Ratio

The Federal Financial Institutions Examination Council defines rate spread as the difference between the APR and the estimated average APR currently offered on comparable mortgage loans ("FFEIC Rate Spread Calculator," 2016). The rate spread captures information regarding the interest rate on a loan and the quality of the pricing an individual is receiving on the loan; if the rate spread is high, the lender is overcharging the borrower.

The loan-to-value ratio calculates the mortgage amount divided by the appraised property value. The LTV ratio is "an assessment of lending risk that financial institutions and other lenders examine before approving a mortgage," making it a useful credit-risk factor to include in the model. Loans with a high LTV ratio typically have higher interest rates (Hayes, 2015).

2.4 Model Methods

The first group of models are ordinary least squares regression models with zip code as the observation, interest rate, rejection rate, and rate spread as response variables, and the remaining variables as the covariates. The models include an interaction term between the Percent Minority and HHI variables, as we seek to understand whether the effect of financial deserts on loan pricing and approval differ in deserts with a high minority population versus those with lower minority populations. The models are listed below:

$$IR = \beta_0 + \beta_1 Credit + \beta_2 Inc + \beta_3 LTV + \beta_4 \%M + \beta_5 \%Pub + \beta_6 HHI + \beta_7 \%Min + \beta_8 HHI * \%Min$$

$$RS = \beta_0 + \beta_1 Credit + \beta_2 Inc + \beta_3 LTV + \beta_4 \%M + \beta_5 \%Pub + \beta_6 HHI + \beta_7 \%Min + \beta_8 HHI * \%Min$$

$$RR = \beta_0 + \beta_1 Credit + \beta_2 Inc + \beta_3 LTV + \beta_4 \%M + \beta_5 \%Pub + \beta_6 HHI + \beta_7 \%Min + \beta_8 HHI * \%Min$$

The second group of models are regularized models. Regularization specifies linear models with an added penalty term which places a constraint on the size of the model parameters (Cremona, 2018). Various regularization methods exist, and this study will use LASSO (Least Absolute Shrinkage and Selection Operator) regularization, which adds the following penalty term to the OLS sum of squared errors function (henceforth known as the loss function):

$$\lambda \sum_{j=1}^{p-1} |\beta_j|$$

where p is the number of parameters in the model, j is the parameter index, and λ is the tuning parameter determining the level of penalization. We use cross-validation to determine the

optimal value of λ, which subsets the data into groups randomly and builds/evaluates models on these different groups (Lewis, 2000). LASSO regularization shrinks coefficients to 0, serving as a model selection technique in eliminating predictor variables which have less effect on the response variable (Cremona, 2018).

The final group of models are regression trees, with the same response variables and covariates as the previous models. These models provide several advantages over OLS models as a more advanced supervised learning technique. The first advantage lies in the efficient mechanism of variable selection, using cross-validation, as the regularization does, to test many models and arrive at the best fit; OLS models do not automatically perform such a model selection process (Lewis, 2000). The second advantage lies in the ability of regression trees to impute data to replace missing values based on predictor variables which are determined to contain similar information (Lewis, 2000). This allows us to avoid discarding zip codes which have missing values from our models, of which there are several, resulting in a potential improvement in model accuracy. Finally, the models are non-parametric unlike OLS models, which will allow more flexibility in determining a model fit compared to a linear fit (Breiman et. al., 1984). Regression tree output works as a classifier, clearly assigning predictions for our response variables of loan pricing and approval based on the predictor variables; this simplifies interpretation of our models compared to OLS regression models.

Chapter 3: Exploratory Data Analysis

Before running statistical models, we first explore how the summary statistics differ between race and ethnic groups for the loan pricing and approval variables. This provides a cursory and immediate glance into any differences that certain groups may face in the mortgage market. We also map our response variables, HHI, and percent minority counts across counties to gain preliminary insights into geographic correlations between these variables. We then explore scatter plots of each response variable on each predictor variable to determine which predictors hold stronger relationships with loan pricing and approval. Finally, we explore correlations between the response variables and the predictor variables.

3.1 Summary Statistics

Table 3.1 provides the average interest rate, average rate spread, and rejection rate by race. Black and African American individuals have a rejection rate of 29.95%, compared to the 17.45% of white individuals. Additionally, we find higher rejection rates for American Indians/Native Americans as well as Pacific Islanders, at 30.75% and 31.63% respectively. All three groups, on average, experience higher loan prices in terms of interest rates and rate spreads, with Black and African American individuals experiencing the largest loan prices.

Race	N	Average Interest Rate	Average Rate Spread	Rejection Rate
Amer. Indian/Native	114770	5.172498	0.8772135	0.3075194
Asian	795837	4.678843	0.3895348	0.1953352
Black/AA	1036912	5.588068	0.9585595	0.2995462
Pacific Islander	51423	4.868369	0.6754440	0.3162787
White	9855889	5.049271	0.6712302	0.1745041

Table 3.1: Summary statistics of loan variables, by race

Table 3.2 provides the average interest rate, average rate spread, and rejection race by ethnicity. Those who identified as not Hispanic or Latino experienced lower interest rates and rate spreads than the other groups, particularly Hispanic/Latino and Mexican individuals. In addition, non-Hispanic/Latino people experience higher rejection rates than other ethnic groups, at 18.25% compared to the highest rates of 40.80% and 40.15% for Cuban and other Hispanic/Latino groups, respectively.

Ethnicity	N	Average Interest Rate	Average Rate Spread	Rejection Rate
Hispanic/Latino	1343854	5.341596	0.9256879	0.2336303
Mexican	45418	5.269539	1.1671613	0.3248492
Puerto Rican	12840	5.072746	0.9663387	0.3602025
Cuban	6321	5.117525	0.9045302	0.4080051
Other Hispanic/Latino	53759	5.048491	0.7755527	0.4014770
Not Hispanic/Latino	10398601	5.024927	0.6421698	0.1824764

Table 3.2: Summary statistics of loan variables, by ethnicity.

The higher rejection rates may not be based on race/ethnicity alone or at all, due to the results found by the NBER researchers in the FinTech paper. However, the paper supports racial/ethnic factors in the fact that loan pricing variables, such as interest rate and rate spread, are higher for minority groups, justifying our exploration of models to test whether this is the case, and if so, the magnitude of this effect on loan pricing.

3.2 Maps

Figure 3.1 displays the map of rejection rates by county across the United States. We observe that rejection rates tend to be highest in Southern areas of the United States, excluding the Southwest. Additionally, rejection rates are generally lower in the Midwest and West.

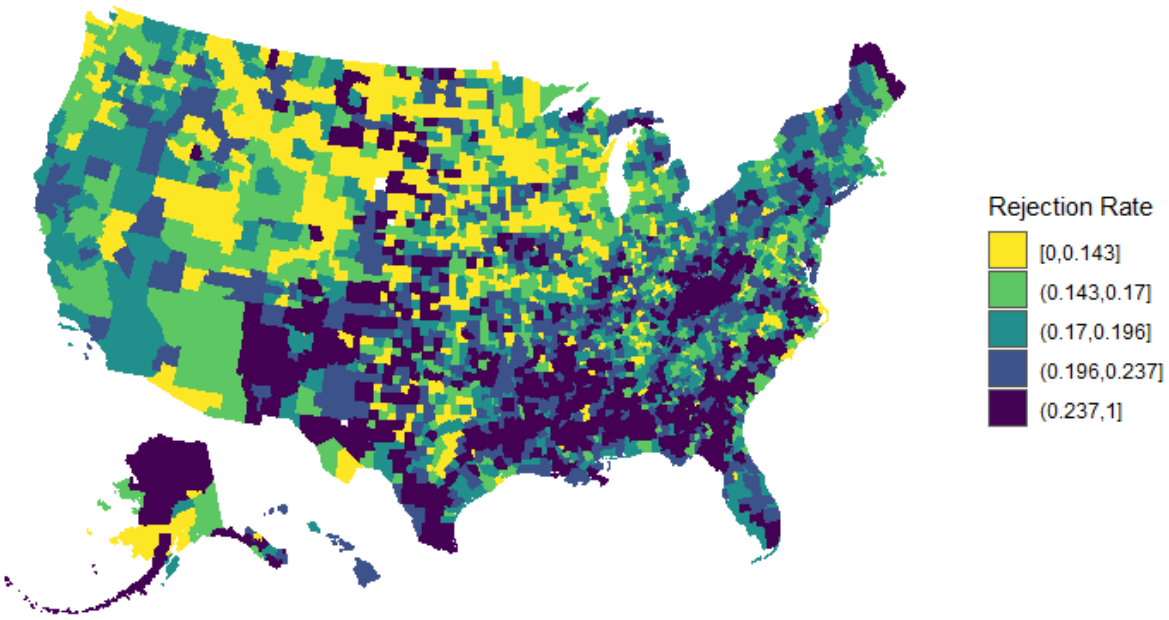

Figure 3.1: Choropleth map of Rejection Rate quintiles in the United States

Figure 3.2 displays the map of median interest rates across the United States, by county. We notice similar patterns to those of the rejection rates, in that the Southern regions tend to have higher interest rates, while being lower in the Midwest and West. We also observe that despite higher rejection rates in areas such as New England, Hawaii, and Alaska, all areas display low interest rates. Finally, we notice that areas in and around Nevada have high interest rates, despite having low rejection rates.

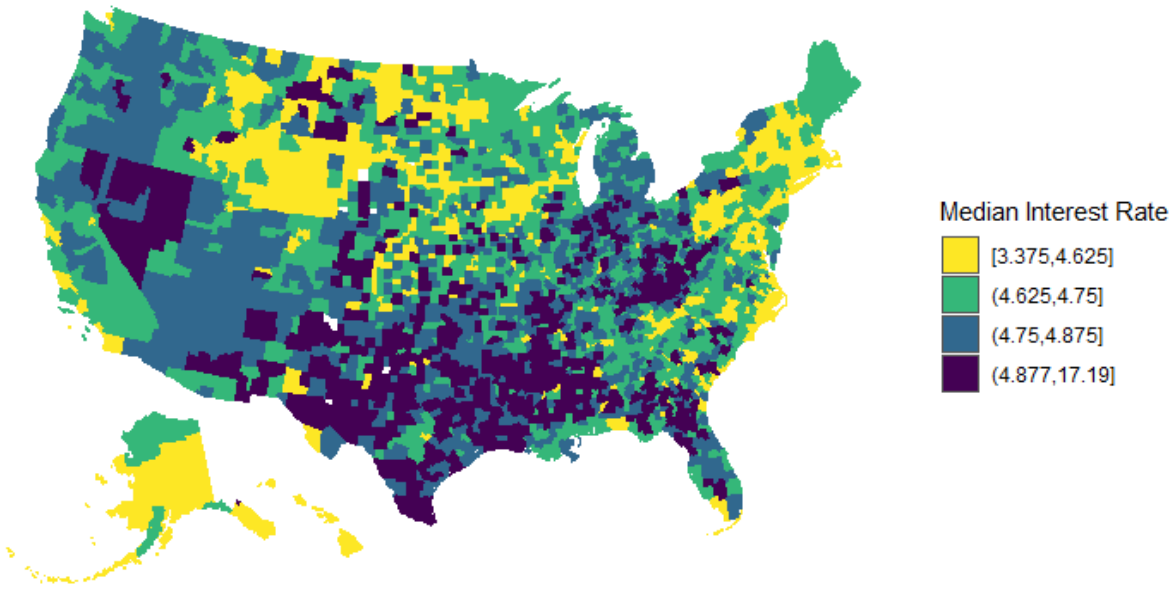

Figure 3.2: Choropleth map of Median Interest Rate quartiles in the United States

Figure 3.3 displays the map of median rate spreads across the United States, by county. The patterns we observe for rate spread are similar to those of interest rates. However, more counties appear to have higher rate spreads, indicating more unfairness in loan pricing than interest rates.

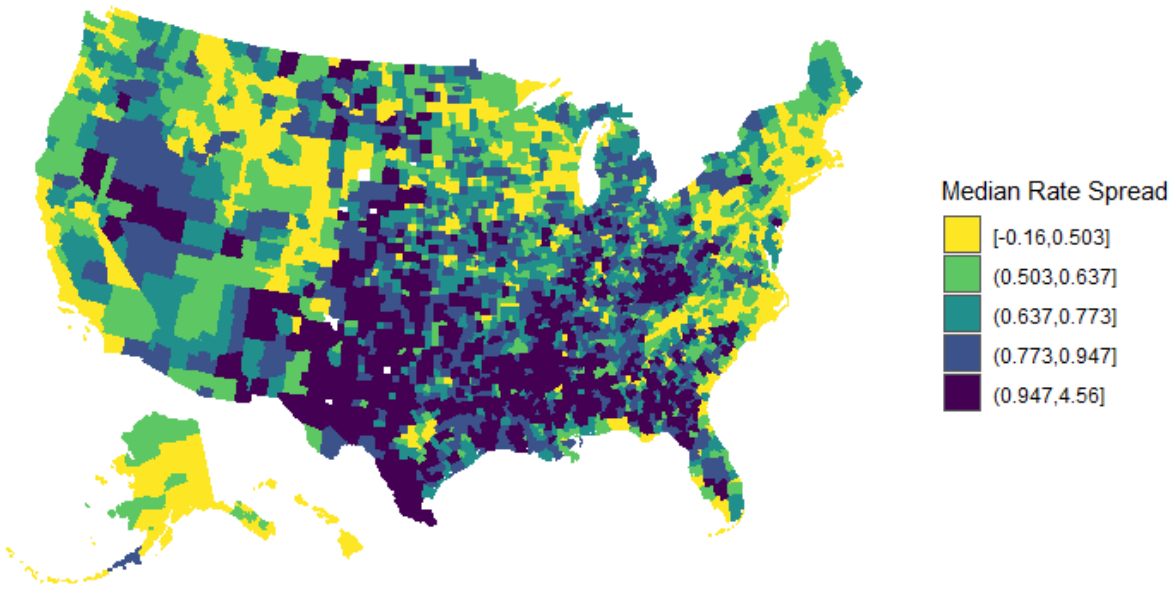

Figure 3.3: Choropleth map of Median Rate Spread quintiles in the United States

Figure 3.4 displays the map of HHIs across the United States, by county. We expect high HHIs to be associated with high loan rejection and pricing; the observed results are mixed. We observe that the Midwest states, the western areas of Texas, Hawaii, and Alaska tend to have the highest market concentration, while the coastal states have lower market concentrations. The coastal areas have low HHIs and low loan rejection and pricing, consistent with our hypothesis. However, the South is mixed in terms of market concentration, and Midwestern areas with low loan rejection and pricing have high market concentrations, contrary to our hypothesis.

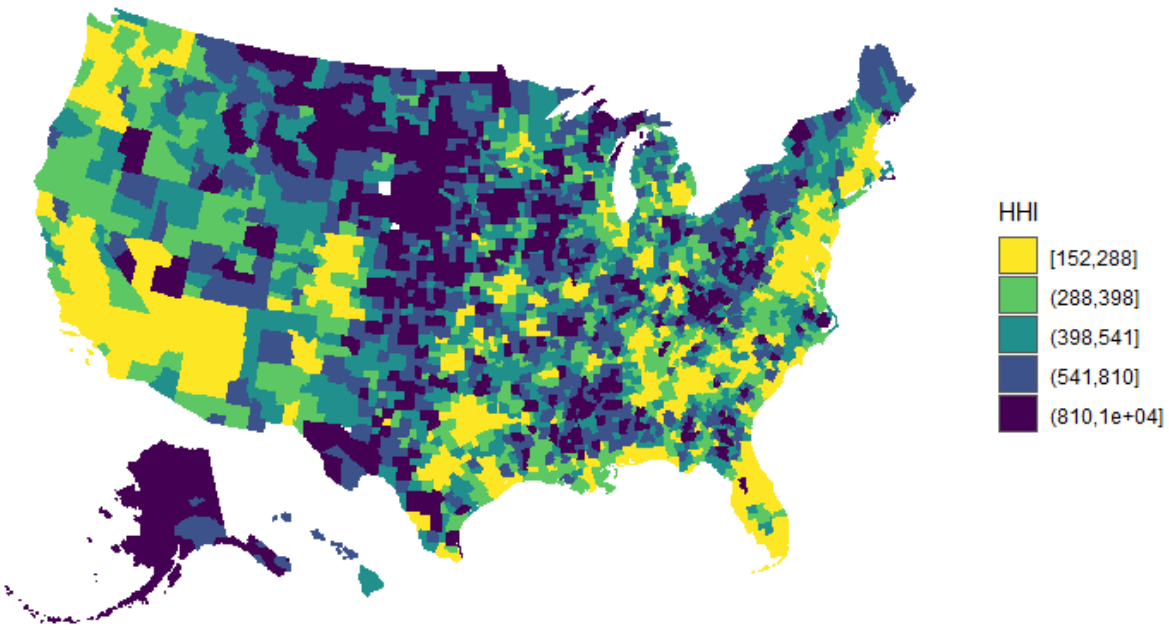

Figure 3.4: Choropleth map of HHI quintiles in the United States

Finally, Figure 3.5 displays the map of the percentage of minorities by county across the United States. We expect areas with high minority concentrations to experience higher loan rejection and pricing. This generally holds, as Southern states have the highest minority concentrations as well as loan rejection and pricing. In addition, Midwestern states and Northern states have the lowest minority concentration, as well as loan rejection and pricing. Hawaii and Alaska also have lower minority concentrations, consistent with low loan pricing.

The maps have given a preliminary glance into geographic correlations of loan pricing and approval with minority percentage and market concentration. We further explore these relationships, and relationships with other predictors, in the following sections.

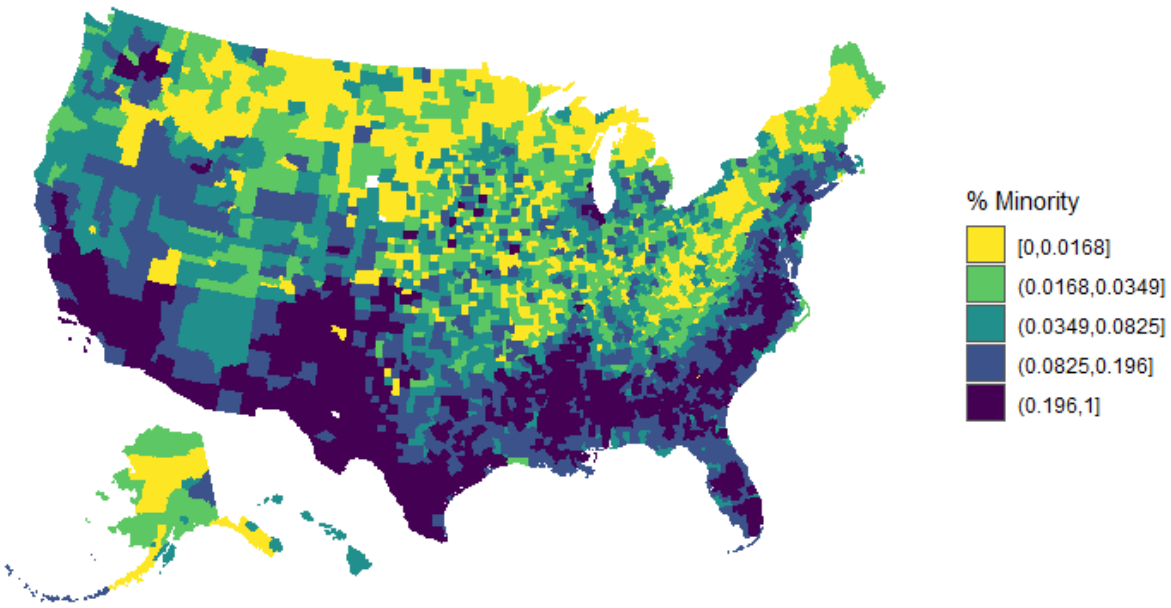

Figure 3.5: Choropleth map of Percent Minority quintiles in the United States

3.3 Scatter Plots

Figure 3.6 displays scatterplots of each predictor variable of interest on the rejection rate for each zip code. First, we observe generally weak relationships with each predictor with rejection rate. However, we observe that demographic variables, such as the percent minority population, percent male, and percent of people in public sector jobs in a zip code shows the strongest relationship with rejection rate, with each having a positive relationship. The credit risk variables seem to have much weaker relationships with rejection rate, as many of the points cluster or scatter randomly. The relationship with the median loan-to-value ratio with the rejection rate of each zip code appears particularly weak. Finally, the zip code HHIs have an unclear relationship with rejection rate; it is possible to argue a positive or no relationship. Given the fact that market

competitiveness would primarily affect pricing, we may expect to see a stronger relationship between the HHIs and the loan pricing variables of interest rates and rate spreads.

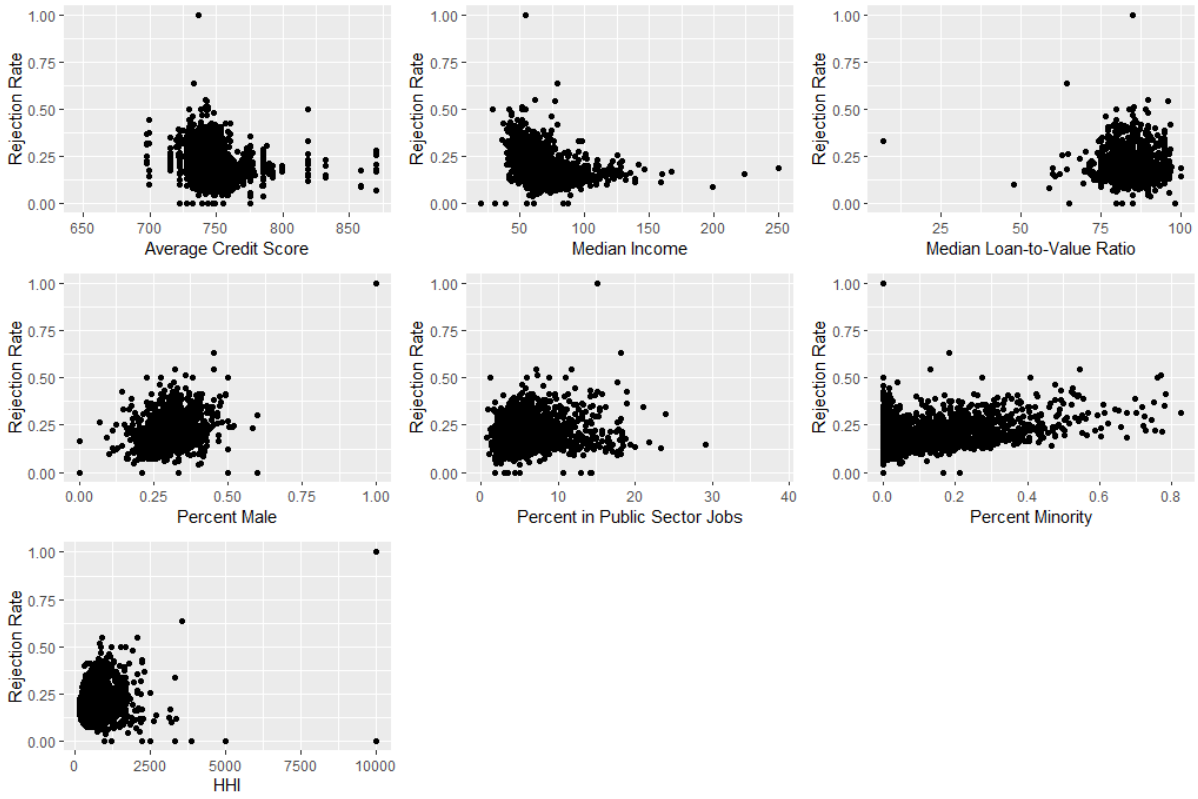

Figure 3.6: Scatter plots of each predictor variable on rejection rate, by zip code

The first loan pricing variable we consider is the median interest rate in each zip code. Figure 3.7 displays scatter plots between this variable and the predictor variables. The demographic variables again show the strongest relationships with the response variable; all relationships are positive. The median income shows a weak negative relationship with interest rates, showing a stronger relationship than with rejection rates. However, other credit risk variables, such as credit score and loan-to-value ratio, do not necessarily display a stronger relationship compared to

rejection rates. Finally, the HHIs display a similar relationship with interest rates compared to rejection rate in that we can argue either a positive or no relationship. This finding aligns with our expectations, as we would expect loan prices to increase as the market competitiveness decreases (in other words, as the HHI increases).

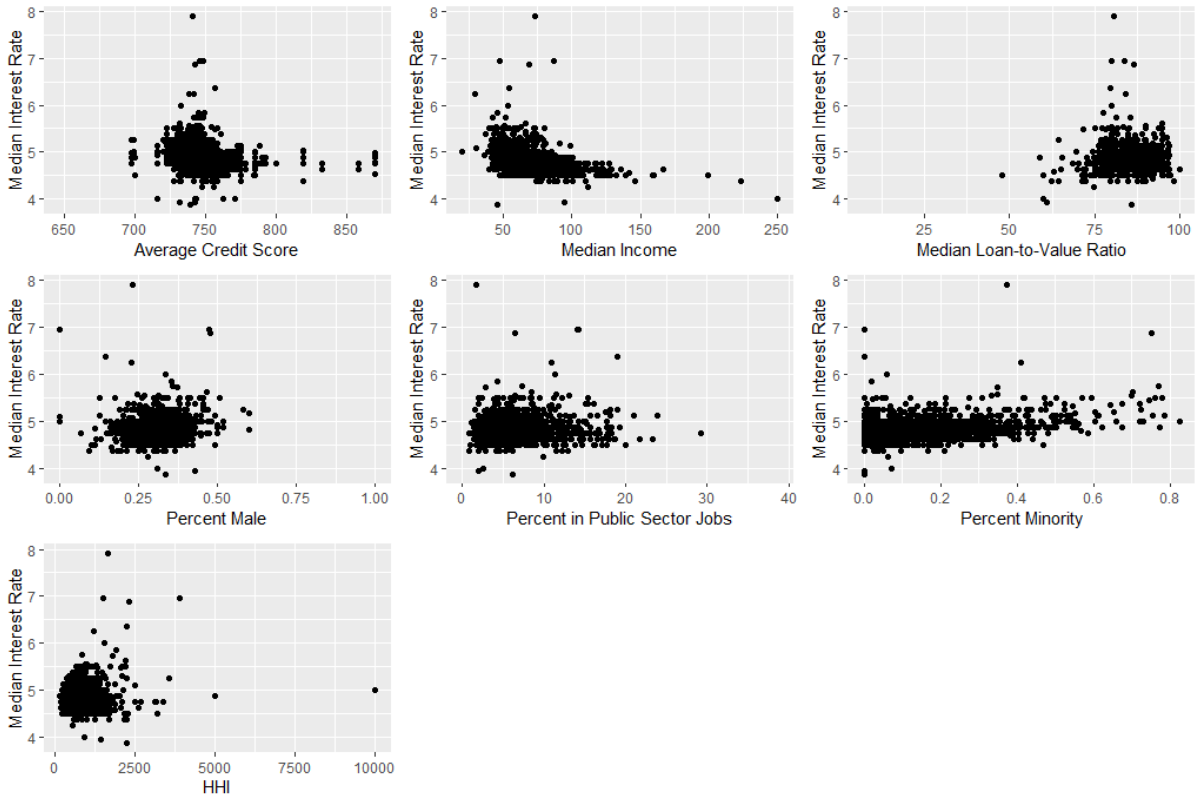

Figure 3.7: Scatter plots of each predictor variable on median interest rate, by zip code

Figure 3.8 displays scatter plots between the median rate spread and the predictor variables. The relationships we observe are similar to those with interest rates in that the strongest relationships are with the demographic variables and median income, with the points scattered slightly more in a random fashion. The relationship with the HHIs appear to have a stronger positive relationship

with rate spread compared to interest rates, which we expect as more competitive areas would price fairer than less competitive areas.

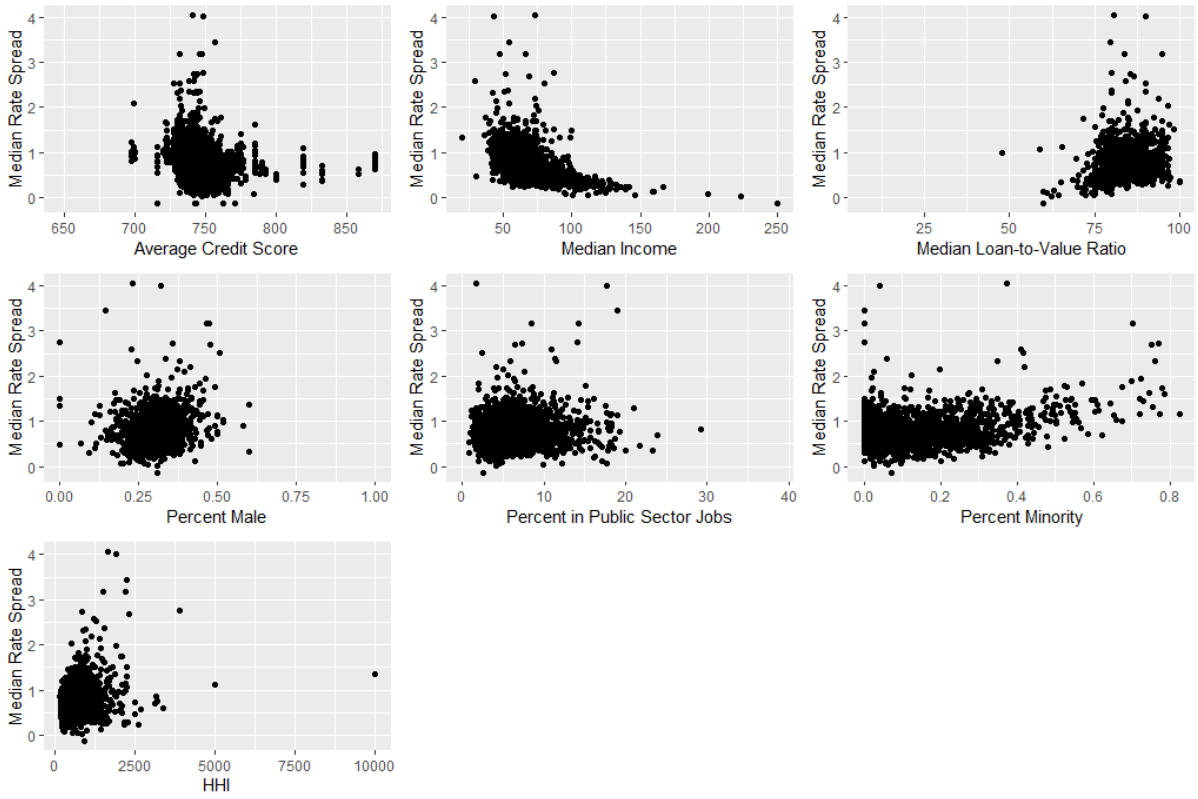

Figure 3.8: Scatter plots of each predictor variable on median rate spread, by zip code

The findings from our scatter plots are in no way conclusive regarding the relationships between the variables. As previously mentioned, the relationships between many of the variables appear ambiguous. In addition, we must consider our predictor variables in the context of the models so that they may control for the effects of each other in stating their own effects; in particular, credit risk variables may prove to account for any relationship we see between the percent minority measure for each zip code and the loan approval/pricing variables. Finally, the relationship

between the HHIs and the response variables may become clearer once it is considered in the context of the percent minority measure by exploring the interaction between the two variables.

3.4 Correlations & Interactions

We generate a correlation plot, shown in Figure 3.9, to quantify the relationships explored in the previous scatterplots. See Table 2.1 for the codebook associating the figure labels with the variables of interest. The plot shows that the correlations between the percent of males in public sector jobs and average credit score, median interest rate, the percent of male in general, and the HHI in each zip code are not statistically significant. Additionally, correlations between the predictor variables appear low, suggesting that our models will not have multicollinearity issues such that the effects of our predictors are cancelled out by each other.

There are two notable relationships: median income and median interest rate (-0.53), as well as median income and median rate spread (-0.61), both of which are negative. This is expected, as income is a commonly known determinant of decreased loan pricing is the income of the borrower. However, the correlations between the response variables and the remaining predictors are lower, particularly with loan-to-value ratio and HHI. These findings support the relationships explored in our scatter plots.

Additionally, there are large correlations between median rate spread and median interest rate (0.83), as well as median rate spread and rejection rate (0.52). This first suggests that interest rates and rate spreads both largely measure loan prices similarly. This also suggests that the factors upon which lenders make loan pricing and approval decisions may be similar as well, as the two variables change with each other, all else held constant.

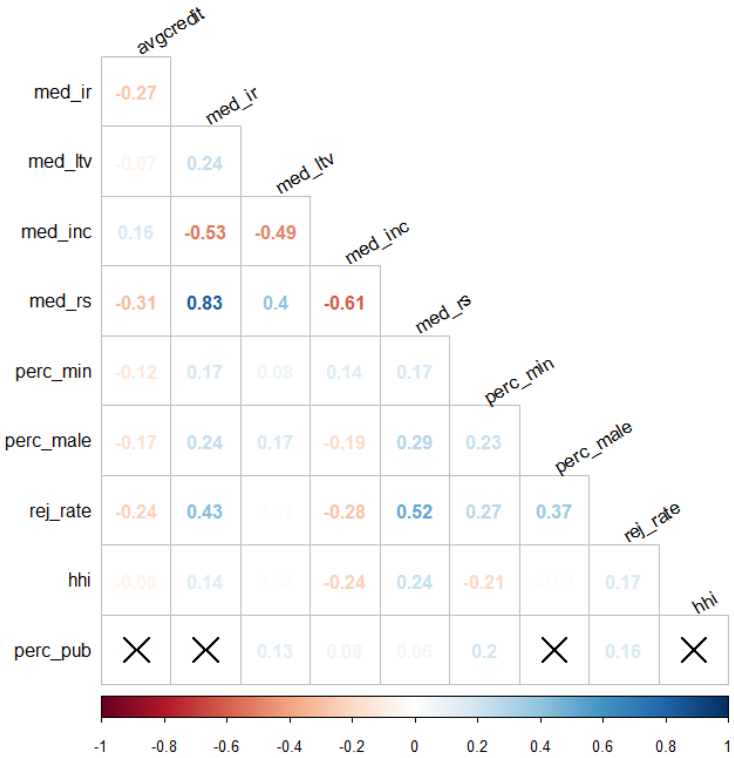

Figure 3.9: Correlation plot between each variable predictor and response variable. The correlations marked with an "X" are not significant at the 5% significance level

Figure 3.10 plots the fits for the loan pricing and approval variables on the percent minority population in each zip code, by HHI categorizations. This plot explores the interaction between the percent minority population and HHI variables, such that we may see any differences in the effect of the percent minority variable on our response variables between different market concentrations. We divide the HHI variable into three categories of market concentration, based on DOJ and FTC guidelines. See Section 2.3.1 for the conditions on which the HHIs were categorized.

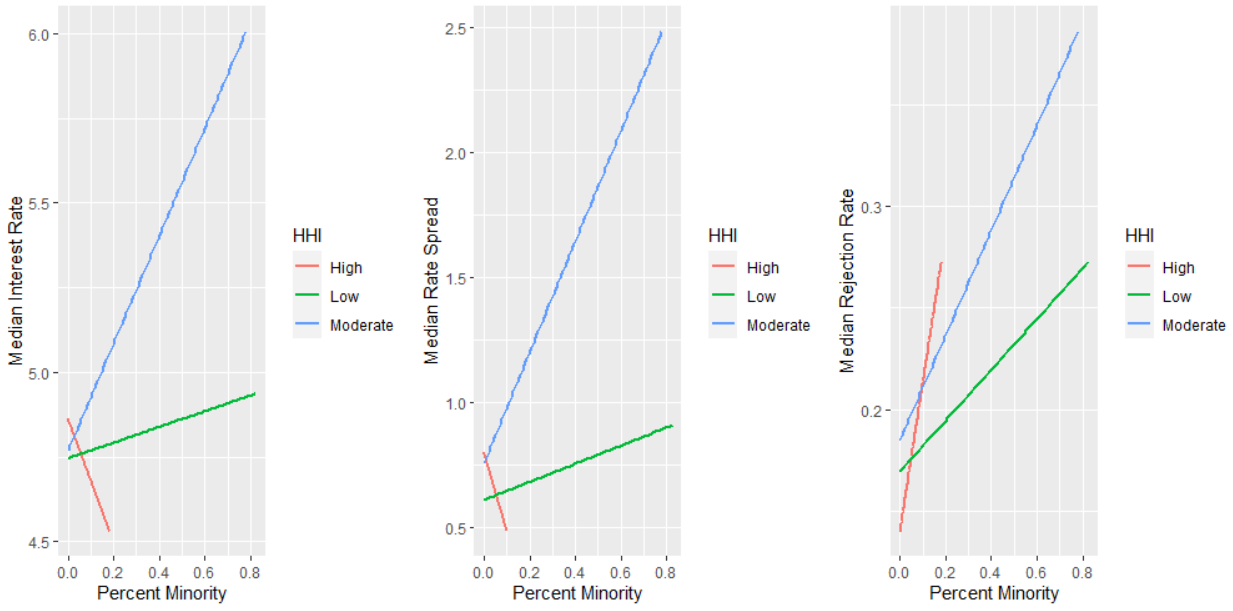

Figure 3.10: Interaction plot displaying the differences in loan price/approval rate changes between mortgage markets with low, moderate, or high levels of competition

The interaction plot supports an interaction between percent minority population and HHI for all three response variables, as the slopes measuring the change in the response variables as the percent minority population increases all visually appear different from each other based on HHI categorization. The most notable finding is that both the loan pricing variables, interest rate and rate spread, decrease as the percent minority population increases in zip codes with a highly concentrated, non-competitive market; however, the opposite is true for zip codes with a less concentrated and more competitive market. This finding is unexpected in the fact that we would expect areas with high market concentrations to discriminate to a greater extent towards minority groups due to the lack of suppliers in the market who would give incentives to behave ethically

and keep loan prices low. We are particularly interested to see whether this phenomenon holds in the formal models.

The other finding from the interaction plot is that as the market concentration increases, the rate at which the rejection rate increases as the percent minority population increases. This supports the claim that lender discrimination against minority groups is exacerbated by the lack of competition in the mortgage market in financial deserts. However, we must look at statistical models to determine whether this interaction truly exists.

Chapter 4: Ordinary Least Squares Regression

We first run three linear regression models using ordinary least squares (OLS) to explore the effects of the percent minority population and HHI on interest rate, rate spread, and rejection rate by zip code, while controlling for the effects of credit risk variables and other demographic variables. These are our "full models," such that each predictor variable is included in the model.

Table 4.1 shows the coefficients for the full models. Coefficients with at least one "*" indicate a statistically significant effect at the 5% level. The intervals given below the coefficients are the 95% confidence intervals for the associated coefficient; any interval containing 0 indicates that we may not conclude that the magnitude of the predictor's effect is different from 0.

We also perform stepwise selection using the Akaike Information Criterion, for the purpose of emphasizing the strongest effects. The stepwise selection builds models using both forward and backward selection; we include forward selection processes to address any concerns that the full models may have given significance to all the predictor variables due to overcomplexity of the models. However, the stepwise selection method produces identical models to the full models.

Overall, the ordinary least squares models concludes that all variables and the interaction influence loan pricing and approval, with the exception of the percent minority count on rate spread. However, this effect on rate spread is still important due to its inclusion in the significant interaction. As the market concentration increases, loan pricing and rejection increases as well. As the percent minority count increases, interest rates and rejection rates increase. Finally, the effect of market concentration on loan pricing and approval increases as the percent minority count increases. These findings support the literature in its justification of utilizing all these factors to control for credit risk and market competitiveness.

	Med. Int. Rate	Med. Rate Spread	Rej. Rate
Avg. Credit Score	-0.0025 ***	-0.0048 ***	-0.0006 ***
	([-0.0026, -0.0024])	([-0.0050, -0.0046])	([-0.0007, -0.0006])
Med. LTV Ratio	-0.0031 ***	0.0052 ***	-0.0033 ***
	([-0.0035, -0.0028])	([0.0047, 0.0056])	([-0.0034, -0.0032])
Med. Income	-0.0045 ***	-0.0067 ***	-0.0009 ***
	([-0.0046, -0.0044])	([-0.0068, -0.0066])	([-0.0010, -0.0009])
% Minority	0.0405 ***	-0.0142	0.0525 ***
	([0.0228, 0.0581])	([-0.0395, 0.0111])	([0.0467, 0.0584])
% Male	0.2449 ***	0.6637 ***	0.3247 ***
	([0.2137, 0.2761])	([0.6189, 0.7084])	([0.3144, 0.3350])
HHI	-0.0000 ***	0.0000 ***	0.0000 ***
	([-0.0000, -0.0000])	([0.0000, 0.0001])	([0.0000, 0.0000])
% Public Sector	0.0022 ***	0.0061 ***	0.0038 ***
	([0.0018, 0.0027])	([0.0055, 0.0068])	([0.0036, 0.0039])
% Minority * HHI	0.0006 ***	0.0011 ***	0.0002 ***
	([0.0006, 0.0007])	([0.0010, 0.0011])	([0.0001, 0.0002])
N	46306	46306	46306
R2	0.3955	0.5228	0.3560

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 4.1: Regression Table of credit risk, demographic variables, and market competitiveness variables on interest rate, rate spread, and rejection rate by zip code

Figure 4.1, Figure 4.2, and Figure 4.3 are residual diagnostic plots for the median interest rate, median rate spread, and rejection rate models, respectively. Statisticians often use diagnostic plots to check the assumptions of their linear models. The residuals should show a linear constant trend (linearity assumption) in the "Residuals vs Fitted" plot, such that a linear fit is in fact the right fit for our data. The residuals should also follow the dotted line in the "Normal Q-Q" plot (normality assumption). Finally, we check that the residuals are equivariant, or equally spread, around the fitted line; the red line in the "Scale-Location" plot should be close to horizontal (equivariance assumption).

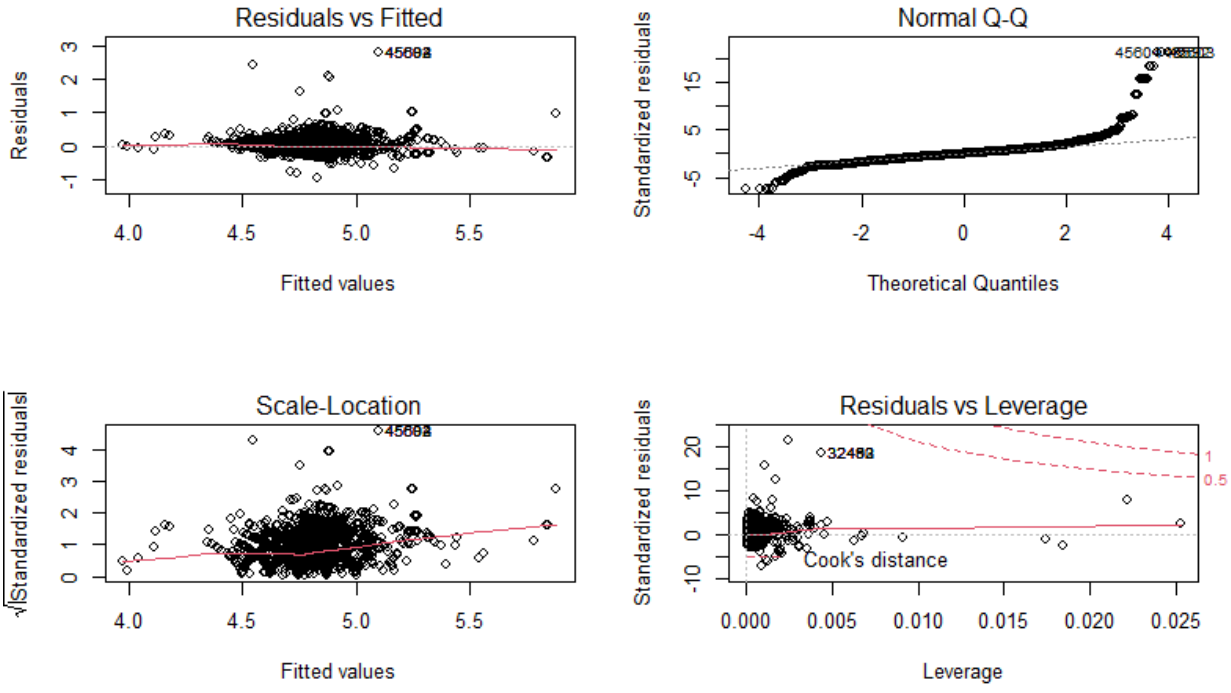

Figure 4.1: Residual Plots for the median interest rate models

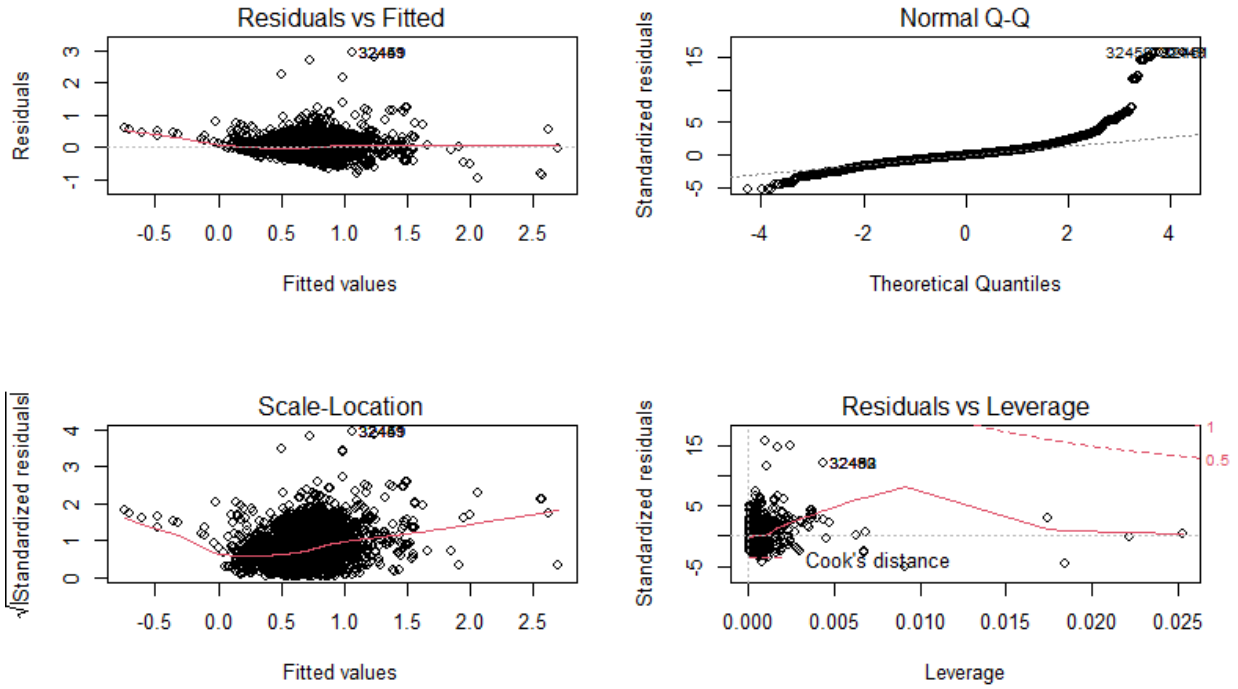

Figure 4.2: Residual Plots for the median rate spread models

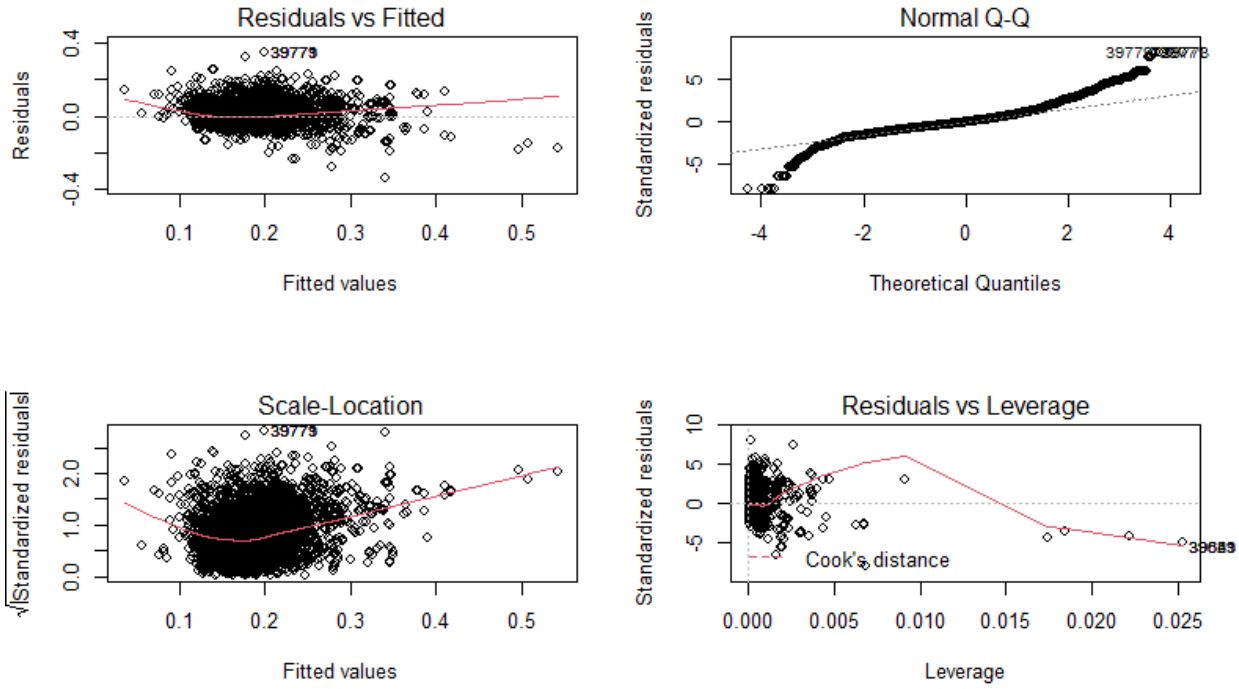

Figure 4.3: Residual Plots for the rejection rate models

The linearity assumption appears to be met in all three models. However, the normality assumption appears to be violated in all models, as well as the equivariance assumption in the loan pricing models. These violations may render our models less useful, as the coefficient estimates may be biased. We may obtain more informative results from using a more flexible fit, such as a regression tree.

We also observe the Residuals vs Leverage for each model, which determines whether any points exist which has largely influenced the model fit. These points will typically appear at the top or bottom right corners of the plot, outside the range of the dotted red lines (Cook's Distance). According to the plots, none of our models contain any influential points of concern which would suggest model instability.

A point of concern with the models is that the observations are not necessarily independent. Zip codes which are close to each other likely have similar demographic characteristics and market conditions. This may explain why our model coefficients are showing a great degree of significance; however, it becomes difficult to research and omit near duplicate zip codes without researching the economic context of every subsection of the United States. Utilizing the regularization and regression tree methods may help in identifying the variables whose effects are more important in explaining loan pricing and approval.

Chapter 5: Regularization

The second group of models are regularized using the LASSO shrinkage method, as outlined in Section 2.4. In analyzing the regularized models, we seek to understand which factors are not important in their effects on loan pricing and approval; these coefficients will shrink to exactly 0. In a ten-fold cross-validation procedure to determine the optimal value of λ (which minimizes the loss function), the optimal parameters for each model results in similar results as the OLS models. The optimal λ values produce the models displayed in Table A.1. All of the coefficients, except for that the percentage minority count, do not shrink to 0; this signifies that the variables significant in the OLS models are all important, and therefore should be kept.

We see that the optimal lambda values do not produce models which address the issue of reducing model complexity to identify the most important effects. Therefore, we use a standard approach of choosing the largest λ for which the cross-validation error will be within 1 standard error of that of the optimal λ (Krstajic et. al.). Larger values of λ increases the magnitude of the penalty term in the loss function, and thus forces more coefficients to shrink to 0. For more information on parameter selection, see Figure A.1, Figure A.2, and Figure A.3 showing the mean-squared errors (CV errors) from the λ values tested in the cross-validation process. The left vertical dashed line denotes the optimal λ, and the right vertical dashed line denotes the λ within one standard error of the minimum CV error.

The new λ values produce the models in Table 5.1. We observe that the rate spread and rejection rate models draws the same conclusions of the OLS models even with the parameter adjustment; however, the interest rate model has been simplified to include less predictors. The interest rate model shows that credit score, income, the percent minority count, the percent male count, and

the interaction between the percent minority count and HHI are the important effects on interest rates in a zip code. Due to the inclusion of the interaction, we would claim that the individual effects are important despite the HHI coefficient shrinking to 0. The model leaves out the percent count of people working in the public sector, indicating that this is not a necessary control for determining the effects of being a minority or market competitiveness on interest rates.

	Med. Int. Rate	Med. Rate Spread	Rej. Rate
Avg. Credit Score	-0.0020501850	-3.768148e-03	-5.553003e-04
Med. LTV Ratio	0	3.576090e-03	-2.423246e-03
Med. Income	-0.0038030062	-6.498355e-03	-7.813823e-04
% Minority	0.0242617603	0	2.968322e-02
% Male	0.1514939442	4.130421e-01	2.958075e-01
HHI	0	1.229922e-05	4.264190e-06
% Public Sector	0	1.438323e-03	2.999564e-03
% Minority * HHI	0.0005508956	9.517396e-04	1.669574e-04
N	46306	46306	46306

Table 5.1: Coefficient results from LASSO shrinkage methods, using the maximum lambda producing a CV error within 1 standard error

The regularized models serve as a robustness check for the OLS models. We can draw the same conclusion that loan pricing and rejection increases as the percent minority count or market concentration increases. These models also support the conclusion that the effects of these two variables of interest exacerbate each other.

Chapter 6: Regression Tree

The final group of models are regression trees, as outlined in Section 2.4. The full regression trees appear in Appendix B: as Figure B.1, Figure B.2, and Figure B.3. Each oval represents a node of the tree, with the top number representing the loan pricing or approval prediction given the classification at the node and the bottom number representing the percentage of the sample falling in the given classification. For example, in Figure B.1, we can interpret the leftmost terminal node by following the branches connecting to it: if a zip code median income is greater than or equal to 61, greater than or equal to 103, and the HHI is greater than or equal to 526, we can predict the median interest rate to be 4.1%, with 1% of the sample falling within this classification.

These trees display possible overfitting issues, as some of the nodes contain close to 0% of the sample. The cross-validation process of regression tree analysis can help with this issue; similarly to the regularization models, we may examine how the model error changes as the complexity parameter changes. The complexity parameter is similar to λ from the regularization methods in determining the severity of the penalty term of the loss function. As the complexity parameter increases, the model will simplify, and thus generalize to more applicable conclusions. As in the previous section, we seek to find the model with the largest complexity parameter giving an error within 1 standard error of the optimal model; we do so by finding the complexity parameter cutoffs denoted by the dashed horizontal line in Figure B.4, Figure B.5, Figure B.6, and applying these cutoffs to a tree pruning mechanism.

Figure 6.1 displays the pruned regression tree for predicting interest rates. We still observe that there may be issues with overfitting, as 3 of the terminal nodes contain close to 0% of the

observations. However, we still observe that the percent minority count and HHI positively affect the interest rates in most cases. If we observe terminal nodes 5-8 (from the left), we notice that zip codes with high minority percentages see interest rates that can be exacerbated by market concentration, and vice versa. Given that the median income is greater than or equal to 103,000 and the average credit score is greater than or euqal to 744, a market with a minority percentage of over 5.9% is predicted to see 4.8% interest rates in competitive markets (HHI < 1120), whereas a non-competitive market (HHI >= 1120) is predicted to have interest rates ranging from 5-7.4%, within which the prediction falls in the higher side of the range with increased minority percentages.

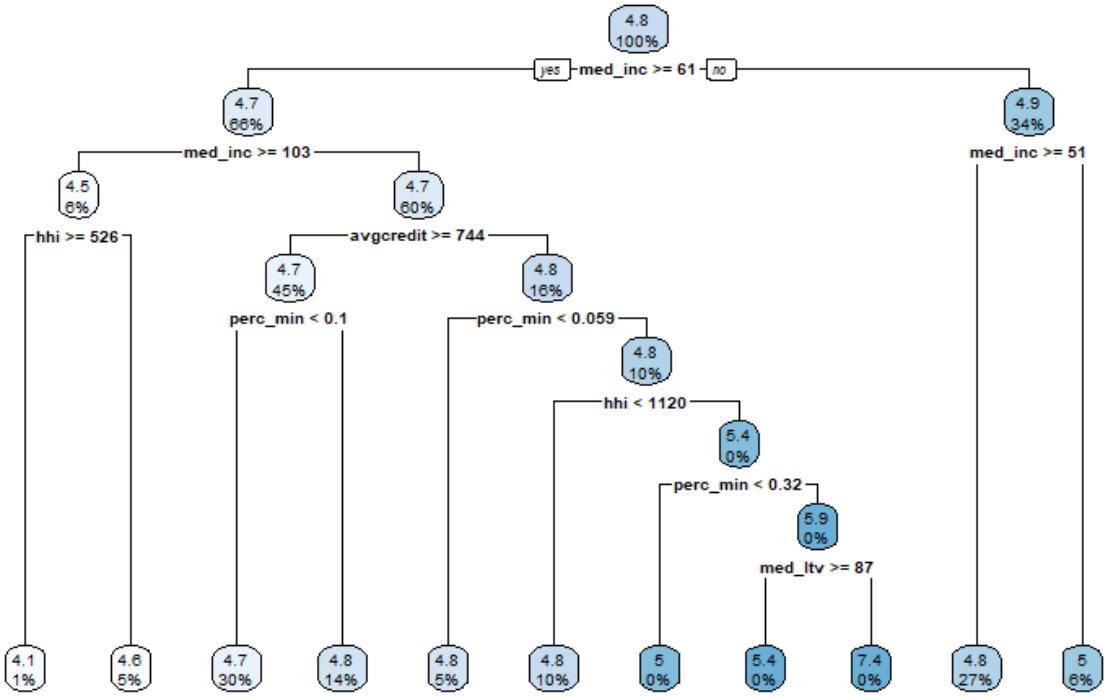

Figure 6.1: Pruned Regression Tree for interest rate

We seek to verify the effects on loan pricing through Figure 6.2, which displays the pruned regression tree for rate spread. As in the previous group of models, we observe that the zip code

minority percentage does not have an effect on determining rate spread, and therefore no interaction exists between minority percentage and market competitiveness. However, the market concentration has a positive effect on rate spread, as given that the median income is less than 55,000 and the percent of people in public sector jobs is greater than 7.1%, the rate spread for a market with an HHI < 934 is predicted to be 1%, whereas this increases to 1.6-3.8% for markets with a larger HHI. We should note that these results may not generalize, as the sample percentages in these classifications are small.

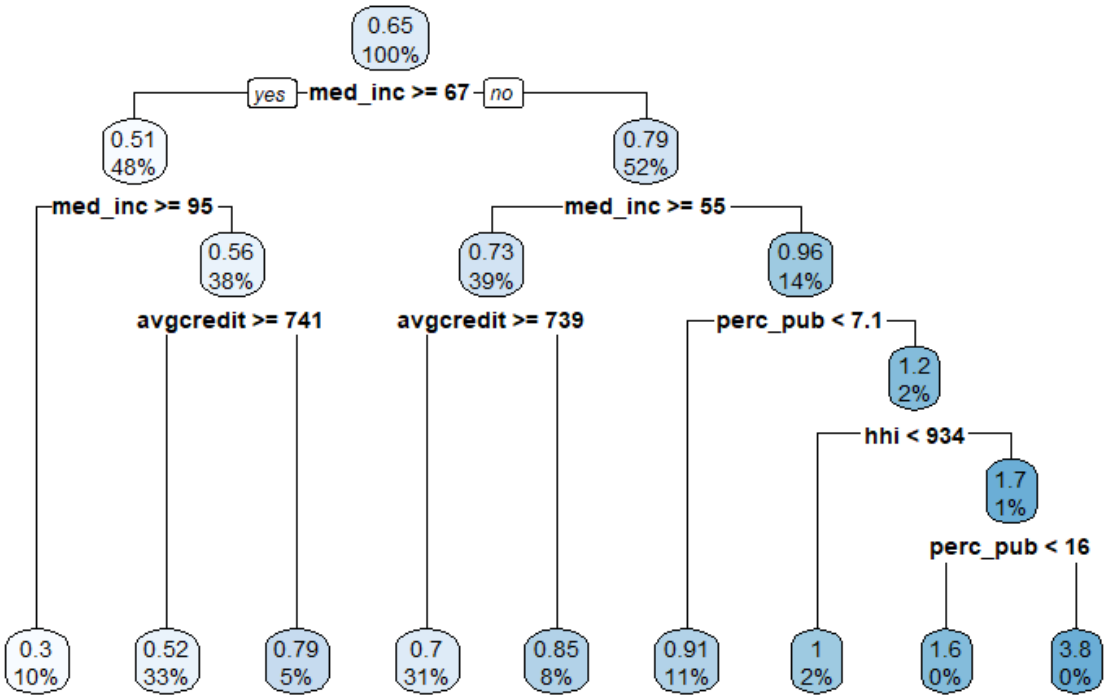

Figure 6.2: Pruned Regression Tree for rate spread

Finally, we analyze the effects of minority percentage and market competitiveness on rejection rates through Figure 6.3. The first finding from this model is that market competitiveness does not affect loan approval, and thus there is no interaction between minority percentage and market

competitiveness. This contradicts our regression models, but this may mean that market competitiveness in the mortgage market plays less of a role in market entry than market pricing. The second finding from the model is that the percent minority count generally has mixed effects on loan approval. If the median income is less than 55,000, the zip codes with a minority percentage greater than 17% see predicted rejection rates ranging from 27-34%. If the median income is greater than or equal to 55,000 and the percentage of males is less than 29%, the markets with a minority percentage greater than 14% observe a rejection rate of 18%, compared to 14-17% for markets with a smaller minority percentage. Due to the mixed ranges for markets with a median income greater than 55,000, we cannot necessarily conclude loan approval discrimination.

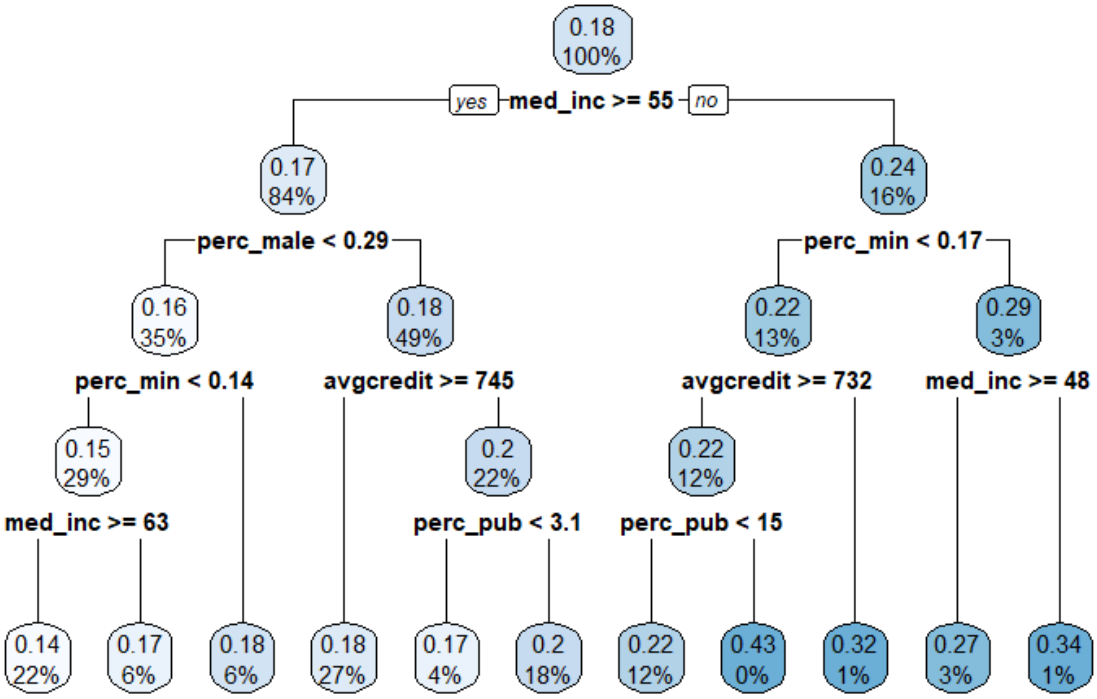

Figure 6.3: Pruned Regression Tree for rejection rate

Chapter 7: Conclusion

The findings from our models provide mixed support to our hypotheses as well as previous findings from the NBER FinTech study. First, we find that the minority percentage and market competitiveness are associated with increases in interest rates. Loan pricing discrimination in terms of interest rates also appears to be exacerbated by areas that are financial deserts, indicating that there is validity in the NBER hypothesis that financial deserts serve as an inadvertent avenue of discrimination.

However, these relationships do not hold with rate spreads; in fact, there is no effect of minorities on rate spread. This may indicate that rate spread may not be as good of an indicator of loan pricing as expected. If this is the assumption we can make, then we may generally claim the being a minority and living in a financial desert has a positive relationship with loan pricing. Future research should explore the causes of this discrepancy and explore the validity of putting more weight on the conclusions of the interest rate models.

The lack of a relationship between being a minority and rate spread may also be valid and a topic worth studying. This outcome could make sense considering the fact that algorithmic lending has been implemented with the intention to reduce discrimination, due to the incentives caused by laws forbidding discriminatory behavior. However, if this were the case, one might expect this to apply to interest rates as well. Ultimately, the effects of race/ethnicity or financial deserts may not be the most important in determining loan prices as shown by the regression trees, including the one for interest rates. These trees show that income and credit score play the largest role in loan pricing, which is ideal and the behavior we want to observe.

Finally, we find mixed results on the relationship between being a minority and living in a financial desert on loan approval. The OLS models and regularized models support discrimination based on race and ethnicity as well as its interaction with living in a financial desert, but the regression tree models support do not. We are inclined to draw from the conclusions of the regression trees, as these models show more flexibility and support the conclusions from the NBER FinTech study.

There exist several limitations to this study. The first is the fact that we used data aggregation to build the data set. This lowers the sample size, as well as erases information present in the original data. The second limitation is the possibility of non-independence, as we expect zip codes in close proximity to share similar demographic, financial, and market characteristics, overstating the magnitude of the observed relationships. The third limitation is that the OLS and regularized models may ignore information from less competitive areas due to missing data. Data that is missing likely indicates a lack of infrastructure in the given zip codes to record this information. The fourth limitation is the possibility of overfitting – models may not generalize. This limitation is less of an issue since we seek to observe general effects, not make precise predictions. The fifth limitation is that single regression trees may not be robust on their own, as the method is fairly prone to providing mixed results. Finally, this paper is an observational study, meaning that we observe characteristics without being able to implement treatments onto the subjects or zip codes. This issue is common in most problems of economics, and thus we cannot conclude causality due to the fact that we cannot perfectly control for external factors.

Future replications of this research should seek to gain access to more data such that loan characteristics and performance, borrower characteristics, and credit risk variables can be combined on a loan-level. This will increase the sample size and information of the data,

increasing the power and accuracy of our models. Loan-level data would potentially address independence issues as well since individuals are likely less similar to each other than zip codes. In addition, future studies should produce models of regression forests or bootstrap aggregated trees to balance the results towards a consistent outcome. Both considerations would potentially address many of the limitations and allow our conclusions to be more robust.

Appendix A: Regularization Additional Materials

	Med. Int. Rate	Med. Rate Spread	Rej. Rate
Avg. Credit Score	-2.499298e-03	-4.749602e-03	-6.361335e-04
Med. LTV Ratio	-2.991260e-03	5.062625e-03	-3.281839e-03
Med. Income	-4.481967e-03	-6.711906e-03	-9.289250e-04
% Minority	4.129268e-02	0	5.167387e-02
% Male	2.420526e-01	6.503845e-01	3.234249e-01
HHI	-2.056199e-05	4.769185e-05	9.195660e-06
% Public Sector	2.134775e-03	5.916132e-03	3.726848e-03
% Minority * HHI	6.301051e-04	1.038211e-03	1.529569e-04
N	46306	46306	46306

Table A.1: Coefficient results from LASSO shrinkage methods, using the optimal lambda minimizing CV error

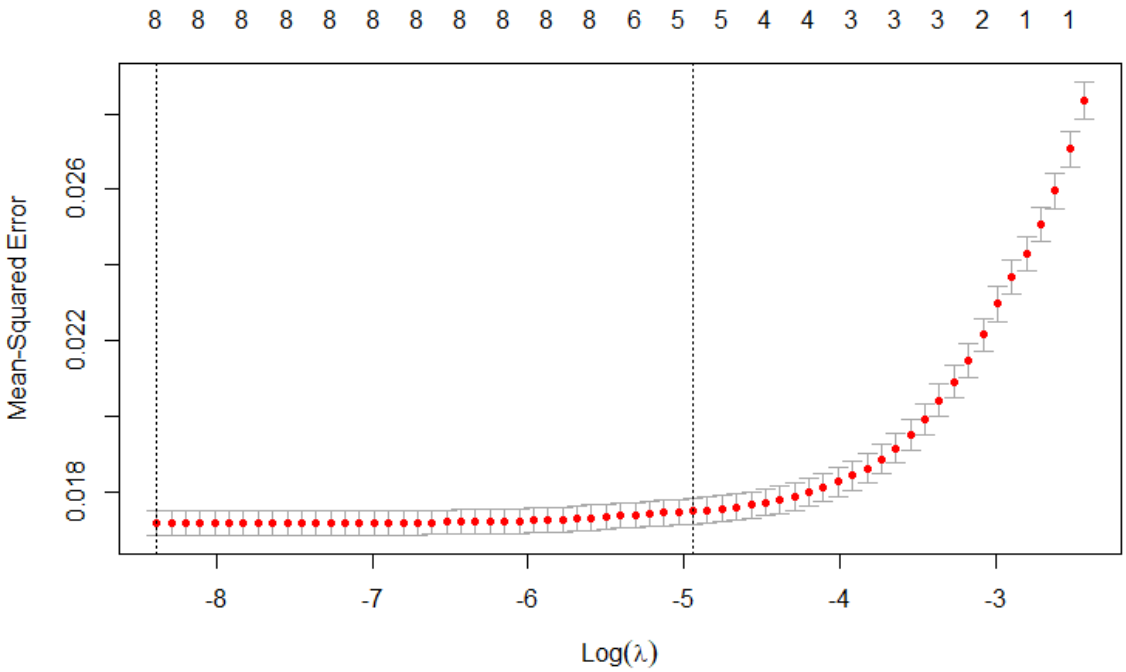

Figure A.1: Cross-validation plot for the interest rate model

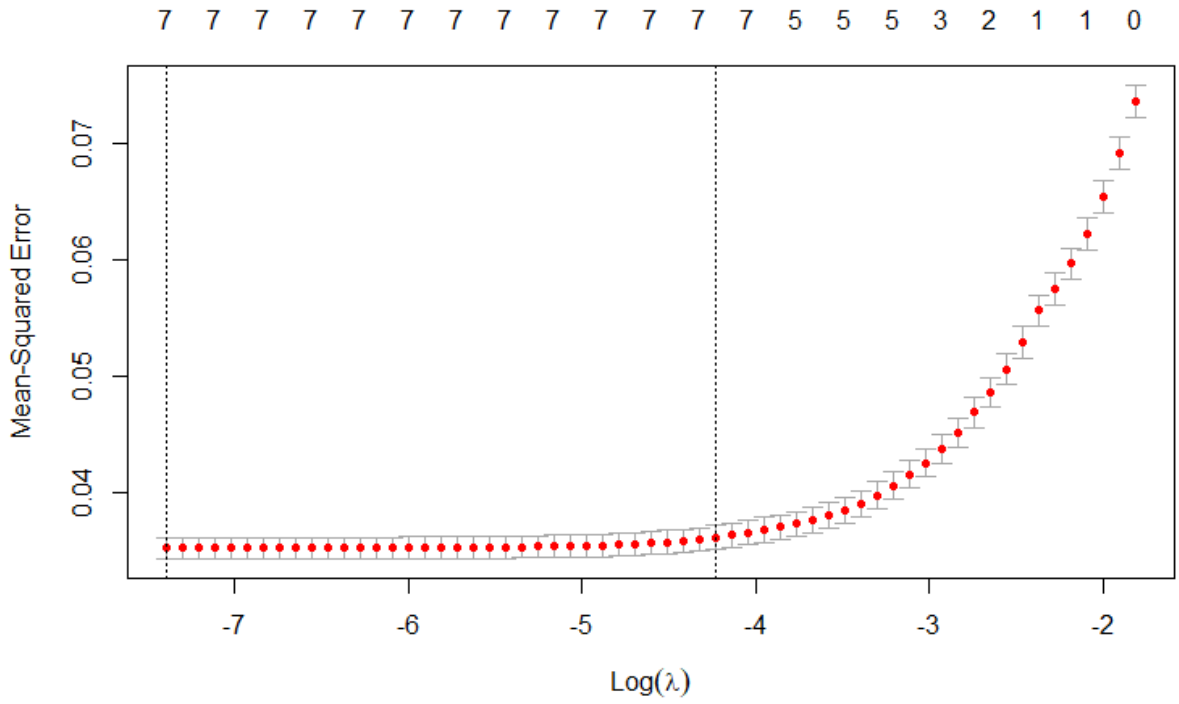

Figure A.2: Cross-validation plot for the rate spread model

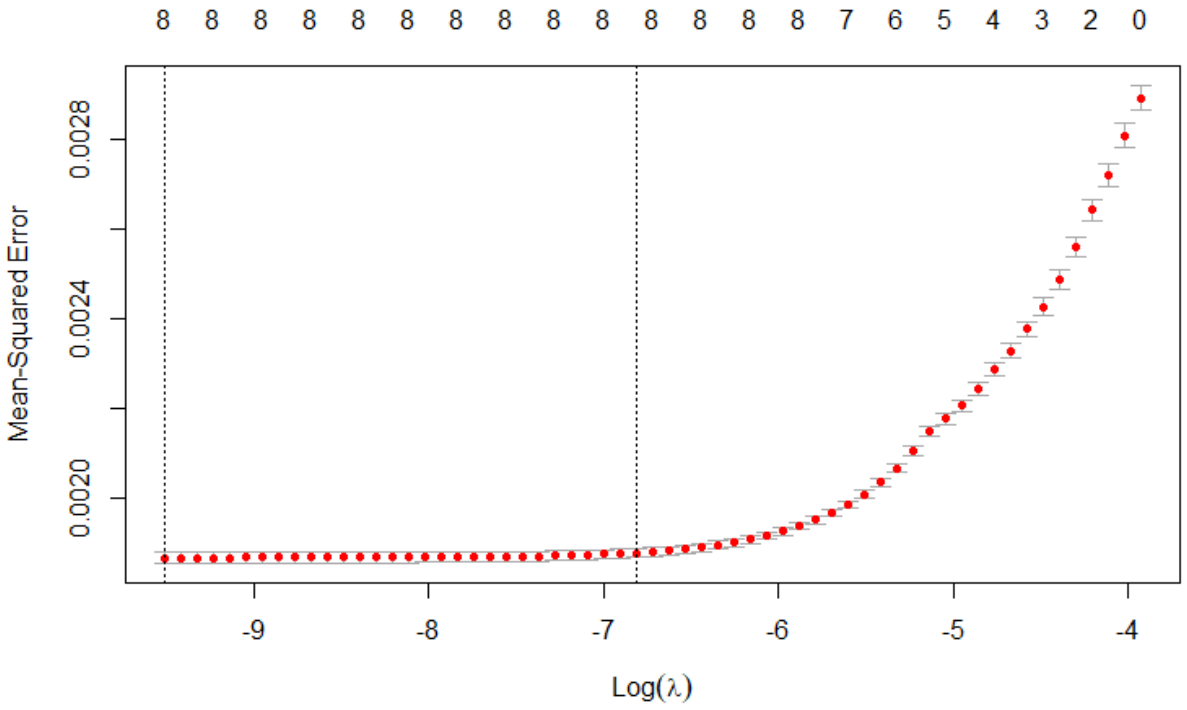

Figure A.3: Cross-validation plot for the rejection rate model

Appendix B: Regression Tree Additional Materials

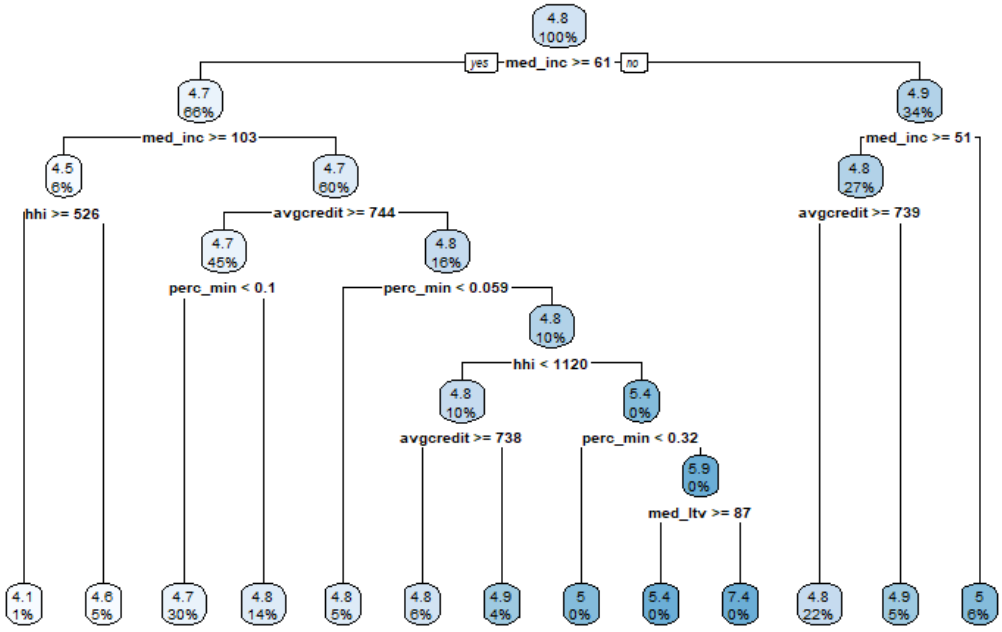

Figure B.1: Full Regression Tree for interest rate

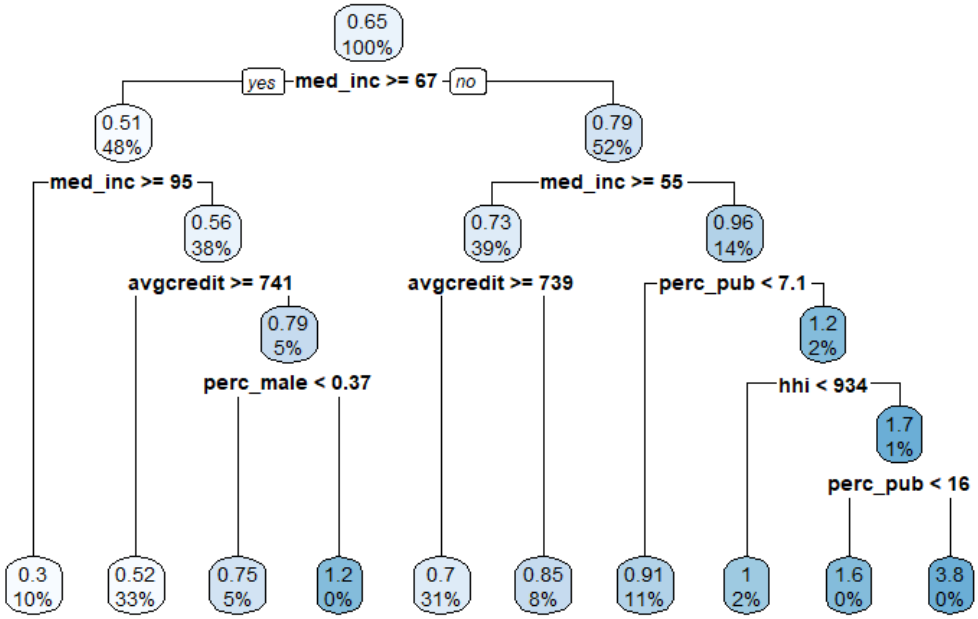

Figure B.2: Full Regression Tree for rate spread

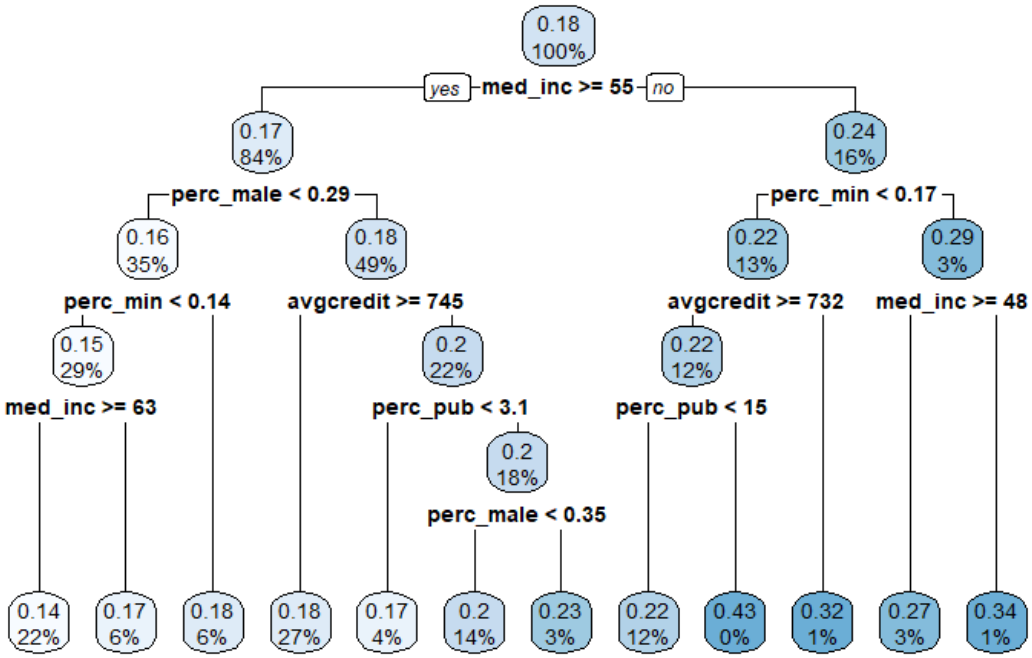

Figure B.3: Full Regression Tree for rejection rate

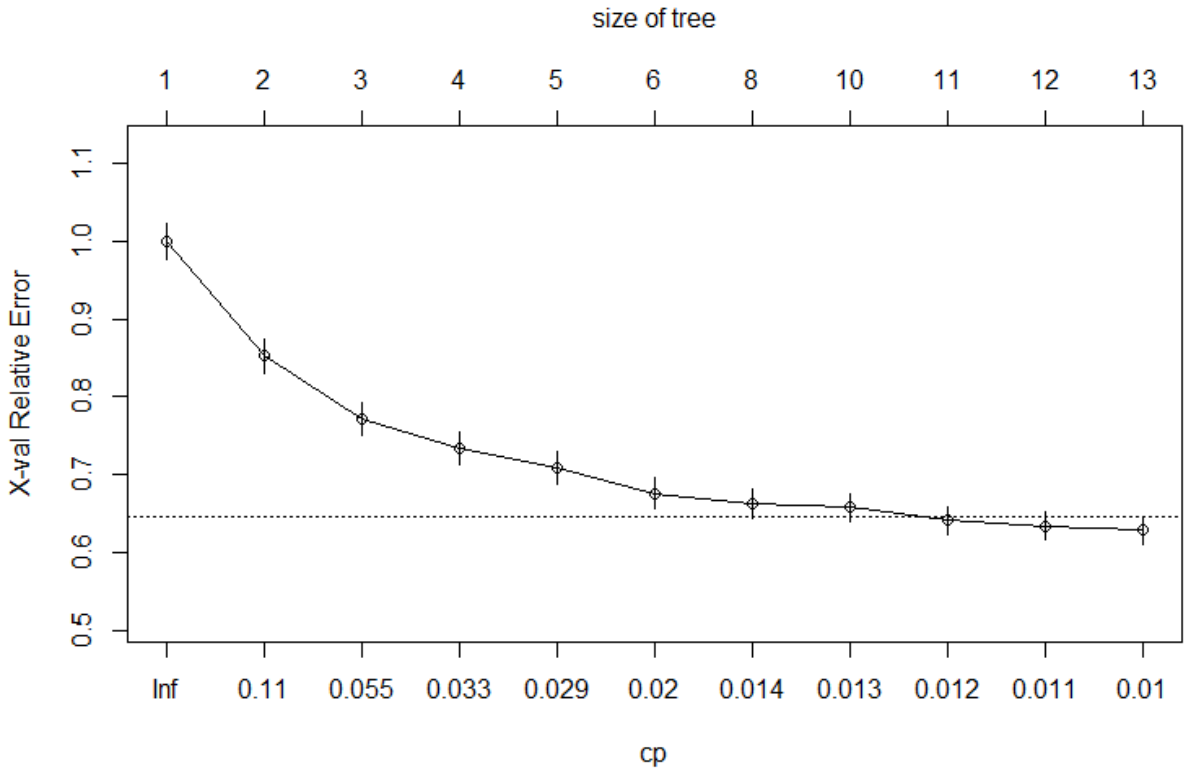

Figure B.4: Plot of errors from different complexity parameters in the cross-validation process for interest rate

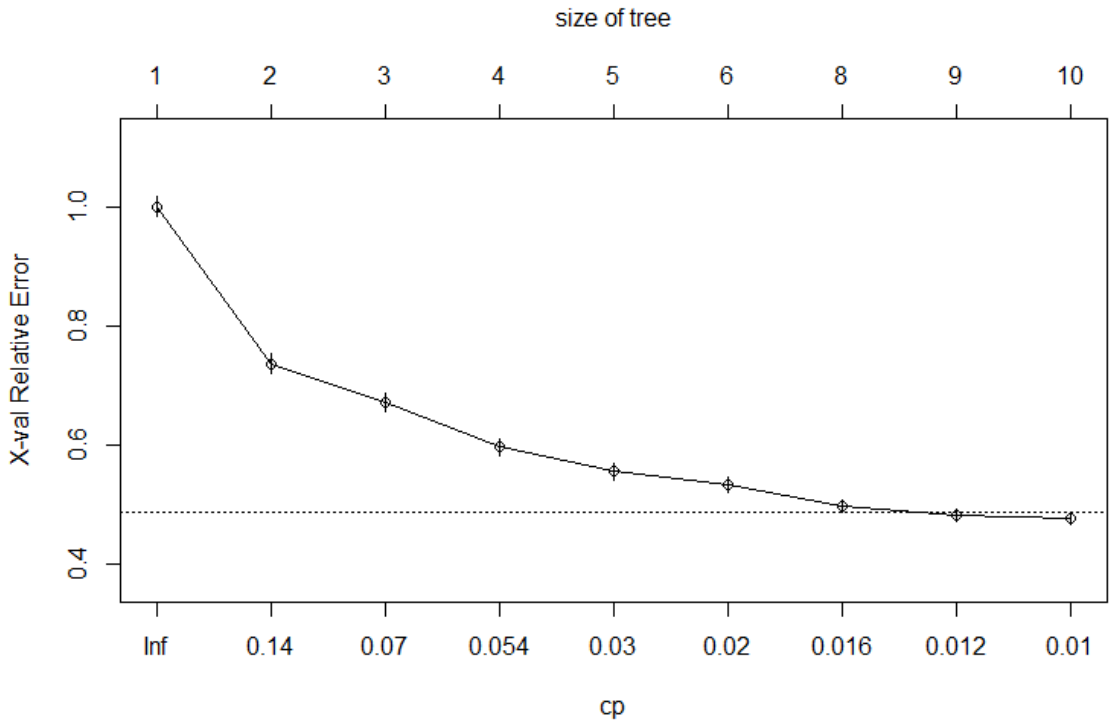

Figure B.5: Plot of errors from different complexity parameters in the cross-validation process for rate spread

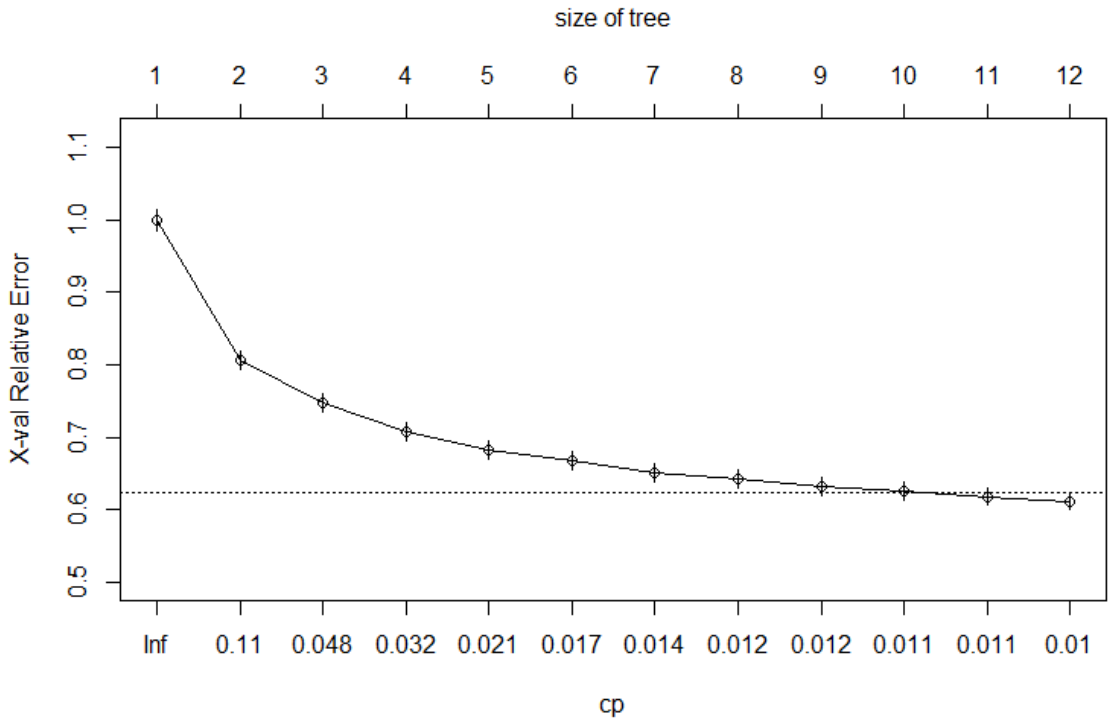

Figure B.6: Plot of errors from different complexity parameters in the cross-validation process for rejection rate

Bibliography

Aaronson, D., Hartley, D., & Mazumder, B. (2017). The effects of the 1930s HOLC "redlining" maps. *EconStor, Working Paper*(No. 2017-12).

About Freddie Mac. (2021). Retrieved March 05, 2021, from http://www.freddiemac.com/about/#

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). Consumer-lending discrimination in the fintech era. *NBER Working Paper Series*. doi:10.3386/w25943

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall, Wadsworth, New York.

Chapman, J. M. (1940). Factors Affecting Credit Risk in Personal Lending. *National Bureau of Economic Research, Commercial Banks and Consumer Instalment Credit*, 109-139.

Cremona, M. A. (2018, November 05). *Shrinkage methods: Ridge regression, Lasso*. Lecture, State College, PA.

FFIEC Rate Spread Calculator. (2016, September 19). Retrieved March 05, 2021, from https://www.ffiec.gov/ratespread/newcalchelp.aspx

Hayes, A. (2020, October 15). Loan-to-Value (LTV) Ratio. Retrieved March 05, 2021, from https://www.investopedia.com/terms/l/loantovalue.asp

Hollister, M. (2011). Employment stability in the U.S. labor Market: Rhetoric versus reality. *Annual Review of Sociology, 37*(1), 305-324. doi:10.1146/annurev-soc-081309-150042

How are fico scores calculated? (2019, November 19). Retrieved March 05, 2021, from https://www.myfico.com/credit-education/whats-in-your-credit-score

Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform*. 2014;6(1):10. Published 2014 Mar 29. doi:10.1186/1758-2946-6-10

Lewis, R. J. (2000). *An Introduction to Classification and Regression Tree (CART) Analysis*. Lecture presented at Annual Meeting of the Society for Academic Emergency Medicine, San Francisco, California.

Mortgage Data (HMDA). (n.d.). Retrieved March 05, 2021, from https://www.consumerfinance.gov/data-research/hmda

United States, Department of Justice and Federal Trade Commission. (2010). *Horizontal merger guidelines* (pp. 18-19). Washington, D.C.: U.S. Dept. of Justice.

Academic Vita of Deric Liang

EDUCATION

The Pennsylvania State University, ***Schreyer Honors College,*** University Park, PA Graduation Date: May 2021
Bachelor of Science in Applied Statistics, Economics | Minor in Mathematics
The President's Freshman Award Recipient, May 2018

PROFESSIONAL EXPERIENCE

The Pennsylvania State University Department of Economics, University Park, PA Jan 2021-Present
Game Theory Undergraduate TA

- Assess homework assignments for class of 80 students biweekly
- Contribute to homework solutions compilation through accuracy checks
- Maintain database of student grades and release student grades in an independent manner

Bates White Economic Consulting, Washington, D.C. Jun 2020-Aug 2020
Summer Consultant

- Performed data analysis and document review to evaluate competitiveness of 2 healthcare markets for clients
- Led data processing and modeling for a price-fixing case study with 5 other Summer Consultants
- Organized a weeklong firmwide charity event to raise ~$3000 for the D.C. Education Fund

The Pennsylvania State University Department of Economics Bates White REU, University Park, PA Sep 2019-May 2020
Undergraduate Researcher

- Analyzed mortgage data and evaluated observed and predicted lender rejection rates to gain insight into lender behavior
- Collaborated with a faculty advisor to ensure model validity and improve research techniques

Dordt University National Science Foundation Ukraine REU, Sioux Center, IA May 2019-Jul 2019
Undergraduate Researcher

- Explored risk factors for suicide ideation and regional differences in mental health outcomes in Ukraine
- Analyzed data taken from a World Mental Health survey of 6445 individuals in Ukraine with R, SPSS
- Collaborated with peers/faculty in 2 interdisciplinary teams to write papers for journal submission

LEADERSHIP EXPERIENCE/INVOLVEMENT

Penn State Oriana and Glee Club Benefiting THON, University Park, PA Oct 2017-Present
Donor and Alumni Relations Chair, Apr 2020-Present

- Engage with organization alumni to encourage continued participation in fundraising efforts
- Implement online donation platforms to widen donor network

Primary Chair, Mar 2018-Mar 2020

- Managed fundraisers involving ~25 students for children fighting cancer, totaling ~$17,000
- Standardized organization structure and procedures through documentation

Penn State Glee Club, University Park, PA Sep 2017-Present
President, Apr 2020-Present

- Coordinate with faculty director to manage choir operations, engagement, and recruitment
- Preside over the student Executive Board to ensure timely completion of tasks
- Received the Bruce Trinkley Service Award*,* a $300 scholarship for exceptional contribution to the Glee Club

SKILLS

Software & Programming Languages: R, Python, STATA, Minitab, SPSS, SAS
Statistical Knowledge: Regression Analysis, Probability Theory, Time Series, Survey Sampling, ANOVA, Non-Parametric Statistics, Supervised/Unsupervised Learning, Deep Learning/Neural Networks
Economic Knowledge: Microeconomics, Macroeconomics, Econometrics/Forecasting, Labor Economics, Growth & Development Economics, Environmental Economics, Game Theory
Mathematical Knowledge: Real/Complex Analysis, Matrix Algebra, Multivariable Calculus, Vector Calculus
Spoken Languages: Basic Mandarin Chinese, Basic French