TO: Professor NaLette Brodnax

FROM: Allen Church

DATE: May 4, 2020

SUBJECT: Home Mortgage Disclosure Act Data and Examining 2018 Mortgage Rejection Rates Between Races

**Executive Summary**

The following analysis seeks to address and quantify if a mortgage applicant’s race influences loan approval or denial. Using data provided by the Home Mortgage Disclosure Act, the analysis fits a parametric (logistic classifier) and non-parametric (decision tree classifier) statistical learning model to assess this issue. The results of the analysis did not conclusively find discrimination against African American applicants but underscores the importance of analyzing such granular datasets. Namely, results from both models indicate that debt to income ratio – among the variables chosen – had the highest influence on applicant denial.

**Background**

*Mortgage Approval Rates in 2018*

How does a mortgage applicant’s race influence the probability that a financial institution rejects or approves their application? This report will address, using a logistic model, if there are statistically significant differences across races in American mortgage approval rates. On top of this, the report will study if there are statistically significant differences in approval for refinancing loan applications between races, which may address financial trends across races or geographic areas. Using 2018 loan-level data from the United States Consumer Financial Protection Bureau, the report will leverage detailed applicant information – such as race, loan purpose, action taken, and reasons for denial – to study if there are significant differences in probability of loan rejection across different races, but with similar incomes and within the same geographic region. Moreover, the detailed state, county, and census tract breakdown will enable granular analysis of data with a particular state or metropolitan area. In the broader context, the report will address and quantify the American legacy of housing-market discrimination (Holloway, 1998, 252). In a historical context, examining rates of mortgage refinancing in 2018 may expose a disparity between geographic areas or races in the United States.

**Data**

*Home Mortgage Disclosure Act Data*

The primary dataset uses over 1 million mortgage application records from the year 2018, provided by the United States Consumer Financial Protection Bureau under the Home Mortgage Disclosure Act (HMDA). This dataset provides both applicant and loan-specific information, collected from nine United States federal agencies, and contains 99 variables. The 2018 HMDA data includes more detailed applicant information than previous versions. Specifically, the applicant-specific variables include race, age, income level, debt to income ratio, and reason for loan denial. Additionally, the HMDA data include loan amount, loan type, interest rate, and loan term in months. Namely, the data can attempt to answer whether patterns of lending exist that target specific racial groups (Schwemm & Taren, 2010, 378). While the HMDA data provide applicant ethnicity (Hispanic or Not Hispanic), it fails to provide the language in which the application was filed. This information would be helpful to combine with the “reason for denial” section of the data, which provides a “Credit application incomplete” field.

Due to computational limitations, data prepared for this report include a randomized selection of 25,540 observations, out of the total 15,119,652 observations.

**Table of Descriptive Statistics for HMDA Data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Action Taken** | **Loan Amount (Thousands)** | **Income** | **Minority Population %** | **Median Family Income** |
| **Count** | 25,540 | 25,540 | 25,540 | 25,540 | 25,540 |
| **Mean** | 1 | 202 | 156 | 29 | 73,374 |
| **Std** | 1 | 192 | 2,995 | 25 | 17,325 |
| **Min** | 1 | 5 | -133 | 0 | 0 |
| **25%** | 1 | 85 | 51 | 10 | 64,000 |
| **50%** | 1 | 155 | 79 | 21 | 72,600 |
| **75%** | 1 | 255 | 125 | 41 | 81,300 |
| **Max** | 3 | 5,625 | 381,235 | 100 | 134,800 |

**Graphical Summaries of HMDA Data**

**A screenshot of a cell phone

Description automatically generatedGraph 1: Distribution of Applicants by Race**

**A screenshot of a cell phone

Description automatically generatedGraph 2: Distribution of Applicant by Sex**

**A screenshot of a cell phone

Description automatically generatedGraph 3: Distribution of Race Among Denied Applicants**

**A screenshot of a cell phone

Description automatically generatedGraph 4: Distribution of Debt to Income Ratio by Applicant Race**

As the visual summaries above indicate, a large majority of applicants are White, followed by Race Not Available. Following this logic, the majority of denied applicants are also White, followed by Race Not Available. Interestingly, the sex of the majority of applicants is Joint, followed by Male. Lastly, the distribution of debt to income ratio shows that the median value for African Americans is 42, while the median value for White is 38.

**Methodology**

*Parametric model – Logistic Regression*

The chosen parametric technique is logistic regression, which is a probabilistic classifier used on a binary outcome. In the HMDA data, this involves converting the target variable into a binary outcome – loan originated or application denied by financial institution. Logistic regression allows us to infer the effects that a one-unit increase in an independent variable has on the dependent variable of application approval or denial. Specifically, it achieves this by relating the independent variables to the log odds of the dependent variable (James et al., 2017, p. 133).

In Barbic et al. (2016, p. 319), researchers used a logistic regression model to analyze the effects of financial literacy and planning for retirement in Croatia. Specifically, their logistic results indicated that financial literacy has a positive effect on attitudes towards retirement. (Barbic et al., 2016, p. 327). However, it should be noted that logistic regression assumes a linear decision boundary and will fit the data accordingly if this assumption holds (James et al., 2017, p. 153). Conversely when the data does not fit this assumption, or in instances with high correlation between predictors, logistic regression will not fit the data well, and other models should be considered (James et al., 2017, p. 136).

*Non-parametric model – Decision Tree Classifier*

The chosen non-parametric technique is a decision tree classifier, to be used on the qualitative response of loan originated or application denied. Specifically, this classification tree works by assigning an observation to “the most commonly occurring class of training observations in the region to which it belongs” (James et al., 2017, p. 311). The decision tree’s main advantage is in interpreting results. When interpreting output from a decision tree classifier (presented in the Analysis section below), one reads the top of the tree as the most important factor in determining the outcome variable (James et al., 2017, p. 305).

In an example of a decision tree classifier, an author used a model to predict whether or not a particular firm was bankrupt. Specifically, the researcher used input variables such as debt-repaying ability, earning ability, and cash flow to predict financial distress for 100 Taiwanese firms (Chen, 2011, p. 11264). Interestingly, the author obtained a 97% accuracy score when limiting data to shortly before a period of financial crisis (Chen, 2011, p. 11272)

**Variable Selection**

*Target Variable*

Action Taken – To analyze possible discrimination in application approval or denial between races, the data includes this as a binary target variable. However, the HMDA (HMDA 2020) includes the following values for this variable:

1. Loan originated
2. Application approved but not accepted
3. Application denied
4. Application withdrawn by applicant
5. File closed for incompleteness
6. Purchased loan
7. Preapproval request denied
8. Preapproval request approved but not accepted

*Feature Selection*

The models include the following features:

* *Derived Race* – This variable contained nine values, but for the purpose of this analysis, only includes Black or African American and White.
* *Loan Amount* – The total amount applied for, in dollars. In Mendez et al. (2011, p. 105), the authors use this as a covariate in developing a redlining index. In addition to this, the authors cite four previous HMDA studies that include this as a covariate. (p. 105).
* *Income* – The gross annual income, in thousands of dollars. Similar to Loan Amount, Mendez et al. cite four previous HMDA studies that use Income as a covariate (2011, p. 105).
* *Loan Term* – The length of mortgage applied for, in months. The analysis includes this variable to control for mortgage applications of varying terms.
* *Debt to Income Ratio* – The percentage of monthly debt to monthly income. Courchane & Zorn (2008, p. 6) use debt to income ratio when analyzing differing HMDA interest rates between Whites and African Americans. Specifically, the authors argue that debt to income ratio is a “key underwriting factor… [and] directly affects credit risk” (p. 6).
* *Tract Minority Population Percent* – The percentage of minority that comprises a census population tract. In Guy et al. (1982, p. 289), the authors analyzed HMDA data and found a relationship between the racial composition of census tracts in Memphis and total loan amount.
* *Median Family Income* – Adjusted median family income for the census tract, in dollars. Similar to the previous variable, Guy et al. (1982, p. 289) found a relationship between median family income and awarded loan amount, with differing values between African American and White people.
* *Derived Sex* – The sex of the applicant, divided into male, female, joint, or not available. The model includes this variable to assess if applicant sex has a significant relationship in determining loan approval or denial. However, it should be noted that this variable fails to include those who do not identify on the binary gender system.
* *Applicant Age* – The age, divided into bins, of the applicant. The model includes this variable to assess if applicant age is a statistically significant determinant of loan approval or denial.
* *State Code* – The two-letter United States state code. The analysis includes this variable to control for location-specific housing discrimination against African Americans.

**Analysis**

*Logistic Classification Results*

As discussed in the Methodology section above, the logistic classifier allows us to infer the effect of the input variables on the output variable. As seen in the table of the coefficient estimates below, the most substantively significant variable in loan denial is Debt to Income Ratio. While one cannot directly interpret these coefficients, as in ordinary least squares, we can infer that the below coefficients produce a net positive effect in probability of application denial. Lastly, the logistic classifier had an overall performance score of 80%, meaning that it correctly classified predictions 80% of the time.

|  |  |
| --- | --- |
| **Variable** | **Coefficient Estimate** |
| Debt to Income Ratio | 0.0462 |
| Tract Minority Population % | 0.0031 |
| Applicant Age 55-64 | 0.0002 |
| Derived Sex - Female | 0.0002 |
| State Code - CA | 0.0002 |
| Derived Sex - Male | 0.0002 |

**Table 1: Logistic Classifier Coefficients**

In addition to this, the confusion matrix below displays the logistic model’s performance with regards to correct classification. Specifically, the matrix divides classifications into True Positive, True Negative, False Positive, and False Negative. The matrix below indicates that the logistic classifier classified 19,253 True Positives, 977 True Negatives, 446 False Positives, and 4,864 False Negatives.

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | 19,253 | 446 |
| Negative | 4,864 | 977 |

**Table 2: Confusion Matrix for Logistic Classifier**

*Decision Tree Classifier Results*

As discussed in the Methodology section above, a decision tree classifier assigns an observation to “the most commonly occurring class of training observations in the region to which it belongs” (James et al., 2017, p. 311). In the attached decision tree graphic, the top leaf represents the most important feature in determining probability of application approval or denial. Similar to the logistic model, the most significant feature in the decision tree is also Debt to Income Ratio. Specifically, the tree model identifies a cut-off point of 57.5% debt to income ratio as determining whether or not an institution approves or denies a mortgage application. When compared to the logistic classifier above, and using a maximum depth of five, the decision tree slightly out-performs and produces a mean accuracy score of 82.8%.

In addition to showing the most significant features in order, the tree also produces a Gini index, which represents variance across classes. Specifically, the “Gini index is referred to as a measure of node purity – a small value indicates … observations from a single class” (James et al., 2017, p. 312). As seen in the decision tree output, the Gini index for this model is relatively low across nodes, indicating heterogenous classes. In addition to this, the confusion matrix for the decision tree also indicates better classification when compared to the logistic model. The matrix below indicates that the logistic classifier classified 19,351 True Positives, 1,943 True Negatives, 348 False Positives, and 3,898 False Negatives.

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | 19,351 | 348 |
| Negative | 3,898 | 1,943 |

**Table 3: Confusion Matrix for Decision Tree Classifier**

**Conclusion**

As reviewed above, a number of authors have used HMDA data to analyze and quantify levels of discrimination against African Americans in mortgage applications. While the two models used in this analysis did not conclusively find mortgage discrimination against African Americans, it decisively found a common variable in application denial. In addition to this, the FDIC Statements of Policy states that “HMDA data alone do not prove lending discrimination” (Federal Deposit Insurance Corporation 2019). It is important to note that these results do not disprove the presence of institutional racism in mortgage applications. Instead, it identifies a relevant variable and contributes to the contemporary discussion on the role that burdensome debt plays in applying for a home loan. Since HMDA data provides granular loan-level data, future analyses should assess whether differences exist in interest rates or loan amounts between races – as past studies have concluded. In the face of a pandemic that has exposed and widened racial health and economic disparities (Perry et al., 2020), it is increasingly important to analyze the micro and macro-level effects of institutional racism in the housing market, and beyond.

**References**

Barbic, D., Palic, I., & Bahovec, V. (2016). Logistic regression analysis of financial literacy implications for retirement planning in Croatia. *Croatian Operational Research Review*, 7(2), 319–331. <https://doi.org/10.17535/crorr.2016.0022>.

Chen, M. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems With Applications*, 38(9), 11261– 11272. <https://doi.org/10.1016/j.eswa.2011.02.173>.

Consumer Financial Protection Bureau. (2018). *HMDA – Home Mortgage Disclosure Act*. Retrieved May 1, 2020, from <https://ffiec.cfpb.gov/data-browser>.

Consumer Financial Protection Bureau. (2020). *Public HMDA data fields with values and definitions*. Retrieved May 1, 2020, from <https://ffiec.cfpb.gov/data-browser>.

Federal Deposit Insurance Corporation. (2019). *FDIC Law, Regulations, Related Acts – Statements of Policy*. Retrieved May 1, 2020, from <https://www.fdic.gov/regulations/laws/rules/5000-3860.html>.

Guy, R., Pol, L., & Ryker, R. (1982). Discrimination in mortgage lending: The Home Mortgage Disclosure Act. *Population Research and Policy Review*, 1(3), 283–296. <https://doi.org/10.1007/BF00140097>.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to Statistical Learning: with applications in R*. New York: Springer.

Mendez, D., Hogan, V., & Culhane, J. (2011). Institutional Racism and Pregnancy Health: Using Home Mortgage Disclosure Act Data to Develop an Index for Mortgage Discrimination at the Community Level. *Public Health Reports*, 126, 102–114.

Perry, A., Harshbarger, D., & Romer, C. (2020). Mapping racial inequity amid COVID-19 underscores policy discriminations against Black Americans. *The Brookings Institution*. Retrieved May 1, 2020, from [https://www.brookings.edu/blog/the avenue/2020/04/16/mapping-racial-inequity-amid-the-spread-of-covid-19/](https://www.brookings.edu/blog/the%09%09%09%09avenue/2020/04/16/mapping-racial-inequity-amid-the-spread-of-covid-19/).

Schwemm, R. G., & Taren, J. L. (2010). Discretionary Pricing, Mortgage Discrimination, and the Fair Housing Act. *Harvard Civil Rights-Civil Liberties Law Review, 45*(2), 375–433.