Using Big Data Systems to Analyze Big (not Jumbo) Mortgage Data

Cody Gilbert
NYU Computer Science

Fang Han
NYU Computer Science

Jeremy Lao
NYU Computer Science

July 18, 2019

Abstract

Banking professionals are required to submit data to Federal regulators for the purposese of monitoring the health and safety of the financial system and individual banks. However, banks are also required by law to help promote growth in their local economies through lending. The Home Mortgage Disclosure acts requires banks and lenders to provide low-level mortgage application data to the Consumer Finance Protection Bureau (CFPB). Federal banking regulators analyze the data to discern economic trends and monitor for unfair lending practices. Our analysis utilizes big data architecture, namely Spark, to dig deeper into the numbers to analyze denial rates by race group, gender, and various borrower characteristics. Our visualization application will serve as a tool for both regulators and lenders to help identify possible red flags in lending practices.

1 Introduction

2 Objectives

3 Data, Analysis, and Methodology

This section discusses the analysis of the data set and our modelling methodology. The time frame of the data set spans from 2007-2017 and our analysis focuses on the denial rate of mortgage applications by borrow characteristics such as race, ethnicity, and gender. We will also analyze denial rate by geographic features and at the institution level. We will also model whether borrower characteristics impact denial rates through controlling for factors in a regression analysis. Ultimately our model will show how the denial rate, controlled for borrower characteristics, varies by race, ethnicity, state, and lending institution.

3.1 Average Denial Rate

Our work begins with studying the denial rate of one-to-four family, manufactured, and multifamily housing for the various race and ethnic groups reported in the HMDA data.

The following ethnic and race groups are reported in the HMDA data:

- White
- Hispanic
- Asian
- African American

- Native Hawaiian or Pacific Islander
- American Indian or Alaska Native

Our analysis excludes line items where the race and ethnicity are not known and reported as *Not Applicable, Blank, or "Cannot be Reached by Phone,Internet,Fax, Mail "*. The percentage of items where ethnicity is not identified is $\approx 21\%$ and the percentage of items where race is not reported is only $\approx 11\%$. Given the high percentage of data where race and ethnicity are specified, we will have a significantly large data set. The intent of analyzing denial rate by race and ethnicity is to compare mortgage loan denial rate of non-white loan applicants to white loan applicants (the control group).

Previous research has been done in this area and it is widely acknowledged that white loan applicants as a whole have the lowest denial rates. [1] However, we will begin our analysis with the overall denial rate and further analysis will show the comparative denial rate when borrower characteristics are taken into account.

We will calculate the average denial rate, D by race r, represented as D_r. The denial rate D_r by race is caluclated as the number of approvals, a_r, divided by the number of loans applied, n_r, for given a race group.

$$a = \begin{cases} 1, & \text{if loan approved.} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The average denial rate by race, D_r, is calculated as:

$$D_r = \frac{1}{n_r}\sum_{i=1}^{n_r} a_i = \frac{a_{r_1} + a_{r_2} + \cdots + a_{r_n}}{n_r} \tag{2}$$

3.2 Discussion of Methodology

4 Analytic/Visualization Application

5 Further Work

6 Conclusion

Reducing the sparsity of the matrices generated from CountVectorizer by stacking the documents greatly improved the results.

A Appendix

A.1 What is HMDA

References

[1] *Opportunities and Issues in Using HMDA Data.* Robert Avery, Kenneth Brevoort, Glenn Canner. JRER Vol. 29, April 2007.