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

3.1 Denial Rate Analysis

3.1.1

Our work includes 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. In our initial study of denial rates, we will look at the denial rate for the following ethnic/race groups r:

- White
- Hispanic
- Asian
- African American
- Native Hawaiian or Pacific Islander
- American Indian or Alaska Native

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:

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

3.2 Data

3.3 Apply ML to Sparse Matrices generated from CountVectorizer

3.3.1 Training, Testing, and Determining the Efficacy of the Models

4 Fruther Work

5 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] tem