Analsyis_of_bankruptcies_by_ZIP

Jacob Steimer

2023-07-27

What causes high bankruptcy rates in some U.S. ZIP codes?

Introduction

During the height of the COVID-19 pandemic, the number of personal bankruptcy filings dropped significantly across the country. However, some parts of the country continue to see far more bankruptcies than others, in a way that doesn't seem clearly correlated with any one factor, according to data from the Federal Judicial Center. Also, people's choices of type of bankruptcy — Chapter 7 or Chapter 13 — vary widely.

In this analysis, I explore the relationship between bankruptcy rates in U.S. ZIP codes and various demographic variables pulled from the U.S. Census Bureau. Using machine learning techniques, I explore the importance of the different variables and build models that can predict bankruptcy rates given Census data.

Data Overview

(See lines 1-28 in "CodeToFollow.R")

I obtained bankruptcy data from the Federal Judicial Center website and demographic data from the U.S. Census Bureau. Because the FJC and Census website links are subject to change, I uploaded the relevant data from both websites to GitHub and wrote code for its download.

The bankruptcy data includes one row for every bankruptcy filed during 2022. Columns included the type of bankruptcy and the ZIP code connected with each debtor.

The Census Data included the following data for each ZIP code:

- The Labor Force Participation Rate (the percentage of adults who are either working or looking for work)
- The Unemployment Rate (The percentage of the Labor Force that isn't working)
- The Homeownership Rate (the percentage of households that own their residence)
- The percentage of residents with bachelor's degrees
- The percentage of residents who are Black
- The population and median income in each ZIP

I chose most of these factors because they're important economic characteristics. I chose the percentage of Black residents because prior research has shown African Americans are more likely to file Chapter 13 than Chapter 7.

Methods/Analysis

Data Preparation

(See lines 29-68 in "CodeToFollow.R")

I cleaned the dataset by removing duplicate cases and filtering it to include only those filed in 2022. I then determined the count of Chapter 13 bankruptcies, count of total bankruptcies and the percentage of Chapter 13 bankruptcies using "summarize" and "mutate". Additionally, I transformed certain variables for better interpretability. For instance, I divided Median Income by 10,000 because a difference in $10,000 between ZIPs is more understandable than a $1 difference.

I then split the data into a holdout set, a training set and a testing set. Since the data only included about 18,000 observations, I used a 10% partition each time, to ensure both the holdout and testing sets were large enough.

Modeling with linear regression

(See lines 69-97 in "CodeToFollow.R")

To determine which demographic characteristics correlate most with elevated bankruptcy rates, I started with linear models. I fit separate linear models for total bankruptcies per 1,000 residents (fitlm), Chapter 13 bankruptcies per 1,000 (fitlm3) and the percentage of bankruptcies that are Chapter 13 (fitlm5). I experimented with various subsets of variables along the way (fitlm2, fitlm4 and fitlm6).

When fitting models that could predict the sum of Chapter 13 bankruptcies and of total bankruptcies, race, homeownership rate, and bachelor's degrees proved the strongest predictors. For the percentage of bankruptcies that are Chapter 13, college education lost most of its predictive power, and both the unemployment rate and the labor force participation rate became more important.

Modeling with Random Forrest

(see lines 121-144 in code)

Next, I applied the Random Forrest algorithm to predict the same outcomes as in the linear regression models. The Random Forrest's variable importance measurements did not highlight the same variables that jumped out in our linear models. Also, while the linear models consistently saw large effects caused by similar variables, the Random Forrest importance rankings were much different from each other, depending on whether total bankruptcies, Chapter 13 bankruptcies or percentage of Chapter 13 bankruptcies were being predicted.

For total bankruptcies, it listed the bachelor's degree and race variables as most important and homeownership rate as least important. For Chapter 13 bankruptcies, it listed the labor force participation rate as most important and race as least important. And for the percentage of bankruptcies that are Chapter 13, it listed homeownership rate and race as most important and median income and labor force participation rate as least important.

Results

Linear models performance

(see lines 98-120 in code)

My linear models had by far their lowest RMSEs (around .25) when predicting the percentage of bankruptcies in a given ZIP that were Chapter 13 (fitlm5 and fitlm6). It struggled the most in predicting the total number of bankruptcies in a ZIP (fitlm and fitlm2). But even there, the RMSE was relatively low (around 1.6).

Random Forrest performance

(see lines 145-158 in code)

All three of our Random Forrest models performed slightly better — in terms of RMSE — than their linear model counterparts. However, since the differences were small, we shouldn't necessarily consider the Random Forrest results more important than the linear model results.

Aside #1: KNN model and performance

(see lines 159-165 in code)

I also attempted to use K-nearest neighbors (KNN) for prediction and analysis, but the results showed higher root mean square errors (RMSE) compared to linear regression and Random Forrest. Also, I realized that for the purposes of this analysis, KNN is not the best type of model, since it makes it hard to interpret the importance of certain variables. Consequently, I decided to focus on linear regression and Random Forrest.

Aside #2: Using the bankruptcy data itself to predict Chapter type

(see lines 166-197 in code)

As I've continued to research the causes of someone filing a Chapter 13 instead of Chapter 7, I decided to analyze this data in one more way. The bankruptcy data itself contains far more information than Chapter and ZIP. So, I wanted to analyze which variables within that dataset can help us accurately predict Chapter choice.

To do this, I pared the data down into include 9 variables that seemed like they could affect Chapter choice: Total Assets (TOTASSTS), Real Property (REALPROP), Personal Property (PERSPROP), Total Liabilities (TOTLBLTS), Secured Claims (SECURED), Unsecured Priority Claims (UNSECPR), Unsecured Nonpriority Claims UNSECNPR, Total Debt (TOTDBT), and Current Monthly Income (CNTMNTHI). I then made Chapter choice a binary variable.

I then chopped the dataset up into four parts: A final holdout set, a testing set, a training set and a small training set. I created the small training set because the data has over 388,000 observations.

Using the small training set, I ran a linear regression to find the most important variables. These were real property, secured debt and current monthly income.

Then, using just these variables, I ran a linear regression on the larger training set. The RMSE turned out quite small, which gives me confidence that these three factors play large roles in the Chapter someone chooses.

Conclusions

This project primarily accomplished three tasks:

1. It demonstrated the ability of multiple machine learning techniques to predict bankruptcy rates (and types) using Census data. These techniques achieved low RMSEs.

2. It proved that the percentage of Black residents in a given ZIP plays a major role in the number and type of bankruptcies there. In fact, it seems to play a larger role than any of the economic or educational characteristics we evaluated. This further proves the previously documented role of racial bias in the bankruptcy system. This finding informed a news article, which can be found here: https://mlk50.com/2023/09/05/shelby-county-ranks-no-1-in-bankruptcy-heres-why/.

3. It also showed that the homeownership rate, college education, and labor force participation rate are better predictors of bankruptcy than median income or the unemployment rate.

These results have potential implications for policymakers looking to address systemic racism and researchers interested in understanding why some parts of the country have much higher bankruptcy rates than others.