Factors Related to Income During Economic Recession

Rachel Rumas (501211310)

CIND 820 – Big Data Analytics Project

Dr. Ceni Babaoglu

October 24 2022

# Revised Abstract

Income inequality in the United States has worsened dramatically over the last 40 years. With the looming threat of a recession, poverty rates are also on the rise and both the government and non-profit sector will need targeted interventions to aid the most vulnerable. Traditionally, socioeconomics has used tried and true statistical methods to determine who is most vulnerable, but this project will attempt to answer whether machine learning algorithms offer improved effectiveness at answering these questions on historical data. This project will use the “Census Income” dataset donated to the University of California Irvine’s Machine Learning Repository, a dataset which offers a subset of 13 variables from the United States Census in 1994. This dataset aligns with the aftermath of the early 1990s recession which saw peaks in poverty and income inequality. The results of this data will be compared with existing literature on determinants of low-income and poverty to identify if any differences exist during a recession versus in general.

# Literature Review

Income inequality is a major and ever-growing topic of conversation in today’s globalized society. Over the last 40 years, income inequality has worsened in many developed countries around the world, particularly in the United States. Greater income inequality within a country is associated with a host of negative socioeconomic outcomes, including slowed economic growth, increased financial crises, and a decreased stability in democratic systems (Alesina & Rodrick, 1994; Muller, 1988). There are many metrics to measure income inequality, but the Gini coefficient is a popular statistical measure used by the World Bank and other international financial institutions to quantify wealth inequality within a nation. The Gini coefficient of the United States has risen from 34.5 in 1979 to 41.5 in 2019 (World Bank, 2022), with higher scores representing greater income inequality. The United States Gini coefficient currently ranks as one of the highest amongst developed nations, ranking 5th among OECD (Organization for Economic Cooperation and Development) member countries, behind Costa Rica, Chile, Mexico, and Türkiye (OECD, 2022).

Another measure which helps contextualize the economic landscape of a country is its poverty rate. Over the last several decades, the poverty rate in the United States has steadily declined, but has seen peaks corresponding with economic recessions in the early 1980s, early 1990s, and the 2008 recession (Hoynes et al., 2006; U.S. Census Bureau, 2022). Most recently, the poverty rate fell to a record low in 2019 at 10.5% (U.S. Census Bureau, 2022). However, income inequality and poverty rates have been pushed to the forefront of the public mind due to recent events. The economic fallout and rampant job loss that occurred globally in response to the COVID-19 pandemic has been accompanied by an economic instability that has continued up until the present day for many developed nations, particularly in the United States. Income inequality has continued to rise. In 2021, popular news outlets circulated stories about American billionaires adding over $1 trillion dollars to their collective net worth during the 2020 fiscal year while 20 million Americans lost their jobs (Peterson-Withorn, 2021; Ingraham, 2021). Poverty rates increased for the first time in 5 years in 2020 and remained at elevated levels in 2021 (U.S. Census Bureau, 2022). Consequently, many economists have reported that the United States may be headed towards a recession if the country is not in one already (Torry & DeBarros, 2022). By the traditional definition of two consecutive fiscal quarters with negative economic growth, the United States entered a recession in the Summer of 2022.

As income inequality and poverty affect the most vulnerable of a country’s population, it is therefore prudent to determine factors that have historically predicted low-income or poverty during times of economic recession in order to better target social safety programs in the future, particularly in the United States where these social safety programs are often difficult to access or the required infrastructure is not in place.

Much research has already been done in the fields of economics, global development, and politics on factors which contribute to poverty and low-income and can in some cases be intuited. Education for example, is a barrier to escaping poverty in the United States. Post-secondary tuition fees in many cases can leave a household in debt even with financial aid, and higher paying jobs often require post-secondary degrees or diplomas or even graduate degrees. As another example, women have both a decreased median income compared to men working the same job as well as a 35% increased likelihood of living in poverty than men (Fins, 2020). The aim of this paper will be to determine which of these factors are most important specifically in times of recession, and if these factors differ in importance during non-recession times, based on previous literature. As we head further into economic uncertainty, this information may prove to be valuable moving forward.

The methods with which this project aims to answer these questions are through machine learning algorithms. A separate gap in the literature that this project intends to address is the use of machine learning in the field of economics. Machine learning has been readily embraced in various fields of research such as clinical science, cognitive science, and market research, but is not as prevalent in the social sciences, though is beginning to gain traction. The social sciences most regularly rely on traditional statistical metrics that have stood the test of academic time. One of the most common statistical methods in econometrics is regression analysis (Schneider, Hommel, & Blettner, 2010; Ramcharan, 2006). Traditionally, simple linear or binary regression is used to explain the effect of one or more independent variables on a dependent variable of interest in the field of economics, as well as in sociology, psychology, amongst others. Machine learning algorithms can also be applied to explain the variance of independent variables on a dependent variable of interest, but are mostly spoken of in terms of classification models. These models are fed testing data and are trained to detect patterns in the data as accurately as possible and then the models are applied to a dataset to classify a dependent variable into its constituent categories or along a line of best fit. An advantage of using machine learning algorithms in social science fields where the methods are not as prominent is that it affords the opportunity to scale up analyses, reduce computational time, potentially improve their effectiveness, and be applied to updated versions of the dataset as well as similar datasets which explore similar topics. Machine learning approaches have been shown to have greater effectiveness on econometric data than simple linear regression techniques before (Lazar, 2004; Mullainathan & Spiess, 2017), and so it is worth exploring this finding on similar datasets under similar circumstances, both to validate the results of these previous studies and to suggest conclusions to our own research question in light of both our own findings and the previous literature.

For this project, a variety of popular classification algorithms will be employed that have been selected from literature from research on similar topics. A similar project conducted by Matkowski (2021) compared 8 models, ranging from linear regression to Random Forest and K-Nearest Neighbours, two popular machine learning algorithms. Other projects focus on comparing linear regression to Naïve-Bayes, another popular algorithm (Karim & Rahman, 2013). The details of the model selection for comparison for this project will be outlined further in the Tentative Methodology section.

The dataset that will be utilizing in this project is the “Census Income” dataset, a subset of the 1994 United States Census which was retrieved from the University of California Irvine’s Machine Learning repository, which is a resource which maintains hundreds of datasets for the purpose of machine learning research and education. The “Census Income” dataset contains 13 demographic variables across 32562 datapoints which represent the working United States population in 1994 when their sample weights are applied. The demographic variables include age, sex, race, education level, marital status, household relationship, native country, occupation, work class, number of hours worked per week, capital gains, and capital losses. The predictor variable in this dataset is a binary variable which holds values of either “<=50k” or “>50k” denoting the yearly income of each case.

This dataset is ideal for this project for a multitude of reasons. First, it provides precisely the information we need to answer our first research question: which demographic variables are most strongly related to predicting income in the wake of an economic recession. 1994 corresponds with the period directly following the recession which occurred in the United States during the early 1990s where poverty rates were at peaks that would not be seen until the aftermath of the 2008 recession. This data therefore provides an ideal timeframe. A second reason this dataset is ideal for the purposes of this project is that the database is readily available through the UCI’s Machine Learning repository. Because of this, the data is well-maintained and therefore limited preprocessing steps will be required. This is also convenient because it means this dataset is well-documented throughout the literature. Since the dataset was donated in 1996 by Kohavi and Becker, it has been used for a wide variety of publication topics including the study of data mining, comparing Bayesian network classifiers, and comparing clustering analysis methods (Kohavi, 1996; Cheng & Greiner, 1999; Gionis, Mannila, & Tsaparas, 2005).

Previous literature using this database is restricted to specific methods and the development of new proposed models which hope to add tools for machine learning classification, often in research papers that are now 15-25 years old. As this data contains demographic information approaching 30 years in age, recent analysis using this dataset has been limited. This does not pose an issue for the current project, as historical data, particularly recession data, is the focus of our analysis. This also benefits the current project as it affords the opportunity to run the gamut of the best available machine learning classification tools, many of which did not exist or were not as well-regarded when this dataset was originally donated. The history of Random Forest, for example, which is a machine learning method which involves constructing multiple decision trees, can be dated to 1995 at its earliest, and was refined further and did not reach greater popularity in research until 2006 when it was trademarked.

As previously alluded, this project intends to address two primary gaps in the literature, one related to the methods and one related to the applications. The methods-based gap is that the current gold standard of classification models has not yet been applied onto the “Census Income” dataset and compared amongst each other. It will be compelling to determine if significantly different patterns emerge using more current classification models compared to the results from older publications. The application-based gap in the literature refers to the real-life implications which the data represents. According to economists, the United States is either in a recession at the moment, will be in a recession in the near future, or failing all else, will most likely be in a recession at some point in the future. This necessitates the collection of information on how best to prepare for the aftermath and the better target interventions which will protect the most vulnerable, in a climate where income inequality is progressively increasing already.

# Summary Statistics

It is necessary to examine the summary statistics of the variables in the dataset to ensure the analysis can be interpreted effectively or if extra preprocessing steps will be needed, particularly with regards to the dependent variable, income. After removing missing variables from the data, the descriptive data for the variable ‘income’ is as follows:

|  |  |
| --- | --- |
| ‘income’ | count |
| <=50k | 22654 |
| >50k | 7508 |

*Table 1. Summary of Target Variable (‘income’)*

It should be noted that there is an imbalanced dependent variable as the <=50k class has significantly greater data points than the >50k class. This indicates that during the experiment stage, it will be necessary to balance the classes using SMOTE oversampling, a widely-used method for class imbalance.

In order to better contextualize the data, the descriptive statistics for the numeric variables are listed below. It should however be noted that the aggregated statistics such as mean, median, and mode are not to be interpreted as the mean age, mean capital gain, median highest level of education, etc., of the United States population, as these variables are given a sampling weight by the variable ‘fnlwgt’ which is not taken into account in the descriptive statistics. This will be factored in during the analysis stage.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| stat | age | education-num | capital-gain | capital-loss | hours-per-week |
| count | 30162 | 30162 | 30162 | 30162 | 30162 |
| mean | 38.4379 | 10.1213 | 1092.01 | 88.3725 | 40.9312 |
| std | 13.1347 | 2.55 | 7406.35 | 404.298 | 11.98 |
| min | 17 | 1 | 0 | 0 | 1 |
| 25% | 28 | 9 | 0 | 0 | 40 |
| 50% | 37 | 10 | 0 | 0 | 40 |
| 75% | 47 | 13 | 0 | 0 | 45 |
| max | 90 | 16 | 99999 | 4356 | 99 |

*Table 2: Descriptive Statistics of Numeric Variables*

# Tentative Methodology

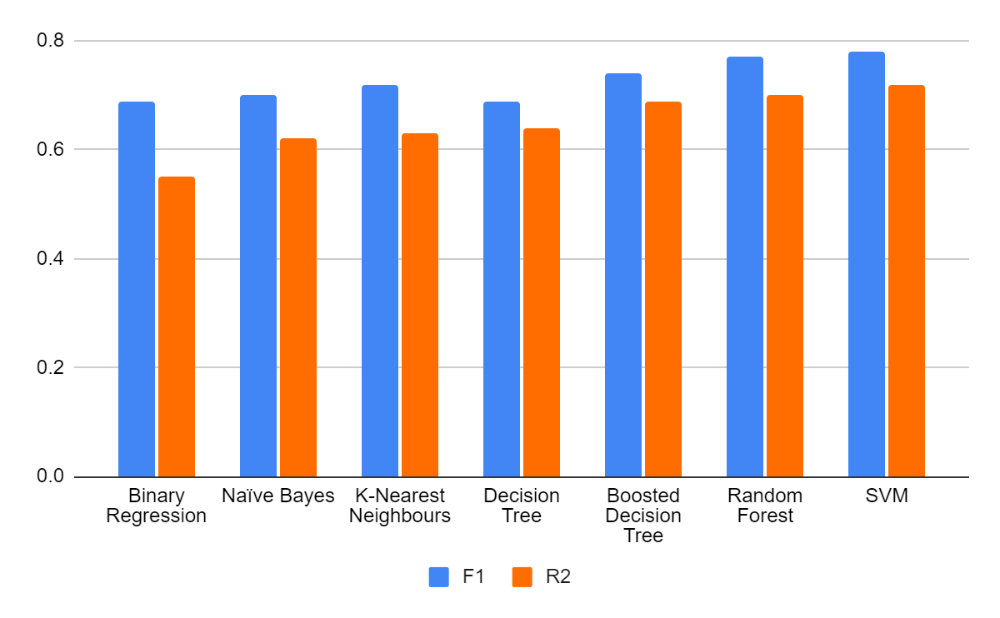
This project will employ a wide variety of methods to perform a simple binary classification outcome. The question that will be asked of each of these methods is such: how accurately can each model identify which datapoints represent household incomes below and above $50,000? And more broadly, which variables are most important in the models’ decision-making?

A selection of 7 models will be tested on this dataset, ranging from the simplest models most widely used in the economic sector to tackle these sorts of problems, to the gold standard gamut of common machine learning binary classification methods available today. Specifically, the project will employ the 1) binary logistic regression, 2) Naïve Bayes, 3) K-Nearest Neighbours, 4) Decision Tree, 5) Boosted Decision Tree, 6) Random Forest, and 7) Support Vector Machine. Each of these methods will be measured and compared using traditional classification fitness metrics (accuracy, precision, recall, f1, and R2 where applicable for each method).

The methodology proposed above will produce a results table and chart that will look similar to the figure below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 | R2 |
| Binary Regression | *0.62* | *0.52* | *0.66* | *0.64* | *0.51* |
| Naïve Bayes | *0.69* | *0.63* | *0.71* | *0.70* | *0.62* |
| K-Nearest Neighbours | *0.75* | *0.53* | *0.90* | *0.72* | *0.63* |
| Decision Tree | *0.80* | *0.76* | *0.83* | *0.69* | *0.64* |
| Boosted Decision Tree | *0.82* | *0.81* | *0.82* | *0.74* | *0.69* |
| Random Forest | *0.80* | *0.75* | *0.83* | *0.77* | *0.70* |
| SVM | *0.81* | *0.69* | *0.90* | *0.78* | *0.72* |

*Table 3: Potential Results with Sample Values (Fitness Metrics)*



*Figure 1: Visualization of Key Metrics from Table 3*

# GitHub Link

The link to the project GitHub, which currently contains the dataset as well as the coding for the data preprocessing and descriptive statistics, is linked below.

<https://github.com/rrumas/CIND_820_Final_Project.git>

# References

Alesina, A. & Rodrick, D. (1994). Distributive politics and economic growth. *Quarterly Journal of Economics*. *109*(2), 465–90. doi:10.2307/2118470

Cheng, J. & Greiner, R. (1999, Jul 30-Aug 1). *Comparing Bayesian network classifiers* [Paper presentation]. Fifteenth Conference on Uncertainty in Artificial Intelligence: Stockholm, Sweden.

Fins, A. (2020). National snapshot: Poverty among women & families, 2020. *National Women’s Law Center*. Retrieved from: https://nwlc.org/wp-content/uploads/2020/12/PovertySnapshot2020.pdf

Gionis, A., Mannila, H., & Tsaparas, P. (2005, Apr 5-8). *Clustering aggregation* [Paper presentation]. Twenty-first International Conference on Data Engineering: Tokyo, Japan.

Hoynes, H. W., Page, M. E., & Stevens, A. H. (2006). Poverty in America, trends and explanations, *Journal of Economic Perspectives*, *20*(1), 47-68.

Ingraham, C. (2021, Jan 1). World’s richest men added billions to their fortunes last year as others struggled. *The Washington Post*. Retrieved from: https://www.washingtonpost.com/business/2021/01/01/bezos-musk-wealth-pandemic/

Karim, M. & Rahman, R. (2013). Decision tree and Naïve-Bayes algorithm for classification and generation of actionable knowledge for direct marketing. *Journal for Software Engineering and Applications*, *6*(4), 196-206.

Kohavi, R. (1996, Aug 2-4). *Scaling up the accuracy of Naïve-Bayes classifiers: A decision-tree hybrid* [Paper presentation]. Second International Conference on Knowledge Discovery and Data: Portland, OR, United States.

Lazar, A. (2004, Dec 16-18). *Income prediction via support vector machine* [Paper presentation]. 2004 International Conference on Machine Learning and Applications: Louisville, KY, United States.

Matkowski, M. (2021). *Prediction of individual level income: A machine learning approach* [Unpublished undergraduate honours thesis]. Bryant University.

Mullainathan, S. & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87-106.

Muller, E. (1988). Democracy, economic development, and income inequality. *American Sociological Review, 53*(1), 50-68.

OECD (2022). *Income inequality* [Data set]. doi:10.1787/459aa7f1-en (Accessed on 26 September 2022).

Peterson-Withorn, C. (2021, Apr 30). How much money America’s billionaires have made during the COVID-19 pandemic. *Forbes.* Retrieved from: https://www.forbes.com/sites/chasewithorn/2021/04/30/american-billionaires-have-gotten-12-trillion-richer-during-the-pandemic/?sh=6c886386f557

Ramcharan, R. (2006). Regressions: Why are economists obsessed with them? *International Monetary Fund*, *43*(1).

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: Part 14 on a series of evaluation of scientific publications. *Deutsches Artzeblatt Interational*, *107*(44). 776-782. doi: 10.3238/arztebl.2010.0776.

Torry, H. & DeBarros, A. (2022, Oct 22). Economists now expect a recession, job losses by next year. *The Wall Street Journal.* Retrieved from https://www.wsj.com/articles/economists-now-expect-a-recession-job-losses-by-next-year-11665859869

U.S. Census Bureau (2022). *Income, poverty, and health insurance coverage in the United States: 2021*. Retrieved from https://www.census.gov/newsroom/press-releases/2022/income-poverty-health-insurance-coverage.html

World Bank, World Development Indicators (2022). *Gini Index* [Data file]. Retrieved from https://data.worldbank.org/indicator/SI.POV.GINI.