Determining Factors Related to Poverty During Economic Recession  
  
Rachel Rumas (ID: 501211310)

Toronto Metropolitan University

CIND 820: Big Data Analytics Project

September 26, 2022

Abstract

A) Income inequality in the United States has worsened dramatically over the last 40 years. 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 currently ranks as one of the highest amongst developed nations, ranking 5th among OECD countries, behind Costa Rica, Chile, Mexico, and Türkiye(OECD, 2022). Similarly, according to the 2021 Census, 37 million people, or 11.6% of the U.S. population now live below the poverty line (U.S. Census Bureau, 2022). Poverty rates in the U.S. have seen peaks in the last 40 years corresponding with economic recessions in the early 1980s, early 1990s, and the 2008 recession (Hoynes et al., 2006; U.S. Census Bureau, 2022). However, poverty rates increased for the first time in 5 years in 2020 and remained elevated at 11.6% in 2021 (U.S. Census Bureau, 2022) As the United States lacks some elements of social safety nets implemented by other OECD countries such as socialized medicine, B) it is therefore prudent to determine factors that have historically predicted low-income or poverty in the wake of economic hardships in order to better target social safety programs in the future. This will be especially important in the near future as many economists have reported that the U.S. may be headed towards a recession or may already be in one, and so we may expect poverty rates and the Gini index to continue to rise. C) 1994 was one of such peaks in poverty rates due to economic recession, and as such I will be using the “Census Income” dataset, a subset of the 1994 U.S. Census lifted from the UCI Machine Learning Repository, to address potential demographic variables of interest in determining low-income in the aftermath of a recession. D) I am proposing to use a classification model (specifically a Gaussian Naïve Bayes classifier) to determine which of the dataset variables are most important in predicting a census household income below $50,000/year. This method will employ a train/test/split method and will be evaluated using traditional classification fitness metrics (accuracy, precision, recall, f1). I intend to use Python’s scikit-learn toolkit (Pedregosa et al., 2011) to address this question.

References

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

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

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

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.