PPOL 670 Project Proposal | Immigrant Labor Force Participation in the United States

Ravneet Kaur

March 25, 2020

Research Question

The goal of this project is to explore the determinants of the US immigrant population's labor force participation and develop a model which can explain the varying levels of foreign-born residents' labor force participation between 2016 and 2018. The dependent variable used in the machine learning component is the foreign-born labor force participation rate (%) and the determinants, the independent variables, will include education level, household type, occupation, poverty status, regional price parities, and region of birth. The current political climate focuses on the negative impacts of US immigration policy, specifically that foreign-born residents are taking jobs that otherwise would be occupied by natural citizens. By doing research on the immigrant labor force participation rate, policy makers can use this analysis to dispell misinformation and shed light on the importance of immigrants in the US.

Data Sources

I will analyze the data collected at the state-year level. The majority of the data for the analysis will come from the American Community Survey (ACS) conducted annually by the U.S. Census Bureau and captures demographic, social, economic and housing data. To easily retrieve this data through APIs, I will use the 'tidycensus' package and plan to conduct analysis at the state-year level. In order to measure purchasing power parity (PPP) by state, I will utilize data from the Bureau of Economic Analysis (BEA) that is readily available in .csv format. In order to conduct research across the data sets, I will utilize a 'full-join' to combine the ACS survey data and the BEA purchasing power data sets at the state-year level of analysis.

Methods of Analysis

As stated above, the main data wrangling component will be combining two different data sets (BEA and ACS) between 2016 and 2018. However, it is important to note that the

BEA purchasing power parity data set only has information available up to 2017. In order to complete the data set, I will utilize the same values reported in 2017 to represent data for 2018 PPP. Upon reviewing the data set, this is the best solution since across states, on average, the PPP did not change or did so at a very marginal rate.

As for data visualizations, my goal is to create a geospatial analysis of the number of foreign-born residents across the United States. Another geospatial analysis that can provide helpful insight would be mapping the varying levels of PPP across the United States. To demonstrate varying wealth levels, I will generate bar graphs showing poverty status across foreign-born residents and by age group. In addition, I will generate a plot analyzing education levels across the different age groups. Time permitting, I would be curious to see the average income of foreign-born residents across the different occupations captured int the ACS survey. At this time, my model does not include a lense of gender, but this could be an important determinant and something I will explore in the data set before deciding to add it to the final model.

The machine learning component will explore the relationship between foreign-born labor force participation and its determinants, the independent variables stated initially. To determine a best fit to explain the participation rate, I will use multiple algorithms to explore the relationships, including multiple linear regression model, random forest, and K-nearest neighbors. This will require splitting the final joint data set into training and testing sets, with 25% of the data randomly allocated to the testing set. As discussed in class, this is an important step to make sure our algorithms are not simply fitting the test data and can be applied out of sample.

Defining Success

A successful project would entail cleaning the data in an accurate and efficient manner to best conduct the analysis of the determinants of foreign-born labor force participation in the US. It will require filtering out unnecessary data point and harnessing the correct variables in the correct unit of analysis (state-year). If there are any missing values, it will be critical to consider if the missing values can be estimated accurately or if doing so would seriously undermine the final model. In addition, I am estimating there will be a fair amount of data pre-processing given that the model contains nominal, categorical and dummy variables. I will need to do further research on how the variables are stored and consider how to scale the different variables.

To determine the success of my project, I will compare the root square mean error (RMSE) across the different models (multiple linear regression model, random forest, and K-nearest neighbors) and select the model with the lowest RMSE. After selecting the model, I will look into the top three variables of importance to determine what factors best explain the variance in foreign-born labor force participation. Using this information, the research can help policy makers with an interest in labor participation determine the best way to ensure successful integration of foreign-born residents and allocate funding accordingly.