Determining the effects of population density, age, household and worker income, race, workforce distribution, and education on unemployment Rate in various cities in the United States

Aathithya Selvam[1], Andrew Cheng[1], Ayush Saha[1], Pranav Damal[1], and Advisor: Suresh Subramaniam[1]

[1]Aspiring Scholars Directed Research Program, 43505 Mission Blvd, Fremont, CA 94539

Understanding a city's economy is key for any individual who is trying to be a successful participant in that economy. Crucial to understanding the economy are the effects of socioeconomic factors on unemployment. Using data science principles with the publicity of data, it is important to establish a system by which individuals can understand the factors that play a role in the changes in unemployment, and thereby the economy. Through our project we have devised a system for predicting a city's unemployment rate based on the socioeconomic factors of population density, age, worker earning, education, worker industry distribution, and race within 50 states of the United States. We used a Random Forest model to observe correlations between these factors and unemployment rate within the 50 states, training and testing the data in order to build a classification model for unemployment rates based on these factors. We established our three strongest correlations to unemployment: percent of the populations who were of Caucasian descent, percent of individuals involved in the education work industry, and the median worker earning with unemployment rate respectively. They were all negative correlations.

Unemployment Rate | Socioeconomic Factors | Random Forest Model | Classification

Introduction

Within the United States today, many people face issues with financial security, as evidenced by a Google search for "when to move to a new state" or "when to find a new job." Individuals have managed to combine poor judgement and decision-making with their personal situations to produce a poverty rate of 11.8% in 2018 (1). In addition, according to a summary study conducted on the US Bureau of Labor Statistics, the correlation between unemployment rates and specific degrees attained, whether they may be high school graduates or higher, have a strong negative correlation with unemployment (2). As the degree is more prestigious, the unemployment rate for individuals with that degree sharply seems to decrease. Unemployment and a lack of financial security seem to be inextricably linked to socioeconomic factors, either in or out of control of an individual. Using a Random Forest algorithm to help determine which factors have the strongest influence on the unemployment rate in the United States helps show the link between socioeconomic factors and unemployment rate. These factors, namely population density, age, salary, education, race, and industry distribution, without bias from political sources, and their correlation with unemployment rates, can help inform individuals on their potential for career success. The random forest algorithm works by synthesizing decision trees to classify outcomes by variables within the dataset. Random forest is the synthesis of these decision trees. Based on the mode of the classification, or the most commonly occurring class, the algorithm outputs a predicted value for the data. Each factor that is used in the study is independent of the other factors, which provides the ability to isolate one variable as one with strong correlation as a result of its own effect on the unemployment rate. The data used was collected from the official Census Bureau website, using the 2018 American Community Survey 5-Year Estimates (3). The data collected uses approximately 6 cities from each state in order to provide an extensive range of data. The specific data used to build the model and map out correlations include: population density, median age, percentage of the population over 65 years old, median individual worker earning, percentage of population with a Bachelor's degree or higher, percentage of population that passed high school, percentage of population working in education, percentage of population working in science and engineering, percentage of population of Caucasian descent, percentage of population of African American descent, and percentage of population of other races. Based on the data collected, we built a Random Forest classification model that can assess the unemployment rate of an area based on the values of these factors.

Methods

Data Collection. In order to collect the necessary data, we used the Census Bureau's website, with the aforementioned 2018 5-Year American Community Survey Estimates (3). A specific place would be selected within a state in the United States, and based on the category the data were part of, either social, economic, housing, or

demographic, a specific table was generated for that city by the Census website upon clicking the link. Within this table, the factors selected for the model were recorded.

Data Cleaning and Computation. Once the raw data was collected from the Census Bureau, there were multiple formulas that needed to be processed in order to output the factors that were used in the study. For example, the population density needed to be calculated from the land area and the population by dividing the population by land area in square miles. All percentages were calculated by dividing the number of individuals within the respective group by the total population of the city. However, these values used in the raw data were removed when determining features' effects on unemployment, as they were not a part of the target factors. One key procedure when formatting the data was the replacement of empty data. The Census Bureau does not account for all cities in the United States, nor does it consider every large populated area a city. Therefore, more than 10 new cities had to be found to replace these cities, whose data was unavailable due to the Census' lack of listing them. When cities contained blank data, these cities were replaced with data from a different random city within the same state.

Building the Model. In order to build the model, the Python library scikit learn (sklearn) was used, specifically the RandomForestClassifier module. Once this module was used to instantiate a Random Forest model, the explanatory variables were assigned as such and unemployment rate was assigned as the response variable. Varying the test size of the data changed the confusion matrices and the accuracy scores drastically. The test size was adjusted in order to produce an optimal result, until a percentage of 10% was decided as a good balance between training size and test size. Confusion matrices were plotted, showing how accurate the model was when given a specific amount of data to be trained, as against some other value of data.

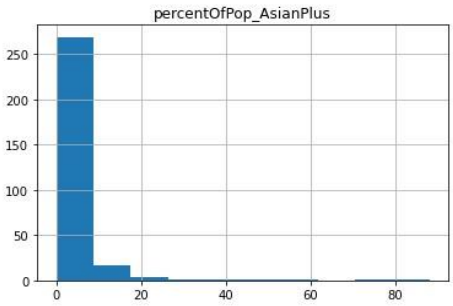

Fig. 1. Histogram depicting the distribution of the percentage of Asian people in each of the cities inside our data set. Note that the over 90% of the data lies in the 0% to 9% range

Testing. Over time, multiple factors were deemed unnecessary because, even within a randomized set ofcities, skewed data caused either an extremely low effect on the data, i.e. in the case of the percentage of the population that was Asian 1. This factor had to be removed due to the lack of distribution which led to a low importance. In addition, while completing the model, in order to improve the accuracy of the model, all factors that could be represented by another concise metric while preserving other major factors were replaced. For example, individuals with Caucasian descent, and African American descent were the final race factors, rather than Caucasian, African American, Native American, Pacific Islander, Asian, and other. A similar procedure was performed on the industry distribution. While building the model, multiple adjustments were made in order to improve the accuracy of the model, in addition to collection of more data. At first, the use of no classification of unemployment rates caused accuracy percentages from 0 to 10%. However, when classifications were implemented, the accuracy of the model nearly tripled to about 28% on average. These classifications allowed for a range of correct outcomes, since the Census Bureau reports an error margin in their own unemployment rate estimates. In order to continue to improve the accuracy of the model, multiple types of intervals were assigned, such as 8 intervals of 1.25% unemployment rate up to 10% and 1 interval for values greater than 10%. However, this did not create the desired response, as the values that were closest to the limits of the intervals were highly erratic. As an adjustment, a weighted system of 3 classifications was used, so that a relatively even number of cities could be a part of each group. These intervals were 0 to 3%, 3 to 4.5%, and 4.5% and higher. Using this method of classification, the results improved significantly again, showing a spike in accuracy percentage by about 20%, and resulted in the highest and final accuracy percentage reached.

Results

Feature	Importance
percentOfPop_White	0.113240
percentOfPopInEducation	0.111286
medianWorkerEarning	0.111102
percentOfPop_passedHighSchool	0.104387
percentOfPop_Black	0.100932
percentOfPopWithBADegreeOrHigher	0.097438
percentOfPopOver65	0.095569
medianAge	0.090154
percentOfPopInSciOrEngineering	0.088702
percentOfPop_Other	0.087192

Fig. 2. This chart depicts the importance value that the trained model assigns

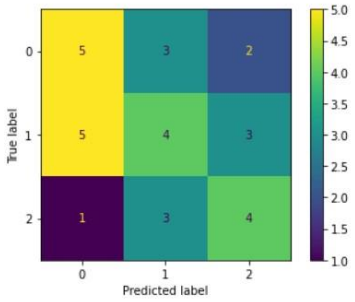

Fig. 3. Confusion matrix showing a representative sample of the accuracy of a sample model. Sample shows an accuracy of 43%, while the program's average accuracy is 47% over the span of 100 models.

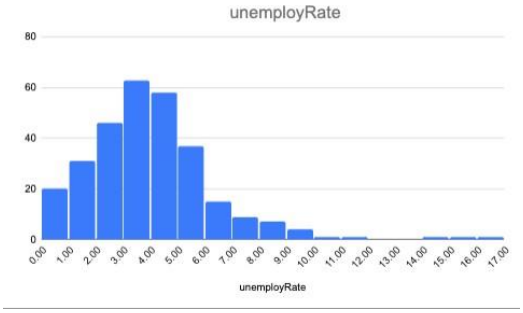

Fig. 4. Histogram depicting the distribution of unemployment rate across the cites in our data set

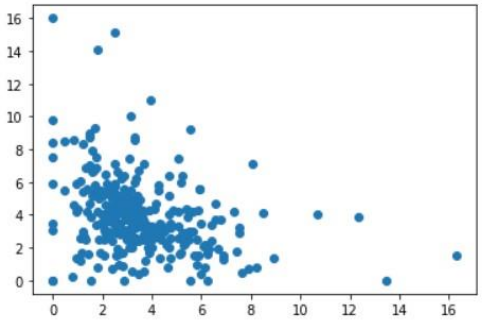

Fig. 5. Scatter Plot depicts the percentage of white on the x-axis and the unemployment rate on the y-axis

Discussion

The results of the project show correlations between each of the factors and unemployment rate, with the strongest of these correlations being between unemployment rate and percentage of population in the education work force, median worker earning, and percentage of population that is of Caucasian descent2. Unemployment rate and percent of population in the education industry are negatively correlated, meaning that an increase of one causes a decrease in the other5. The correlation between median worker earning and unemployment rate is also negative, as is the correlation between the percentage of the population who is of Caucasian descent67. Notice that the model only predicts a low unemployment rate for a true high and a high unemployment rate for a true low for 3 out of 30 times, indicating a better accuracy than just the 47% mean accuracy would seem to indicate3. Additionally, the data is concentrated at the border between low and medium unemployment rates4. While the correlations between these factors are not strong enough to prove causation, they can be used to show some patterns. The percentage of individuals working in education may indicate an increase in student learning, thereby increasing the chance that students will find jobs later. Additionally, a higher average salary may lure individuals to an area where jobs are plentiful. This could in turn correlate with a decrease in unemployment rate. The concentration of Caucasian individuals in the Midwest coincides with the area being sparsely populated, which thereby produces lower

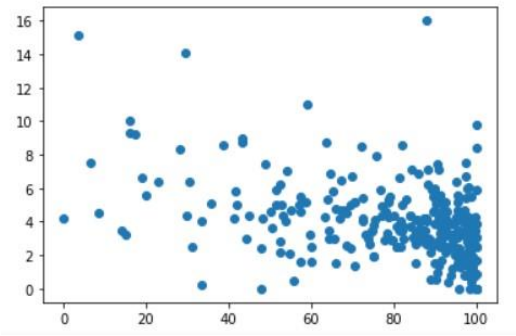

Fig. 6. Scatter Plot depicts the percentage of Caucasians on the x-axis and the unemployment rate on the y-axis

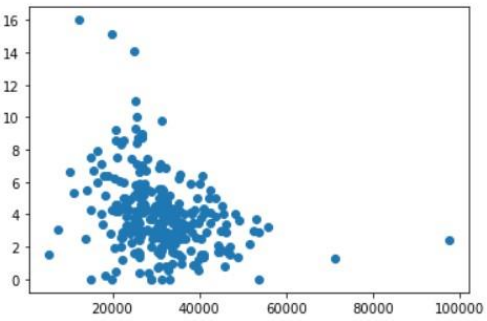

Fig. 7. Scatter Plot depicts the median worker earning on the x-axis and the unemployment rate on the y-axis

unemployment rate.

Limitations. The main limitation of our study is the lack of variety and volume of data. If more data were collected, especially data for higher unemployment rates, the accuracy percentage should be much better, but because most of the randomly chosen data concentrated the unemployment rates at lower rates, the model continued to predict incorrect rates for new conditions. Because of the lack of data, the model's accuracy was not as strong as possible, and the model may have been able to show causal relationships if given an optimal amount of data.

Comparisons. A study conducted by the Educational Department of Development of California defined projections for jobs in Orange County for the future based solely on past job growth(4). With more factors than just the previous data for one factor, our model uses multiple possible causes of growth and decline in employment. However, due to a lack of data, we were unable to create a model like the one defined in the EDD's projections.

Conclusion

All in all, this study was able to establish negative correlations between unemployment rate these 3 factors: the percentage of individuals who were of Caucasian descent, the percentage of individuals involved in the education work force, and the median worker earning. Our study confirms that the percentage of population in

the education industry is an important factor in the unemployment rate, and that the number of people within the area is key as well. Generally, as fewer people populate an area, less competition exists for jobs, as seen with the majority Caucasian population in the Midwest. In order to improve this project, the data set would have to be expanded to encompass multiple cities of high unemployment rates, which would in turn improve the accuracy of the model in projecting those high unemployment rates.

Author Information

Address. Aspiring Scholars Directed Research Program, 43505 Mission Blvd, Fremont, CA 94539

Author Contributions. The work and paper was done through the contribution of all authors.

Acknowledgements

We would like to thank our advisor Mr. Suresh Subramaniam for his excellent advice and assistance in this project. We would also like to thank ASDRP for allowing us to have this opportunity.

Bibliography

1. Jessica Semega, Melissa Kollar, John Creamer, and Abinash Mohanty. *Income And Poverty in the United States: 2018*, 2020 (accessed June 26, 2020). https://www.census.gov/ /publications/2019/demo/p60-266.html#:%7E:text=The%20official%20poverty%20rate%20in, 14.8%20percent% 20to%2011.8%20percent.
2. Elka Torpey. *Education Pays*, 2019 (accessed June 26, 2020). https://www.bls. gov/careeroutlook/ 2019/data-on-display/education_pays.htm#:%7E: text=As%20that%20chart%20shows%2C% 20the,unemployment%20rate% 20of%203.2%20percent.
3. United States Census Bureau. *Explore Census Data*, n.d. (accessed June 26, 2020). https://data.census.gov/cedsci/.
4. Employment Development Department. *Employment Projections,* n.d. (accessed June 26, 2020). https://www.labormarketinfo.edd.ca.gov/data/ employment-projections.html.