**County Job Growth Analysis**

**Overview**

My task was to compute the job growth rate in the US by county between 2015 and 2020 and use linear regression to identify the factors most associated with those changes.

**Methodology**

My input data came from a County Health Rankings dataset provided by EDai, and 2015 and 2020 employment counts from field DP03\_0004E from the Census Bureau's 5-year American Community Survey. The DP03\_0004E field is a count of the population aged 16 and older who are actively employed in a civilian job. I saved the County Health Rankings dataset as a comma separated value file, and then imported it into Python. For the census data, I created a function in python to make an API call to the census bureau, and saved tables containing each year's employment numbers. I merged the 2015 and 2020 employment numbers into the County Health Data. I cleaned the merged data, checking for duplicate rows and columns, and then handled null values.  I decided not to use any column where greater than 30% of the rows had a null value. For all other null values, I substituted the average of the values that were present in the data. I then performed a multiple linear regression in Python and also in R to determine which of the factors/features are most correlated with the employment growth rate. In the multiple regression I used a significance level of 0.01. I then compared the resulting p-values to the significance level, and those factors whose p-value is below 0.01 are considered to be significant.

**Findings**

The significant factors associated with the Employment Growth Rate are:

Chart

Description automatically generated

The r squared value from the analysis is 25.1%, which means that the model would be a weak predictor of the employment growth rate.

**Limitations**

The two greatest limitations I can see for this analysis are the broadness of the dataset and the volume of missing data. The dataset is very diverse - the analysis includes 77 different independent factors that were used to predict the job growth rate in more than 3,000 counties across the US. That is likely why the accuracy is low, because there are so many things going on that it is hard to predict well using the entire dataset. Also, many of the columns are missing data. I filled in the average for those where the columns contained data for at least 70% of the rows to preserve the information that was provided in those fields, but filling in more of the missing values would likely result in a more robust and accurate model. Also, we need to remember that no matter how many columns we have, we will never be able to account for all of the factors that are involved in the rate of job growth, but helpfully we can evaluate how good our predictions are.

**Proposed Improvements**

Two ways to improve the accuracy of the predictions are through breaking the data into smaller, more homogenous subsets, and also by trying other predictive models.  For example, when I reran the regression with data broken into 7 ranges based on the percentage rural, the accuracy improved for all but the most rural 2 groups, with rates of 49.8% to 67.3%, a great improvement in accuracy. I also used unsupervised learning to create 7 clusters, and when I ran those separately, the accuracy improved to as much as 56.4%. Also, when the data is broken into subsets, the significant factors which emerge reveal trends in the individual subgroups which are not apparent when you look at the data as a whole.

Additionally, other types of machine learning analysis could also be helpful, such as random forest models or neural network models. I set up models for each of these types with the whole dataset and their outcomes were only slightly improved, with an accuracy of 26.8% for the random forest model and 25.2% for the best of the neural networks.

A third way to improve accuracy might be to have fewer features in the model by combining factors from the data. For example, a column for health-related features could be created by combining health related fields, and similarly this could be done for other areas. A model like this would show broader results by category of factor, and some detail would be lost, but also probably a lot of noise would be eliminated from the data. I didn’t perform this analysis, but it would be interesting to run it and see what the impact on the accuracy would be.

As to which model should be selected, you should be guided by the accuracy of the model and by what you are seeking from the data. If you are really interested in creating a predictor of job growth, you need to choose the model that is the most accurate. If you are more interested in finding what factors are significant, accuracy is less important, because the regression will still likely return the factors with the most importance. You need to choose your data on the right scale as well – if you are looking for predictors or significant factors analysis for large urban areas, the above model of the whole US is not the best choice. From these models, I would choose one of the rural vs. urban subset groups as a predictor based on the location of the requested analysis, unless the request was for information on a national level.