**Random Forest Model Analysis**

**Overview**

A random forest machine learning algorithm combines many small decision tree predictors to create a complex ensemble model that can be very powerful. The combination of many small samples of the data to build the predictions can lead to much greater accuracy, and because each decision tree is based on only a small bit of the underlying data, it provides protection against overfitting the data. (Overfitting is when the model fits the data too well, and can predict that dataset, but performs poorly when presented with new data). Random forest models make categorical predictions instead of the continuous numeric predictions returned by a linear regression.

**Methodology**

I created a random forest model to predict job growth. Since the random forest model returns a categorical prediction, I converted the employment growth rate into bands 2.5% wide, with all growth rates greater than 15% going into the top band and all rates below -10% going into the lowest band. I converted the state field to 50 Boolean fields using a function called get\_dummies. For example, the new variable state\_virginia is 1 when the county is in Virginia, and 0 for all other states. I then scaled the data to eliminate noise from very large and very small values in the data, and fit and tested the model and ran the predictions.

**Results**

The random forest model performed slightly better than the multiple regression, with an accuracy of 26.8%. One of the interesting outputs from the random forest model is a report of feature importance, which is similar to the significant factors that emerge from regression analysis. As with the significant factors from regression, these factors can be used to narrow down your model or to simply inform you of what features in the data have the most weight in the prediction. Here are the top 20 features from the analysis:

A picture containing text

Description automatically generated

Interestingly, only two factors appear on both the list for regression significant factors and the random forest feature importances list: Primary Care Physician Ratio and % Less than 18 years of age. That suggests that these two factors are important no matter how you look at the data. The differences in other factors reported is likely due to the differences in the models and the large amount of noise in the data. Figuring out why the factors that emerge are different would be an interesting topic to explore.