Electricity has always been at the heart of residents’ daily life, industrial production, and transportation. Thus, the price of it can affect citizen’s wellbeing and economic development deeply. However, the unforeseeable fluctuation of electricity price can greatly disturb the normal conduct of residents and businesses by increased their cost. A reliable prediction of electricity price can help people and businesses to plan, avoid unnecessary cost and maximize output. This project aims to collect data on electricity generation, fuel consumption and weather conditions of state of Virginia to form a model to predict future electricity price in Virginia.

According to the Energy Informatio Administration, fuel, transportation, power plant cost, weather condition and regulation have great impact on electricity price.

This project will employ both traditional machine learning techniques such as linear regression, KNN and random forest and Time Series methods to generate a plausible model for predicting future electricity price and gain insights from the models.

This project will employ datasets such as monthly electricity price of state Virginia, monthly generation of electricity by source, the consumption of fossil fuels when generating electricity by source and average temperature and precipitation by month. There are two other variable called internet search and covid. They reflect public perception of the electricity price and whether there’s covid. These data will be organized in a single data frame with month/year as index and other features as columns. Due to the availability of data, this project would only include data from 2001 to 2020.

The model that I plan to use are as follows: traditional machine learning models includes Linear Model

KNN with K tuning parameters

Decision Tree regression with the Max Depth parameters

The Bagging decision tree regression model and

Random forest with the N Estimators and max\_depth tuning param

For Time Seires model, I plan to use ARIMA.

The rationale behind my choices of two different model types is electricity price is subject to many factors such as its generation, weather condition etc but it is also a variable that is highly correlated with its previous self. AKA autocorrelation. Building two types of model can yield a more comprehensive view and can show better improvement

So here’s some preliminary results. The best performing model is the RandomForestRegressor with a max depth of 5 and 300 estimators. The R square socre is almost at 0.9. And here’s the plot.

As we can see from the Permutation Importance plot natural gas consumption, coal consumption and internet search are the top three important features. And others have surprisingly small contribution to the prediction. In the ICE plots, we can see that there are heteroscedasticity to some extent in the internet search variable. Maybe there are some subsets of data that need to be distinguished.

In the arima model, we can see the general trend in the first plot. After we separate the trend, seasonality and residuals, We can see each part more clearly. From the trend, we can see that there is a significant bump around 2008. And of course electricity price is heavily affected by seasonal change, so we can observe clear seasonality patterns.

In the next plot, I used 2015 and after as test data. The prediction is pretty good, but only on the basis that the later prediction rest on the actual data of the previous time, not on the predicted ones.

As for future efforts, First, I would like to find a way to boost up the importance of weather conditions. It is because electricity does vary in different weather setting in data. Secondly, natural gas consumption and coal consumption contribute a lot in the prediction, which is not surprising because they are still the two main source of electricity generation. I would probably add the prices of coal and gas as well for they can raise the price of electricity if they are high. Thirdly, from the trend we can see that there is a price bump around 2008, probably because the 2008 financial crisis. Considering this, I would probably add another dummy variable to control for the financial crisis.