Capstone Project Summary

Capstone Project Name: Econometric Modeling of Power Transmission Rates in the United Kingdom

Student: Supervisor:

Table of Contents

Project Description

The goal of this project is to forecast Transmission Network Use of System tariff in the case of United Kingdom(refer to glossary). This tariff is paid by the suppliers and generators. The supplier tariff is also called demand tariff and this tariff is divided into 2 parts: Half Hourly Zonal tariff and Non Half Hourly Zonal tariff. After carefully discussing with the host organization, the tariff that will be analyzed in this project will be the Half Hourly tariff. Moreover, the Half-Hourly zonal tariff is location-based, so the amount that the end-user will pay depends on their location.

Note: One demand zone(Northern Scotland) is dropped from the analysis, due to some unexpected political changes(refer to glossary).

Implementation

Data collection

After carefully reading about Half Hourly demand tariff, me and the project members discussed the features needed for the analysis. The features(variables) collected during this projects were: Historical TNUoS tariff(for 12 years), National Demand, Financial information retrieved from National Grid reports, Stock data from National Grid and Inflation index from 2007 until 2019.

Explanatory Data Analysis

The data after collection, was grouped and resampled yearly in a data frame and it was from 2007 to 2019. The main issue was that the tariffs were yearly which meant that I only had 12 data points for all 14 demand zones(locations). After plotting half hourly tariff for all 14 zones, I came to realization that this tariff is increasing every year, however in the case of the Scottish demand zones(especially Northern Scotland), the half hourly zonal tariff was increasing until 2016 after which it decreased. Another important realization was that the National Demand is decreasing throughout the years as opposed to the Half Hourly tariff. From the financial situation of the National Grid, I came to realize that their revenue, amortization costs and operation costs are increasing, however the National Demand is decreasing. The National Grid must pay their workers and their projects for improving and optimizing the grid and this tariff is used to recover the cost and constrain the profit, therefore regardless of the demand, these costs must be recovered.

Moreover, the conclusion from the EDA was that the more southern the demand zone, the higher Half Hourly tariff is(Southern Scotland being the lowest and London being the highest). The features that were retrieved were highly correlated with average Pearson coefficient of 0.8. Furthermore, multiple scatter plots were created to see the relationship between the Half-Hourly tariff and the other features. The distribution of almost all features were either normal or log normal(in the case of national demand). Please refer to the glossary for visualizations.

In order to see the relationship between the features, I performed grangers analysis (refer to glossary), and target feature was Half-Hourly zonal tariff, for each of the demand zones. Moreover as good predictors I assumed the features with p-value below 0.05. After which these selected features were fed into Random Forest in order to see their

importance. However, after discussing with the Senior Data Scientist in the host organization, the grangers analysis was not accurate, and therefore dropped from the analysis.

Regularization and OLS

After realizing that coefficients are correlated, I continued the analysis with regularization(LASSO) and performing OLS (Ordinary Least Squares) linear model for features that were not shrank to zero for each of the demand zone. The LASSO(Least Absolute Shrinkage and Selection Operator) was performed with aim to select features. As conclusion the most repeating features with p-value less than 0.05(from the OLS results) were: operation cost of National Grid, the Non Half Hourly tariff and the average adjusted close price of the National Grid stock. I need to mention that the regression is done with small sample data, however after discussions with experts in this field, the results are expected, and with this analysis they can quantify the relationship between the features.

Feature importance with Random Forest

I also created Feature Importance plot with the help of the Random Forest algorithm for all of the demand zones in order to see which features are good predictors of the Half Hourly Zonal tariff. The 5 best reoccurring features(for the 14 demand zones) are: revenue of National Grid, the Non Half Hourly zonal tariff, National Demand, the adjusted close price of the National Grid stock and liabilities from National Grid. There is a theme that the state of the National Grid is affecting the Half-Hourly zonal tariff. The only issue is that lack of data, however after consulting with professionals in the field, I was reassured that the results are as expected. Moreover the same method was used for the features that were not shrank to zero in the LASSO Regression, and similar results were produced.

Time-Series models

After feature selection, I continued with time-series models. For the time series models, I did not use differencing because of not having enough data, therefore the time series were not stationary(mean and variance not constant over time) which constraints the use of more time-series models.

Models that were created for this problem

1. Facebook's Prophet model
2. Holt's model
3. Dynamic Regression model

Important to note is that the models require more data and more models were attempt(Box-Jenkins models, Simple Exponential Smoothing, VAR) but poor results were produced.

The prophet model was created with interval width of 0.95 and the results show that each of the demand zone tariff will increase in the future. The model was evaluated on the National Grid forecasts and the results have small RMSE. The distributions of the residuals are normal, with a bump from the Northern Scotland zone, which was requested from the host company to overlook that and not to spend time on that zone.

As next model, I decided to create a simple Holt's model with smoothing level of 0.3 and smoothing slope of 0.2 and also the errors are normally distributed with a small bump from the Northern Scotland demand zone. The coefficients were retrieved with trial and error.

The last model I choose was Bayesian dynamic linear model (Harrison and West, 1999) with two trend components and one auto-regressive component, after trial and error. This choice yielded the best results and the residuals were noramlly distributed(mean of zero and constant variance).

For results please refer to glossary.

Challenges

The first challenge experienced was the lack of data. There are not enough data points to follow the standard Box-Jenkins methodology, and basically there is nothing to be done. I was left to try every model that I was familiar to utilize and produce results. The models specified above were just the models that yielded results similar to the National Grid forecasts. The famous ARIMA, AR, MA models could not be used because stationarity was an issue and also the results produced from these models were extremely inaccurate (residuals were not Gaussian , some forecasts were even negative).

Moreover, extracting features from the National Grid reports was a challenge, mainly because the reports are not consistent throughout the years, and the packages like PyPDF and tabula could not extract the features. The reports were yearly and their length was around 110 pages. Therefore, reading all the reports was extremely slow and inaccurate almost always, even with complex regular expressions.

Bayesian techniques were new to me, therefore I was in need to learn new methods, however because of the time frame that I had and complications from the new global emergency, this was harder than expected. Moreover I attempted to use the dynamic linear model with some state components and learn that by actually performing it on the data.

Lessons Learned

The most important lesson that was learned was the importance of domain knowledge. For the first time I was faced with electricity market in the UK. In my opinion, domain knowledge accounts for large part (even bigger than the models and analysis) in one project. If I had been more familiar with electricity market, the project might have been more accurate.

The next lesson was the importance of having clean and structured data. In this project, I collected data from public sources, because of some complications in the organization that were result from the Corona virus. The initial plan was for the company to provide internal data. Nevertheless, the organization was extremely helpful with providing advice on which data to collect and documentations to read about the electricity market. However, because of more data sources, cleaning and merging data was more difficult, especially reading data from yearly results, on which I failed to retrieve automatically.

Moreover, learning how to communicate with people who are not in the field of data science was a challenging lesson to me, because I was encouraged to explain the techniques and methodology used in the project. Learning this skill will be useful in my future projects. The project's members were extremely patient,understanding and eager to learn the technicalities of some methods. In my opinion this kind of behavior will motivate to perform better in projects. Furthermore, the freedom that was offered by the host organization was more than enough and this concept is also motivating and results in better performance of the analysis. To sum up the knowledge and experience are important, however the people that are included in a project are also making a strong impact on the analysis.

Planning and making small goals in the project's lifetime is crucial for the success of the projects. This methodology was encouraged by the project's members and this concept will be used in my future projects.

Glossary

Granger Analysis

Granger causality is a statistical concept of causality based on prediction. If a signal X_1 "Granger-causes"(not causes) a signal X_2, then past values of X_1 should contain information that helps predict X_2 above and beyond the information contained in past values of X_2 alone. Just to note that this is not the true causality. The analysis contains two hypothesis(one that X_2 does not cause and one that it causes). According to the p-value, a variable can be choosen, in this case it was used the famous p-value of 0.05.

Visualizations

- Distribution of variables

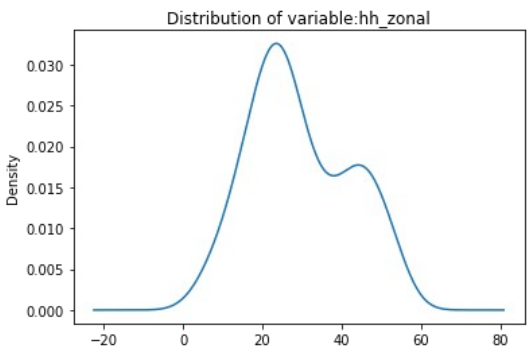

Illustration 3: Distribution of Half Hourly Zonal Tariff for all UK

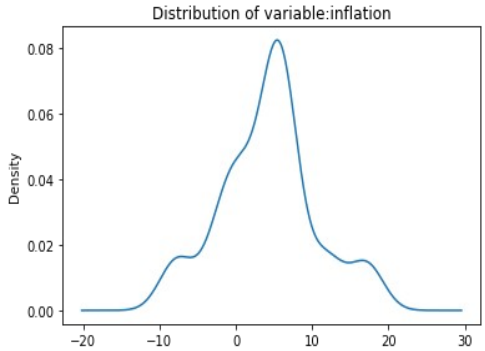

Illustration 1: Distribution of CPI Inflation

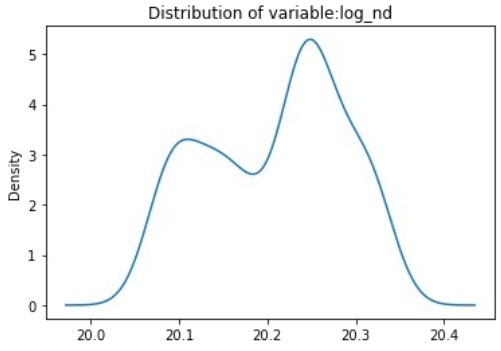

Illustration 2: Distribution of National Demand(natural log taken before)

- Plotting of Half Hourly Zonal Tariff for all 14 demand zones

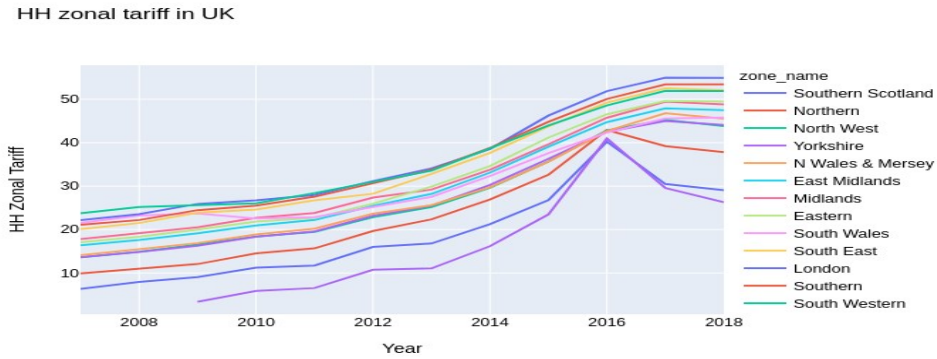

Illustration 4: Half Hourly Zonal Tariff for the 14 demand zones

- Correlation heatmap with Pearson correlation coefficients

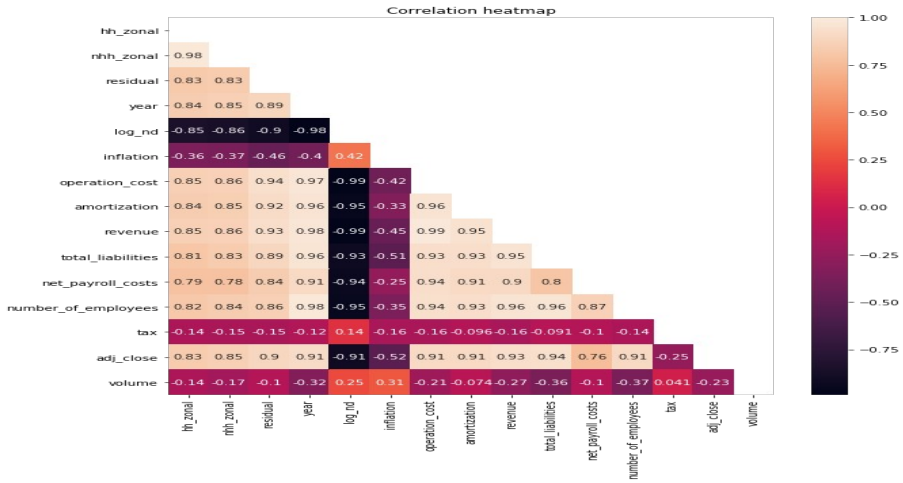

Illustration 5: Correlation heatmap with Pearson correlation coefficient

- Mean Half Hourly Zonal Tariff by demand zone

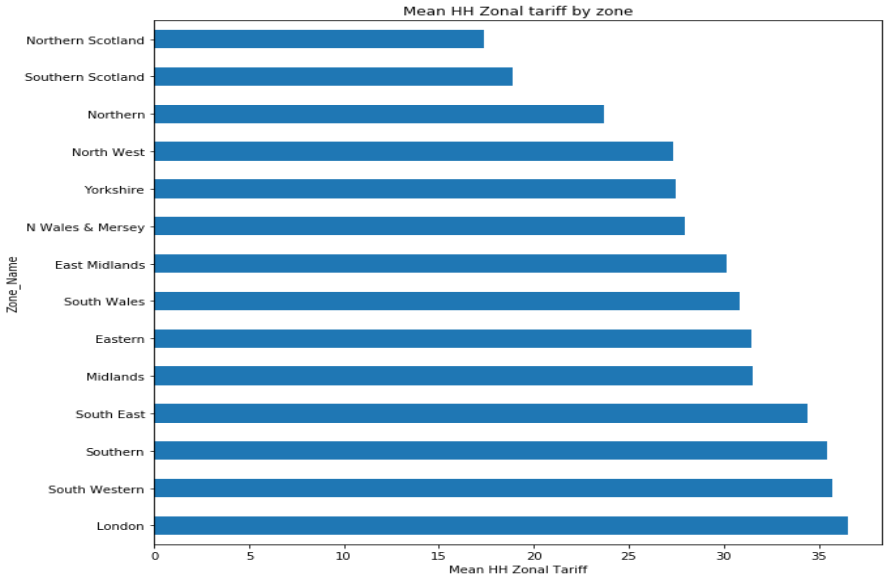

Illustration 6: Mean Half Hourly zonal tariff across years by demand zone

Random Forest

Random Forest is a machine learning algorithm that uses decision trees for predictions. A Decision tree is a graphical representation of possible solutions to a decision based on certain condition. In the first step, one variable is selected(where the split has lowest error on predictions), and divided in 2 parts. After this step, the 2 nodes are also divided into 4 nodes(each node into 2 parts), again depending where the error is lowest. This algorithm builds more trees like that, and in order not to have same trees(because data is not changing, every tree will be the same as the previous one) there is a method that makes the trees uncorrelated,hence the name Random Forest. According to these trees, important variables can be selected(which variable was more times selected throughout the trees.)

Lasso Regression

Lasso Regression is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. In this project the LASSO is penalizing some variables (if variables are correlated) and this method will output coefficients that are zero and non-zero. The zero coefficient variables are the ones that LASSO penalized.

Prophet Model

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data[1]. This model is made by Facebook.

Holt's Model

Holt's two-parameter model, also known as linear exponential smoothing, is a popular smoothing model for forecasting data with trend. Holt's model has three separate equations that work together to generate a final forecast. The first is a basic smoothing equation that directly adjusts the last smoothed value for last period's trend. The trend itself is updated over time through the second equation, where the trend is expressed as the difference between the last two smoothed values. Finally, the third equation is used to generate the final forecast. Holt's model uses two parameters, one for the overall smoothing and the other for the trend smoothing equation. The method is also called double exponential smoothing or trend-enhanced exponential smoothing. [2]

Dynamic Linear Model

Dynamic Linear Models are a special case of general state-space models where the state and the observation equations are linear, and the distributions follow a normal law. They are also referred to as gaussian linear state-space models. Generalized DLMs relax the assumption of normality by allowing the distribution to be any of the exponential family of functions (which includes the Bernoulli, binomial and Poisson distributions, useful in particular for count data).

TNUoS Tariff

The price, to create energy, is decreasing, however the energy bills are increasing at a fast rate across the globe. The energy costs can be divided into two parts: commodity costs and non-commodity costs. The commodity costs are the actual price of creating energy or the actual cost of the fuel consumed which is a tradable commodity – gas or electricity. Non-commodity costs are charges added to an energy bill which originate from the government and third parties such as distribution companies. By and large these non-commodity costs are used to cover the cost of system and network charges involved in the running of the distribution and transmission network. There are many types of non-commodity costs such as TNUoS (Transmission Network Use of System), DUoS (Distribution Use of System) Charges, BSUoS (Balancing Services Use of System), however this project is focusing on the TNUoS charges which covers the cost of using the transmission network. This cost or tariff as mentioned from the National Grid(transmission network operator in the UK) are paid by generators(those who produce electricity) and suppliers(those who distribute electricity). The TNUoS tariff is determined based on the location of the users and whether they are generators or suppliers. Moreover the businesses(mainly industrial/commercial) which are directly connected to the network are also liable to pay this tariff(those companies that need high-voltage electricity).

The tariff, this project is focused on, is the supplier tariff which is also called demand tariff. The area of UK is divided into 14 demand zones(locations) and each of the demand zone pays different amounts. The demand tariff is also divided into 2 parts which are the following:

1 https://facebook.github.io/prophet/

2 https://link.springer.com/referenceworkentry/10.1007%2F1-4020-0612-8_409

- Half Hourly metered demand for users directly connected, mostly commercial/industrial users.
- Non Half-Hourly demand is for small users, mostly domestic and small businesses

The project will be about forecasting Half Hourly Zonal tariff.

For better understanding there is a graph I created:

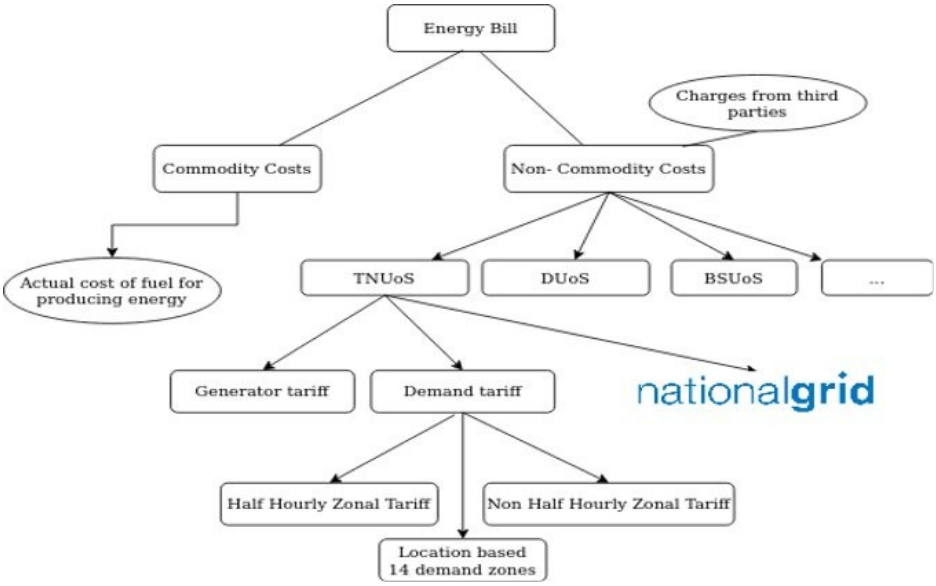

Illustration 7: Graph of Electricity Non-Commodity Costs

Results

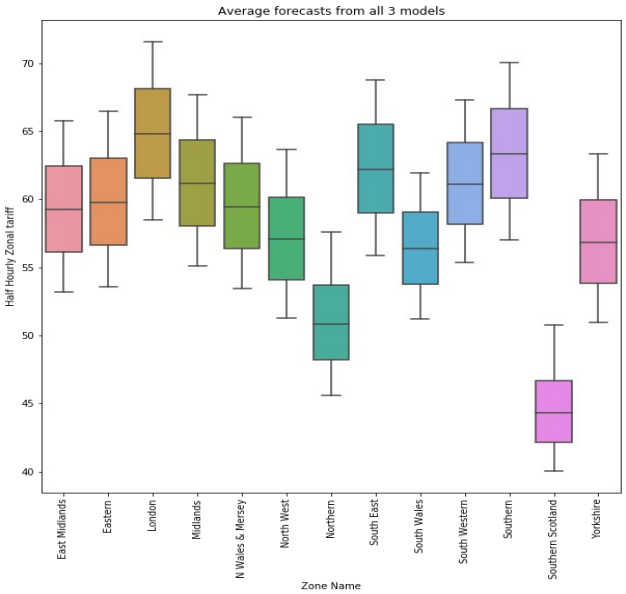

Illustration 9: Distribution of the average forecasts for each of the demand zone

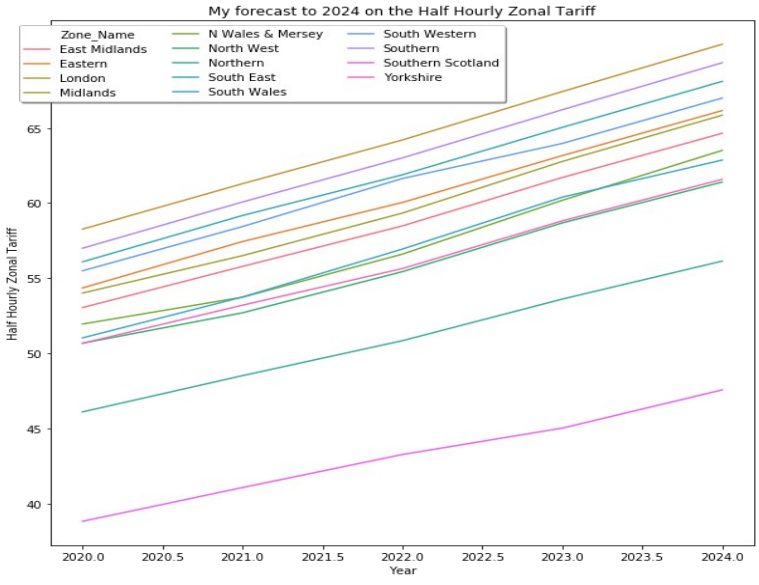

Illustration 8: Average forecast for 2020-2024 for each of the demand zone

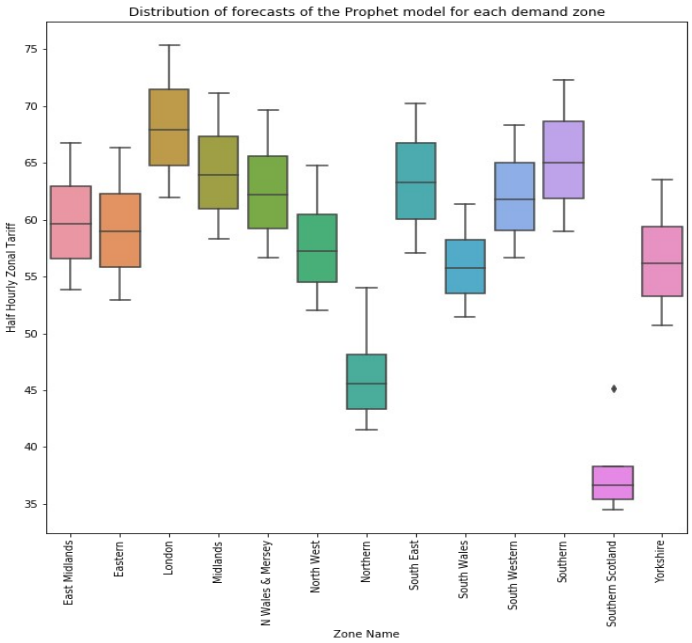

Illustration 10: Prophet forecast distribution of each demand zones

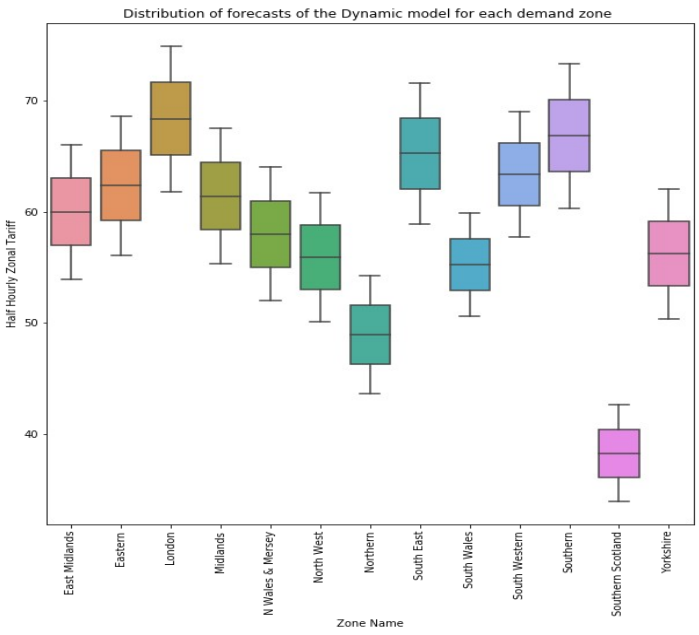

Illustration 11: Dynamic Linear Model forecast distribution for each of the demand zone

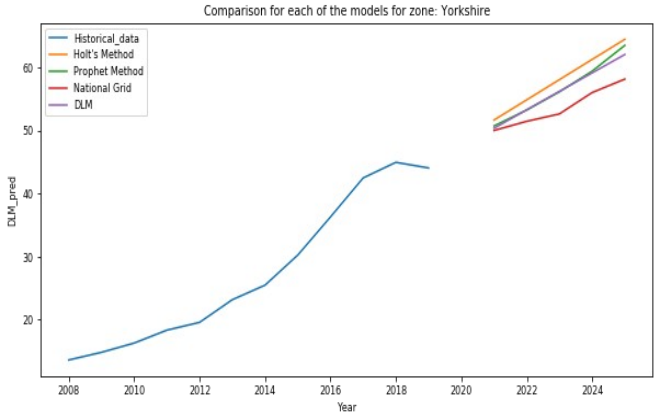

Illustration 13: Forecast for Yorkshire demand zone

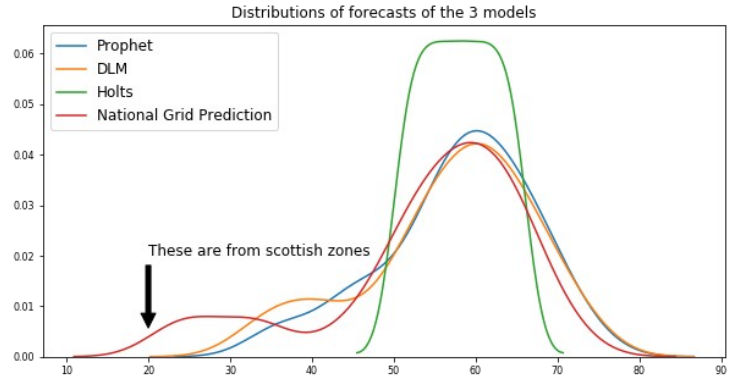

Illustration 12: Distributions of forecasts of the 3 models and the National Grid Predictions

From the plots above, it can be concluded that Northern and Southern Scotland have the lowest TNUoS Half Hourly zonal tariffs, and the London demand zone is expected to have the

highest Half Hourly tariff. In the Illustration 11, it is showed the forecast of the Yorkshire demand zone, and the plots are similar for the other demand zones.

My conclusion would be that the Half-Hourly tariff will increase, as it was increasing for the past years, however the confidence intervals also suggests that it may decrease. After analyzing the confidence intervals, I realized that the intervals are larger on the positive side, which makes it more probable that the Half-Hourly tariff will increase.

Furthermore, this tariff are susceptible to be changed by the regulatory organization and this change is hard to model and predict. For example, the Northern Scotland demand zone after 2016, changed because of some political reasons(as discussed with an expert in this field).

From Illustration 12, the distributions of the forecasts from the 3 models, it can be seen that they follow Gaussian distribution, and the errors that can be seen are from the Scottish zones(excluding Northern Scotland demand zone).

Therefore, these models are not capable of forecasting political changes. I would like to bring one quote from British Independent Utilities:

" The ageing DUoS and TUoS systems are likely to require significant investment within the next decade. This could be brought sooner with the influx of Renewable energy onto the local grids and forecasts could change depending on the result of the next general election."

"There is still a lot of uncertainty around the future prices of the non-commodity charges, but one thing is for certain they will be increasing"

Furthermore, the impact of COVID-19 was not taken into consideration,therefore the forecasts might be inaccurate.

Recommendations

The results from the project are not statistically significant, and have huge confidence intervals. Moreover, the changes that are happening due to some political changes could not be captures. Therefore I recommend collecting additional data sources on news articles or some consultation documents from Ofgem.

In terms of modeling, numerous models could be created, however I had limited knowledge when it comes to time-series. I would assume that some state-space models could predict and model changes and trend better than the models utilized here.

In terms of data collection, the way of collecting data from reports was manually, therefore, more experienced analyst can overcome this issue, and read data automatically from the raw reports from National Grid.

The code that was used to perform this project, is done as an ad-hoc analysis, and therefore, the code is not for deployment purposes, thus refactoring is needed. The notebooks and helper functions can be found in the GitHub repository.

For more information, please feel free to contact me on: sebair_selmani@yahoo.com