**5- Modelling**

The data is almost ready for use in machine learning models. Before running any models, the following steps must be considered:

* Feature engineering can be used to intuitively add features that may be beneficial as the dataset has 4 features which will not be enough for a model this dynamic.
* Cross validate models where possible.
* Compare the best model of each model
* Dividing the training and testing set.

The dataset contains observation from **2015 to 2018**. **2015 to 2017** was used as the training set and **2018** was used as the testing set.

1. **Feature Engineering**

**Set 1: Current Features (All aggregate weekly**

1. Maximum temperature
2. Minimum temperature
3. Mean temperature
4. Mean pressure
5. Mean humidity
6. Maximum windspeed

These are directly weather related from the dataset.

**Features that were added**

**Set 2** = Percentage Changes from Week to Week of all the weather features in set 1. Qty = 6

**Set 3** = Datetime features. (Week of year, Month of year and Quarter of year). Qty = 3

**Set 4** = Difference of demand between current week and n-weeks ago. (1 & 2 weeks used). Qty =2

**Set 5** = Difference of max, min and mean temperatures between current week and 1 week ago. Qty=3

**Total features= 6 (Original) + 14 (Engineered)**

1. **Time Series Cross Validation Techniques**

A screenshot of a social media post

Description automatically generated

To cross validate a Time Series, the training set must be split using TimeSeriesSplit(n\_splits=n).

Traditional splits will not work as each datapoint is dependant on the previous.

The figure visualizes the difference between TimeSeriesSplit() and Traditional Train-Test Splits.

There will be many combination of parameters to loop through. Thankfully there are tools that help such as

* Randomized Search CV. **Inputs**:
  + Estimator (Random Forest in this example)
  + Range of Parameters to randomly search through.
  + Cross validator (TimeSeriesSplit)
  + Number of iterations
* Grid Search CV: **Inputs**:
  + Estimator (Random Forest in this example)
  + Range of Parameters to randomly search through.
  + Cross validator (TimeSeriesSplit)

|  |  |
| --- | --- |
| **Randomized Search CV** | **Grid Search CV** |
| Relatively Faster Computation | Slow/Expensive Computation |
| May not get the best parameters | Better to minimize loss function |
| Reasonable to accommodate larger range of values | A large range of values can cause memory problems |

Most of this project utilized Randomized Search CV but Grid Search was used only once just for testing purposes as it takes much longer to fit a model.

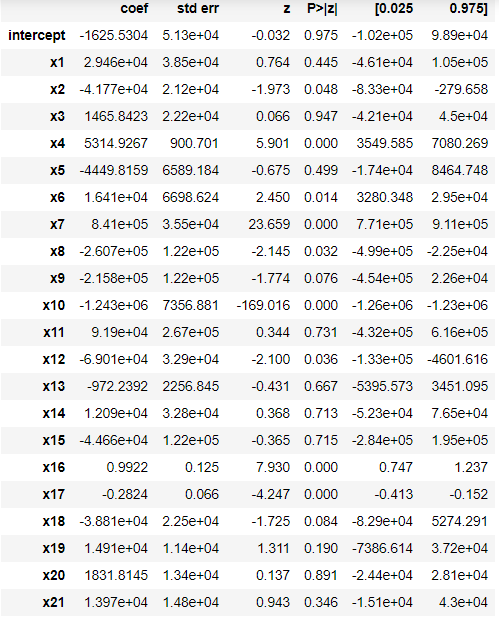
1. **Modelling**
2. **SARIMAX**
   1. **Parameter Choice**

* Parameters = (p,d,q)(P,D,Q,m)
* q/Q = Trend/Seasonal autoregressive Order
* d/D = Trend/Seasonal differencing order
* p/P = Trend/Seasonal Moving average order
* m= Seasonal cycle = 52 weeks
* All orders are integers

D is the significance of seasonality and was tested using the dicky fuller test. It returned a low p-value and negative test statistic which implies the differencing order can be set to 0.

The other parameters can range from 0 to infinity. But our limits will be set between 0 and 5 and the model that return the lowest AIC score (Akaike information criterion).

Pyramid Arima has provided Auto Arima which runs through the different combinations of parameters retrieving the best model. After the first model was run, features and residuals were analyzed using some summary statistics. After running the test the best model with the lowest AIC retrieved was **(3,0,0)(1,0,0,52).**



* 1. **Feature Choice**

Part of SARIMAX summary statistics retrieves the table on the right. The column of interest is **P>|z|.** This signifies how important the features is in prediction. Features with values greater than 0.05 were removed from the analysis and the model was reran.

The model with less featured performed better on the training set but fell behind the original model.

* 1. **Residual Analysis**

The following are the results available to analyze residuals.

1. Prob(Q) & Prob(JB)

**Prob(Q)** = p-value for the null hypothesis that residuals are uncorrelated. If greater than 0.05 then the residuals uncorrelated.

**Prob (JB)** = p-value for the null hypothesis that residuals are normally distributed. It is greater than 0.05, then the residuals are indeed normally distributed.

1. Residual Plots

Residual distribution(Orange) vs

normal distribution (Green).

- Mean is at 0

- Std is narrower than normal

This is acceptable

Residuals over time showing no pattern which is good

A close up of a map

Description automatically generated

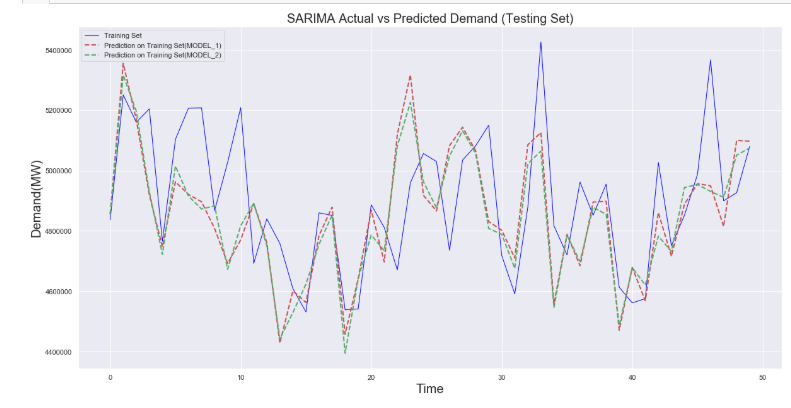
Residuals do not show significance with each other so they are uncorrelated

This is good

Residuals lining up well with normal distribution line

Residuals showing no reasons for concerns about prediction. Parameter choice looks like the best choice with current dataset and computation.

* 1. **Results**



Testing Set

Model\_1 Prediction(Testing)

Model\_2 Prediction(Testing)

The graph above shows the testing set(Blue line), Model\_1(Dotted Red line), is the initial model with all features and Model\_2 with reduced features.

Model\_2 returned the better results on the training and testing sets. Here are the results on the training and testing sets.

***Testing Set***

RMSE = 196,000 MW

= **3.99%**

***Training Set***

The RMSE is 98,000 MW.

= **2.04%**

Cross validation could not be performed on this SARIMA model as parameter choice was not complex.

1. **Random Forest Regression**
   1. **Parameter Tuning**

* n\_estimators = Number of trees use in the model
* max\_depth = Max depth of each decision tree used
* max\_features = Max number of features used to decide each split

|  |  |
| --- | --- |
| **RANDOMIZED SEARCH CV** | |
| n\_estimators | Range(100,800) . Interval = 50 |
| max\_depth | Range(1,150). Interval = 1 |
| max\_features | Range(2,15). Interval = 1 |

The above ranges were used for Randomized Search CV. Grid Search CV could not handle running these ranges so they were reduced to the below table.

|  |  |
| --- | --- |
| **GRID SEARCH CV** | |
| n\_estimators | Range(100,500) . Interval = 50 |
| max\_depth | Range(1, 50). Interval = 2 |
| max\_features | Range(2,12). Interval = 1 |

* 1. **Results**

1. *Best Parameters*

|  |  |
| --- | --- |
| n\_estimators | 100 |
| max\_depth | 9 |
| max\_features | 4 |

1. *Errors*

The following results are those on the testing set. Both models have achieve the exact same parameters. Even though Grid Search took less time to run, less range of values were tested due to memory issues.

|  |  |
| --- | --- |
| **Randomized Search CV (Iterations=3000)** | **Grid Search CV** |
| Run Time = **1 hour 53 minutes** | Run Time = **50 minutes** |
| RMSE = **171,600 MW** | RMSE =**171,600 MW** |
|  |  |

1. **Gradient Boost Regression**
   1. **Parameter Tuning**

n\_estimators = Number of trees use in the model

max\_depth = Max depth of each decision tree used

max\_features = Max number of features used to decide each split

subsample = Fraction of training data to be randomly samples for each tree.

learning\_rate = Controls weight of new trees added to the model based on errors

|  |  |
| --- | --- |
| n\_estimators | Range(100,800) . Interval = 50 |
| max\_depth | Range(1,150). Interval = 1 |
| max\_features | Range(2,15). Interval = 1 |
| subsample | Range(0.001,0.99). Number of points = 50 |
| learning\_rate | Range(0.001,1). Number of points = 30 |

Randomized Search CV was run at 3000 iterations and the below results were found.

* 1. **Results**

1. *Best Parameters*

|  |  |
| --- | --- |
| n\_estimators | 600 |
| max\_depth | 52 |
| max\_features | 3 |
| subsample | 0.057 |
| learning\_rate | 0.03 |

1. *Errors*

|  |  |
| --- | --- |
| **Training - Iterations=3000** | **Testing - Iterations=3000** |
| RMSE = **99,300 MW** | RMSE = **154,800 MW** |
| **2.07%** | **3.16%** |

1. **XG Boost Regression**
   1. **Parameter Tuning**

n\_estimators = Number of trees use in the model

max\_depth = Max depth of each decision tree used

max\_features = Max number of features used to decide each split

subsample = Fraction of training data to be randomly samples for each tree.

learning\_rate = Controls weight of new trees added to the model based on errors

colsample\_bynode = subsample ratio of columns for each split

|  |  |
| --- | --- |
| n\_estimators | Range(100,800) . Interval = 50 |
| max\_depth | Range(1,150). Interval = 1 |
| max\_features | Range(2,15). Interval = 1 |
| subsample | Range(0.001,0.99). Number of points = 50 |
| learning\_rate | Range(0.001,1). Number of points = 30 |
| colsample\_bynode | Range(0, 1). Interval = 0.05 |

* 1. **Results**

1. *Best Parameters*

|  |  |
| --- | --- |
| n\_estimators | 620 |
| max\_depth | 69 |
| max\_features | 9 |
| subsample | 0.119 |
| learning\_rate | 0.09 |

1. *Errors*

|  |  |
| --- | --- |
| **Training - Randomized Search CV (Iterations=3000)** | **Testing - Randomized Search CV (Iterations=3000)** |
| RMSE = **23,100 MW** | RMSE = **170,000** **MW** |
| **0.48%** | **3.47 %** |

1. **Best Model Found**

For best results, the best model out of all was remodelled and ran at 10,000 iteration(3.3x original iterations). The results are shown below.

|  |
| --- |
| **Randomized Search CV (Iterations=10,000)** |
| RMSE = |
|  |

|  |  |
| --- | --- |
| n\_estimators | 775 |
| max\_depth | 147 |
| max\_features | 14 |
| subsample | 0.28 |
| learning\_rate | 0.069 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **RMSE Training (MW)** | **RMSE/Mean**  **Training (%)** | **RMSE Testing (MW)** | **RMSE/Mean**  **Testing**  **(%)** |
| *SARIMAX* | 97,842 | 2.04 | 195,663 | 3.99 |
| *RFR* | 76,179 | 1.59 | 171,575 | 3.5 |
| *Gradient Boost* | 99,304 | 2.07 | 154,831 | 3.16 |
| *XG Boost* | 23,096 | 0.48 | 169,981 | 3.47 |
| *Gradient Boost – 10,000 runs* | 46 | 0 | 146,529 | 2.99 |

1. **Limitations**

* The training set only contained data for 3 years. That is only 3 seasonal cycles. Given more data, the results will improve.
* Computational power is not strong to tune more model parameters as well as a wider range of parameter range.