Springboard --- DSC

Capstone Project 2

Predicting the Electric Energy Output of a Combined Cycle Power Plant

By Jefferson Fernandez

April 2021

1. Introduction

The goal if this project is to develop and evaluate regression models to predict the Electric Energy Output of a Combined Cycle Power Plant The machine learning model developed in this project is of interest especially to the Resource Department of the Electric Company that is responsible for balancing the load in their grid. If they could predict the energy output of their powerplants then they could potentially plan for demand ahead of time and foresee and avoid a possible shortfall in electricity supply.

The best XGboost model was then selected as the best model that is accurate to within -+ 2 MW of the actual demand. The link below will direct the reader to all the notebooks related to the project.

https://github.com/petrojeff/Springboard/tree/master/Capstone%20Project%202

2. Approach

2.1 Data Acquisition and Wrangling

The data from this project was collected from the UCI Machine Learning repository which is available in CSV format. This data set contains 9568 rows of data points collected from a Combined Cycle Power Plant over a period of 6 years. There are 4 features available to predict the net hourly electrical energy output. The features are average ambient variables.

- Temperature (T) in the range 1.81°C and 37.11°C,
- Ambient Pressure (AP) in the range 992.89-1033.30 millibar,
- Relative Humidity (RH) in the range 25.56% to 100.16%
- Exhaust Vacuum (V) in the range 25.36-81.56 cm Hg
- Net hourly electrical energy output (EP) 420.26-495.76 MW

	AT	V	AP	RH	PE
0	14.96	41.76	1024.07	73.17	463.26
1	25.18	62.96	1020.04	59.08	444.37
2	5.11	39.40	1012.16	92.14	488.56
3	20.86	57.32	1010.24	76.64	446.48
4	10.82	37.50	1009.23	96.62	473.90

There were no missing values in the dataset and each feature came in the right format (float64) as a result there was no need to clean the data. Statistically describing the value of each column of data it is apparent that the Combined Cycle Power Plant has a mean output of 454 MW. This will be the prediction of the base model.

The Pearson correlation was then employed to find out whether there are any correlated features. On the other hand, from the correlation heatmap there is a strong inverse correlation between electric energy output and the Atmospheric Pressure. This is as is expected since engines are much more efficient the colder the intake air is (lower air temperature results in higher air density which allows for more efficient fuel burn). There are no negative values in dataset which would have represented physically impossible measurements and thus there was no need to drop any row. Different regression algorithms were then used to model electric energy output.

2.2 Storytelling and Inferential Statistics

Below are some of the figures and statistics that shows different relationships between the features and electric energy output, and the relationship among the features.

	AT	V	AP	RH	PE
count	9568.000000	9568.000000	9568.000000	9568.000000	9568.000000
mean	19.651231	54.305804	1013.259078	73.308978	454.365009
std	7.452473	12.707893	5.938784	14.600269	17.066995
min	1.810000	25.360000	992.890000	25.560000	420.260000
25%	13.510000	41.740000	1009.100000	63.327500	439.750000
50%	20.345000	52.080000	1012.940000	74.975000	451.550000
75%	25.720000	66.540000	1017.260000	84.830000	468.430000
max	37.110000	81.560000	1033.300000	100.160000	495.760000

Fig.1. The statistics shows that the mean output of the power plant is 454 MW.

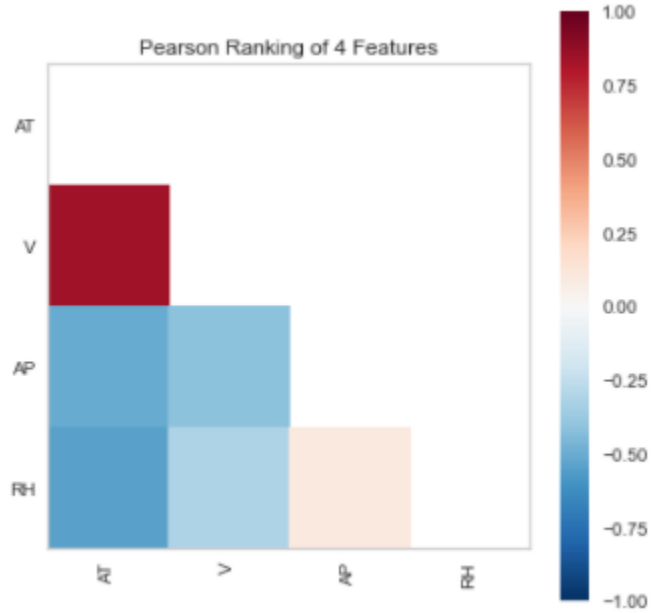


Fig.2. The Pearson Ranking shows the correlation between the different features and as can be inferred from the figure, the exhaust vacuum pressure and temperature are highly correlated.

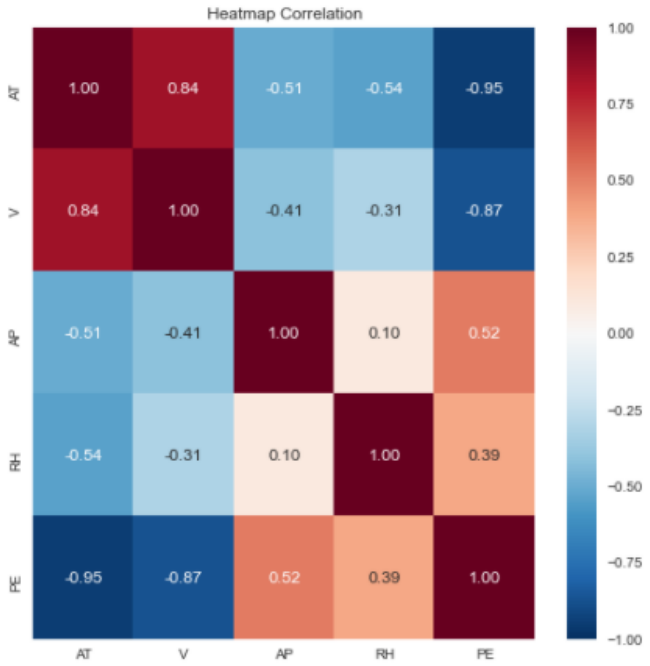


Fig.3. The heatmap correlation shows the relationship between the features and Energy Output which shows an inverse relationship between Atmospheric Temperature and Energy Output.

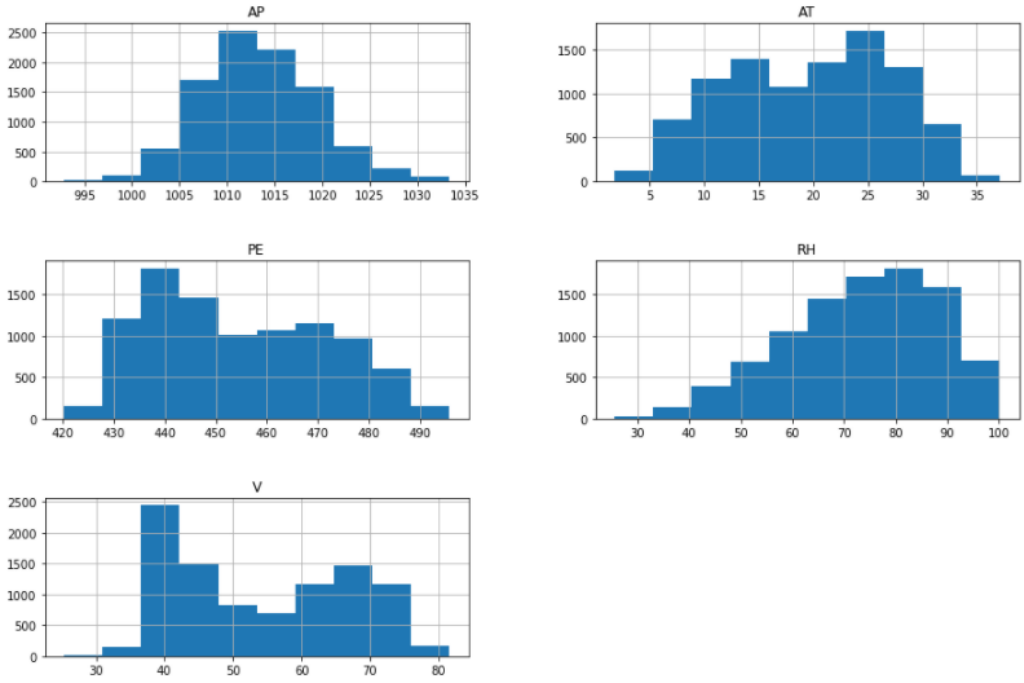


Fig.4. Histogram of each column of data that shows the distribution for each figure.

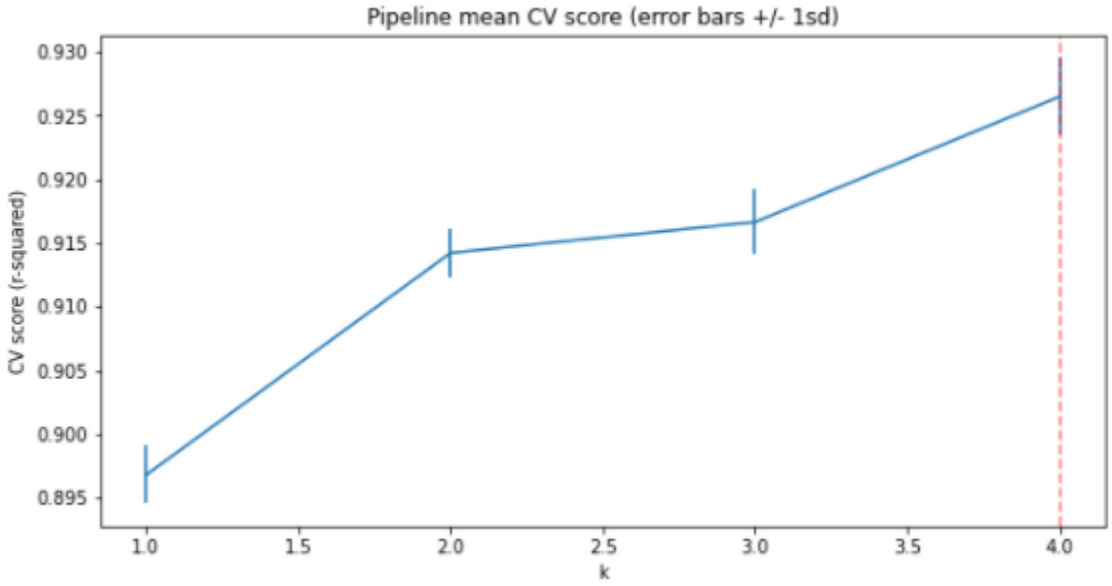
2.3 Baseline Modeling

The baseline model that was used for this project was the dummy regressor model which outputs the average value for every input. Using the test set the dummy regressor results had a Mean Absolute Error 14.7 MW. Clearly this value can be significantly improved with much more sophisticated algorithms.

2.4 Extended Modelling

6 algorithms were employed in this project these are Linear Regression, Random Forest Model, AdaBoost Model, GradientBoost Model, Multi-layer Perceptron and XGBoost. These models were chosen for the following reasons. 1. Linear Regression is easy to train and thus provides the easiest way to compare our baseline model with. 2. Random Forest is a good off the shelf decision tree-based models. 3. Multi-Layer Perceptron is a good model to see whether neural networks are a good option. 4. AdaBoost and XGBoost are Models that are gradient boosted that are very efficient that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler and weaker models.

Via cross-validation and modelling different combinations of hyperparameters via the grid search method over 30+ models where created an analyzed. The best-case model (highest cross-validation score) for each algorithm was then selected and compared with one another. As can be seen I the figure below cross-validation score increases as we add more features. This figure pertains to the linear regression model.



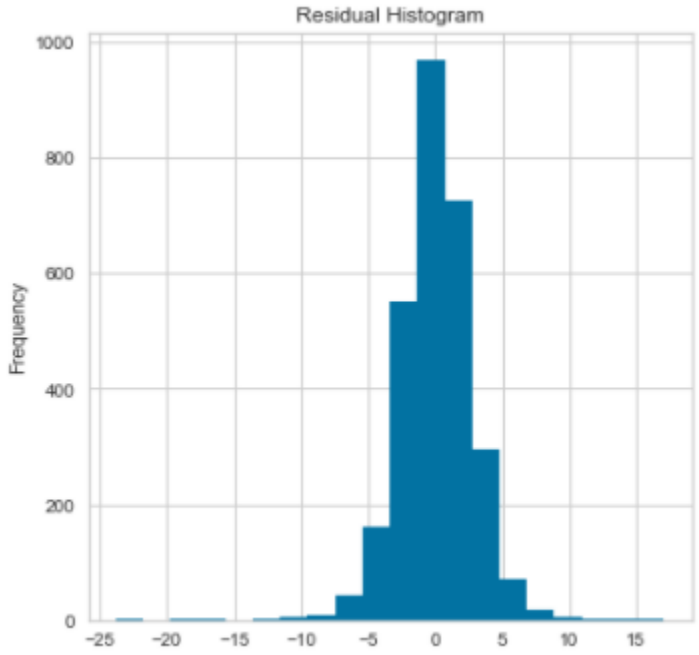
3. Findings

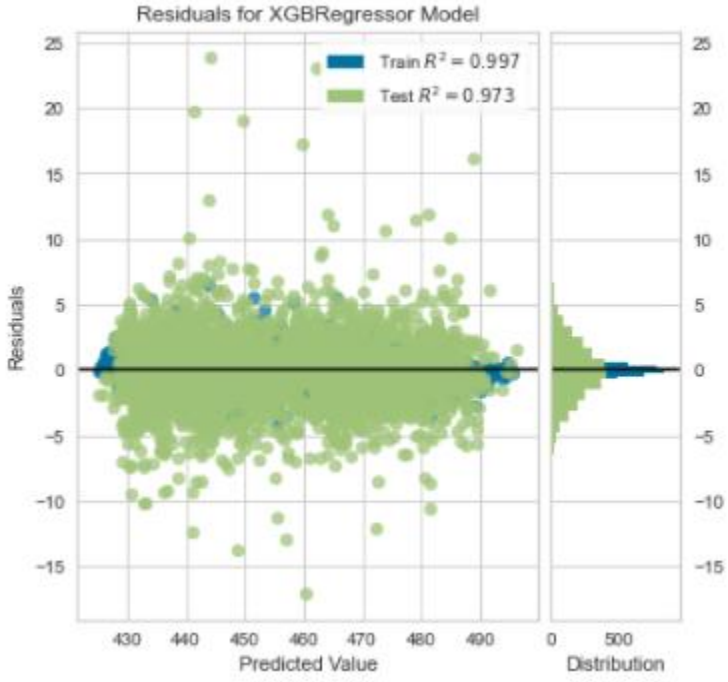
	Model	Average Residuals	Standard Deviation of Residuals	R Squared	Adjusted R Squared	Mean Absolute Error
0	Best Linear Regression Model	0.030635	4.449466	0.933215	0.933175	3.587932
1	Best Random Forest Model	0.019321	0.019321	0.964996	0.964975	2.354754
2	Best AdaBoost Model	0.455474	0.455474	0.923163	0.923117	3.793316
3	Best GradientBoost Model	0.021790	0.021790	0.970713	0.970696	2.131491
4	Best Multi-layer Perceptron Model	0.042563	0.042563	0.948247	0.948216	3.050394
5	Best XGBoost Model	0.064444	0.064444	0.973104	0.973087	2.034630

Using different metrics, a comparison can be made of the performance of each model. The average residuals metric is simply the average of the residuals from the test set. This metric cancels out the negative residuals and thus does not give us the full picture of how good each prediction is. The Standard Deviation of Residuals shows the spread of the residuals. The R Squared or Coefficient of Determination helps us compared the model with a constant baseline and tells us how much our model is better. The Adjusted R squared is an improvement of the R squared metric and takes care of the problem of the score improving as we add more terms even though the model itself is not improving.

The best model as can be seen is the XGBoost Model. It scored the highest in every single metric. With a Mean Absolute Score of 2 compared to our base model's MAE of 14.7 that is a reduction in error by 85%. This is a significant because each Mega Watt can approximately provide power for 800 houses. To check whether collecting more data will further improve the performance of our model a learning curve was built. The learning curve plots the training and cross-validation score as the model is created with more samples. As can be seen from the figure below collecting more data will not significantly increase the cross-validation score as it is already at 96%.

The figures below show that the model is somewhat normal. This is an important aspect of the model as it shows that our model exhibits homoscedasticity. This implies that the variance is the same for all values of targets regardless of the input and that our model is unbiased. To test this, we used the Kolmogorov-Smirnov test. The p-value for our model is below <0.05 and thus our model does not exhibit homoscedasticity. This can be seen from the figures below where there is a small number of residuals around -25.





4. Conclusions and Future Work

In conclusion we created a machine learning model that is very accurate for this Combined Cycle Powerplant. What we have shown is that if we have data available for every single powerplant that the company operates we can potentially predict the electric supply capacity of the company and therefore predict in advance whether there will be a surplus or deficit of electricity within the grid. This is critical for any electric company as this allow the company to buy electricity at a lower rate instead of buying electricity in the spot market which is significantly more expensive.

5. Recommendations for Clients

In this section we would like to recommend to the client the following.

- Deploy the best XGboost Machine Learning that way the resource department could utilize it in predicting the electric output of the powerplant.
- Start the conversation about building a similar model for every single power plant owned by the company.
- Since the data collected comes from the combined cycle power plant at full load, the powerplant maintenance department could potentially use this model to figure out whether the plant is operating in an optimal and predict the need for maintenance (i.e., The model predicted an output of 500 MW but the actual output is only 300 MW.).

6. Consulted Resources

The following resources for this project.

- Springboard Materials
- Machine Learning Pocket Reference by Matt Harrison
- Hands-On Machine Leaning with Scikit-Learn Keras and TensorFlow by Geron
- Practical Statistics for Data Scientist by Bruce and Bruce