Predicting the Energy Output of a Combined Cycle Power Plant

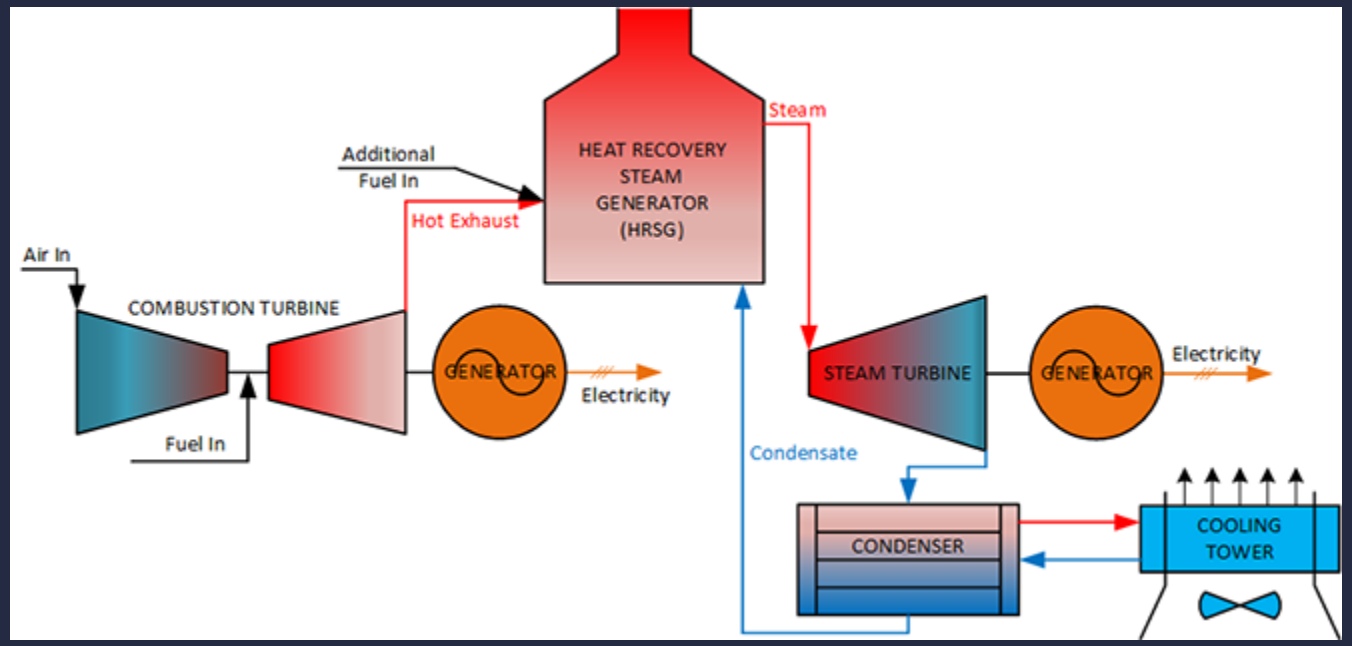
By:

Jefferson Fernandez

1. Introduction

A combined-cycle power plant uses both a gas and a steam turbine together to produce up to 50% more electricity from the same fuel than a traditional simple-cycle plant

Source: GE



2. Problem Definition

How do we predict the energy out of a combined cycle power plant at full power given for features, which are Ambient Temperature, Ambient Pressure, Relative Humidity, and Exhaust Vacuum?

3. Data

CSV file from UCI Machine Learning Repository

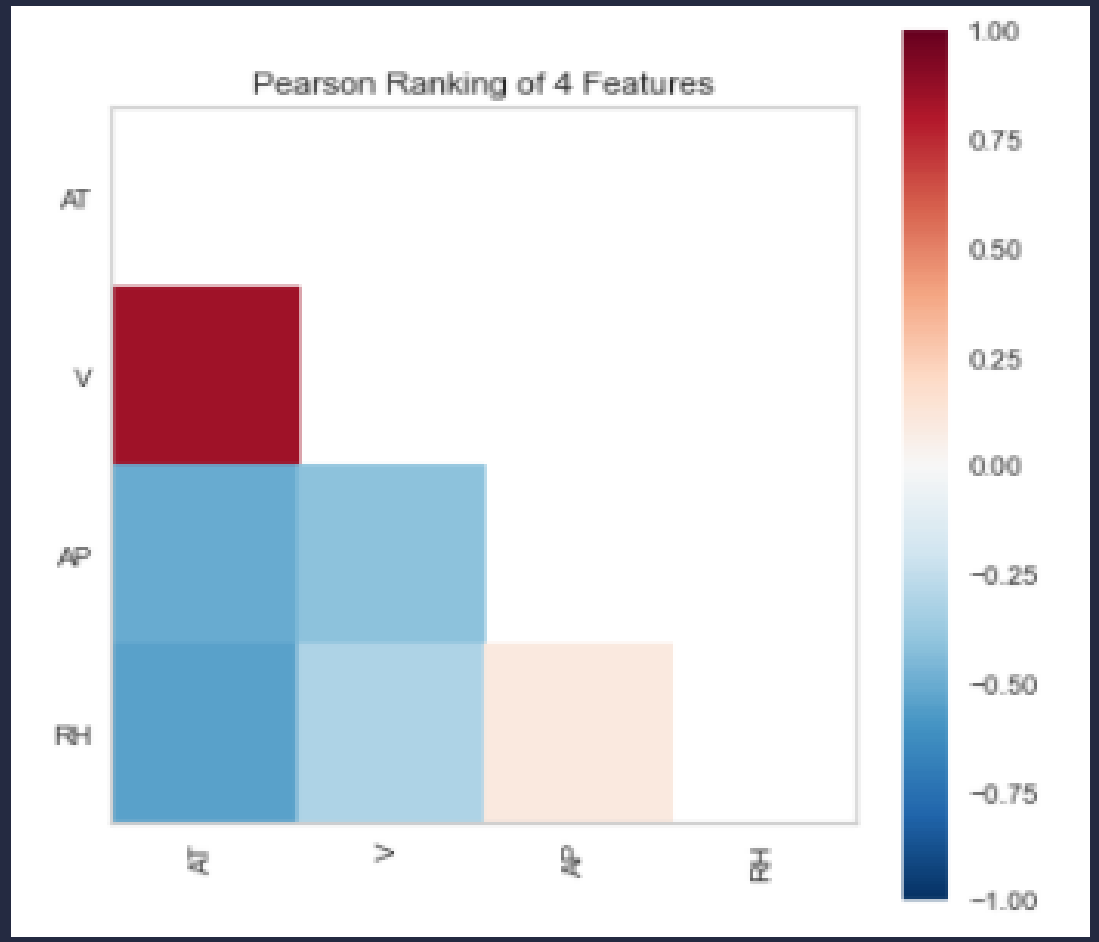
- 9568 rows
- Collected over a period of 6 years
- Temperature (T) in the range 1.81 C and 37.11 C
- Ambient Pressure (AP) in the range 992.89 1033.3 mmilibar,
- Relative Humidity (RH) in the range 25.56% to 100.16%
- Exhaust Vacuum (V) in the range 25.36-81.56 cm Hg
- Net hourly electrical energy output (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

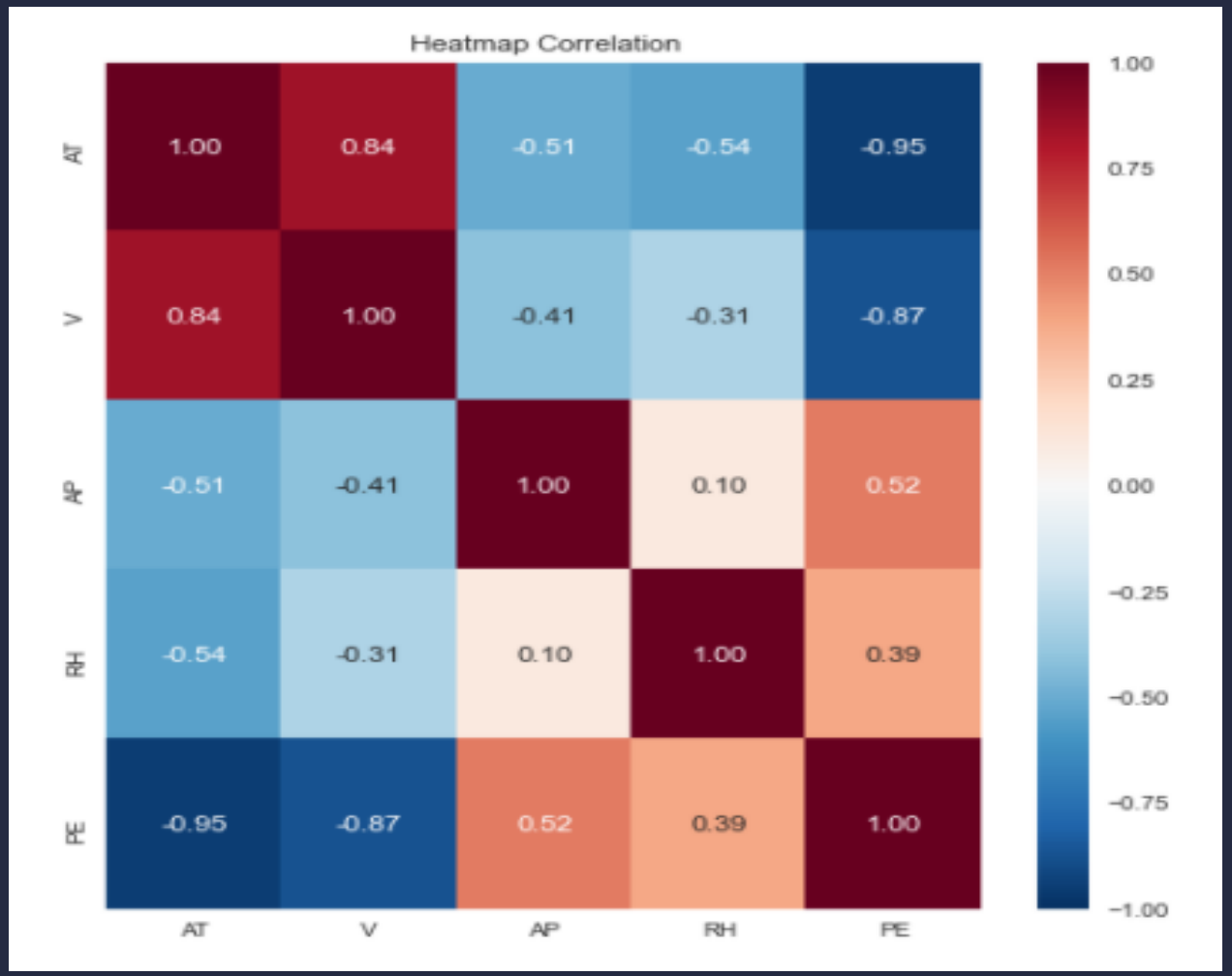
4. Data Wrangling and Exploratory Data Analysis

- No missing value in the dataset
- No negative value in the dataset
- The mean output of the power plant at full load is at 454 MW
- Data does not need to be prepared as it already comes at the right format

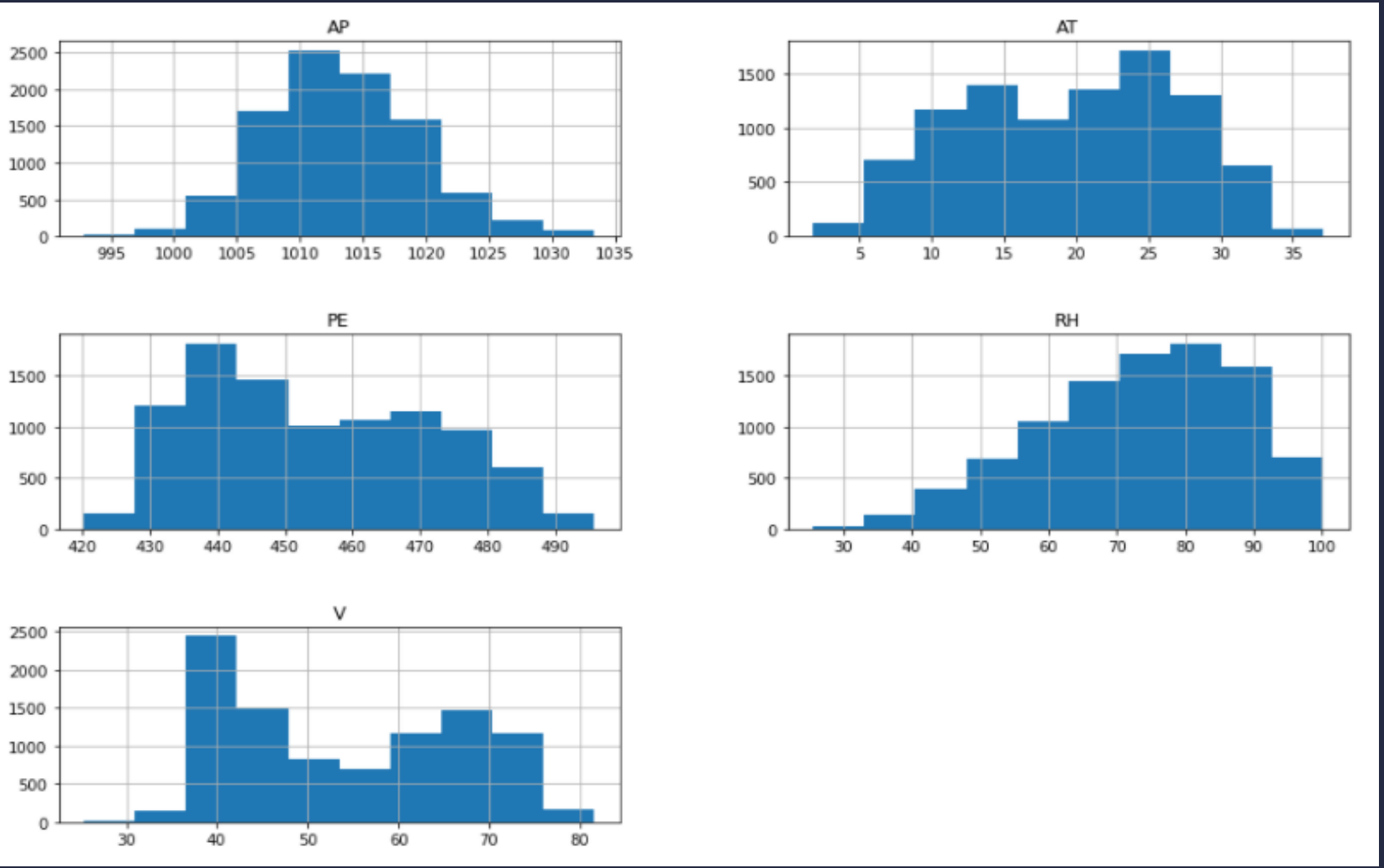
	AT	V	AP	RH	PE
count	9568.000000	9568.000000	9568.000000	9568.000000	9568.000000
mean	19.651231	54.305804	1013.259078	73.308978	454.365009
std	7.452473	12.707893	5.938784	14.600269	17.066995
min	1.810000	25.360000	992.890000	25.560000	420.260000
25%	13.510000	41.740000	1009.100000	63.327500	439.750000
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



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.

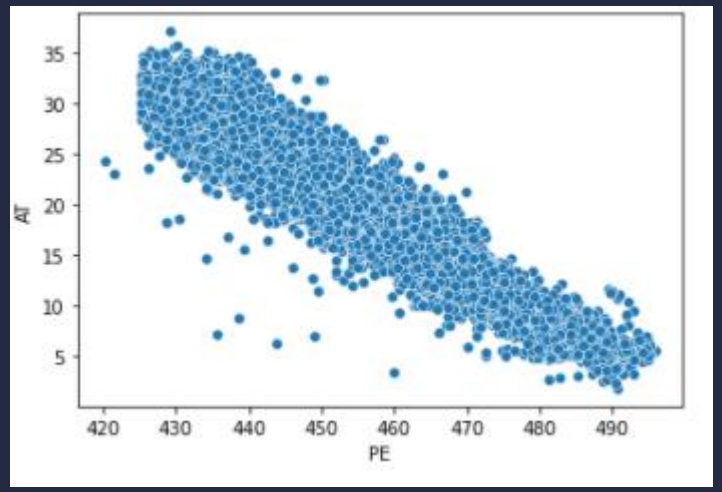


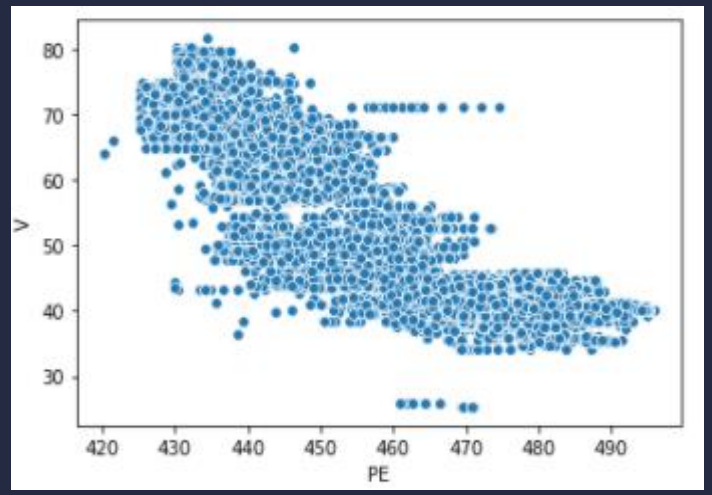
The heatmap correlation shows the relationship between the features and Energy Output which shows an inverse relationship between Atmospheric Temperature and Energy Output.

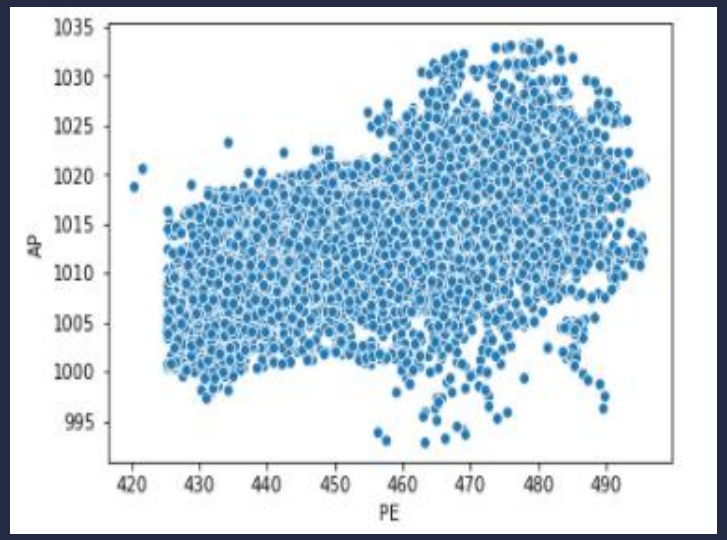


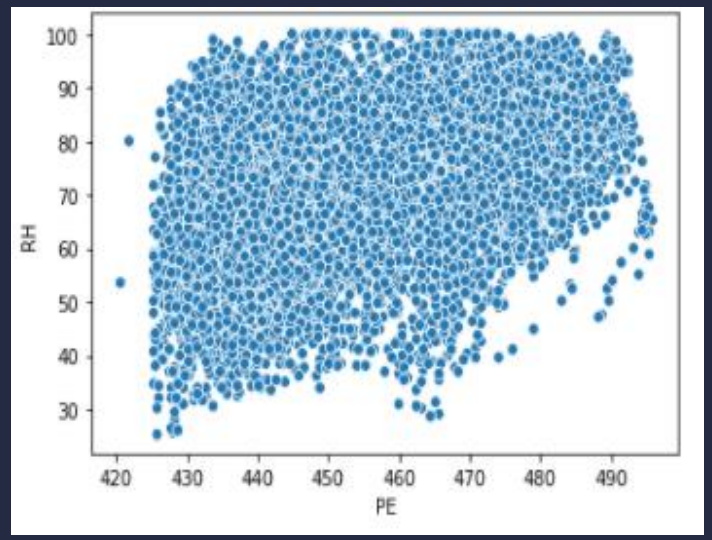
Histogram of each column of data that shows the distribution for each figure.

- Plotting each variable against the Electric Energy Output (PE) we can see an inverse relationship between Atmospheric Temperature and PE.









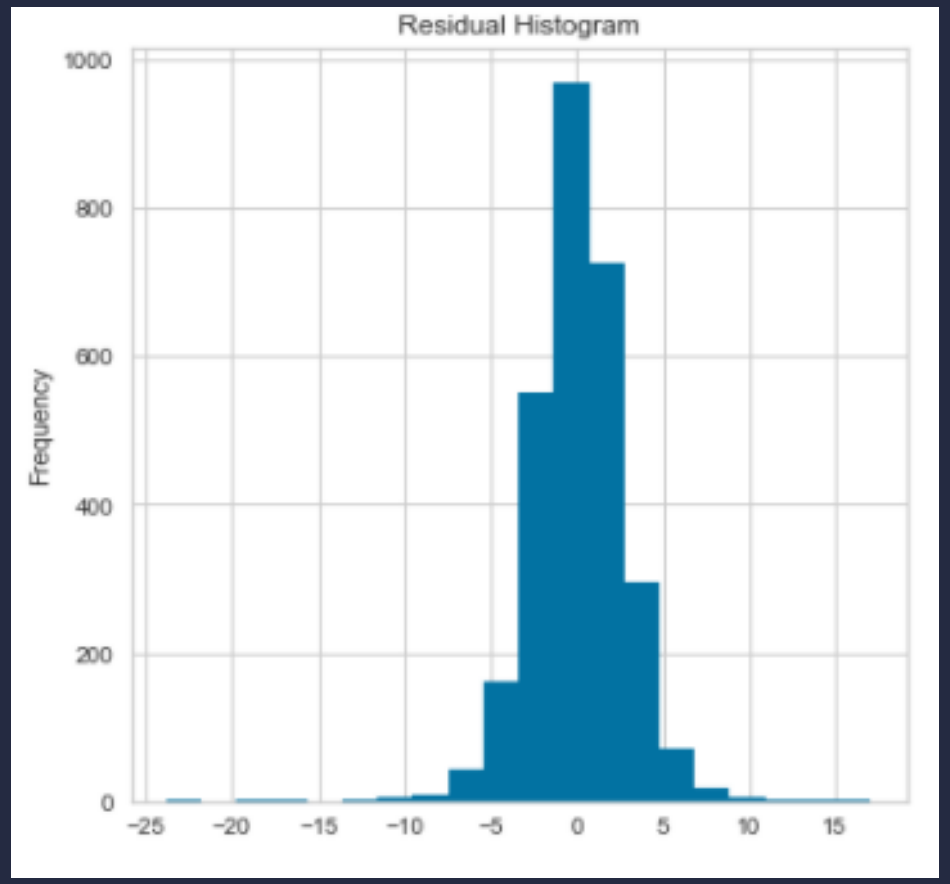
5. Preprocessing and Training

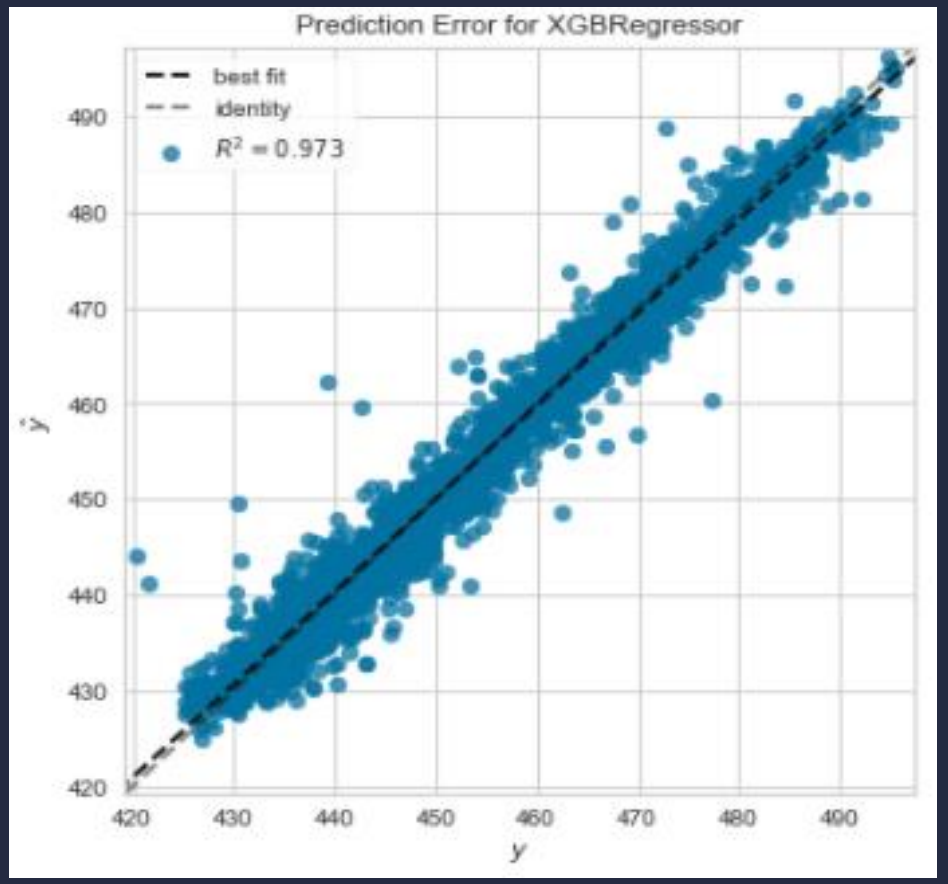
- 6 different regression machine learning algorithms were chosen, and bench marked against our base model (mean output).
- These models are linear regression, random forest, AdaBoost, GradientBoost, Multi-layer Perceptron, XGBoost.
- 70% of data set assigned as learning set and the remaining set assigned as test set.
- The mean absolute error of the base model is 15 MW. This is significant error because a megawatt can typically power 800 households and in our case under predicting or over predicting power for 12,000 households.
- Cross-Validation was used for each model to avoid overfitting.
- Hyperparameter tuning was employed for each model in order to find the best case for each machine learning model.
- Over 30+ models were examined and the best model for each algorithm was picked.

6. Modelling

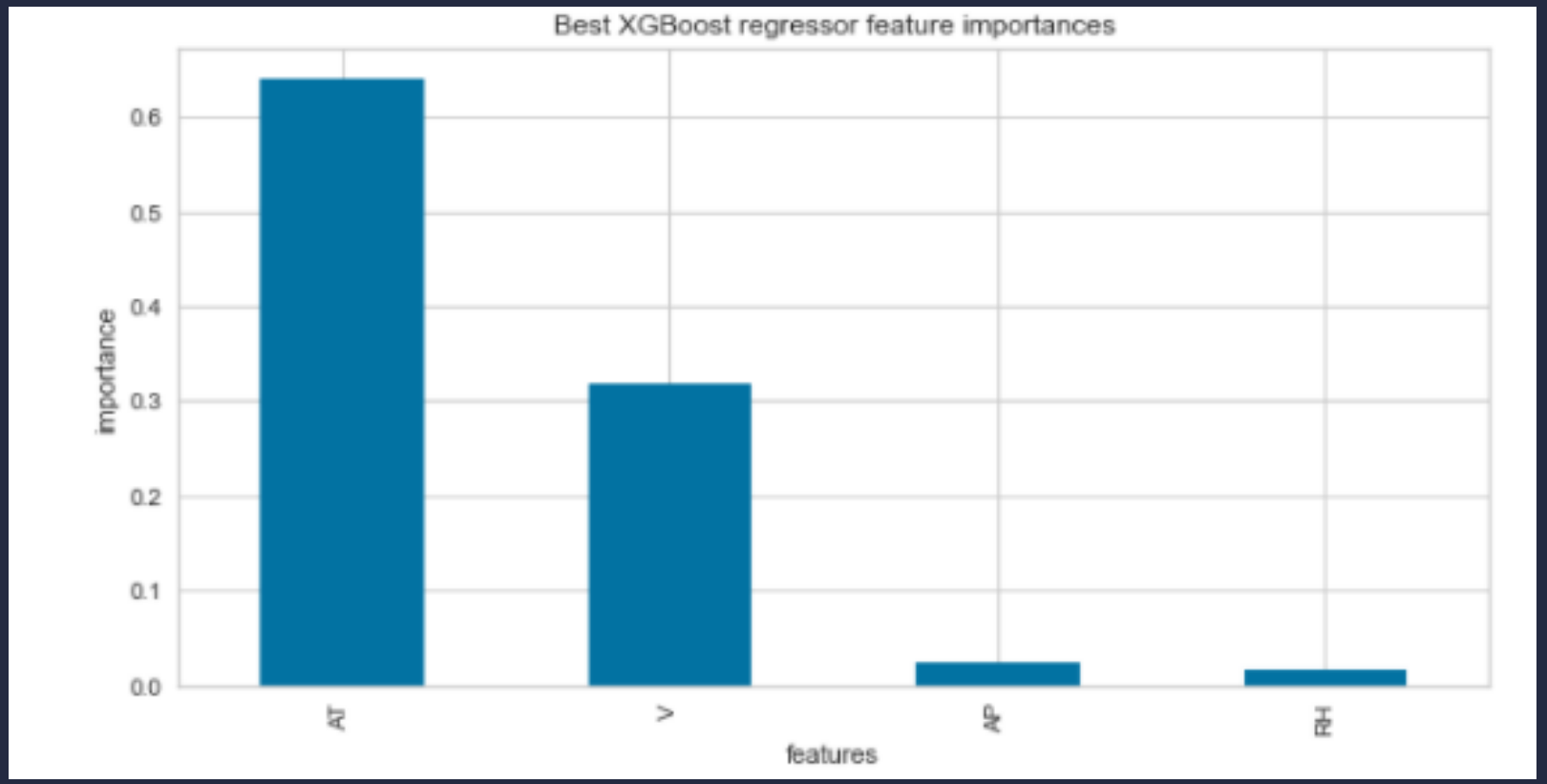
	Model	Average Residuals	Standard Deviation of Residuals	R Squared	Adjusted R Squared	Mean Absolute Error
0	Best Linear Regression Model	0.030635	4.449466	0.933215	0.933175	3.587932
1	Best Random Forest Model	0.019321	0.019321	0.964996	0.964975	2.354754
2	Best AdaBoost Model	0.455474	0.455474	0.923163	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

- XGBoost Model is the best performing model with the lowest Mean Absolute Error, and highest Adjusted R Squared value.

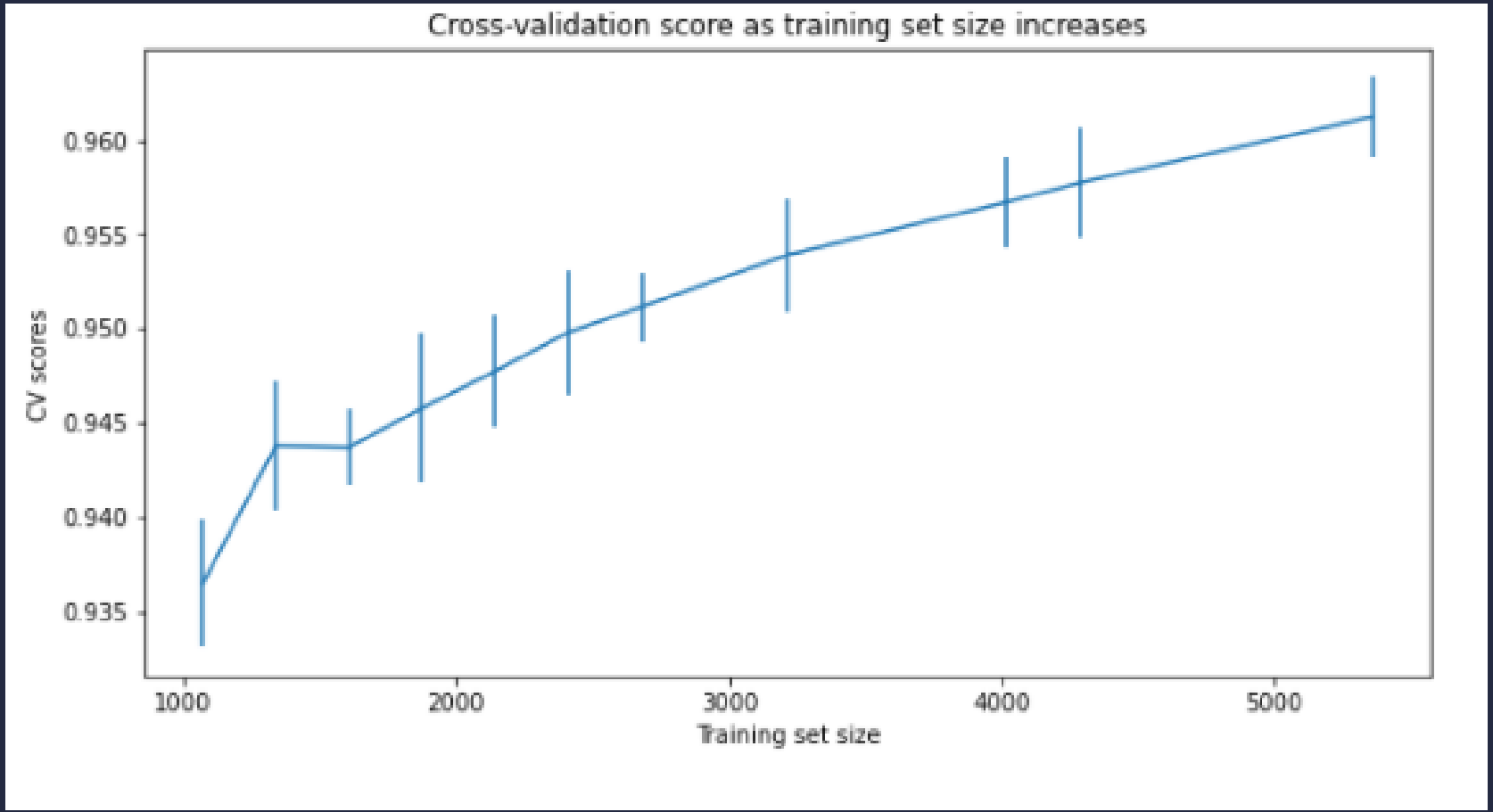




- The histogram of the XGBoost Model residuals from the test set shows that it looks normal, with a prediction error value between -7 and 7.



- As expected, the most important feature is the atmospheric temperature is the most important feature as expected. Colder temperature is correlated with the performance of a turbine engine since colder air is denser thus leading to a higher air fuel ratio and increased turbine efficiency or power output.



- 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%.

7. Conclusion and Recommendations

The XGBoost Machine Learning Model is significantly more accurate than simply using the average value of the electric energy output. The model is 85% more accurate compared to the baseline model. Below are our Recommendations.

- 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.).

8. Consulted Resources

- 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