# Output of Power Plant Energy

Hedyeh Erfani

DSC 680

Abstract

When we think of the things that keep our country running, we might quickly start thinking of things such as money. While this is definitely important and essential, one thing that I feel is a driving force is power plants. Going to the core of what is needed for day-to-day operations, power plants help ensure that everything we have works. Without them, none of our infrastructure, technology, or way of life could exist. Power plants are also a driving force behind our agriculture (food) as well. So, for this project, I’ve decided to go ahead and take a look at creating a model that can accurately predict power plant output.



Data

The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant. These features were set to the following listed below:

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

These measurements were averages that were taken by sensors. The sensors picked up on the measurements every second.

Below is a demonstration of a power plant, so you can see the various components from which the features were measured.

Diagram

Description automatically generated

Walkthrough

The data was taken from the Excel file provided by the researchers who collected information regarding the power plants. I’ve decided to use Python to run my analysis. Here is the data from plant.head() to see a short summary of what the variables look like. After taking a glance at the data, I checked for any null variables, which did not exist in this dataset.

Table

Description automatically generated

From there I looked at a pairplot, created by the Seaborn library. This showed me various histograms and scatterplots that I used for my analysis. I was looking at the interactions between the variables.

Qr code

Description automatically generated

From there, I decided to take a look at Pearson Correlation Coefficient. Table

Description automatically generated

I noticed that there existed a high score between AT and PE. I decided to then perform linear regression to test the relationship between the two variables. After doing an 80/20 train test split, I ran a few more tests to assess the prediction. The square root of mean squared error between y test and prediction was 5.45108120605578. The r squared between y test and prediction was 0.8969276808302301.

I repeated that process but instead used AT, V, AP, RH against PE. I received similar results, with the square root of mean squared error being roughly 4.89 and the r squared being around 0.92.

From there I repeated that process but decided to test a Decision Tree Regressor instead of a Linear Regressor. The results were pretty similar. The square root of mean squared error was roughly 5.33 and the r squared was roughly around 0.90.

Conclusion

Between the Linear Regression and the Decision Tree Regression, there wasn’t too much of a difference between the two. Looking at the results, I will go ahead and conclude that the feature “AT”, referred to above as “T” had the greatest significance in power plant output.

Reference

https://archive.ics.uci.edu/ml/datasets/combined+cycle+power+plant