**Combined Cycle Power Plant Power**

Scott Turro

**Introduction**

A power plant looks to maximize their revenue by creating a model that predicts their energy output. The plant uses a system of turbines to generate electricity for their customers. If the plant can produce more energy in a given hour, they will be able to sell more of their product. With a good model, the plant knows what to fix in order to increase their energy output.

Turbines are the main source of the plant’s power. The system is composed of two gas turbines and one steam turbine. The gas turbines produce heat while running. A device attached to each gas turbine collects the heat and produces steam. The steam from both systems is then funneled into a single steam turbine to reduce waste.

Initial data over a six-year period is split into test data and training data. Three variables were measured on the gas turbines. Temperature (T) in Celsius, Ambient Pressure (AP) in units millibar, and Relative Humidity (RH) which is a percentage. Exhaust Vacuum (V) is collected for the steam turbine. Finally, the plant has collected the net hourly Energy generated by the plant (EP). This is the target variable to be maximized.

**Exploratory Data Analysis**:

First glance in the training data shows significant trends between EP and the predictors. The two plots in Figure A show good evidence that EP can to be predicted. The plot of EP vs T has a very linear shape and is similar to the graph of EP vs V. As you increase T or V, in both cases EP decreases.

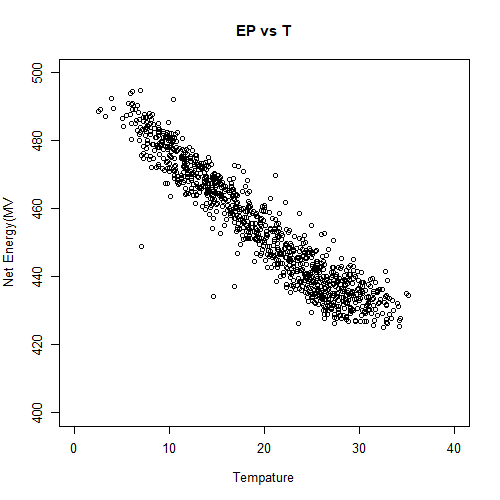
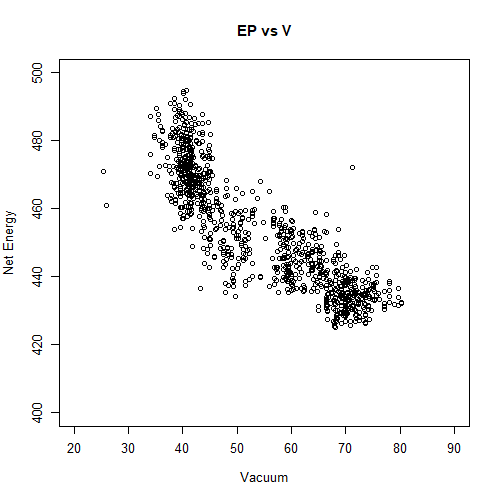


Figure A| Plotting the target variable against predictors.

Another observation from Figure A is that the Exhaust Vacuum data seems to be grouped instead of evenly spaced over the domain. This observation is present in any graph that has V as a variable. There is a small gap in the middle each of these plots which means the data was not sampled uniformly. This could be an indication that a non-linear relationship involving V should be used to describe the data.

While performing multivariate regression it is assumed the variables are independent. In other words, the data should be checked to ensure none of the variables do not have trends among another. All of the variables are continuous so the correlation coefficient can be calculated between any pair. A high correlation is defined to be above .8. Dependent variables tend to have a high correlation coefficient.

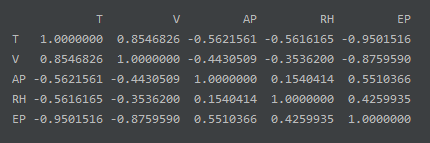


Figure B. Correlation Table

Matching with what the graphs showed, EP and T have the highest correlation among the values. The second highest correlation is the correlation between V and EP. It’s important to note that V and T are also highly correlated, which means these two variables may not be independent. The model should pay attention to these variables T being correlated. Regression should be run on their sum. This treats the two variables as a single number and attempts to reach the assumption of independence.

**Simple Model**

The simplest model has the following mathematical structure. The coefficients of the model are found using regression on the training data. In specific the function lm in R was used to run the regression and perform all other calculations presented in this paper. The predicted parameters are shown in Figure C.

**EP = β₀ + β₁ \* (T + V) + β₂ \* AP + β₃ \* RH**

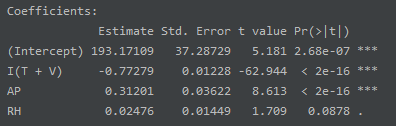
****

Figure C| Coefficients of full simple Linear Model

The rightmost column shows how significant each parameter was in the model. A low value is desirable. The value for RH is relatively high which says there is less statistical evidence that the variable has an effect on EP. Terms are typically dropped if they exceed a limit of .05. In this case the probability is magnitudes higher than the rest. The contribution is dropped for this model. Re-running this regression without including RH gives a more accurate model.

**EP = β₀ + β₁ \* (T + V) + β₂ \* AP**

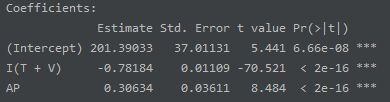
****

Figure D | Coefficients of reduced simple Linear Model

The final simple model has highly significant parameters which is a indication of a good model. Another assumption of linear regression is that the residuals are randomly distributed centered around zero. A good check for this assumption is to plot the residuals vs the predictors and look for any signs of trends.

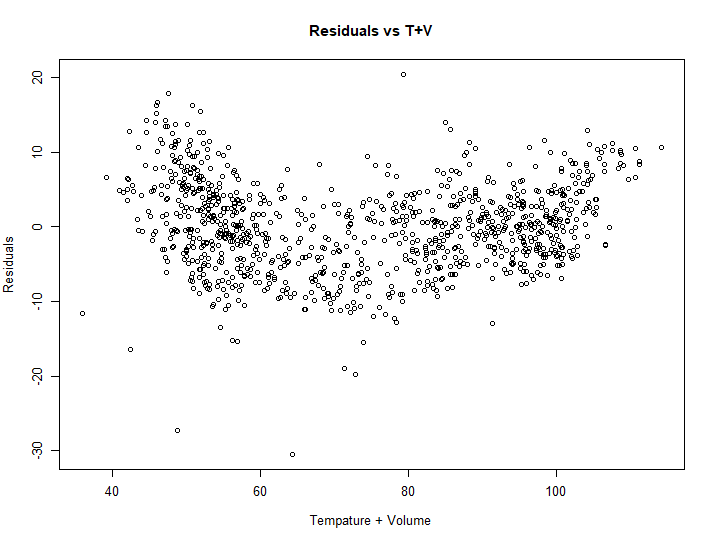


Figure E | Residuals of reduced simple model plotted against T+V

A significant trend can be seen between the residuals and T+V. In specific the residuals seem to be higher on the sides and lower in the middle, giving a parabolic shape. It is not safe to assume that the residuals were independently distributed with constant variance.

With this in mind we can still see how well this model

By only performing regression on a subspace of the total data, there is plenty of left-over data to test the model on. There are many measures of how well a model fits the test data. The Root mean square deviation is one such value. A lower value means the model fits the data better. The value for the simple reduced model is 5.927. Even though there are patterns in the residuals graph. The model predicted the testing data well. This shows promising signs that the model is accurate. There is room for improvement but a good starting score.

**Quadratic Model**

The new model will consider the squared value of T+V in attempts to fix the curved pattern in the residual graph. Initially RH is considered as a factor in this new model. As in the previous model, the term is dropped due to less degree of significance. The final mathematical equation for the quadratic model is formed, regression is again performed, and estimates are obtained in Figure F.

**EP = β₀ + β₁ \* (T + V) + β₂ \* (T + V)² + β₃ \* AP**

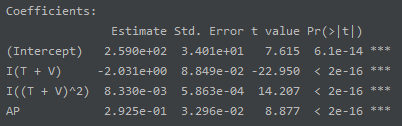
****

Figure F | Coefficients of quadratic Model

The quadratic term is very significant telling us it is worth including this term in the model. All terms have high significance and are all relatively the same. No further changes should be made to this model. Testing this new model against the testing data results in a better RMSE value. The value for the quadratic model is 5.370. By acknowledging the pattern in the residual graph, a better model was able to be produced.

**Categorical Model**

The third model will account for the grouped data present in the Vacuum graph. A new variable replaces V in the regression. It is defined as 0 if the vacuum value is above the mean vacuum level or 1 if it is less than the mean. The mean vacuum level is 54.072 and was determined based on the training data. The only interaction term that was not significant was the one between RH and V. This interaction term will suffice as acknowledging the correlation between those two variables. The other interaction terms were left out of the model.

**EP = β₀ + β₁ \* V + β₂ \* T + β₃ \* AP + β₄ \* RH + β₅ \* (T\*V) + β₆ \* (AP\*V)**

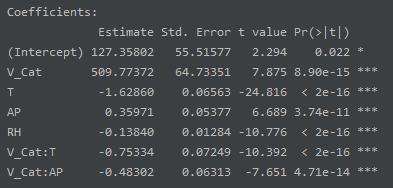
****

Figure G | Coefficients of Categorical Model

This model’s terms are all highly significant. The least significant term was the intercept β₀. It does not make sense to drop the intercept since it can be considered as the base energy produced when Vaccuum is high. The contribution of the intercept could be overshowed from the magnitude of β₁. In this model the RH term was significant enough to include it in the model. When predicting the testing data the mean Vacuum level is defined as the mean of the training data. This model outperformed the rest with a RMSD of 4.502723. The biggest advantage of this model is the ability to account for the Relative humidity in a significant way.

**Results and Conclusions:**

The training data was able to produce promising models to predict the power the plant produces over an hour. Modeling the testing data gave good results. The best model includes all variables and treated Vacuum as a categorical variable. The model did not consider the physical design of the system. A better model can possibly be developed by defining the power output from each turbine as a separate random variable and the sum as EP.

Even though the Temperature and Vacuum data came from different types of turbines, they seem to be correlated. When considering these as a single variable, Humidity was no longer significant. Grouping the Vacuum data into two groups allowed for the inclusion of Humidity and gave rise to the best model. This paper calls for the investigation into why temperature and Vacuum are so correlated. This question may be entangled with why Vacuum data is grouped in the first place.

* Good evidence that net hourly Energy (EP) can be well predicted.
* Temperature (T) and Exhaust Vacuum (V) seem to be correlated.
* The Exhaust Vacuum (V) data is not sampled uniformly.