Statistical Modeling for Business Analytics – MBA652A – Project 1

Multiple Linear Regression - Combined Cycle Power Plant

Submitted To:
Prof. (Dr.) Devlina Chatterjee

Submitted By: Group 5

1. Ashish Tiwari (21129004)
2. Jyoti Sharma (21129265)
3. Shiv Shakti Singh (21129024)
4. Shreeyash Nitin Malode (20214271)

Outline of the Presentation

- Introduction
- Objective
- Hypothesis & Data
- Descriptive Statistics
- Correlation & Co-efficient of correlation
- Multiple Linear Regression Modeling
- Interpretation of Results
- Inference & Conclusion

Main Reference -

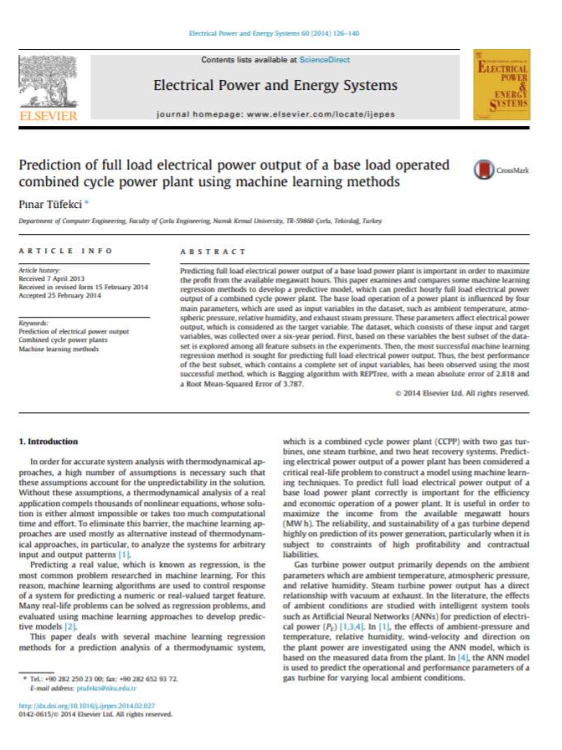
Tüfekci, Pınar. "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. " *International Journal of Electrical Power & Energy Systems* 60 (2014): 126-140.

Dataset Source -

UCI Machine Learning Repository

Software used -

R & Excel



Electrical Power and Energy Systems 60 (2014) 126–140

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods

Pınar Tüfekci *

Department of Computer Engineering, Faculty of Çorlu Engineering, Namık Kemal University, TR-59860 Çorlu, Tekirdağ, Turkey

ARTICLE INFO

Article history:
Received 7 April 2013
Received in revised form 15 February 2014
Accepted 25 February 2014

Keywords:
Prediction of electrical power output
Combined cycle power plants
Machine learning methods

ABSTRACT

Predicting full load electrical power output of a base load power plant is important in order to maximize the profit from the available megawatt hours. This paper examines and compares some machine learning regression methods to develop a predictive model, which can predict hourly full load electrical power output of a combined cycle power plant. The base load operation of a power plant is influenced by four main parameters, which are used as input variables in the dataset, such as ambient temperature, atmospheric pressure, relative humidity, and exhaust steam pressure. These parameters affect electrical power output, which is considered as the target variable. The dataset, which consists of these input and target variables, was collected over a six-year period. First, based on these variables the best subset of the dataset is explored among all feature subsets in the experiments. Then, the most successful machine learning regression method is sought for predicting full load electrical power output. Thus, the best performance of the best subset, which contains a complete set of input variables, has been observed using the most successful method, which is Bagging algorithm with REPTree, with a mean absolute error of 2.818 and a Root Mean-Squared Error of 3.787.

1. Introduction

In order for accurate system analysis with thermodynamical approaches, a high number of assumptions is necessary such that these assumptions account for the unpredictability in the solution. Without these assumptions, a thermodynamical analysis of a real application compels thousands of nonlinear equations, whose solution is either almost impossible or takes too much computational time and effort. To eliminate this barrier, the machine learning approaches are used mostly as alternative instead of thermodynamical approaches, in particular, to analyze the systems for arbitrary input and output patterns [1].

Predicting a real value, which is known as regression, is the most common problem researched in machine learning. For this reason, machine learning algorithms are used to control response of a system for predicting a numeric or real-valued target feature. Many real-life problems can be solved as regression problems, and evaluated using machine learning approaches to develop predictive models [2].

This paper deals with several machine learning regression methods for a prediction analysis of a thermodynamic system, which is a combined cycle power plant (CCPP) with two gas turbines, one steam turbine, and two heat recovery systems. Predicting electrical power output of a power plant has been considered a critical real-life problem to construct a model using machine learning techniques. To predict full load electrical power output of a base load power plant correctly is important for the efficiency and economic operation of a power plant. It is useful in order to maximize the income from the available megawatt hours (MW h). The reliability, and sustainability of a gas turbine depend highly on prediction of its power generation, particularly when it is subject to constraints of high profitability and contractual liabilities.

Gas turbine power output primarily depends on the ambient parameters which are ambient temperature, atmospheric pressure, and relative humidity. Steam turbine power output has a direct relationship with vacuum at exhaust. In the literature, the effects of ambient conditions are studied with intelligent system tools such as Artificial Neural Networks (ANNs) for prediction of electrical power (P_E) [1,3,4]. In [1], the effects of ambient-pressure and temperature, relative humidity, wind-velocity and direction on the plant power are investigated using the ANN model, which is based on the measured data from the plant. In [4], the ANN model is used to predict the operational and performance parameters of a gas turbine for varying local ambient conditions.

* Tel.: +90 282 250 23 00; fax: +90 282 652 93 72.
E-mail address: ptufekci@nku.edu.tr

http://dx.doi.org/10.1016/j.ijepes.2014.02.027
0142-0615/© 2014 Elsevier Ltd. All rights reserved.

UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

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

Introduction

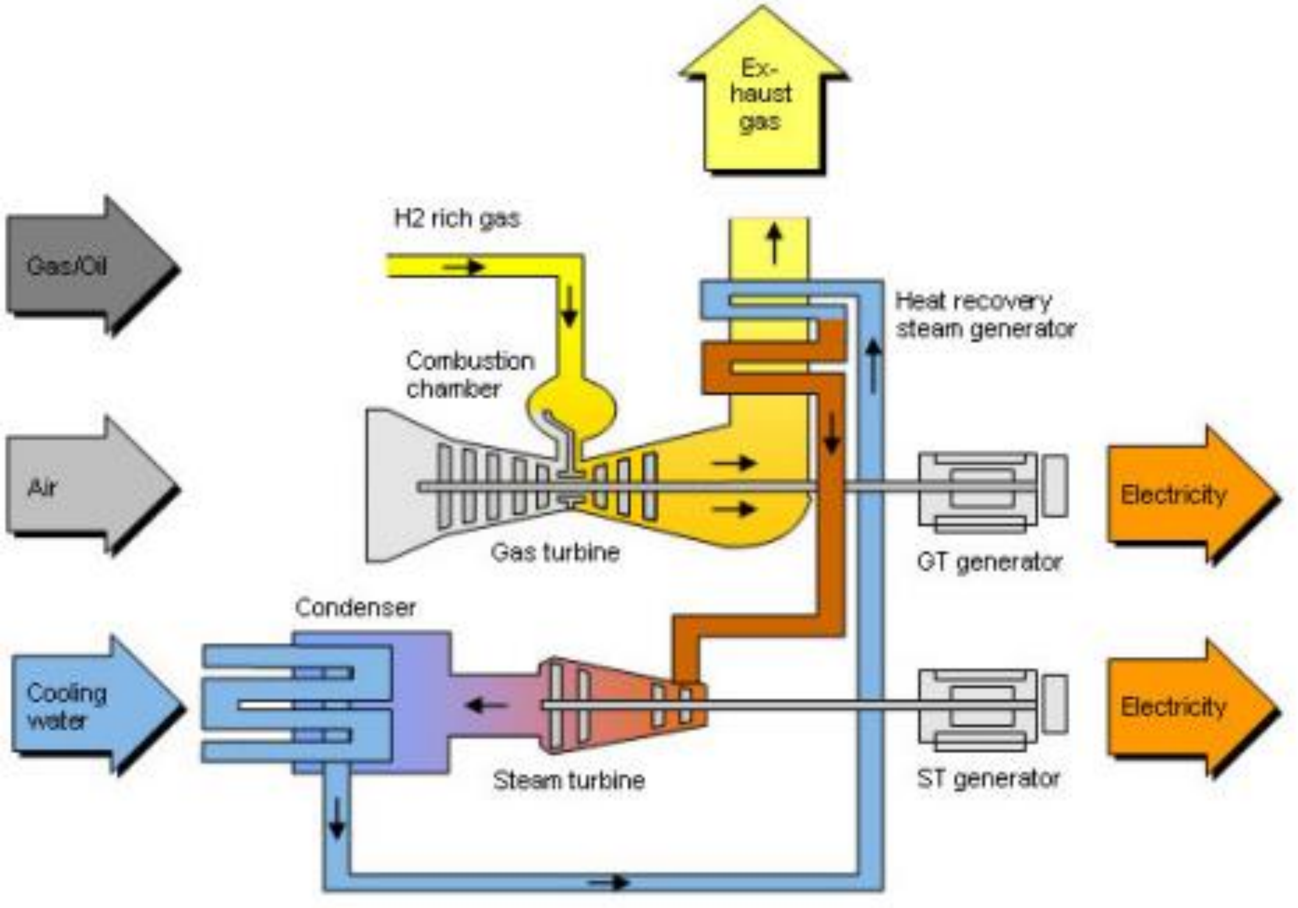


Figure 1. Schematic diagram of a Combined cycle gas power plant

- Gas turbines and steam turbines are used in conjunction to ensure maximum energy extraction on burning fossil fuels, viz. hazardous to the environment.
- Conventionally, thermodynamic approach is used to calculate the energy output with many assumptions.
- However, these assumptions account for unprecedented output of the system.
- To eliminate this barrier, we try to establish a regression model to predict the real values taking real variable account for energy output.

Image Source - http://www.zeroco2.no/capture/sources-of-co2/combined-cycle-power-plant/

Objective

- In this analysis, we are trying to understand the unexplained relationship between the ambient conditions of a power plant and the quantity of electrical energy it produces, with an objective to predict an optimal geographic location for a combined cycle power plant harnessing more output from same level of input.

- We have divided the above objective in following sub tasks-
  - To predict the performance parameter of a gas turbine for varying ambient parameters when it is running at full load capacity.
  - To develop a regression model to predict the output of a thermodynamic system.
  - To investigate which parameter or combination of parameter most affect the output performance of the system.

Hypothesis

Null Hypothesis(H_0): No significant relationship exists between output power (dependent variable) and ambient variables (independent variable).

Alternate Hypothesis(H_a): Significant linear relationship exists between output power (dependent variable) and ambient variables (independent variable).

Data Structure: Full load working Combined Cycle Power Plant data over 6 years (2006-2011) based in Turkey.

```
> str(CCPPDATA1)
'data.frame':   9568 obs. of  5 variables:
 $ AT: num  14.96 25.18 5.11 20.86 10.82 ...
 $ V : num  41.8 63 39.4 57.3 37.5 ...
 $ AP: num  1024 1020 1012 1010 1009 ...
 $ RH: num  73.2 59.1 92.1 76.6 96.6 ...
 $ PE: num  463 444 489 446 474 ...
```

AT: Ambient Temperature(^0C)

V: Exhaust steam pressure (Vacuum) (cm Hg)

AP: Ambient Pressure(mbar)

RH: Relative Humidity(%)

PE: Energy Output(MW)

Source – Computed; **Image Source –** R Output

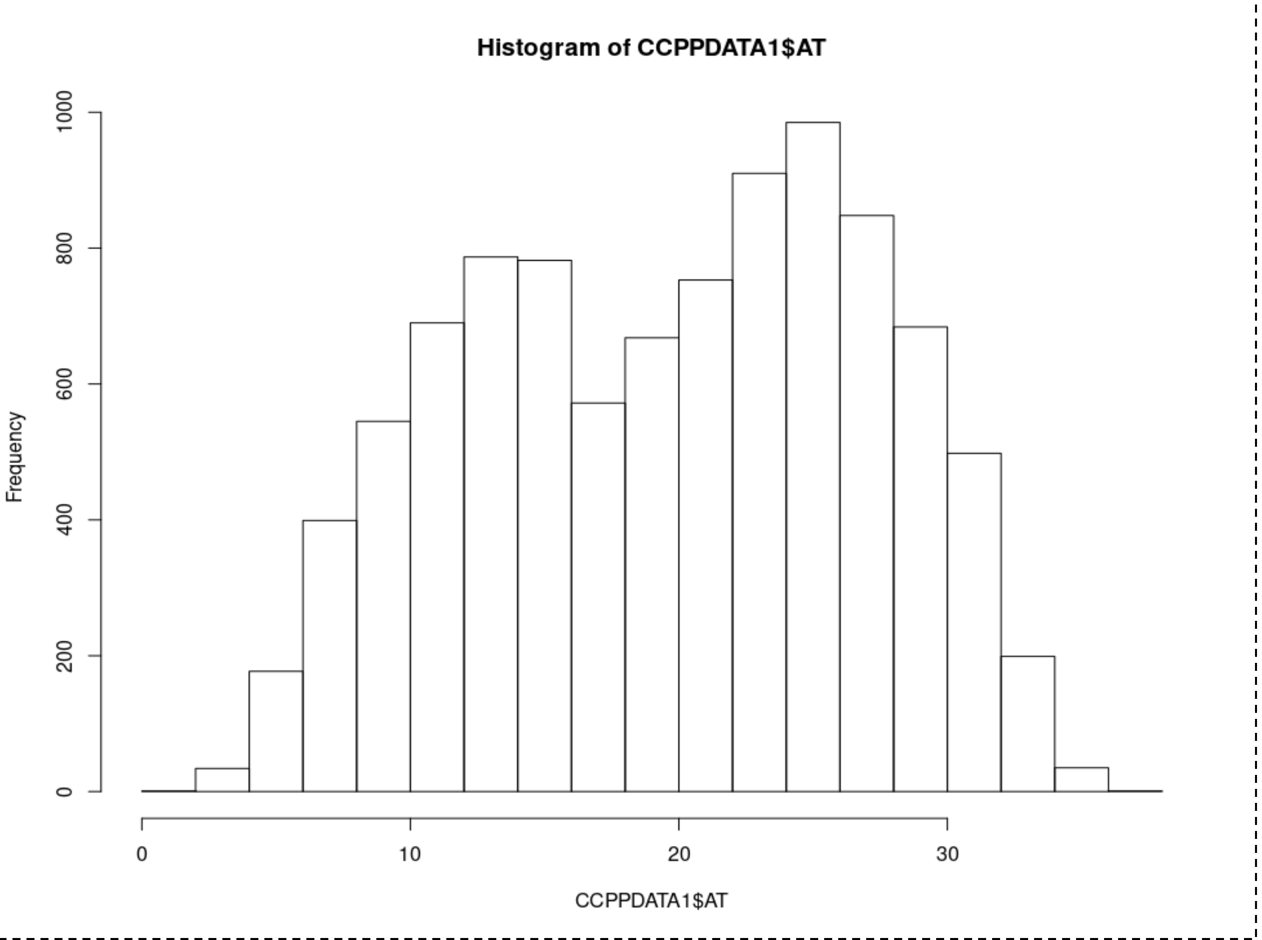
Data Snapshot

	AT	V	AP	RH	PE
Mean	19.65	54.31	1013.26	73.31	454.37
Standard Error	0.08	0.13	0.06	0.15	0.17
Mode	25.21	41.17	1013.88	100.09	468.80
Median	20.35	52.08	1012.94	74.98	451.55
First Quartile	13.51	41.74	1009.10	63.33	439.75
Third Quartile	25.72	66.54	1017.26	84.83	468.43
Variance	55.54	161.49	35.27	213.17	291.28
Standard Deviation	7.45	12.71	5.94	14.60	17.07
Skewness	-0.14	0.20	0.27	-0.43	0.31
Range	35.30	56.20	40.41	74.60	75.50
Minimum	1.81	25.36	992.89	25.56	420.26
Maximum	37.11	81.56	1033.30	100.16	495.76
Sum	188022.98	519597.93	9694862.86	701420.30	4347364.41
Count	9568.00	9568.00	9568.00	9568.00	9568.00

Source - Computed

Variable Exploration

(1) Ambient Temperature:

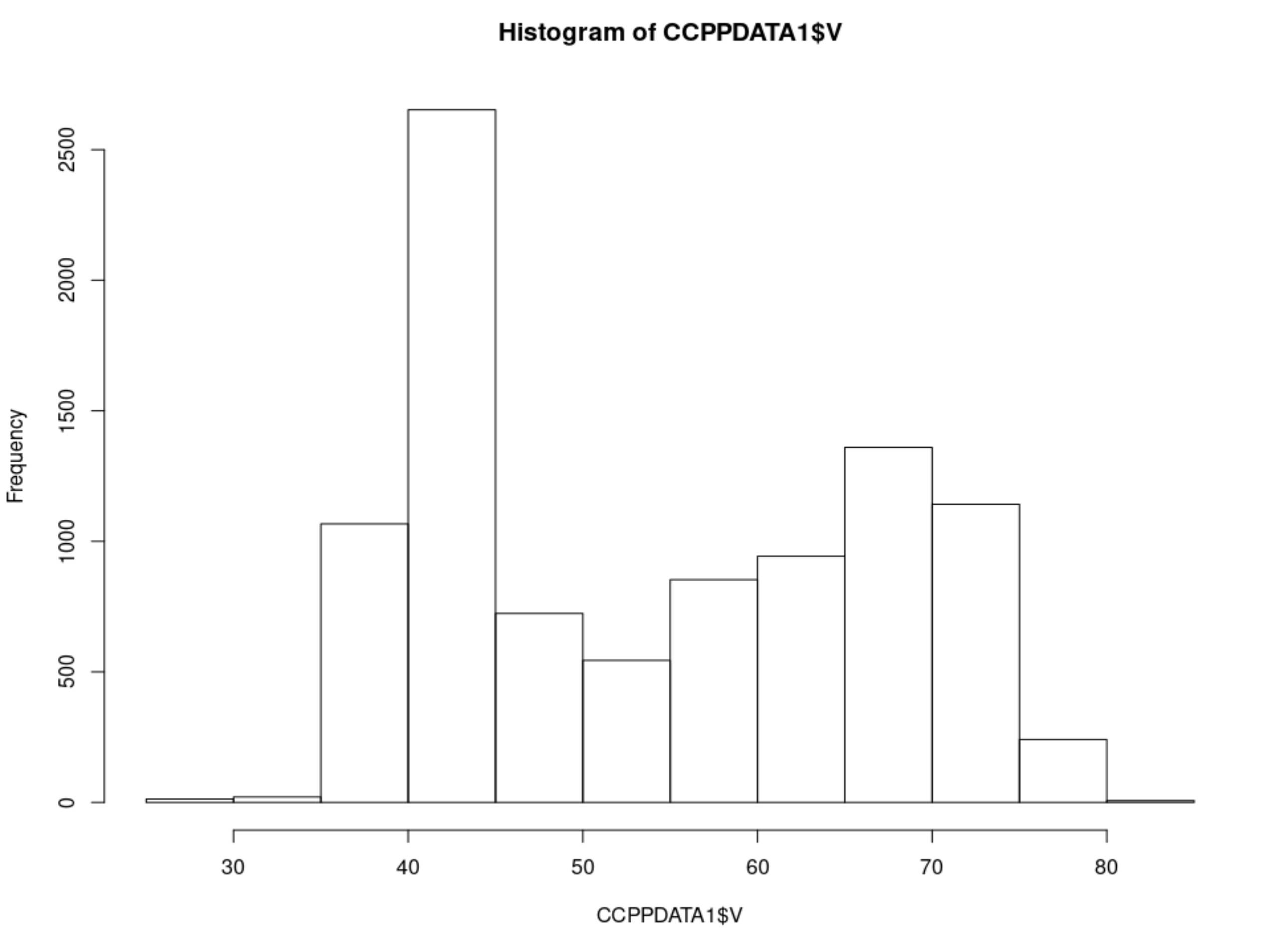


	AT
Mean	19.65
Standard Error	0.08
Mode	25.21
Median	20.35
First Quartile	13.51
Third Quartile	25.72
Variance	55.54
Standard Deviation	7.45
Skewness	-0.14
Range	35.30
Minimum	1.81
Maximum	37.11
Sum	188022.98
Count	9568.00

Source – Computed; **Image Source** – R Output

Variable Exploration (cond.)

2. Exhaust Vacuum:

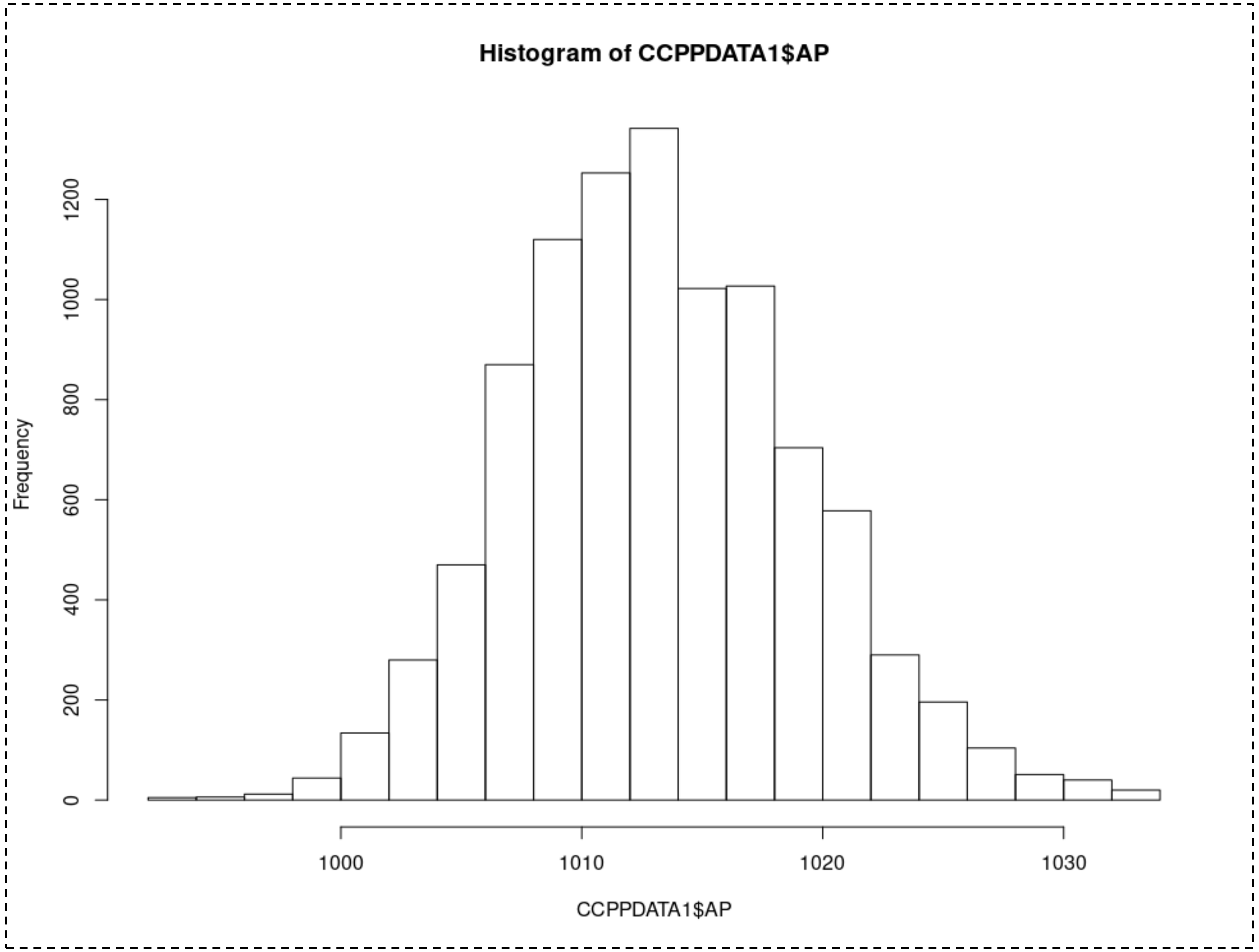


	V
Mean	54.31
Standard Error	0.13
Mode	70.32
Median	52.08
First Quartile	41.74
Third Quartile	66.54
Variance	161.49
Standard Deviation	12.71
Skewness	0.20
Range	56.20
Minimum	25.36
Maximum	81.56
Sum	519597.93
Count	9568.00

Source – Computed; **Image Source** – R Output

Variable Exploration (cond.)

3. Ambient Pressure:

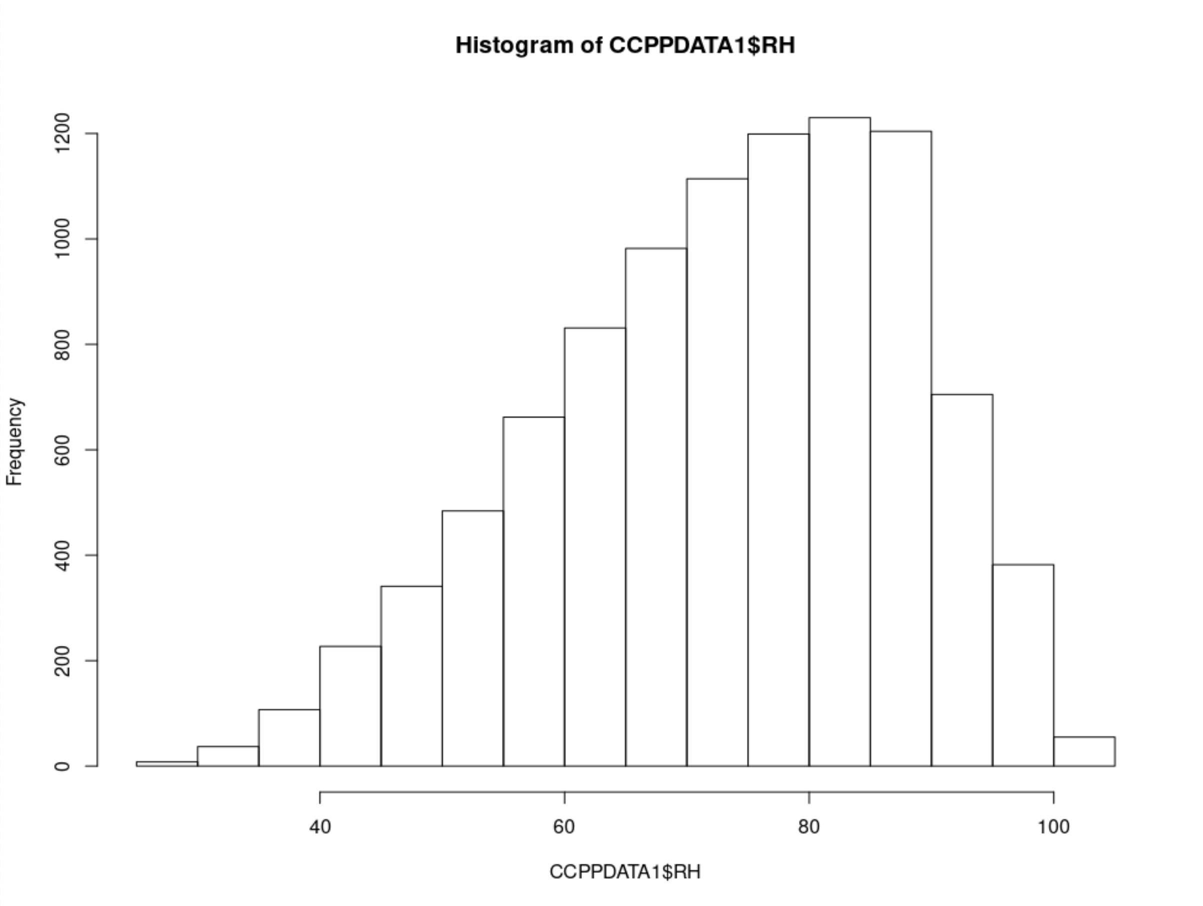


	AP
Mean	1013.26
Standard Error	0.06
Mode	1013.88
Median	1012.94
First Quartile	1009.10
Third Quartile	1017.26
Variance	35.27
Standard Deviation	5.94
Skewness	0.27
Range	40.41
Minimum	992.89
Maximum	1033.30
Sum	9694862.86
Count	9568.00

Source – Computed; **Image Source** – R Output

Variable Exploration (cond.)

4. Relative Humidity:

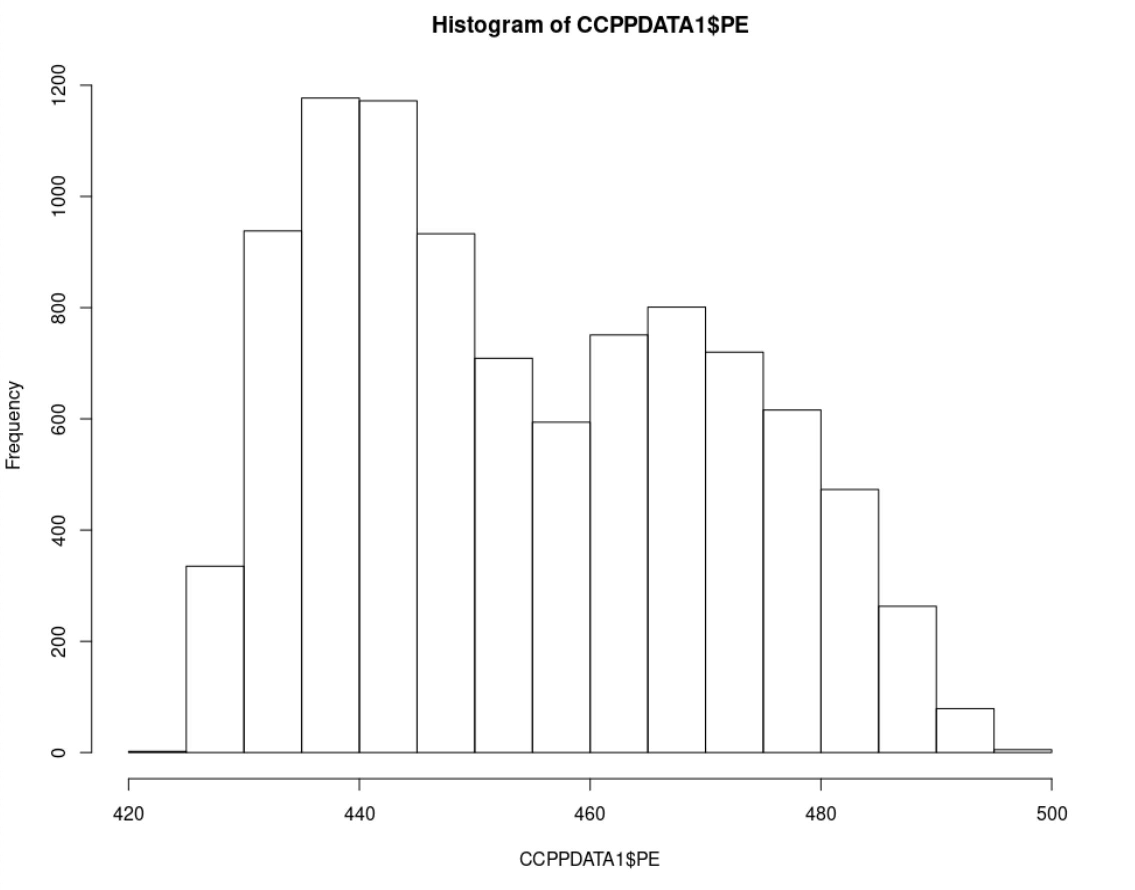


	RH
Mean	73.31
Standard Error	0.15
Mode	100.09
Median	74.98
First Quartile	63.33
Third Quartile	84.83
Variance	213.17
Standard Deviation	14.60
Skewness	-0.43
Range	74.60
Minimum	25.56
Maximum	100.16
Sum	701420.30
Count	9568.00

Source – Computed; **Image Source** – R Output

Variable Exploration (cond.)

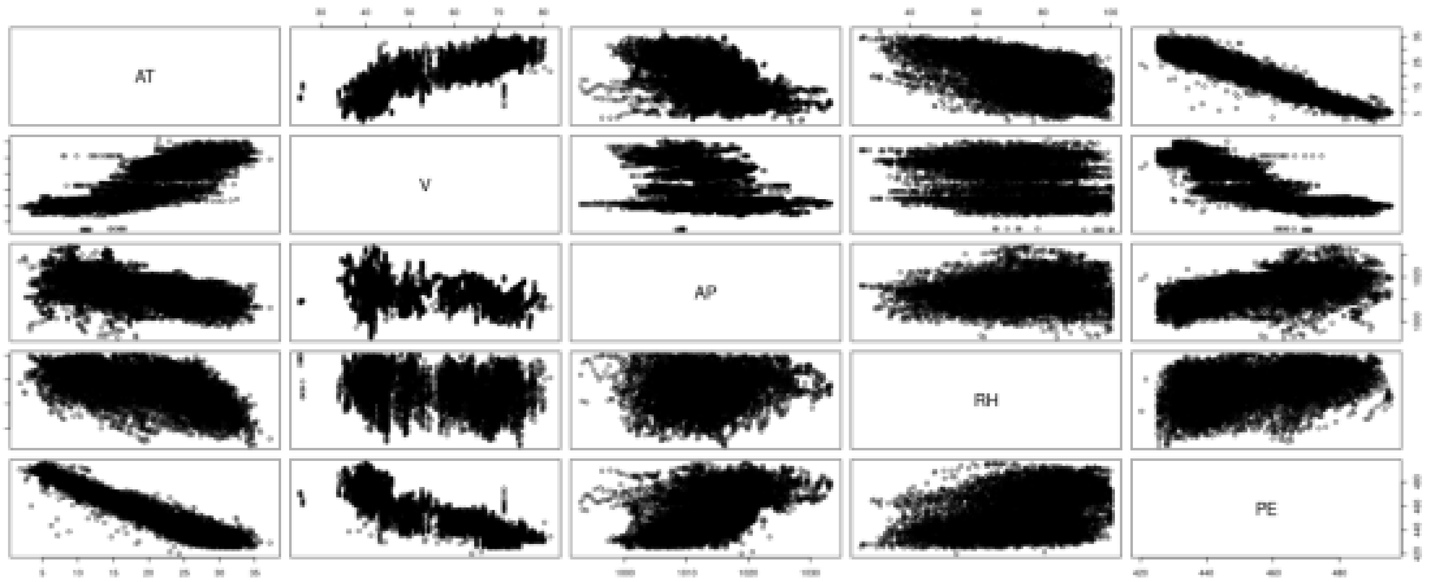
5. Energy Output:



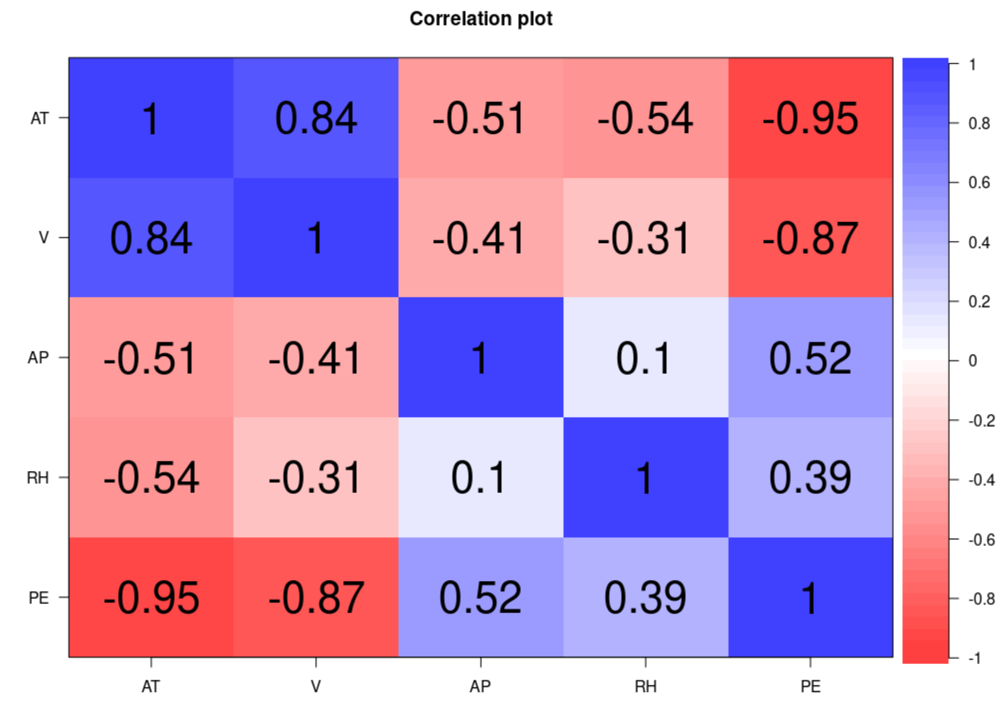
	PE
Mean	454.37
Standard Error	0.17
Mode	468.80
Median	451.55
First Quartile	439.75
Third Quartile	468.43
Variance	291.28
Standard Deviation	17.07
Skewness	0.31
Range	75.50
Minimum	420.26
Maximum	495.76
Sum	4347364.41
Count	9568.00

Source – Computed; **Image Source** – R Output

Correlation Matrix

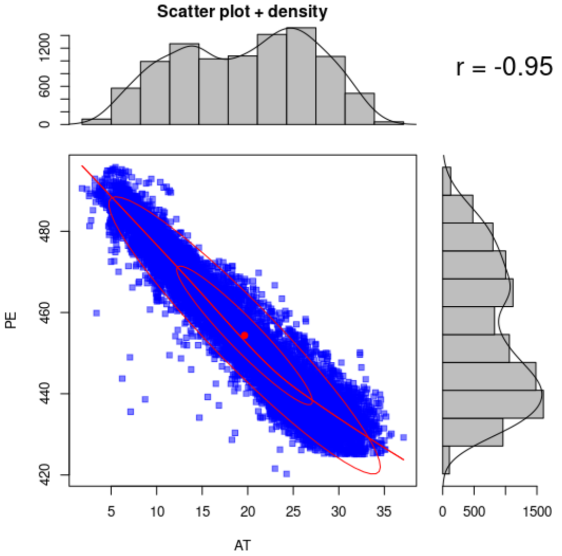


	AT	V	AP	RH	PE
AT	1.0000000	0.8441067	-0.50754934	-0.54253465	-0.9481285
V	0.8441067	1.0000000	-0.41350216	-0.31218728	-0.8697803
AP	-0.5075493	-0.4135022	1.00000000	0.09957432	0.5184290
RH	-0.5425347	-0.3121873	0.09957432	1.00000000	0.3897941
PE	-0.9481285	-0.8697803	0.51842903	0.38979410	1.0000000

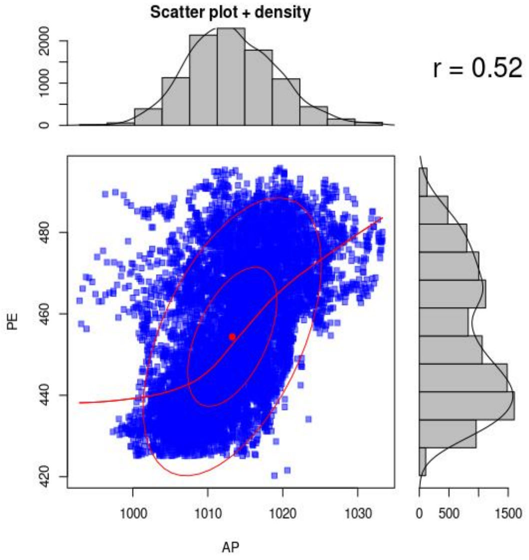


Source – Computed; **Image Source** – R Output

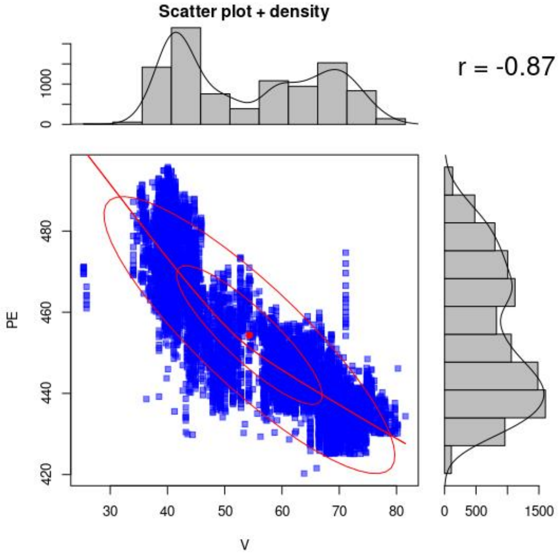
Scatter Plot



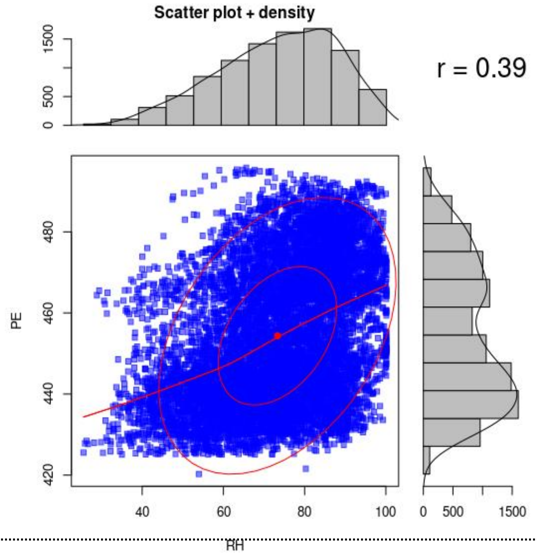
PE v/s AT



PE v/s AP



PE v/s V



PE v/s RH

Source – Computed; **Image Source** – R Output

Modelling & Interpretation of results

Formula (IV & DV)

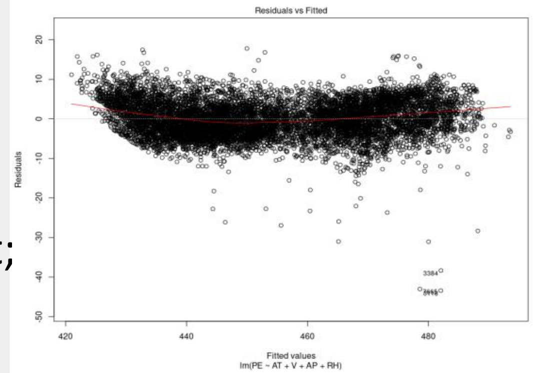
```
> model13 <- lm(formula = PE ~ AT + V + AP + RH, CCPPDATA1)
> summary(model13)

Call:
lm(formula = PE ~ AT + V + AP + RH, data = CCPPDATA1)

Residuals:
    Min      1Q  Median      3Q     Max 
-43.435  -3.166  -0.118   3.201  17.778 

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept) 454.609274   9.748512   46.634  < 2e-16 ***
AT           -1.977513   0.015289 -129.342  < 2e-16 ***
V            -0.233916   0.007282  -32.122  < 2e-16 ***
AP            0.062083   0.009458    6.564 5.51e-11 ***
RH           -0.158054   0.004168  -37.918  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.558 on 9563 degrees of freedom
Multiple R-squared:  0.9287,	Adjusted R-squared:  0.9287 
F-statistic: 3.114e+04 on 4 and 9563 DF,  p-value: < 2.2e-16
```



Difference between Predicted & Actual Values

Shows symmetrical or not; data above or below line

Betas

Standard deviation of coefficient;

Represent significance of coefficient

P<0.05

Coefficient not zero

Estimate/std. Error – how small or large is std error – should be large number

R^2 fraction of variance Y explained by X; Data points explains 92.87% of variation in energy output (dependent variable);

PE = 454 – 1.97*AT – 0.23*V + 0.06*AP – 0.15*RH

Source – Computed; **Image Source** – R Output; Emphasis added by researchers

Short listing of best model

Model	Variable Retained	R^2	Adjusted R^2
Model 1	Ambient Temperature	0.8989	0.8989
Model 2	Vacuum	0.7565	0.7565
Model 3	Atmospheric Pressure	0.2688	0.2687
Model 4	Relative Humidity	0.1519	0.1519
Model 5	**Ambient Temperature, Atmospheric Pressure**	**0.9008**	**0.9008**
Model 6	**Ambient Temperature, Relative Humidity**	**0.9209**	**0.9209**
Model 7	Vacuum, Atmospheric Pressure	0.7869	0.7869
Model 8	Vacuum, Relative Humidity	0.7720	0.7720
Model 9	Atmospheric Pressure, Relative Humidity	0.3843	0.3841
Model 10	**Ambient Temperature, Vacuum, Atmospheric Pressure**	**0.9180**	**0.9179**
Model 11	**Ambient Temperature, Vacuum, Relative Humidity**	**0.9284**	**0.9284**
Model 12	Vacuum, Atmospheric Pressure, Relative Humidity	0.8040	0.8039
Model 13	**Ambient Temperature, Vacuum, Atmospheric Pressure, Relative Humidity**	**0.9287**	**0.9287**
Model 14	**Ambient Temperature, Atmospheric Pressure, Relative Humidity**	**0.9210**	**0.9210**

Source - Computed

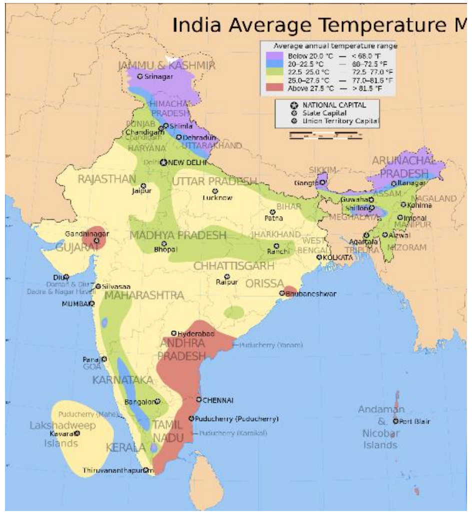
Multicollinearity Test for Independent Variable (VIF Factor)

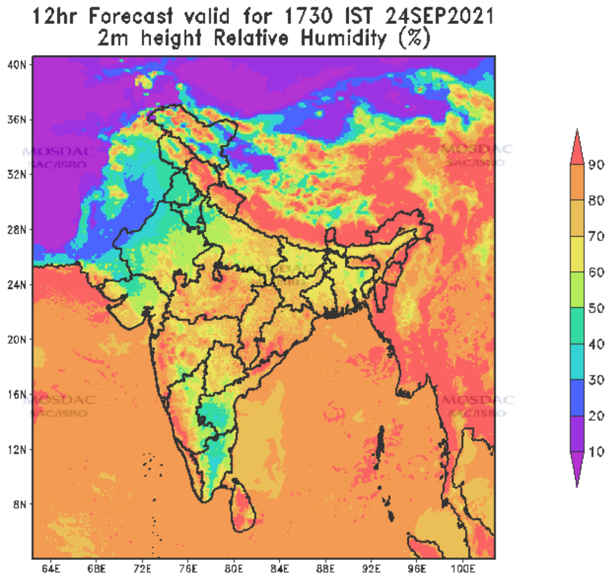
	Model 5	Model 6	Model 10	Model 11	Model 13	Model 14
Ambient Temperature (AT)	1.34	1.41	3.88	4.96	5.97	2.00
Vacuum (V)			3.48	3.88	3.94	
Atmospheric Pressure (AP)	1.34		1.34		1.45	1.43
Relative Humidity (RH)		1.41		1.58	1.70	1.50
Multicollinearity	**Not exist**	**Not exist**	**Not exist**	**Exist**	**Exist**	**Not Exist**

Source – Computed and Tabulated

Conclusion

- Independent variable taken into consideration does affect energy output.
- Correlation coefficients shows that strong correlation exists between temperature and vacuum, leading to multicollinearity while modelling.
- Possible omitted variable – Exhaust Vacuum
- Model 14, which consist ambient temperature, relative humidity and atmospheric pressure is best suited for the prediction of energy output.
- Based on available Average Temperature, Pressure and Relative Humidity Data we can shortlist candidate geographical locations for setting up these power plants.





Source – Meteonorm website

References

[1] Pınar Tüfekci, Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods, International Journal of Electrical Power & Energy Systems, Volume 60, September 2014, Pages 126-140, ISSN 0142-0615.

[2] Heysem Kaya, Pınar Tüfekci , Sadık Fikret Gürgen: Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine, Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE 2012, pp. 13-18 (Mar. 2012, Dubai).

[3] UCL Machine Learning Repository

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

[4] Lantz, Brett. *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.

[5] https://medium.com/analytics-vidhya/prediction-of-the-output-power-of-a-combined-cycle-power-plant-using-machine-learning-a2ca01848eea

[6] https://risk-engineering.org/notebook/regression-CCPP.html

[7] https://www.slideshare.net/JyothiLakshmi12/analytics-project-combined-cycle-power-plant

[8] http://rstudio-pubs-static.s3.amazonaws.com/269645_4a16828a78fd44bdad4bc0481d5ac0bc.html

Thank you