SDA Project Report

Power Plant Dataset

Group No - 53

Deepesh Bhageria(S20170010041)
Aman Kumar(S20170010010)
Rahul prasad(S20170010118)
Vineet Sharma(S20170010179)

Content

Abstract

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). An appropriate dataset must be accessible to the system under analysis to train. The purpose of this work is to do Exploratory Data Analysis and apply a Linear regression model and test how good the model is. We also checked the linear regression basic assumptions and based on that we apply Principal Component Analysis (PCA) with 2 principal components.

Introduction

One of the most important resources for human activities is electricity. In order to provide human populations with the necessary amount of energy, a power plant has been set up. For several factors, including environmental conditions, the power supplied from power plants fluctuates throughout the year. A high number of parameters and assumptions are needed to accurately analyze thermodynamic power plants using mathematical models in order to reflect the real unpredictability of the system. Regression model methods can be used for modeling the thermodynamics of the device.Environmental conditions are studied as inputs of the model, and the generated power as the output of the model. Using this model, given the environmental conditions, we can predict the plant's production capacity. The extensive analysis of the regression model is analyzed in this work.

Methodology

1. EDA:
    - Used box plot for determining whether data contains outliers or not.
    - We plot scatter matrix and confusion matrix to understand the correlation between dependent and independent variables.
    - We then normalize the data.
    - To check whether data is stationary or not, we used the Augmented Dickey Fuller Test.
    - We then used the Ordinary Least Squares (OLS) regression model to determine which independent variable is significant and also plotted independent variable vs dependent variable plot.

2. Model Building

- Split the data into training and test sets in the ratio 7:3.
- We then build the regression model using the formula shown below.

$$\theta = (X^T X)^{-1} X^T y$$

- For evaluating the model, we calculated the Mean Squared Error and R-Square value.
- In order to validate the model we checked a few assumptions of the linear regression model.
- Based on the assumptions we then applied PCA. For this we first used scree plot to determine the number of principal components and used these principal components for further analysis.

Dataset

The dataset was collected from a Combined Cycle Power Plant over 6 years (2006-2011) when the power plant was set to work with a 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 (PE) of the plant. A combined-cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST), and heat recovery steam generators.

RangeIndex: 9568 entries, 0 to 9567

Data columns (total 5 columns):

```
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   AT      9568 non-null   float64
 1   V       9568 non-null   float64
 2   AP      9568 non-null   float64
 3   RH      9568 non-null   float64
 4   PE      9568 non-null   float64
```

dtypes: float64(5)
memory usage: 373.9 KB

First five samples of the dataset:

```
     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
```

- All the independent variables are of numerical data type and it doesn't have any null values.
- The dependent variable is PE - Electrical energy output is also a numeric type without null values.

Description of Data

Attribute Information: Features consist of hourly average ambient variables:

- Temperature (AT) 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
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

Null or Missing value checking

Is there a null value in Train data?	No
Is there any missing value in the dataset?	No

Exploratory Data Analysis

Box plot of the independent variables

Boxplot

A boxplot is a standardized way of displaying the dataset based on a five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles.

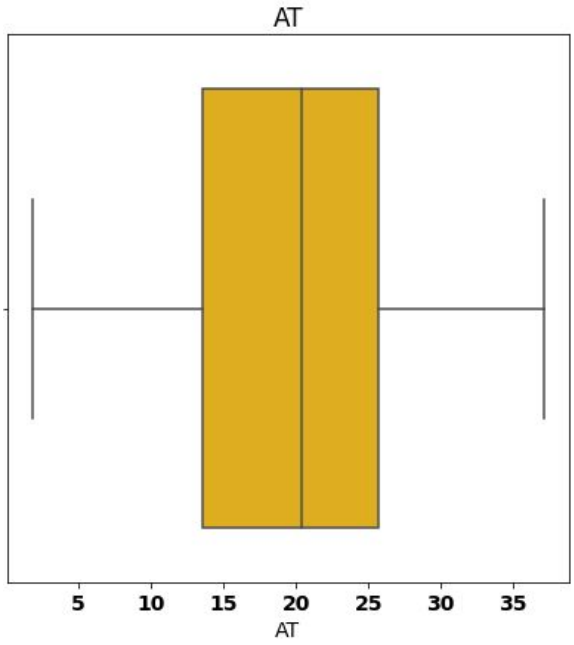
Elements of Boxplot

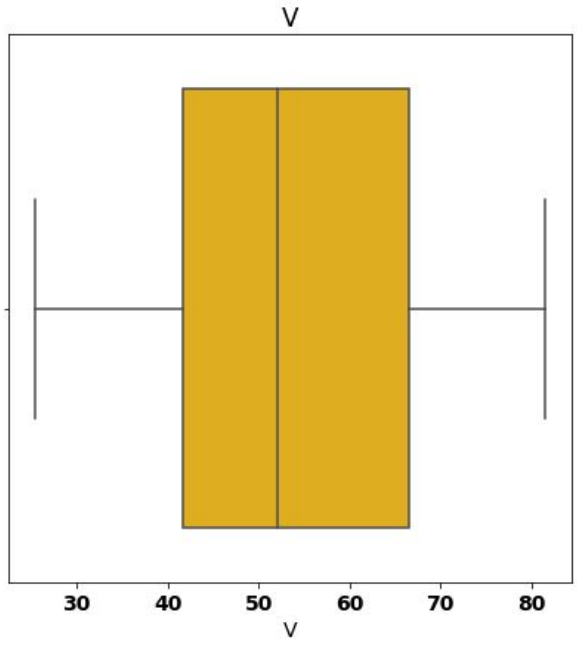
- **Minimum**
  - The lowest data point excluding any outliers.
- **Maximum**
  - The largest data point excluding any outliers.
- **Average (Q2 / 50th percentile)**
  - The middle value of the dataset.
- **First quartile (Q1 / 25th percentile)**
  - Also known as the lower quartile qn(0.25)
  - It is the median of the lower half of the dataset.
- **Third quartile (Q3 / 75th percentile)**
  - Also known as the upper quartile qn(0.75)
  - It is the median of the upper half of the dataset
- **Interquartile range (IQR)**
  - It is the distance between the upper and lower quartiles.

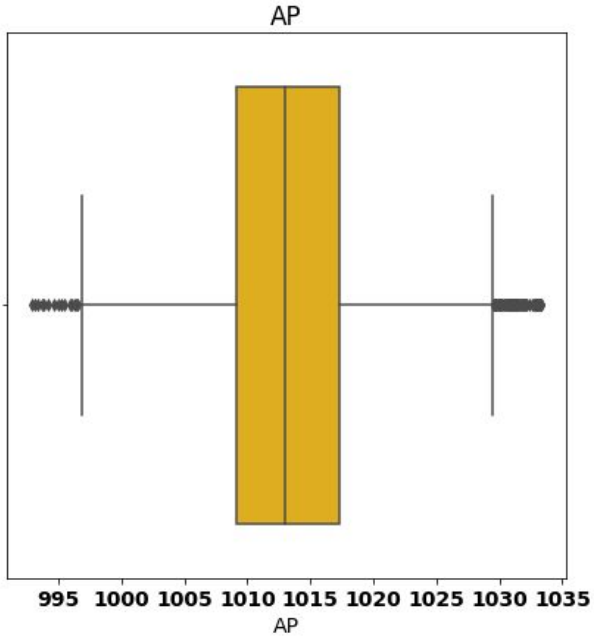
$$\mathrm{IQR} = Q_3 - Q_1 = q_n(0.75) - q_n(0.25)$$

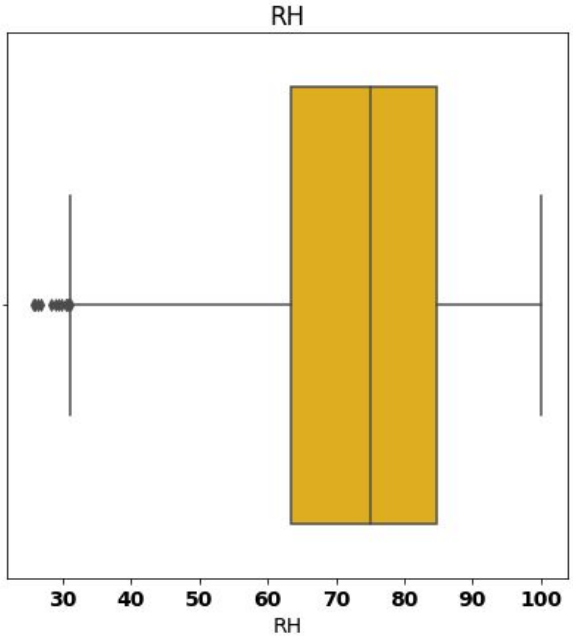
A box plot is constructed of two parts, a box and a set of whiskers. The lowest point is the minimum of the data set and the highest point is the maximum of the data set. The box is drawn from Q1 to Q3 with a horizontal line drawn in the middle to denote the median.

Any data not included between the whiskers should be plotted as an outlier with a dot, small circle, or star, but occasionally this is not done.









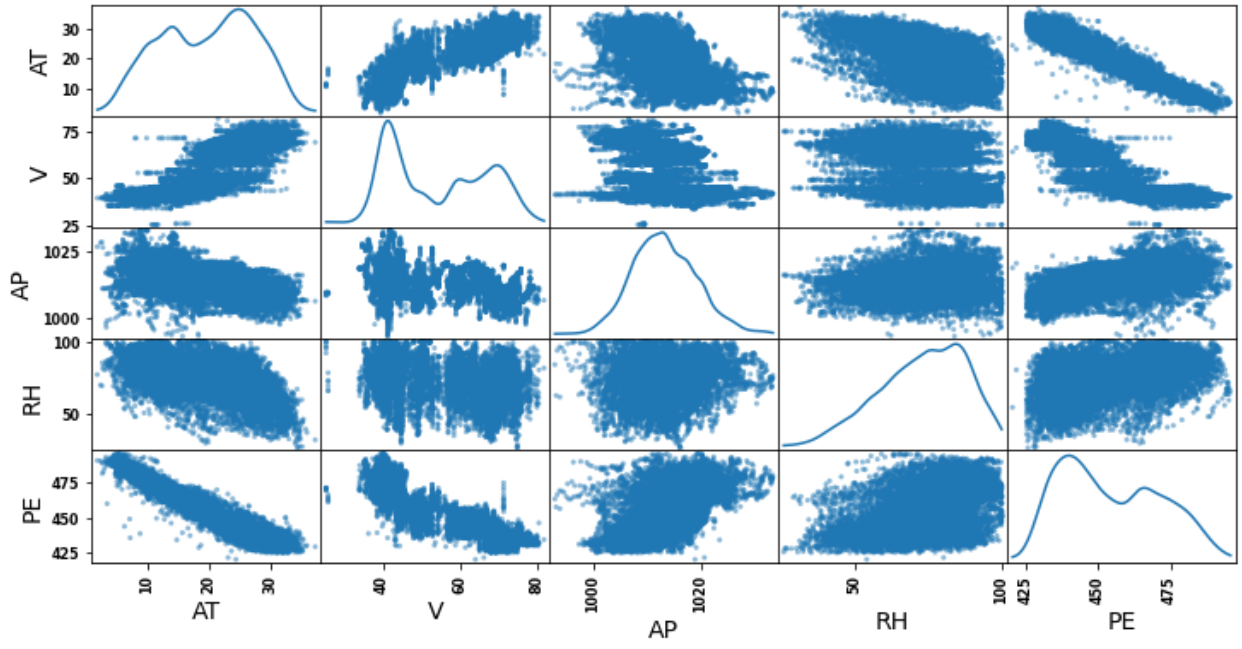
- No outliers for predictor AT
- No outliers for predictor V.
- There are outliers for predictor RH. Outliers are the values less than 30.
- There are outliers for predictor AP. Outliers are the values less than 997 and greater than 1030.

Scatter Plot

About Scatter Plot

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

A scatter plot can be used either when one continuous variable that is under the control of the experimenter and the other depends on it or when both continuous variables are independent. If a parameter exists that is systematically incremented and/or decremented by the other, it is called the control parameter or independent variable and is customarily plotted along the horizontal axis. The measured or dependent variable is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis and a scatter plot will illustrate only the degree of correlation (not causation) between two variables.



In this matrix, the diagonal contains a plot of the distribution of each variable. We observe that:

- There is an approximately linear relationship between PE and the negative of AT
- there is an approximately linear relationship between PE and negative of V

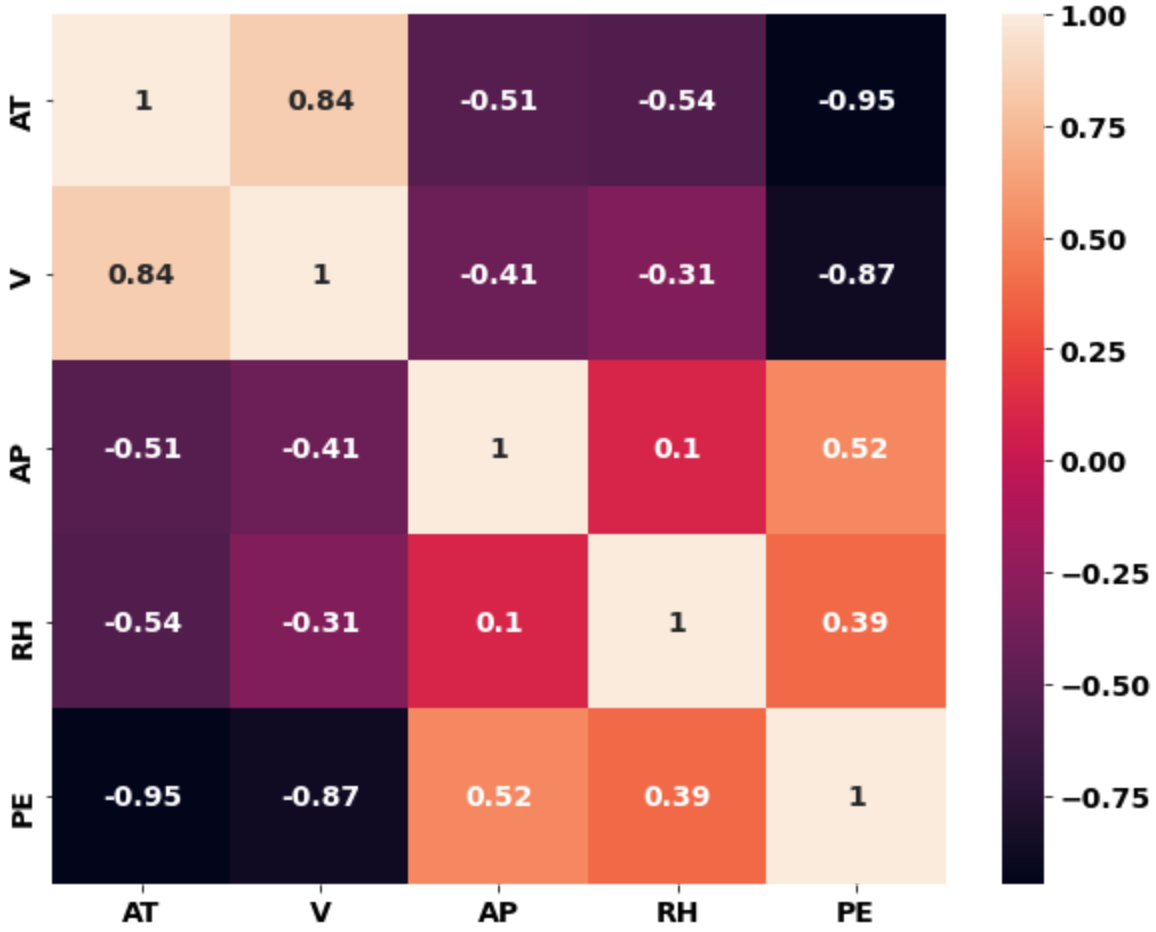
It can be inferred from the scatter plots that the feature AT and V have a significant linear relationship with the response variable. The features AP and RH are not that closely linearly related to the response.

- AT - Temperature is more or less normally distributed.

- AP - Pressure is normally distributed
- RH - Humidity is left skewed.
- PE - It has 2 peaks with normally distributed plot.

Correlation between the attributes(Correlation Matrix)

Heatmap



The Relationship between AT vs V has a strong positive correlation as 0.84.

Augmented Dickey Fuller Test to check whether PE is time dependent or not

Augmented Dickey Fuller test (ADF Test) is a statistical test used to test whether a given Time series is stationary or not. It is from the test statistic and the p-value, we can make an inference as to whether a given series is stationary or not.

Null hypothesis	PE is time dependent(non-stationary)
Alternate hypothesis	PE is time independent(stationary)

If the p-value is less than the significance level, we reject the null hypothesis and infer that PE is stationary.

Result

p-value	0.000000
Critical Values	1%: -3.431 5%: -2.862 10%: -2.567

0 p-value, lower than 1% crit val. We can safely reject the null-hypothesis that the PE is time dependent.

Checking Independent Variable Significance

For AT

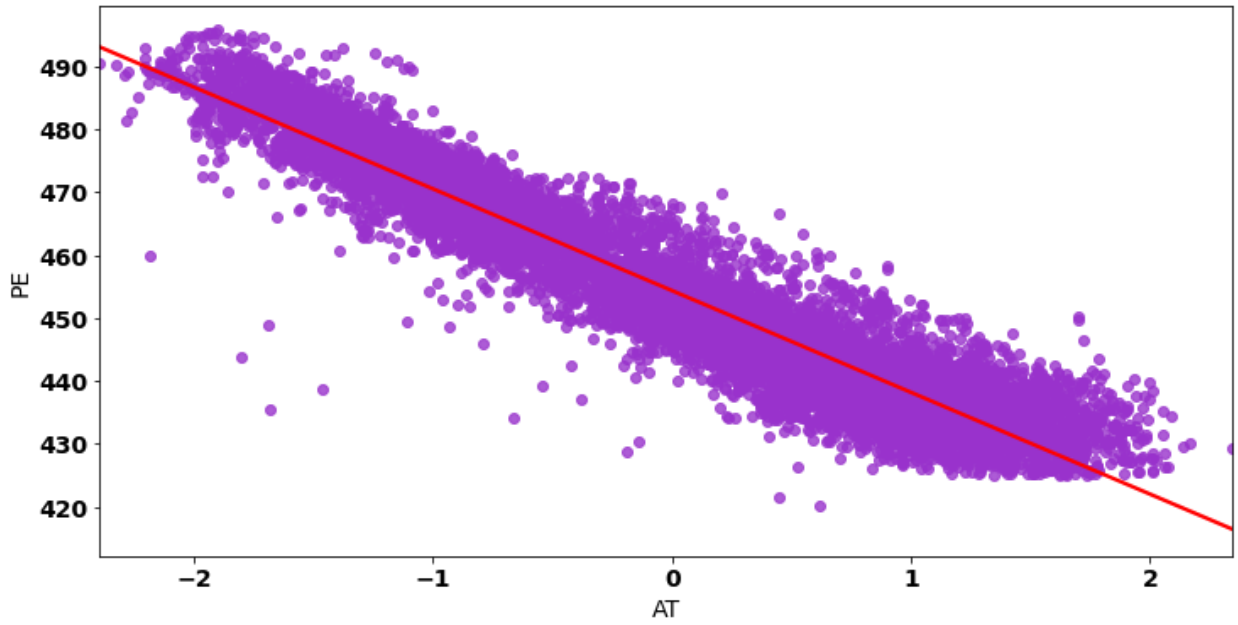
OLS Regression Results

Dep. Variable:	PE	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.899
Method:	Least Squares	F-statistic:	8.510e+04
Date:	Sat, 05 Dec 2020	Prob (F-statistic):	0.00
Time:	12:10:22	Log-Likelihood:	-29756.
No. Observations:	9568	AIC:	5.952e+04
Df Residuals:	9566	BIC:	5.953e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	454.3650	0.055	8191.480	0.000	454.256	454.474
AT	-16.1809	0.055	-291.715	0.000	-16.290	-16.072

Omnibus:	417.457	Durbin-Watson:	2.033
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1117.844
Skew:	-0.209	Prob(JB):	1.83e-243
Kurtosis:	4.621	Cond. No.	1.00

- From the above summary, this can be inferred that AT is a significant feature as the p-value for AT is less than 0.05.
- R_squared=0.899



For V

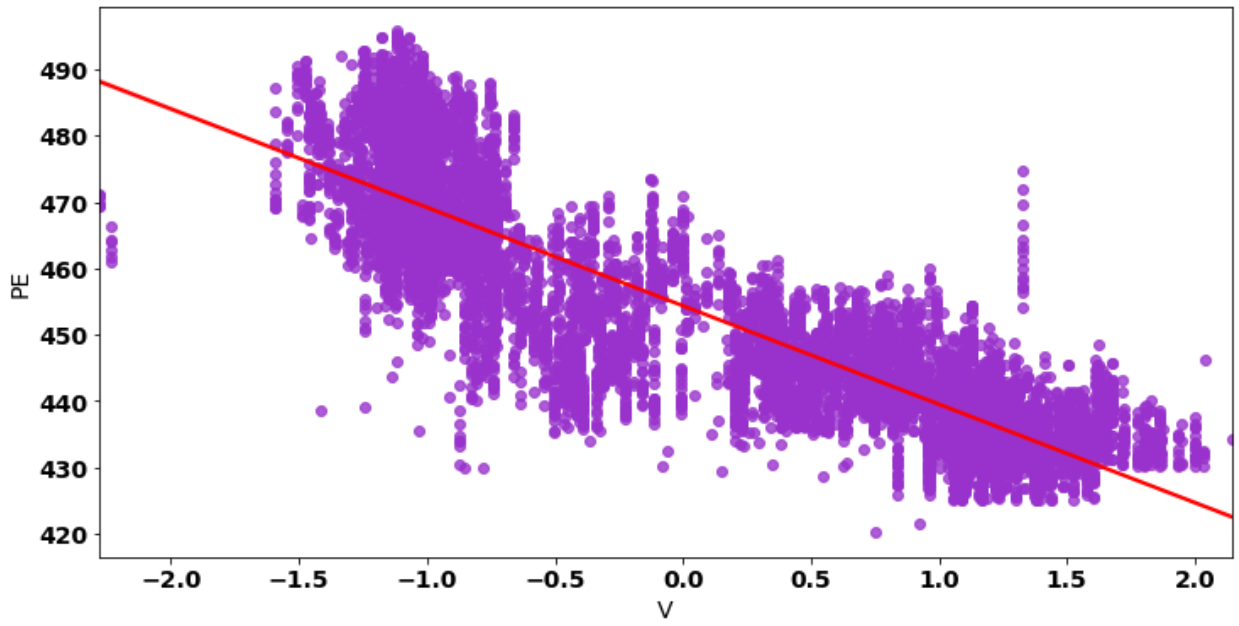
OLS Regression Results

Dep. Variable:	PE	R-squared:	0.757
Model:	OLS	Adj. R-squared:	0.756
Method:	Least Squares	F-statistic:	2.972e+04
Date:	Sat, 05 Dec 2020	Prob (F-statistic):	0.00
Time:	12:10:23	Log-Likelihood:	-33963.
No. Observations:	9568	AIC:	6.793e+04
Df Residuals:	9566	BIC:	6.794e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	454.3650	0.086	5277.182	0.000	454.196	454.534
V	-14.8438	0.086	-172.402	0.000	-15.013	-14.675

Omnibus:	77.693	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	109.571
Skew:	-0.097	Prob(JB):	1.61e-24
Kurtosis:	3.487	Cond. No.	1.00

- From the above summary, this can be inferred that V is a significant feature as the p-value for AT is less than 0.05.
- R_squared=0.757



For AP

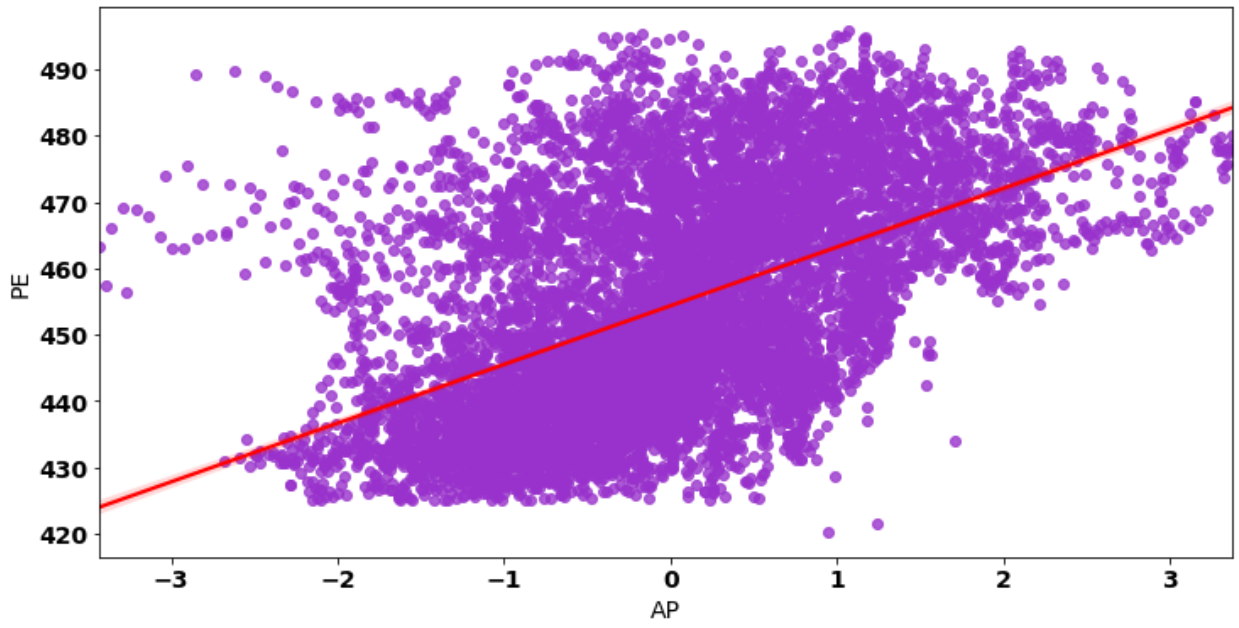
OLS Regression Results

Dep. Variable:	PE	R-squared:	0.269
Model:	OLS	Adj. R-squared:	0.269
Method:	Least Squares	F-statistic:	3516.
Date:	Sat, 05 Dec 2020	Prob (F-statistic):	0.00
Time:	12:10:24	Log-Likelihood:	-39224.
No. Observations:	9568	AIC:	7.845e+04
Df Residuals:	9566	BIC:	7.847e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	454.3650	0.149	3045.147	0.000	454.073	454.657
AP	8.8476	0.149	59.296	0.000	8.555	9.140

Omnibus:	525.438	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	612.290
Skew:	0.616	Prob(JB):	1.10e-133
Kurtosis:	2.859	Cond. No.	1.00

- From the above summary, this can be inferred that AP is a significant feature as the p-value for AT is less than 0.05.
- R_squared=0.269



For RH

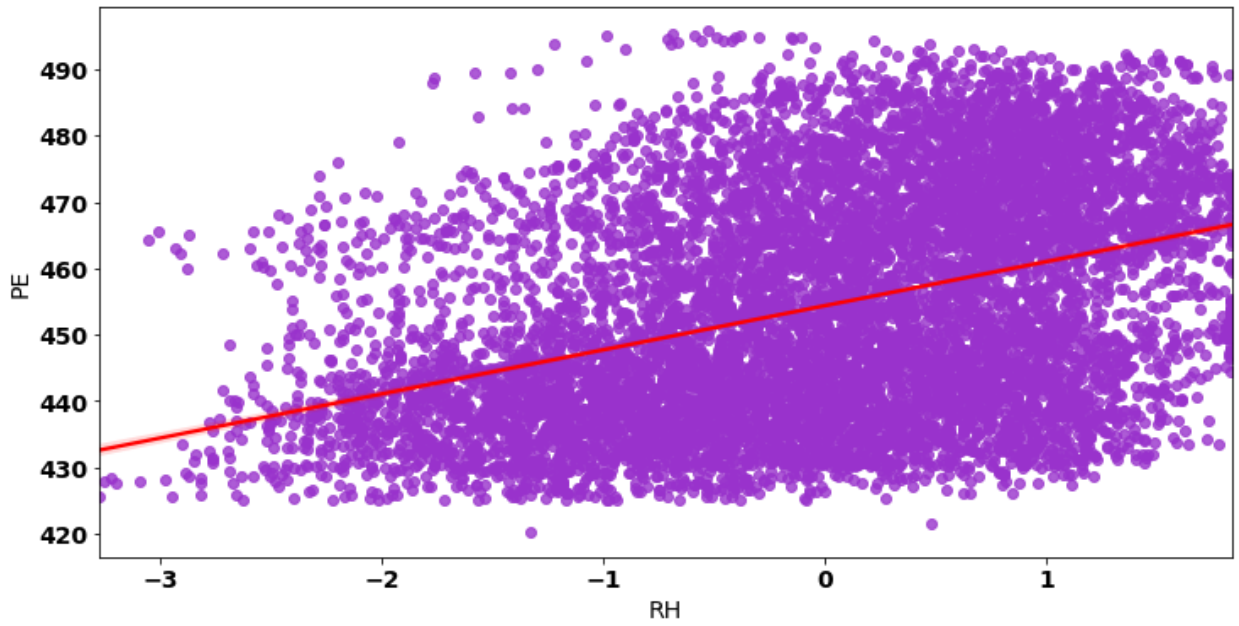
OLS Regression Results

Dep. Variable:	PE	R-squared:	0.152
Model:	OLS	Adj. R-squared:	0.152
Method:	Least Squares	F-statistic:	1714.
Date:	Sat, 05 Dec 2020	Prob (F-statistic):	0.00
Time:	12:10:25	Log-Likelihood:	-39933.
No. Observations:	9568	AIC:	7.987e+04
Df Residuals:	9566	BIC:	7.988e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	454.3650	0.161	2827.628	0.000	454.050	454.680
RH	6.6523	0.161	41.399	0.000	6.337	6.967

Omnibus:	772.278	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	319.245
Skew:	0.231	Prob(JB):	4.75e-70
Kurtosis:	2.234	Cond. No.	1.00

- From the above summary, this can be inferred that RH is a significant feature as the p-value for AT is less than 0.05.
- R_squared=0.152



For [AT, V, AP, RH]

OLS Regression Results

Dep. Variable:	PE	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	3.114e+04
Date:	Sat, 05 Dec 2020	Prob (F-statistic):	0.00
Time:	12:10:26	Log-Likelihood:	-28088.
No. Observations:	9568	AIC:	5.619e+04
Df Residuals:	9563	BIC:	5.622e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	454.3650	0.047	9750.142	0.000	454.274	454.456
AT	-14.7366	0.114	-129.342	0.000	-14.960	-14.513
V	-2.9724	0.093	-32.122	0.000	-3.154	-2.791
AP	0.3687	0.056	6.564	0.000	0.259	0.479
RH	-2.3075	0.061	-37.918	0.000	-2.427	-2.188

Omnibus:	892.002	Durbin-Watson:	2.033
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4086.777
Skew:	-0.352	Prob(JB):	0.00
Kurtosis:	6.123	Cond. No.	4.88

- P-value of all the independent variables is less than 0.05, So all are significant.
- R_Squared=0.929

Model Building

In this step build model using our linear regression equation:-

$$\theta = (X^T X)^{-1} X^T y$$

In first step we need to add a feature $x_0 = 1$ to our original data set.

Parameter	Θ	
x_0	Θ_0	454.343488
x_1	Θ_1	-14.561446
x_2	Θ_2	-3.092156

x_3	Θ_3	0.393396
x_4	Θ_4	-2.233959

Model Evaluation

We will predict the value for the target variable by using our model parameter for the test data set. Then compare the predicted value with the actual value in the test set. We compute Mean Square Error using formula:-

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

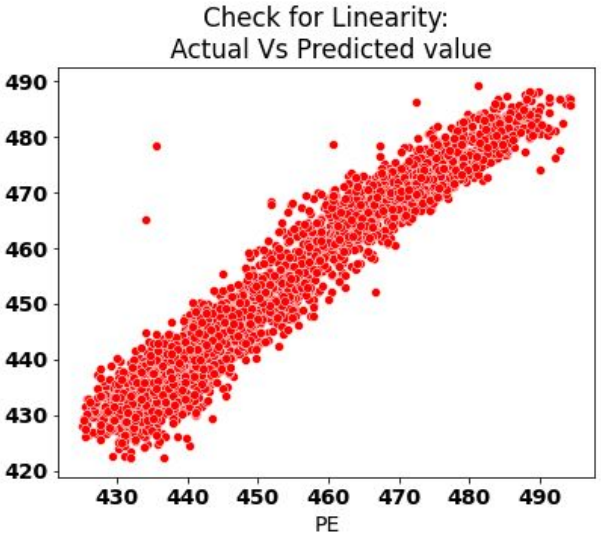
R^2 is a statistical measure of how close data are to the fitted regression line. R^2 is always between 0 to 100%. 0% indicated that the model explains none of the variability of the response data around it's mean. 100% indicated that the model explains all the variability of the response data around the mean.

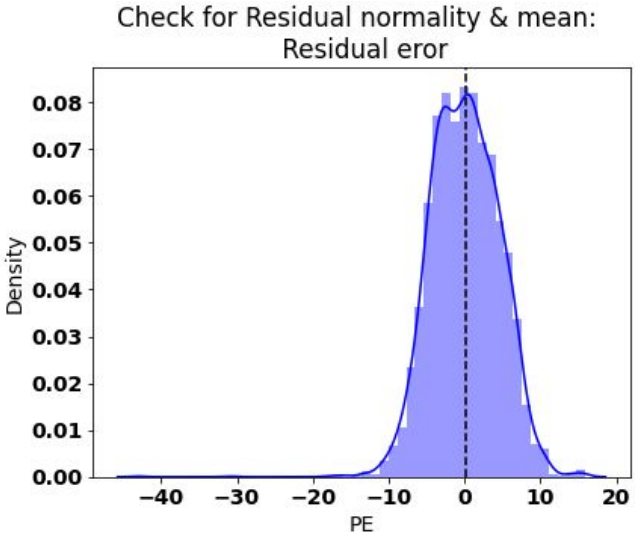
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

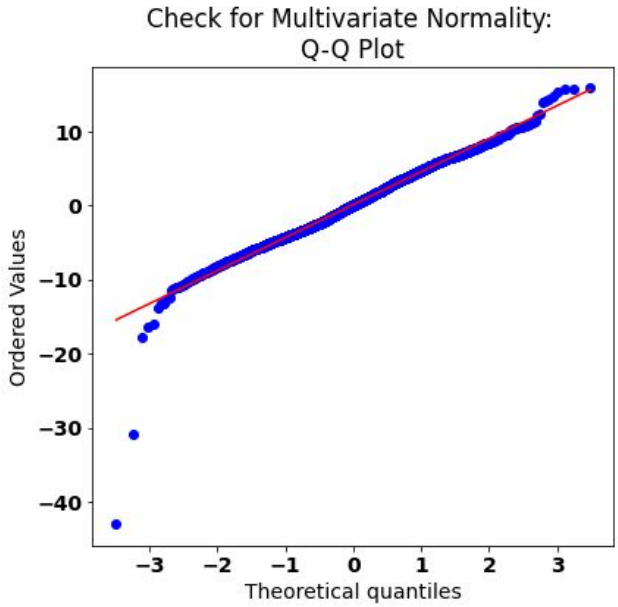
MSE	20.200280077790055
R^2	0.9315977159014461

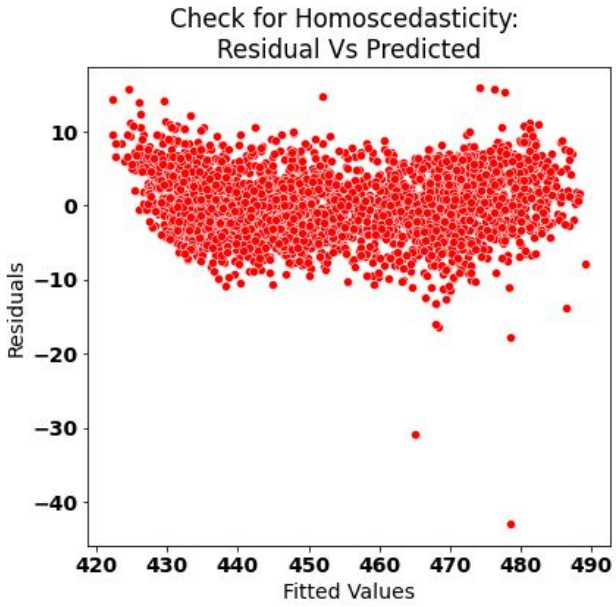
Validating Regression Model

- Linear Relationship: In linear regression the relationship between the dependent and independent variable to be linear. This can be checked by scatter plotting Actual value Vs Predicted value
- The residual error plot should be normally distributed.
- The mean of residual error should be 0 or close to 0 as much as possible
- Linear regression requires all variables to be multivariate normal. This assumption can best be checked with a Q-Q plot.
- Linear regression assumes that there is little or no Multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. The variance inflation factor VIF* identifies correlation between independent variables and strength of that correlation. VIF=1/(1−R2) , If VIF >1 & VIF <5 moderate correlation, VIF > 5 critical level of multicollinearity.
- Homoscedasticity: The data are homoscedastic meaning the residuals are equal across the regression line. We can look at residual Vs fitted value scatter plots. The heteroscedastic plot would exhibit a funnel shape pattern.







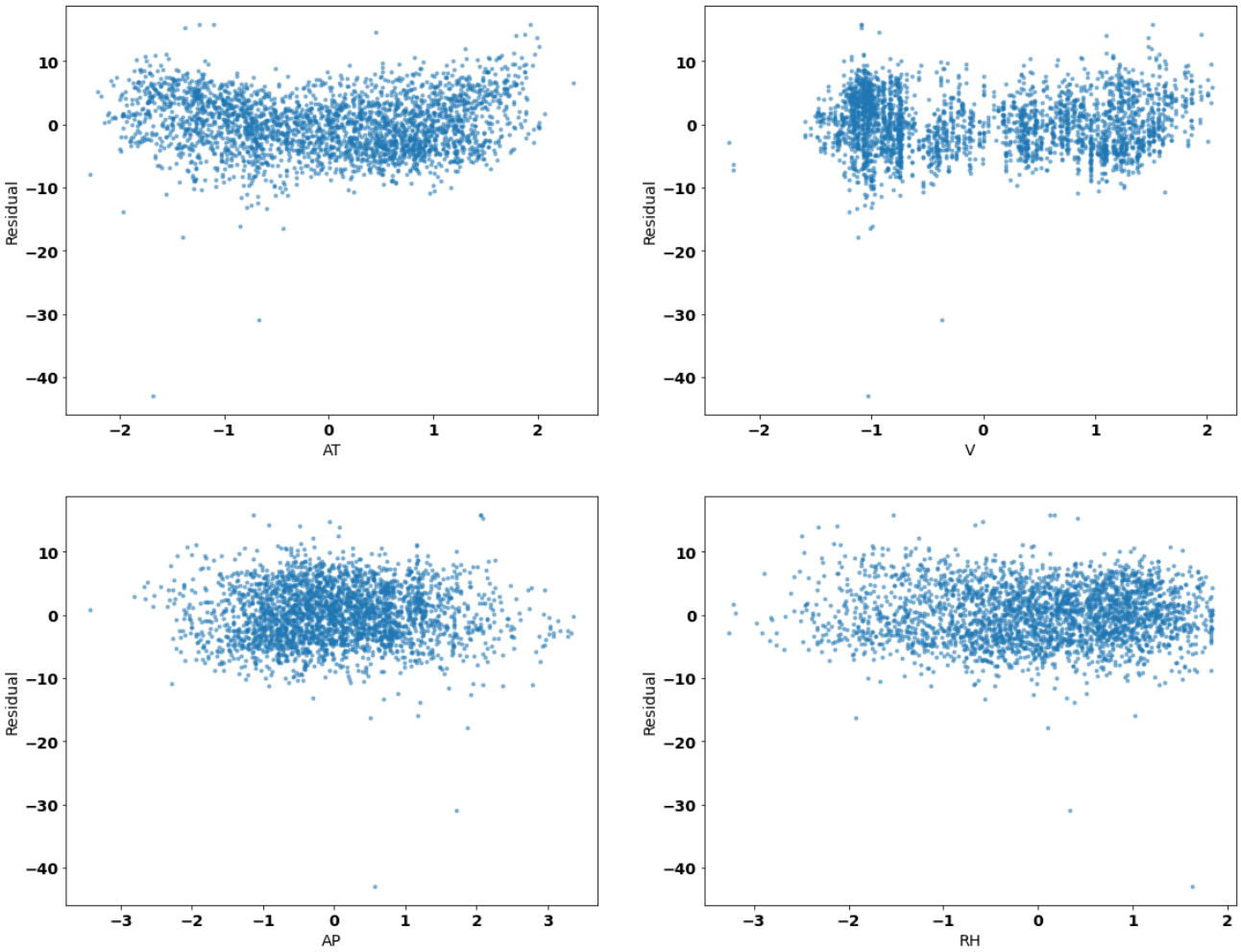


Check for MultiCollinearity

$$VIF(Variance\ Inflation\ Factor) = 1/(1-R^2) = 14.61939$$

- In our model the actual vs predicted plot is almost linear.
- The residual mean is zero.
- Q-Q plot is not showing an almost straight line which means both sets of Quantiles come from the same distribution.
- The plot is homoscedastic.
- Variance inflation factor value is greater than 5, so multicollearity.

Residual Plots

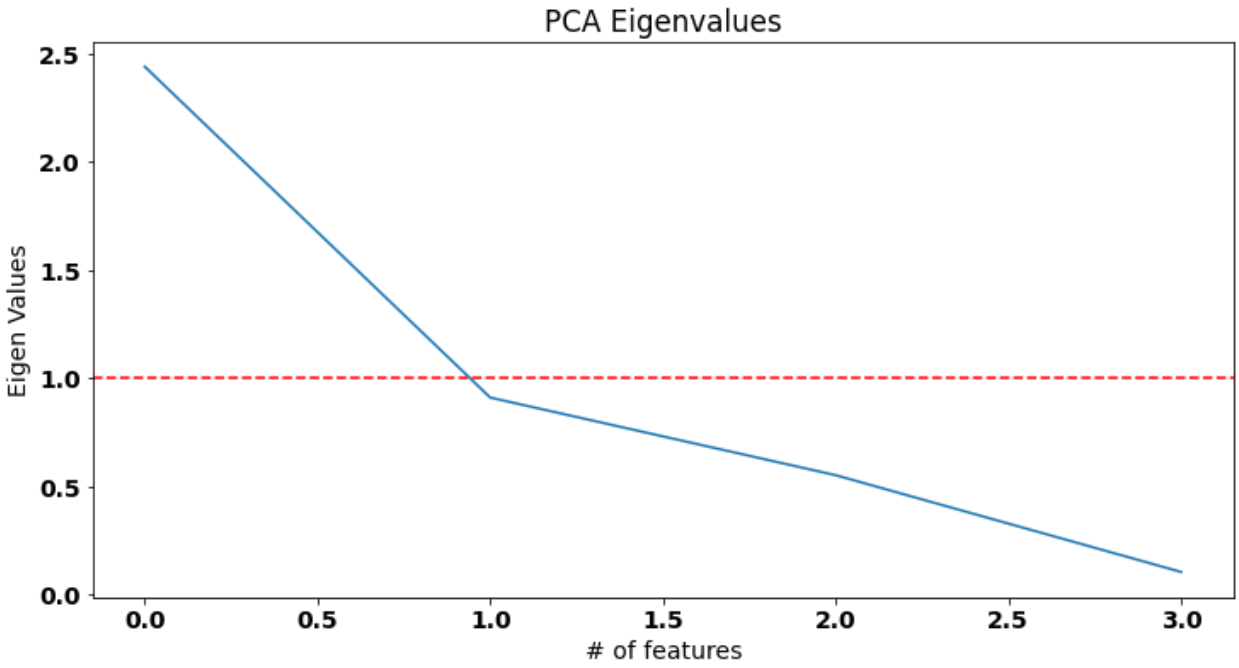


Plot between Residual and Independent variables

Indeed, except for a minor quadratic shape to the residuals of variable AT, the residuals look random, without any systematic feature.

PCA

Scree plot for finding number of PCs to be retained



- Eigen value 1 is greater than 1
- Eigen value 2 is approximately 1.
- Eigen value 3 and 4 are much less than 1.

This suggests that 2 principal components should be preserved.

Applying PCA with n_components=2

The Mean Square Error(MSE) or J(theta)	46.02471677751573
R square	0.8441508860051645
VIF	6.416462528193378

After Applying PCA with n_components=2, the VIF is reduced from 14.62 to 6.41, Which shows that multicollinearity decreased. However the R_square value decreased and MSE value increased.

Conclusion

Primary observations and EDA on the dataset found that the dataset had only numerical variables and there are variables which are directly involved in calculation of Energy output(PE). Some of the variables contain outliers. Based on scatter plot and confusion matrix we found out that some variables are highly correlated. We used the Augmented Dickey Fuller test and observed that the data is stationary. The OLS regression model suggests that all the given factors are significant for determining the dependent variable. The mean squared error calculated on the regression model was 20.200 and R-square value was 0.931. While validating the regression model, we found out that the independent variables are highly correlated as VIF was 14.619. To deal with multicollinearity, we first used scree plot which suggests that the number of principal components should be 2. We then used these principal components to reduce the multicollinearity. After applying PCA, MSE was 46.024, R_square was 0.844 and VIF was 6.416. VIF=6.416 suggests multicollinearity reduced by a lot.

References

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, [Web Link]. ([Web Link])

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)