

# Parameters Tuning and Feature Selection

October 2, 2019

## 1 CGC Total Power Prediction

### 1.1 Data Information

#### 1.1.1 Description

The cracked gas compressor (CGC) is the most critical unit in the entire ethylene production facility, It Consisting 5 stages.

### 1.2 Attribute information

The dataset is time series (10/Jun/2014 - 31/Dec/2017) data of various online sensors and other measured attributes for a CGC unit.

The dataset attributes properties are as follows:

Numer – Name – Data Type – Measurement – Description

0	1st Stage Suction Temperature	quantitative	°C	input variable
1	1st Stage Suction Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
2	1st Stage Discharge Temperature	quantitative	°C	input variable
3	1st Stage Discharge Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
4	2nd Stage Suction Temperature	quantitative	°C	input variable
5	2nd Stage Suction Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
6	2nd Stage Discharge Temperature	quantitative	°C	input variable
7	2nd Stage Discharge Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
8	3rd Stage Suction Temperature	quantitative	°C	input variable
9	3rd Stage Suction Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
10	3rd Stage Discharge Temperature	quantitative	°C	input variable
11	3rd Stage Discharge Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
12	4th Stage Suction Temperature	quantitative	°C	input variable
13	4th Stage Suction Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
14	4th Stage Discharge Temperature	quantitative	°C	input variable
15	4th Stage Discharge Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
16	5th Stage Suction Temperature	quantitative	°C	input variable
17	5th Stage Suction Pressure	quantitative	kNm <sup>3</sup> /hr	input variable
18	5th Stage Discharge Temperature	quantitative	°C	input variable
19	5th Stage Discharge Pressure	quantitative	kg/cm <sup>2</sup> g	input variable
20	1st Stage Discharge Flow	quantitative	kNm <sup>3</sup> /hr	input variable
21	3rd Stage Discharge Flow	quantitative	kNm <sup>3</sup> /hr	input variable
22	5th Stage Discharge Flow	quantitative	kNm <sup>3</sup> /hr	input variable

23	C3 Splitter Purge to 4th Stage Suction	quantitative	Nm3/hr	input variable
24	C2 Splitter Purge to 5th Stage Suction	quantitative	Nm3/hr	input variable
25	CW Supply Temperature	quantitative	°C	input variable
26	CW Flow to Olefins	quantitative	tons/hr	input variable
27	CW Pressure	quantitative	kg/cm2g	input variable
28	Compressor Speed	quantitative	rpm	input variable
29	UHP Steam Flow to KT-1	quantitative	tons/hr	input variable
30	UHP Steam Pressure	quantitative	kg/cm2g	input variable
31	UHP Steam Temperature	quantitative	°C	input variable
32	HP Steam Extraction Flow from KT-1	quantitative	tons/hr	input variable
33	HP Steam Extraction Pressure	quantitative	kg/cm2g	input variable
34	HP Steam Extraction Temperature	quantitative	°C	input variable
35	E-24 PG Inlet Temperature	quantitative	°C	input variable
36	Total Power	quantitative	MW	output variable
37	Date/Time	DateTime	Date & H:M:s	input variable

### 1.3 Notebook Setup

```
[1]: # Common imports
import sys
import os
import numpy as np

#Python Data Analysis Library
import pandas as pd

#Data visualization
%matplotlib inline
#sets the backend of matplotlib to the 'inline' backend
#%matplotlib notebook
import matplotlib
import seaborn as sns
import matplotlib.pyplot as plt

# Ignore useless warnings (see SciPy issue #5998)
import warnings
warnings.filterwarnings(action="ignore", message="^internal gelsd")

#Data pretty printer
from pprint import pprint
```

### 1.4 Data Import

```
[2]: df = pd.read_excel('E:/nikhitha/datasets/CGC/CGC Total Power consumption.xlsx', ↴sheetname='Sheet1')
```

E:\nikhitha\Anaconda\lib\site-packages\pandas\util\\_decorators.py:188:

```
FutureWarning: The `sheetname` keyword is deprecated, use `sheet_name` instead
return func(*args, **kwargs)
```

### 1.4.1 Viewing the Dataset

```
[3]: df.head()
```

```
[3]:
```

	Description	UOM	Type of Data	Tag	\
0	1st Stage Suction Temperature	°C	Measured	01TI2356-PV	
1	1st Stage Suction Pressure	kg/cm2g	Measured	01PC2358-PV	
2	1st Stage Discharge Temperature	°C	Measured	01TI2378-PV	
3	1st Stage Discharge Pressure	kg/cm2g	Measured	01PI2364-PV	
4	2nd Stage Suction Temperature	°C	Measured	01TI2402-PV	

	2014-06-10 00:00:00	2014-06-11 00:00:00	2014-06-12 00:00:00	\
0	30.245354	33.657	32.857358	
1	0.420737	0.354	0.392626	
2	88.910154	91.477	92.121653	
3	2.006712	2.007	2.038484	
4	32.622496	32.967	34.982900	

	2014-06-13 00:00:00	2014-06-14 00:00:00	2014-06-15 00:00:00	...	\
0	31.50	30.3	31.922800	...	
1	0.40	0.4	0.424725	...	
2	89.07	87.3	88.611175	...	
3	2.00	2.0	2.017133	...	
4	34.40	33.0	34.457996	...	

	2017-12-22 00:00:00	2017-12-23 00:00:00	2017-12-24 00:00:00	\
0	35.223558	34.082387	34.465854	
1	0.372329	0.366712	0.442096	
2	85.545846	84.807633	84.338225	
3	2.410392	2.387933	2.368546	
4	37.430446	35.956879	36.735058	

	2017-12-25 00:00:00	2017-12-26 00:00:00	2017-12-27 00:00:00	\
0	32.872308	33.156150	32.496662	
1	0.353642	0.397717	0.449746	
2	82.447150	83.877837	84.337496	
3	2.328996	2.327996	2.335800	
4	33.056492	33.887275	35.249033	

	2017-12-28 00:00:00	2017-12-29 00:00:00	2017-12-30 00:00:00	\
0	33.565096	34.323767	35.719146	
1	0.431175	0.428258	0.491954	
2	84.488404	84.339037	90.680967	
3	2.342221	2.345983	2.349912	

```
4           36.143021          35.829054          37.922896  
2017-12-31 00:00:00  
0           33.044412  
1           0.499692  
2          91.099225  
3          2.348679  
4          34.827075  
  
[5 rows x 1184 columns]
```

[4]: df.info()

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 37 entries, 0 to 36  
Columns: 1184 entries, Description to 2017-12-31 00:00:00  
dtypes: float64(1178), object(6)  
memory usage: 342.3+ KB
```

#### 1.4.2 Removing the Unnecessary columns

[5]: df1 = df.drop(columns = ["UOM", "Type of Data", "Tag"])

#### 1.4.3 Finding Dimensions of data

[6]: df1.shape

[6]: (37, 1181)

This data consisting of 37 rows and 1181 columns, which is unstructured form, so we need to make them into structured form by converting rows to columns and vice versa

#### 1.4.4 Transpose of data

[7]: df1 = df1.T.reset\_index() # transpose the dataset converting rows to columns and columns to rows  
df1.head()

```
[7]:      index          0  \  
0    Description  1st Stage Suction Temperature  
1  2014-06-10 00:00:00          30.2454  
2  2014-06-11 00:00:00          33.657  
3  2014-06-12 00:00:00          32.8574  
4  2014-06-13 00:00:00          31.5  
  
              1          2  \  
0  1st Stage Suction Pressure  1st Stage Discharge Temperature
```

1	0.420737	88.9102		
2	0.354	91.477		
3	0.392626	92.1217		
4	0.4	89.07		
	3	4 \		
0	1st Stage Discharge Pressure	2nd Stage Suction Temperature		
1	2.00671	32.6225		
2	2.007	32.967		
3	2.03848	34.9829		
4	2	34.4		
	5	6 \		
0	2nd Stage Suction Pressure	2nd Stage Discharge Temperature		
1	2.0983	91.5992		
2	1.934	91.405		
3	2.07911	94.0512		
4	2.11	92.03		
	7	8 ... \		
0	2nd Stage Discharge Pressure	3rd Stage Suction Temperature	...	
1	5.83635	34.8865	...	
2	5.407	34.709	...	
3	5.82529	37.0767	...	
4	6	36.9	...	
	27	28	29	30 \
0	CW Pressure	Compressor Speed	UHP Steam Flow to KT-1	UHP Steam Pressure
1	4.1175	4554.04	94.6215	105.489
2	4.075	4546.4	98.9756	105.696
3	4.05212	4601.99	103.414	105.709
4	4.075	4508.8	111.3	105.7
	31		32 \	
0	UHP Steam Temperature	HP Steam Extraction Flow from KT-1		
1	464.566	25.5534		
2	459.615	39.6966		
3	459.159	40.9557		
4	451.9	39.88		
	33		34 \	
0	HP Steam Extraction Pressure	HP Steam Extraction Temperature		
1	45.3681	368.727		
2	45.3875	368.215		
3	45.6778	367.63		
4	45.87	361.19		

	35	36
0	E-24 PG Inlet Temperature	Total Power
1	33.0434	13.8885
2	33.399	13.5798
3	36.6697	14.6584
4	36.9915	14.7133

[5 rows x 38 columns]

### Assigning the first row of the dataset to Column header

```
[8]: header = df1.iloc[0] # assigning the first row of the dataset to header
header
```

```
[8]: index                      Description
0               1st Stage Suction Temperature
1               1st Stage Suction Pressure
2               1st Stage Discharge Temperature
3               1st Stage Discharge Pressure
4               2nd Stage Suction Temperature
5               2nd Stage Suction Pressure
6               2nd Stage Discharge Temperature
7               2nd Stage Discharge Pressure
8               3rd Stage Suction Temperature
9               3rd Stage Suction Pressure
10              3rd Stage Discharge Temperature
11              3rd Stage Discharge Pressure
12              4th Stage Suction Temperature
13              4th Stage Suction Pressure
14              4th Stage Discharge Temperature
15              4th Stage Discharge Pressure
16              5th Stage Suction Temperature
17              5th Stage Suction Pressure
18              5th Stage Discharge Temperature
19              5th Stage Discharge Pressure
20              1st Stage Discharge Flow
21              3rd Stage Discharge Flow
22              5th Stage Discharge Flow
23              C3 Splitter Purge to 4th Stage Suction
24              C2 Splitter Purge to 5th Stage Suction
25              CW Supply Temperature
26              CW Flow to Olefins
27              CW Pressure
28              Compressor Speed
29              UHP Steam Flow to KT-1
30              UHP Steam Pressure
31              UHP Steam Temperature
```

```

32          HP Steam Extraction Flow from KT-1
33          HP Steam Extraction Pressure
34          HP Steam Extraction Temperature
35          E-24 PG Inlet Temperature
36          Total Power
Name: 0, dtype: object

```

### Loading the data of all rows and columns

```
[9]: df1 = df1[1:] # loading the data of all rows and columns
df1.head()
```

```
[9]:
      index      0      1      2      3      4      5 \
1  2014-06-10 00:00:00  30.2454  0.420737  88.9102  2.00671  32.6225  2.0983
2  2014-06-11 00:00:00   33.657   0.354   91.477   2.007   32.967   1.934
3  2014-06-12 00:00:00  32.8574  0.392626  92.1217  2.03848  34.9829  2.07911
4  2014-06-13 00:00:00     31.5     0.4     89.07     2     34.4     2.11
5  2014-06-14 00:00:00    30.3     0.4     87.3     2     33     2.11

      6      7      8 ...      27      28      29      30 \
1  91.5992  5.83635  34.8865 ...  4.1175  4554.04  94.6215  105.489
2  91.405   5.407   34.709 ...  4.075   4546.4   98.9756  105.696
3  94.0512  5.82529  37.0767 ...  4.05212  4601.99  103.414  105.709
4  92.03     6     36.9 ...  4.075   4508.8   111.3   105.7
5  90.64     6     35.3 ...  4.05212  4488.2   124     105.7

      31      32      33      34      35      36
1  464.566  25.5534  45.3681  368.727  33.0434  13.8885
2  459.615  39.6966  45.3875  368.215  33.399   13.5798
3  459.159  40.9557  45.6778  367.63   36.6697  14.6584
4  451.9    39.88   45.87   361.19   36.9915  14.7133
5  448.5    53.68   45.89   356.98   35.3933  14.7618

[5 rows x 38 columns]
```

### Assigning the column names to the transposed data set

```
[10]: df1 = df1.rename(columns = header) # Assigning the column names to the
      ↪ transposed data set
df1.head()
```

```
[10]:
      Description 1st Stage Suction Temperature \
1  2014-06-10 00:00:00                      30.2454
2  2014-06-11 00:00:00                      33.657
3  2014-06-12 00:00:00                      32.8574
4  2014-06-13 00:00:00                      31.5
5  2014-06-14 00:00:00                      30.3
```

1st Stage Suction Pressure		1st Stage Discharge Temperature	\
1	0.420737	88.9102	
2	0.354	91.477	
3	0.392626	92.1217	
4	0.4	89.07	
5	0.4	87.3	
1st Stage Discharge Pressure		2nd Stage Suction Temperature	\
1	2.00671	32.6225	
2	2.007	32.967	
3	2.03848	34.9829	
4	2	34.4	
5	2	33	
2nd Stage Suction Pressure		2nd Stage Discharge Temperature	\
1	2.0983	91.5992	
2	1.934	91.405	
3	2.07911	94.0512	
4	2.11	92.03	
5	2.11	90.64	
2nd Stage Discharge Pressure		3rd Stage Suction Temperature	... CW Pressure \
1	5.83635	34.8865	...
2	5.407	34.709	...
3	5.82529	37.0767	...
4	6	36.9	...
5	6	35.3	...
			4.1175
			4.075
			4.05212
			4.075
			4.05212
Compressor Speed UHP Steam Flow to KT-1		UHP Steam Pressure	\
1	4554.04	94.6215	105.489
2	4546.4	98.9756	105.696
3	4601.99	103.414	105.709
4	4508.8	111.3	105.7
5	4488.2	124	105.7
UHP Steam Temperature HP Steam Extraction Flow from KT-1		\	
1	464.566	25.5534	
2	459.615	39.6966	
3	459.159	40.9557	
4	451.9	39.88	
5	448.5	53.68	
HP Steam Extraction Pressure HP Steam Extraction Temperature		\	
1	45.3681	368.727	
2	45.3875	368.215	
3	45.6778	367.63	

```
4          45.87          361.19
5          45.89          356.98
```

```
E-24 PG Inlet Temperature Total Power
1          33.0434      13.8885
2          33.399       13.5798
3          36.6697      14.6584
4          36.9915      14.7133
5          35.3933      14.7618
```

```
[5 rows x 38 columns]
```

## 1.5 Data Engineering and Visualization

### 1.6 Renaming the Column names

```
[11]: df1=df1.rename(columns={"Description": "DateTime"})
```

```
[12]: df1.columns = df1.columns.str.replace(' ', '_')
```

```
[13]: df1.columns = df1.columns.str.replace('-', '_')
```

### Finding the Structure of data after transpose

```
[14]: df1. shape
```

```
[14]: (1180, 38)
```

### 1.7 Finding the datatype of variables

```
[15]: df1.dtypes
```

```
[15]: DateTime                  object
1stStageSuctionTemperature    object
1stStageSuctionPressure       object
1stStageDischargeTemperature   object
1stStageDischargePressure     object
2ndStageSuctionTemperature    object
2ndStageSuctionPressure       object
2ndStageDischargeTemperature   object
2ndStageDischargePressure     object
3rdStageSuctionTemperature    object
3rdStageSuctionPressure       object
3rdStageDischargeTemperature   object
3rdStageDischargePressure     object
4thStageSuctionTemperature    object
4thStageSuctionPressure       object
```

```

4thStageDischargeTemperature      object
4thStageDischargePressure        object
5thStageSuctionTemperature       object
5thStageSuctionPressure         object
5thStageDischargeTemperature    object
5thStageDischargePressure       object
1stStageDischargeFlow           object
3rdStageDischargeFlow          object
5thStageDischargeFlow          object
C3SplitterPurgeto4thStageSuction object
C2SplitterPurgeto5thStageSuction object
CWSupplyTemperature             object
CWFlowtoOlefins                object
CWPressure                      object
CompressorSpeed                 object
UHPSteamFlowtoKT1               object
UHPSteamPressure                object
UHPSteamTemperature              object
HPSteamExtractionFlowfromKT1   object
HPSteamExtractionPressure       object
HPSteamExtractionTemperature    object
E24PGInletTemperature          object
TotalPower                      object
dtype: object

```

## 1.8 Datatype Conversion

```
[16]: a = [
    '1stStageSuctionTemperature', '1stStageSuctionPressure', '1stStageDischargeTemperature', '1stStageDischargePressure',
    '4thStageDischargeTemperature', '5thStageSuctionTemperature', '5thStageSuctionPressure', '5thStageDischargeTemperature',
    '3rdStageDischargeFlow', '5thStageDischargeFlow', 'C3SplitterPurgeto4thStageSuction', 'C2SplitterPurgeto5thStageSuction',
    'HPSteamExtractionPressure', 'HPSteamExtractionTemperature', 'E24PGInletTemperature', 'TotalPower'
]
for i in a:
    # df1[i] = df1[i].astype(float)
    df1[i] = pd.to_numeric(df1[i], errors = "Coerce")
```

```
[17]: df1['DateTime'] = pd.to_datetime(df1['DateTime'], format='%Y%m%d:%H:%M:%S.%f')
```

```
[18]: num_cols = df1._get_numeric_data().columns #finding the numerical data types in
      ↪the dataset
num_cols
```

```
[18]: Index(['1stStageSuctionTemperature', '1stStageSuctionPressure',
       '1stStageDischargeTemperature', '1stStageDischargePressure',
       '2ndStageSuctionTemperature', '2ndStageSuctionPressure',
       '2ndStageDischargeTemperature', '2ndStageDischargePressure',
```

```
'3rdStageSuctionTemperature', '3rdStageSuctionPressure',
'3rdStageDischargeTemperature', '3rdStageDischargePressure',
'4thStageSuctionTemperature', '4thStageSuctionPressure',
'4thStageDischargeTemperature', '4thStageDischargePressure',
'5thStageSuctionTemperature', '5thStageSuctionPressure',
'5thStageDischargeTemperature', '5thStageDischargePressure',
'1stStageDischargeFlow', '3rdStageDischargeFlow',
'5thStageDischargeFlow', 'C3SplitterPurgeto4thStageSuction',
'C2SplitterPurgeto5thStageSuction', 'CWSupplyTemperature',
'CWFlowtoOlefins', 'CWPressure', 'CompressorSpeed', 'UHPSteamFlowtoKT1',
'UHPSteamPressure', 'UHPSteamTemperature',
'HPSteamExtractionFlowfromKT1', 'HPSteamExtractionPressure',
'HPSteamExtractionTemperature', 'E24PGInletTemperature', 'TotalPower'],
dtype='object')
```

## 1.9 Finding the Missing Values

[19]: df1.isnull().sum()

[19]:	DateTime	0
	1stStageSuctionTemperature	0
	1stStageSuctionPressure	0
	1stStageDischargeTemperature	0
	1stStageDischargePressure	0
	2ndStageSuctionTemperature	0
	2ndStageSuctionPressure	0
	2ndStageDischargeTemperature	0
	2ndStageDischargePressure	0
	3rdStageSuctionTemperature	0
	3rdStageSuctionPressure	0
	3rdStageDischargeTemperature	0
	3rdStageDischargePressure	0
	4thStageSuctionTemperature	0
	4thStageSuctionPressure	0
	4thStageDischargeTemperature	0
	4thStageDischargePressure	0
	5thStageSuctionTemperature	0
	5thStageSuctionPressure	0
	5thStageDischargeTemperature	0
	5thStageDischargePressure	0
	1stStageDischargeFlow	0
	3rdStageDischargeFlow	0
	5thStageDischargeFlow	0
	C3SplitterPurgeto4thStageSuction	9
	C2SplitterPurgeto5thStageSuction	9
	CWSupplyTemperature	126
	CWFlowtoOlefins	61

```

CWPRESSURE          65
CompressorSpeed     9
UHPSteamFlowtoKT1   20
UHPSteamPressure    9
UHPSteamTemperature  9
HPSteamExtractionFlowfromKT1 11
HPSteamExtractionPressure 9
HPSteamExtractionTemperature 9
E24PGInletTemperature 13
TotalPower          0
dtype: int64

```

## 1.10 Drop the NA values

Removing the null values by dropping

```
[20]: df1 = df1.dropna()
```

```
[21]: df1.isnull().sum()
```

```

[21]: DateTime          0
       1stStageSuctionTemperature 0
       1stStageSuctionPressure   0
       1stStageDischargeTemperature 0
       1stStageDischargePressure  0
       2ndStageSuctionTemperature 0
       2ndStageSuctionPressure   0
       2ndStageDischargeTemperature 0
       2ndStageDischargePressure  0
       3rdStageSuctionTemperature 0
       3rdStageSuctionPressure   0
       3rdStageDischargeTemperature 0
       3rdStageDischargePressure  0
       4thStageSuctionTemperature 0
       4thStageSuctionPressure   0
       4thStageDischargeTemperature 0
       4thStageDischargePressure  0
       5thStageSuctionTemperature 0
       5thStageSuctionPressure   0
       5thStageDischargeTemperature 0
       5thStageDischargePressure  0
       1stStageDischargeFlow      0
       3rdStageDischargeFlow      0
       5thStageDischargeFlow      0
       C3SplitterPurgeto4thStageSuction 0
       C2SplitterPurgeto5thStageSuction 0
       CWSupplyTemperature        0

```

```
CWFlowtoOlefins          0  
CWPressure               0  
CompressorSpeed          0  
UHPSteamFlowtoKT1         0  
UHPSteamPressure          0  
UHPSteamTemperature       0  
HPSteamExtractionFlowfromKT1 0  
HPSteamExtractionPressure 0  
HPSteamExtractionTemperature 0  
E24PGInletTemperature    0  
TotalPower                0  
dtype: int64
```

```
[22]: df1 =df1.drop(columns = ["DateTime"])
```

## 1.11 Descriptive statistics

Finding the Statistics of a variables

```
[23]: df1.describe()
```

```
[23]:      1stStageSuctionTemperature  1stStageSuctionPressure  \  
count           1052.000000          1052.000000  
mean            34.176146           0.407002  
std             2.248426           0.047696  
min             29.067408           0.275196  
25%            32.495414           0.375643  
50%            33.873740           0.401821  
75%            35.791180           0.440996  
max             43.193750           0.500850  
  
      1stStageDischargeTemperature  1stStageDischargePressure  \  
count           1052.000000          1052.000000  
mean            89.003059           2.206902  
std             3.062452           0.131811  
min             80.248479           1.832733  
25%            86.829975           2.110726  
50%            89.087515           2.203019  
75%            91.188673           2.300000  
max             96.194108           2.478983  
  
      2ndStageSuctionTemperature  2ndStageSuctionPressure  \  
count           1052.000000          1052.000000  
mean            35.646605           2.130069  
std             1.707685           0.117609  
min             29.600000           1.750804  
25%            34.484008           2.056935
```

50%	35.598692	2.119512
75%	36.858337	2.207049
max	40.765908	2.397146

	2ndStageDischargeTemperature	2ndStageDischargePressure	\
count	1052.000000	1052.000000	
mean	91.078895	6.215063	
std	2.312947	0.316423	
min	81.611038	5.190508	
25%	89.465036	6.000000	
50%	91.172954	6.168448	
75%	92.824479	6.414383	
max	95.903882	6.984396	

	3rdStageSuctionTemperature	3rdStageSuctionPressure	...	CWPressure	\
count	1052.000000	1052.000000	...	1052.000000	
mean	41.449776	6.144479	...	4.133374	
std	3.306384	0.305581	...	0.156669	
min	34.047133	5.088039	...	3.906527	
25%	38.859876	5.934025	...	4.044953	
50%	41.521540	6.092201	...	4.066325	
75%	43.700996	6.334749	...	4.120564	
max	50.149942	6.926439	...	4.529100	

	CompressorSpeed	UHPSteamFlowtoKT1	UHPSteamPressure	\
count	1052.000000	1052.000000	1052.000000	
mean	4672.900737	125.049858	105.862761	
std	78.496642	16.847428	0.115327	
min	4349.316983	55.095658	104.709571	
25%	4618.242649	117.241789	105.895295	
50%	4664.601746	127.427488	105.905567	
75%	4723.358593	134.967433	105.912120	
max	4920.829988	158.700000	106.194775	

	UHPSteamTemperature	HPSteamExtractionFlowfromKT1	\
count	1052.000000	1052.000000	
mean	450.480235	67.139369	
std	6.874233	19.722090	
min	432.830092	0.000000	
25%	445.365307	63.839020	
50%	450.538090	74.181414	
75%	455.294639	78.392570	
max	469.813531	89.351296	

	HPSteamExtractionPressure	HPSteamExtractionTemperature	\
count	1052.000000	1052.000000	
mean	46.860645	356.128045	

```
      std          0.709262          12.680941
      min          43.232963         258.314796
      25%          46.479201         351.956441
      50%          46.786329         357.109449
      75%          47.496215         361.810971
      max          48.216304         379.786356
```

```
      E24PGInletTemperature  TotalPower
count          1052.000000  1052.000000
mean           42.243892   16.404868
std            4.985009    1.354001
min            33.043400   12.120520
25%            38.613957   15.477312
50%            41.437683   16.101638
75%            45.319932   17.248492
max            57.809096   20.593777
```

```
[8 rows x 37 columns]
```

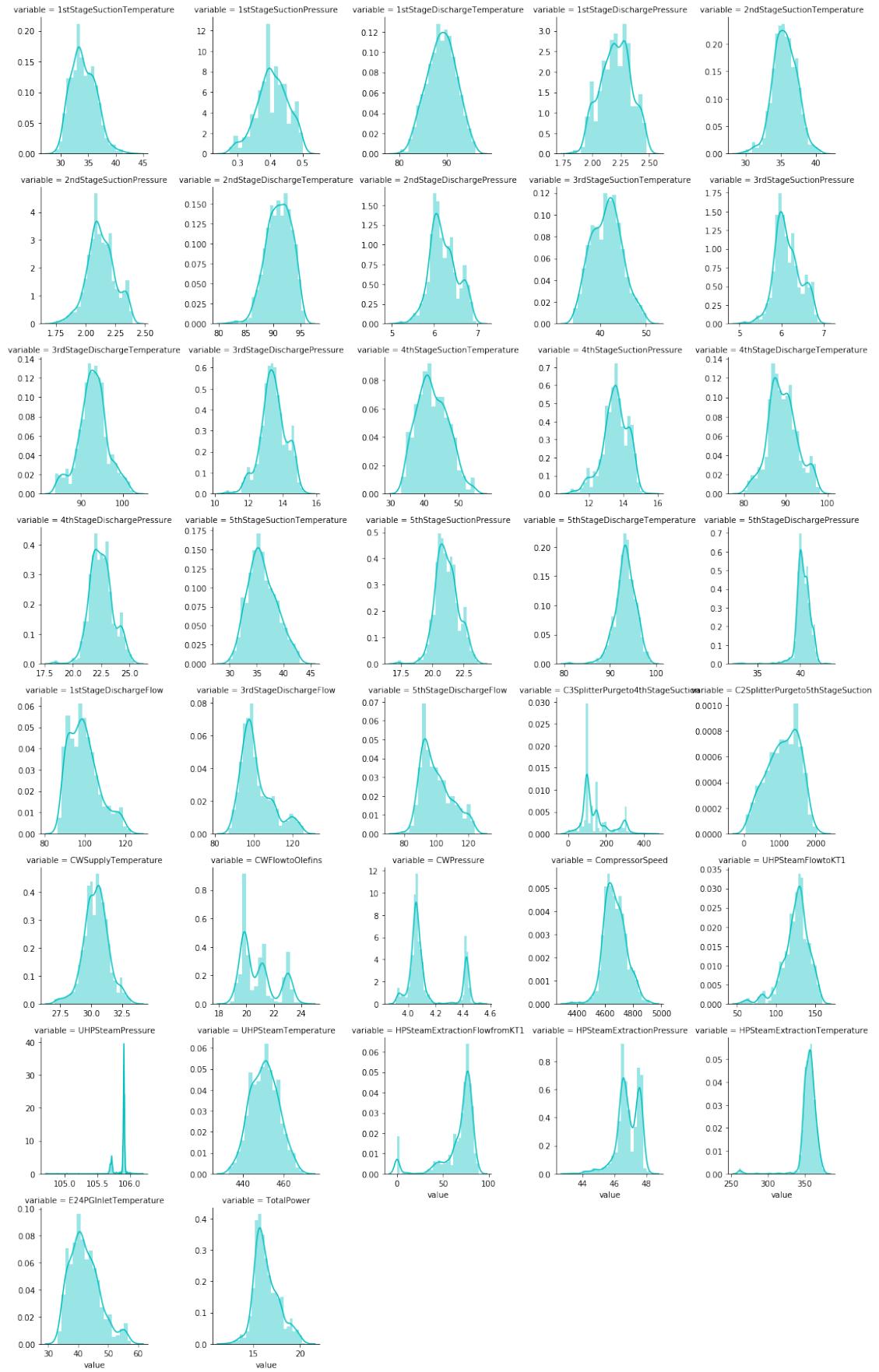
## 2 Histograms for numeric

```
[24]: num_col= df1._get_numeric_data().columns
```

```
[25]: def hist_plots(data, numeric,col_no):
        data_fig = pd.melt(data, value_vars= numeric)
        fig = sns.FacetGrid(data_fig, col = "variable", col_wrap = col_no,sharex=False, sharey = False)
        fig = fig.map(sns.distplot, "value", color = 'c')
        return fig
```

```
[26]: hist_plots(df1,num_col,5)
```

```
[26]: <seaborn.axisgrid.FacetGrid at 0xb826e80>
```



### 3 Correlation matrix

Correlation Matrix gives the relation ship between independent variables and dependent variable

```
[27]: corr = df1.corr()  
corr
```

```
[27]:
```

	1stStageSuctionTemperature \
1stStageSuctionTemperature	1.000000
1stStageSuctionPressure	0.141168
1stStageDischargeTemperature	-0.007672
1stStageDischargePressure	0.234413
2ndStageSuctionTemperature	0.699401
2ndStageSuctionPressure	0.165737
2ndStageDischargeTemperature	-0.017209
2ndStageDischargePressure	0.211601
3rdStageSuctionTemperature	0.641027
3rdStageSuctionPressure	0.214852
3rdStageDischargeTemperature	0.295165
3rdStageDischargePressure	0.283867
4thStageSuctionTemperature	0.317601
4thStageSuctionPressure	0.358797
4thStageDischargeTemperature	0.051518
4thStageDischargePressure	0.458405
5thStageSuctionTemperature	0.334865
5thStageSuctionPressure	0.478413
5thStageDischargeTemperature	0.196970
5thStageDischargePressure	0.364350
1stStageDischargeFlow	0.164147
3rdStageDischargeFlow	0.237205
5thStageDischargeFlow	0.200692
C3SplitterPurgeto4thStageSuction	-0.402859
C2SplitterPurgeto5thStageSuction	-0.431625
CWSupplyTemperature	0.559498
CWFlowtoOlefins	0.582143
CWPressure	0.413435
CompressorSpeed	0.221732
UHPSteamFlowtoKT1	0.275636
UHPSteamPressure	-0.187639
UHPSteamTemperature	-0.084481
HPSteamExtractionFlowfromKT1	0.019901
HPSteamExtractionPressure	0.414389
HPSteamExtractionTemperature	0.057769
E24PGInletTemperature	0.109235

TotalPower	0.171169
1stStageSuctionPressure	\
1stStageSuctionTemperature	0.141168
1stStageSuctionPressure	1.000000
1stStageDischargeTemperature	0.103621
1stStageDischargePressure	0.769631
2ndStageSuctionTemperature	0.331987
2ndStageSuctionPressure	0.936810
2ndStageDischargeTemperature	0.434185
2ndStageDischargePressure	0.834932
3rdStageSuctionTemperature	0.283198
3rdStageSuctionPressure	0.836773
3rdStageDischargeTemperature	0.405154
3rdStageDischargePressure	0.667343
4thStageSuctionTemperature	0.295354
4thStageSuctionPressure	0.520309
4thStageDischargeTemperature	0.228300
4thStageDischargePressure	0.610020
5thStageSuctionTemperature	0.302819
5thStageSuctionPressure	0.570757
5thStageDischargeTemperature	0.144302
5thStageDischargePressure	0.404343
1stStageDischargeFlow	0.654033
3rdStageDischargeFlow	0.626214
5thStageDischargeFlow	0.697764
C3SplitterPurgeto4thStageSuction	-0.066083
C2SplitterPurgeto5thStageSuction	-0.325755
CWSupplyTemperature	0.115196
CWFlowtoOlefins	0.050102
CWPressure	0.105540
CompressorSpeed	0.490429
UHPSteamFlowtoKT1	0.455168
UHPSteamPressure	-0.070814
UHPSteamTemperature	-0.217090
HPSteamExtractionFlowfromKT1	0.257504
HPSteamExtractionPressure	0.148129
HPSteamExtractionTemperature	-0.055860
E24PGInletTemperature	0.224945
TotalPower	0.621677
1stStageDischargeTemperature	\
1stStageSuctionTemperature	-0.007672
1stStageSuctionPressure	0.103621
1stStageDischargeTemperature	1.000000
1stStageDischargePressure	-0.040862
2ndStageSuctionTemperature	0.100243

2ndStageSuctionPressure	0.041322
2ndStageDischargeTemperature	0.474855
2ndStageDischargePressure	0.167152
3rdStageSuctionTemperature	0.015288
3rdStageSuctionPressure	0.146578
3rdStageDischargeTemperature	0.322330
3rdStageDischargePressure	0.188328
4thStageSuctionTemperature	0.237327
4thStageSuctionPressure	0.102008
4thStageDischargeTemperature	0.541149
4thStageDischargePressure	-0.099754
5thStageSuctionTemperature	0.314959
5thStageSuctionPressure	-0.144361
5thStageDischargeTemperature	0.411791
5thStageDischargePressure	0.004344
1stStageDischargeFlow	0.158000
3rdStageDischargeFlow	0.145907
5thStageDischargeFlow	0.156748
C3SplitterPurgeto4thStageSuction	0.238692
C2SplitterPurgeto5thStageSuction	0.248583
CWSupplyTemperature	0.195979
CWFlowtoOlefins	-0.386329
CWPressure	-0.206330
CompressorSpeed	0.184308
UHPSteamFlowtoKT1	-0.054271
UHPSteamPressure	0.224509
UHPSteamTemperature	0.061934
HPSteamExtractionFlowfromKT1	0.103289
HPSteamExtractionPressure	-0.171993
HPSteamExtractionTemperature	0.068533
E24PGInletTemperature	0.352288
TotalPower	0.325785

	1stStageDischargePressure \
1stStageSuctionTemperature	0.234413
1stStageSuctionPressure	0.769631
1stStageDischargeTemperature	-0.040862
1stStageDischargePressure	1.000000
2ndStageSuctionTemperature	0.414007
2ndStageSuctionPressure	0.830207
2ndStageDischargeTemperature	0.287768
2ndStageDischargePressure	0.789620
3rdStageSuctionTemperature	0.483778
3rdStageSuctionPressure	0.774854
3rdStageDischargeTemperature	0.372478
3rdStageDischargePressure	0.613094
4thStageSuctionTemperature	0.431523

4thStageSuctionPressure	0.562084
4thStageDischargeTemperature	0.207010
4thStageDischargePressure	0.646457
5thStageSuctionTemperature	0.339037
5thStageSuctionPressure	0.603355
5thStageDischargeTemperature	0.120193
5thStageDischargePressure	0.494945
1stStageDischargeFlow	0.645181
3rdStageDischargeFlow	0.631591
5thStageDischargeFlow	0.695118
C3SplitterPurgeto4thStageSuction	-0.294630
C2SplitterPurgeto5thStageSuction	-0.358145
CWSupplyTemperature	0.173266
CWFlowtoOlefins	0.225894
CWPressure	0.268147
CompressorSpeed	0.581689
UHPSteamFlowtoKT1	0.477319
UHPSteamPressure	-0.149670
UHPSteamTemperature	-0.249723
HPSteamExtractionFlowfromKT1	0.286000
HPSteamExtractionPressure	0.248509
HPSteamExtractionTemperature	-0.051586
E24PGInletTemperature	0.253732
TotalPower	0.609290

2ndStageSuctionTemperature \	
1stStageSuctionTemperature	0.699401
1stStageSuctionPressure	0.331987
1stStageDischargeTemperature	0.100243
1stStageDischargePressure	0.414007
2ndStageSuctionTemperature	1.000000
2ndStageSuctionPressure	0.370896
2ndStageDischargeTemperature	0.365512
2ndStageDischargePressure	0.374057
3rdStageSuctionTemperature	0.681781
3rdStageSuctionPressure	0.368412
3rdStageDischargeTemperature	0.407570
3rdStageDischargePressure	0.340995
4thStageSuctionTemperature	0.472394
4thStageSuctionPressure	0.346017
4thStageDischargeTemperature	0.327550
4thStageDischargePressure	0.347742
5thStageSuctionTemperature	0.443132
5thStageSuctionPressure	0.323397
5thStageDischargeTemperature	0.244647
5thStageDischargePressure	0.381283
1stStageDischargeFlow	0.324309

3rdStageDischargeFlow	0.328426
5thStageDischargeFlow	0.340198
C3SplitterPurgeto4thStageSuction	-0.287842
C2SplitterPurgeto5thStageSuction	-0.137189
CWSupplyTemperature	0.707227
CWFlowtoOlefins	0.293691
CWPressure	0.217351
CompressorSpeed	0.441901
UHPSteamFlowtoKT1	0.217596
UHPSteamPressure	-0.106692
UHPSteamTemperature	-0.083702
HPSteamExtractionFlowfromKT1	0.081467
HPSteamExtractionPressure	0.281862
HPSteamExtractionTemperature	0.084233
E24PGInletTemperature	0.303740
TotalPower	0.313414

2ndStageSuctionPressure \

1stStageSuctionTemperature	0.165737
1stStageSuctionPressure	0.936810
1stStageDischargeTemperature	0.041322
1stStageDischargePressure	0.830207
2ndStageSuctionTemperature	0.370896
2ndStageSuctionPressure	1.000000
2ndStageDischargeTemperature	0.457002
2ndStageDischargePressure	0.922069
3rdStageSuctionTemperature	0.338739
3rdStageSuctionPressure	0.927558
3rdStageDischargeTemperature	0.466670
3rdStageDischargePressure	0.764641
4thStageSuctionTemperature	0.310683
4thStageSuctionPressure	0.612464
4thStageDischargeTemperature	0.253338
4thStageDischargePressure	0.711898
5thStageSuctionTemperature	0.293198
5thStageSuctionPressure	0.673264
5thStageDischargeTemperature	0.199392
5thStageDischargePressure	0.501852
1stStageDischargeFlow	0.630039
3rdStageDischargeFlow	0.687736
5thStageDischargeFlow	0.755947
C3SplitterPurgeto4thStageSuction	-0.132595
C2SplitterPurgeto5thStageSuction	-0.423247
CWSupplyTemperature	0.102290
CWFlowtoOlefins	0.160046
CWPressure	0.167331
CompressorSpeed	0.529811

UHPSteamFlowtoKT1	0.579376
UHPSteamPressure	-0.089306
UHPSteamTemperature	-0.231475
HPSteamExtractionFlowfromKT1	0.330396
HPSteamExtractionPressure	0.250741
HPSteamExtractionTemperature	-0.043621
E24PGInletTemperature	0.194948
TotalPower	0.664264
2ndStageDischargeTemperature \	
1stStageSuctionTemperature	-0.017209
1stStageSuctionPressure	0.434185
1stStageDischargeTemperature	0.474855
1stStageDischargePressure	0.287768
2ndStageSuctionTemperature	0.365512
2ndStageSuctionPressure	0.457002
2ndStageDischargeTemperature	1.000000
2ndStageDischargePressure	0.457887
3rdStageSuctionTemperature	0.071394
3rdStageSuctionPressure	0.456597
3rdStageDischargeTemperature	0.479112
3rdStageDischargePressure	0.381498
4thStageSuctionTemperature	0.236811
4thStageSuctionPressure	0.195534
4thStageDischargeTemperature	0.481766
4thStageDischargePressure	0.165871
5thStageSuctionTemperature	0.287039
5thStageSuctionPressure	0.118749
5thStageDischargeTemperature	0.502844
5thStageDischargePressure	0.189780
1stStageDischargeFlow	0.297755
3rdStageDischargeFlow	0.309866
5thStageDischargeFlow	0.372578
C3SplitterPurgeto4thStageSuction	0.229988
C2SplitterPurgeto5thStageSuction	0.061874
CWSupplyTemperature	0.213657
CWFlowtoOlefins	-0.318767
CWPressure	-0.210474
CompressorSpeed	0.348005
UHPSteamFlowtoKT1	0.161968
UHPSteamPressure	0.246153
UHPSteamTemperature	0.014033
HPSteamExtractionFlowfromKT1	0.153182
HPSteamExtractionPressure	0.012500
HPSteamExtractionTemperature	0.101342
E24PGInletTemperature	0.319771
TotalPower	0.401877

	2ndStageDischargePressure	\
1stStageSuctionTemperature	0.211601	
1stStageSuctionPressure	0.834932	
1stStageDischargeTemperature	0.167152	
1stStageDischargePressure	0.789620	
2ndStageSuctionTemperature	0.374057	
2ndStageSuctionPressure	0.922069	
2ndStageDischargeTemperature	0.457887	
2ndStageDischargePressure	1.000000	
3rdStageSuctionTemperature	0.435628	
3rdStageSuctionPressure	0.989710	
3rdStageDischargeTemperature	0.615334	
3rdStageDischargePressure	0.918472	
4thStageSuctionTemperature	0.481931	
4thStageSuctionPressure	0.761185	
4thStageDischargeTemperature	0.456605	
4thStageDischargePressure	0.766421	
5thStageSuctionTemperature	0.474424	
5thStageSuctionPressure	0.710366	
5thStageDischargeTemperature	0.300223	
5thStageDischargePressure	0.543924	
1stStageDischargeFlow	0.708340	
3rdStageDischargeFlow	0.805876	
5thStageDischargeFlow	0.864511	
C3SplitterPurgeto4thStageSuction	-0.155754	
C2SplitterPurgeto5thStageSuction	-0.432364	
CWSupplyTemperature	0.148739	
CWFlowtoOlefins	0.192097	
CWPressure	0.217724	
CompressorSpeed	0.563452	
UHPSteamFlowtoKT1	0.644726	
UHPSteamPressure	-0.072955	
UHPSteamTemperature	-0.268258	
HPSteamExtractionFlowfromKT1	0.419191	
HPSteamExtractionPressure	0.323897	
HPSteamExtractionTemperature	-0.038107	
E24PGInletTemperature	0.365239	
TotalPower	0.806745	
	3rdStageSuctionTemperature	\
1stStageSuctionTemperature	0.641027	
1stStageSuctionPressure	0.283198	
1stStageDischargeTemperature	0.015288	
1stStageDischargePressure	0.483778	
2ndStageSuctionTemperature	0.681781	
2ndStageSuctionPressure	0.338739	

2ndStageDischargeTemperature	0.071394
2ndStageDischargePressure	0.435628
3rdStageSuctionTemperature	1.000000
3rdStageSuctionPressure	0.409503
3rdStageDischargeTemperature	0.607220
3rdStageDischargePressure	0.453904
4thStageSuctionTemperature	0.824021
4thStageSuctionPressure	0.494453
4thStageDischargeTemperature	0.434855
4thStageDischargePressure	0.564576
5thStageSuctionTemperature	0.688573
5thStageSuctionPressure	0.537534
5thStageDischargeTemperature	0.347949
5thStageDischargePressure	0.226694
1stStageDischargeFlow	0.444905
3rdStageDischargeFlow	0.501935
5thStageDischargeFlow	0.492269
C3SplitterPurgeto4thStageSuction	-0.361036
C2SplitterPurgeto5thStageSuction	-0.356780
CWSupplyTemperature	0.515835
CWFlowtoOlefins	0.570351
CWPressure	0.595674
CompressorSpeed	0.494759
UHPSteamFlowtoKT1	0.303537
UHPSteamPressure	-0.314934
UHPSteamTemperature	-0.280596
HPSteamExtractionFlowfromKT1	0.125849
HPSteamExtractionPressure	0.280320
HPSteamExtractionTemperature	-0.010786
E24PGInletTemperature	0.522456
TotalPower	0.479314
	3rdStageSuctionPressure ... CWPressure \
1stStageSuctionTemperature	0.214852 ... 0.413435
1stStageSuctionPressure	0.836773 ... 0.105540
1stStageDischargeTemperature	0.146578 ... -0.206330
1stStageDischargePressure	0.774854 ... 0.268147
2ndStageSuctionTemperature	0.368412 ... 0.217351
2ndStageSuctionPressure	0.927558 ... 0.167331
2ndStageDischargeTemperature	0.456597 ... -0.210474
2ndStageDischargePressure	0.989710 ... 0.217724
3rdStageSuctionTemperature	0.409503 ... 0.595674
3rdStageSuctionPressure	1.000000 ... 0.220395
3rdStageDischargeTemperature	0.611337 ... 0.458126
3rdStageDischargePressure	0.921002 ... 0.269353
4thStageSuctionTemperature	0.441636 ... 0.356410
4thStageSuctionPressure	0.761750 ... 0.374312

4thStageDischargeTemperature	0.410738	...	0.046352
4thStageDischargePressure	0.787714	...	0.498139
5thStageSuctionTemperature	0.434340	...	0.214141
5thStageSuctionPressure	0.739009	...	0.516061
5thStageDischargeTemperature	0.278238	...	0.038399
5thStageDischargePressure	0.569898	...	0.090769
1stStageDischargeFlow	0.674901	...	0.278106
3rdStageDischargeFlow	0.782130	...	0.362086
5thStageDischargeFlow	0.849632	...	0.316639
C3SplitterPurgeto4thStageSuction	-0.183793	...	-0.247472
C2SplitterPurgeto5thStageSuction	-0.473909	...	-0.492187
CWSupplyTemperature	0.144426	...	0.082384
CWFlowtoOlefins	0.219728	...	0.790773
CWPressure	0.220395	...	1.000000
CompressorSpeed	0.517014	...	0.275043
UHPSteamFlowtoKT1	0.669973	...	0.373674
UHPSteamPressure	-0.089645	...	-0.486830
UHPSteamTemperature	-0.238366	...	-0.088633
HPSteamExtractionFlowfromKT1	0.399752	...	0.021404
HPSteamExtractionPressure	0.341360	...	0.192641
HPSteamExtractionTemperature	-0.033947	...	-0.001378
E24PGInletTemperature	0.318035	...	0.054917
TotalPower	0.768394	...	0.252765

	CompressorSpeed	UHPSteamFlowtoKT1	\
1stStageSuctionTemperature	0.221732	0.275636	
1stStageSuctionPressure	0.490429	0.455168	
1stStageDischargeTemperature	0.184308	-0.054271	
1stStageDischargePressure	0.581689	0.477319	
2ndStageSuctionTemperature	0.441901	0.217596	
2ndStageSuctionPressure	0.529811	0.579376	
2ndStageDischargeTemperature	0.348005	0.161968	
2ndStageDischargePressure	0.563452	0.644726	
3rdStageSuctionTemperature	0.494759	0.303537	
3rdStageSuctionPressure	0.517014	0.669973	
3rdStageDischargeTemperature	0.486353	0.488066	
3rdStageDischargePressure	0.419251	0.735187	
4thStageSuctionTemperature	0.434645	0.207810	
4thStageSuctionPressure	0.356741	0.709708	
4thStageDischargeTemperature	0.472608	0.075553	
4thStageDischargePressure	0.376745	0.769693	
5thStageSuctionTemperature	0.343193	0.177896	
5thStageSuctionPressure	0.315842	0.771192	
5thStageDischargeTemperature	0.233641	0.186844	
5thStageDischargePressure	0.241421	0.567850	
1stStageDischargeFlow	0.689766	0.336263	
3rdStageDischargeFlow	0.660622	0.511459	

5thStageDischargeFlow	0.654932	0.607127
C3SplitterPurgeto4thStageSuction	-0.218012	-0.094103
C2SplitterPurgeto5thStageSuction	-0.072465	-0.625432
CWSupplyTemperature	0.053548	0.050040
CWFlowtoOlefins	0.155935	0.493764
CWPressure	0.275043	0.373674
CompressorSpeed	1.000000	0.198302
UHPSteamFlowtoKT1	0.198302	1.000000
UHPSteamPressure	-0.165147	-0.148030
UHPSteamTemperature	-0.127680	-0.112471
HPSteamExtractionFlowfromKT1	0.121442	0.588734
HPSteamExtractionPressure	-0.009479	0.582360
HPSteamExtractionTemperature	-0.011960	0.108449
E24PGInletTemperature	0.305513	0.024882
TotalPower	0.709981	0.492128
1stStageSuctionTemperature	UHPSteamPressure	UHPSteamTemperature \
1stStageSuctionPressure	-0.187639	-0.084481
1stStageDischargeTemperature	-0.070814	-0.217090
1stStageDischargePressure	0.224509	0.061934
2ndStageSuctionTemperature	-0.149670	-0.249723
2ndStageSuctionPressure	-0.106692	-0.083702
2ndStageDischargeTemperature	-0.089306	-0.231475
2ndStageDischargePressure	0.246153	0.014033
3rdStageSuctionTemperature	-0.072955	-0.268258
3rdStageSuctionPressure	-0.314934	-0.280596
3rdStageDischargeTemperature	-0.089645	-0.238366
3rdStageDischargePressure	-0.279734	-0.055197
4thStageSuctionTemperature	-0.068103	-0.307194
4thStageSuctionPressure	-0.093473	-0.369571
4thStageDischargeTemperature	-0.168906	-0.237568
4thStageDischargePressure	0.112187	-0.163140
5thStageSuctionTemperature	-0.249363	-0.315023
5thStageSuctionPressure	0.055029	-0.410184
5thStageDischargeTemperature	-0.273388	-0.281430
5thStageDischargePressure	0.156917	-0.147680
1stStageDischargeFlow	-0.006564	-0.173402
3rdStageDischargeFlow	-0.220097	-0.151362
5thStageDischargeFlow	-0.248542	-0.180092
C3SplitterPurgeto4thStageSuction	-0.154596	-0.279550
C2SplitterPurgeto5thStageSuction	0.237895	0.032113
CWSupplyTemperature	0.301163	0.197242
CWFlowtoOlefins	0.036152	-0.018181
CWPressure	-0.486329	-0.111517
CompressorSpeed	-0.486830	-0.088633
UHPSteamFlowtoKT1	-0.165147	-0.127680
	-0.148030	-0.112471

UHPSteamPressure	1.000000	-0.127572
UHPSteamTemperature	-0.127572	1.000000
HPSteamExtractionFlowfromKT1	0.327228	-0.117567
HPSteamExtractionPressure	0.146666	-0.323983
HPSteamExtractionTemperature	0.330751	0.354137
E24PGInletTemperature	0.150343	-0.345512
TotalPower	-0.065722	-0.235469
	HPSteamExtractionFlowfromKT1 \	
1stStageSuctionTemperature	0.019901	
1stStageSuctionPressure	0.257504	
1stStageDischargeTemperature	0.103289	
1stStageDischargePressure	0.286000	
2ndStageSuctionTemperature	0.081467	
2ndStageSuctionPressure	0.330396	
2ndStageDischargeTemperature	0.153182	
2ndStageDischargePressure	0.419191	
3rdStageSuctionTemperature	0.125849	
3rdStageSuctionPressure	0.399752	
3rdStageDischargeTemperature	0.201579	
3rdStageDischargePressure	0.461721	
4thStageSuctionTemperature	0.255689	
4thStageSuctionPressure	0.447949	
4thStageDischargeTemperature	0.218306	
4thStageDischargePressure	0.356338	
5thStageSuctionTemperature	0.262321	
5thStageSuctionPressure	0.323953	
5thStageDischargeTemperature	0.226511	
5thStageDischargePressure	0.312255	
1stStageDischargeFlow	0.208696	
3rdStageDischargeFlow	0.300621	
5thStageDischargeFlow	0.404084	
C3SplitterPurgeto4thStageSuction	0.080281	
C2SplitterPurgeto5thStageSuction	-0.161144	
CWSupplyTemperature	0.049854	
CWFlowtoOlefins	0.031951	
CWPressure	0.021404	
CompressorSpeed	0.121442	
UHPSteamFlowtoKT1	0.588734	
UHPSteamPressure	0.327228	
UHPSteamTemperature	-0.117567	
HPSteamExtractionFlowfromKT1	1.000000	
HPSteamExtractionPressure	0.342632	
HPSteamExtractionTemperature	0.341049	
E24PGInletTemperature	0.241795	
TotalPower	0.386332	

	HPSteamExtractionPressure \
1stStageSuctionTemperature	0.414389
1stStageSuctionPressure	0.148129
1stStageDischargeTemperature	-0.171993
1stStageDischargePressure	0.248509
2ndStageSuctionTemperature	0.281862
2ndStageSuctionPressure	0.250741
2ndStageDischargeTemperature	0.012500
2ndStageDischargePressure	0.323897
3rdStageSuctionTemperature	0.280320
3rdStageSuctionPressure	0.341360
3rdStageDischargeTemperature	0.169140
3rdStageDischargePressure	0.498524
4thStageSuctionTemperature	0.147265
4thStageSuctionPressure	0.514439
4thStageDischargeTemperature	-0.029181
4thStageDischargePressure	0.567899
5thStageSuctionTemperature	0.166964
5thStageSuctionPressure	0.574749
5thStageDischargeTemperature	0.149059
5thStageDischargePressure	0.548415
1stStageDischargeFlow	-0.032156
3rdStageDischargeFlow	0.140455
5thStageDischargeFlow	0.250815
C3SplitterPurgeto4thStageSuction	-0.183823
C2SplitterPurgeto5thStageSuction	-0.515526
CWSupplyTemperature	0.134297
CWFlowtoOlefins	0.459288
CWPressure	0.192641
CompressorSpeed	-0.009479
UHPSteamFlowtoKT1	0.582360
UHPSteamPressure	0.146666
UHPSteamTemperature	-0.323983
HPSteamExtractionFlowfromKT1	0.342632
HPSteamExtractionPressure	1.000000
HPSteamExtractionTemperature	0.100955
E24PGInletTemperature	0.030580
TotalPower	0.152691

	HPSteamExtractionTemperature \
1stStageSuctionTemperature	0.057769
1stStageSuctionPressure	-0.055860
1stStageDischargeTemperature	0.068533
1stStageDischargePressure	-0.051586
2ndStageSuctionTemperature	0.084233
2ndStageSuctionPressure	-0.043621
2ndStageDischargeTemperature	0.101342

2ndStageDischargePressure	-0.038107
3rdStageSuctionTemperature	-0.010786
3rdStageSuctionPressure	-0.033947
3rdStageDischargeTemperature	0.038939
3rdStageDischargePressure	-0.032126
4thStageSuctionTemperature	-0.046648
4thStageSuctionPressure	-0.015274
4thStageDischargeTemperature	0.031778
4thStageDischargePressure	-0.055635
5thStageSuctionTemperature	-0.063756
5thStageSuctionPressure	-0.044325
5thStageDischargeTemperature	0.087975
5thStageDischargePressure	-0.020683
1stStageDischargeFlow	-0.096434
3rdStageDischargeFlow	-0.066581
5thStageDischargeFlow	-0.066545
C3SplitterPurgeto4thStageSuction	0.071374
C2SplitterPurgeto5thStageSuction	0.013538
CWSupplyTemperature	0.123952
CWFlowtoOlefins	0.006618
CWPressure	-0.001378
CompressorSpeed	-0.011960
UHPSteamFlowtoKT1	0.108449
UHPSteamPressure	0.330751
UHPSteamTemperature	0.354137
HPSteamExtractionFlowfromKT1	0.341049
HPSteamExtractionPressure	0.100955
HPSteamExtractionTemperature	1.000000
E24PGInletTemperature	-0.066252
TotalPower	0.007597

	E24PGInletTemperature	TotalPower
1stStageSuctionTemperature	0.109235	0.171169
1stStageSuctionPressure	0.224945	0.621677
1stStageDischargeTemperature	0.352288	0.325785
1stStageDischargePressure	0.253732	0.609290
2ndStageSuctionTemperature	0.303740	0.313414
2ndStageSuctionPressure	0.194948	0.664264
2ndStageDischargeTemperature	0.319771	0.401877
2ndStageDischargePressure	0.365239	0.806745
3rdStageSuctionTemperature	0.522456	0.479314
3rdStageSuctionPressure	0.318035	0.768394
3rdStageDischargeTemperature	0.390526	0.662108
3rdStageDischargePressure	0.376037	0.759285
4thStageSuctionTemperature	0.871080	0.540366
4thStageSuctionPressure	0.302973	0.632415
4thStageDischargeTemperature	0.723102	0.544685

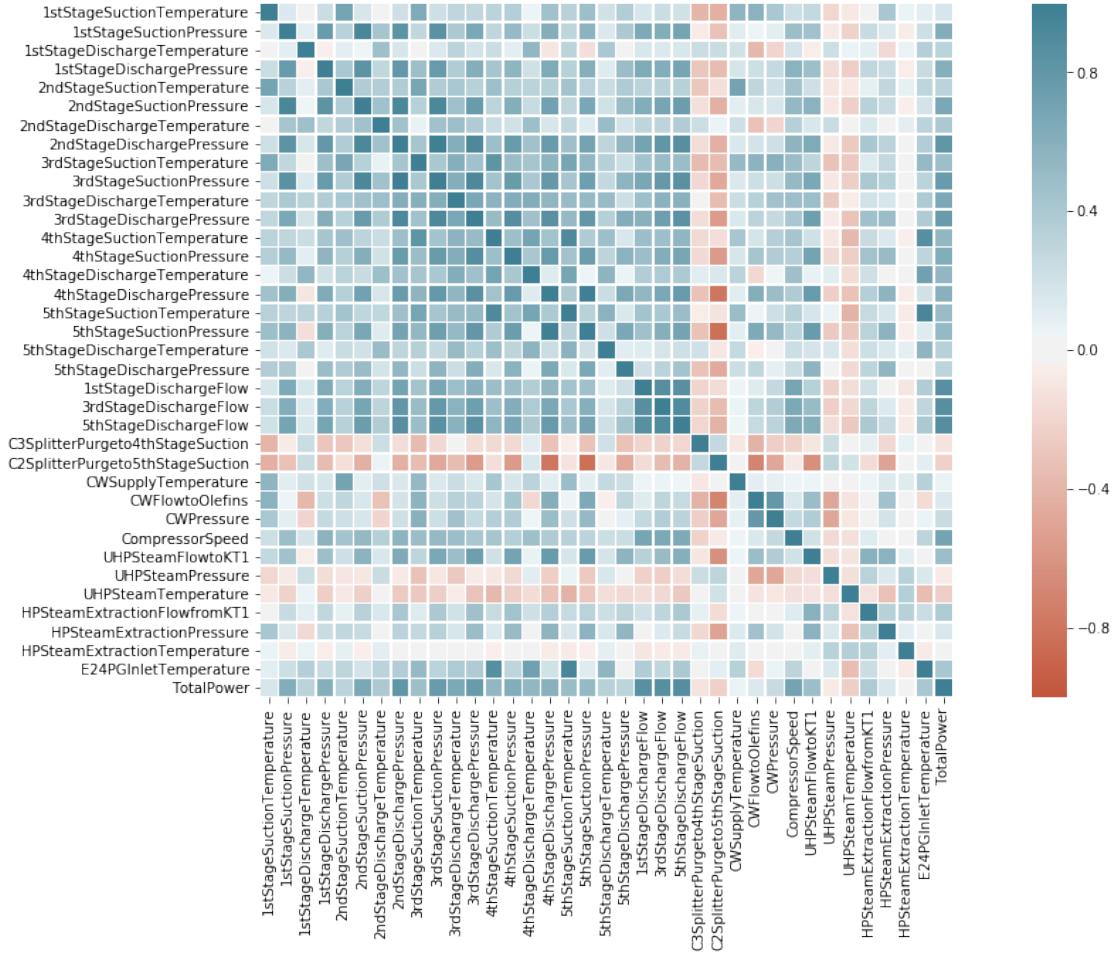
4thStageDischargePressure	0.198147	0.598248
5thStageSuctionTemperature	0.939058	0.504974
5thStageSuctionPressure	0.118165	0.522238
5thStageDischargeTemperature	0.554888	0.342142
5thStageDischargePressure	0.038373	0.305859
1stStageDischargeFlow	0.374282	0.855854
3rdStageDischargeFlow	0.290867	0.867900
5thStageDischargeFlow	0.399253	0.879583
C3SplitterPurgeto4thStageSuction	0.020030	-0.114082
C2SplitterPurgeto5thStageSuction	0.103290	-0.218828
CWSupplyTemperature	0.340670	0.067238
CWFlowtoOlefins	-0.159572	0.144430
CWPressure	0.054917	0.252765
CompressorSpeed	0.305513	0.709981
UHPSteamFlowtoKT1	0.024882	0.492128
UHPSteamPressure	0.150343	-0.065722
UHPSteamTemperature	-0.345512	-0.235469
HPSteamExtractionFlowfromKT1	0.241795	0.386332
HPSteamExtractionPressure	0.030580	0.152691
HPSteamExtractionTemperature	-0.066252	0.007597
E24PGInletTemperature	1.000000	0.416587
TotalPower	0.416587	1.000000

[37 rows x 37 columns]

```
[28]: def corr_plot(data_corr):
    palette = sns.diverging_palette(20, 220, as_cmap=True)
    fig, (ax) = plt.subplots(1, 1, figsize=(20,9))
    fig = sns.heatmap(data_corr,
                       ax=ax, # Axes in which to draw the plot, otherwise use the
                       # currently-active Axes.
                       vmin = -1.0,
                       vmax = 1.0,
                       cmap=palette, # Color Map.
                       square=True, # If True, set the Axes aspect to "equal" so each
                       # cell will be square-shaped.
                       annot=False,
                       fmt='.{2f}', # String formatting code to use when adding
                       # annotations.
                       annot_kws={"size": 14},
                       linewidths=.5)
    return fig
```

```
[29]: corr_plot(corr)
```

```
[29]: <matplotlib.axes._subplots.AxesSubplot at 0xd272d68>
```

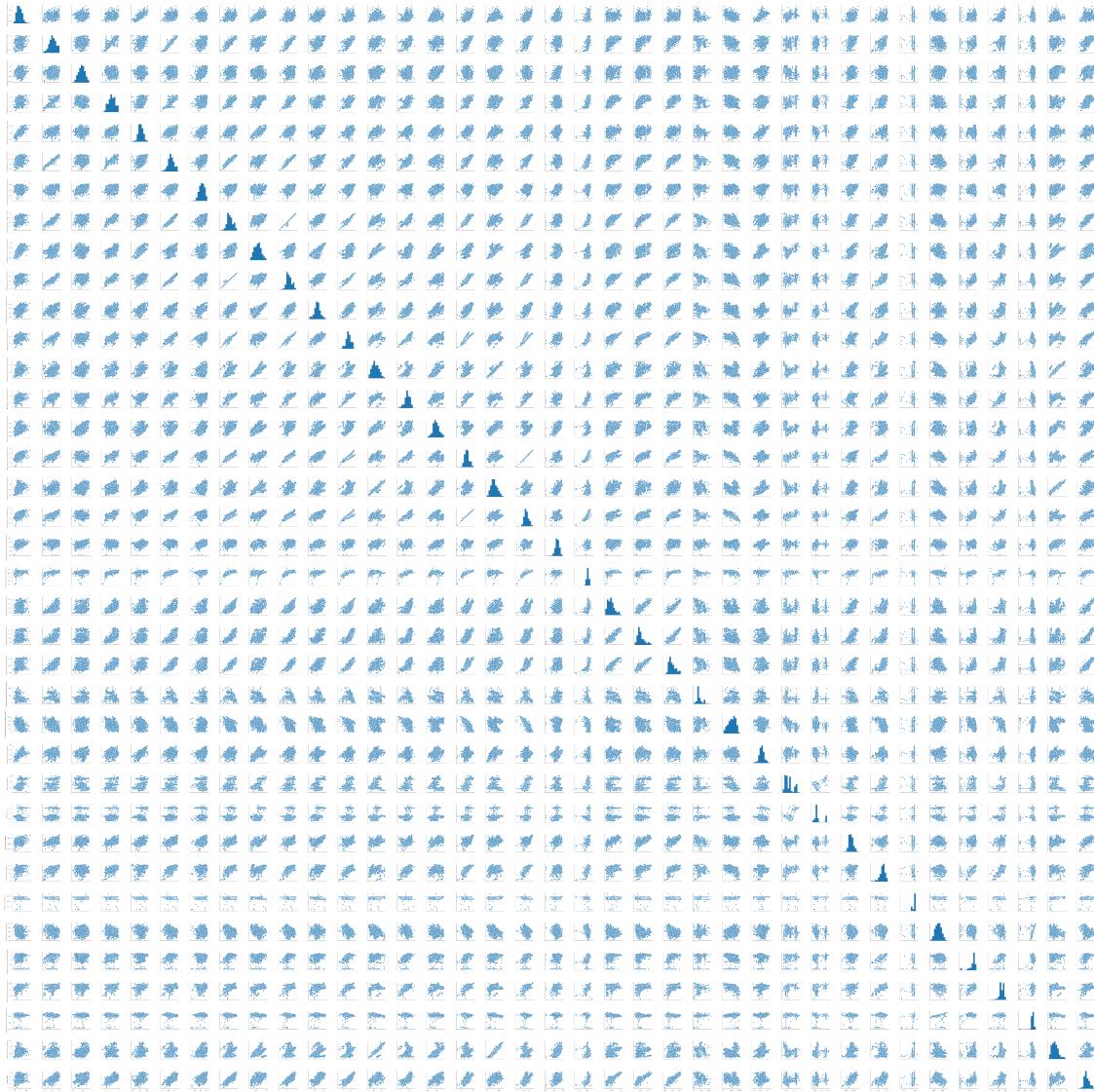


### 3.0.1 Pair Plot

Pairplots gives the relation ship between single variable and also between two variables

```
[30]: # Seaborn visualization library
import seaborn as sns
# Create the default pairplot
sns.pairplot(df1)
```

```
[30]: <seaborn.axisgrid.PairGrid at 0xd78c668>
```



## 4 Model development

### 4.0.1 Train and test split

For any machine learning model, we always need train and test datasets. We will be building the model on the train dataset and test the performance on the test dataset.

```
[ ]: from sklearn.model_selection import train_test_split # splitting the dataset into train and test
y=df1.TotalPower
x=df1.drop('TotalPower',axis=1)

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2)
```

## 4.1 Pre processing

```
[32]: numeric = list(x_train.select_dtypes(include=['float64','int64']).columns)

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaler.fit(x_train[numeric])
scaler.transform(x_train[numeric])
x_train = pd.DataFrame(data = scaler.transform(x_train[numeric]), columns = ↴
    ↪numeric)
x_train.head()
```

```
[32]:   1stStageSuctionTemperature  1stStageSuctionPressure \
0           1.682394            -0.360194
1           1.019781            -1.195528
2          -1.422716            1.453704
3          -0.409894            1.407180
4           0.124901            -0.568489

   1stStageDischargeTemperature  1stStageDischargePressure \
0          -0.954903             0.729175
1           0.329442            -0.832738
2           0.663584            -1.734957
3          -0.304849             1.536364
4          -0.583558            -0.268177

   2ndStageSuctionTemperature  2ndStageSuctionPressure \
0           0.265775            -0.107004
1           0.862834            -0.897796
2          -1.350218             0.940155
3           0.909419             1.670912
4          -0.568817            -0.265087

   2ndStageDischargeTemperature  2ndStageDischargePressure \
0          -1.407548             0.217723
1          -0.295067            -0.885116
2           0.551823            -0.619771
3           0.948326             1.077354
4          -0.077157            -0.295904

   3rdStageSuctionTemperature  3rdStageSuctionPressure ...  CWFflowtoOlefins \
0           1.302189             0.216255 ...           1.450350
1           0.440012            -0.857817 ...           0.389374
2          -1.553466            -0.478201 ...          -0.606966
3           0.372364             1.163350 ...           0.168090
4          -0.590240            -0.221751 ...           0.317252
```

```

      CWPressure  CompressorSpeed  UHPSteamFlowtoKT1  UHPSteamPressure  \
0    1.918443      -0.070404        0.414727      -1.162822
1   -0.587993      -0.251669        0.175895       0.347895
2   -0.513216      -0.898008        0.819324       0.490559
3   -0.324715      0.599911        0.742029       0.352997
4   -1.213771      -0.367939        0.168487       0.415342

      UHPSteamTemperature  HPSteamExtractionFlowfromKT1  \
0            -0.254657      -0.019582
1          -1.788837       0.057160
2           0.352329       0.731283
3          -1.191622       0.602085
4           1.089855       0.367286

      HPSteamExtractionPressure  HPSteamExtractionTemperature  \
0            -0.387101       0.029115
1           1.004033      -0.612115
2          -0.468431       0.062245
3           1.280877      -0.444283
4           0.765613       0.774115

      E24PGInletTemperature
0            0.728212
1           -0.593009
2           -0.854374
3            0.467983
4           -0.277159

[5 rows x 36 columns]

```

```
[33]: x_test = pd.DataFrame(data = scaler.transform(x_test[numERIC]), columns = 
                           ↪numERIC)
x_test.head()
```

```
[33]: 1stStageSuctionTemperature  1stStageSuctionPressure  \
0            -0.107406       1.214029
1             0.383982      -0.404278
2            2.256321      -0.344425
3            -0.467744      -0.905670
4            -0.130030      -0.640119

      1stStageDischargeTemperature  1stStageDischargePressure  \
0              1.049673       1.430645
1            -0.892708       1.634579
2            -0.440087      -1.039320
3            -0.106388      -1.542563
```

4	0.416289	-0.622971	
	2ndStageSuctionTemperature	2ndStageSuctionPressure	\
0	0.216063	0.698998	
1	1.806142	0.118131	
2	-0.023952	-0.783609	
3	0.161437	-1.372504	
4	-0.118658	-0.692132	
	2ndStageDischargeTemperature	2ndStageDischargePressure	\
0	1.144235	0.638080	
1	-1.832977	-0.011019	
2	-1.200412	-0.742226	
3	-0.829344	-1.076548	
4	-0.241405	-0.475219	
	3rdStageSuctionTemperature	3rdStageSuctionPressure	... CWFflowtoOlefins \
0	0.961607	0.464581	... -1.249437
1	2.272773	-0.033197	... 1.504541
2	0.210401	-0.725551	... 1.646051
3	-0.604720	-1.157402	... -0.785038
4	-0.599126	-0.431350	... -0.816383
	CWPressure	CompressorSpeed	UHPSteamFlowtoKT1 UHPSteamPressure \
0	-0.228748	0.967328	-0.211789 0.563557
1	1.781540	0.893514	1.132333 -1.238007
2	1.765406	-0.904208	-0.805725 -1.244776
3	-0.651010	-0.497596	-1.189178 0.486012
4	-0.447056	-0.779405	-0.188562 0.433910
	UHPSteamTemperature	HPSteamExtractionFlowfromKT1	\
0	-1.121840	0.672853	
1	0.844791	0.872693	
2	-1.075208	-3.359332	
3	2.106814	0.259322	
4	2.351745	-0.162101	
	HPSteamExtractionPressure	HPSteamExtractionTemperature	\
0	-0.016629	-0.548677	
1	-0.014047	0.467097	
2	0.574499	-0.285671	
3	-0.267235	1.312885	
4	0.584727	1.522358	
	E24PGInletTemperature		
0	2.348736		
1	-1.351580		

```
2          -0.488801
3           0.856027
4          -0.503329
```

[5 rows x 36 columns]

#### 4.1.1 XGboosting

##### Building Model

```
[34]: import xgboost as xgb
```

```
[35]: xg_reg = xgb.XGBRegressor(objective ='reg:linear', colsample_bytree = 0.3,
   ~learning_rate = 0.1,
   max_depth = 5, alpha = 10, n_estimators = 40)
```

```
[36]: xg_reg.fit(x_train,y_train)
```

[15:41:25] WARNING: src/objective/regression\_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
Series.base is deprecated and will be removed in a future version  
if getattr(data, 'base', None) is not None and \

```
[36]: XGBRegressor(alpha=10, base_score=0.5, booster='gbtree', colsample_bylevel=1,
   colsample_bynode=1, colsample_bytree=0.3, gamma=0,
   importance_type='gain', learning_rate=0.1, max_delta_step=0,
   max_depth=5, min_child_weight=1, missing=None, n_estimators=40,
   n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
   reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
   silent=None, subsample=1, verbosity=1)
```

Here we do the usual, use the trained model to make predictions on the test dataset.

```
[37]: y_pred = xg_reg.predict(x_test)
```

## 4.2 Model Performance Evaluation

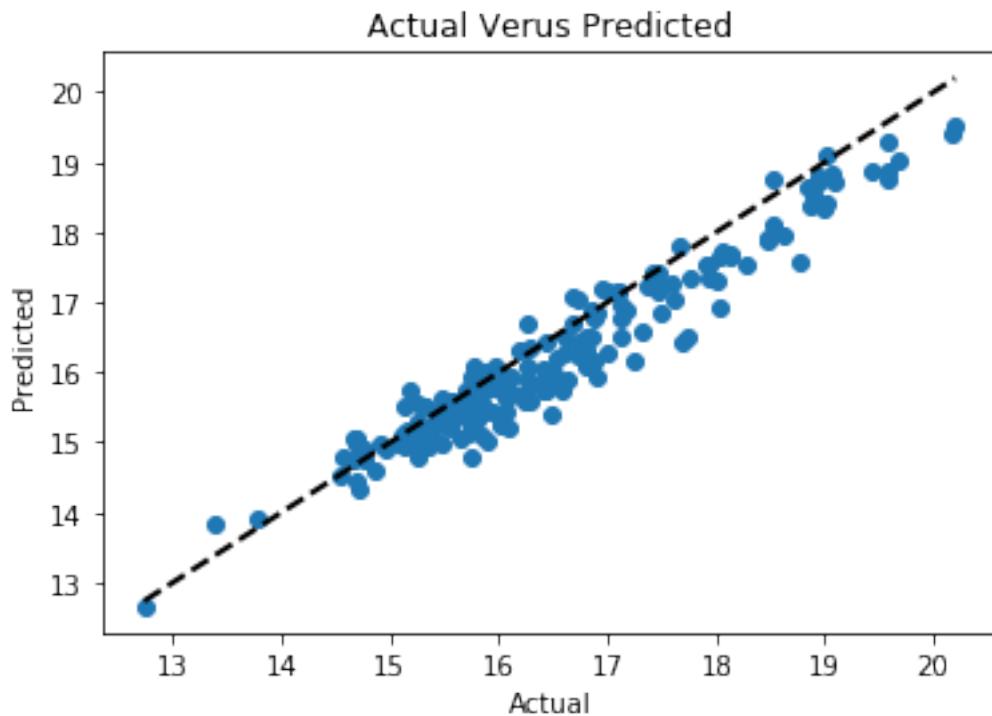
Time to put the model to the test! Let's evaluate how our model has performed with its predictions on the test data

```
[38]: from sklearn.metrics import mean_squared_error

xgb_mse = mean_squared_error(y_test, y_pred)
xgb_rmse = np.sqrt(xgb_mse)
xgb_rmse
```

```
[38]: 0.4386666503229113
```

```
[39]: fig, ax = plt.subplots()
plt.scatter(y_test,y_pred)
ax.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=2)
ax.set_xlabel('Actual')
ax.set_ylabel('Predicted')
ax.set_title("Actual Verus Predicted")
plt.show()
```



#### 4.2.1 Cross validation scores

```
[41]: from sklearn.model_selection import cross_val_score

xgb_scores = cross_val_score(xg_reg, x, y,
                             scoring="neg_mean_squared_error", cv=10)
xgb_rmse_scores = np.sqrt(-xgb_scores)
```

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
 Series.base is deprecated and will be removed in a future version  
 if getattr(data, 'base', None) is not None and \  
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
 Series.base is deprecated and will be removed in a future version  
 if getattr(data, 'base', None) is not None and \  
[15:41:26] WARNING: src/objective/regression\_obj.cu:152: reg:linear is now

```
deprecated in favor of reg:squarederror.  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
  
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
Series.base is deprecated and will be removed in a future version  
    if getattr(data, 'base', None) is not None and \  
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
Series.base is deprecated and will be removed in a future version  
    if getattr(data, 'base', None) is not None and \  
  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
  
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
Series.base is deprecated and will be removed in a future version  
    if getattr(data, 'base', None) is not None and \  
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
Series.base is deprecated and will be removed in a future version  
    if getattr(data, 'base', None) is not None and \  
  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
  
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
Series.base is deprecated and will be removed in a future version  
    if getattr(data, 'base', None) is not None and \  
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
Series.base is deprecated and will be removed in a future version  
    if getattr(data, 'base', None) is not None and \  
  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
[15:41:27] WARNING: src/objective/regression_obj.cu:152: reg:linear is now  
deprecated in favor of reg:squarederror.  
  
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:  
Series.base is deprecated and will be removed in a future version  
    if getattr(data, 'base', None) is not None and \  

```

```
[42]: def display_scores(scores):
    print("Scores:", scores)
    print("The mean cross-validation score and the 95% confidence interval: ")
    print("Accuracy: %.2f (+/- %.2f)" % (scores.mean(), scores.std() * 2))

display_scores(xgb_rmse_scores)
```

Scores: [0.67608098 1.00137683 1.05684433 1.02094805 0.67531627 0.41781056  
0.28047138 0.38708964 0.50012679 0.72822538]  
The mean cross-validation score and the 95% confidence interval:  
Accuracy: 0.67 (+/- 0.53)

**Hyperparameter tuning** We will try and adjust following hyperparameters:

n\_estimators = number of trees in the forest  
subsample = max number of features considered for splitting a node  
max\_depth = max number of levels in each decision tree  
bootstrap = method for sampling data points (with or without replacement)  
colsample\_bytree = min number of data points placed in a node before the node is split  
min\_child\_weight = min number of data points allowed in a leaf node

### Random Hyperparameter Grid

```
[43]: from sklearn.model_selection import RandomizedSearchCV

n_estimators = [int(x) for x in np.linspace(start = 10, stop = 50, num = 10)]

max_depth = [int(x) for x in np.linspace(10, 120, num = 5)]

min_child_weight = [2, 5, 10]

learning_rate = [float(x) for x in np.linspace(0.01, 1, num = 4)]

subsample = [float(x) for x in np.linspace(0.1, 1, num = 4)]
colsample_bytree = [float(x) for x in np.linspace(0.1, 1, num = 4)]

random_grid = {'n_estimators': n_estimators,
               'max_depth': max_depth,
               'min_child_weight': min_child_weight,
               'learning_rate': learning_rate,
               'subsample': subsample,
               'colsample_bytree': colsample_bytree }

print(random_grid)

{'n_estimators': [10, 14, 18, 23, 27, 32, 36, 41, 45, 50], 'max_depth': [10, 37, 65, 92, 120], 'min_child_weight': [2, 5, 10], 'learning_rate': [0.01, 0.34, 0.67, 1.0], 'subsample': [0.1, 0.4, 0.7, 1.0], 'colsample_bytree': [0.1, 0.4, 0.7, 1.0]}
```

The most important arguments in RandomizedSearchCV are n\_iter, which controls the number of different combinations to try, and cv which is the number of folds to use for cross validation (we use 200 and 3 respectively). More iterations will cover a wider search space and more cv folds reduces the chances of overfitting, but raising each will increase the run time.

```
[44]: model = xgb.XGBRegressor()
xgb_reg_random = RandomizedSearchCV(estimator = model,
                                      param_distributions = random_grid,
                                      n_iter = 200, cv = 3,
                                      verbose=2, random_state=101,
                                      n_jobs = -1, iid = True)

xgb_reg_random.fit(x_train, y_train)
```

```
Fitting 3 folds for each of 200 candidates, totalling 600 fits

[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done 33 tasks      | elapsed:   3.9s
[Parallel(n_jobs=-1)]: Done 479 tasks      | elapsed:  17.5s
[Parallel(n_jobs=-1)]: Done 600 out of 600 | elapsed:  20.5s finished
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
  if getattr(data, 'base', None) is not None and \
[15:48:23] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.
```

```
[44]: RandomizedSearchCV(cv=3, error_score='raise-deprecating',
                         estimator=XGBRegressor(base_score=0.5, booster='gbtree',
                         colsample_bylevel=1,
                         colsample_bynode=1, colsample_bytree=1, gamma=0,
                         importance_type='gain', learning_rate=0.1, max_delta_step=0,
                         max_depth=3, min_child_weight=1, missing=None, n_estimators=100,
                         n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
                         reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
                         silent=None, subsample=1, verbosity=1),
                         fit_params=None, iid=True, n_iter=200, n_jobs=-1,
                         param_distributions={'n_estimators': [10, 14, 18, 23, 27, 32, 36, 41,
                         45, 50], 'max_depth': [10, 37, 65, 92, 120], 'min_child_weight': [2, 5, 10],
                         'learning_rate': [0.01, 0.34, 0.67, 1.0], 'subsample': [0.1, 0.4, 0.7, 1.0]},
                         'colsample_bytree': [0.1, 0.4, 0.7, 1.0]}, pre_dispatch='2*n_jobs', random_state=101, refit=True,
                         return_train_score='warn', scoring=None, verbose=2)
```

```
[45]: xgb_reg_random.best_params_
```

```
[45]: {'subsample': 0.7,
       'n_estimators': 41,
       'min_child_weight': 10,
```

```

'max_depth': 120,
'learning_rate': 0.34,
'colsample_bytree': 1.0}

[46]: xgb_scores_random = cross_val_score(xgb_reg_random.best_estimator_,
                                         x_train, y_train,
                                         scoring="neg_mean_squared_error", cv=10)

xgb_rmse_scores_random = np.sqrt(-xgb_scores_random)

display_scores(xgb_rmse_scores_random)

```

[15:49:38] WARNING: src/objective/regression\_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version  
if getattr(data, 'base', None) is not None and \

[15:49:39] WARNING: src/objective/regression\_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version  
if getattr(data, 'base', None) is not None and \

[15:49:39] WARNING: src/objective/regression\_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version  
if getattr(data, 'base', None) is not None and \

[15:49:39] WARNING: src/objective/regression\_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version  
if getattr(data, 'base', None) is not None and \

[15:49:39] WARNING: src/objective/regression\_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version  
if getattr(data, 'base', None) is not None and \

[15:49:40] WARNING: src/objective/regression\_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning: Series.base is deprecated and will be removed in a future version  
if getattr(data, 'base', None) is not None and \

```

[15:49:40] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:49:40] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:49:40] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:49:41] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.
Scores: [0.39328818 0.49111933 0.41919351 0.43900958 0.349456   0.39684117
         0.33727915 0.37024969 0.30823367 0.33607725]
The mean cross-validation score and the 95% confidence interval:
Accuracy: 0.38 (+/- 0.11)

```

#### 4.2.2 Grid Search with Cross Validation

Random search allowed us to narrow down the range for each hyperparameter. Now, we tune the parameters between the range got from the randomised hyperparameters

```
[49]: from sklearn.model_selection import GridSearchCV

param_grid = {
    'subsample': [0.4, 0.7, 1.0],
    'n_estimators': [20, 40, 50],
    'min_child_weight': [3, 8, 11],
    'max_depth': [60, 90, 125],
    'learning_rate': [0.01, 0.05, 0.1],
    'colsample_bytree': [0.1, 0.3, 0.5]}

model = xgb.XGBRegressor()

grid_search = GridSearchCV(estimator = model, param_grid = param_grid,
                           cv = 3, n_jobs = -1, verbose = 2)

print(param_grid)
```

```
{'subsample': [0.4, 0.7, 1.0], 'n_estimators': [20, 40, 50], 'min_child_weight': [3, 8, 11], 'max_depth': [60, 90, 125], 'learning_rate': [0.01, 0.05, 0.1], 'colsample_bytree': [0.1, 0.3, 0.5]}
```

```
[50]: grid_search.fit(x_train, y_train)

grid_search.best_params_
```

Fitting 3 folds for each of 729 candidates, totalling 2187 fits

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  96 tasks      | elapsed:   1.8s
[Parallel(n_jobs=-1)]: Done 580 tasks      | elapsed:   7.2s
[Parallel(n_jobs=-1)]: Done 1392 tasks     | elapsed:  20.1s
[Parallel(n_jobs=-1)]: Done 2120 tasks     | elapsed:  34.9s
[Parallel(n_jobs=-1)]: Done 2180 out of 2187 | elapsed:  36.4s remaining:
0.0s
```

```
[15:58:15] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.
```

```
[Parallel(n_jobs=-1)]: Done 2187 out of 2187 | elapsed:  36.6s finished
```

```
[50]: {'colsample_bytree': 0.5,
        'learning_rate': 0.1,
        'max_depth': 60,
        'min_child_weight': 8,
        'n_estimators': 50,
        'subsample': 1.0}
```

```
[51]: xgb_scores_grid = cross_val_score(grid_search.best_estimator_,
                                         x_train, y_train,
                                         scoring="neg_mean_squared_error", cv=10)

xgb_rmse_scores_grid = np.sqrt(-xgb_scores_grid)

display_scores(xgb_rmse_scores_grid)
```

```
[15:59:39] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.
```

```
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
  if getattr(data, 'base', None) is not None and \
```

```
[15:59:39] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.
```

```
E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
  if getattr(data, 'base', None) is not None and \
```

```
[15:59:40] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:59:40] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:59:40] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:59:41] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:59:41] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:59:41] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

E:\nikhitha\Anaconda\lib\site-packages\xgboost\core.py:587: FutureWarning:
Series.base is deprecated and will be removed in a future version
    if getattr(data, 'base', None) is not None and \

[15:59:42] WARNING: src/objective/regression_obj.cu:152: reg:linear is now
deprecated in favor of reg:squarederror.

Scores: [0.39496716 0.50647906 0.41001441 0.40169811 0.32672285 0.3681499
        0.32783326 0.35162986 0.30647063 0.32648544]
The mean cross-validation score and the 95% confidence interval:
Accuracy: 0.37 (+/- 0.11)
```

### 4.3 Model Performance Evaluation

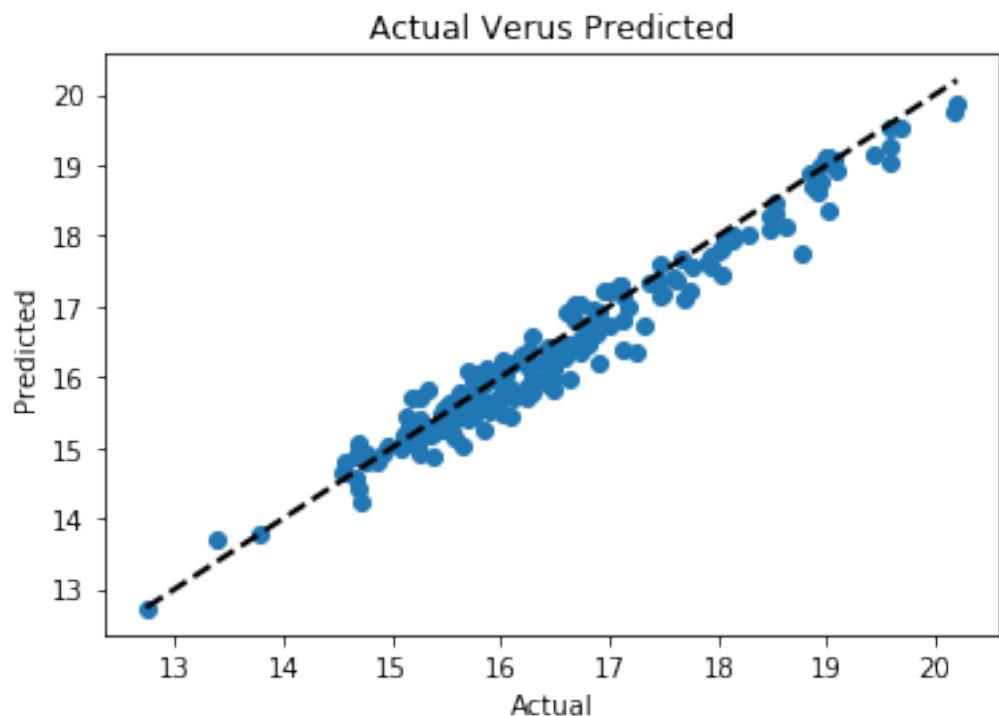
Time to put the model to the test! Let's evaluate how our model has performed with its predictions on the test data

```
[52]: y_pred = grid_search.best_estimator_.predict(x_test)

xgb_mse = mean_squared_error(y_test, y_pred)
xgb_rmse = np.sqrt(xgb_mse)
xgb_rmse
```

```
[52]: 0.28712032059523485
```

```
[53]: fig, ax = plt.subplots()
plt.scatter(y_test,y_pred)
ax.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=2)
ax.set_xlabel('Actual')
ax.set_ylabel('Predicted')
ax.set_title("Actual Verus Predicted")
plt.show()
```



### 4.3.1 Feature Importance

```
[54]: feature_importances = pd.DataFrame(grid_search.best_estimator_.
    ↪feature_importances_,
                                         index = x_train.columns,
                                         columns=['importance']).
    ↪sort_values('importance',
    ↪ascending=False)
feature_importances
```

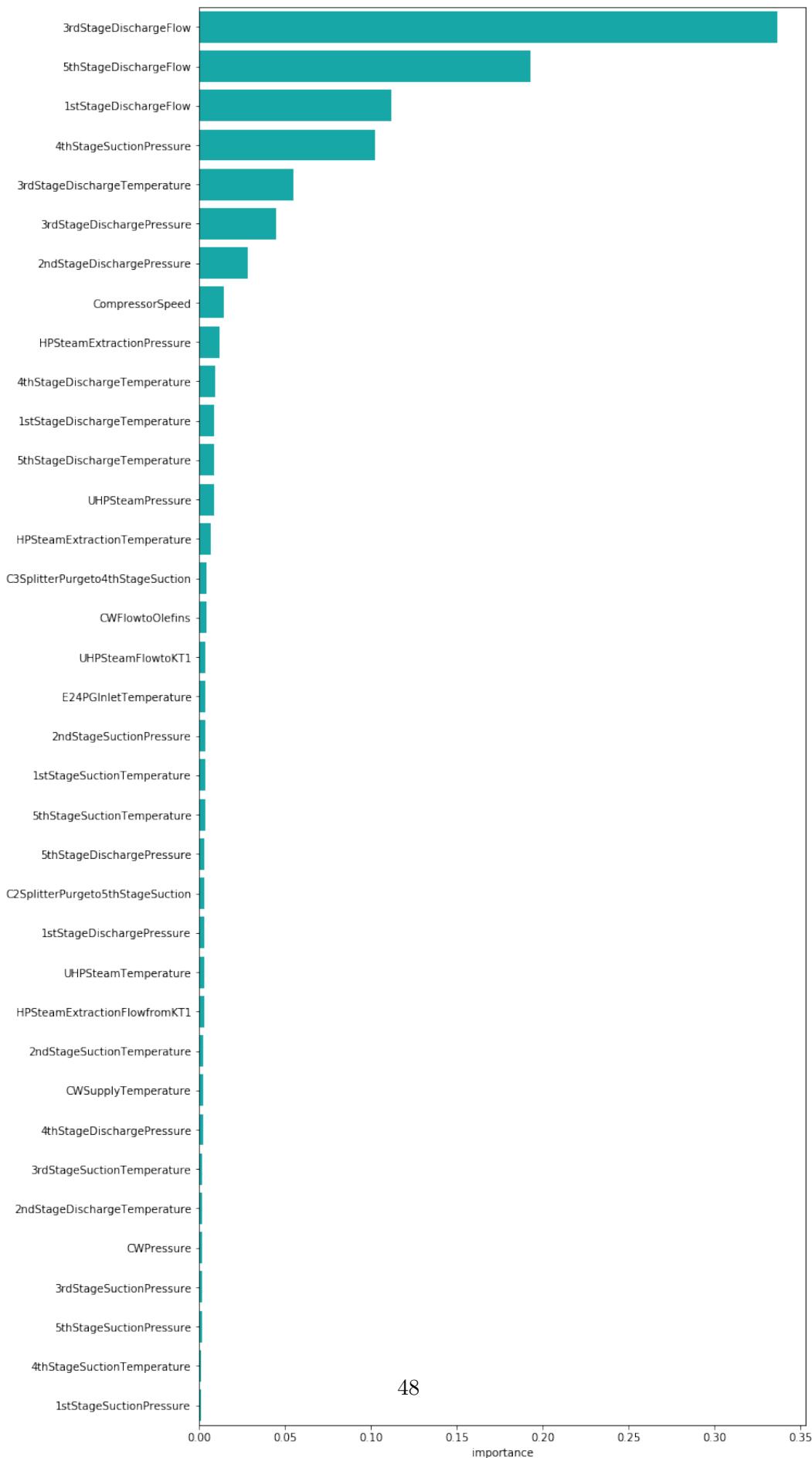
```
[54]:          importance
3rdStageDischargeFlow      0.336238
5thStageDischargeFlow       0.193085
1stStageDischargeFlow       0.111603
4thStageSuctionPressure     0.102406
3rdStageDischargeTemperature 0.054922
3rdStageDischargePressure    0.044591
2ndStageDischargePressure    0.028386
CompressorSpeed             0.014240
HPSteamExtractionPressure    0.012191
4thStageDischargeTemperature 0.009445
1stStageDischargeTemperature 0.008902
5thStageDischargeTemperature 0.008626
UHPSteamPressure              0.008613
HPSteamExtractionTemperature 0.006552
C3SplitterPurgeto4thStageSuction 0.004197
CWFlowtoOlefins              0.004070
UHPSteamFlowtoKT1             0.003864
E24PGInletTemperature        0.003771
2ndStageSuctionPressure       0.003705
1stStageSuctionTemperature    0.003699
5thStageSuctionTemperature    0.003698
5thStageDischargePressure     0.003301
C2SplitterPurgeto5thStageSuction 0.003246
1stStageDischargePressure     0.003168
UHPSteamTemperature            0.002741
HPSteamExtractionFlowfromKT1 0.002736
2ndStageSuctionTemperature    0.002609
CWSupplyTemperature            0.002378
4thStageDischargePressure     0.002124
3rdStageSuctionTemperature    0.001851
2ndStageDischargeTemperature 0.001778
CWPressure                    0.001629
3rdStageSuctionPressure       0.001601
5thStageSuctionPressure       0.001504
4thStageSuctionTemperature    0.001437
```

```
1stStageSuctionPressure          0.001091
```

```
[55]: fig, (ax) = plt.subplots(1, 1, figsize=(10,24))

sns.barplot(x='importance', y=feature_importances.index,
            data=feature_importances, color="c")
```

```
[55]: <matplotlib.axes._subplots.AxesSubplot at 0x6108390>
```



[ ]: