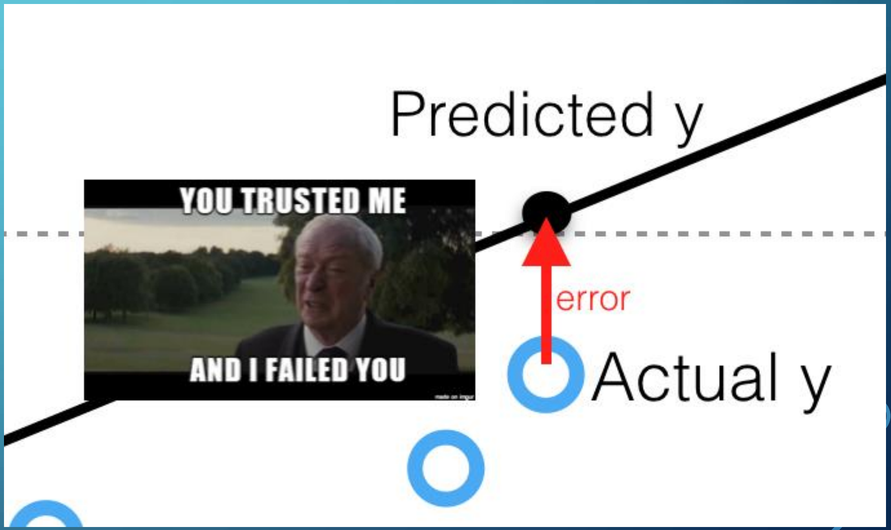
PREDICTION OF TURBINE ENERGY YIELD

LINEAR REGRESSION WITH FEATURE SELECTION

Machine Learning Foundation

By Chiranjit Pathak

TURBINE ENERGY YIELD PREDICTION



PROBLEM STATEMENT

The goal is to predict Turbine Energy Yield (TEY) using ambient variables as features

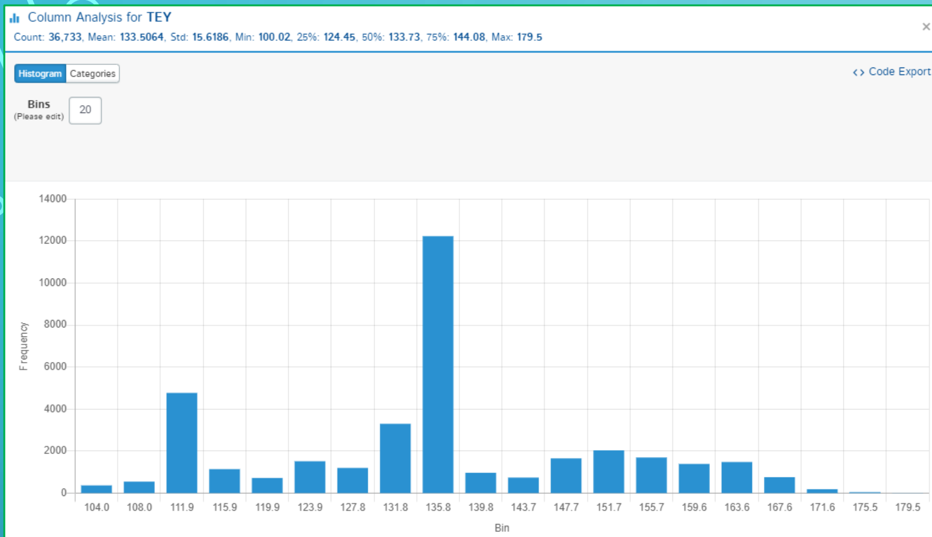
DATA LOADING AND DESCRIPTION

Data Source:
https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set

Variable (Abbr.)	Unit	Min	Max	Mean
Ambient temperature (AT)	C	6.23	37.10	17.71
Ambient pressure (AP)	mbar	985.85	1036.56	1013.07
Ambient humidity (AH)	(%)	24.08	100.20	77.87
Air filter difference pressure (AFDP)	mbar	2.09	7.61	3.93
Gas turbine exhaust pressure (GTEP)	mbar	17.70	40.72	25.56
Turbine inlet temperature (TIT)	C	1000.85	1100.89	1081.43
Turbine after temperature (TAT)	C	511.04	550.61	546.16
Compressor discharge pressure (CDP)	mbar	9.85	15.16	12.06
Turbine energy yield (TEY)	MWH	100.02	179.50	133.51
Carbon monoxide (CO)	mg/m3	0.00	44.10	2.37
Nitrogen oxides (NOx)	mg/m3	25.90	119.91	65.29

The dataset comprises of **36733 instances of 11 sensor measures**. Above is a table showing names of all the columns and their description.

EXPLORATORY DATA ANALYSIS : TARGET

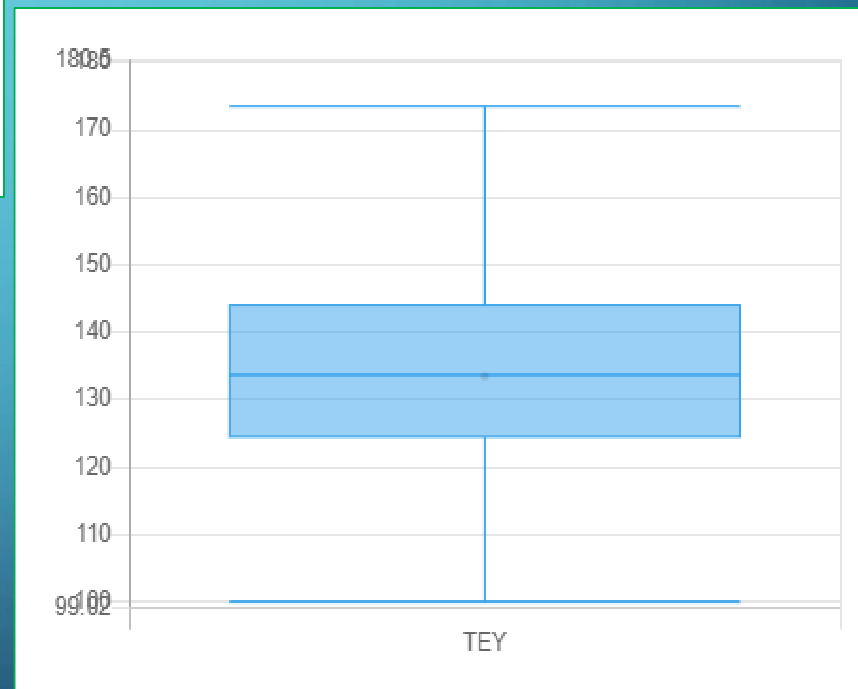


- Target is normally distributed ;
- No outliers

TEY (float64)

Describe | Histogram | Categories

- **Total Rows:** 36,733
  - **Count (non-nan):** 36,733
  - **Count (missing):** 0
  - **% Missing:** 0
- **top:** 133.78
  - **Frequency:** 185
- **25%:** 124.45
- **50%:** 133.73
- **75%:** 144.08
- **kurt:** -0.5 (platykurtic)
- **max:** 179.5
- **mean:** 133.5064
- **median:** 133.73



EXPLORATORY DATA ANALYSIS: FEATURES

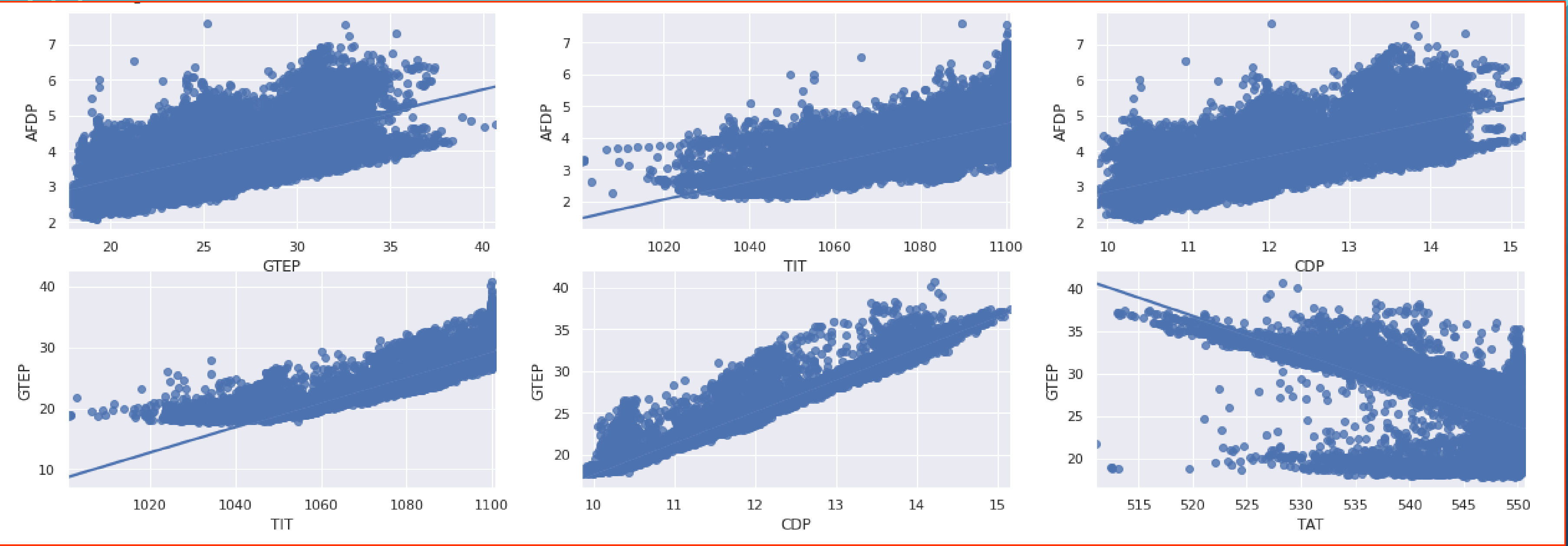
Overview

Overview | Warnings 4 | Reproduction

Warnings

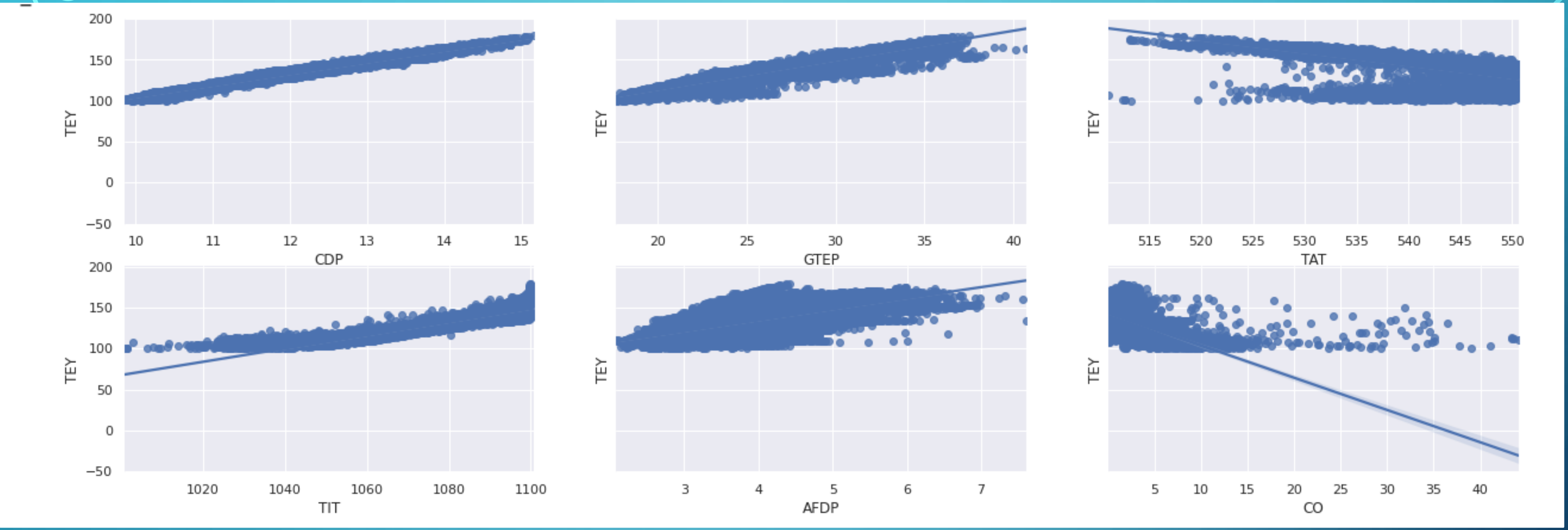
- TEY is highly correlated with GTEP and 2 other fields
- GTEP is highly correlated with TEY and 1 other fields
- TIT is highly correlated with TEY and 1 other fields
- CDP is highly correlated with GTEP and 2 other fields

EXPLORATORY DATA ANALYSIS: HIGH CORRELATION AMONG FEATURES



AFDP, GTEP, CDP, TAT, TIT are highly correlated among each others

EXPLORATORY DATA ANALYSIS: TARGET VS FEATURES

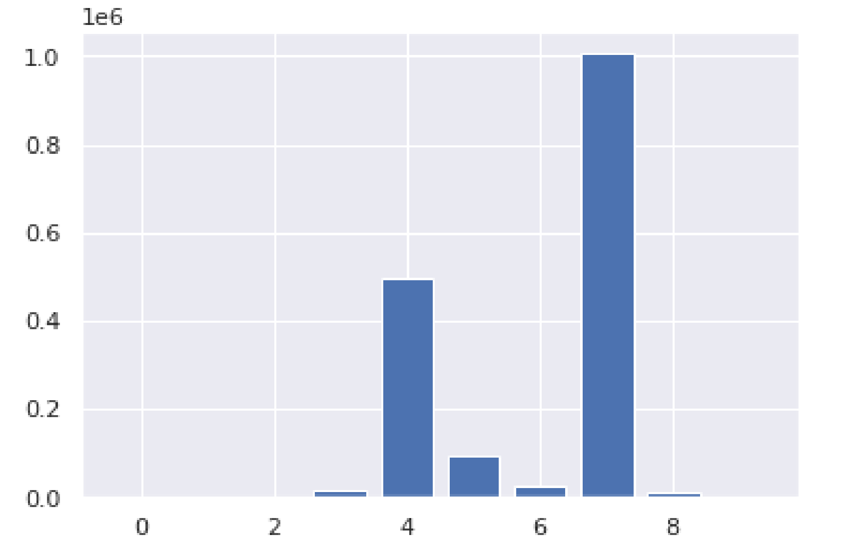


AFDP, GTEP, CDP, TIT, TAT are highly correlated with TEY

FEATURE SELECTION

FEATURE SELECTION:
USING CORRELATION STATISTICS

import SelectKBest and f_regression

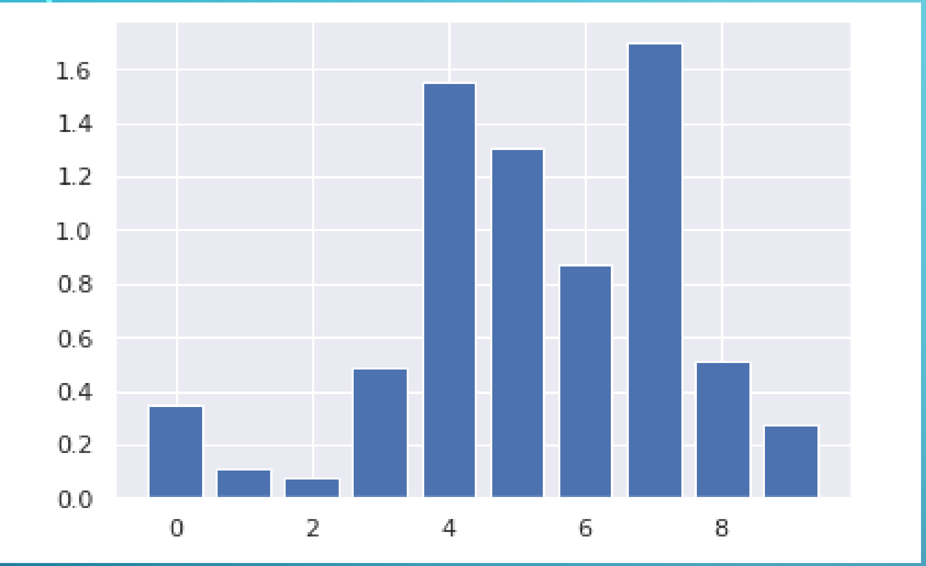



```
Feature 0: 860.624004
Feature 1: 718.013902
Feature 2: 293.925853
Feature 3: 15789.036161
Feature 4: 496027.685586
Feature 5: 95289.882724
Feature 6: 23891.083505
Feature 7: 1005912.467434
Feature 8: 11633.388290
Feature 9: 67.526325
```

So 3 to 4 features are having hi impact on the model

FEATURE SELECTION:
USING MUTUAL INFORMATION THEORY

import SelectKBest and mutual_info_regression

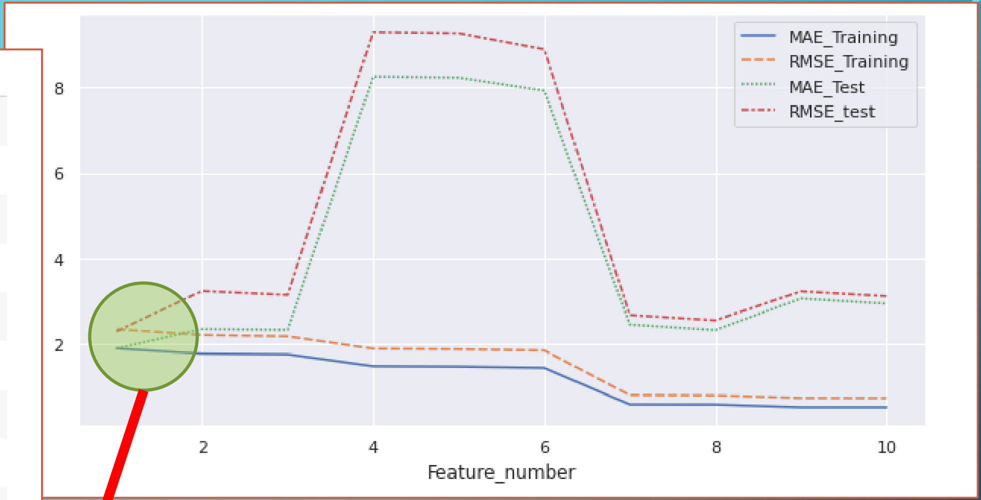



```
Feature 0: 0.350200
Feature 1: 0.109574
Feature 2: 0.073342
Feature 3: 0.484026
Feature 4: 1.552312
Feature 5: 1.304582
Feature 6: 0.874015
Feature 7: 1.699997
Feature 8: 0.510088
Feature 9: 0.272602
```

So 3 to 5 features are having hi impact on the model

FEATURE SELECTION: USING GRIDSEARCH WITH MUTUAL INFORMATION THEORY

	Feature_number	MAE_Training	RMSE_Training	MAE_Test	RMSE_test
0	1	1.905447	2.354340	1.903045	2.302691
1	2	1.782063	2.218478	2.356738	3.244547
2	3	1.765120	2.187249	2.338517	3.154224
3	4	1.490215	1.900632	8.242611	9.281714
4	5	1.480846	1.888095	8.220992	9.255912
5	6	1.451452	1.859275	7.916408	8.881514
6	7	0.597309	0.817165	2.461045	2.682705
7	8	0.595421	0.807878	2.333199	2.559623
8	9	0.531623	0.746790	3.070164	3.237879
9	10	0.531585	0.744256	2.951597	3.124839



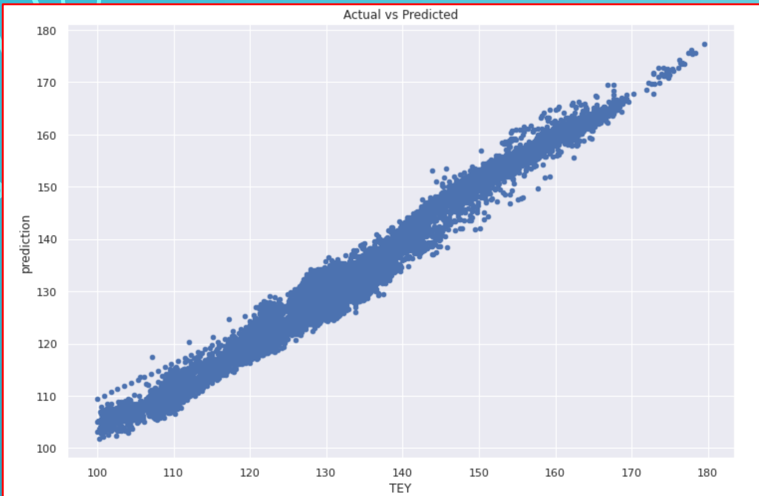
MAE/RMSE are in same level for both the Train & Test datasets together with a single feature

TEY = -38.03 + 14.227 * CDP

How do we interpret the coefficient (+14.227)

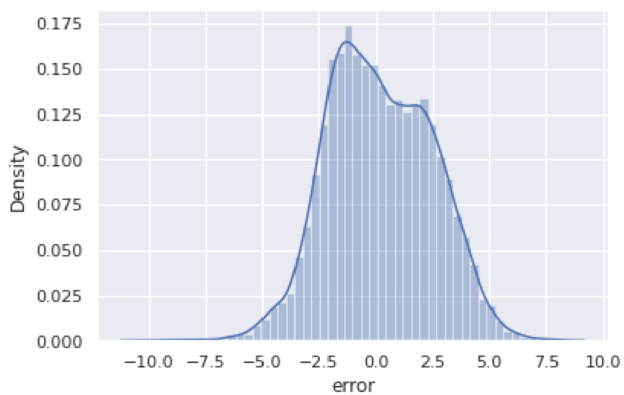
- A "unit" increase in **CDP** is **associated with** a **"14.227** unit" increase in **TEY**.

MODEL EVALUATION

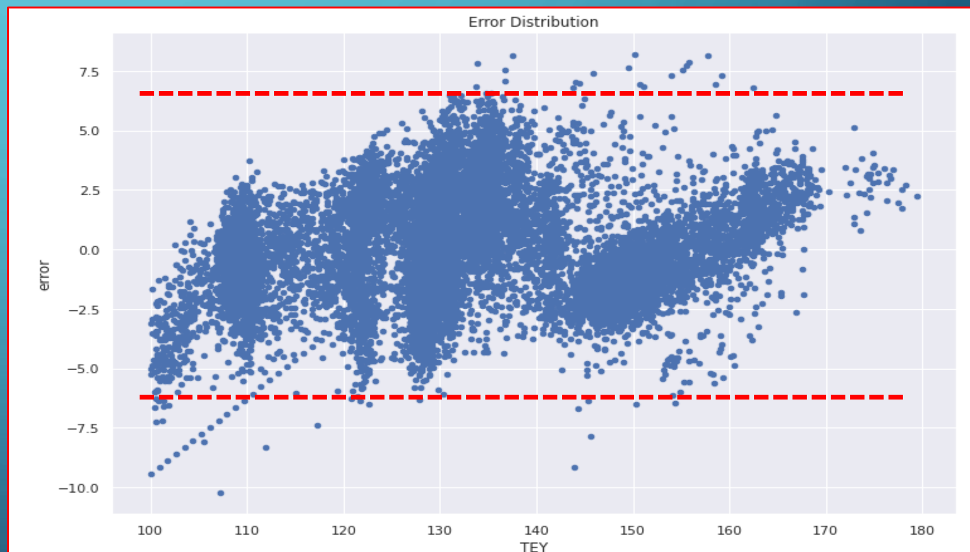



```
[ ]  sns.distplot(a['error'])
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f0d4badbcf8>



Homoscedasticity has been observed.



CONCLUSION

- Feature selection is an important criterion when they are strongly correlated among each other.
- Search technique using RMSE and MAE for different number of features in train & test dataset is also an important factor while selecting a model.
- Homoscedasticity observed for the errors.
- This case study can also be referred for feature engineering having more number of features.

THANKS FOR READING

Lets collaborate and happy to receive any feedback/suggestion/comment at…….

pathak.chiranjit@gmail.com