Capstone Project

Appliances Energy Prediction

G. V. Kapeesh Varma

CONTENTS

PROBLEM STATEMENT

In this time of global uncertainty world needs energy and in increasing quantities to support economic and social progress and build a better quality of life, in particular in developing countries. But even today there are many regions which face constant power outages. These outages are primarily caused due to excess load consumed by home appliances. Hence, the ability to predict energy consumption can not only save a lot of money for the end user but also helps to avoid such power outages.

The aim of this project is to analyze data from various sensors and predict the Energy consumption through various Regression models.

APPROACH

In this project, I performed Regression analysis in order to predict Appliance energy usage based on data collected from various sensors. Initially, I imported and analyzed various features of the Appliance Energy dataset through data exploration and visualizations.

In the next part, appropriate Dependent & Independent variables have been picked for Regression. The Data is then preprocessed in order to transform categorical variables. Regression is then performed through Linear Regression, Ridge Regression, Random Forest Regression etc., and the performance metrics of these models have been calculated to find the best fit.

DATA SUMMARY

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   date          19735 non-null  object
 1   Appliances    19735 non-null  int64
 2   lights        19735 non-null  int64
 3   T1            19735 non-null  float64
 4   RH_1          19735 non-null  float64
 5   T2            19735 non-null  float64
 6   RH_2          19735 non-null  float64
 7   T3            19735 non-null  float64
 8   RH_3          19735 non-null  float64
 9   T4            19735 non-null  float64
 10  RH_4          19735 non-null  float64
 11  T5            19735 non-null  float64
 12  RH_5          19735 non-null  float64
 13  T6            19735 non-null  float64
 14  RH_6          19735 non-null  float64
 15  T7            19735 non-null  float64
 16  RH_7          19735 non-null  float64
 17  T8            19735 non-null  float64
 18  RH_8          19735 non-null  float64
 19  T9            19735 non-null  float64
 20  RH_9          19735 non-null  float64
 21  T_out         19735 non-null  float64
 22  Press_mm_hg   19735 non-null  float64
 23  RH_out        19735 non-null  float64
 24  Windspeed     19735 non-null  float64
 25  Visibility    19735 non-null  float64
 26  Tdewpoint     19735 non-null  float64
 27  rv1           19735 non-null  float64
 28  rv2           19735 non-null  float64
dtypes: float64(26), int64(2), object(1)
```

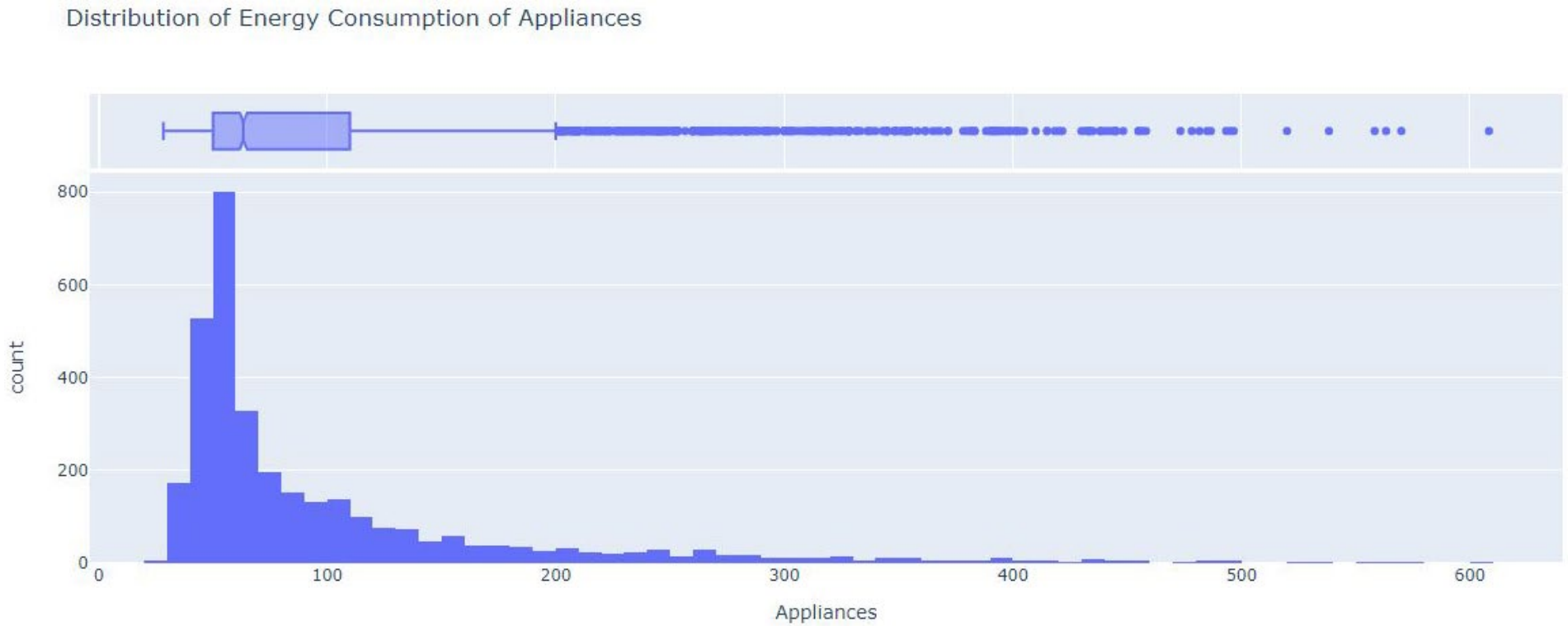
	date	Appliances	lights	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4	T5	RH_5	T6
19730	2016-05-27 17:20:00	100	0	25.566667	46.560000	25.890000	42.025714	27.200000	41.163333	24.7	45.590000	23.20	52.400000	24.796667
19731	2016-05-27 17:30:00	90	0	25.500000	46.500000	25.754000	42.080000	27.133333	41.223333	24.7	45.590000	23.23	52.326667	24.196667
19732	2016-05-27 17:40:00	270	10	25.500000	46.596667	25.628571	42.768571	27.050000	41.690000	24.7	45.730000	23.23	52.266667	23.626667
19733	2016-05-27 17:50:00	420	10	25.500000	46.990000	25.414000	43.036000	26.890000	41.290000	24.7	45.790000	23.20	52.200000	22.433333
19734	2016-05-27 18:00:00	430	10	25.500000	46.600000	25.264286	42.971429	26.823333	41.156667	24.7	45.963333	23.20	52.200000	21.026667

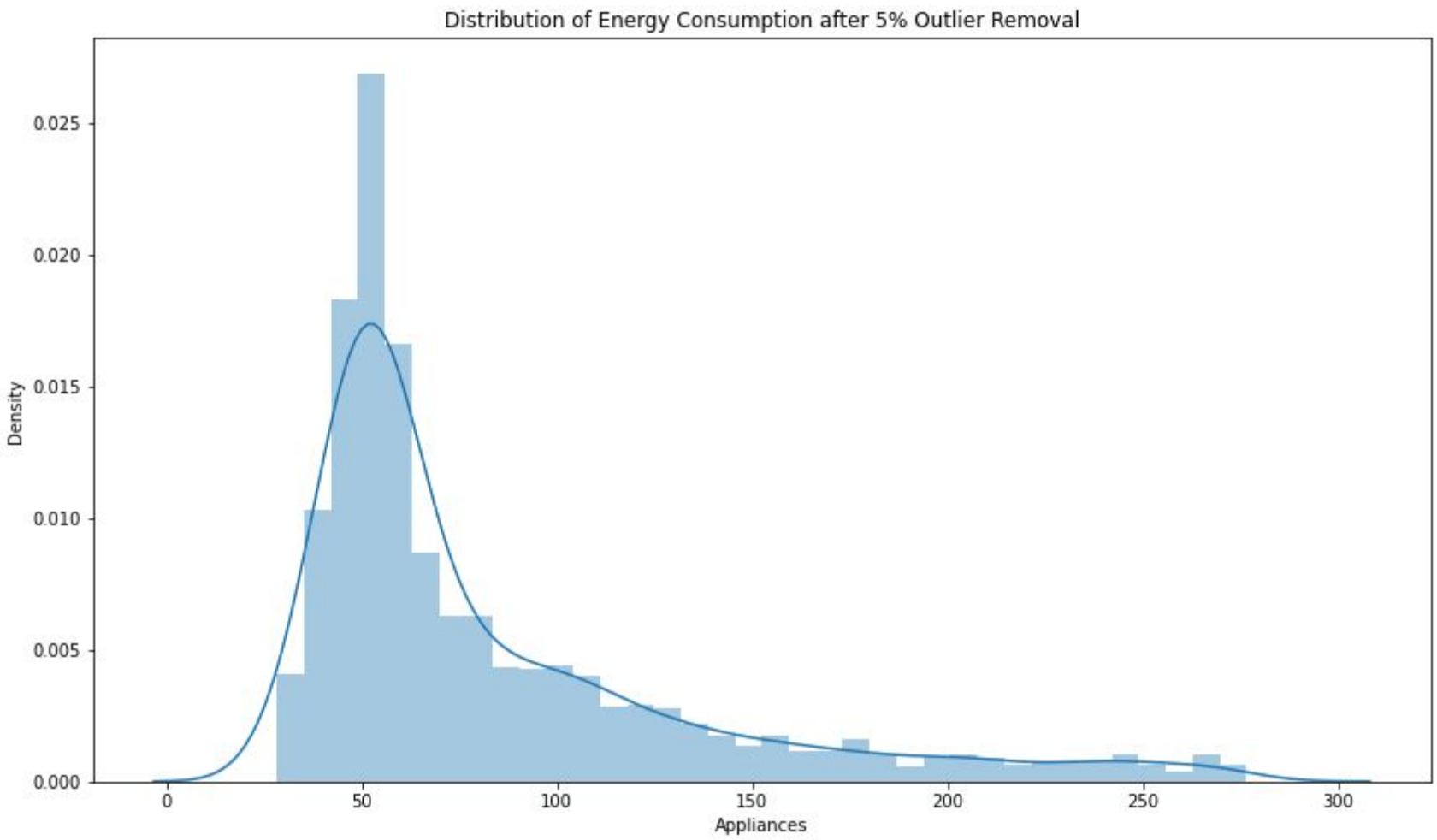
Appliance Energy DataFrame

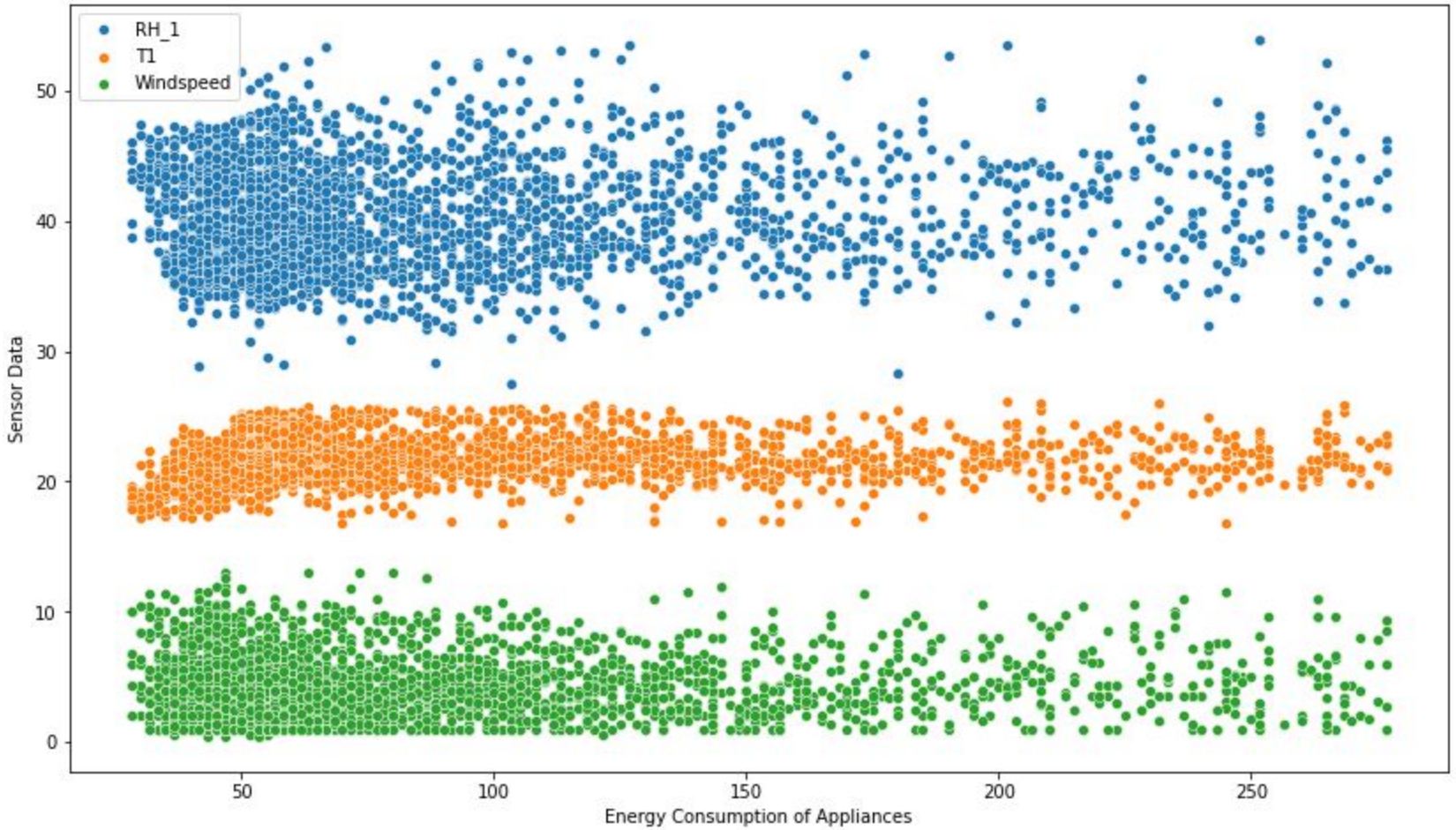
	Appliances	lights	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4	T5
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000
mean	97.694958	3.801875	21.686571	40.259739	20.341219	40.420420	22.267611	39.242500	20.855335	39.026904	19.592106
std	102.524891	7.935988	1.606066	3.979299	2.192974	4.069813	2.006111	3.254576	2.042884	4.341321	1.844623
min	10.000000	0.000000	16.790000	27.023333	16.100000	20.463333	17.200000	28.766667	15.100000	27.660000	15.330000
25%	50.000000	0.000000	20.760000	37.333333	18.790000	37.900000	20.790000	36.900000	19.530000	35.530000	18.277500
50%	60.000000	0.000000	21.600000	39.656667	20.000000	40.500000	22.100000	38.530000	20.666667	38.400000	19.390000
75%	100.000000	0.000000	22.600000	43.066667	21.500000	43.260000	23.290000	41.760000	22.100000	42.156667	20.619643
max	1080.000000	70.000000	26.260000	63.360000	29.856667	56.026667	29.236000	50.163333	26.200000	51.090000	25.795000

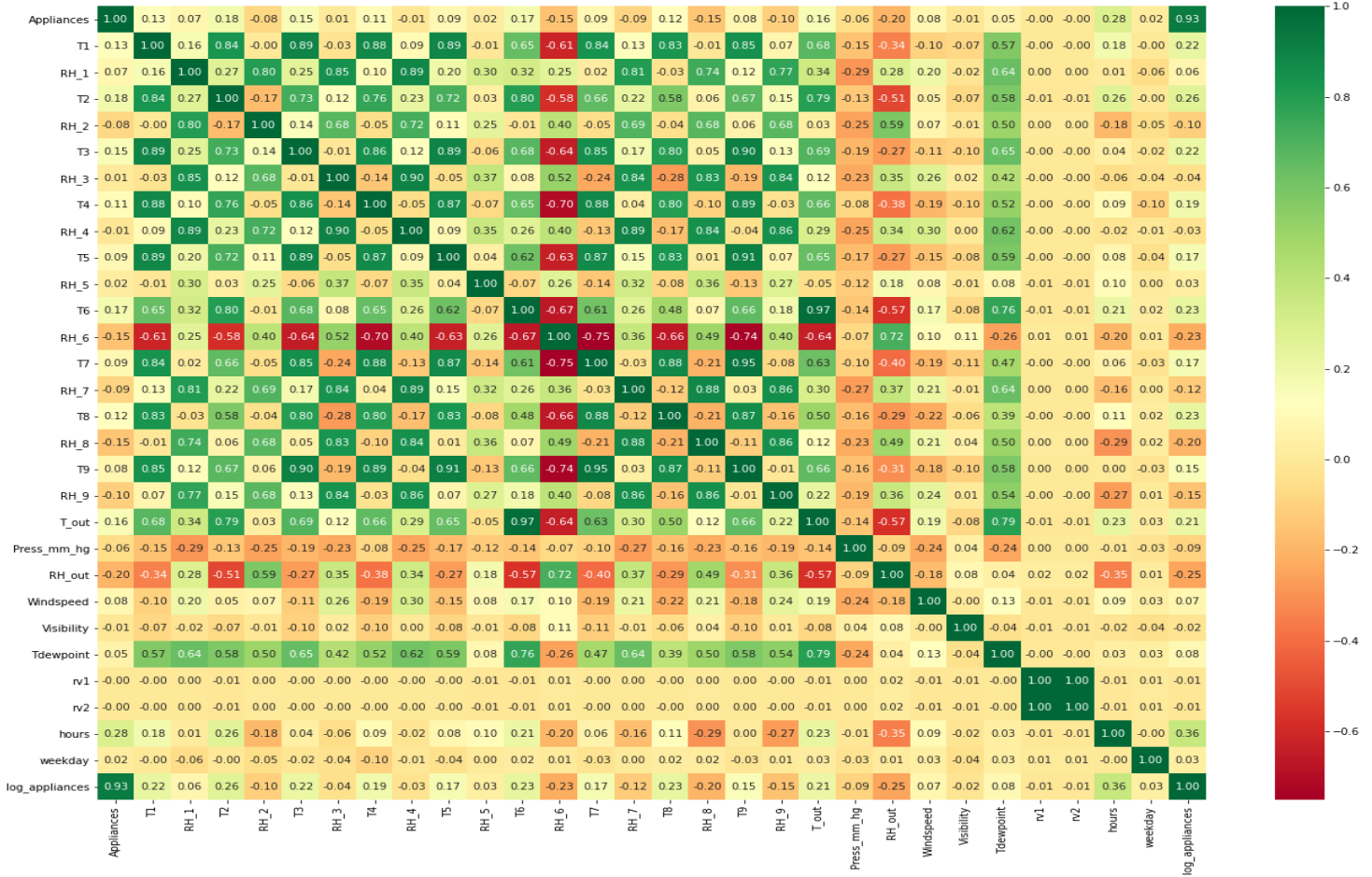
Descriptive Statistics

DATA VISUALIZATIONS




Distribution of Energy Consumption after 5% Outlier Removal
Density
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
Appliances


RH_1
T1
Windspeed
Sensor Data
0
10
20
30
40
50
50
100
150
200
250
Energy Consumption of Appliances

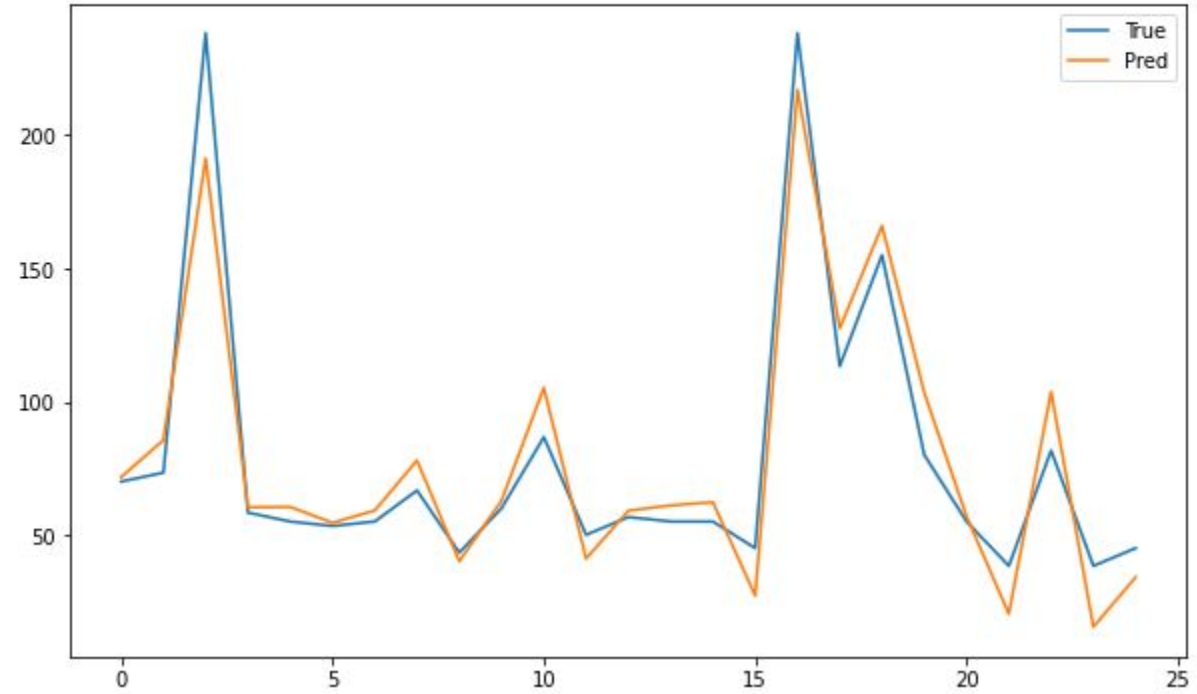


Correlation Plot

REGRESSION MODELLING

PERFORMANCE METRICS FOR LINEAR REGRESSION

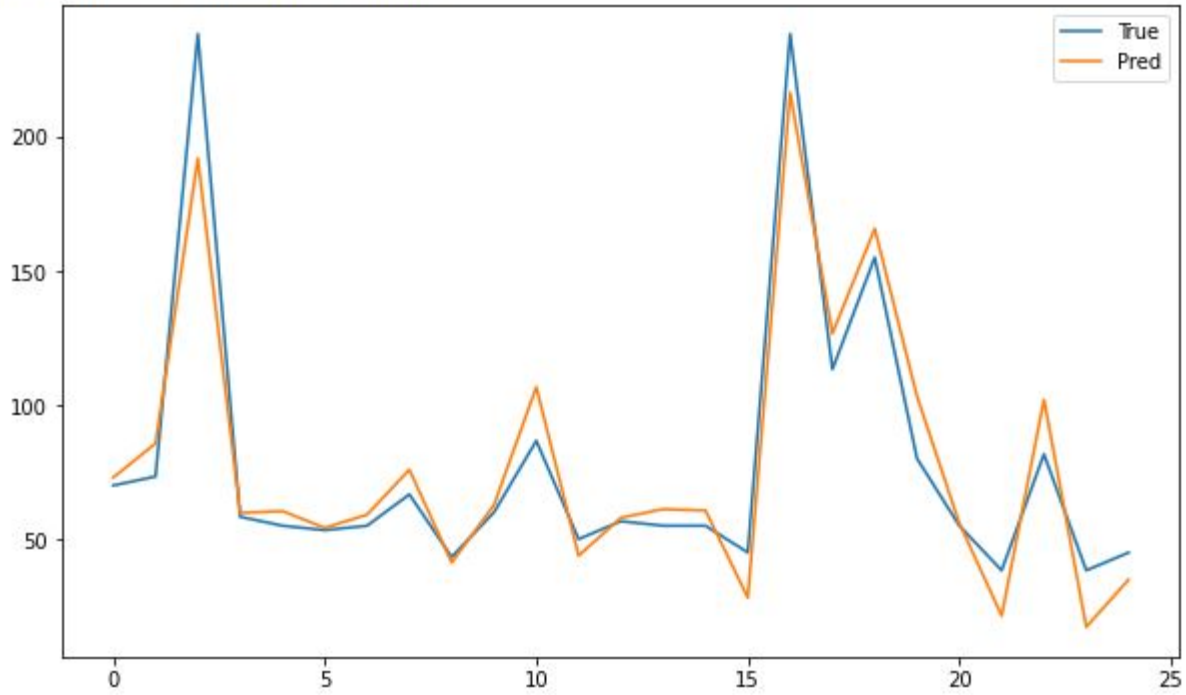
MSE: 305.9395537435371
R2_Score: 0.8982294119683846



PERFORMANCE METRICS FOR RIDGE REGRESSION

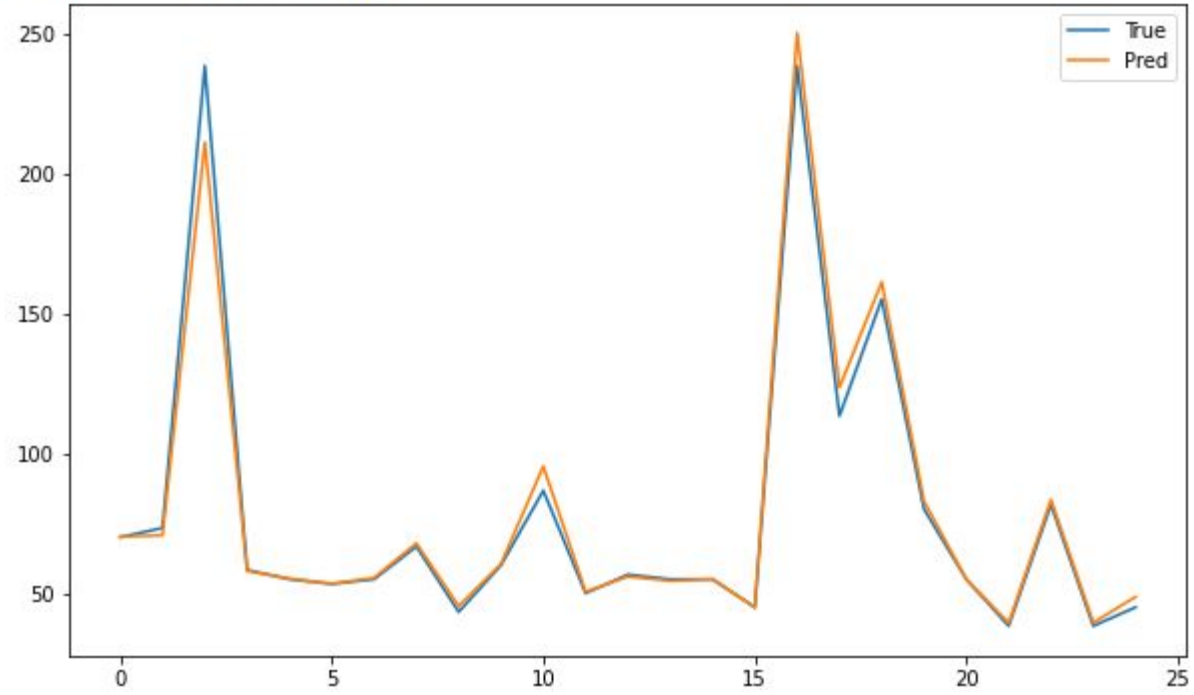
Best Alpha value for Ridge Regression: {'alpha': 5}

MSE: 298.55207446862335
R2_Score: 0.9006868520106408



PERFORMANCE METRICS FOR REGRESSION WITH EXTREME GRADIENT BOOST

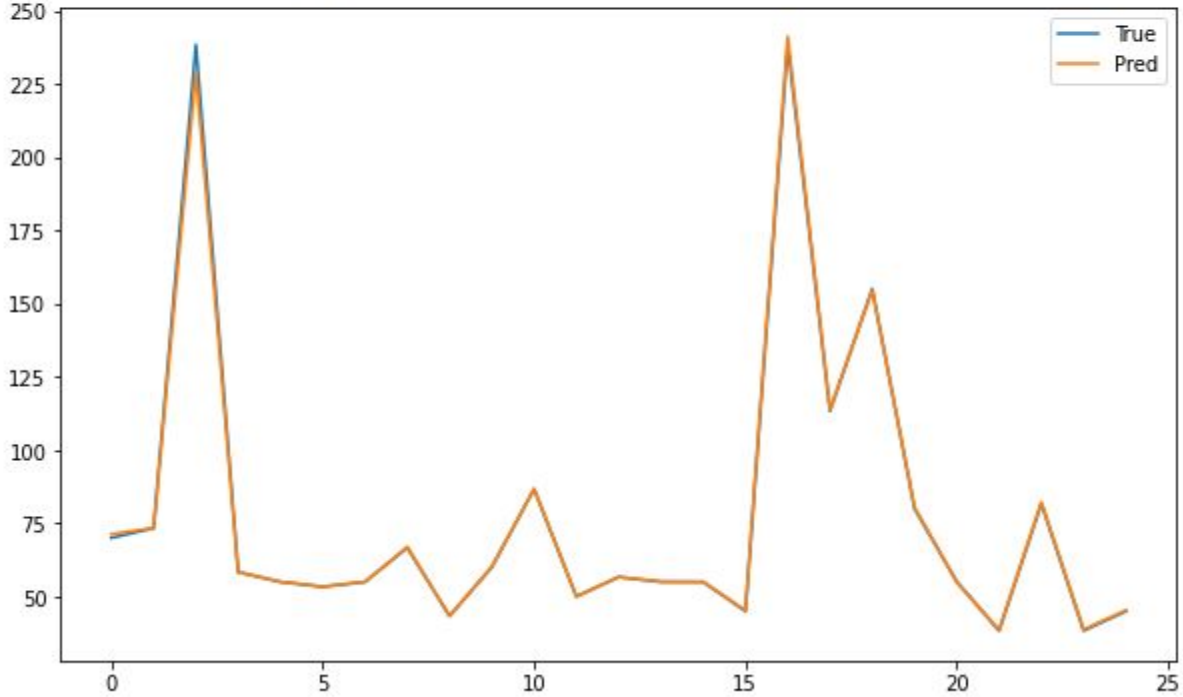
MSE: 140.76694659472426
R2_Score: 0.953173969318236



PERFORMANCE METRICS FOR REGRESSION WITH RANDOM FORESTS

Best Number of Estimators for Random Forest Regression: {'n_estimators': 100}

MSE: 22.245189918352818
R2_Score: 0.9926001524446116



CONCLUSIONS

1. The "lights" column has been dropped as it is of least importance in this model.
2. "Hours" is the most influencing data column i.e.; it has the highest correlation with Appliance Energy consumption.
3. Weekends (Saturdays and Sundays) are observed to have higher Energy consumption than Weekdays.
4. Energy consumption is higher during evenings.
5. Random Forest Regressor (with n_estimators = 100) has the most efficient performance metrics (MSE = 22.24, R2_Score = 0.99) among the different models used in this prediction.