Capstone Project-2

Supervised ML Regression

Appliances Energy Prediction

Sammed Naresh Majalekar
Individual Capstone Project

Introduction

Problem Statement:

What we have to do is to predict the energy consumptions in an house by using given data considering time in a regular day, temperature at that time, humidity at that time and which day it is, which week it is or which month it is.

Dataset:

Our dataset consist of 19735 records and 29 features in which one is dependent feature.

Feature Information

1. date : time year-month-day hour:minute:second
2. lights : energy use of light fixtures in the house in Wh
3. T1 : Temperature in kitchen area, in Celsius
4. T2 : Temperature in living room area, in Celsius
5. T3 : Temperature in laundry room area
6. T4 : Temperature in office room, in Celsius
7. T5 : Temperature in bathroom, in Celsius
8. T6 : Temperature outside the building (north side), in Celsius
9. T7 : Temperature in ironing room, in Celsius
10. T8 : Temperature in teenager room 2, in Celsius
11. T9 : Temperature in parents' room, in Celsius
12. T_out : Temperature outside (from Chievres weather station), in Celsius
13. Tdewpoint : (from Chievres weather station), Â°C
14. RH_1 : Humidity in kitchen area, in %
15. RH_2 : Humidity in living room area, in %
16. RH_3 : Humidity in laundry room area, in %
17. RH_4 : Humidity in office room, in %
18. RH_5 : Humidity in bathroom, in %
19. RH_6 : Humidity outside the building (north side), in %
20. RH_7 : Humidity in ironing room, in %
21. RH_8 : Humidity in teenager room 2, in %
22. RH_9 : Humidity in parents' room, in %
23. RH_out :Humidity outside (from Chievres weather station), in %
24. Pressure : (from Chievres weather station), in mm Hg
25. Wind speed: (from Chievres weather station), in m/s
26. Visibility :(from Chievres weather station), in km
27. Rv1 :Random variable 1, non-dimensional
28. Rv2 :Random variable 2, non-dimensional
29. Appliances : Total energy used by appliances, in Wh

Steps Followed

- **Exploratory Data Analysis**
  - a. Understanding the data
  - b. Null Value treatment
  - c. Outlier treatment
  - d. Applying log transformation on distribution of dependent variable
  - e. Feature engineering for data visualization
- **Data Visualization**
- **Feature Engineering**
- **Model Building**
- **Model Evaluation**
- **Hyper parameter Tuning**
- **Conclusion**

Exploratory Data Analysis

a. Understanding the data –

After loading the dataset, explored the data, features, head, tail, data types, info of dataset, descriptive statistics

b. Null/Duplicate value-

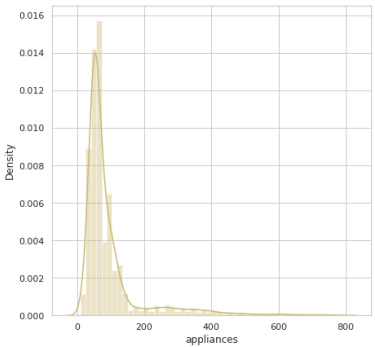
In our dataset we don’t have any null or duplicate record.

c. Outlier removal-

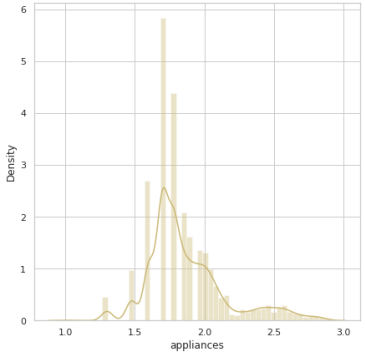
We decided to remove top 0.1% of records considering target variable. So we get the threshold of 790 WH.

d. Applying log transformation

Dependent variable

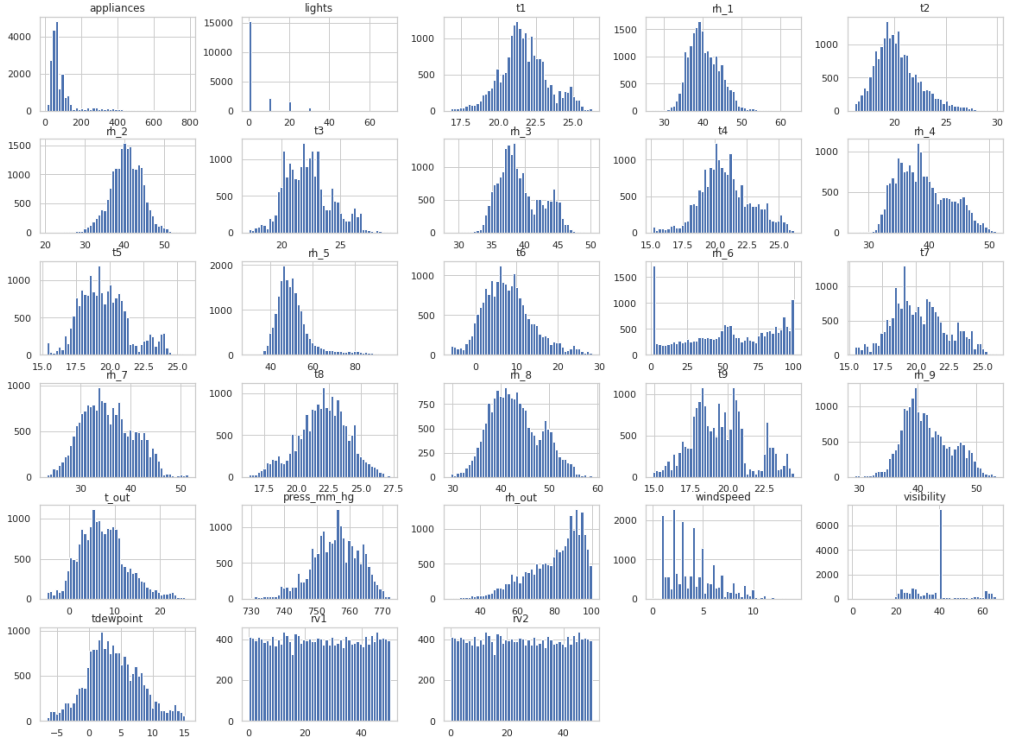


Before



After

Independent Variable

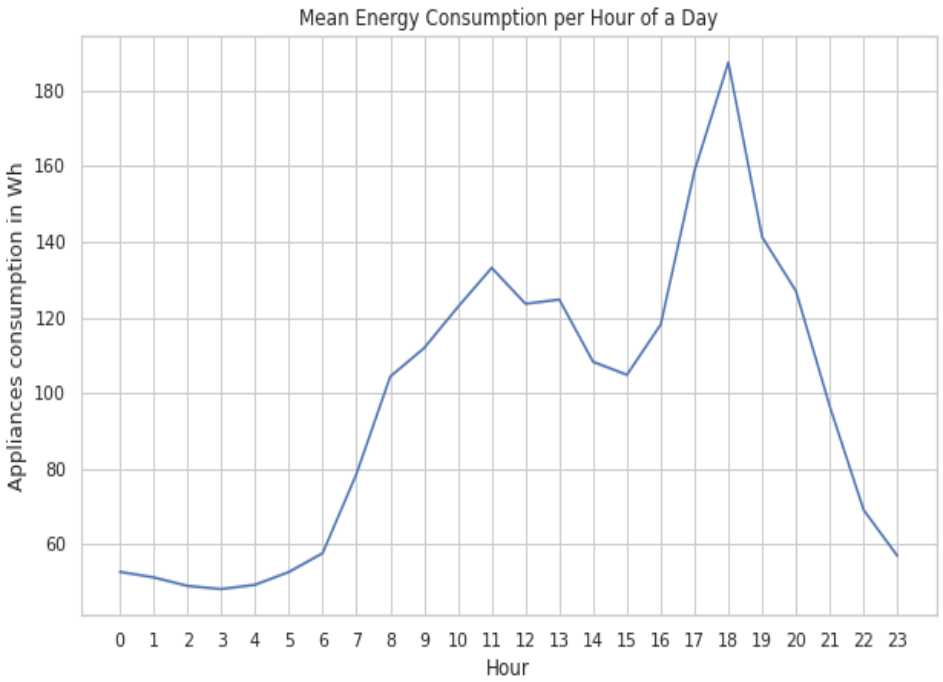


e. Feature Engineering for Data Visualization

- First I applied log transformation on target variable and created a new column of it for further use.
- I created two more columns for average house temperature and for average house humidity.
- Then took product temperature and humidity to remove additional effects.
- Created other columns using date, hour and month columns.
- Created 30 minutes dataset and 1 hour dataset.

Data Visualization

Mean Energy Consumption per hour in a day



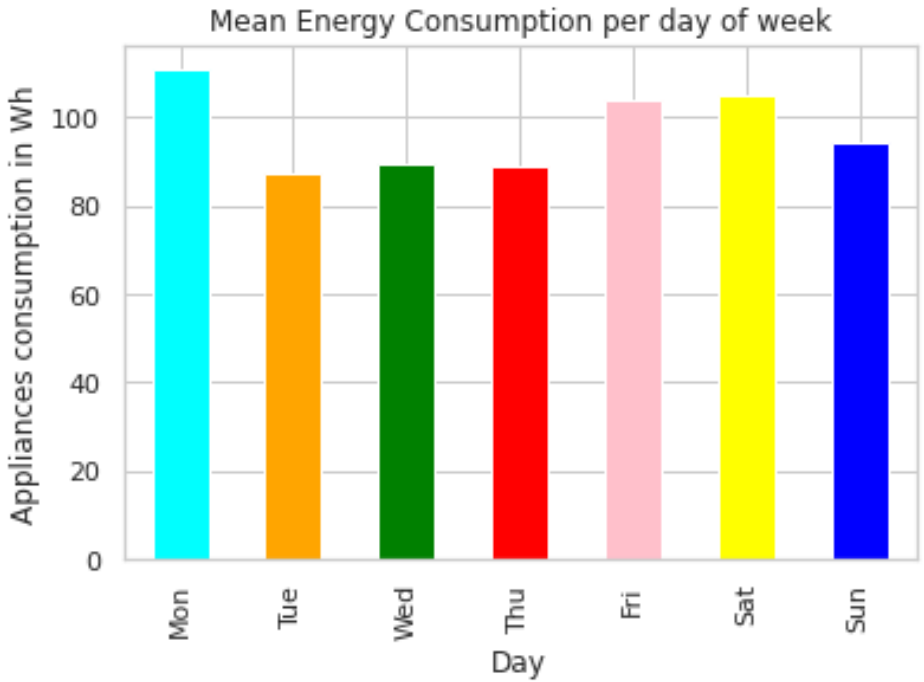
We can clearly see that from 7 AM to 11 AM there is clear spike in energy consumption (from 80 wh to 135 wh), because in morning session we use lot of appliances for various purpose.

Then from 11 AM to 4 PM the energy consumption is ranges between 105 wh to 135 wh.

From 4 pm to 6 pm we can spot a sudden hike (from 120 wh to 185 wh) in energy consumption following by a sudden fall from 185 wh to 70 wh during 6 PM to 10 pm in night.

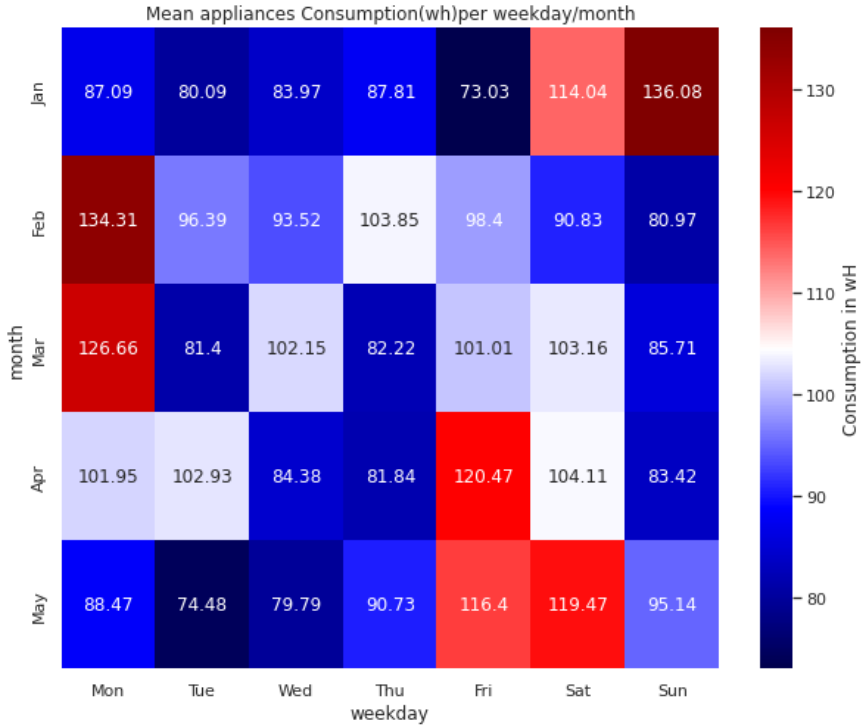
And then from 10 PM in night to 6 AM in the morning there is low energy consumption ranging from 50-60 wh as there is negligible use of energy or appliances.

Mean Energy consumption per day of a week



The power load is a bit higher on Monday, Friday, Saturday and Sunday than the other days.

Energy consumptions on week day in given months



In January, Saturday and Sundays have more energy consumption and Friday has lowest energy consumption.

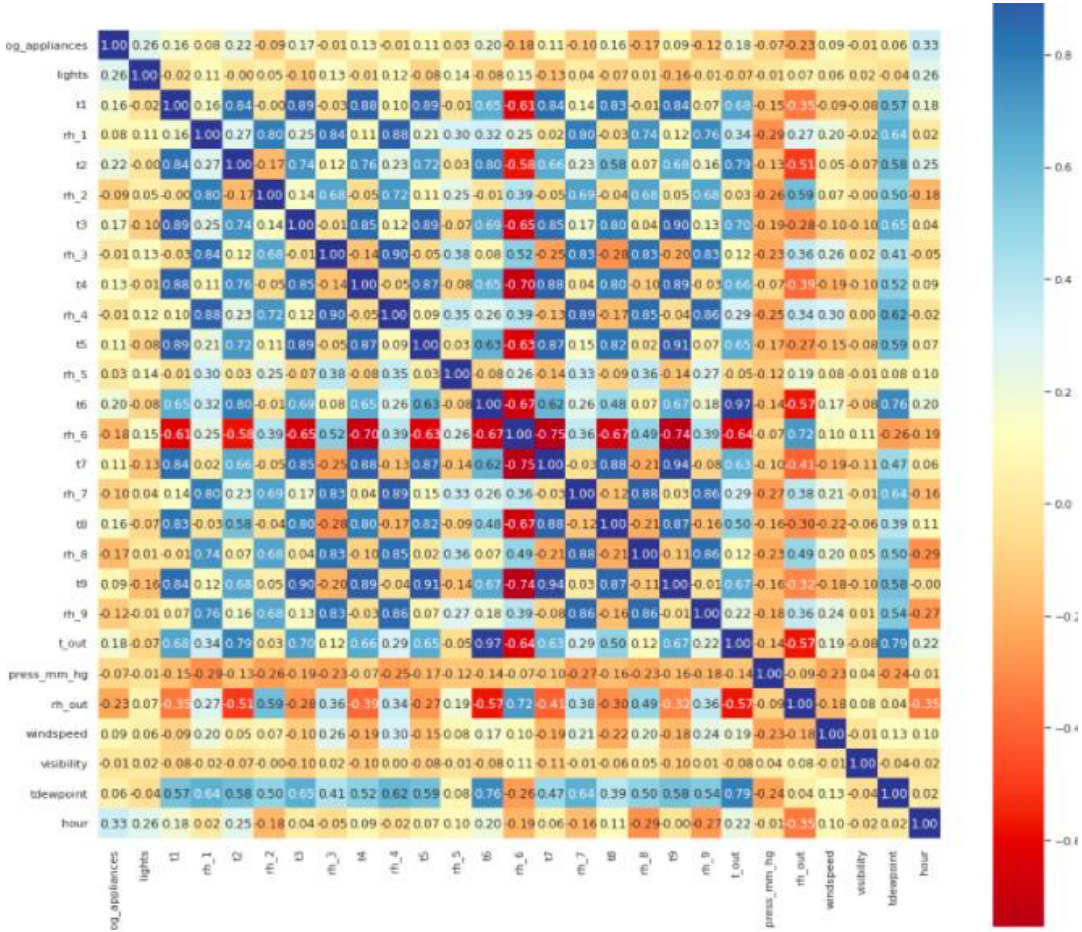
In February, Monday and Thursday having high consumptions.

March has Monday and Saturday as high energy consumers with Tuesday as low consumer.

In April, Friday has more energy consumption and Thursday has the least.

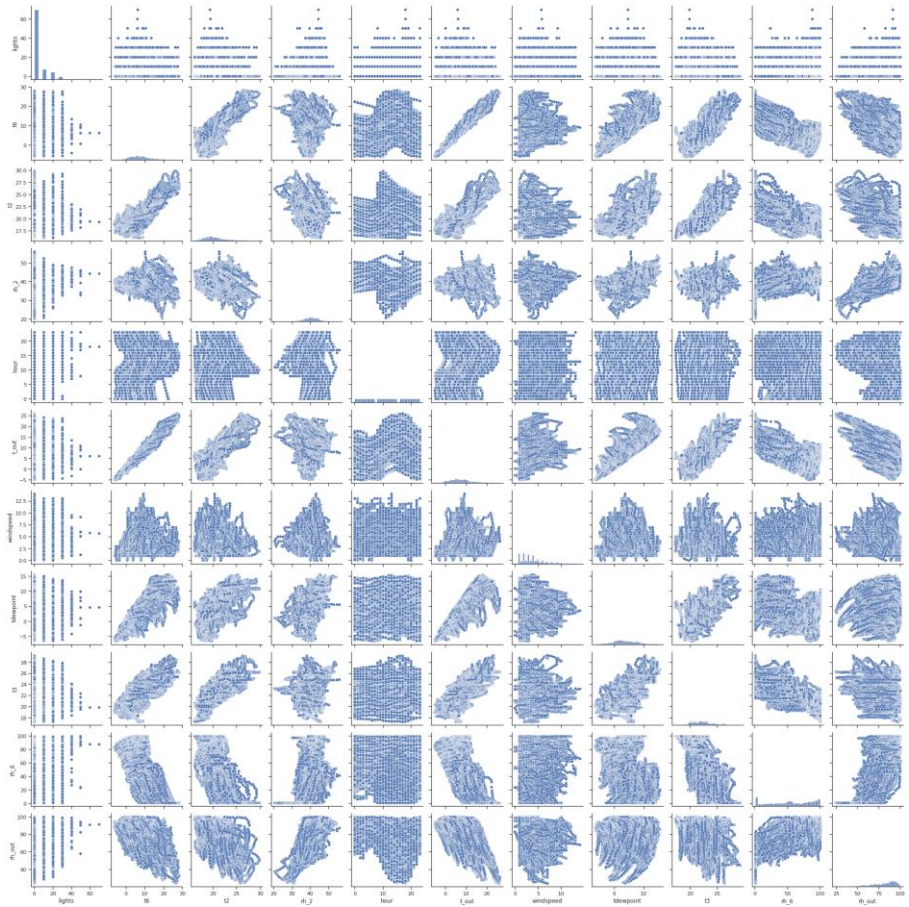
In May Fridays and Saturdays have highest consumption where Tuesdays has least consumption.

Identifying Linear Relationships



- features that have high correlation with dependent variable (log_appliances) are hour, t_out, lights, t6, rh_6, t2, t3, rh_out, rh_8, wind speed

Pair plot



- From above pair plot 't_out' has linear relationship with 't dew point', 't2', 't6' so we are keeping only temperature outside the home 't6'.
- Also 't2' has linear relationship with 't6'.
- we will not use the following features 't2', 't3', 'rh_2', 't_out', 't dew point'

Feature Engineering

For Linear Regression

- From above correlation heat map and pair plot I selected following features for our Linear Regression model:-

  'low_consum','high_consum','hour','lights','t6','rh_6','windspeed','t6rh6'

- Then I transformed some of the categorical variables into numerical variables using ‘get dummies’.

For Random Forest Regressor

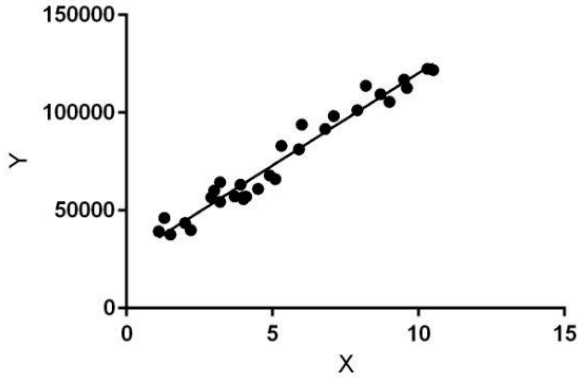
For random forest regressor I fitted the model on training data first then I calculated feature importance score for all features. Then I selected some of the top features for our model as below:-

'rh_3','t5','t8','press_mm_hg', 'hour', 'house_temp','t3rh3', 'hour*lights', 'hour avg', 'low consum', 'high consum'

Model Building

1.Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.



Train-Test Split
Splits the data into 80:20 ratio for training and testing purpose respectively, using ‘sklearn’ train test split metrics.

Model Fitting
After that fitted the model on train data.

Prediction
Then predicted on test data.

Model Evaluation
For model evaluation I defined a metrics which calculates the average error, R2 score and Accuracy.

```
LinearRegression()

Average Error        : 0.3333 degrees
Variance score R^2   : 37.25%
Accuracy             : 92.40%
```

Mean Absolute Error

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

MAE 0.33325880611131997

Mean Squared Error

It represents the squared distance between actual and predicted values. We perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.

MSE 0.19464764211275365

Root Mean Squared Error

It is a simple square root of mean squared error.

RMSE 0.4411888961802571

R Squared (R2)

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit.

R2 score is 0.37249438014797187

Adjusted R2

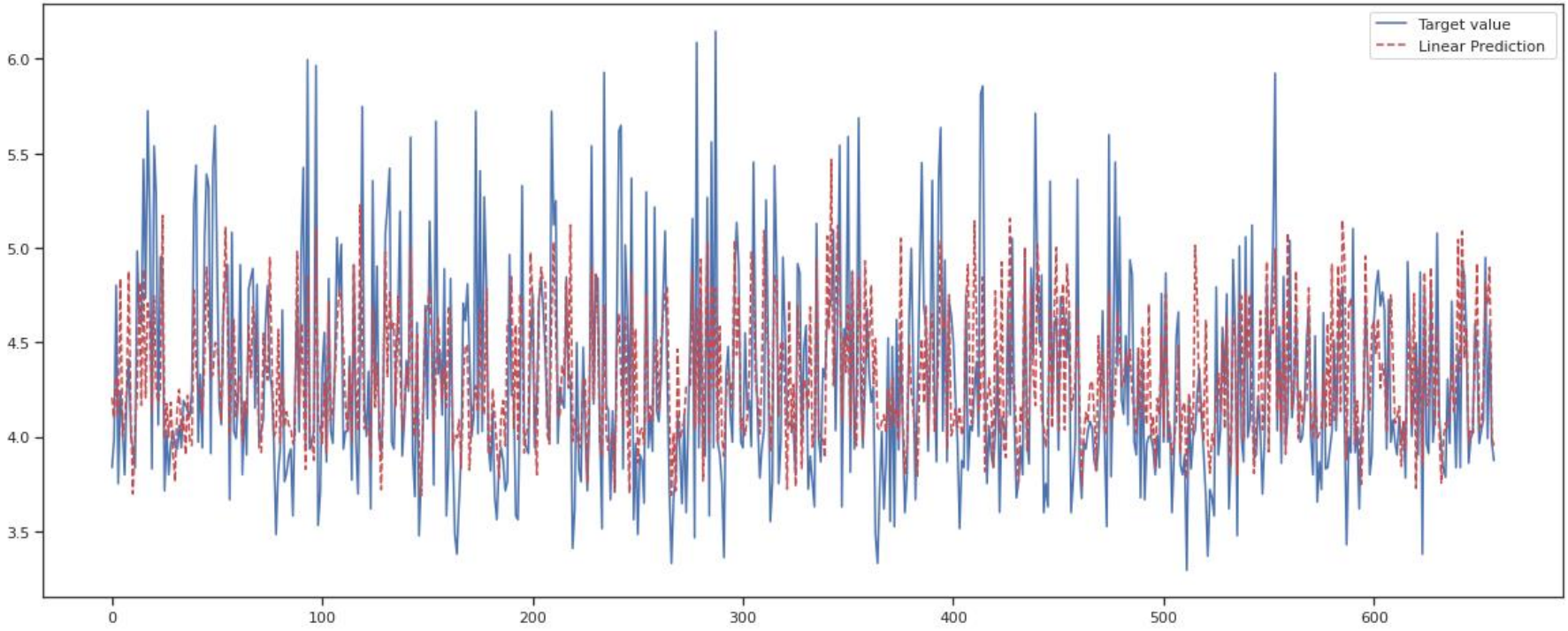
The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because it assumes that while adding more data variance of data increases.

But the problem is when we add an irrelevant feature in the dataset then at that time R2 sometimes starts increasing which is incorrect.

Hence, to control this situation Adjusted R Squared came into existence.

Adjusted R2 score 0.37096434511023446

Actual VS Predicted Values



Random Forest Regressor

Model Fitting

First I calculated feature importance, and according to that score I selected the features and then fitted the model on training data.

Prediction

Predicted the variable after fitting on test data.

Model Evaluation

```
RandomForestRegressor(random_state=0)

Average Error        : 24.2275 degrees
Variance score R^2   : 63.81%
Accuracy             : 77.54%
```

- In random forest regressor we are having R2 score around 64% that means our model is 64% accurately predicting the actual values than linear regression.

Hyperparameter Tuning

For hyper-parameter tuning I used Grid Search CV.

```
RandomForestRegressor(max_depth=800, min_samples_leaf=5, min_samples_split=5,
                      random_state=1)
{'max_depth': 800, 'min_samples_leaf': 5, 'min_samples_split': 5, 'n_estimators': 100, 'random_state': 1}
```

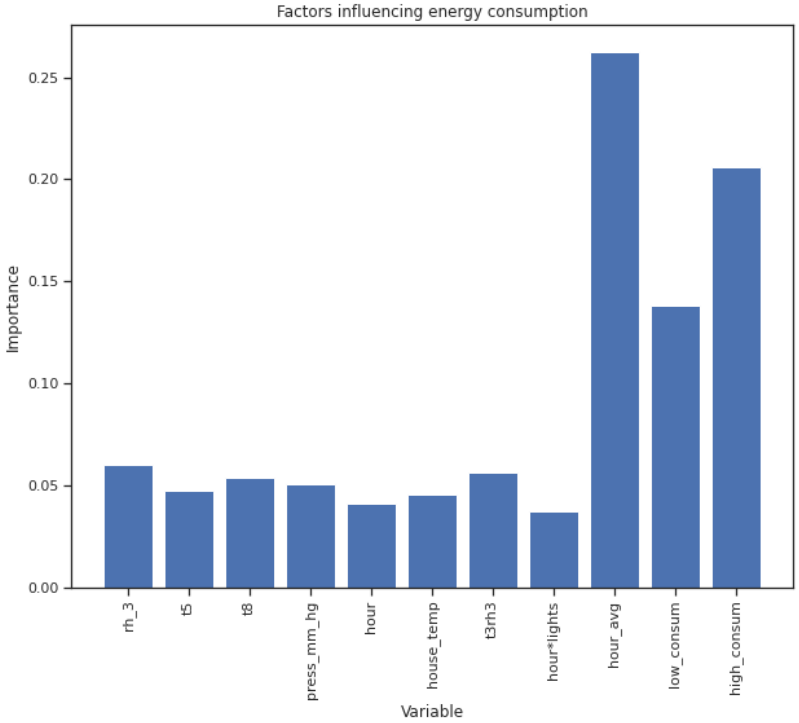
Evaluation after hyper-parameter tuning

```
Average Error        : 24.3381 degrees
Variance score R^2   : 62.93%
Accuracy             : 77.70%
```

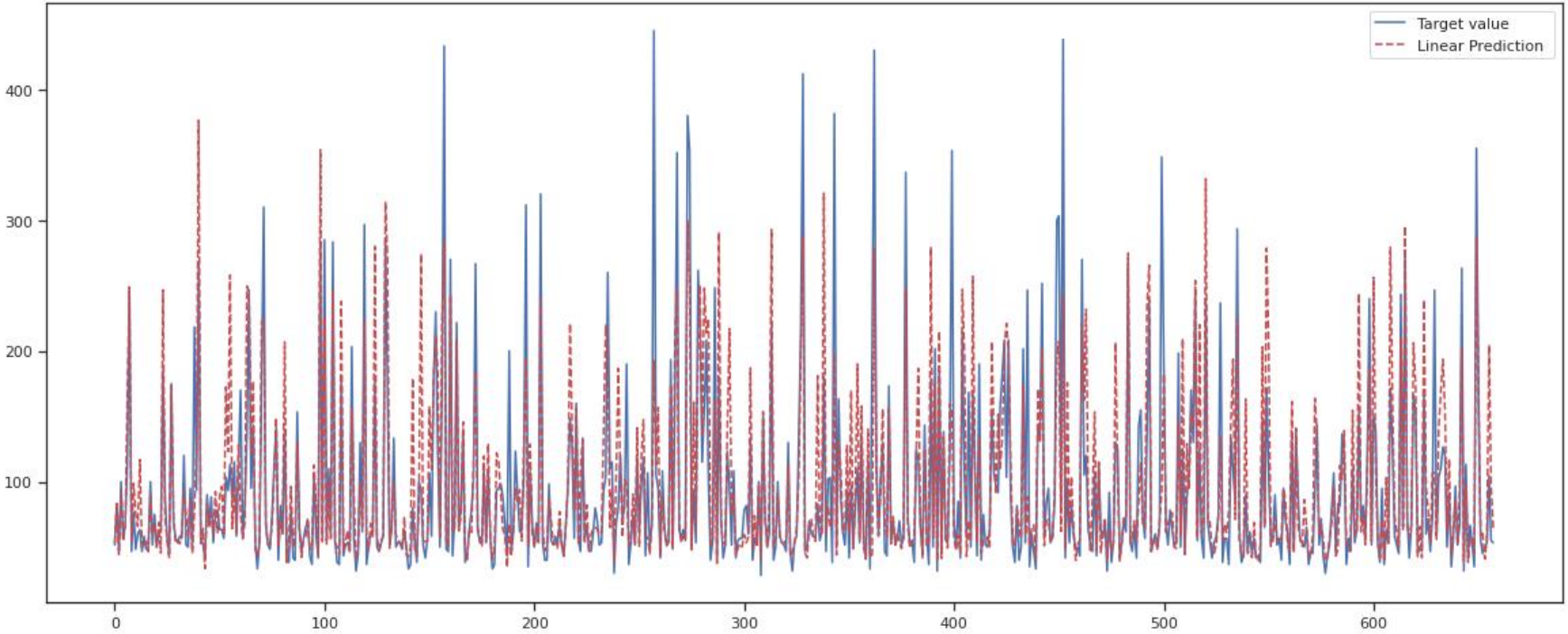
After hyper-parameter tuning we are getting a slight low R2 score so we will continue to use our first set of model.

Feature Importance

```
Variable: hour_avg          Importance: 0.26
Variable: high_consum       Importance: 0.21
Variable: low_consum        Importance: 0.14
Variable: rh_3              Importance: 0.06
Variable: t3rh3             Importance: 0.06
Variable: t5                Importance: 0.05
Variable: t8                Importance: 0.05
Variable: press_mm_hg       Importance: 0.05
Variable: house_temp        Importance: 0.05
Variable: hour              Importance: 0.04
Variable: hour*lights       Importance: 0.04
```



Actual VS Predicted



Conclusion

- The power load is a bit higher on Monday, Friday, Saturday and Sunday than the other days.
- On daily basis, from 7 AM to 11 AM there is clear spike in energy consumption (from 80 wh to 135 wh), because in morning session we use lot of appliances for various purpose.
- From 4 pm to 6 pm there is a sudden hike (from 120 wh to 185 wh) in energy consumption following by a sudden fall from 185 wh to 70 wh during 6 PM to 10 pm in night.
- In linear regression model our R2 score is around 37% which means our model is capturing only 37% variance. We can say it’s not having enough R2 score.
- In random forest regressor our R2 score is around 0.639 which means our model is predicting 64% accurately. And it is much accurate than linear regression model.

Challenges

- Linear regression model has low R2 score, may be because of poor feature selection.
- Had difficulties while choosing features for linear regression model.
- Another regression model can work even better than linear regression and random forest.

Future Scope

- We can select features for linear regression more precisely, so that model can work even better.

- Another regression model can work better on the dataset, so we can try that.

Thank You !