Capstone Project

Supervised Learning : Regression
Appliance Energy Prediction

By - Dipanshu Kumar

Steps Involved:

- Problem Statement
- Exploratory Data Analysis
- Feature Engineering
- Feature Selection for Modelling
- Applying various Models
- Model Validation
- Model Explainability

Problem Statements:

- The availability of energy has transformed the course of humanity over the last few centuries. Not only have new sources of energy been unlocked – first fossil fuels, followed by a diversification to nuclear, hydropower and now other renewable technologies – but also in the quantity we can produce and consume.
- When it comes to residential energy consumption, people are constantly striving for ways to reduce their monthly bills and energy usage and wastage. We use energy every day in a variety of areas of our daily lives. In this project we are considering appliance usage by analyzing the data driven from home sensors.

Problem Statements
(continued):

- All readings are taken at 10 mins intervals for 4.5 months. The goal is to predict energy consumption by appliances. In the age of smart homes, ability to predict energy consumption can not only save money for end user but can also help in generating money for user by giving excess energy back to Grid (in case of solar panels usage).
- In this case of regression analysis will be used to predict Appliance energy usage based on data collected from various sensors. The main objective is to build a predictive model, which could help them in predicting the Energy usage proactively.

Introduction :

- The dataset is used for data driven predictive modelling of the energy usage. Data used here contains assessment of the temperature and humidity using wireless sensors, weathering conditions like pressure, windspeed, visibility is taken from a nearby airport
- It is important to understand the energy consumption behaviors in the residential areas and predict the energy usage by home appliances to decide the energy management and reduce the consumptions. This project focuses on predicting the energy consumption of appliances based on temperature, pressure, humidity, windspeed, visibility.

Dataset Contains :

The dataset from above understanding consists of 19735 rows and 29 columns with no duplicate and no null values. The features i.e. the columns are selected based on the observation from ZigBee Wireless sensors and weather from an airport weather station apart from this we also have included two random variables for testing and filtering out nonpredictive attributes all these observations are combined together for our further process

Purpose of this Project:

The main objective of this project is to predict the energy consumption by the home appliances. With the oncoming of smart homes and the rising need for energy management, existing smart systems can benefit us with accurate prediction. If the energy usage can be predicted for every possible state of appliances, then device control can be optimized for energy savings as well.

Data Set Description:

The dataset that we are using consists of a number of features:

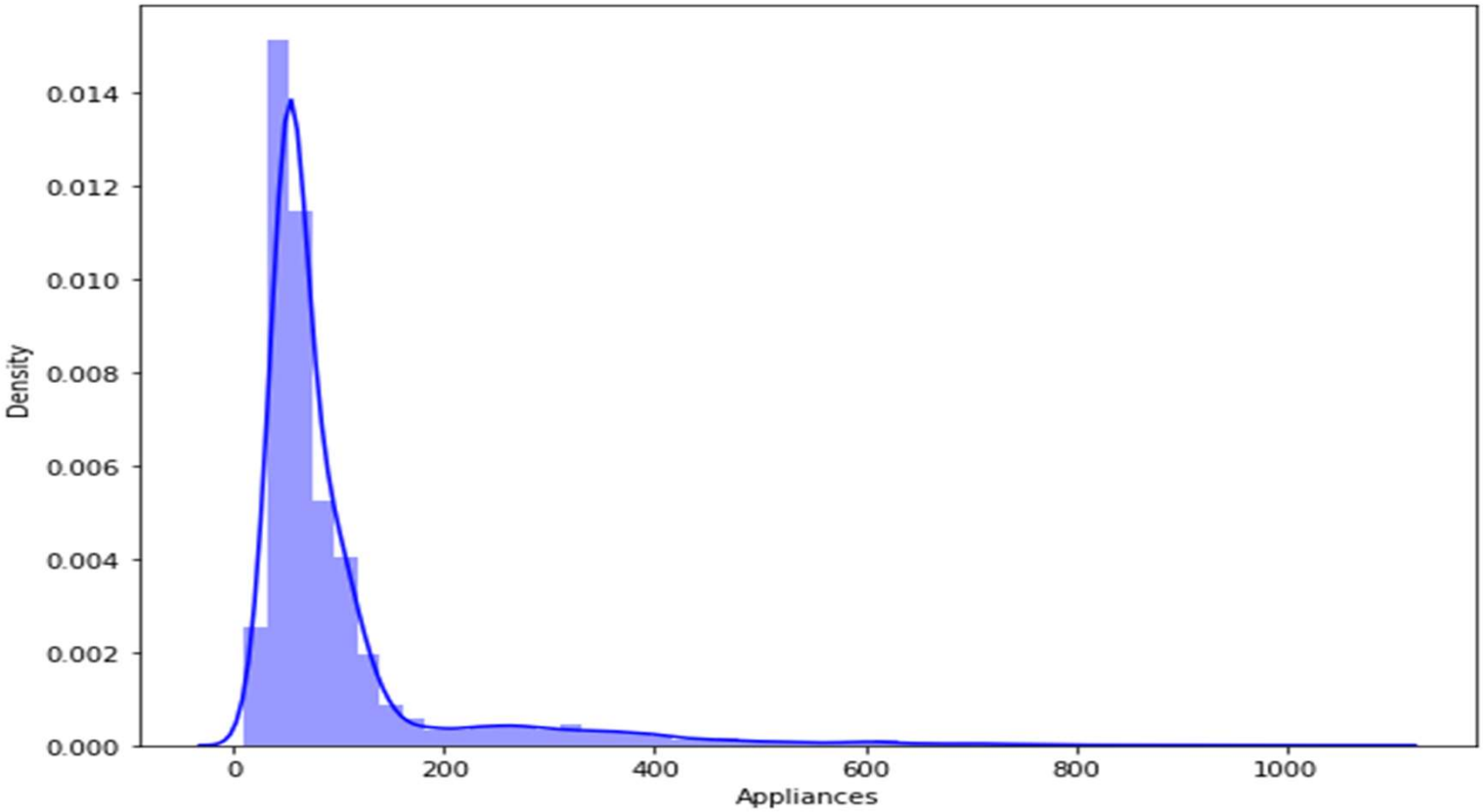
- **Temperature:** The temperatures are recorded for various places (using sensors) like living room, bathroom, kitchen, laundry and also outside area.
- **Humidity:** Similar to temperature, humidity is also measured for living room, bathroom, kitchen, laundry and outside area.
- **Pressure:** The pressures are recorded in mmHg.
- **Visibility:** The visibility is present in km.
- **T Dewpoint:** This tells us about the dew point temperature.
- **Appliances:** The energy consumed in Wh. This column is our dependent variable.

Data Insights:

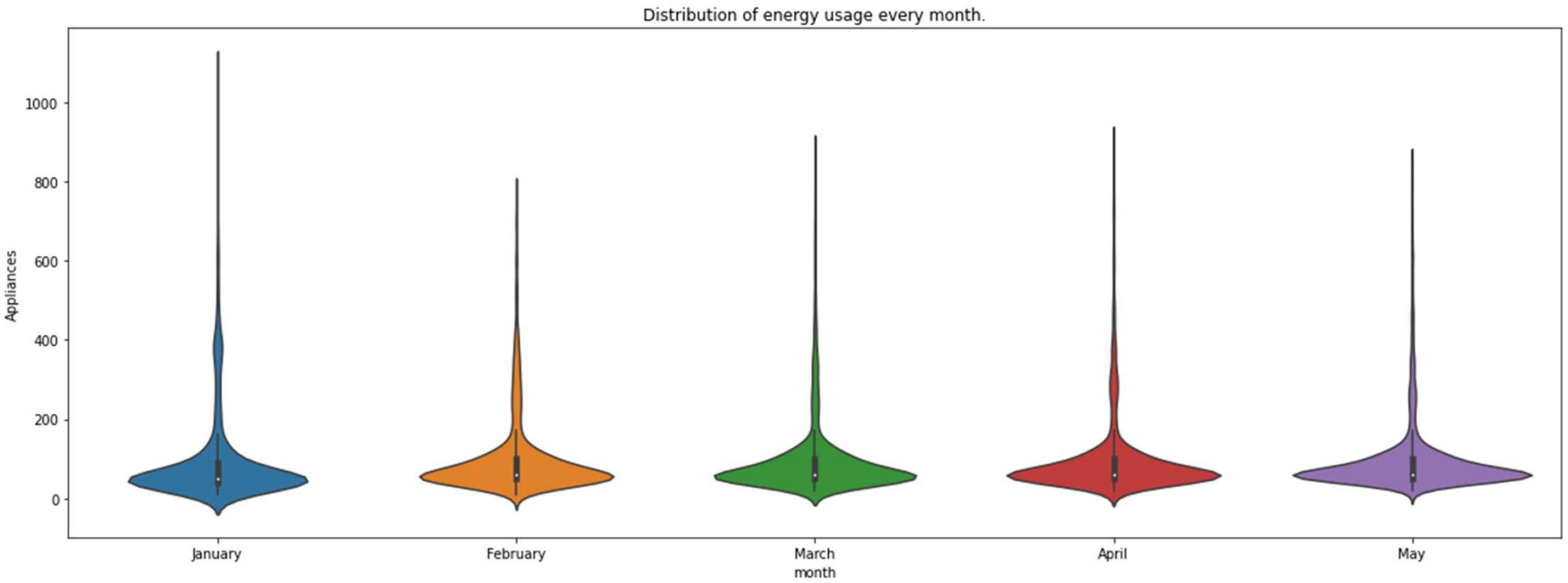
From the description of data we came with the following observations:

- There are 29 columns and 19735 rows in our dataset.
- The maximum energy consumption of the appliance is 1080 watts, while the minimum is 10 watts.
- The majority of the data in the light contains no values.
- The maximum pressure outside the home is 772.3 mmHg.
- There are no categorical columns in the dataset other than the date column.
- There are no null or missing values.
- Outside humidity is higher than inside humidity.
- The maximum wind speed is 14 m/s.

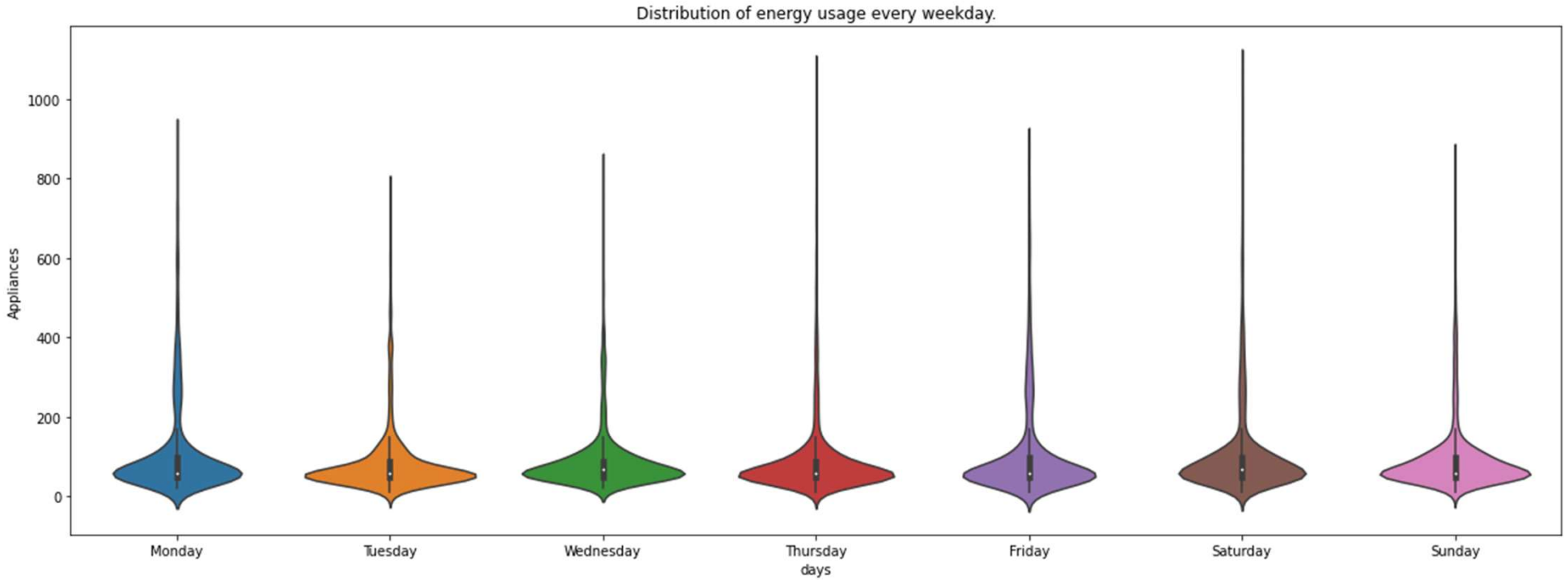
Appliance Energy Distribution:



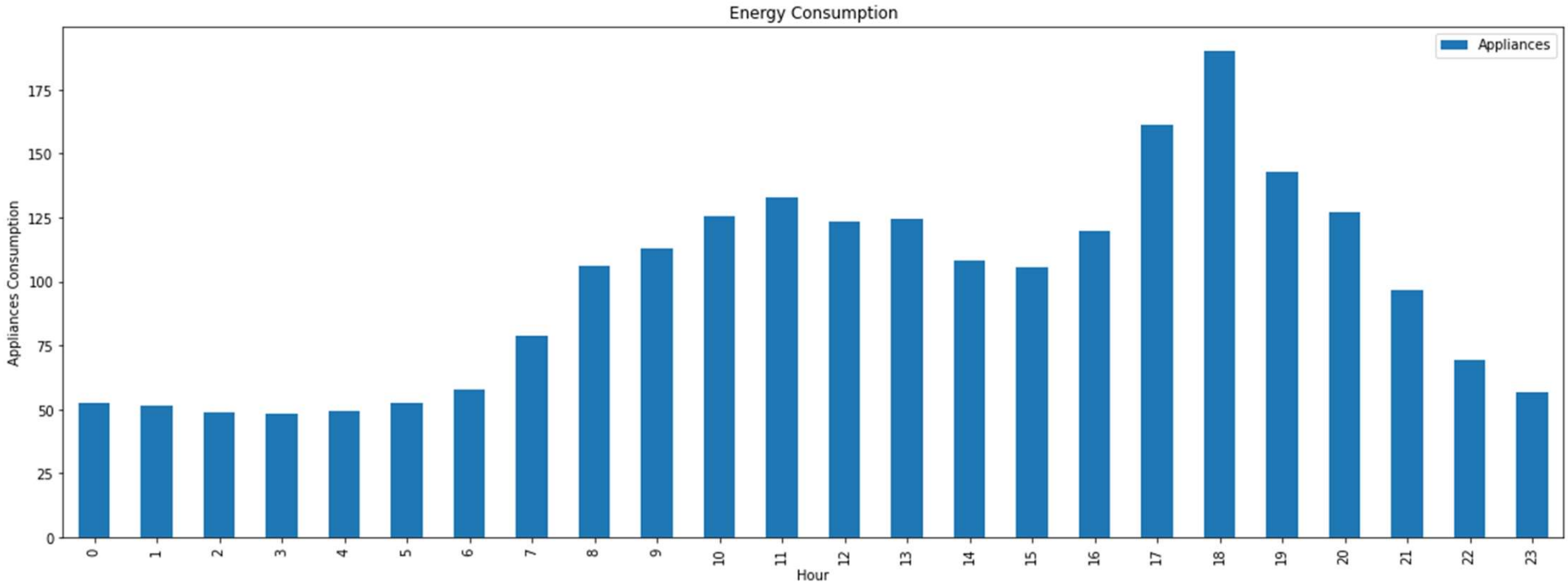
Distribution of Energy Usage Every Month:



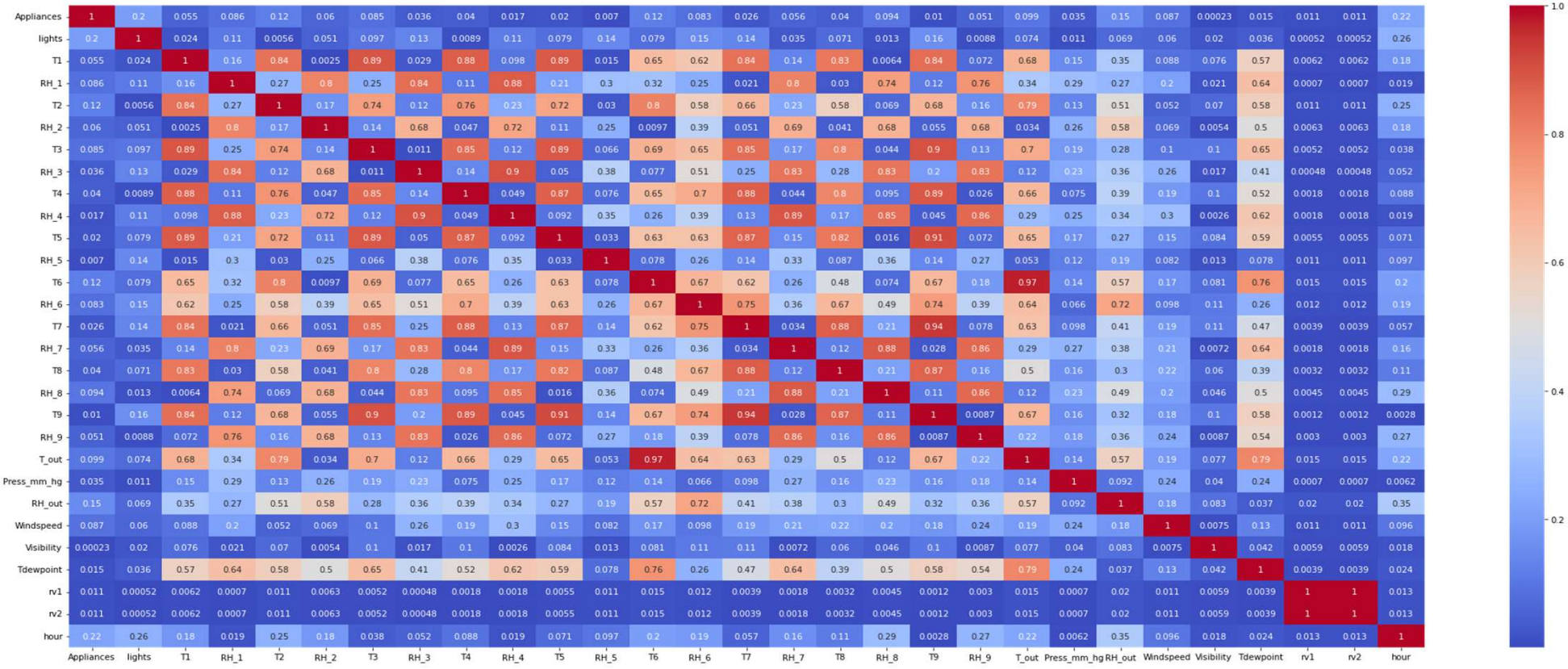
Distribution of Energy Usage Every Weekdays:



Distribution of Energy Usage Every Hour:



Heatmap, Determining Correlations:



Preparing dataset for Modelling:

- Train, test :-(80% and 20%)

- X_Train set :- (10936, 24)
  X_Test set :- (2734, 24)

- Y_Train set :- (10936,1)
  Y_Test set :- (2734,1)

- Dependent or Target Variable :- Appliances

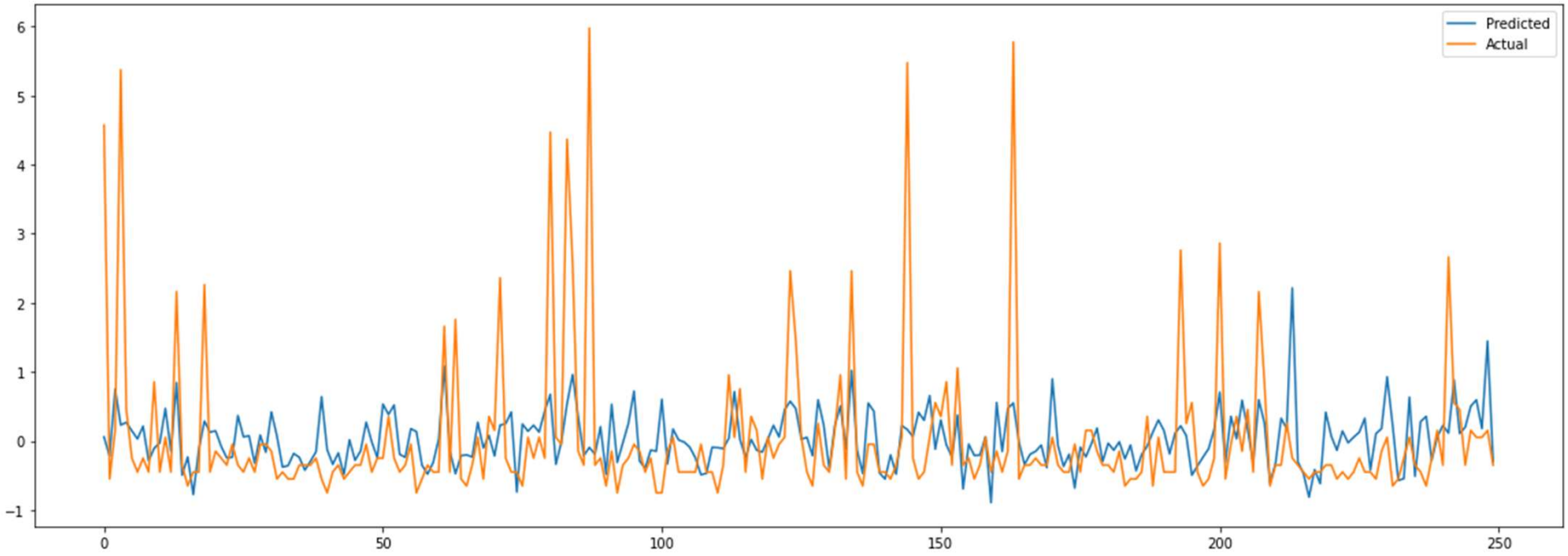
- Due to different ranges of features it is possible that some of the features will dominate the regression algorithm. To avoid this , we have scaled all the features.

```
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()

x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)

y_train = scaler.fit_transform(y_train)
y_test = scaler.fit_transform(y_test)
```

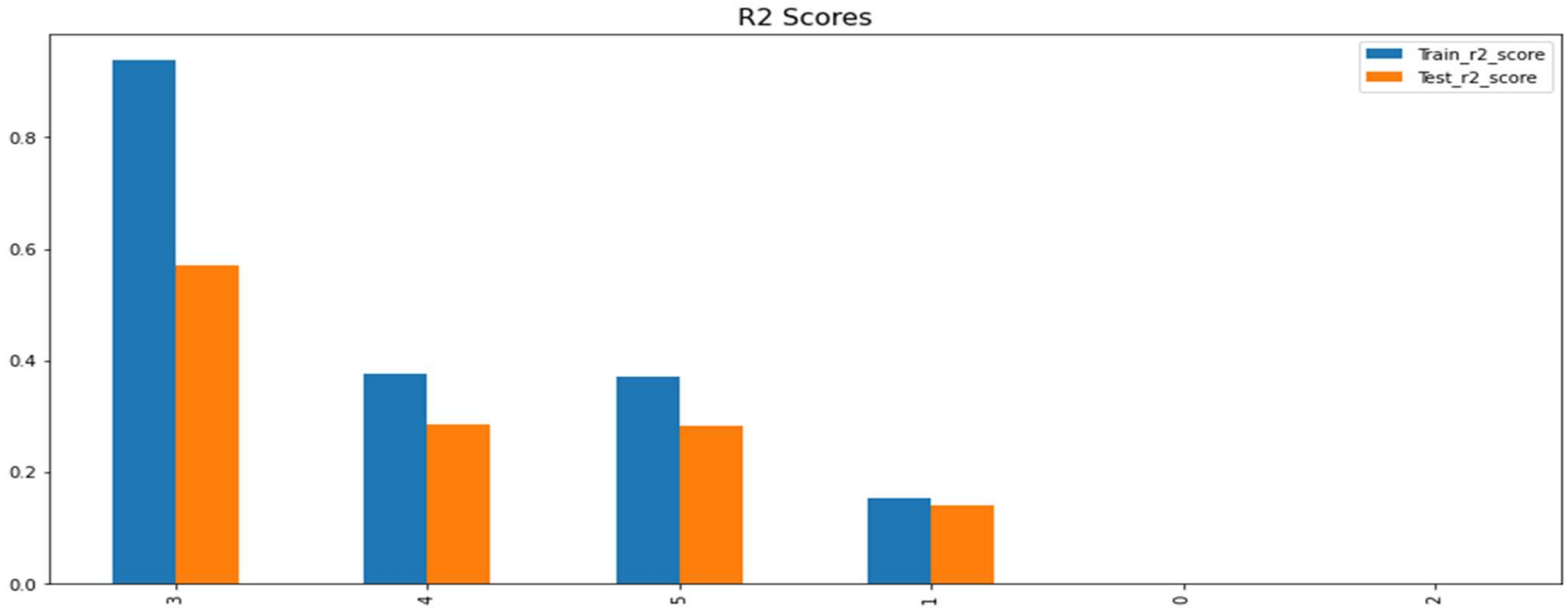
Linear Regression Result Analysis:



Analysis of various Models Results:

	Name	Train_r2_score	Test_r2_score	Train_MSE	Test_MSE	Train_RMSE	Test_RMSE
3	RandomForest :	0.938141	0.569606	0.061859	0.430394	0.248715	0.656044
4	Gradientboosting :	0.377732	0.284910	0.622268	0.715090	0.788840	0.845630
5	Xgboost :	0.370691	0.283070	0.629309	0.716930	0.793290	0.846717
1	Ridge :	0.153399	0.140678	0.846601	0.859322	0.920109	0.926996
0	Lasso :	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
2	ElasticNet :	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000

R2 Scores Compression:



Cross Validation and Hyperparameter tuning:

Name	Train_r2_score	Test_r2_score	Train_MSE	Test_MSE	Train_RMSE	Test_RMSE
Enhanced Random Forest:	0.942470	0.595650	0.057520	0.404300	0.239840	0.635870
RandomForest :	0.938141	0.569606	0.061859	0.430394	0.248715	0.656044
Gradientboosting :	0.377732	0.284910	0.622268	0.715090	0.788840	0.845630
Xgboost :	0.370691	0.283070	0.629309	0.716930	0.793290	0.846717
Ridge :	0.153399	0.140678	0.846601	0.859322	0.920109	0.926996
Lasso :	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
ElasticNet :	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000

Model Interpretation Using ELI5:

Weight	Feature
0.0908 ± 0.0702	x22
0.0497 ± 0.0274	x5
0.0494 ± 0.0254	x4
0.0467 ± 0.0260	x14
0.0461 ± 0.0233	x1
0.0460 ± 0.0234	x17
0.0444 ± 0.0245	x10
0.0437 ± 0.0215	x3
0.0429 ± 0.0254	x12
0.0414 ± 0.0241	x18
0.0405 ± 0.0237	x21
0.0400 ± 0.0249	x7
0.0397 ± 0.0240	x2
0.0396 ± 0.0221	x9
0.0395 ± 0.0232	x15
0.0390 ± 0.0206	x16
0.0382 ± 0.0237	x8
0.0381 ± 0.0217	x13
0.0380 ± 0.0254	x6
0.0355 ± 0.0207	x11
... 4 more ...	

Contribution?	Feature	Value
+0.291	RH_3	45.590
+0.212	RH_4	47.030
+0.144	T3	20.500
+0.142	T8	19.290
+0.121	T2	21.000
+0.112	Windspeed	6.833
+0.083	T5	19.223
+0.080	RH_1	44.200
+0.076	RH_7	42.627
+0.074	RH_6	84.060
+0.053	rv_mean	33.423
+0.051	T_out	5.850
+0.021	T4	20.997
+0.015	Tdewpoint	3.783
+0.014	Press_mm_hg	736.217
+0.013	RH_5	51.520
+0.012	Visibility	40.000
+0.011	RH_out	87.167
+0.007	T7	17.700
+0.001	RH_8	49.230
-0.000	<BIAS>	1.000
-0.001	hour	22.000
-0.002	T1	21.600
-0.014	RH_9	44.500
-0.068	RH_2	43.700

Outcomes:

Observation 1:

Implementing the Linear Regression model does not give a good result. The Matrices calculated clearly define that there is a requirement of implementing different models.

Observation 2:

After using different models namely Lasso, Ridge, ElasticNet, Random Forest Regressor, Gradient Boosting and XGBoost the results shows that Random forest Regressor gives the best prediction

Observation 3:

Cross Validation and Hyperparameter Tuning is a process to find the best possible parameters for our model. We have tuned our Random Forest Regressor and got an Enhanced version of it.

Observation 4:

Using ELI5 technique we came to understand the functioning of our model which states that the impact of feature 22 is the highest followed by 5, 4, 14 and so on.

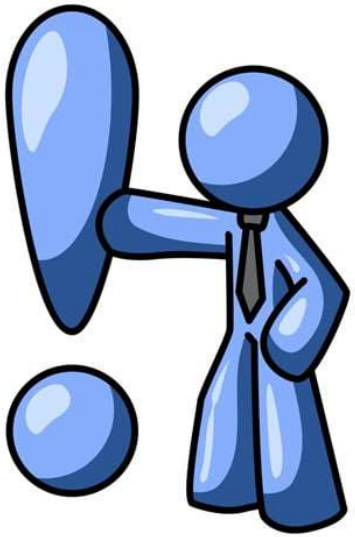
Conclusion:

- **Our main objective is to predict the Energy usage** by the Appliance to achieve this we have Implemented the XGBoosting, Decision tree, Random forest, Gradient Boosting, LinearRegression and Regularized Linear Regression algorithms was done along with cross validation and hyperparameter adjustment .

- **In a comparison of all models, the RandomForest regressor is the best, having a high r2 score, a low MSE, and a low RMSE value(Comparing to other models).** The model explainability Eli5 approach is used to determine which attributes are crucial for predicting output and understanding the model.

Conclusion (continued):

- Tree based models are by far the best model while dealing with this type of data set because of **its ability to stay insulated from the effects of worse features**. For similar reasons, linear models such as linear regression, Ridge and Lasso performances are not up to the mark.



Challenges :

- Features selection was the most challenging task, reason being every feature was important, removal of feature to reduce multicollinearity seem infeasible. Definitely there is a scope for improvement.
- The length of time required especially during Hyperparameter Tuning to perform computational process was consuming.

Thank You