Capstone Project

Appliances Energy Prediction

Chetan Chavan

Contents:

Problem Statement:

- As the energy crisis increasing day by day it affects on various factors. Its not only affecting the world energy requirement issues but also affects the economic and social health of any country, mostly in the countries which are going through developing phase. To overcome several related issues, it is necessary to have the outages in order to compensate the load demand. And hence the prediction of energy used by appliances plays a vital role in this concept.

- This project illustrates the energy consumed by house and the data has been collected with the help of sensors. The readings have been taken for each 10 min interval for consecutive 4.5 months. Saving of energy can be done by controlling the energy usage. And thus, prediction of usage comes into picture. This study can save the money of consumer as well as if extra energy is generated then it can also fed back to Grid(Also called as regeneration). We are going to focus on many machine learning regression technique to find out the energy consumption prediction.

Variables:

- date: time→ given dat time month and day
- lights : energy used by lights in Wh
- T1 : Temperature given in kitchen area, in Celsius
- T2 : Temperature given in living room area, in Celsius
- T3 : Temperature mentioned in laundry room area
- T4 : Temperature of office room, given in CelsiusT5 : Temperature recorded in bathroom area, in Celsius
- T6 : Temperature given outside the building area particularly (north side), in Celsius
- T7 : Temperature provided in ironing room, in Celsius
- T8 : Temperature in teenager room 2, in Celsius
- T9 : Temperature in parents' room, in Celsius

- T_out : Outside temperature (from Chievres weather station), in °C
- Tdewpoint : (from Chievres weather station),
- RH_1 : Kitchen area Humidity %
- RH_2 : Living room area Humidity, in %
- RH_3 : Laundry room area Humidity, in %
- RH_4 : Office room Humidity, in %
- RH_5 : Bathroom area Humidity, in %
- RH_6 :Outside the building Humidity (north side), in %
- RH_7 : Ironing room Humidity, in %
- RH_8 : Teenager room 2 Humidity, in %
- RH_9 : Parents' room Humidity, in %
- RH_out :Outside Humidity (from Chievres weather station), in %
- Pressure : (from Chievres weather station), in mm Hg

- Wind speed: (from Chievres weather station), in m/s
- Visibility :(from Chievres weather station), in km
- Rv1 :Random variable 1, non-dimensional[1]
- Rv2 :Random variable 2, non-dimensional[1]
- Appliances : Total energy used by appliances, in Wh[1]

Energy Crisis and its Solution:

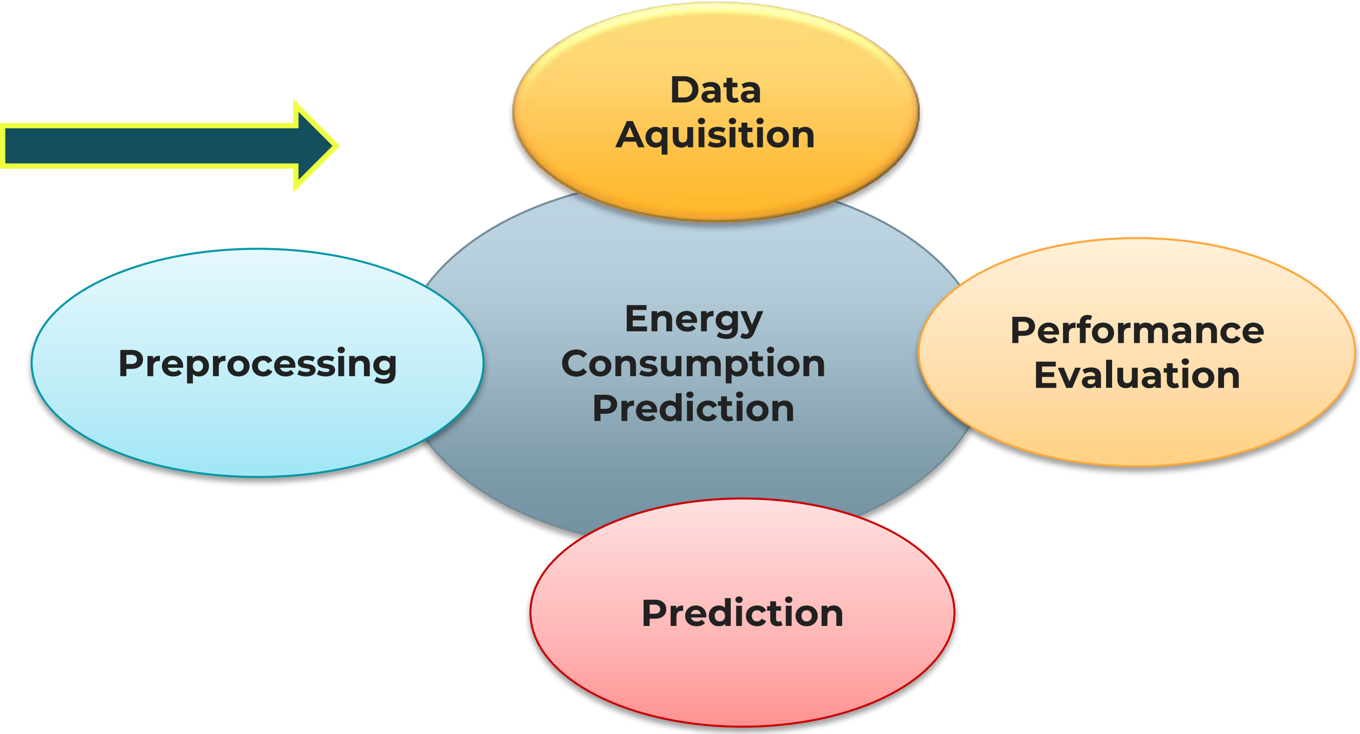


Fig: Prediction methodology for energy consumption using different machine learning algorithms

Data set Division:

There are total 19735 no of rows available. We will go to divide this data into pre-training set, training set, validation set and testing set. The division of data is as follows:

- 75% of data will be put into training set whereas 25% of data will be put into testing set.

- The pre train set was used to find the best models for the given dataset. We have taken best 6 models using pretest set. Their performance will be compared based on their mean absolute errors.

- Once the best 5 models will be obtained, hyper parameters for these models will be tuned and the best parameter will be selected.

Training Process:

We will use following 5 regression techniques to train the data:

1] LASSO Regression: It is the regression which uses the shrinkage technique. Which means data values will be shrunk towards the central point. The Lasso regression is very useful when data parameters are few. The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator.

Lasso solutions are quadratic programming problems, which are best solved with software (like Matlab). The goal of the algorithm is to minimize: lasso regression

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Which is the same as minimizing the sum of squares with constraint Σ |Bj≤ s (Σ = summation notation). Some of the βs are shrunk to exactly zero, resulting in a regression model that's easier to interpret.

2] RIDGE Regression: This regression method is mainly used when data having multi-collinearity. The method performs L2 regularization . Whenever multi-collnearity problem occurs, least-square are unbiased and variance are large. Because of it predicted value being far away from the actual values.

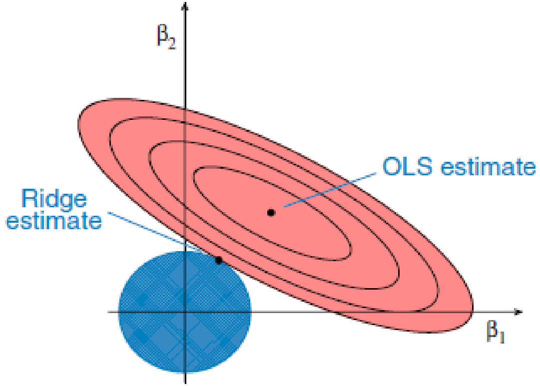


Fig. Ridge Regression

3] Random Forest: Random Forest Regression method is a supervised learning algorithm which uses ensemble learning method for regression technique. Ensemble learning method is nothing but a technique which combines predictions from various machine learning algorithms to prepare a more accurate prediction as compare to the single model.

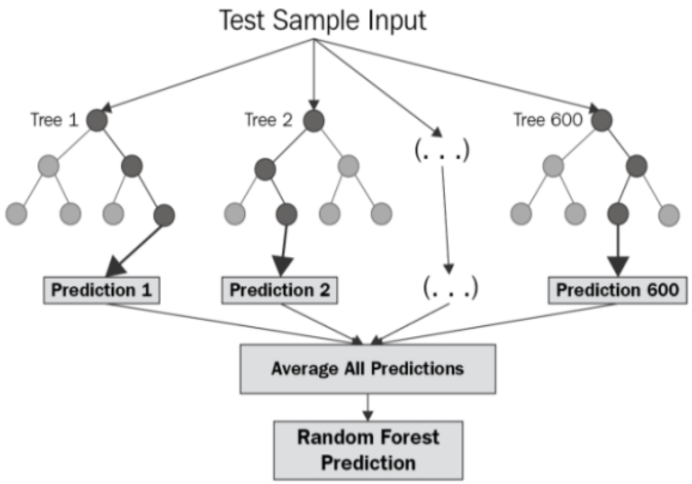


Fig: Random Forest regression

4] Gradient Boosting Classifier: This regression technique calculates the difference between the current predicted value and well known correct target value. This residual is then added to the existing model and this pushes the model towards correct values. To improve the performance of the model we can repeat the process again and again.

5] ExtraTree-regressor: It is a type of ensemble learning technique of regression that adds the results of different de-correlated decision trees which are similar to Random Forest Classifier.Extra Tree can also achieve a good or better prediction than the random forest. The main difference between Random Forest and Extra Tree Classifier is as given below:

• Extra Tree Classifier algorithm never performs bootstrap aggregation as in the random forest. This means, it takes a random subset of data without any replacement. Hence, nodes are always split on random splits but not on best splits.

•In Extra Tree Classifier algorithm randomness doesn't come from bootstrap aggregating but comes from the random splitting of the data.

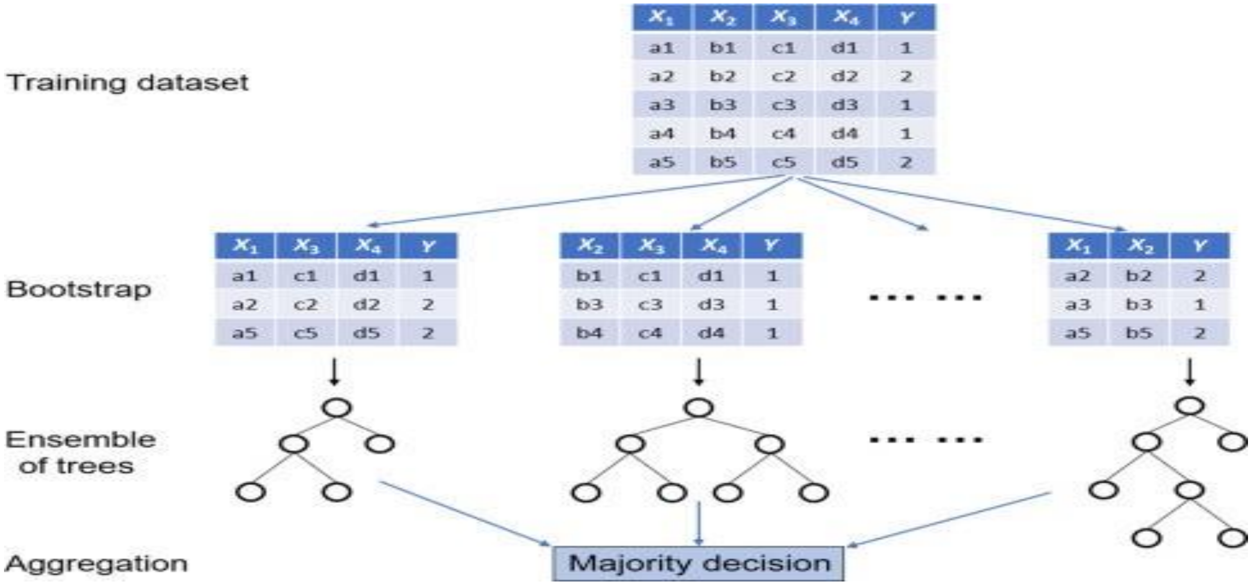
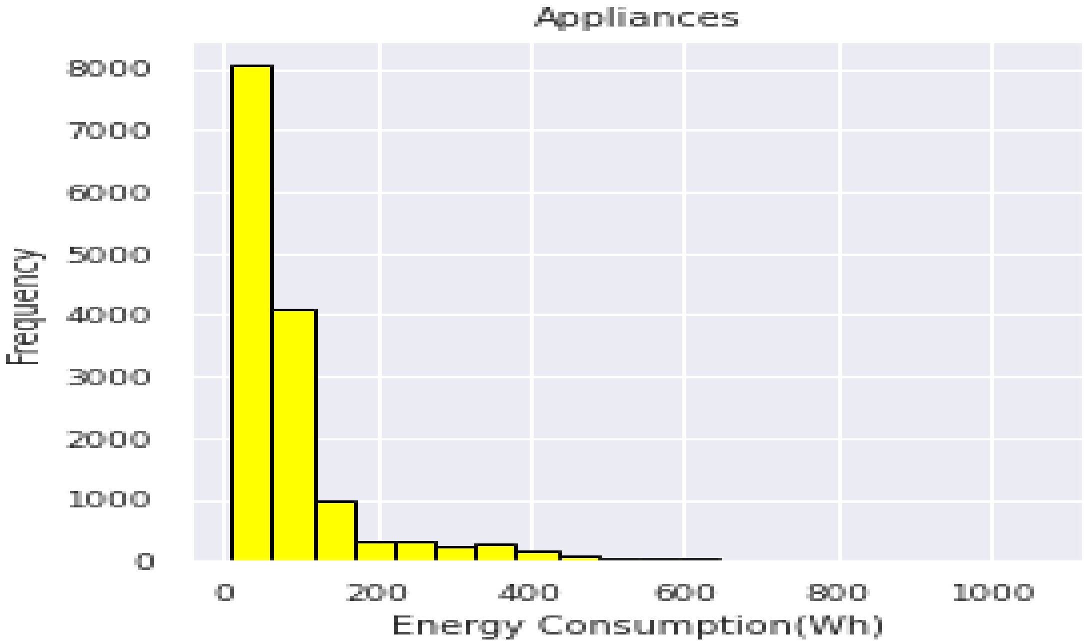


Fig. ExtraRegressor

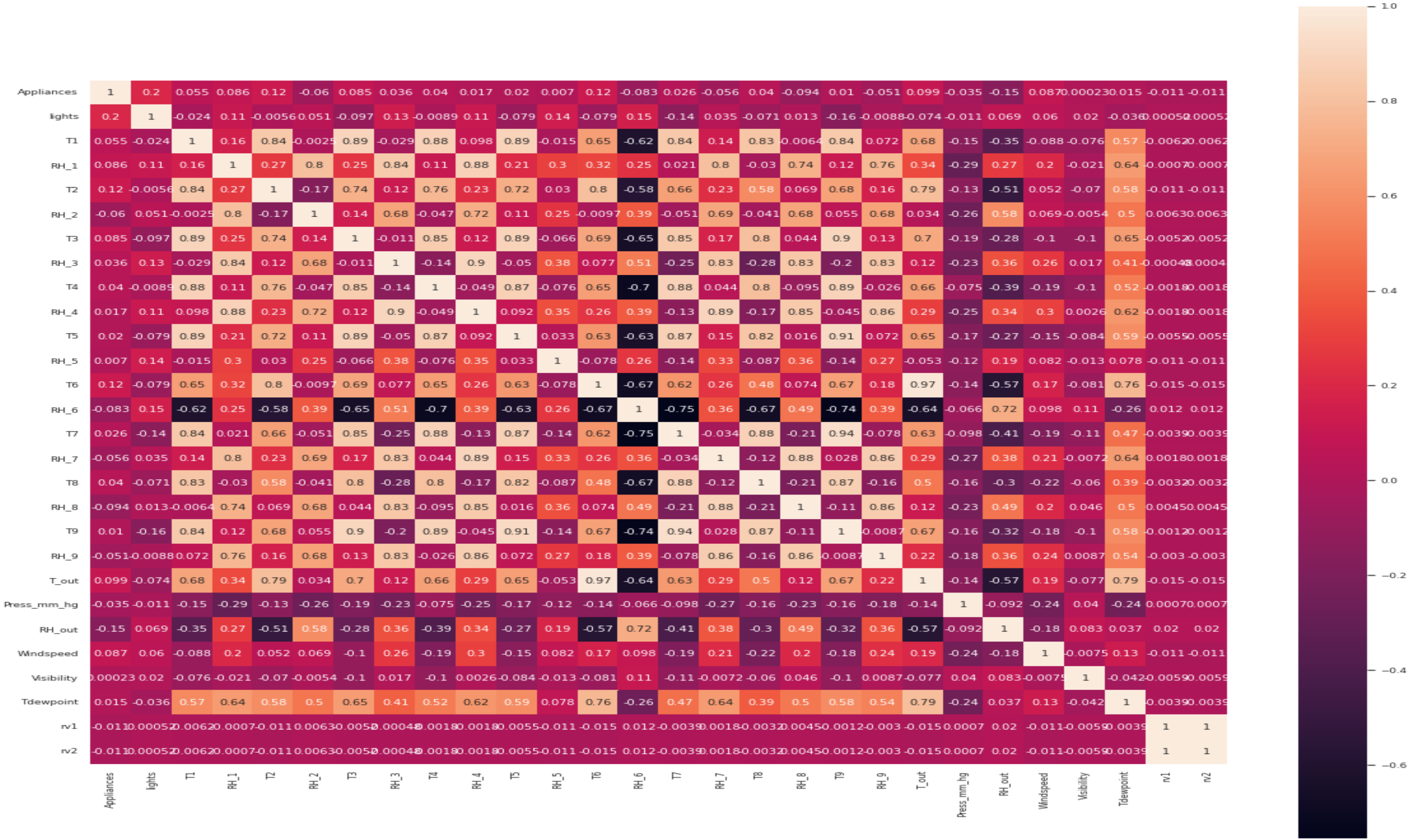
Visualization:

1] Appliance vs. Energy consumption:

When we will plot the appliance against the energy consumption graph, we can see that percentage of appliances consumption is less than 200Wh. The representation is as follows



2] Correlation plot: From the correlation plot, we can see that Temperature values T1-T9 have positive correlational values. For the observation of indoor temperatures, the correlations are looking like high, since the ventilation is characterized by the HRV unit and that minimizes air temperature differences between the rooms. There are four columns which have a high degree of correlation with T9 - T3,T5,T7,T8 also T6 & T_Out has high correlation (both temperatures from outside) . The figure is shown in next slide.



Appliances
lights
T1
RH_1
T2
RH_2
T3
RH_3
T4
RH_4
T5
RH_5
T6
RH_6
T7
RH_7
T8
RH_8
T9
RH_9
T_out
Press_mm_hg
RH_out
Windspeed
Visibility
Tdewpoint
rv1
rv2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

3] Feature Importance: From feature importance graph we can analyzed that most important features are 'RH_out', 'RH_8', 'RH_1', 'T3', 'RH_3'. Whereas least important features will be - 'T7','Tdewpoint','Windspeed','T1','T5'.

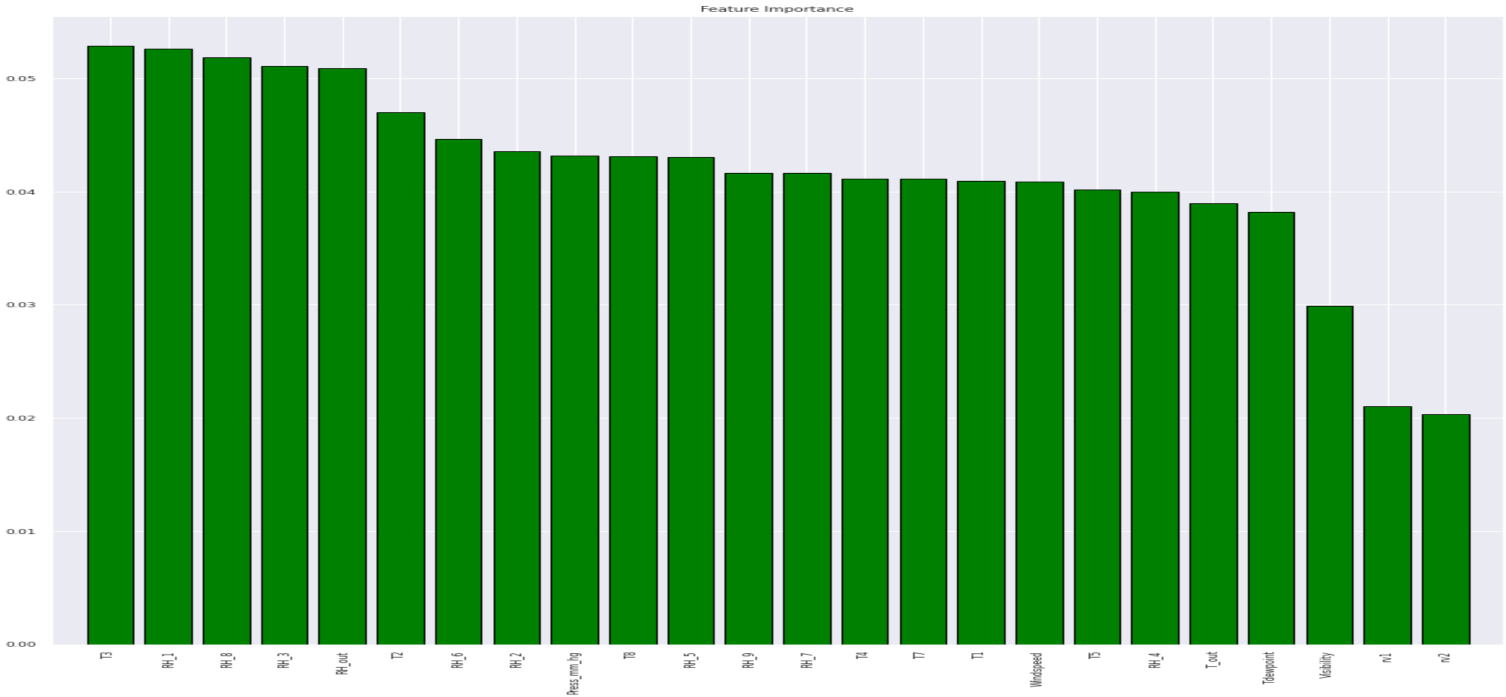
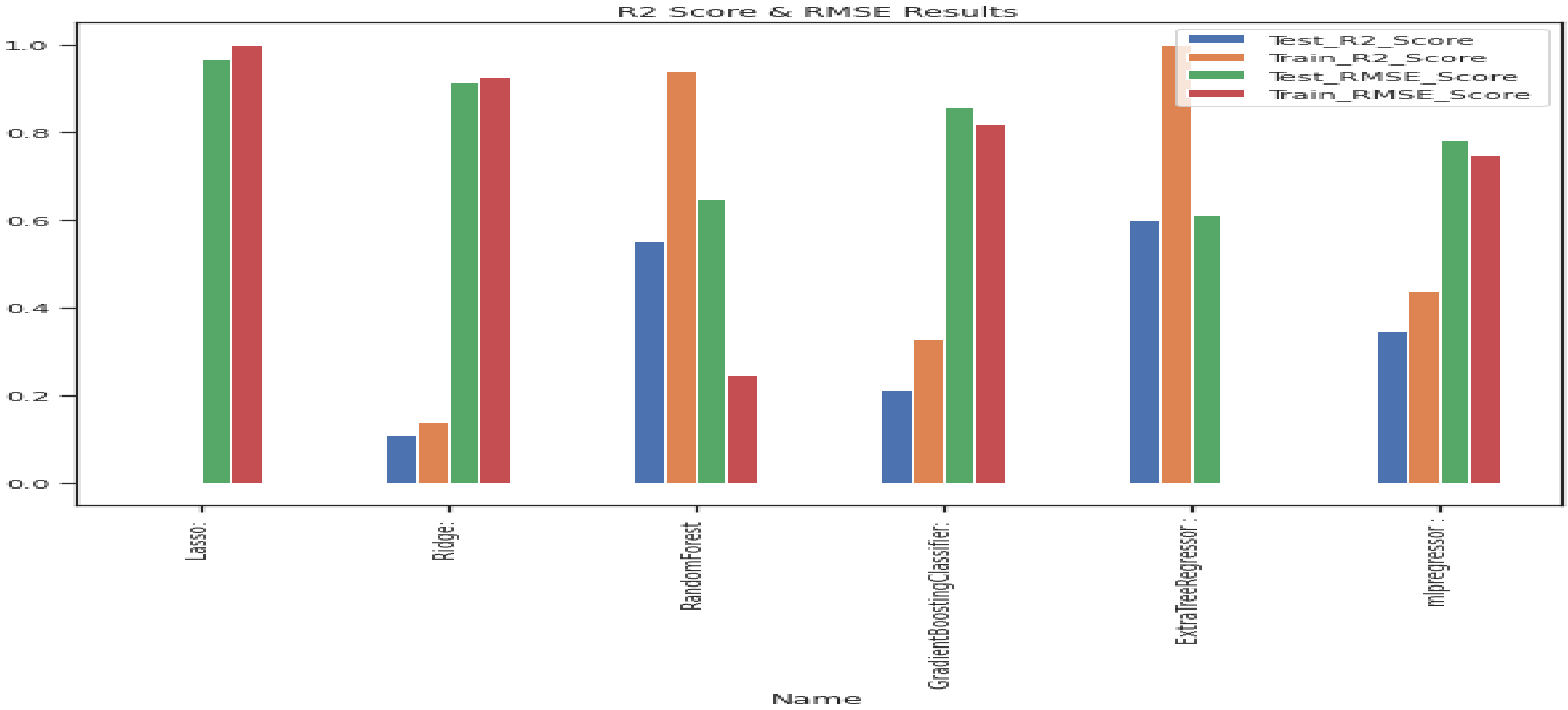


Fig. Feature graph

4] Comparison Graph:



Results:

	Name	Train_R2_Score	Test_R2_Score	Train_RMSE_Score	Test_RMSE_Score
0	Lasso	0.000000	-0.000517	1.000000e+00	0.968239
1	Ridge	0.140903	0.109054	9.268748e-01	0.913684
2	RandomForest	0.939403	0.552971	2.461646e-01	0.647200
3	GradientBoostingClassifier	0.329343	0.214281	8.189363e-01	0.858033
4	ExtraTreeRegressor	1.000000	0.599255	1.326785e-15	0.612780
5	mlpregressor	0.438139	0.346108	7.495736e-01	0.782750

Conclusion:

1) It is clearly seen that best results for test set is being given by Extra Tree Regressor with R2 score of 0.599255

2) Least RMSE score is also by Extra Tree Regressor 0.612780

3) Lasso regression model over Linear regression was not giving good result and hence proven to be the worst model.

Parameter Tuning and observation:

Depending on parameter tuning we can conclude that

- Best possible parameter combination are - 'max_depth' is 100, max_features is 'sqrt', 'n_estimators' is 260 and random state is 40.

- Training set R2 score of 1.0 shows the overfitting on training set.

- Using hyperparameter tuning the R2 score can be improved from 0.59 to 0.60 of the Test set.

- Test set RMSE score is 0.60 which is get improved from 0.61 achieved using hyperparameter tuning.

THANK YOU……..