Capstone Project
Appliances Energy Prediction
(Supervised ML - Regression)

Appliances Energy Prediction

Problem Statement

01

Make predictions

Prediction of the **Energy Consumption by the Appliances** of a household based on factors like **temperature, humidity and pressure.**

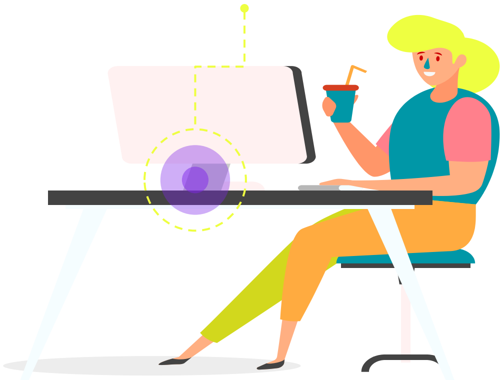
02

Model Development

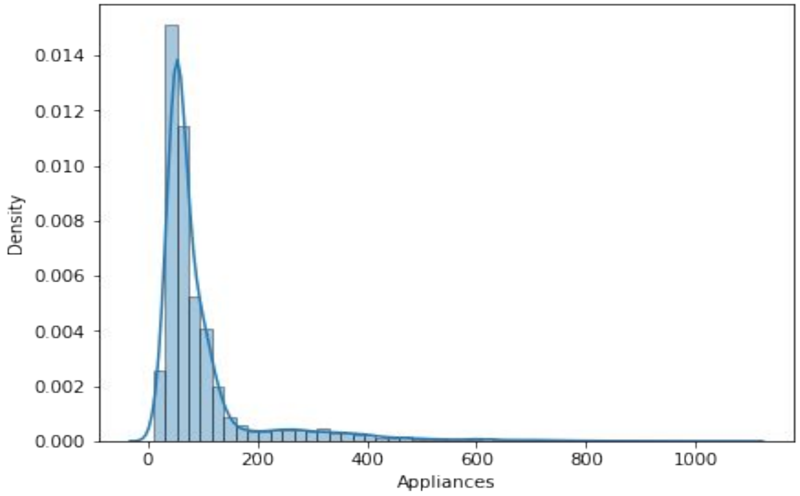
Develop a **Supervised Machine Learning Model** using **regression.**

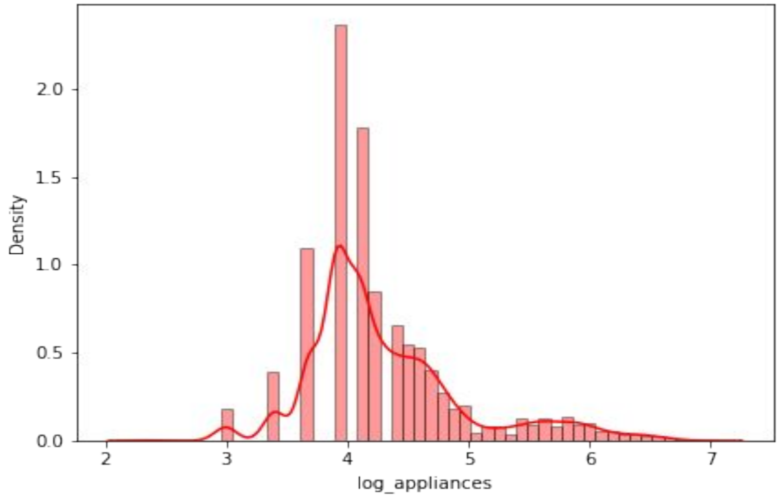
Features

01 date
02 Appliances - Energy in Watt
03 T1 - temperature in kitchen area
04 RH1 - humidity in kitchen area
05 T2 - temperature in living room
06 RH2 -humidity in living room
07 T3-temperature in laundry
08 RH3- humidity in laundry
09 T4- temperature in office room
10 RH4-humidity in office room
11 T5-temperature in bathroom
12 RH5- humidity in bathroom
13 T6-temperature outside the building
14 RH6-humidity outside the building
15 T7-temperature in ironing room



RH7 - Humidity in ironing room 16
T8 - Temperature in teenager room 17
RH8 -Humidity in teenager room 18
T9 - Temperature in parents room 19
RH9 - Humidity in parents room 20
Tout - Temperature outside 21
RHout - Humidity outside 22
Press_mm_hg 23
Visibility 24
Windspeed 25
Tdewpoint 26
rv1 27
rv2 28
lights 29





Appliances

Positively Skewed Distribution

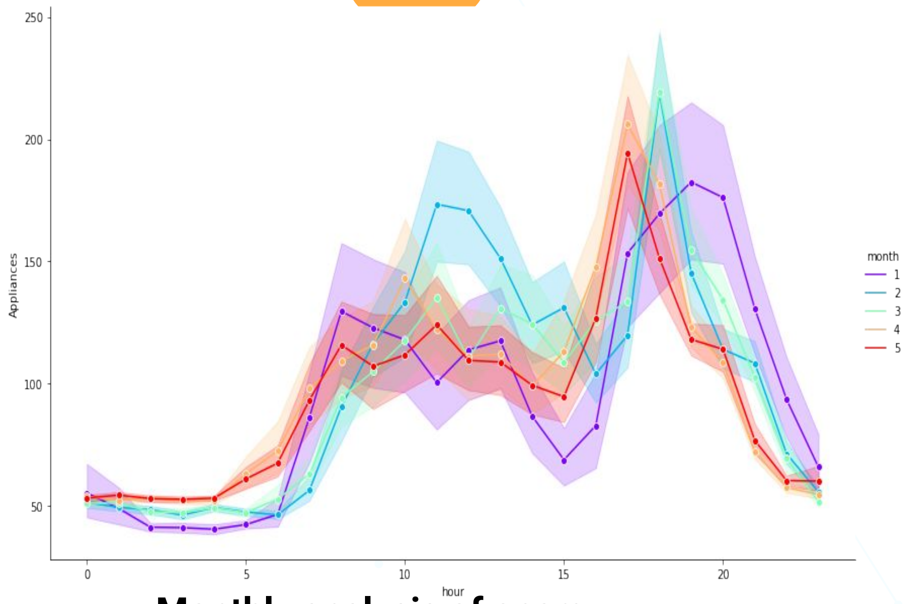
Vs

Log Transformed Appliances

Normal Distribution

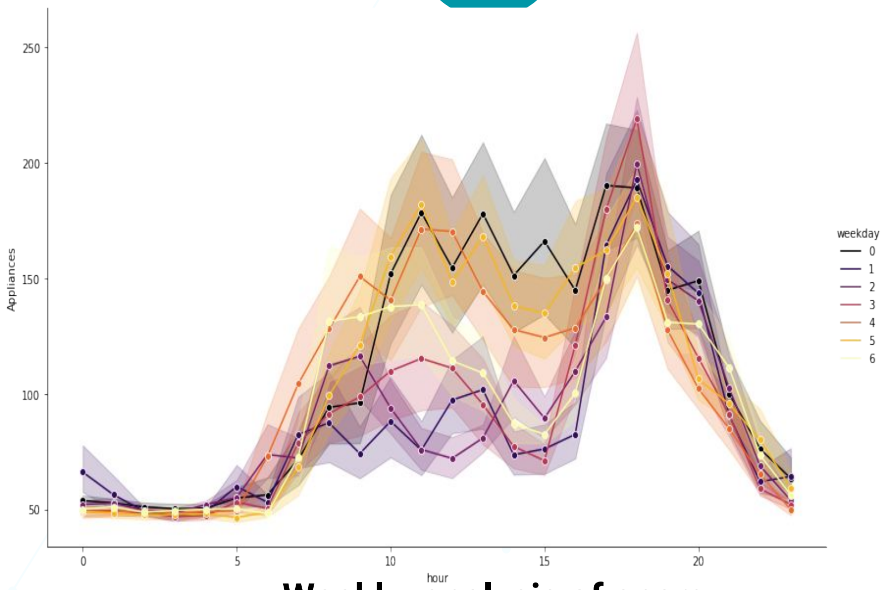
Multivariate analysis with respect to Appliances

01



Monthly analysis of energy consumption on hourly basis.

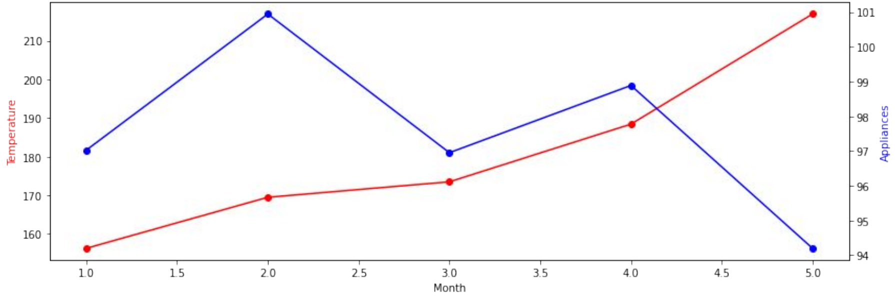
02



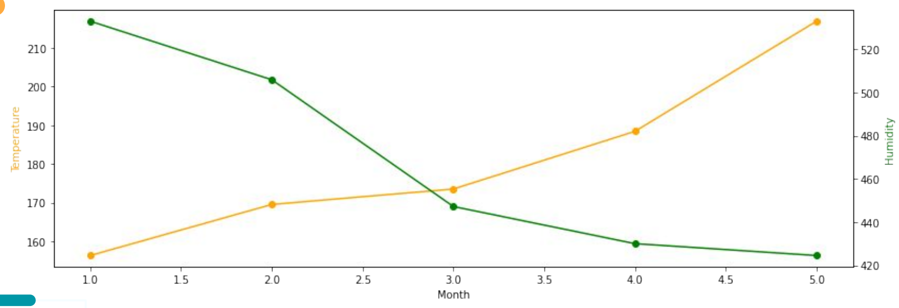
Weekly analysis of energy consumption on hourly basis.

Monthly Analysis of different variables

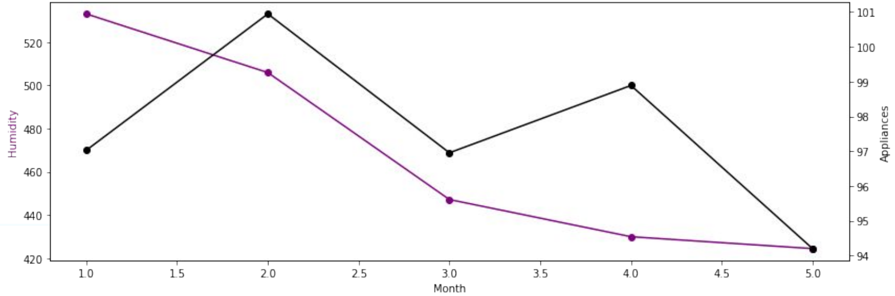
Appliances vs Temperature



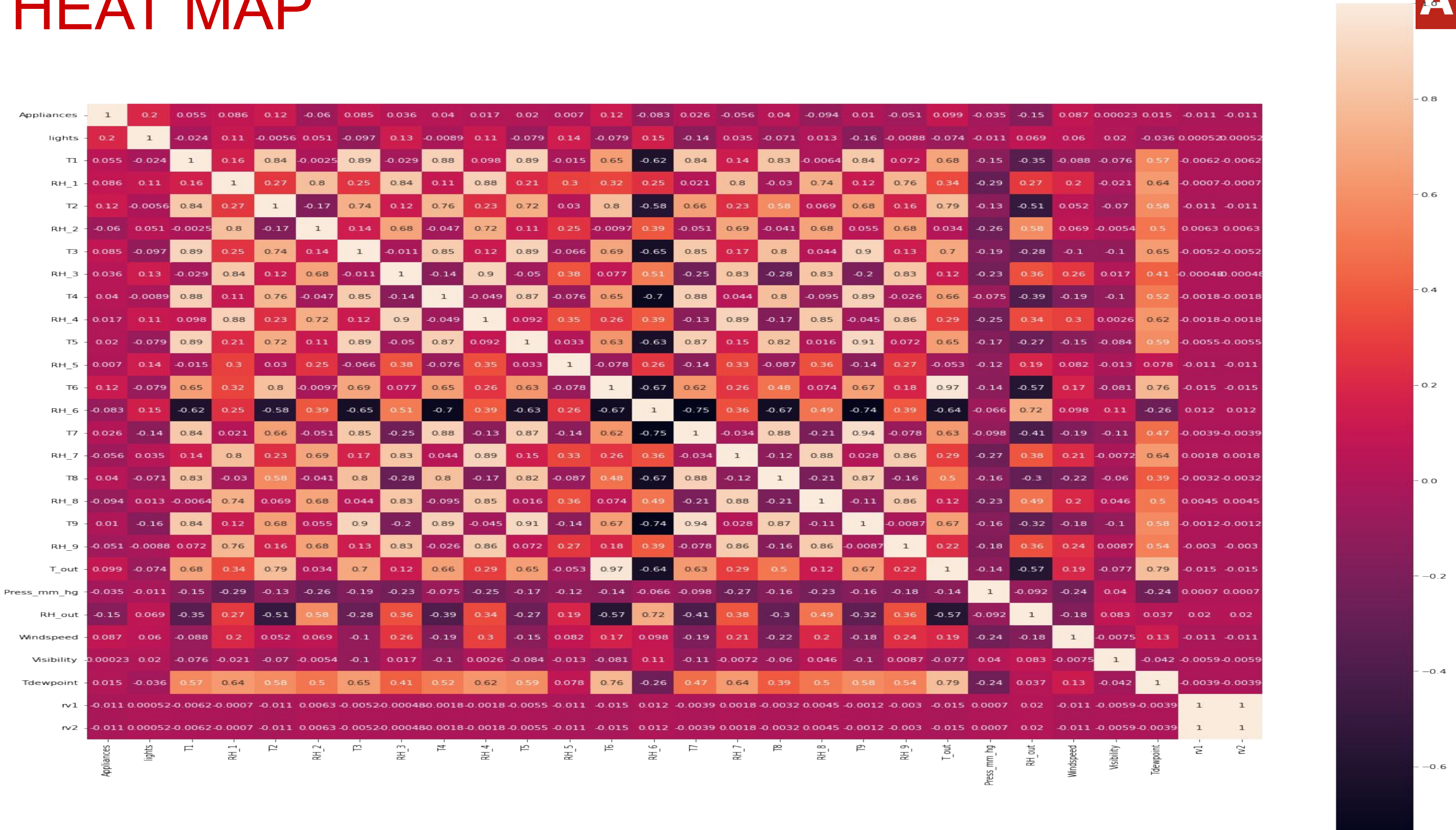
Temperature vs Humidity



Appliances vs Humidity



HEAT MAP


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
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Feature Engineering

```
# Binning of hour column into different buckets
def hourly_basis(hr):
  if 0 <= hr <= 5:
    return 1
  elif 6 <= hr <= 11:
    return 2
  elif 12<= hr <=17:
    return 3
  else:
    return 4
```

```
# creating instance of one-hot-encoder
enc = OneHotEncoder(handle_unknown='ignore')
# passing bridge-types-cat column (label encoded values of bridge_types)
enc_df = pd.DataFrame(enc.fit_transform(df1[['hourly','month','weekday']]).toarray())
# merge with main df1 on key values
df1 = df1.join(enc_df)
df1.head()
```

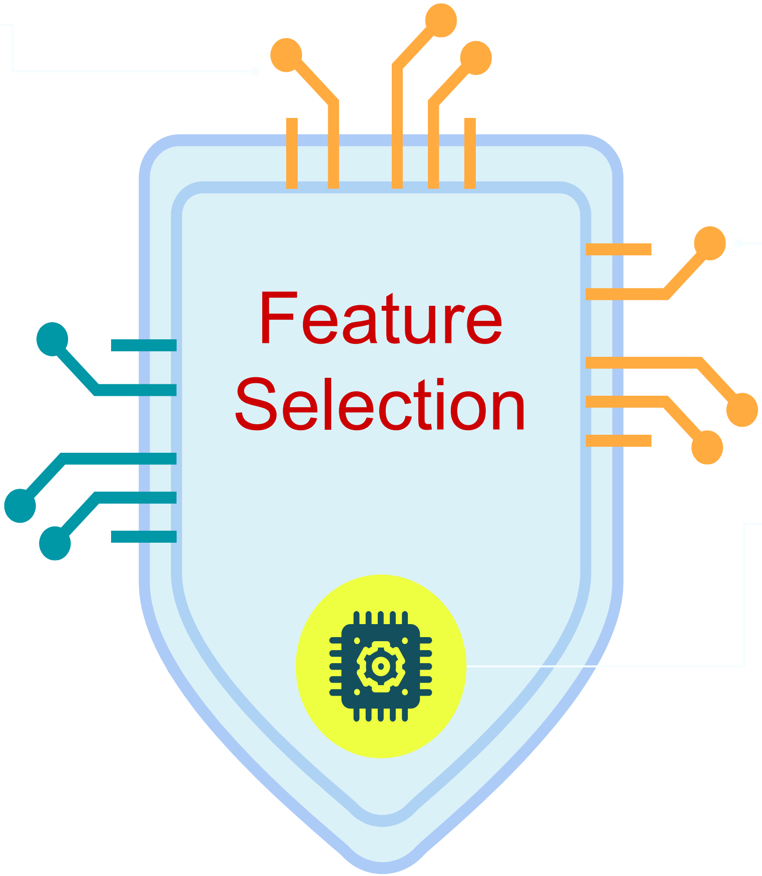
Binning	Hour
One Hot Encoding	Hourly Month Weekday
Skewness	RH_5, RH_6, RH_out, Appliances
Log Transform	RH_5,RH_6,RH_out Appliances

01 Filter Method

Information Gain
Chi-Square test
Fischer’s score

02 Wrapper Method

Forward selection
Backward Elimination
Exhaustive



03 Embedded Method

Regularization
Random Forest
Importance

04 BurotaPy

Feature Selection

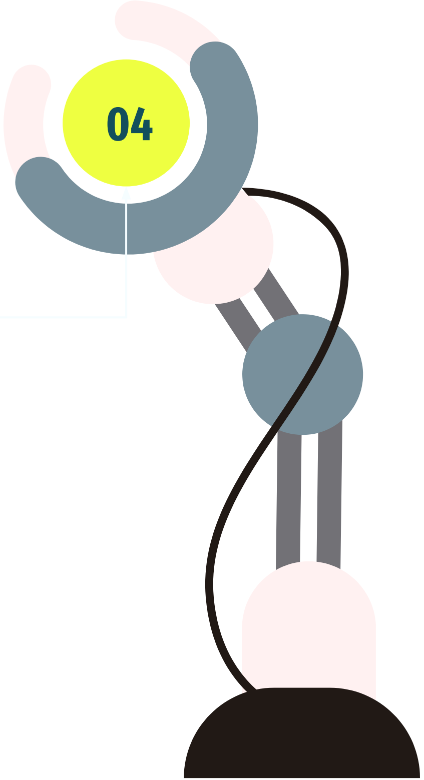
BurotaPy
rv1, rv2, Visibility, Weekday, Lights and month variable.

01 **Shadow Features**

02 **Fitting a Random Forest**

03 **Threshold**

04 **Hits**



Feature Scaling and Its need

Feature Scaling techniques

Standerdization

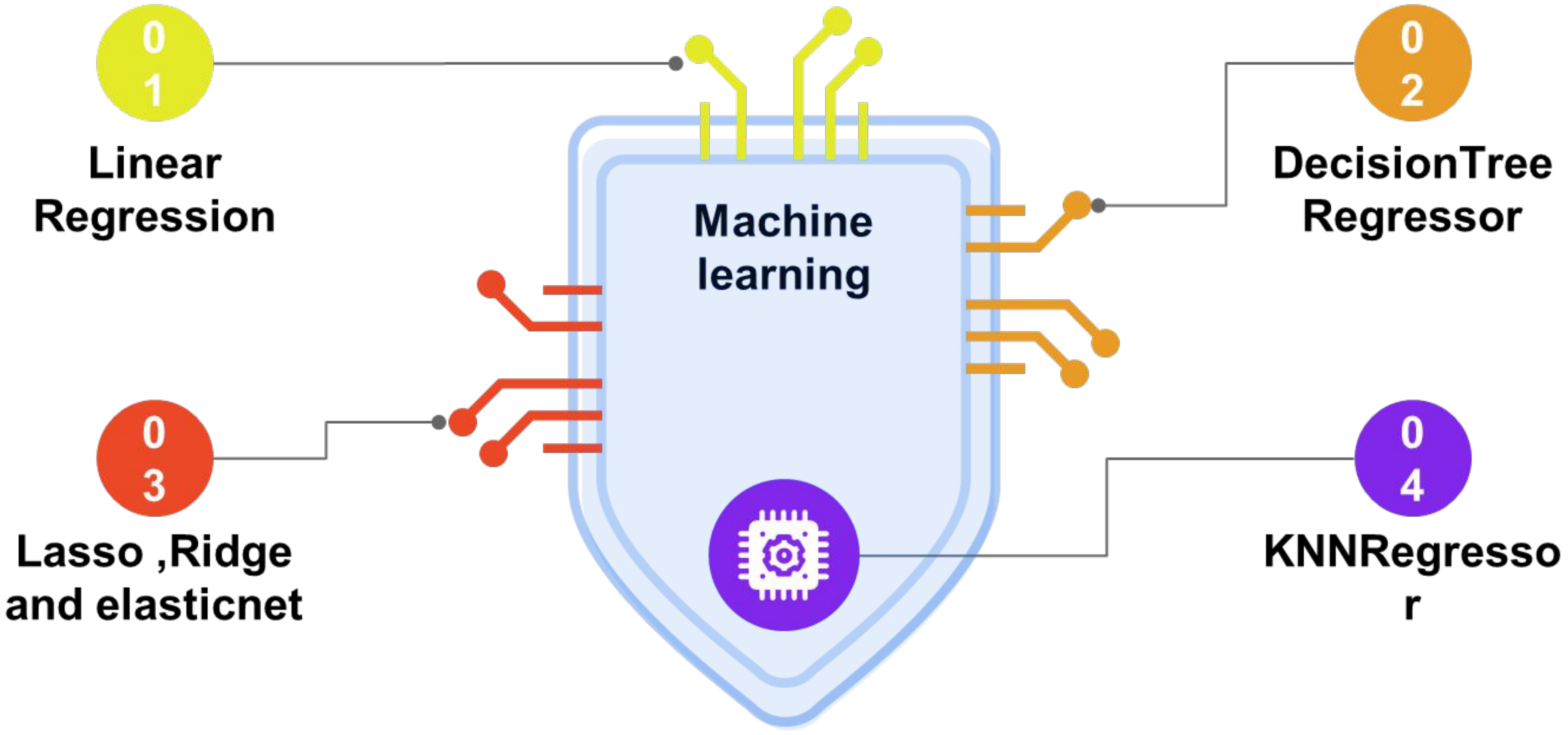
- Values are centred around mean with unit standard deviation
- Values are not restricted to a particular range
- Used when data follows gaussian distribution

Vs

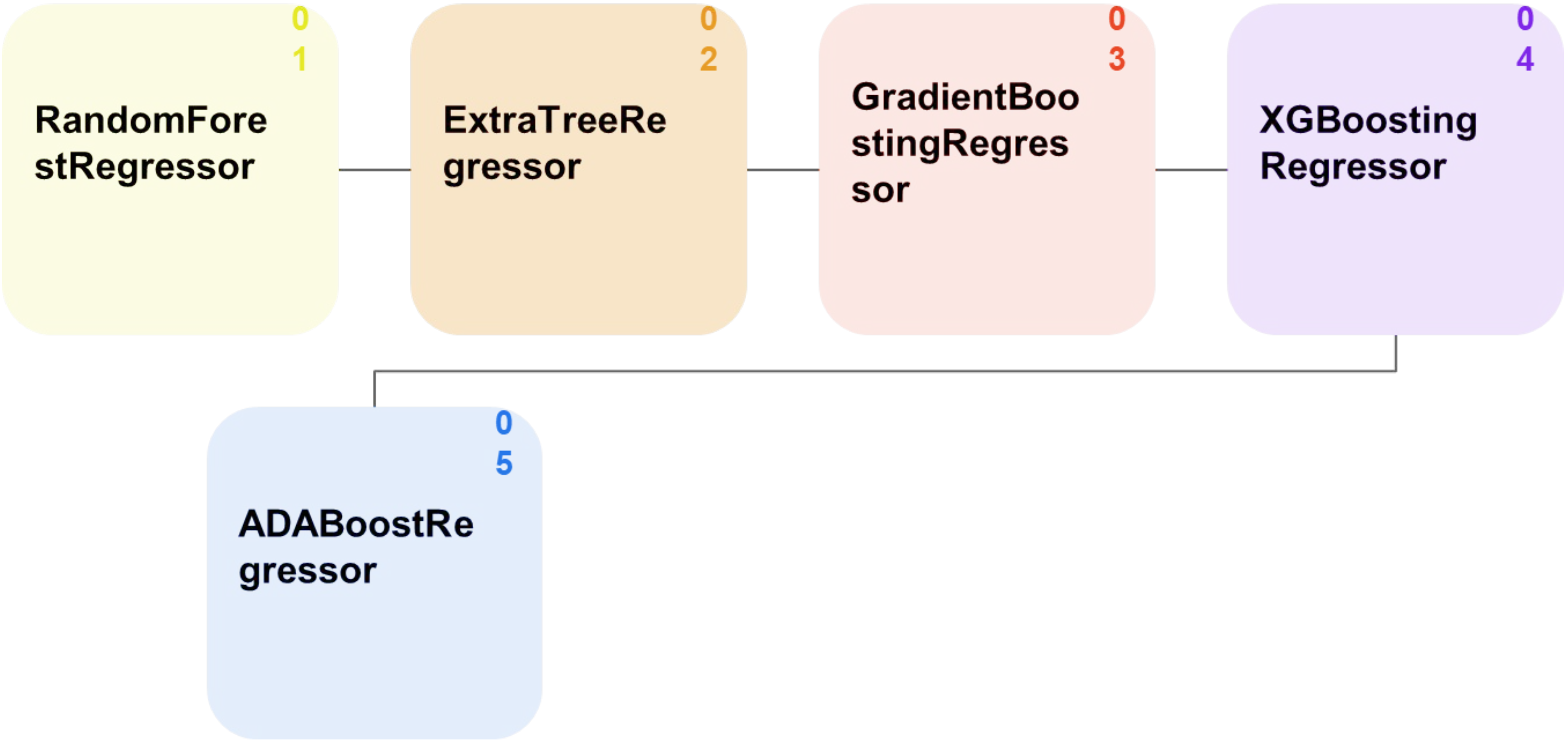
Normalization

- Rescale the values between [0,1]
- More helpful in knn or NN type of algorithms
- Is good to use when data is not following gaussian distribution

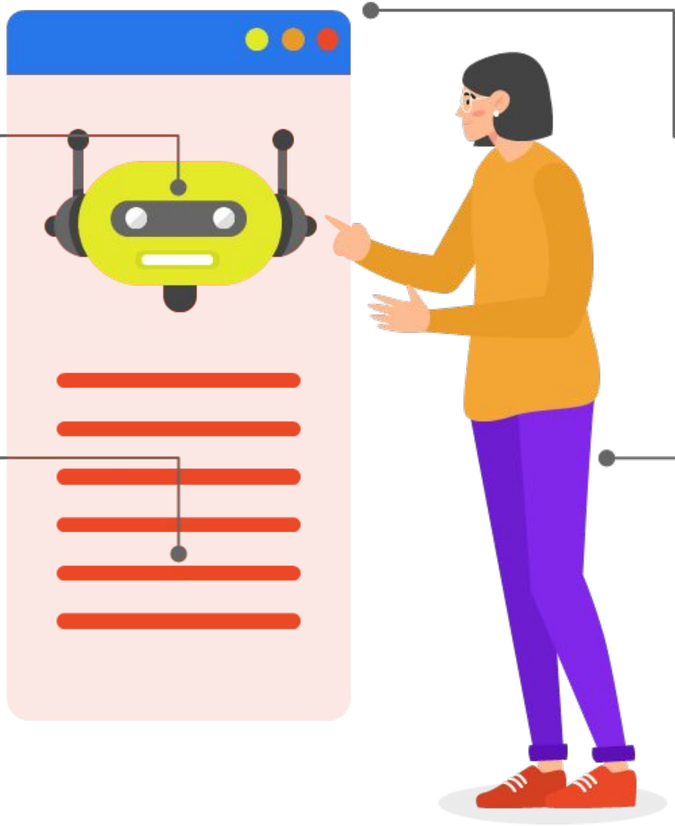
Selecting Best Perfoming Model



Selecting best perfoming model(Ensamble)



Hyperparameter Tunning



01 GridsearchCV

02 Randomsearch CV

03 Bassian Optimization

04 Manual search

EXTRATR
EEREGRE
SSOR

Hyperparameter tunning of extratreeregressor

01

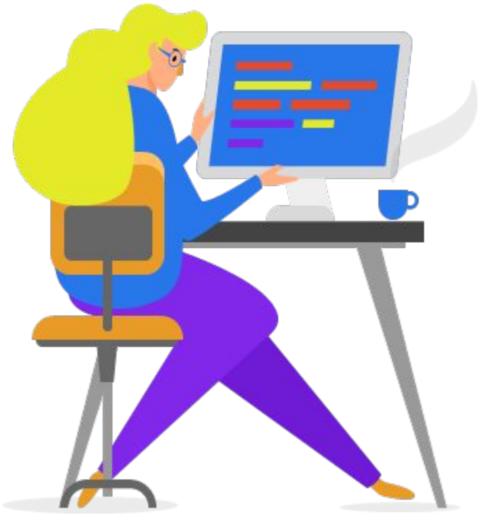
ExtraTreesRegressor(max_depth=80, max_features='sqrt', n_estimators=200, random_state=40)

Make predictions

02

y_pred = reg.predict(X_test_sc)

Model Evaluation Metrics



01 Mean Absolute Error(MAE)

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.(0.65)
we aim to get a minimum MAE because this is a loss.

02 Mean Squared Error(MSE)

It represents the squared distance between actual and predicted values.(1.3220) we perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.

03 Root Mean Squared Error(RMSE)

As RMSE is clear by the name itself, that it is a simple square root of mean squared error.(1.14)

04 Root Mean Squared Log Error(RMSLE)

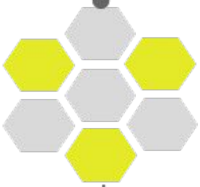
It is the Root Mean Squared Error of the log-transformed predicted and log-transformed actual values.(0.1396) the log of the RMSE metric slows down the scale of error.

05 R Squared

R2 score is a metric that tells the performance of our model.(0.75) when the R2 score is between zero and one like 0.8 which means your model is capable of explaining 80 per cent of the variance of data.

Model Interpretation

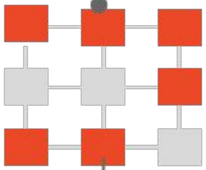
01



Using LIME Library

The output of LIME is a list of explanations, reflecting the contribution of each feature to the prediction of a data sample.

02



Using SHAP Values

The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction.

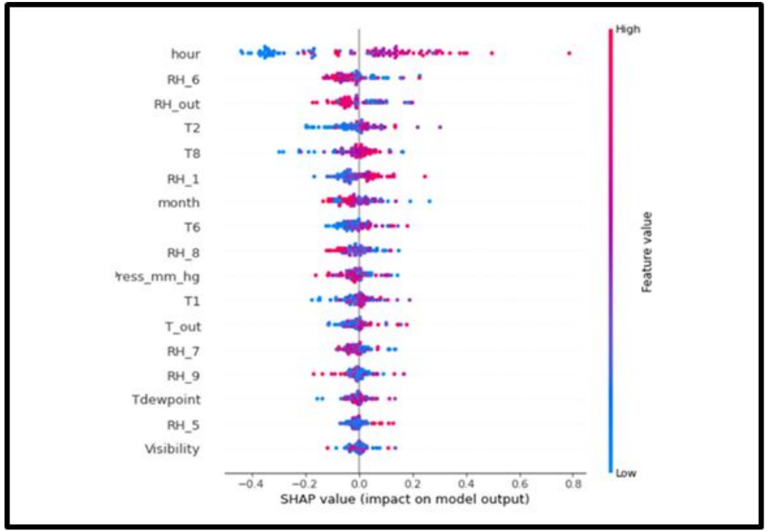


Fig :- SHAP Summary Plot

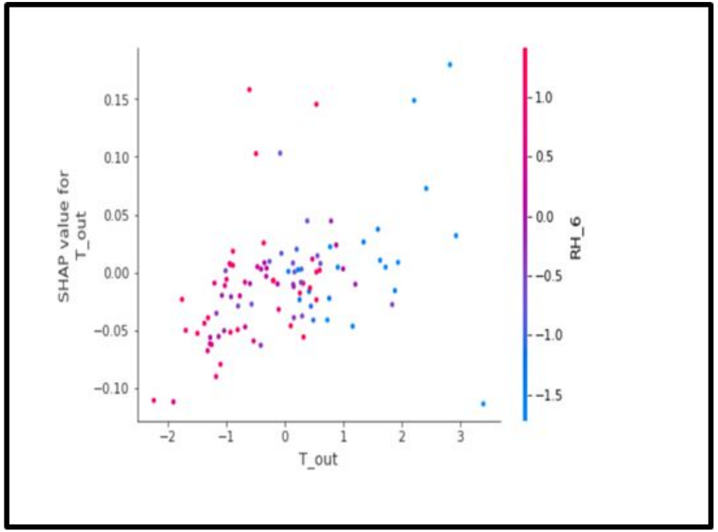


Fig :- SHAP Dependence Plot

The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value of an instance per feature.

A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonic, or more complex.

Conclusion

01 In data visualization, we used log transformation to remove the skewness. And on removal of skewness it is observed that it follows normal distribution.

02 The energy consumption for each day of the week is at its highest during hours between 4 pm to 8 pm while it is at its lowest between 12 am to 5 am . For other times of the day it doesn't follow a general trend, it differs from day to day.

03 As Temperature increases, energy consumption of the Appliances increases from Jan-Feb and March-April. And consumption decreases with increase in temperature for months Feb-March and April-May.

04 1)Filter Method 2) Wrapper Method 3) Embedded Method 4) Burota. We got the best results from the burota method. After performing feature selection using burota five features were dropped - lights, rv1,rv2,weekday, Visibility and month..

05 According to best fit model , the 5 most and least important features The top 3 important features are humidity attributes, which leads to the conclusion that humidity affects power consumption more than temperature. 'Wind Speed' is least important as the speed of wind doesn't affect power consumption inside the house. So controlling humidity inside the house may lead to energy savings.

06 The ExtraTree Regressor model came out to be the best model with an R2 score of 75.10 Extra Trees Regressor, the features and splits are selected at random.

Thank You