Capstone Project

Appliances Energy Prediction

by

Pankaj Beldar

Almabetter Trainee, Bangalore

Problem Statement

- Data-driven prediction of energy use of appliances .The data set is at **10 min** for about **4.5 months**. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with m-bus energy meters. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru) and merged together with the experimental data sets using the date and time column. Two random variables have been included in the data set for testing the regression models. and to filter out non-predictive attributes (parameters).

Introduction

- The understanding of the appliances energy use in buildings has been the subject of numerous research studies , since appliances represent a significant portion (between 20 and 30% of the electrical energy demand .
- Regression models for energy use can help to understand the relationships between different variables and to quantify their impact. Thus, prediction models of electrical energy consumption in buildings can be useful for a number of applications: to determine adequate sizing of photovoltaic and energy storage to diminish power flow into the grid , to detect abnormal energy use patterns , to be part of an energy management system for load control , to model predictive control applications where the loads are needed , for demand side management (DSM) and demand side response (DSR) and as an input for building performance simulation analysis

Data Attributes

- **date:** time year-month-day hour:minute:second
- **Appliances:** energy use in Wh (Dependent variable)
- **lights:** energy use of light fixtures in the house in Wh (Drop this column)
- **T1**, Temperature in kitchen area, in Celsius
- **RH1**: Humidity in kitchen area, in %
- **T2,** Temperature: in living room area, in Celsius
- **RH2**:Humidity in living room area, in %
- **T3:** Temperature in laundry room area
- **RH3:** Humidity in laundry room area, in %
- **T4:** Temperature in office room, in Celsius
- **RH4:**Humidity in office room, in %

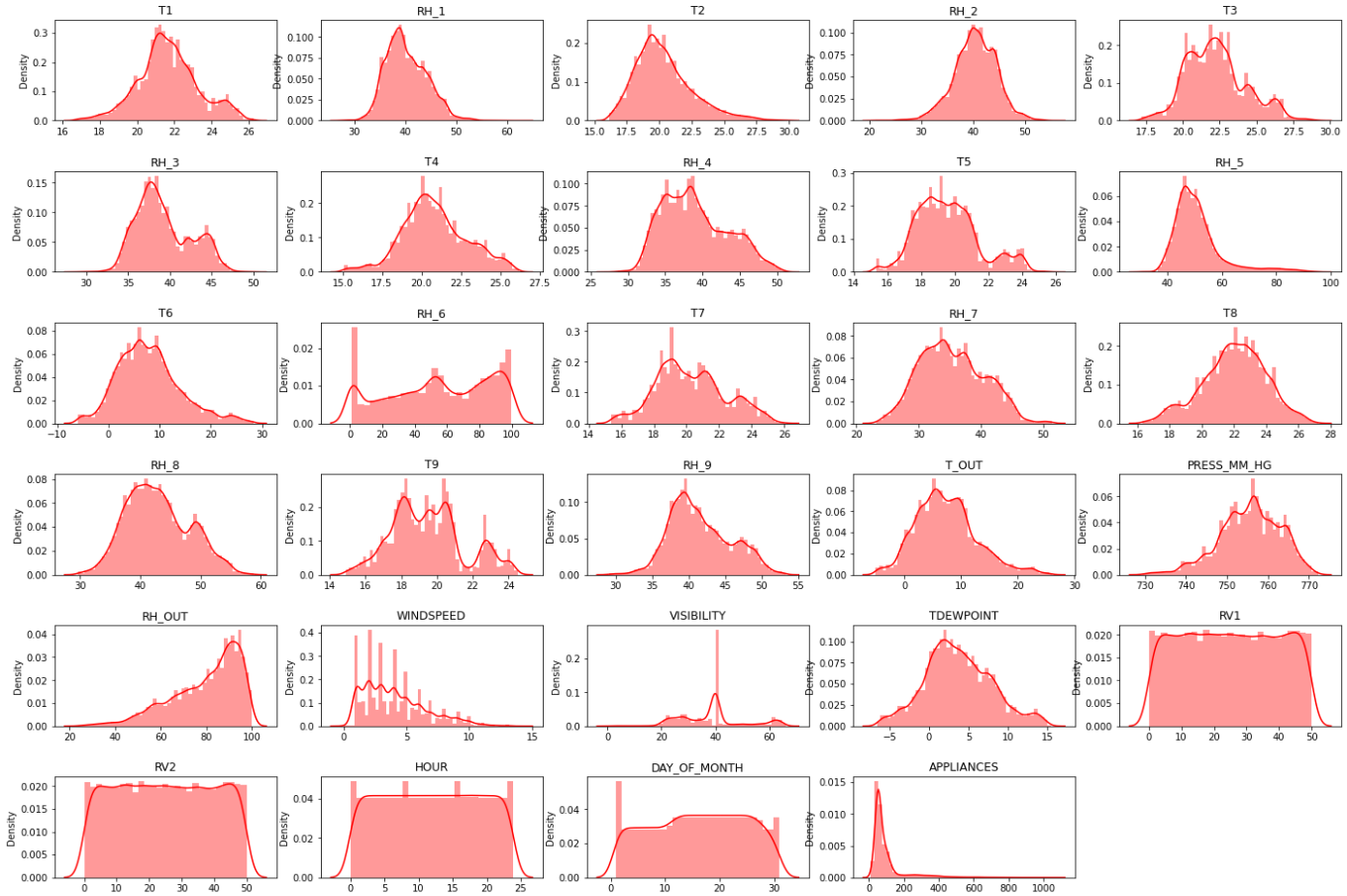
T5: Temperature in bathroom, in Celsius
RH5: Humidity in bathroom, in %
T6, Temperature outside the building (north side), in Celsius
RH6: Humidity outside the building (north side), in %
T7: Temperature in ironing room , in Celsius
RH7: Humidity in ironing room, in %
T8, Temperature in teenager room 2, in Celsius
RH8: Humidity in teenager room 2, in %
T9: Temperature in parents room, in Celsius
RH9: Humidity in parents room, in %
To, Temperature outside (from Chievres weather station)
Pressure (from Chievres weather station), in mm Hg
RHout, Humidity outside (from Chievres weather station), in %
Wind speed: (from Chievres weather station), in m/s
Visibility: (from Chievres weather station), in km
Tdewpoint: (from Chievres weather station), Â°C
rv1: Random variable 1, nondimensional
rv2: Random variable 2, nondimensional

Dataset

Data have 19735 readings and 29 Attributes.
No Missing Value
No Duplicate Value
Include Categorical Feature as Phase of the Day

Time (Hours)	Phase of the Day
6 am to 12 pm	Morning
12 pm to 6 pm	Afternoon
6 pm to 12 pm	Evening
12 pm to 6 am	Night

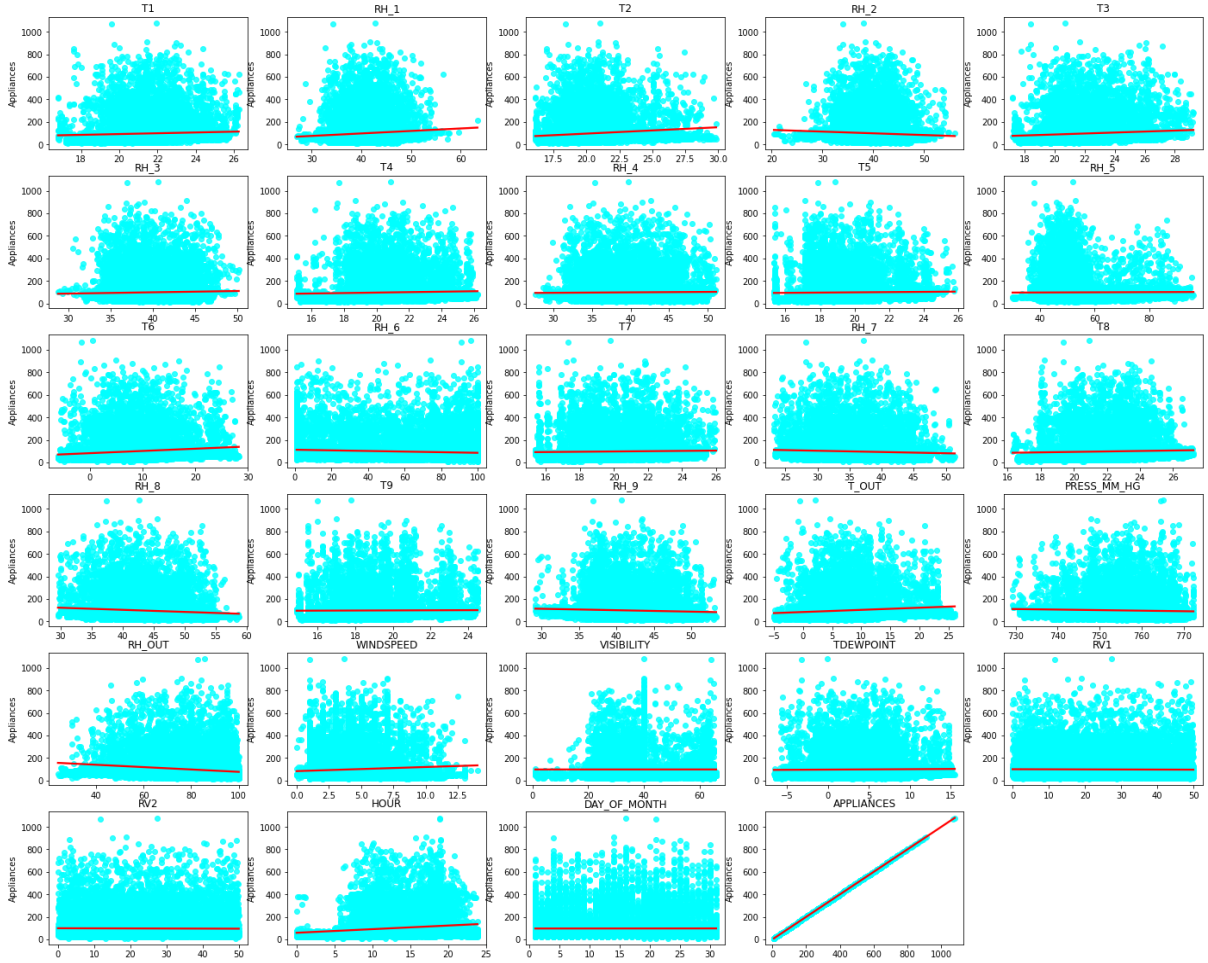
Exploratory Data Analysis-



1. Temperature and Humidity attributes have a Gaussian-like distribution. The target variable 'Appliances' has a skewed Gaussian distribution indicating a wide range of outliers over the 3rd quartile.

2. Variables hour, rv2, rv1, Appliances, rh6, rv1, T9, wind speed, visibility are not normally distributed. Other variables are seeming to be normally distributed.

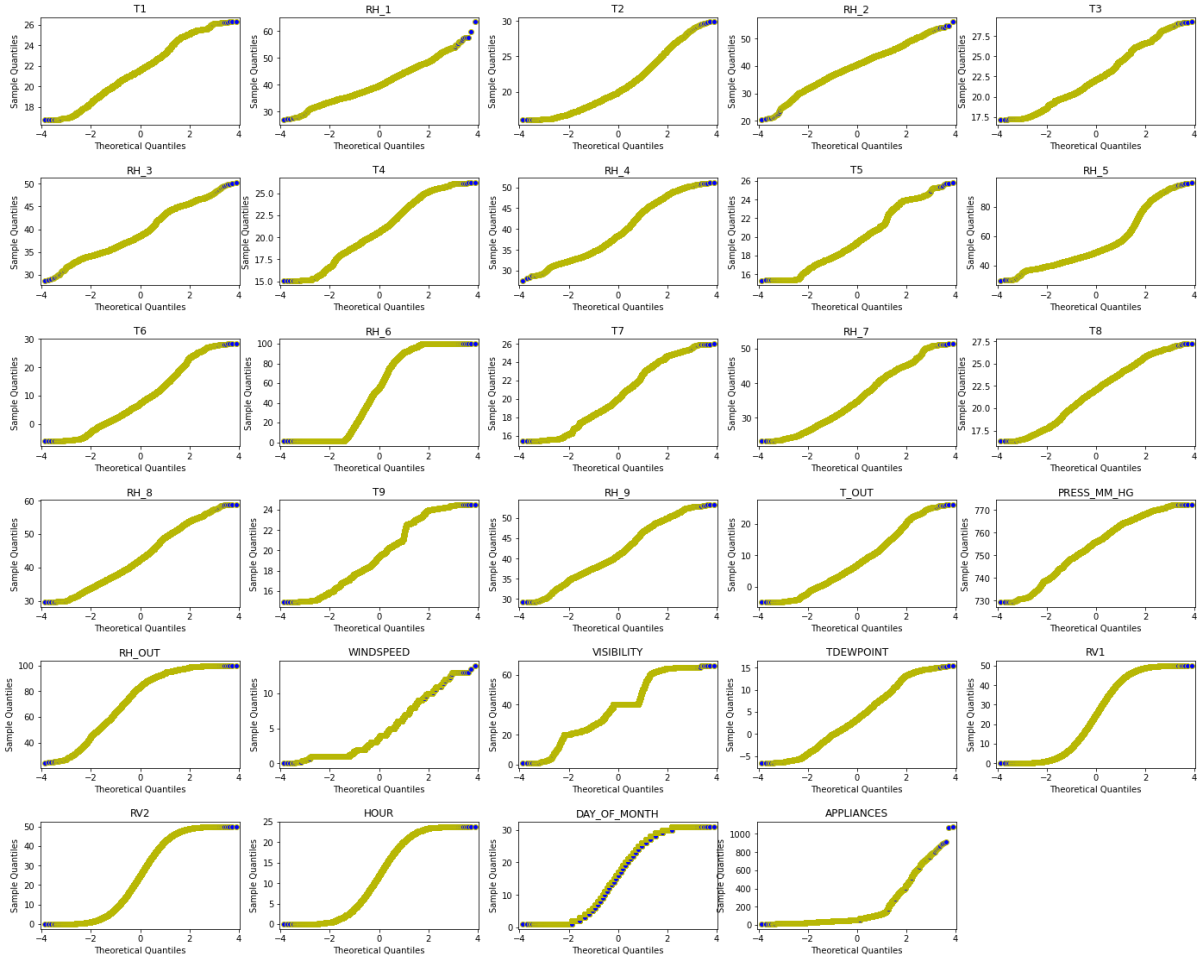
Exploratory Data Analysis-



1. **Data is highly nonlinear with dependent variable (Appliances)**

2. **Regression plot with dependent variables shows very less linear correlation with other variables.**

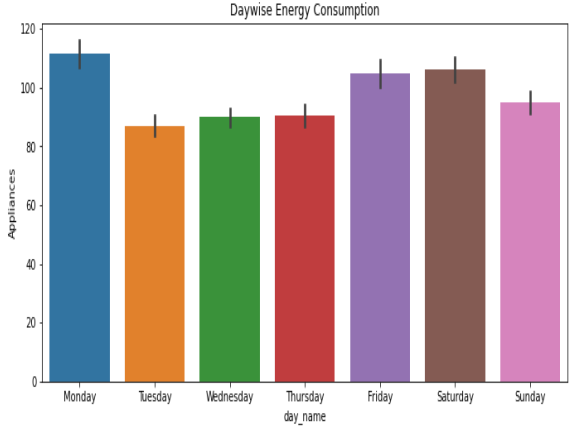
Exploratory Data Analysis-

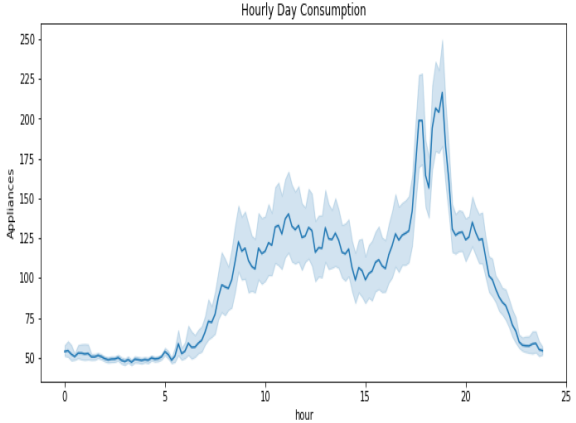


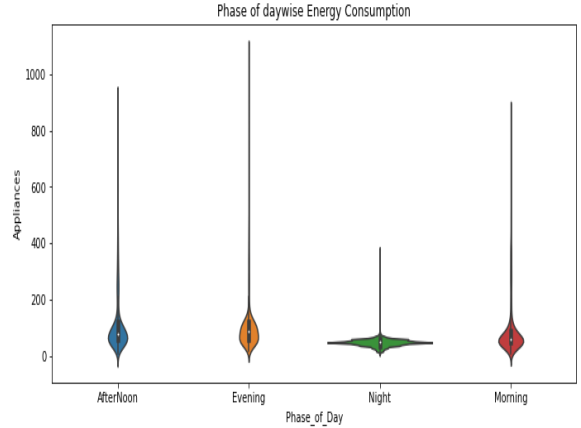
1. The Quantile-Quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

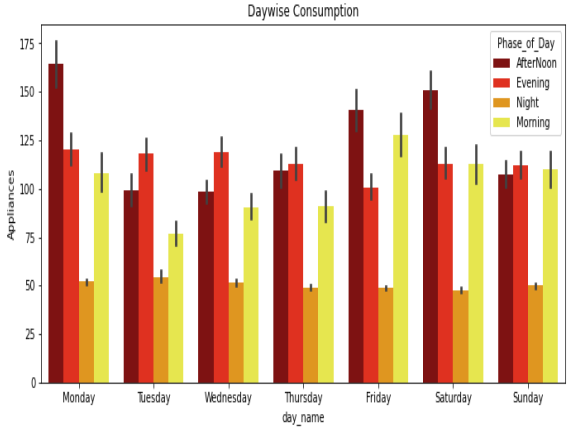
2. There is large variation in QQ plot of variables hour, rv2, day_of_month, appliances, rh6, rv1, T9, T3, windspeed, visibility.

Exploratory Data Analysis-







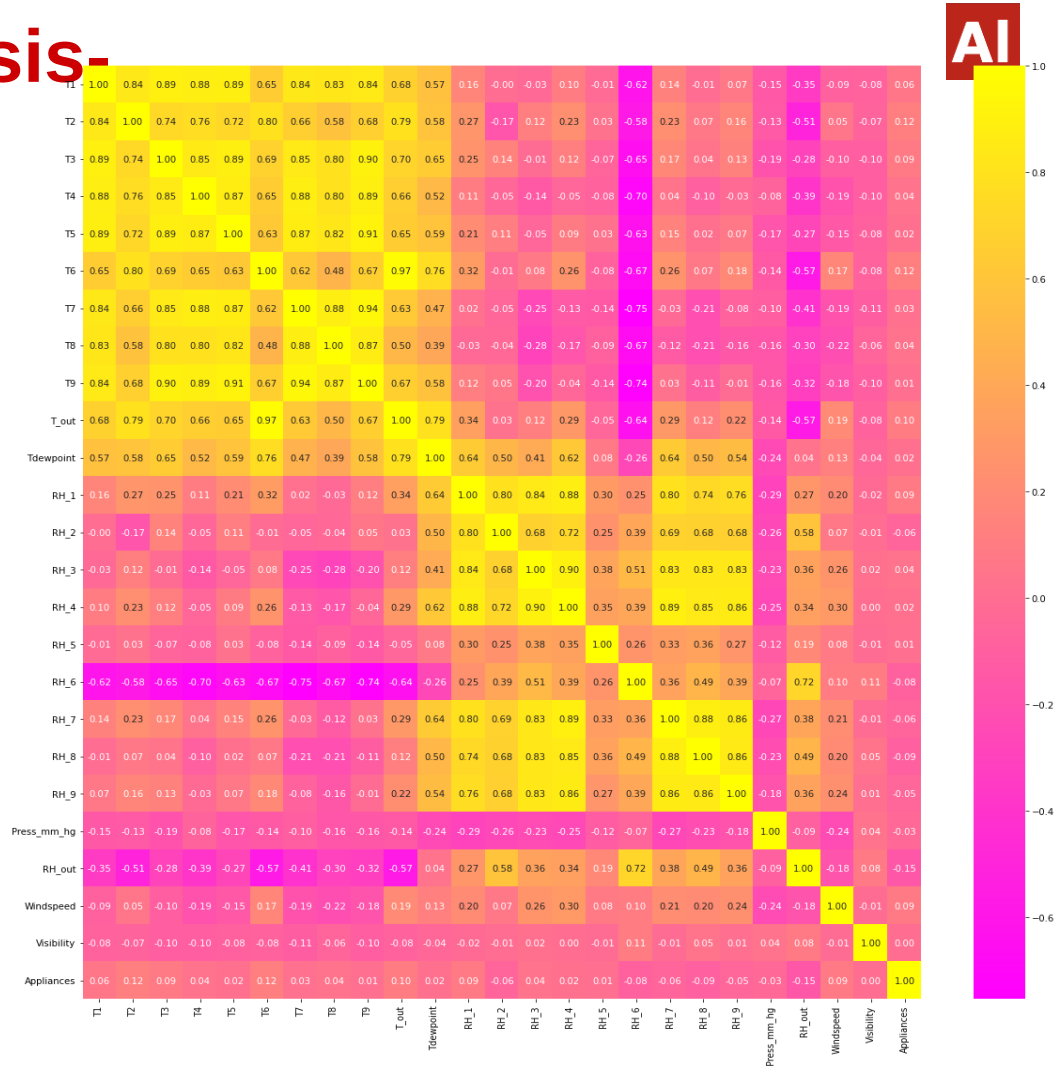


1. Energy consumption at night is less as compared to morning and afternoon on every weekday. Energy consumption is high in the evening.

2. Energy consumption is high on Monday, Friday and Saturday in the afternoon. Most of the appliances consumes around 200 Wh energy

Exploratory Data Analysis-

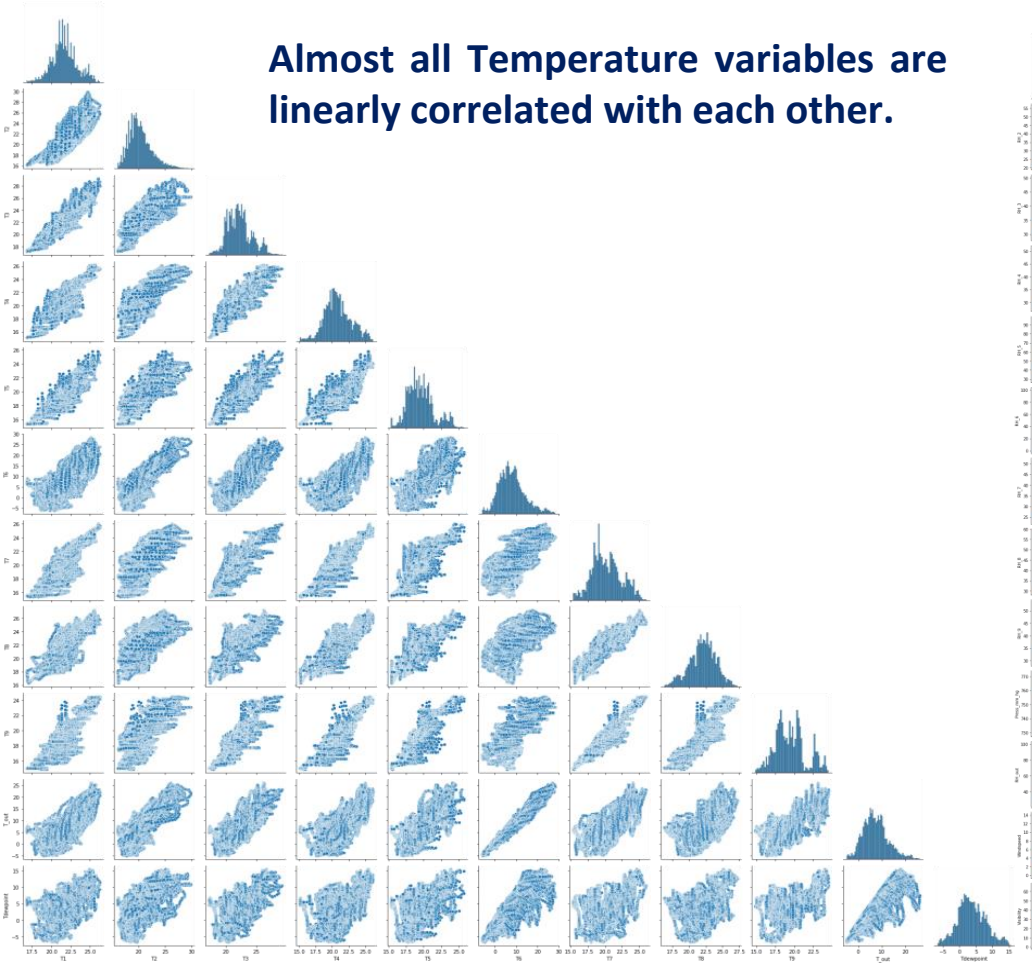
1. T9 ang T7 are highly correlated as 0.94
2. T_out and T6 are as 0.97 correlated as 0.97
3. We see strong correlation among temperature variables as change in outside heat can be experienced by all rooms except when changed only by human intervention such as use of thermostat, heaters etc.
4. There is also a strong inverse correlation observed between RH_6 and all temperature features. This is because RH_6 is the outside humidity.
5. As air temperature increases, air can hold more water molecules, and its relative humidity decreases

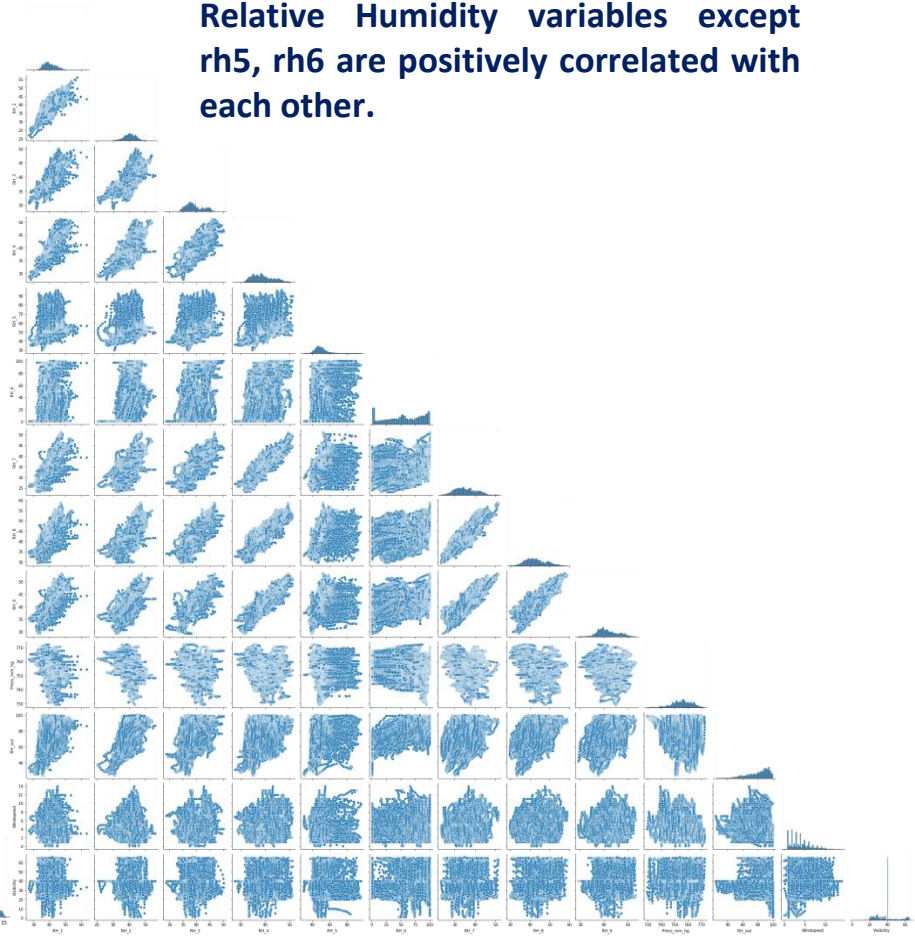


Exploratory Data Analysis-

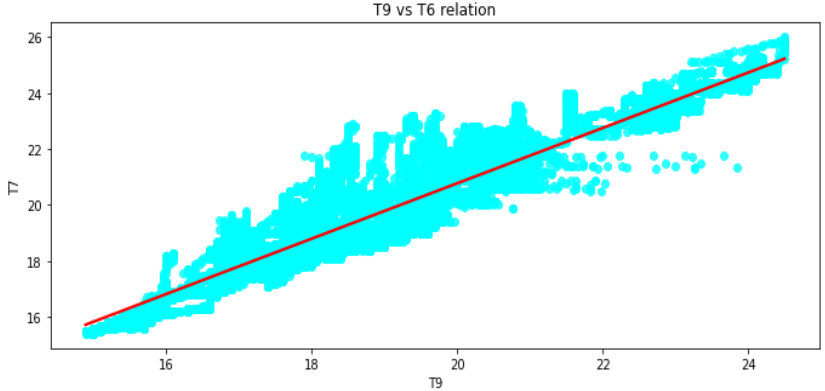
Almost all Temperature variables are linearly correlated with each other.

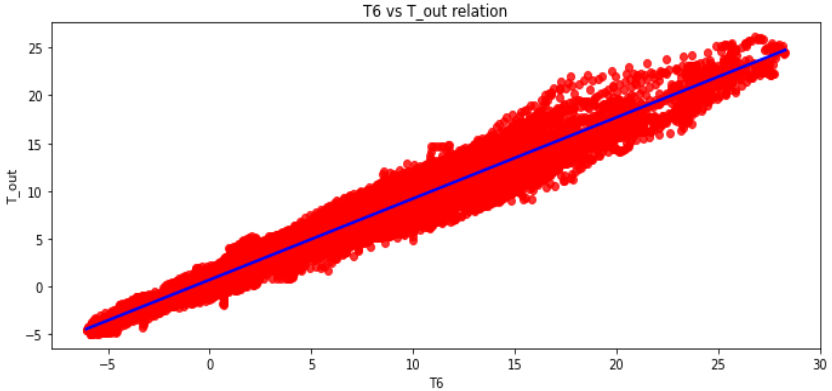
Relative Humidity variables except rh5, rh6 are positively correlated with each other.

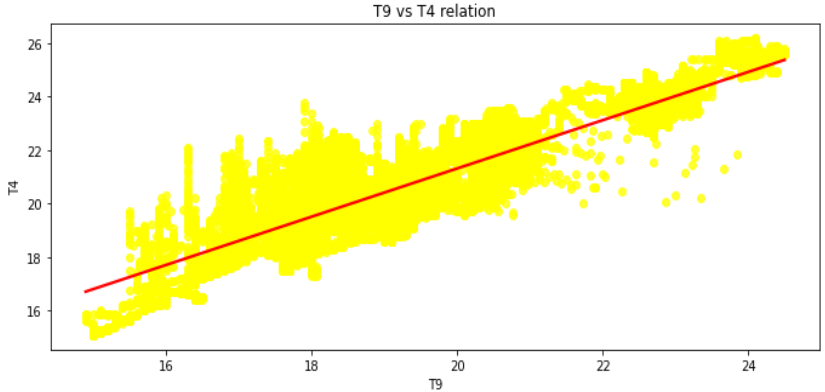


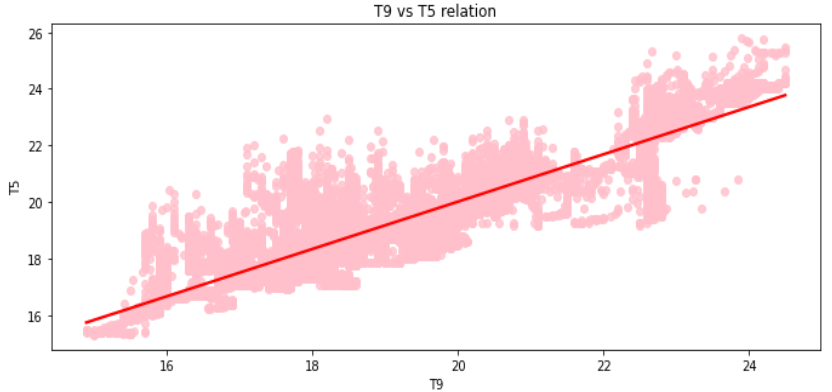


Exploratory Data Analysis-

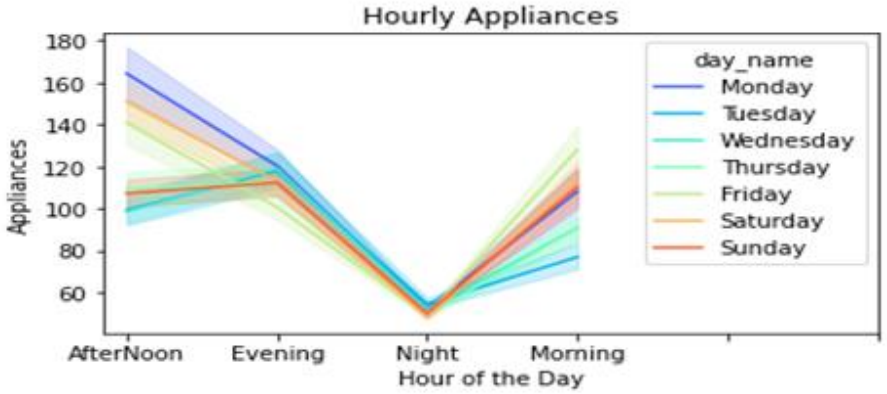


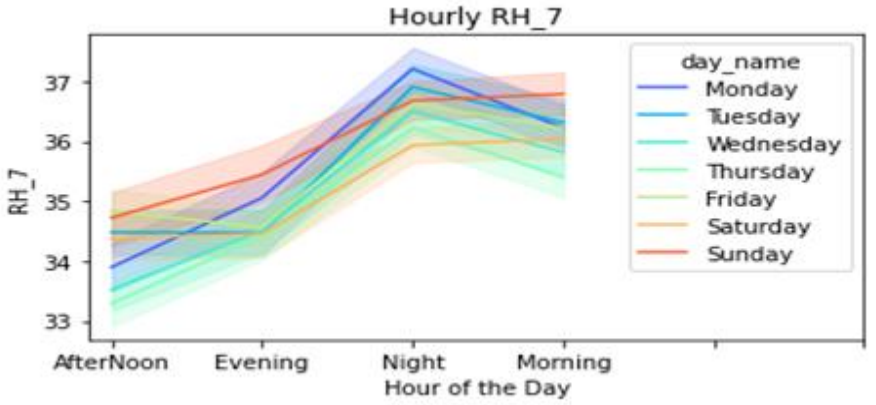


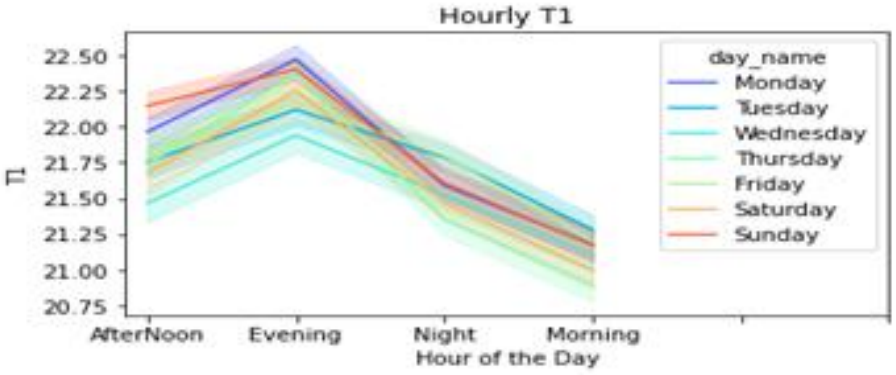


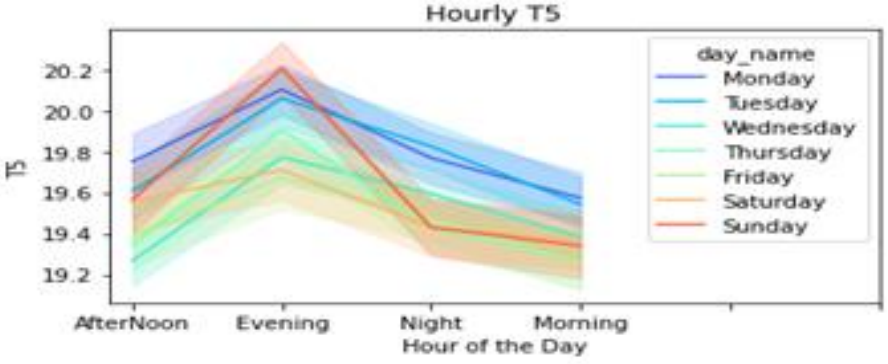


Exploratory Data Analysis-



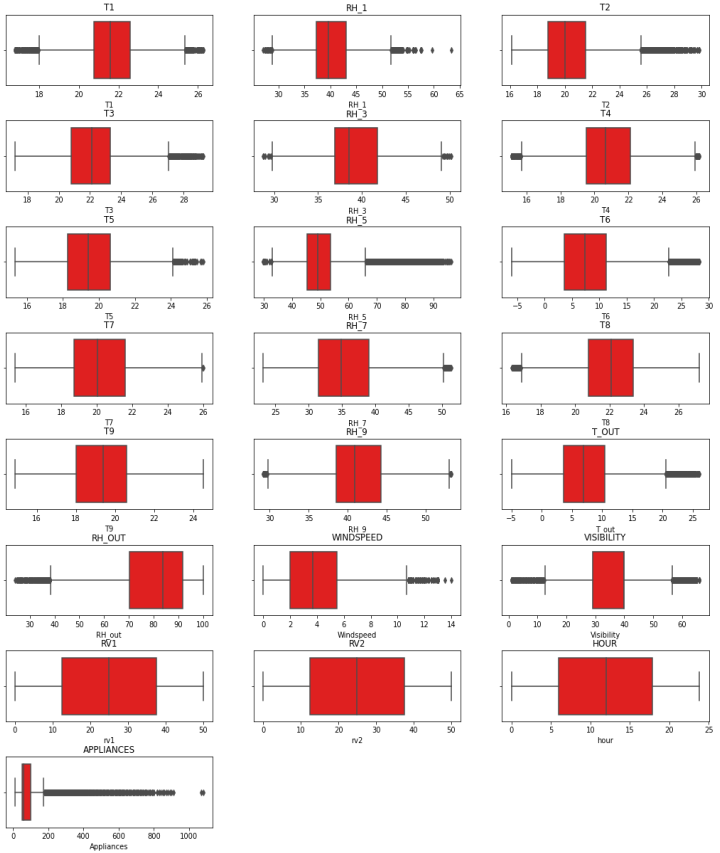




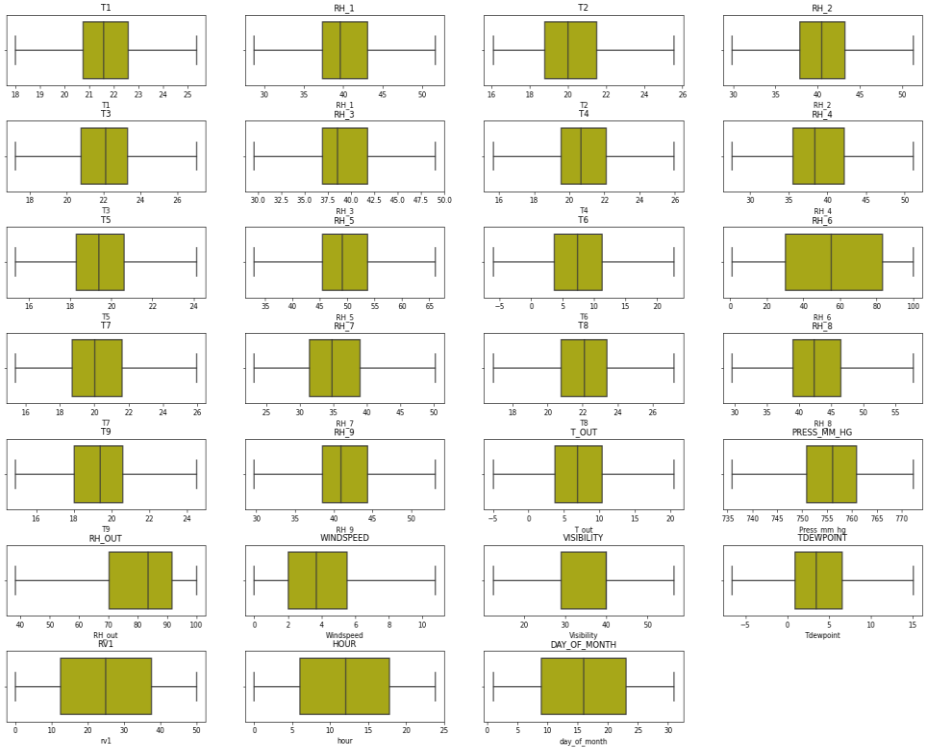


Feature Engineering and Selection

Before removing Outliers



After removing Outliers

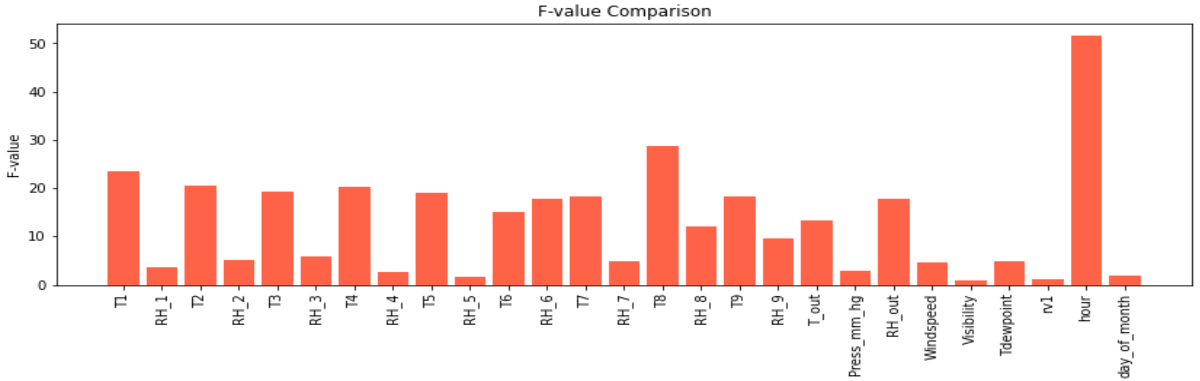


Feature Selection

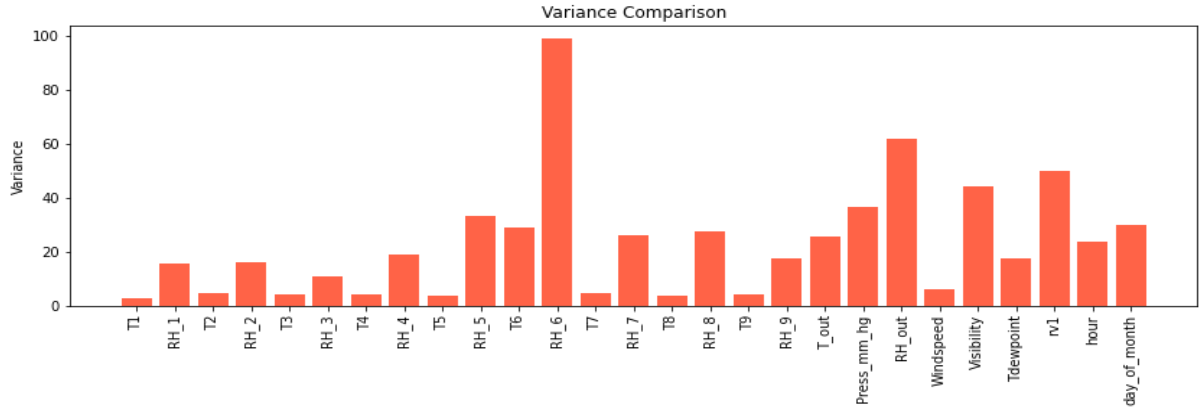
Check Multicollinearity

	variables	VIF
25	rv1	1.002546
23	Visibility	1.050790
0	Appliances	1.213415
10	RH_5	1.534392
20	Press_mm_hg	1.563047
27	day_of_month	1.711111
22	Windspeed	1.717932
18	RH_9	7.886518
7	T4	9.923554

ANOVA F Value

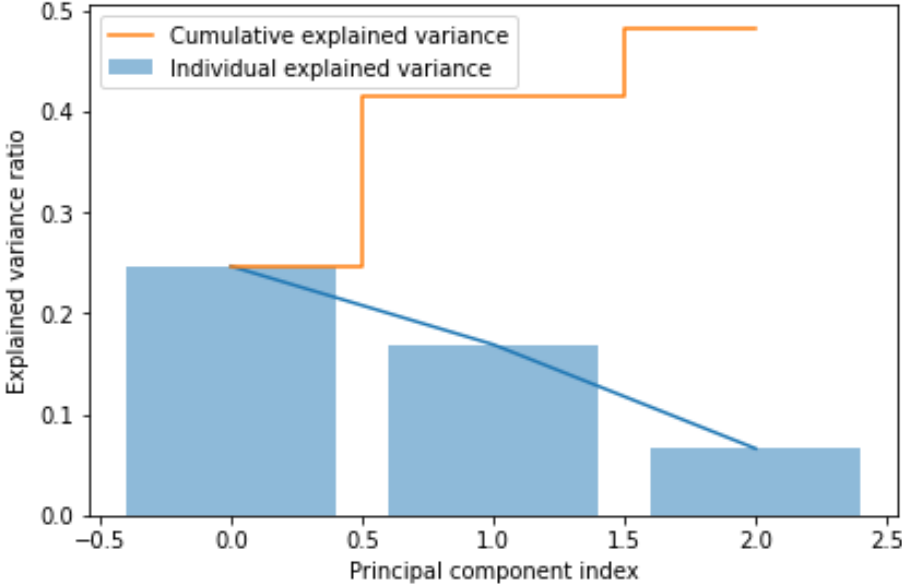


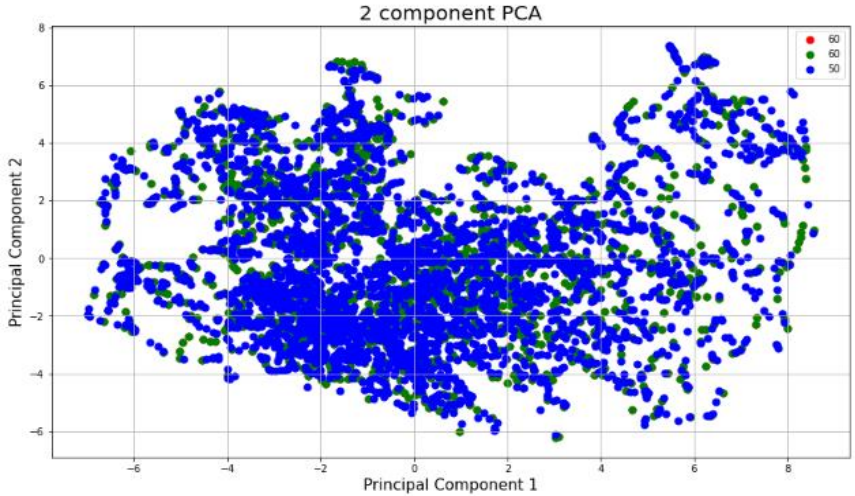
Variance Threshold

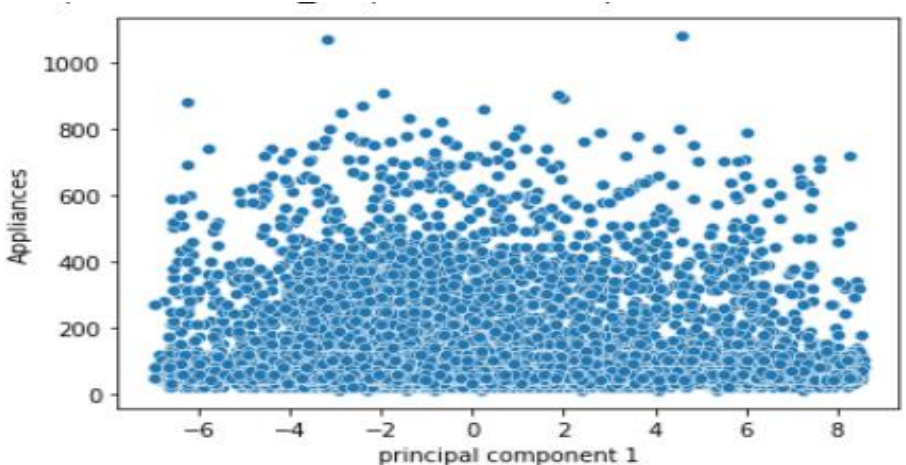


Feature Selection

Principal Component Analysis





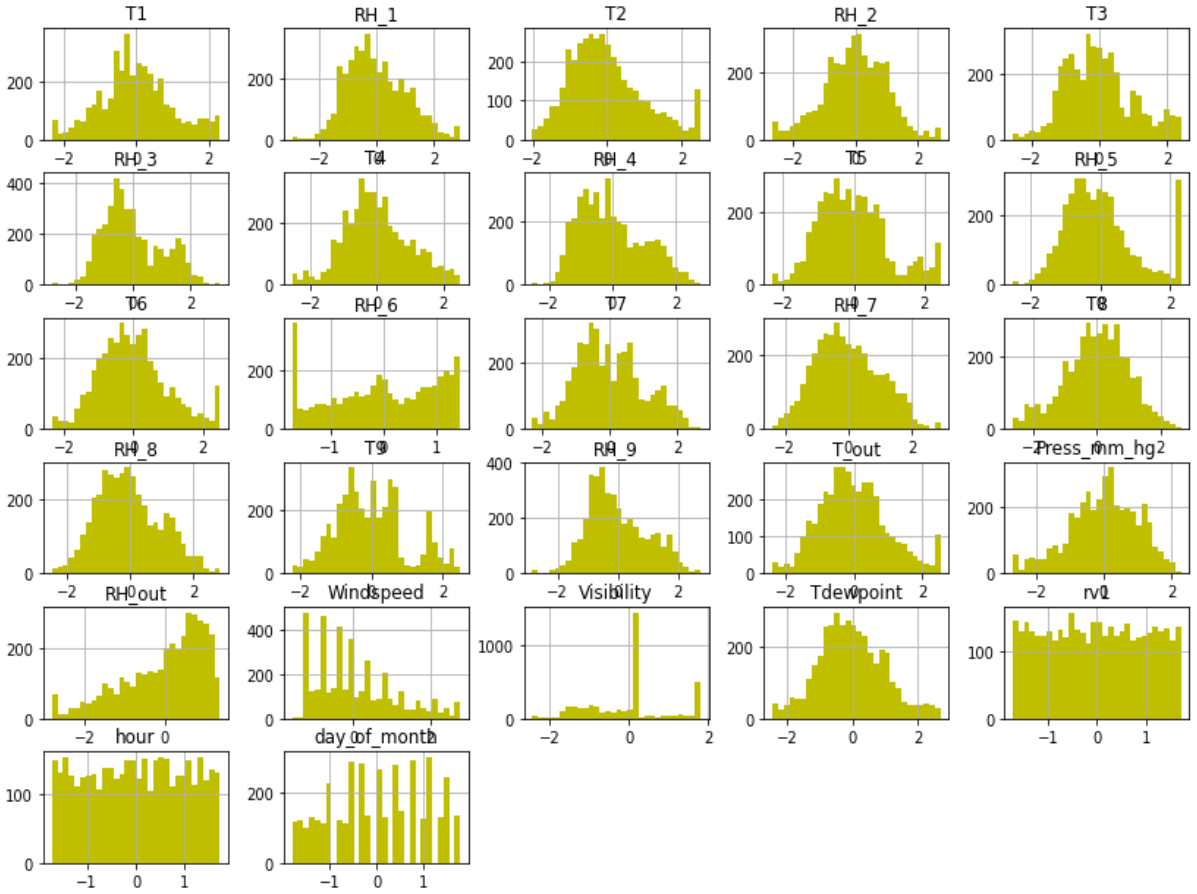


Developing Machine Learning Model

Test Train Split

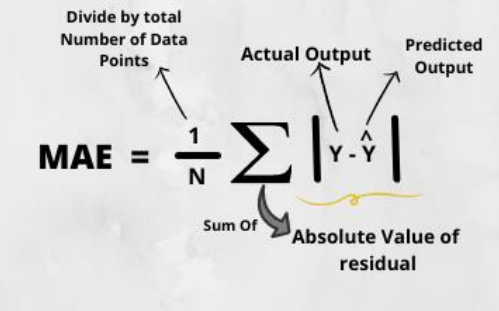
Training Data 80% (Rows, Columns)	Test data 20% (Rows, Columns)
(15788, 45)	(3947, 45)

Scaling Dataset

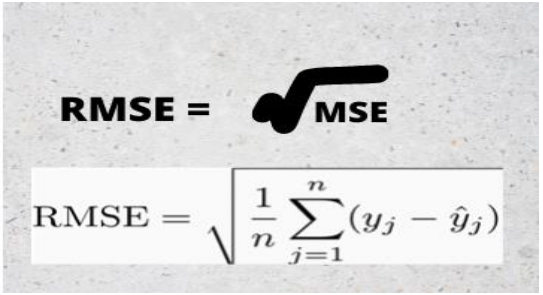


Developing Machine Learning Model

Evaluation Metrics



$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$



R2 Squared = 1 - SSr / SSm

SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2

Developing Machine Learning Model

a) **Hyper Parameter Tuning**

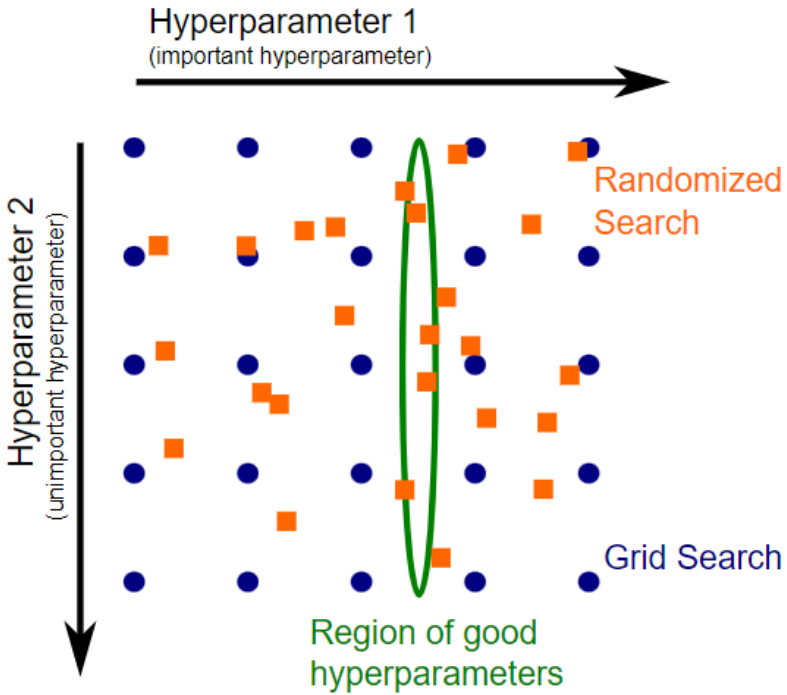
b) **Hyper parameter tuning relies more on experimental results than theory, and thus the best method to determine the optimal settings is to try many different combinations evaluate the performance of each model.**

```
({'n_estimators': 400,
  'max_features': 'sqrt',
  'max_depth': 70,
  'criterion': 'squared_error',
  'bootstrap': True},
 0.5685077929384887)
```

```
The best fit alpha value is found out to be : {'alpha': 0.1}

Using  {'alpha': 0.1}  the negative mean squared error is:  -8819.685818760041
MSE : 8295.74463248603
RMSE : 91.08097843395201
R2 : 0.17101337896202662
Adjusted R2 :  0.16166550317379735
```

Cross Validation (GridSearchCV, RandomizedSearchCV)

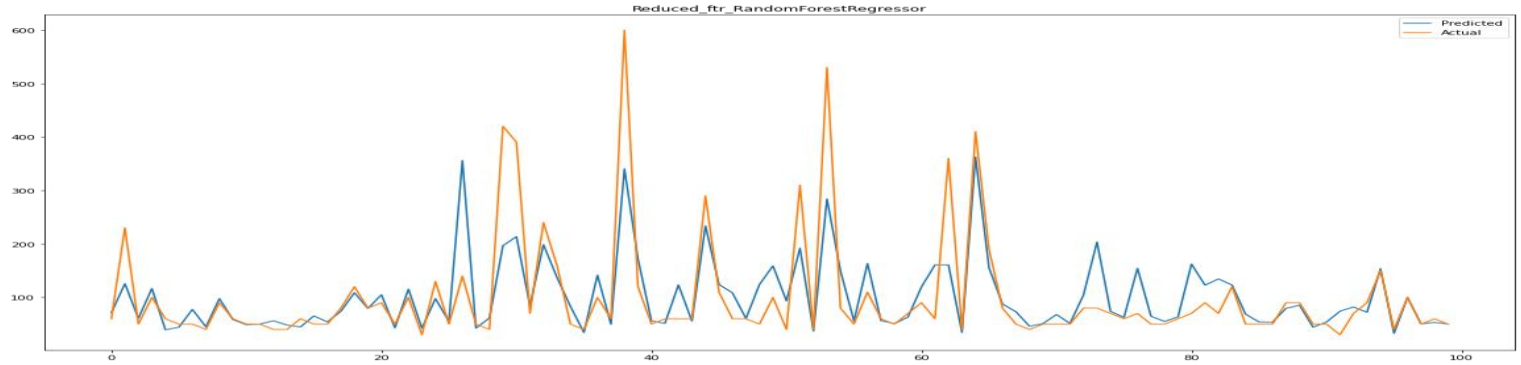



```
seed=5
param_grid = param_grid = { 'bootstrap': [True],
                           'max_depth': [70,80],
                           'criterion' :['squared_error'],
                           'max_features': [ 'log2', 'sqrt'],
                           'n_estimators': [200,400]}
random_forest = RandomForestRegressor(random_state=seed)
kfold = KFold(n_splits= 10)
```

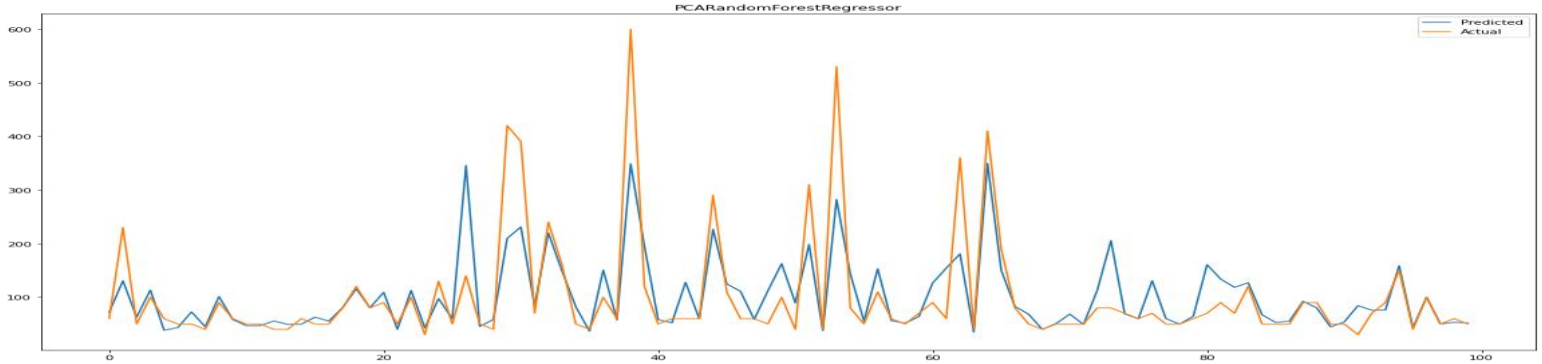
```
({'n_estimators': 400,
  'max_features': 'sqrt',
  'max_depth': 70,
  'criterion': 'squared_error',
  'bootstrap': True},
 0.5685077929384887)
```

Selecting Optimal Model

Random Forest Regressor with Feature reduction and Hyperparameter Tuning



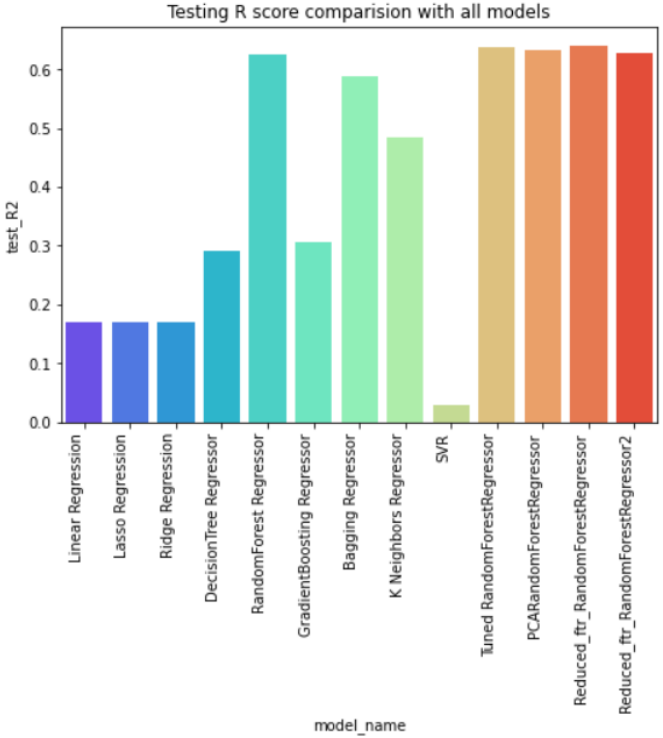
Random Forest Regressor with Principal Component Analysis

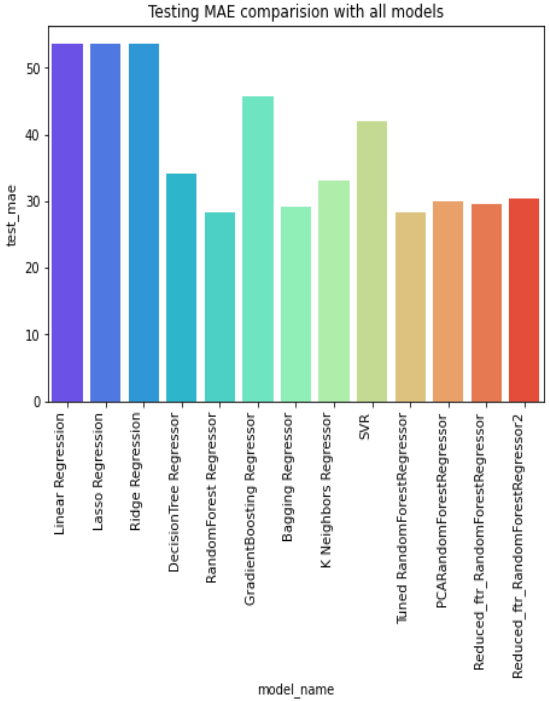


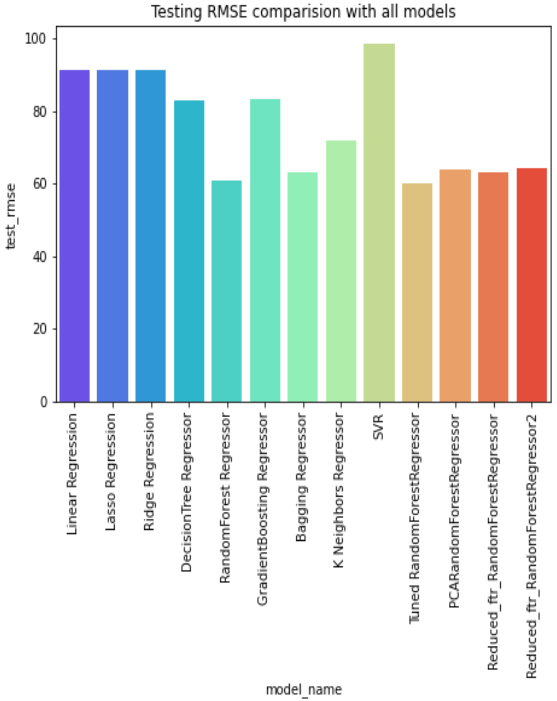
Model Performance Report

	model_type	model_name	test_mse	test_rmse	test_mae	test_R2	test_adjusted R2	training_R2
0	Linear	Linear Regression	8304.561934	91.129369	53.616417	0.170132	0.160774	0.176617
1	Linear	Lasso Regression	8304.550793	91.129308	53.616228	0.170133	0.160776	0.176617
2	Linear	Ridge Regression	8303.598268	91.124082	53.595284	0.170229	0.160872	0.176597
3	Tree	DecisionTree Regressor	6855.434507	82.797551	34.162655	0.314942	0.307217	1.000000
4	Ensemble Method	RandomForest Regressor	3677.424758	60.641774	28.294908	0.632518	0.628374	0.945464
5	Ensemble Method	GradientBoosting Regressor	6948.264897	83.356253	45.701002	0.305666	0.297836	0.365512
6	Ensemble Method	Bagging Regressor	3987.352166	63.145484	29.224474	0.601547	0.597054	0.925809
7	Neighbours	K Neighbors Regressor	5162.465670	71.850300	32.989612	0.484119	0.478302	0.678579
8	SVM	SVR	9722.610917	98.603301	42.019490	0.028428	0.017472	0.019664
9	Ensemble Method	Tuned RandomForestRegressor	3625.693802	60.213734	28.249189	0.637688	0.633508	0.946755
10	Ensemble Method	PCARandomForestRegressor	4093.815912	63.982935	29.901596	0.631854	0.627607	0.945857
11	Ensemble Method	Reduced_ftr_RandomForestRegressor	3994.559141	63.202525	29.487129	0.640780	0.638212	0.946692
12	Ensemble Method	Reduced_ftr_RandomForestRegressor2	4140.840791	64.349365	30.339638	0.627625	0.626299	0.945306

Model Performance Report







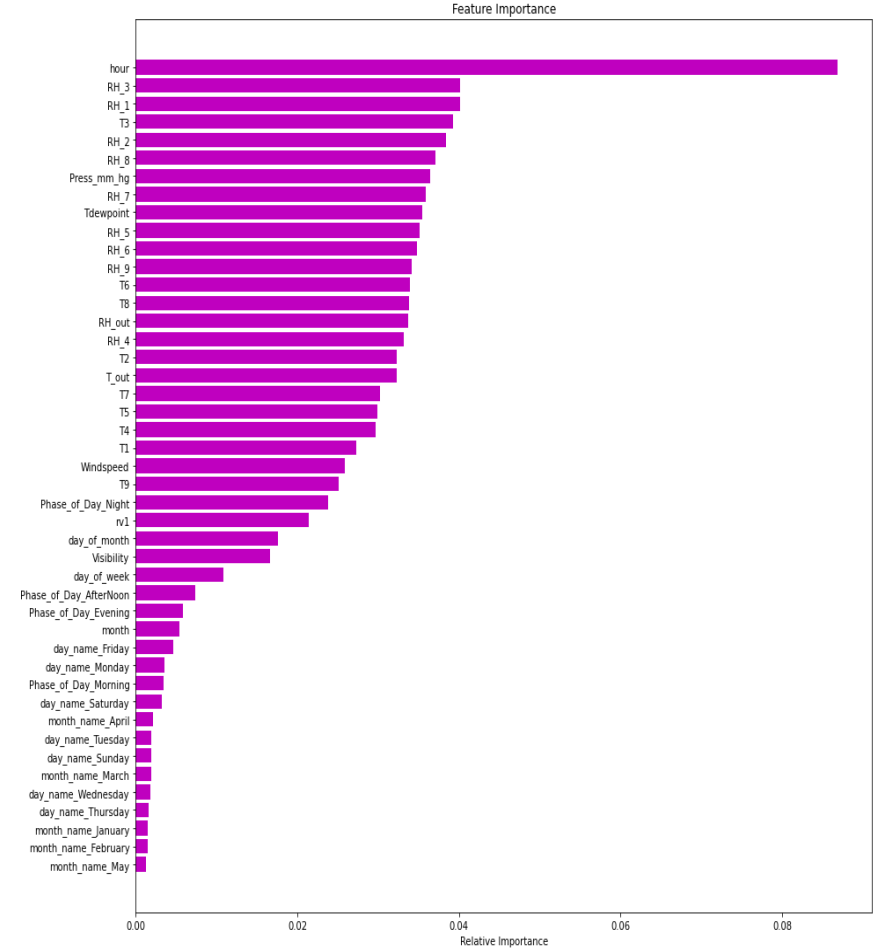
Model Explainability

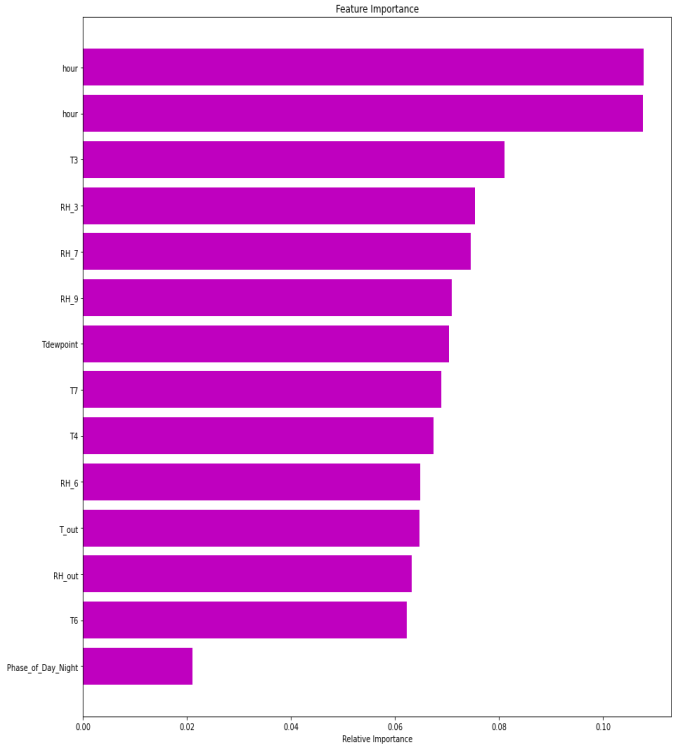
- Model Explainability refers to the concept of being able to understand the machine learning model Importance: Feature Importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Feature importance scores provide insight into the data and the deployed model.

Feature Importance

By looking at the Feature Importance graphs and the contribution chart from ELI5, we can gather that the appliance energy consumption largely depends on the ‘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' ,'Tdewpoint'

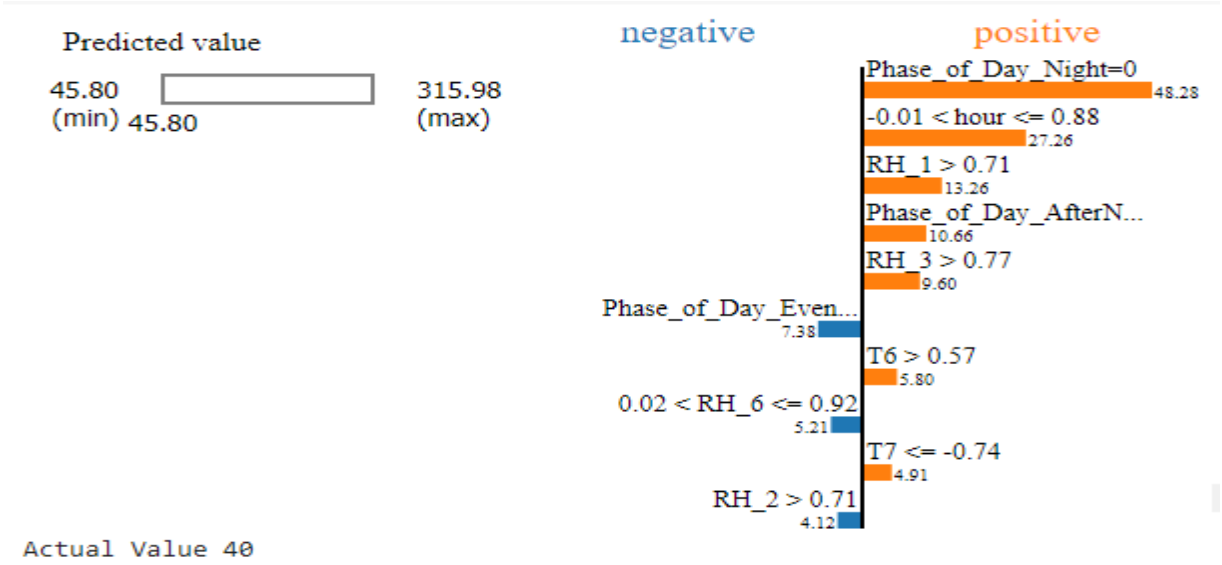
Feature Importance





LIME (Local Interpretable Model-Agnostic Explanations)

- LIME, the acronym for Local Interpretable Model-Agnostic Explanations, is a technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction.



Limitations

1. One of the main limitations of this study is that the analysis was done for only one house.

2. Important information could be found when analyzing several houses, and other relationships can be studied with appliances' energy consumption in combination with: occupant's age, number of occupants, ownership of pets, building's geometry etc.

3. Another research limitation is the length of continuous analyzed data. Different energy use patterns can potentially be found depending on the season of the year.

4. Regarding the weather station, the predictions of appliances energy use could probably be better if the weather station was closer to the house.

5. This research has not looked into the problem of optimal location of the wireless indoor sensors for improvement of the energy prediction. It is also possible that more sensors and better sensor accuracy could help to improve the energy prediction.

Conclusion

1. The household appliance energy consumption prediction models based on Linear Regression, Lasso Regression, Ridge Regression, DecisionTree Regressor Random Forest Regressor, Adaptive Boosting Regressor, Gradient Boosting Regressor, Bagging Regressor, K Neighbors Regressor and Linear SVM are explored.
2. Upon appropriate preprocessing and fittingthe ten models, we compare and evaluate the best model with lowest error and the highest R-squared score.
3. When evaluating the influence of RandomVariable attribute the linear models have assigned near zero weights to the random variable, negating its influence in prediction of the target variable.
4. Random Forest Regressor was found to be the best performing model with an R-squared score of 0.64.
5. After optimizing the hyperparameters of theRandom Forest Regressor, doing principle Component Analysis, its R-squared score increased from 0.62 to 0.64.

Conclusion

- 6. We find that this model's predictions are mainly contributed by the hour, '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' ,'Tdewpoint'. Temperature and Relative Humidity of kitchen, living room, laundry room, Ironing room, outside surrounding are playing important role in Energy Prediction.

- 7. Data from a wireless sensor network that measures humidity and temperature has been proven to increase the prediction accuracy. The data analysis showed that data from the kitchen, laundry room, living room and bathrooms had the most important contributions. Data from the other rooms also helps in the prediction. When looking at the appliances in each room, it can be seen that the laundry, kitchen and living rooms would be expected to have the highest contributions because of the equipment present. The prediction of appliances' consumption with data from the wireless network indicates that it can help to locate where in building the main appliances' energy consumption contributions are found.

Conclusion

- 8. When using all the predictors the light consumption was ranked highly. However, when studying different predictor subsets, removing the light consumption appeared not to have a significant impact. This may be an indication that other features are correlated well with the light energy consumption.
- The possible explanation for why the pressure has a strong prediction power may be related to its influence on the wind speed and higher rainfall probability which could potentially increase the occupancy of the house.

- 9. As this dataset has a time component to it, we believe that better performances can be achieved by using Time Series Analysis concepts.

Future Scope

- This study has found curious relationships between variables. Future work could include considering weather data such as solar radiation and precipitation. Also, occupancy and occupant's activity information could be useful to improve the prediction and find its relationship with other parameters (exterior weather for example). The wireless sensors could also measure CO2 and noise to help in the prediction and to track the occupant's movement from room to room and time spent in each room.

References:

1. Luis M. Candanedo, Véronique Feldheim, Dominique Deramaix, ‘Data driven prediction models of energy use of appliances in a low-energy house’, Energy and Buildings, Thermal Engineering and Combustion Laboratory, University of Mons, Rue de l’ Epargne 56, 7000 Mons, Belgium.
2. https://www.geeksforgeeks.org/python-programming-language/
3. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
4. https://scikit-learn.org/stable/modules/ensemble.html#forest

THANK YOU