ALMABETTER

Appliances Energy Prediction

Supervised Machine Learning Regression

Capstone Project 02

Mr.Pankaj Beldar

2022

ALMABETTER

Supervised Machine Learning Regression Capstone Project: 02

Appliances Energy Prediction

Pankaj Beldar

Abstract- This Capstone Project presents and discusses data-driven predictive models for the energy use of appliances. Data used include measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station and recorded energy use of lighting fixtures. The paper discusses data filtering to remove non-predictive parameters and feature ranking. When using all the predictors. From the wireless network, the data from the kitchen, laundry and living room were ranked the highest in importance for the energy prediction. The prediction models with only the weather data, selected the atmospheric pressure (which is correlated to wind speed) as the most relevant weather data variable in the prediction. Therefore, atmospheric pressure may be important to include in energy prediction models and for building performance modelling.

1. 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. For instance, in a study in the UK for domestic buildings, appliances, such as televisions and consumer electronics operating in standby were attributed to a near about 10% increase in the electricity consumption. 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 photovoltaics 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.

2. 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).

3. Dataset

- Dataset have 19735 rows and 29 attributes.
- Dataset doesn’t have any missing values.
- Dataset doesn’t have any Duplicate values.
- Data Attributes are as follows:

1. **date:** time year-month-day hour:minute:second
2. **Appliances:** energy use in Wh (Dependent variable)
3. **lights:** energy use of light fixtures in the house in Wh
4. **T1**, Temperature in kitchen area, in Celsius
5. **RH1**: Humidity in kitchen area, in %
6. **T2,** Temperature: in living room area, in Celsius
7. **RH2**: Humidity in living room area, in %
8. **T3:** Temperature in laundry room area
9. **RH3:** Humidity in laundry room area, in %
10. **T4:** Temperature in office room, in Celsius
11. **RH4:** Humidity in office room, in %
12. **T5:** Temperature in bathroom, in Celsius
13. **RH5:** Humidity in bathroom, in %
14. **T6**, Temperature outside the building (north side), in Celsius
15. **RH6:** Humidity outside the building (north side), in %
16. **T7:** Temperature in ironing room, in Celsius
17. **RH7:** Humidity in ironing room, in %
18. **T8**, Temperature in teenager room 2, in Celsius
19. **RH8**: Humidity in teenager room 2, in %
20. **T9:** Temperature in parents’ room, in Celsius
21. **RH9:** Humidity in parents’ room, in %
22. **To**, Temperature outside (from Chievres weather station), in Celsius
23. **Pressure** (from Chievres weather station), in mm Hg
24. **RHout**, Humidity outside (from Chievres weather station), in %
25. **Wind speed:** (from Chievres weather station), in m/s
26. **Visibility:** (from Chievres weather station), in km
27. **Tdewpoint:** (from Chievres weather station), Â°C

28. **rv1:** Random variable 1, no dimensional
29. **rv2:** Random variable 2, no dimensional

Where indicated, hourly data (then interpolated) from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis, rp5.ru. Permission was obtained from Reliable Prognosis for the distribution of the 4.5 months of weather data.

I have utilised the 'date' column to extract **hour, day of month, day of week, month, month name, day name** to find out more insights from time series data.

Let's create Categorical feature based on time as the phase of the day as Evening, Night, Morning, Afternoon

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**

Identify the Types of Features eg. numerical features, dependent feature, categorical feature

Numerical Features = ['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', 'hour', 'day_of_month']

Dependent Feature = ['Appliances']

Categorical Features =['Phase_of_Day','day_name','month_name']

4. Exploratory Data Analysis

What is Exploratory Data Analysis?
Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. An EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for a company because it exposes trends, patterns, and relationships that are not readily apparent.
What are the types of exploratory data analysis?
The four types of EDA are

1. Univariate non-graphical,
2. Multivariate non- graphical,
3. Univariate graphical,
4. Multivariate graphical.

5. Lets plot the distribution plot of each variable with distribution plot to check whether all variables are normally distributed or not.

a) Univariate Analysis

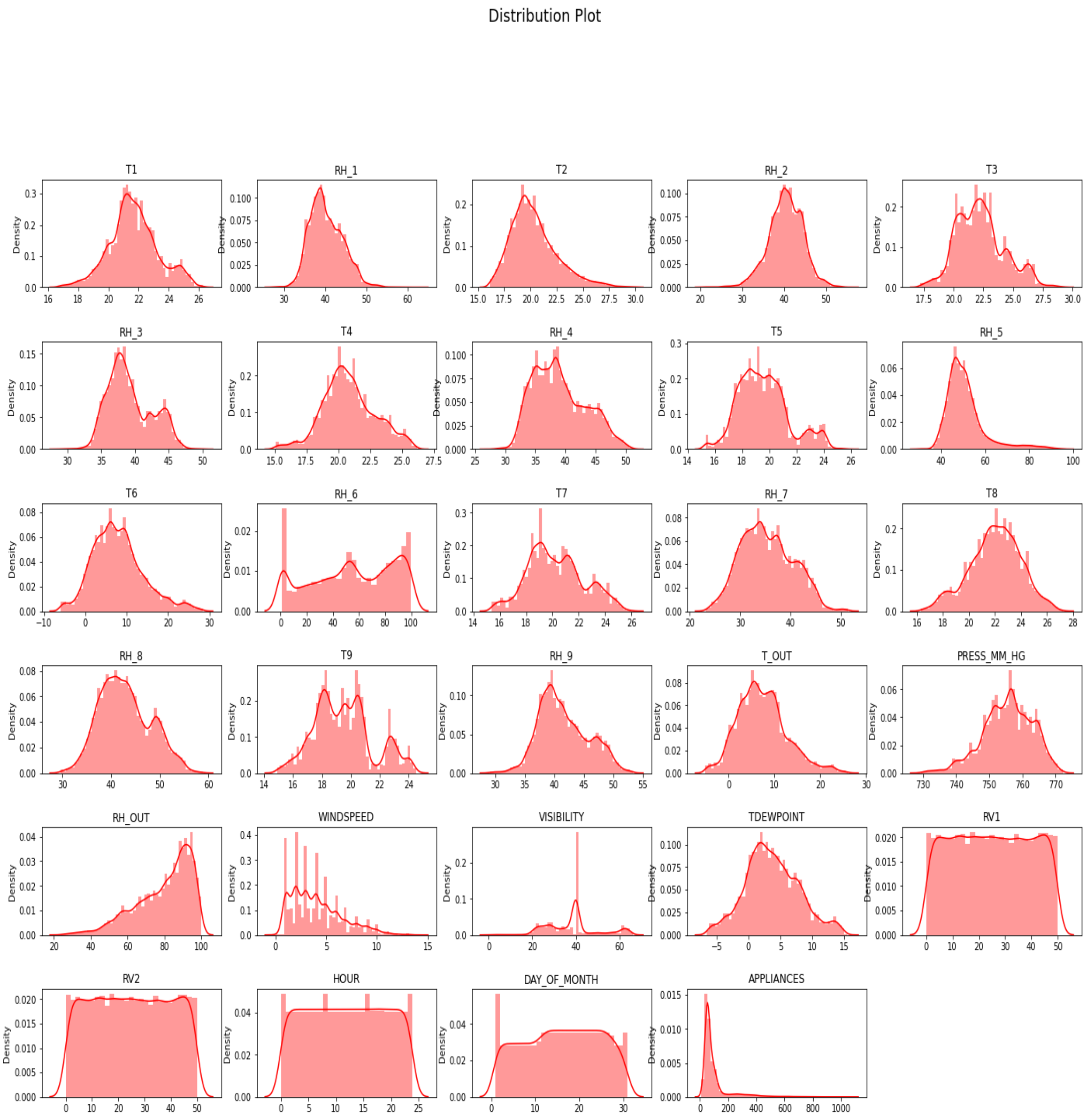


Fig: 01 Distribution Plot of all variables

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. From above distribution plot we can conclude that variables hour, rv2, day_of_month, appliances, rh6, rv1, T9, T3, windspeed, visibility are not normally distributed. other variables are seeming to be normally distributed.

as Appliances is our dependent variable hence to find out the pattern of correlation with dependent variable, we have plotted the scatter plot of each variable with dependent variable.

1. Energy consumption in kitchen area is less as compared to living room area.
2. Energy consumption bathroom, laundry area is more.
3. Maximum Energy consumption is from 12pm to 6pm time period.

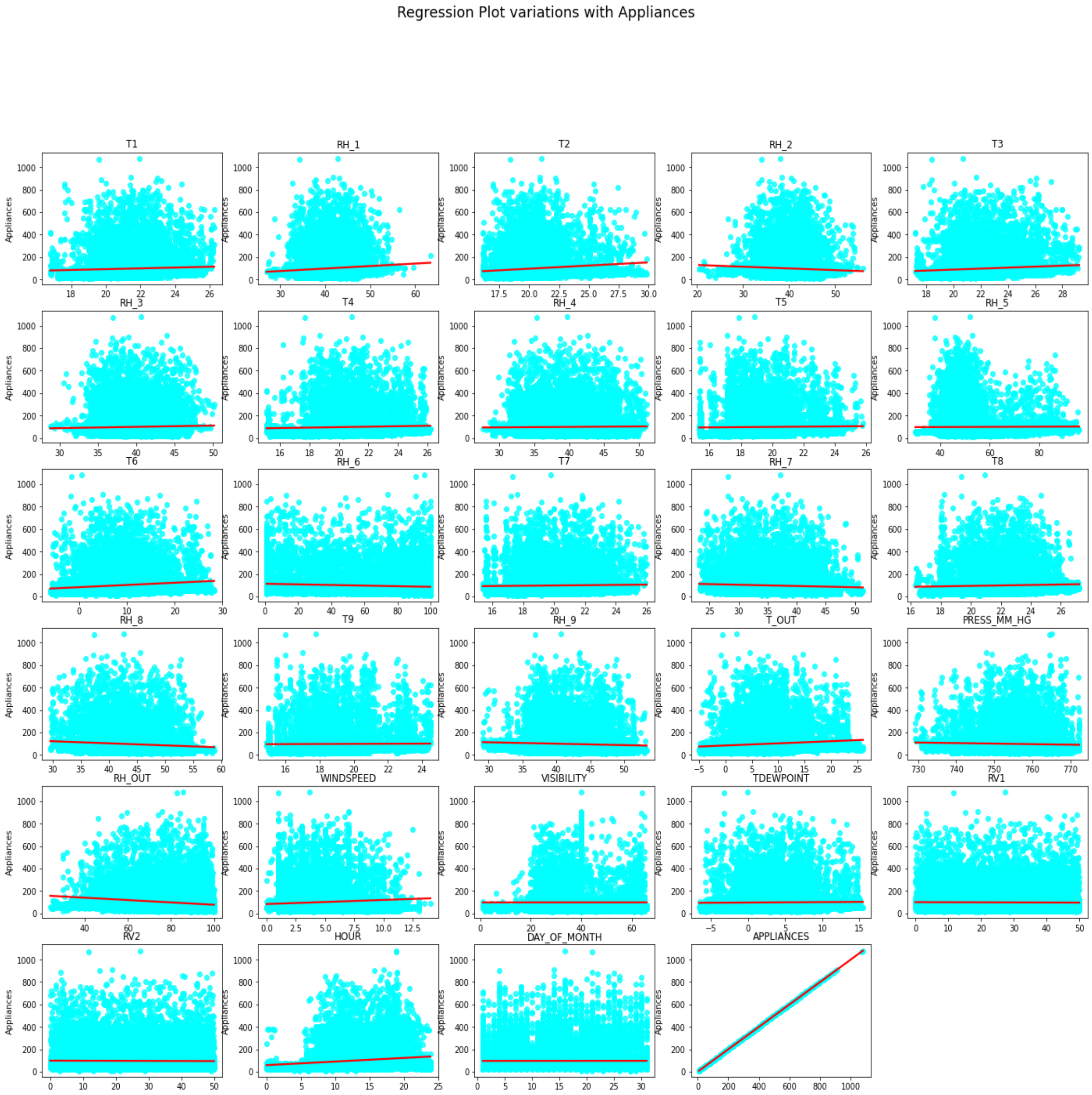


Fig: 02 Regression plot of all variables with dependent variable

When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or QQplot. This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

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

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

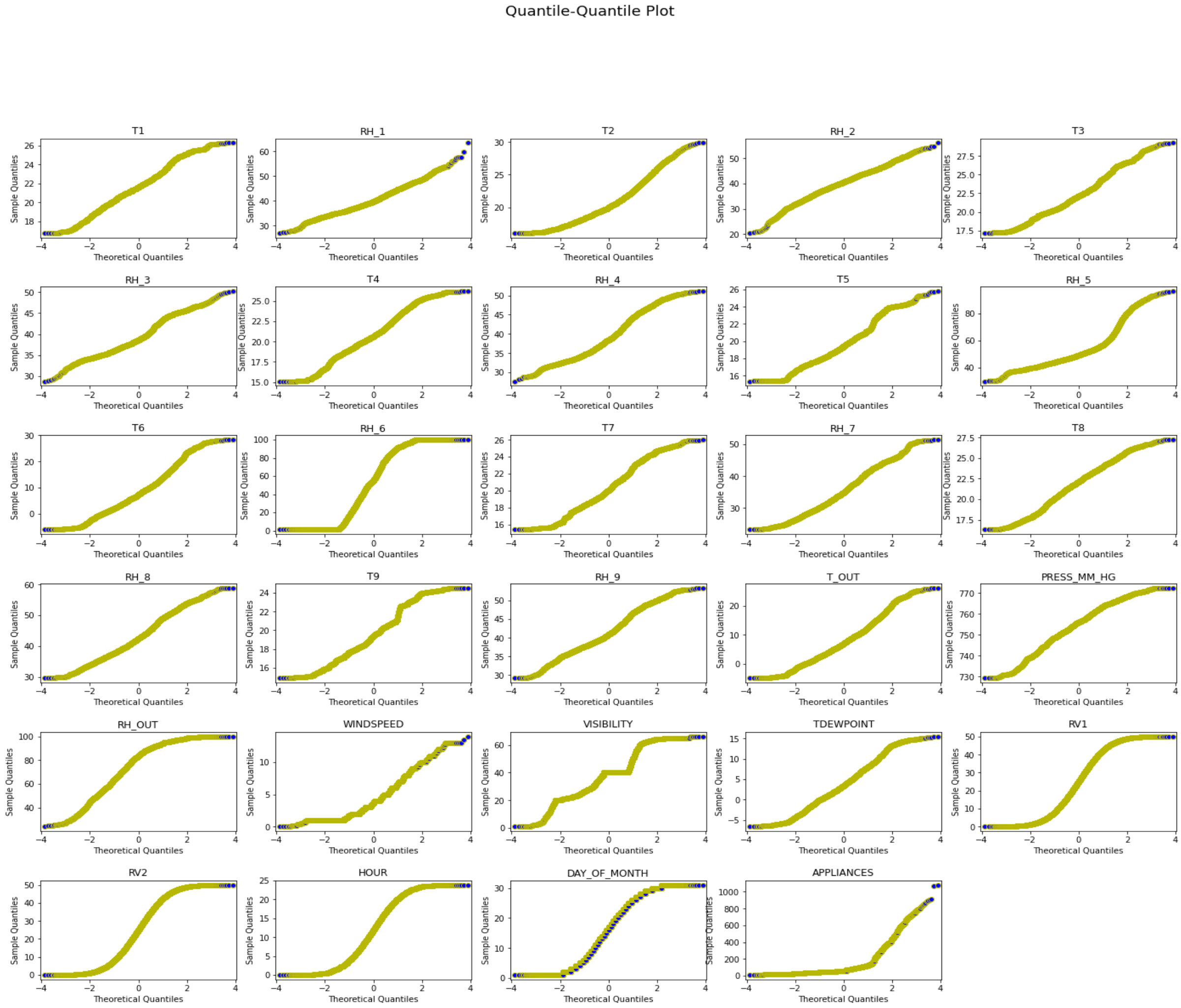


Fig: 03 QQ plot

1. Energy consumption on Monday is high, followed by Saturday, Friday and Sunday.
2. Energy consumption from 9am to 8 pm is high. whereas is less in between 12 am to 6am. its maximum at evening.

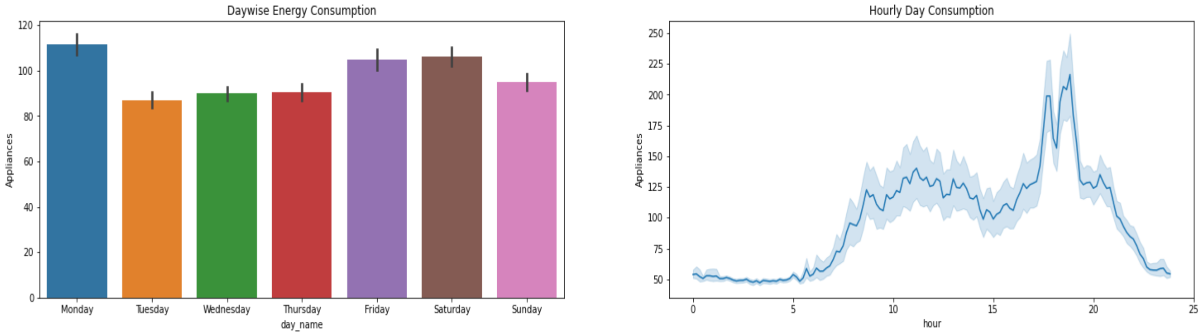


Fig: 04 Day wise and Hourly Energy Consumption

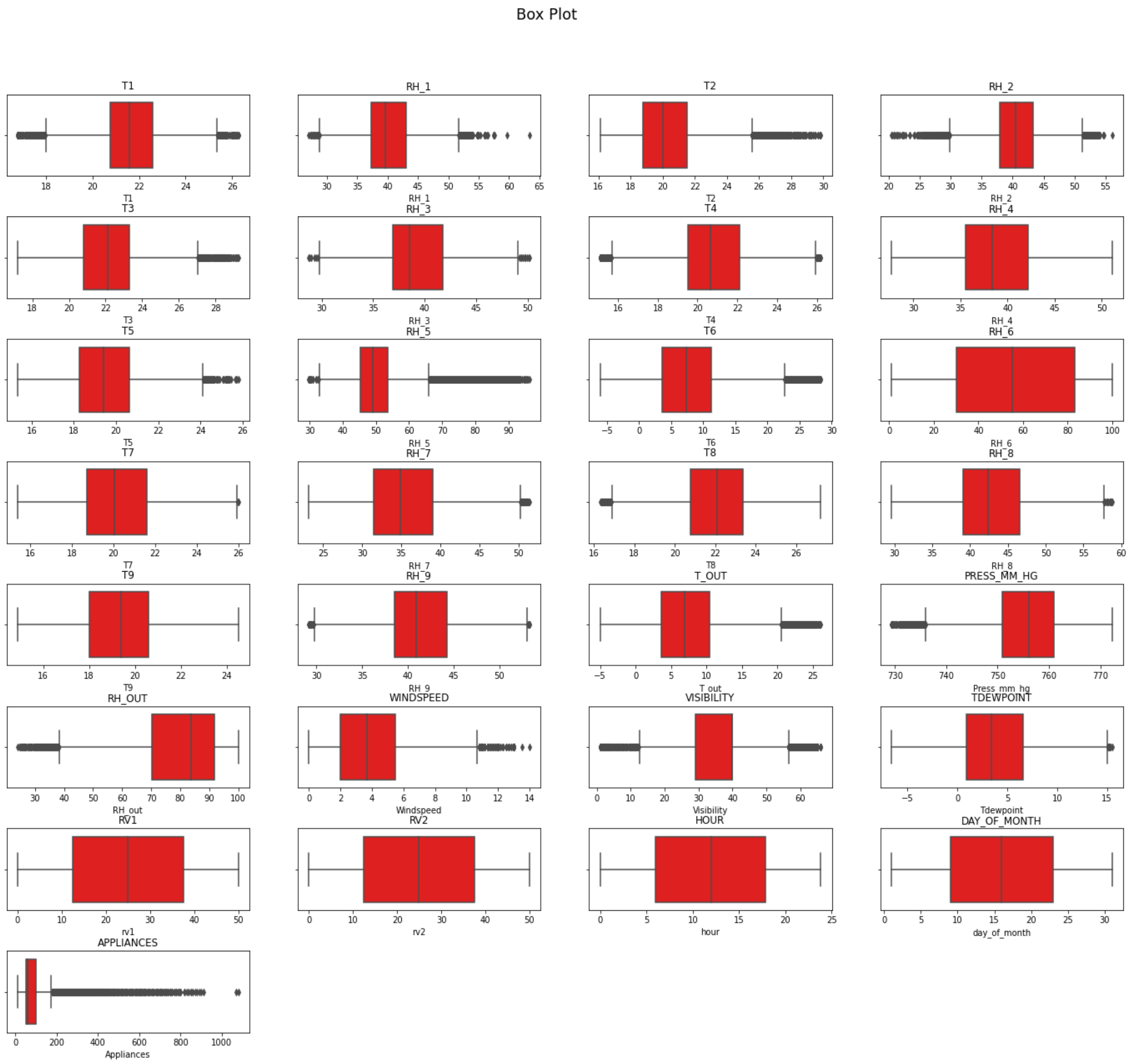


Fig: 05 Box plot

Most of the appliances consume energy in between 0-200 Wh. We can clearly see that many variables consist of outliers. T1, T2, T3, T4, T5, T6, T7, T8 and T_OUT have outliers. RH_1, RH_2, RH_3, RH_5, RH_7, RH_8, RH_9 and RH_OUT have outliers Windspeed, Tdewpoint, Visibility and the target variable Appliances also have outliers.

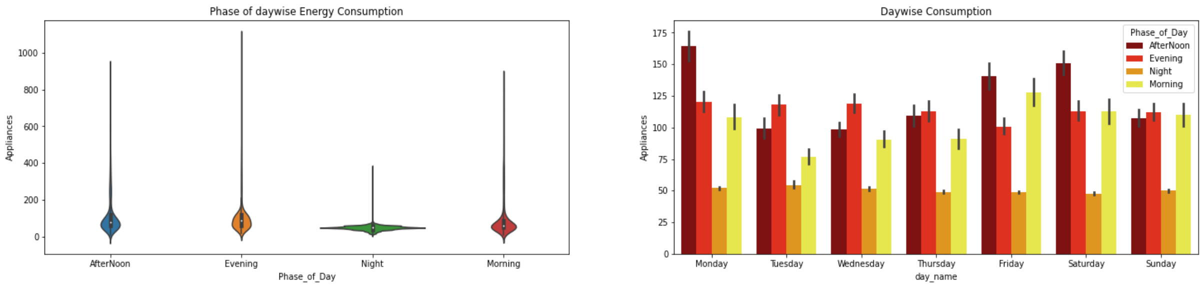
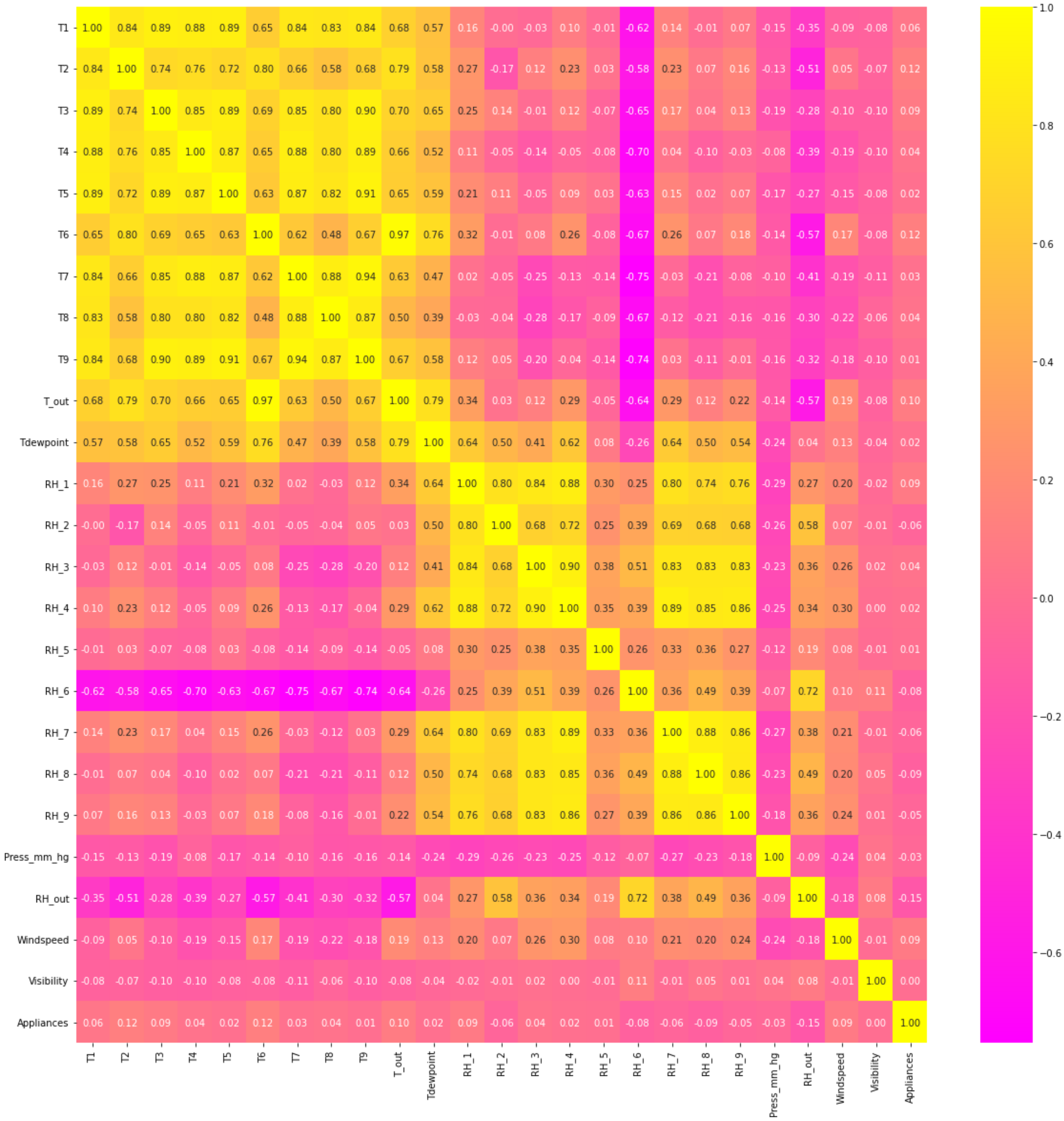


Fig: 06 Day wise and Hourly Energy Consumption

Energy consumption at night is less as compared to morning and afternoon on every weekday. Energy consumption is high in the evening. whereas its high on Monday, Friday and Saturday in the afternoon. Most of the appliances consumes around 200 Wh energy

b) Bivariate Analysis



	T1	T2	T3	T4	T5	T6	T7	T8	T9	T_out	Tdewpoint	RH_1	RH_2	RH_3	RH_4	RH_5	RH_6	RH_7	RH_8	RH_9	Press_mm_hg	RH_out	Windspeed	Visibility	Appliances
T1	1.00	0.84	0.89	0.88	0.89	0.65	0.84	0.83	0.84	0.68	0.57	0.16	-0.00	-0.03	0.10	-0.01	-0.62	0.14	-0.01	0.07	-0.15	-0.35	-0.09	-0.08	0.06
T2	0.84	1.00	0.74	0.76	0.72	0.80	0.66	0.58	0.68	0.79	0.58	0.27	-0.17	0.12	0.23	0.03	-0.58	0.23	0.07	0.16	-0.13	-0.51	0.05	-0.07	0.12
T3	0.89	0.74	1.00	0.85	0.89	0.69	0.85	0.80	0.90	0.70	0.65	0.25	0.14	-0.01	0.12	-0.07	-0.65	0.17	0.04	0.13	-0.19	-0.28	-0.10	-0.10	0.09
T4	0.88	0.76	0.85	1.00	0.87	0.65	0.88	0.80	0.89	0.66	0.52	0.11	-0.05	-0.14	-0.05	-0.08	-0.70	0.04	-0.10	-0.03	-0.08	-0.39	-0.19	-0.10	0.04
T5	0.89	0.72	0.89	0.87	1.00	0.63	0.87	0.82	0.91	0.65	0.59	0.21	0.11	-0.05	0.09	0.03	-0.63	0.15	0.02	0.07	-0.17	-0.27	-0.15	-0.08	0.02
T6	0.65	0.80	0.69	0.65	0.63	1.00	0.62	0.48	0.67	0.97	0.76	0.32	-0.01	0.08	0.26	-0.08	-0.67	0.26	0.07	0.18	-0.14	-0.57	0.17	-0.08	0.12
T7	0.84	0.66	0.85	0.88	0.87	0.62	1.00	0.88	0.94	0.63	0.47	0.02	-0.05	-0.25	-0.13	-0.14	-0.75	-0.03	-0.21	-0.08	-0.10	-0.41	-0.19	-0.11	0.03
T8	0.83	0.58	0.80	0.80	0.82	0.48	0.88	1.00	0.87	0.50	0.39	-0.03	-0.04	-0.28	-0.17	-0.09	-0.67	-0.12	-0.21	-0.16	-0.16	-0.30	-0.22	-0.06	0.04
T9	0.84	0.68	0.90	0.89	0.91	0.67	0.94	0.87	1.00	0.67	0.58	0.12	0.05	-0.20	-0.04	-0.14	-0.74	0.03	-0.11	-0.01	-0.16	-0.32	-0.18	-0.10	0.01
T_out	0.68	0.79	0.70	0.66	0.65	0.97	0.63	0.50	0.67	1.00	0.79	0.34	0.03	0.12	0.29	-0.05	-0.64	0.29	0.12	0.22	-0.14	-0.57	0.19	-0.08	0.10
Tdewpoint	0.57	0.58	0.65	0.52	0.59	0.76	0.47	0.39	0.58	0.79	1.00	0.64	0.50	0.41	0.62	0.08	-0.26	0.64	0.50	0.54	-0.24	0.04	0.13	-0.04	0.02
RH_1	0.16	0.27	0.25	0.11	0.21	0.32	0.02	-0.03	0.12	0.34	0.64	1.00	0.80	0.84	0.88	0.30	0.25	0.80	0.74	0.76	-0.29	0.27	0.20	-0.02	0.09
RH_2	-0.00	-0.17	0.14	-0.05	0.11	-0.01	-0.05	-0.04	0.05	0.03	0.50	0.80	1.00	0.68	0.72	0.25	0.39	0.69	0.68	0.68	-0.26	0.58	0.07	-0.01	-0.06
RH_3	-0.03	0.12	-0.01	-0.14	-0.05	0.08	-0.25	-0.28	-0.20	0.12	0.41	0.84	0.68	1.00	0.90	0.38	0.51	0.83	0.83	0.83	-0.23	0.36	0.26	0.02	0.04
RH_4	0.10	0.23	0.12	-0.05	0.09	0.26	-0.13	-0.17	-0.04	0.29	0.62	0.88	0.72	0.90	1.00	0.35	0.39	0.89	0.85	0.86	-0.25	0.34	0.30	0.00	0.02
RH_5	-0.01	0.03	-0.07	-0.08	0.03	-0.08	-0.14	-0.09	-0.14	-0.05	0.08	0.30	0.25	0.38	0.35	1.00	0.26	0.33	0.36	0.27	-0.12	0.19	0.08	-0.01	0.01
RH_6	-0.62	-0.58	-0.65	-0.70	-0.63	-0.67	-0.75	-0.67	-0.74	-0.64	-0.26	0.25	0.39	0.51	0.39	0.26	1.00	0.36	0.49	0.39	-0.07	0.72	0.10	0.11	-0.08
RH_7	0.14	0.23	0.17	0.04	0.15	0.26	-0.03	-0.12	0.03	0.29	0.64	0.80	0.69	0.83	0.89	0.33	0.36	1.00	0.88	0.86	-0.27	0.38	0.21	-0.01	-0.06
RH_8	-0.01	0.07	0.04	-0.10	0.02	0.07	-0.21	-0.21	-0.11	0.12	0.50	0.74	0.68	0.83	0.85	0.36	0.49	0.88	1.00	0.86	-0.23	0.49	0.20	0.05	-0.09
RH_9	0.07	0.16	0.13	-0.03	0.07	0.18	-0.08	-0.16	-0.01	0.22	0.54	0.76	0.68	0.83	0.86	0.27	0.39	0.86	0.86	1.00	-0.18	0.36	0.24	0.01	-0.05
Press_mm_hg	-0.15	-0.13	-0.19	-0.08	-0.17	-0.14	-0.10	-0.16	-0.16	-0.14	-0.24	-0.29	-0.26	-0.23	-0.25	-0.12	-0.07	-0.27	-0.23	-0.18	1.00	-0.09	-0.24	0.04	-0.03
RH_out	-0.35	-0.51	-0.28	-0.39	-0.27	-0.57	-0.41	-0.30	-0.32	-0.57	0.04	0.27	0.58	0.36	0.34	0.19	0.72	0.38	0.49	0.36	-0.09	1.00	-0.18	0.08	-0.15
Windspeed	-0.09	0.05	-0.10	-0.19	-0.15	0.17	-0.19	-0.22	-0.18	0.19	0.13	0.20	0.07	0.26	0.30	0.08	0.10	0.21	0.20	0.24	-0.24	-0.18	1.00	-0.01	0.09
Visibility	-0.08	-0.07	-0.10	-0.10	-0.08	-0.08	-0.11	-0.06	-0.10	-0.08	-0.04	-0.02	-0.01	0.02	0.00	-0.01	0.11	-0.01	0.05	0.01	0.04	0.08	-0.01	1.00	0.00
Appliances	0.06	0.12	0.09	0.04	0.02	0.12	0.03	0.04	0.01	0.10	0.02	0.09	-0.06	0.04	0.02	0.01	-0.08	-0.06	-0.09	-0.05	-0.03	-0.15	0.09	0.00	1.00

Fig: 07 Heatmap of Correlation

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

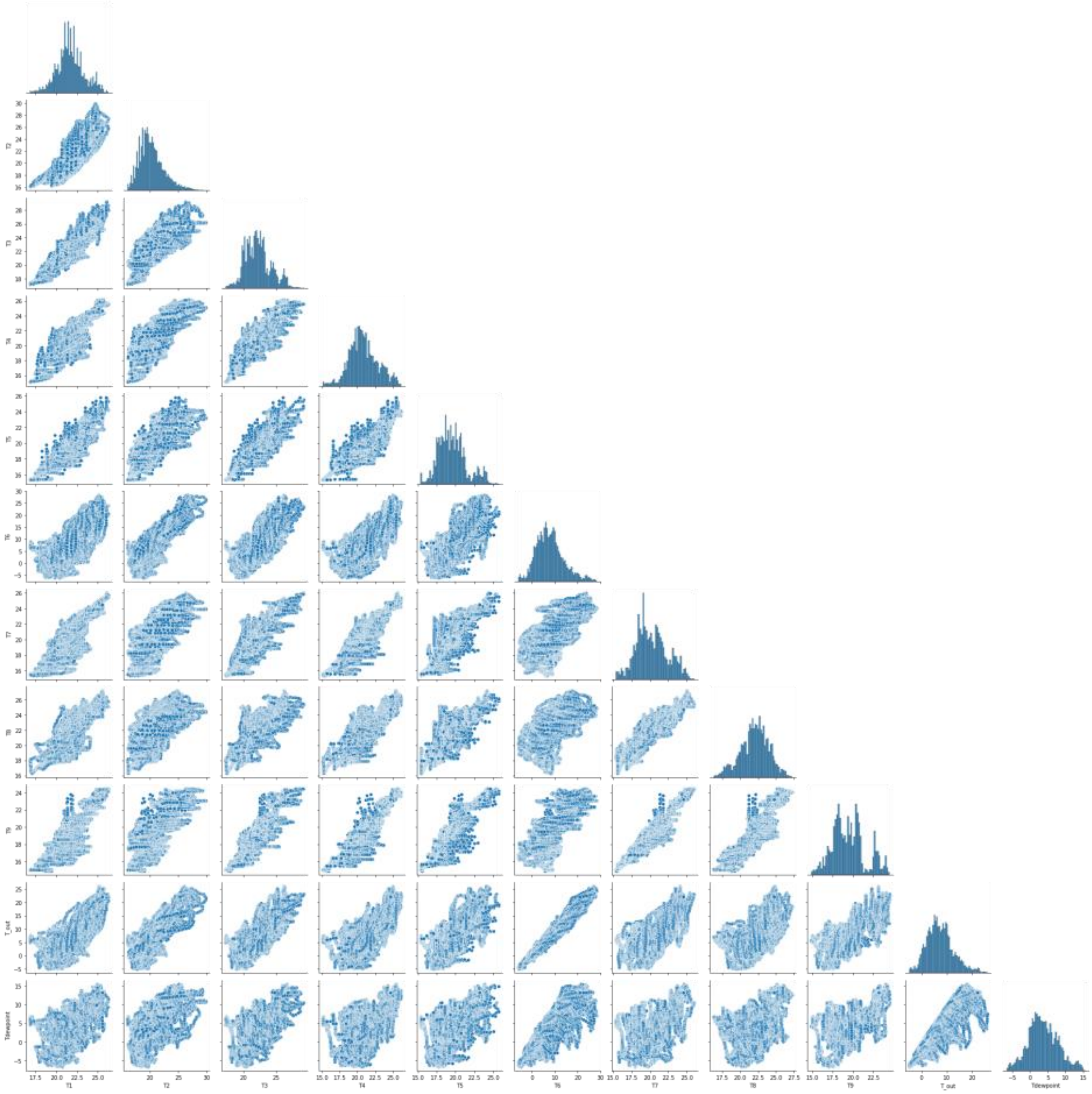


Fig:08 Pair plot of all Temperature Variables

Almost all Temperature variables are linearly correlated with each other.

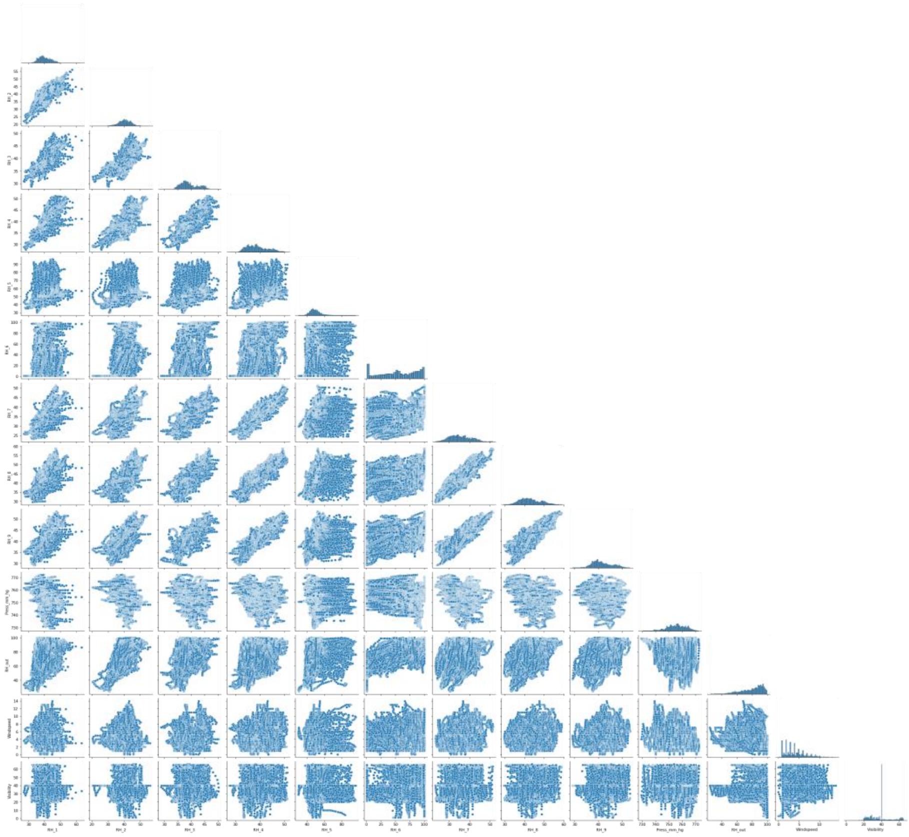


Fig: 09 Pair plot of Relative Humidity

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

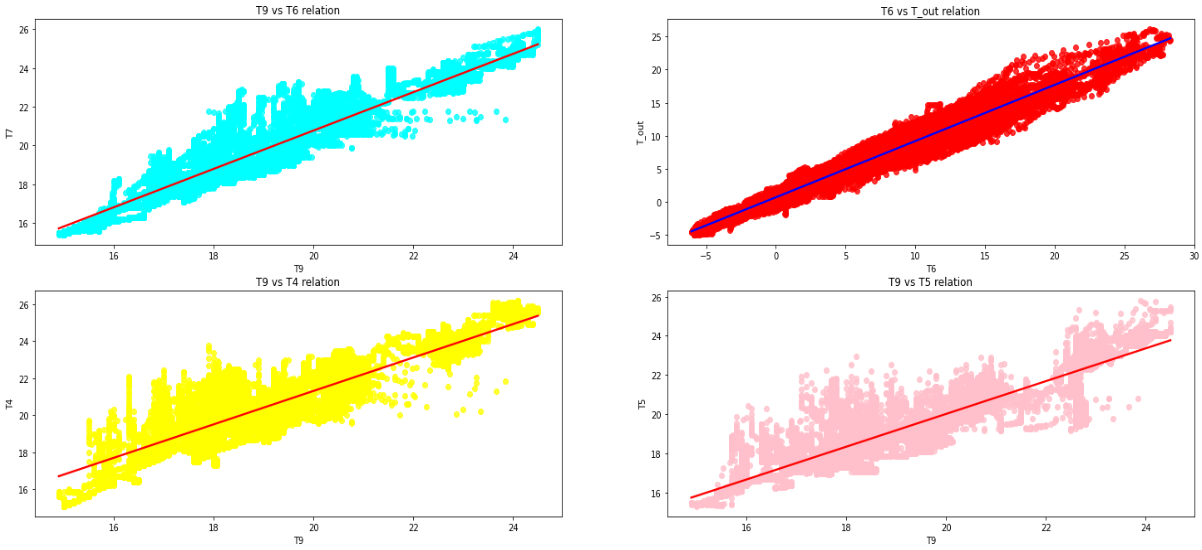


Fig:10 Regression plot of Temperature Variables

T9, T5, T4, T_out, T6 are highly correlated which causes multicollinearity Since 'rv1' and 'rv2' have correlation of 1, we can drop either of them. Since both the features are identical, one of them is eliminated from the dataset. Now let’s check, day wise hourly distribution of data over all variables.Temperature of Kitchen, bathroom, living area, outside the building (north side) and parents’ room is less in night also Energy consumption of appliances is less during night. Relative Humidity is more at night.

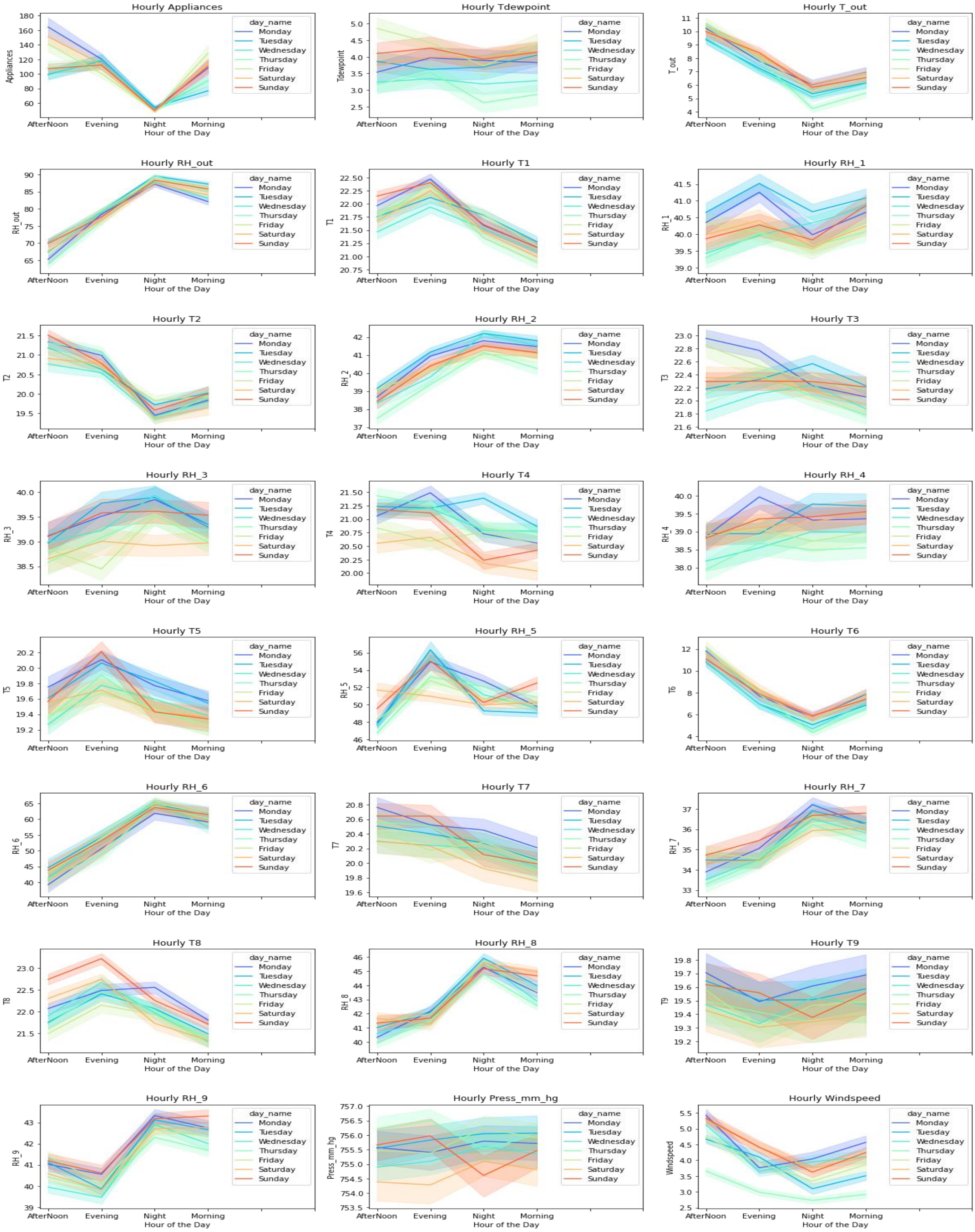


Fig: 11 Line plot of all features with phase of time

6. Feature Engineering and Selection

Removing Outliers-

Outliers are those data points that are significantly different from the rest of the dataset. Machine learning algorithms are susceptible to the statistics and distribution of the input variables. Data outliers can spoil and mislead the training process. That results in longer training times, less accurate models, and poor results. Hence, we have detected the outliers. As discussed in boxplot, many variables are having outliers hence we need to deal with outliers. We can use the IQR method of identifying outliers to set up a "fence" outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. Inter Quantile method is used to remove outliers. Outliers are capped in between lower and upper limit. Figure below shows the Boxplot without outliers.

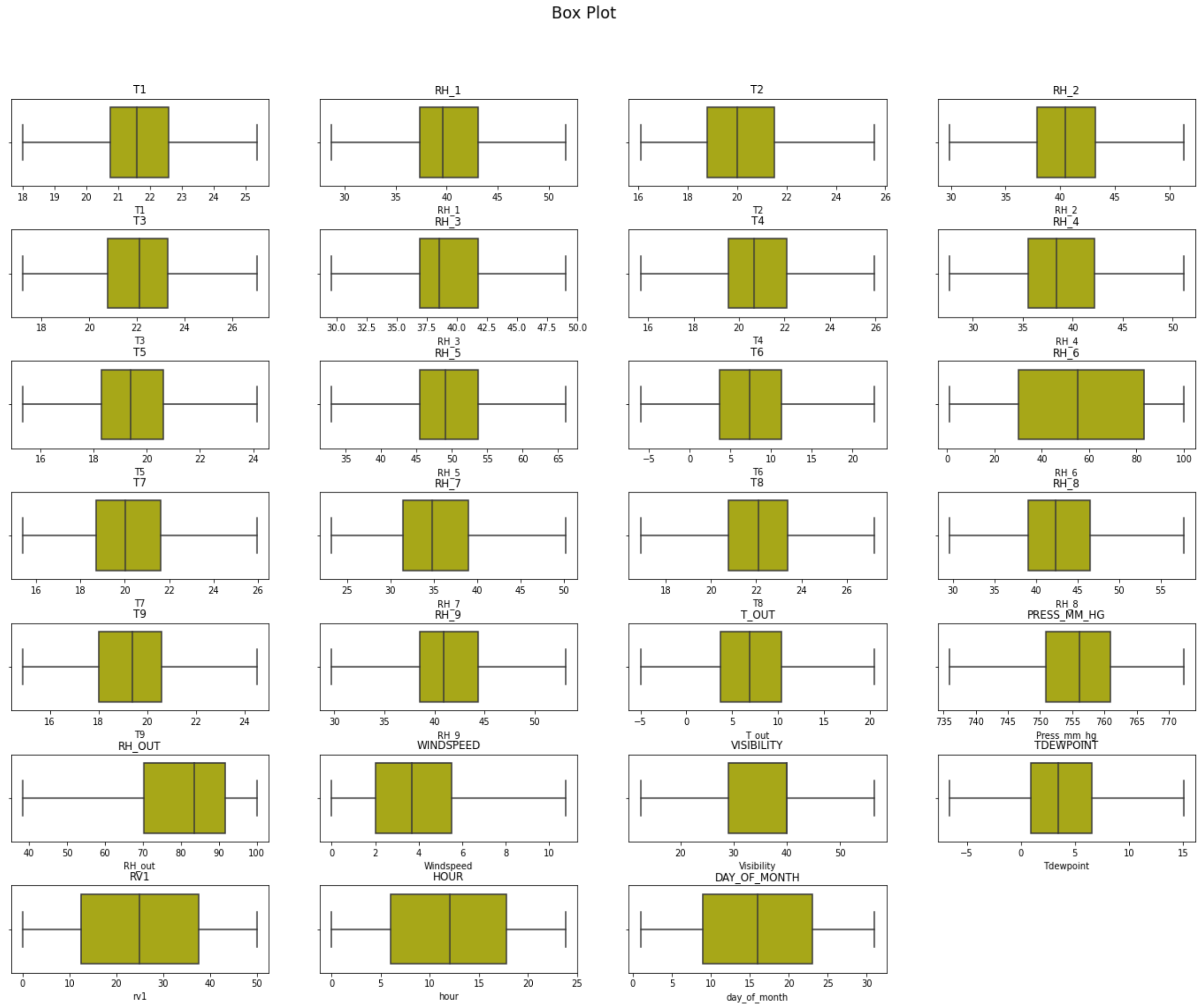


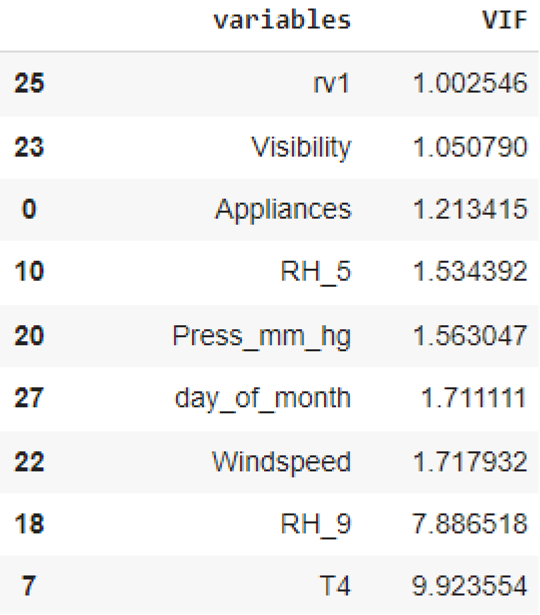
Fig:12 Box plot without Outliers

6. Encoding Categorical Variables

One hot encoding can be defined as the essential process of converting the categorical data variables to be provided to machine and deep learning algorithms which in turn improve predictions as well as classification accuracy of a model.

7. Check Multicollinearity

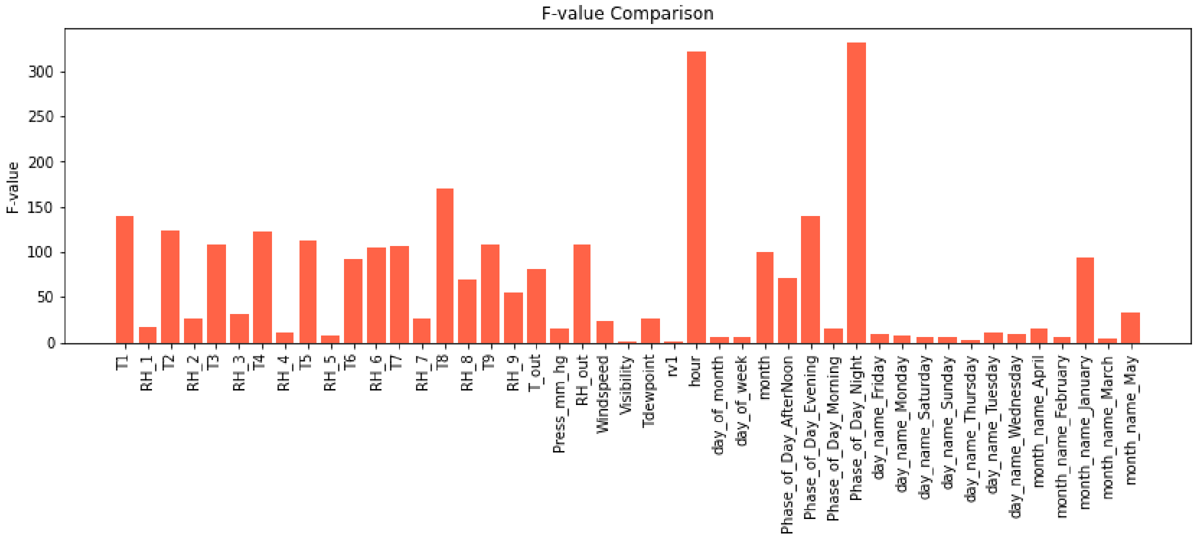
A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. Small VIF values, VIF < 3, indicate low correlation among variables under ideal conditions. The default VIF cutoff value is 5; only variables with a VIF less than 5 will be included in the model. However, note that many sources say that a VIF of less than 10 is acceptable.



	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

c) ANOVA F Value and Variance Threshold

The F value is used in analysis of variance (ANOVA). It is calculated by dividing two mean squares. This calculation determines the ratio of explained variance to unexplained variance.



Fg:13 F Value Comparison of most dominant features 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' are significant features as per F value and variance threshold method. We use above variables for further variables.

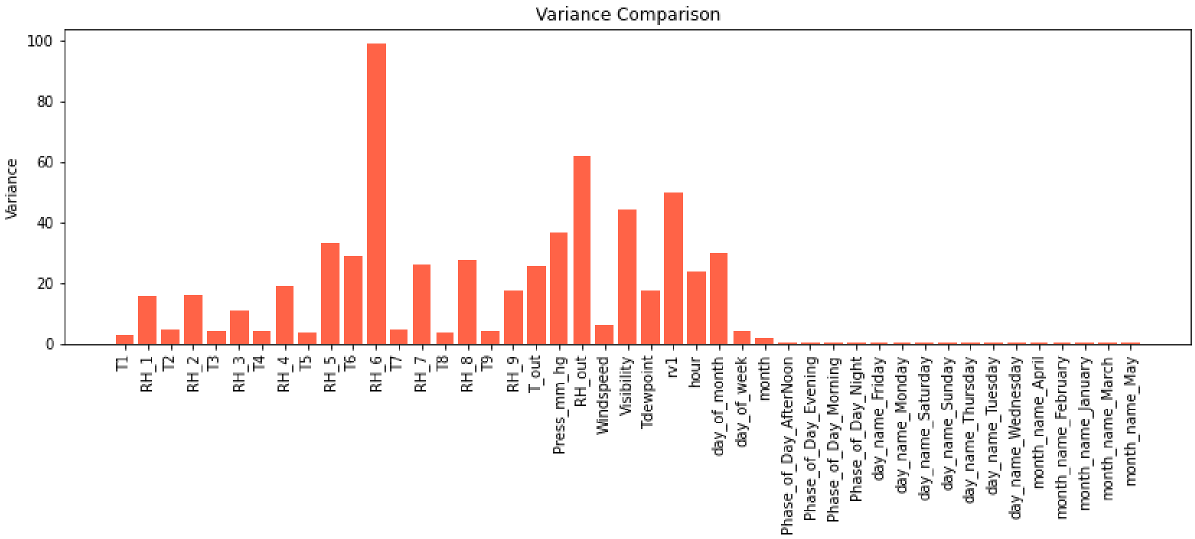


Fig:14 Variance Threshold Comparison

Variance threshold is **a simple baseline approach to feature selection**. It removes all features whose variance doesn't meet some threshold as it is assumed that features with a higher variance may contain more useful information.

8. Developing Machine Learning Model

a) Train Test Split

The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications. This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results.

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

b) Scaling Dataset

Python sklearn library offers us with StandardScaler() function to standardize the data values into a standard format. According to the above syntax, we initially create an object of the StandardScaler() function. Further, we use fit_transform() along with the assigned object to transform the data and standardize it. Figure below shows the Histogram of all scaled variables.

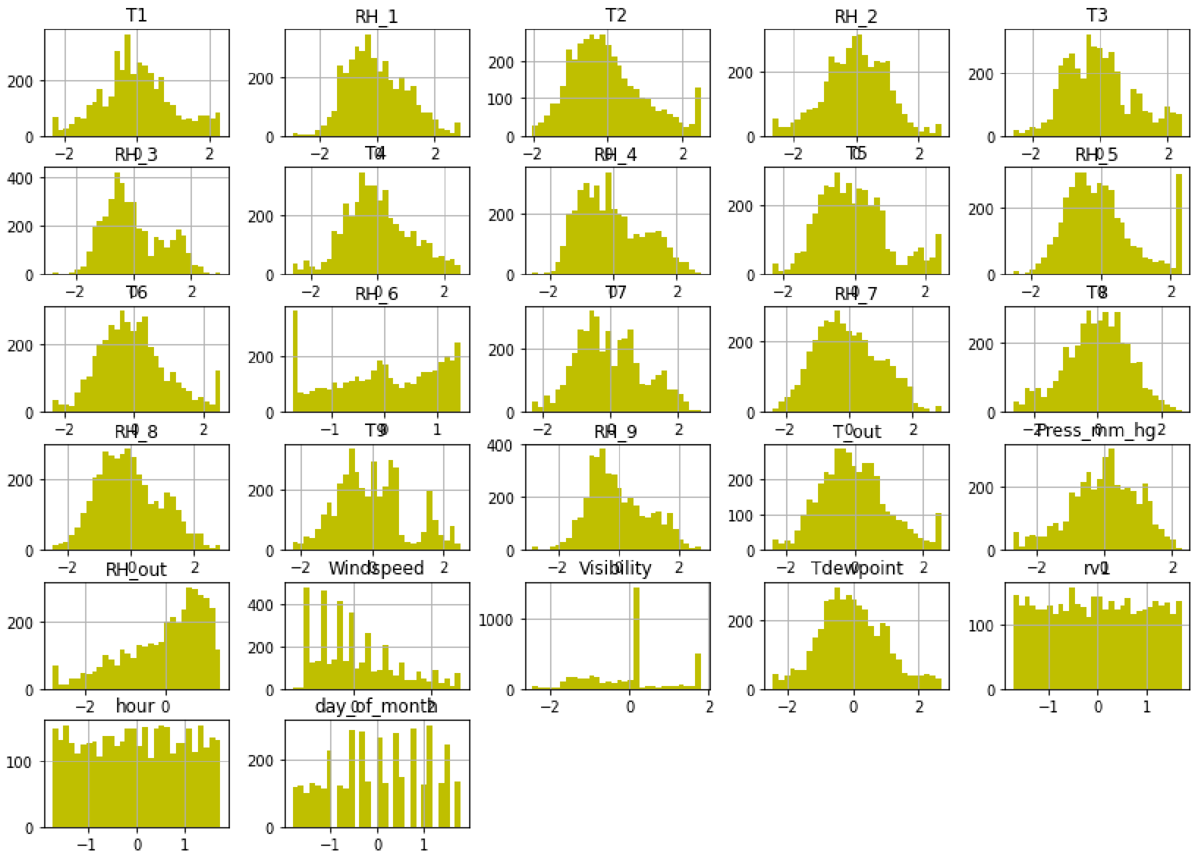


Fig: 15 Histogram of Features with Standard Scaler

c) Functions used for Developing ML Model

All the regression models were trained with 10-fold cross validation to select the best. The first model trained was the multiple linear regression. The multiple linear regression uses all the available predictors and finds the appropriate slope quantifying the effect of each predictor and the response is obvious that the relationship between the variables and the energy consumption of appliances is not well represented by the linear model since the residuals are not normally distributed around the horizontal axis. In order to compare the performance of each of the regression models, different performance evaluation indices are used here: the root mean squared error (RMSE), the coefficient of determination or R-squared/R2, the mean absolute error (MAE) and the mean absolute percentage error (MAPE):

$$\text{RMSE} = \sqrt{\frac{\sum_{n}^{i=1}(Y_i - \hat{Y}_i)^2}{n}}$$

$$R^2 = 1 - \frac{\sum_{n}^{i=1}(Y_i - \hat{Y}_i)^2}{\sum_{n}^{i=1}(Y_i - \bar{Y})^2}$$

$$\text{MAE} = \frac{\sum_{n}^{i=1} \left| Y_i - \hat{Y}_i \right|}{n}$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{\left| Y_i - \hat{Y}_i \right|}{Y_i}$$

Figure below shows the KNN Regressor Model Performance on actual and predicted values.

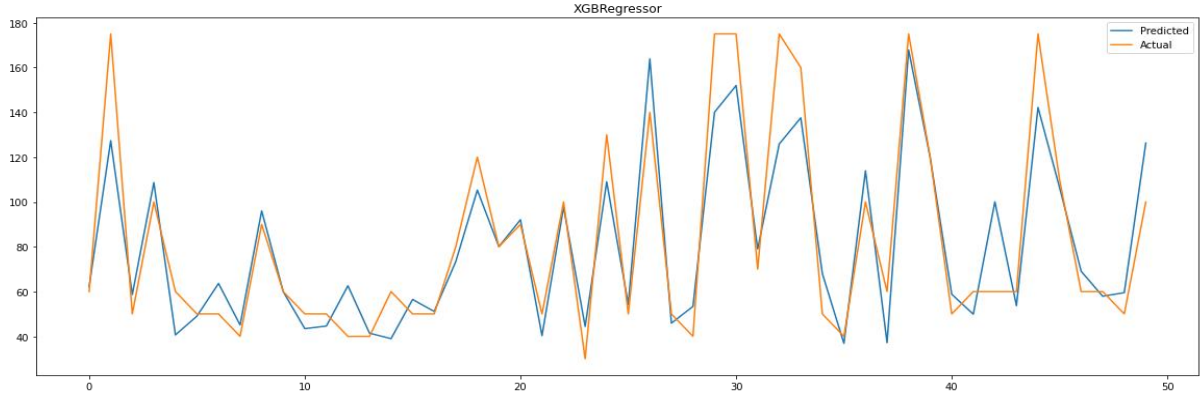


Fig:16 Sample Comparison of Actual and Predicted values of XGBoost Regression

Figure below shows the Random Forest Regressor Model Performance on actual and predicted values.

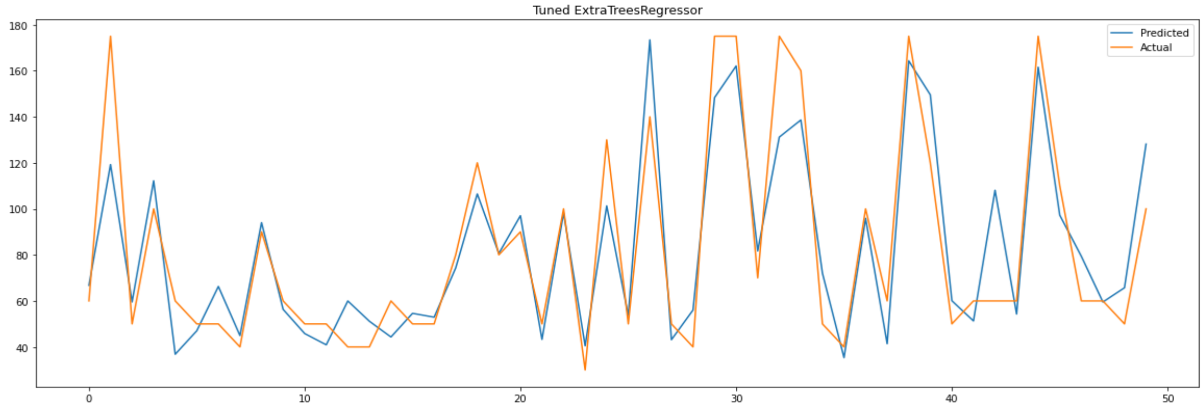
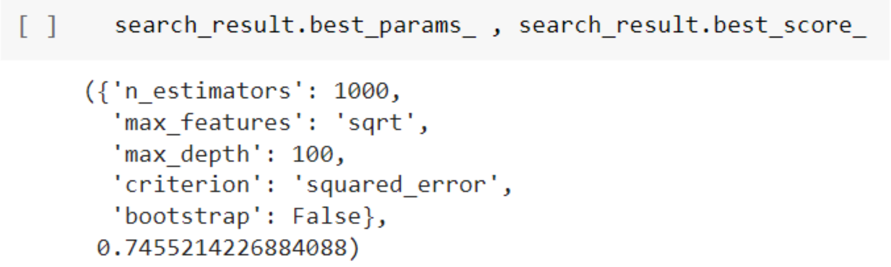


Fig:17 Sample Comparison of Actual and Predicted values of Extra Tree Regression

d) HyperParameter Tuning

Hyperparameter 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. However, evaluating each model only on the training set can lead to one of the most fundamental problems in machine learning overfitting. If we optimize the model for the training data, then our model will score very well on the training set, but will not be able to generalize to new data, such as in a test set. When a model performs highly on the training set but poorly on the test set, this is known as overfitting, or essentially creating a model that knows the training set very well but cannot be applied to new problems. It's like a student who has memorized the simple problems in the textbook but has no idea how to apply concepts in the messy real world.



```
[ ]    search_result.best_params_ , search_result.best_score_

      ({'n_estimators': 1000,
        'max_features': 'sqrt',
        'max_depth': 100,
        'criterion': 'squared_error',
        'bootstrap': False},
       0.7455214226884088)
```

e. Cross Validation

The technique of cross validation (CV) is best explained by example using the most common method, K-Fold CV. When we approach a machine learning problem, we make sure to split our data into a training and a testing set. In K-Fold CV, we further split our training set into K number of subsets, called folds. We then iteratively fit the model K times, each time training the data on K-1 of the folds and evaluating on the Kth fold (called the validation data). As an example, consider fitting a model with K = 5. The first iteration we train on the first four folds and evaluate on the fifth. The second time we train on the first, second, third, and fifth fold and evaluate on the fourth. We repeat this procedure 3 more times, each time evaluating on a different fold. At the very end of training, we average the performance on each of the folds to come up with final validation metrics for the model.

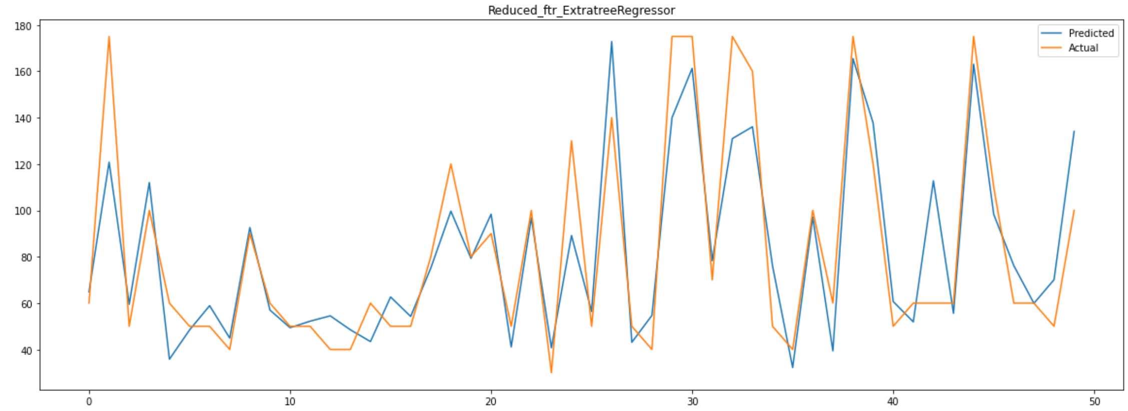
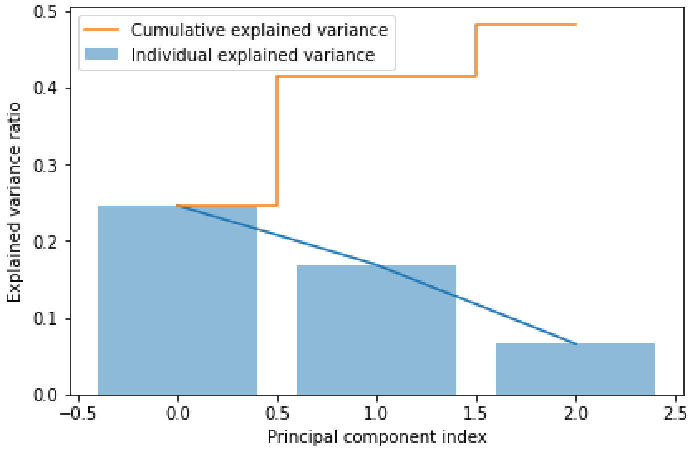
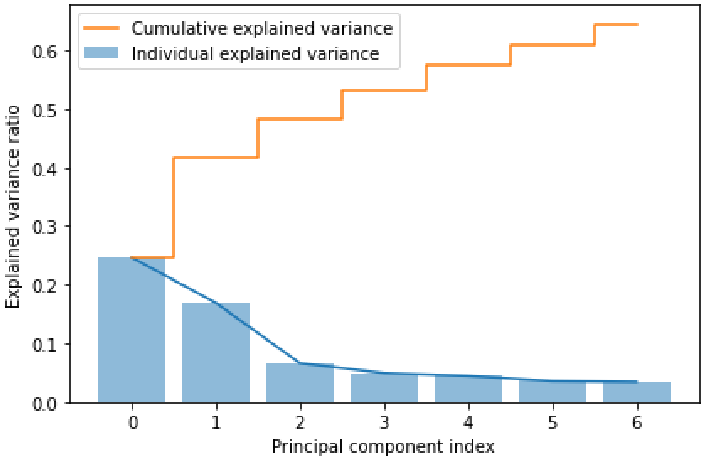


Fig:18 Sample Comparison of Actual and Predicted values of Extra Tree Regression with hyperparameter tuning

f. Model with Principal Component Analysis

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features. Hence it is used for feature reduction. 2 principal components are sufficient for analysis. Over 90% data is influenced by principal component 1 and 2.





g. Optimal Model with Feature Reduction

Figure below shows the optimal regression model with R2 score 0.78 and following features as significant features.

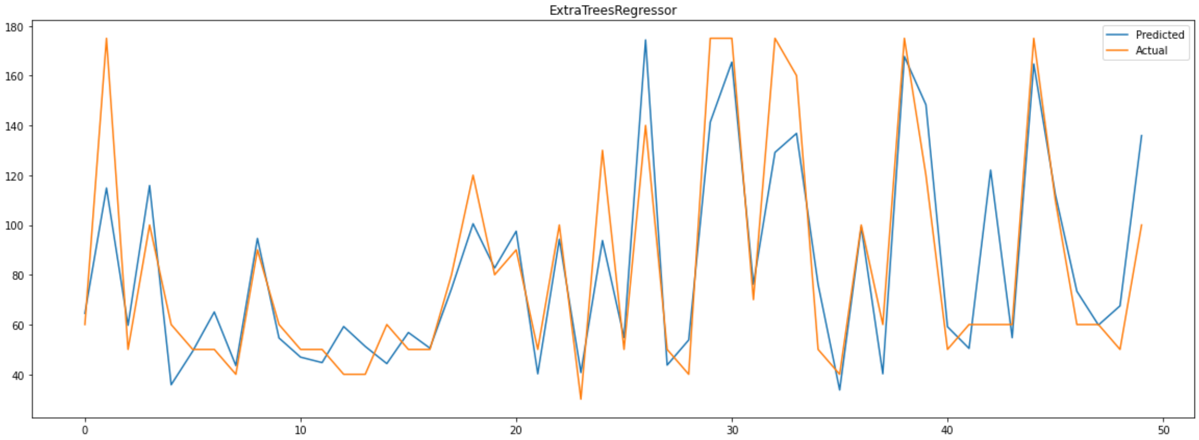
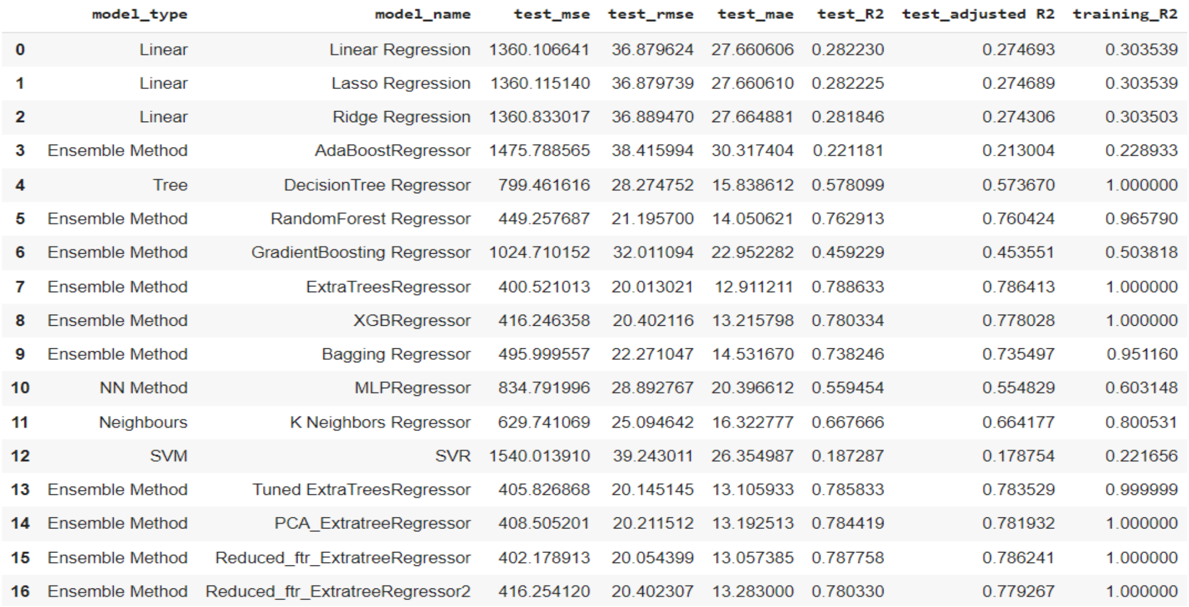


Fig:19 Sample Comparison of Actual and Predicted values of Extra Tree regression

Features=['T1','RH_1','T2','RH_2','T3','RH_3','T4','RH_4','T5','RH_5','RH_6','T7','RH_7','T8','RH_8','RH_9','T_out','Press_mm_hg','RH_out','Windspeed','Tdewpoint','hour','Phase_of_Day_After Noon','Phase_of_Day_Evening','Phase_of_Day_Morning','Phase_of_Day_Night']

Table below shows the Comparison between different trained models with R2score, adjusted R2 score, MAE, RMSE of test data and R2 score of Training data



	model_type	model_name	test_mse	test_rmse	test_mae	test_R2	test_adjusted R2	training_R2
0	Linear	Linear Regression	1360.106641	36.879624	27.660606	0.282230	0.274693	0.303539
1	Linear	Lasso Regression	1360.115140	36.879739	27.660610	0.282225	0.274689	0.303539
2	Linear	Ridge Regression	1360.833017	36.889470	27.664881	0.281846	0.274306	0.303503
3	Ensemble Method	AdaBoostRegressor	1475.788565	38.415994	30.317404	0.221181	0.213004	0.228933
4	Tree	DecisionTree Regressor	799.461616	28.274752	15.838612	0.578099	0.573670	1.000000
5	Ensemble Method	RandomForest Regressor	449.257687	21.195700	14.050621	0.762913	0.760424	0.965790
6	Ensemble Method	GradientBoosting Regressor	1024.710152	32.011094	22.952282	0.459229	0.453551	0.503818
7	Ensemble Method	ExtraTreesRegressor	400.521013	20.013021	12.911211	0.788633	0.786413	1.000000
8	Ensemble Method	XGBRegressor	416.246358	20.402116	13.215798	0.780334	0.778028	1.000000
9	Ensemble Method	Bagging Regressor	495.999557	22.271047	14.531670	0.738246	0.735497	0.951160
10	NN Method	MLPRegressor	834.791996	28.892767	20.396612	0.559454	0.554829	0.603148
11	Neighbours	K Neighbors Regressor	629.741069	25.094642	16.322777	0.667666	0.664177	0.800531
12	SVM	SVR	1540.013910	39.243011	26.354987	0.187287	0.178754	0.221656
13	Ensemble Method	Tuned ExtraTreesRegressor	405.826868	20.145145	13.105933	0.785833	0.783529	0.999999
14	Ensemble Method	PCA_ExtratreeRegressor	408.505201	20.211512	13.192513	0.784419	0.781932	1.000000
15	Ensemble Method	Reduced_ftr_ExtratreeRegressor	402.178913	20.054399	13.057385	0.787758	0.786241	1.000000
16	Ensemble Method	Reduced_ftr_ExtratreeRegressor2	416.254120	20.402307	13.283000	0.780330	0.779267	1.000000

Extra Tree Regression Model with T1','RH_1','T2','RH_2','T3','RH_3','T4','RH_4','T5','RH_5','RH_6','T7','RH_7','T8','RH_8', 'RH_9','T_out','Press_mm_hg','RH_out','Windspeed' ,'Tdewpoint', hour have maximum R2 score as 0.78 after hyper parameter tuning and Less MAE, RMSE.

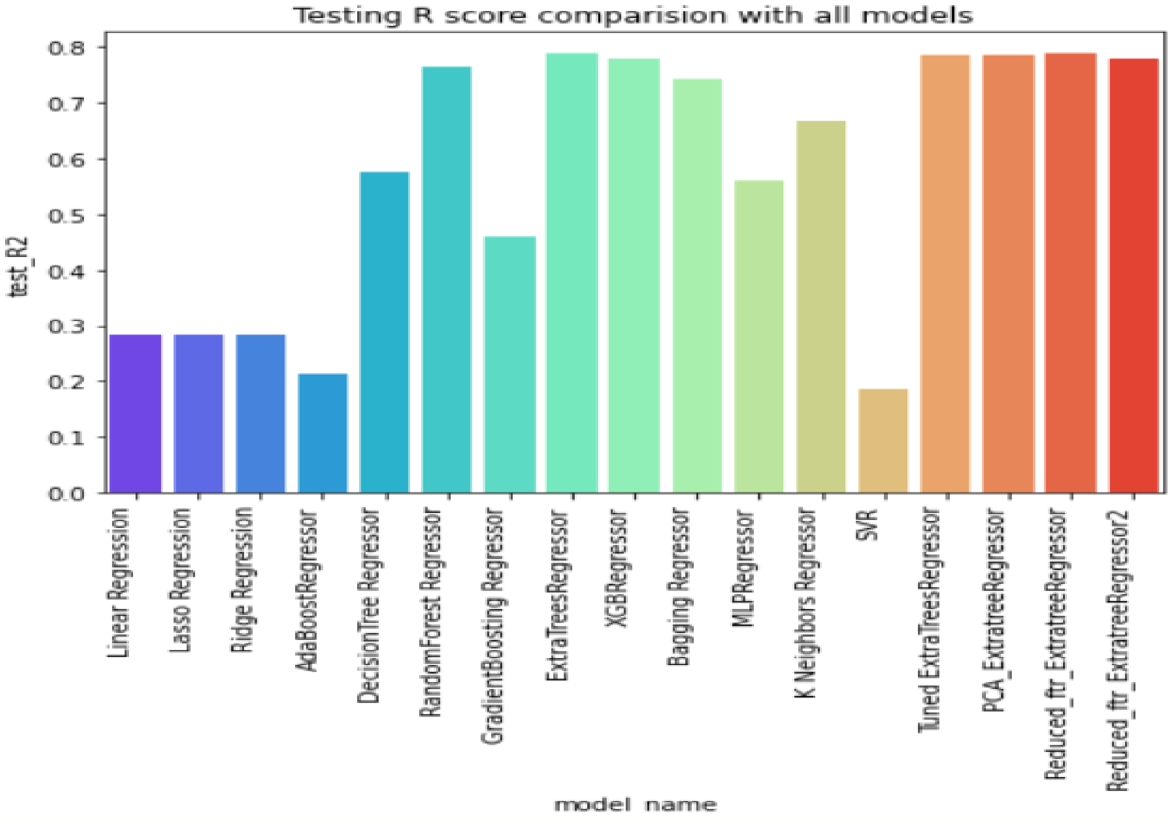


Fig:20 R2 Score Comparison of all ML Models

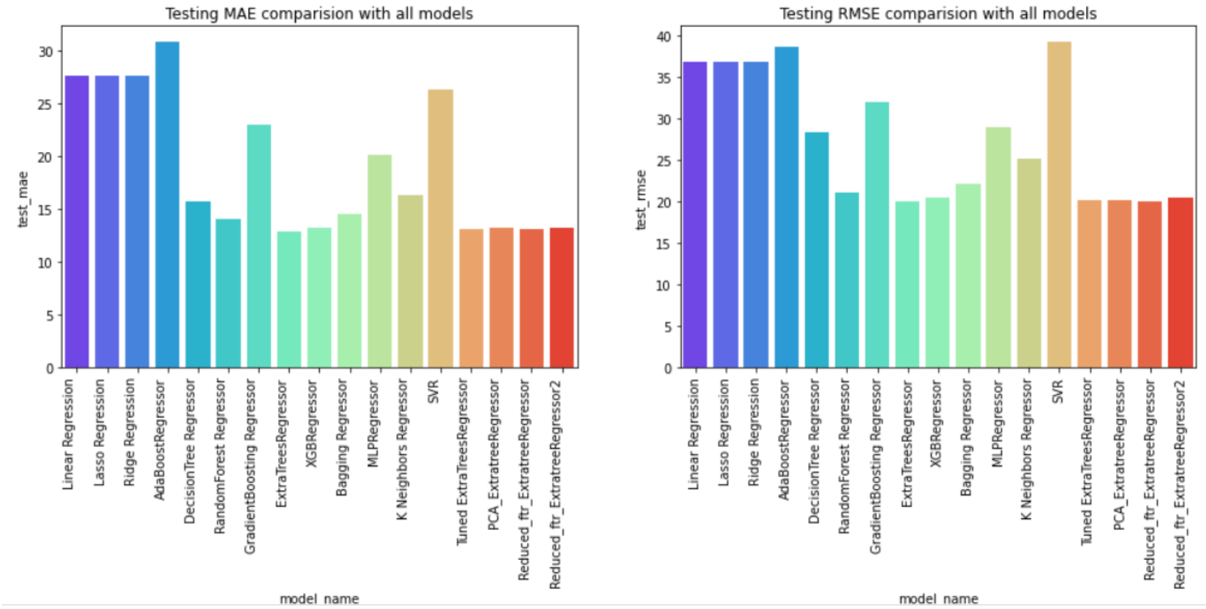


Fig:21 MAE and RMSE Error Comparison of all ML Models

RMSE and MAE should be less for best fitted model. As per above comparison Extra Tree Regression with feature reduction is best fitted model.

h. 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.

i. 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','RH_6','T7','RH_7','T8','RH_8',,'RH_9','T_out','Press_mm_hg','RH_out','Windspeed' ,'Tdewpoint' .

Figure below shows the comparison between random forest regressor model and optimal random forest regressor with feature reduction.

Fig: 22 Feature Importance Comparison

j. LIME

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.

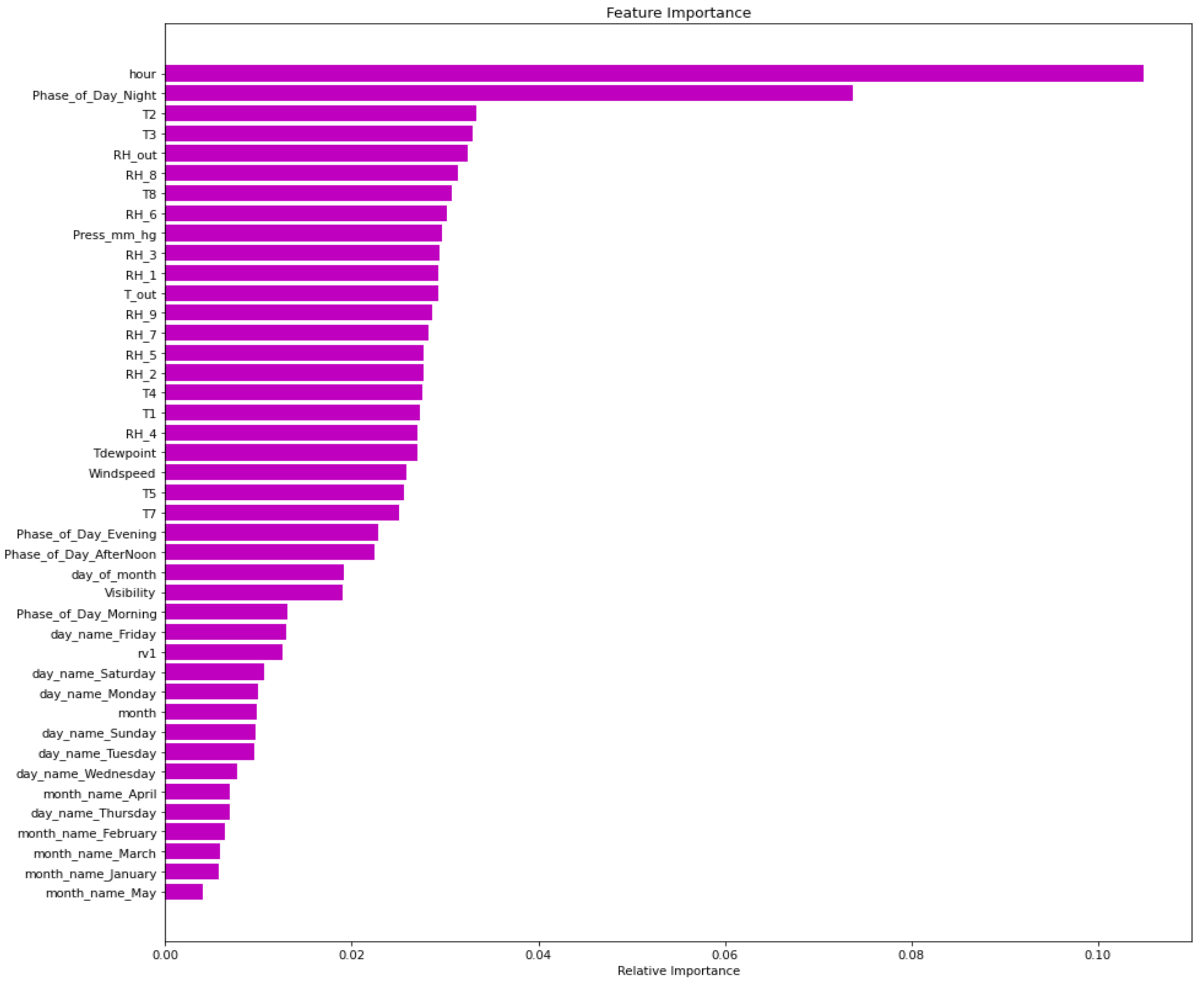


Fig: 22 Feature Importance

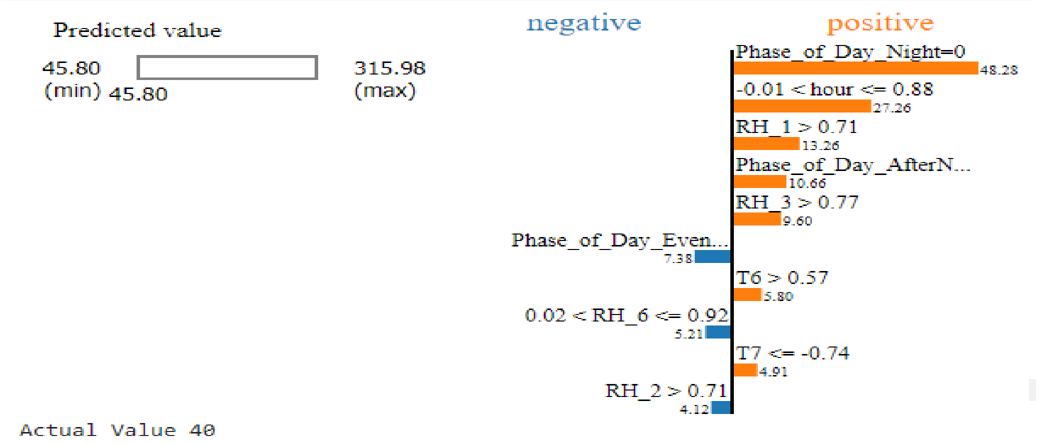
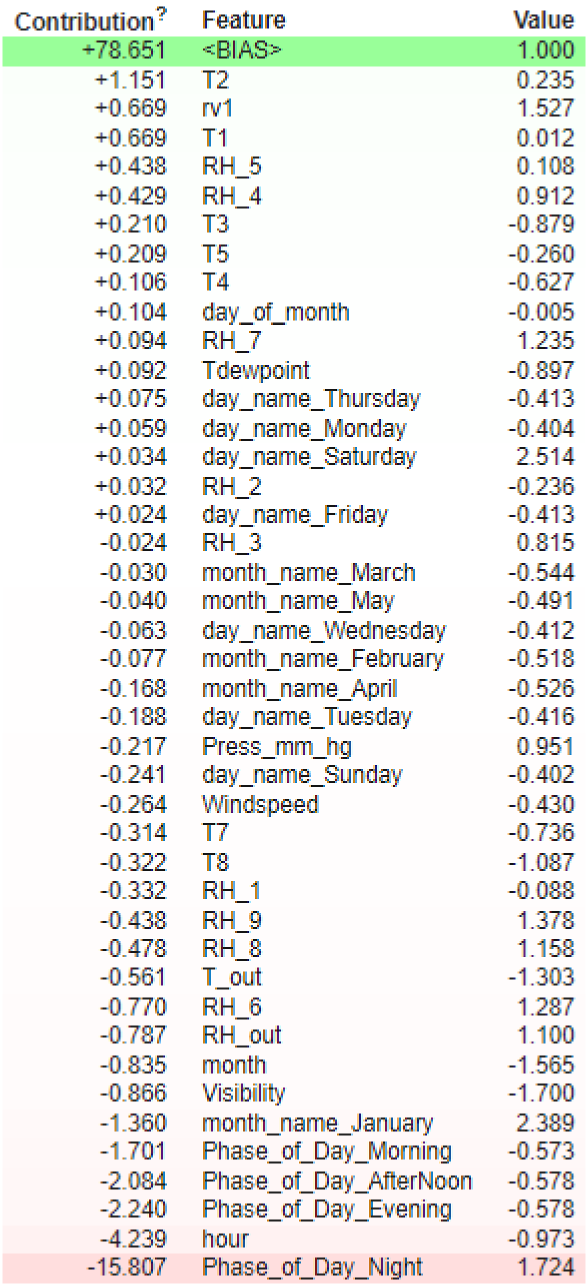


Fig: 23 LIME

The features Phase_of_Day_night, hour, RH_1 , RH_3, T6, T7 are contributing positively for predicting the Energy use. Actual value of this particular instance is 40 and predicted value is 45.80. LIME is used to explain Feature contribution in Black Box Models.

k. Eli5

ELI5 is a python library that gives the flexibility to debug and visualize different Machine Learning models using a single unified API. It provides in-built support for different Machine Learning frameworks and presents a unified way to explain black-box models.



Contribution?	Feature	Value
+78.651	<BIAS>	1.000
+1.151	T2	0.235
+0.669	rv1	1.527
+0.669	T1	0.012
+0.438	RH_5	0.108
+0.429	RH_4	0.912
+0.210	T3	-0.879
+0.209	T5	-0.260
+0.106	T4	-0.627
+0.104	day_of_month	-0.005
+0.094	RH_7	1.235
+0.092	Tdewpoint	-0.897
+0.075	day_name_Thursday	-0.413
+0.059	day_name_Monday	-0.404
+0.034	day_name_Saturday	2.514
+0.032	RH_2	-0.236
+0.024	day_name_Friday	-0.413
-0.024	RH_3	0.815
-0.030	month_name_March	-0.544
-0.040	month_name_May	-0.491
-0.063	day_name_Wednesday	-0.412
-0.077	month_name_February	-0.518
-0.168	month_name_April	-0.526
-0.188	day_name_Tuesday	-0.416
-0.217	Press_mm_hg	0.951
-0.241	day_name_Sunday	-0.402
-0.264	Windspeed	-0.430
-0.314	T7	-0.736
-0.322	T8	-1.087
-0.332	RH_1	-0.088
-0.438	RH_9	1.378
-0.478	RH_8	1.158
-0.561	T_out	-1.303
-0.770	RH_6	1.287
-0.787	RH_out	1.100
-0.835	month	-1.565
-0.866	Visibility	-1.700
-1.360	month_name_January	2.389
-1.701	Phase_of_Day_Morning	-0.573
-2.084	Phase_of_Day_AfterNoon	-0.578
-2.240	Phase_of_Day_Evening	-0.578
-4.239	hour	-0.973
-15.807	Phase_of_Day_Night	1.724

Fig: 24 Eli5

9. Limitations

One of the main limitations of this study is that the analysis was done for only one house. 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. 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. Regarding the weather station, the predictions of appliances energy use could probably be better if the weather station was closer to the house. 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.

10. Conclusion

1. The household appliance energy consumption prediction models based on Linear Regression, Lasso Regression, Ridge Regression, MLP Regressor, Decision Tree Regressor Random Forest Regressor, Adaptive Boosting Regressor, Gradient Boosting Regressor, Bagging Regressor, K Neighbours Regressor and Linear SVM are explored.
2. Upon appropriate pre-processing and fitting the fourteen models, we compare and evaluate the best model with lowest error and the highest R-squared score.
3. The variables T6 and T_out , T9 and T7 has high correlation with each other hence we have dropped T6 and T9. When evaluating the influence of Random Variable attribute the linear models have assigned near zero weights to the random variable, negating its influence in prediction of the target variable.
4. Extra Tree Regressor was found to be the best performing model with an R-squared score of 0.7877.
5. After optimizing the hyperparameters of the Extra Tree Regressor, doing principal Component Analysis, its R-squared score increased from 0.7837 to 0.7877.
6. We find that this model’s predictions are mainly contributed by the 'T1', 'T2', 'T3', 'T4', 'T5','RH_6', 'T7', 'T8', 'RH_8','RH_9', 'T_out', 'RH_out', 'hour', 'Phase_of_Day_AfterNoon','Phase_of_Day_Evening','Phase_of_Day_Night','month_name_January' 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.
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.
9. 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.
10. As this dataset has a time component to it, we believe that better performances can be achieved by using Time Series Analysis concepts.

11. **Future Scope and Suggestions**

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. We Could try NN regressor to analyze the pattern of the data for prediction. Dataset should have the More features to interpret the appliances energy consumption like list of appliances with per hour watt energy consumption.

12.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