**Appliances Energy Prediction**

**Tanmay Bohra, Vikram Pratap**

**Amrutha B.S, Mahima Shree, Purnima Rai, Pooja Rana**

**Data science trainees,**

**AlmaBetter, Bangalore**

### Abstract

### In this time of global unreliability world needs energy to support economic and social progress and build a better quality of life, in particular in developing countries.

### But even in today’s time there are many places particularly developing countries where there are interruption of services. These fallouts are primarily because of excess load consumed by appliances at home. Heating and cooling appliances takes most power in house. In this project we will be analyzing the appliance usage in the house gathered via sensor network. All readings are taken at 10 minutes intervals for 4.5 months. The goal is to predict energy consumption by appliances. It is important to study the energy consumption in the residential sector and predict the energy consumption by home appliances as it consume maximum amount of energy in the residence.

In the era 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 this case Linear Regression, Support Vector Machine, KNN Regression, Ensemble Models will be used to predict Appliance energy usage based on data collected from various sensors.

**Key Words: KNN Regression, Support Vector Machine, Linear Regression, Ensemble Models.**

### 1. Problem Statement:

### We are given a dataset having 19735 rows and 29 columns. A lot of work has been carried out to predict appliance energy usage with the help of this dataset. Different levels of accuracy have been attained using various algorithms which are explained as follows.

### We study various different Machine Learning algorithms that can be used for prediction of energy usage. Research was carried out to study Ensemble methods, KNN Regression, Support Vector Machine, Linear Regression that can be used for prediction and their accuracy were compared. This research concludes that accuracy obtained by Extra Tree Regression was highest, further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

### 2. Introduction

### The work proposed in this paper focus mainly on various data mining practices that are used to predict the appliance energy consumption. The increasing trend in energy consumption is becoming cause of concern for the entire world, as the energy consumption is increasing year after year so is the carbon and greenhouse gas emission, the majority portion of the electricity generated is consumed by industrial sector but a considerable amount is also consumed by residential sector. It is important to study the energy consuming behavior in the residential sector and predict the energy consumption by home appliances as it consume maximum amount of energy in the residence. This project focuses on predicting the energy consumption of home appliances based on humidity and temperature information obtained through sensor network. It has resulted in implementation of five prediction regression models, i.e. multiple regression, lasso regression, ridge regression, SVM regression and Random Forest are developed and results are presented. The dataset for the analysis was taken from a house located in Belgium. 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 regression analysis will be used to predict Appliance energy usage based on data collected from various sensors.

### We should predict Appliance energy consumption for a house based on factors like temperature, humidity & pressure. In order to achieve this, we need to develop a supervised learning model using regression algorithms. Regression algorithms are used as data consist of continuous features and there is no identification of appliances in dataset. The dataset was collected by sensors placed inside the house and outside readings came from the nearby weather station. The main attributes are temperature, humidity and pressure readings. Each observation measures electricity in a 10-minute interval. The temperatures and humidity have been averaged for 10-minute intervals.

**2.1 Application Energy Dataset**

The data set is obtained by observing data from a sensor network at every10 min for about 4.5 months. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set and merged together with the experimental data sets using the date and time column.. The goal is to predict the appliance energy consumption based on conditions of home and outside. It includes over 19735 records and 29 attributes.

Variables – Each attribute is a potential risk factor.

1. Date
2. Target Variable – Appliances, energy use in Watthour.
3. lights, energy use of light fixtures in the house in Wh (Drop this column).
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 %To, Temperature outside (from Chievres weather station), in Celsius Pressure (from Chievres weather station), in mm Hg RHout, Humidity outside (from Chievres weather station), in %.
22. Wind speed (from Chievres weather station), in m/s.
23. Visibility (from Chievres weather station), in km.
24. Tdewpoint (from Chievres weather station), Â°C.
25. rv1, Random variable 1, non-dimensional.
26. rv2, Random variable 2, non-dimensional.

**2.2 Python**

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programming language to select up compared to other language. That’s the most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is straightforward to use. That’s one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open source library is named panda.

Here is the first step to clean the data that will make the results “more” accurate. By finding all unique values of each row the inappropriate values can be identified. Different methods can then be used for removing them or to change those values accordingly to use them to make predictions better. As the proverb goes by saying –

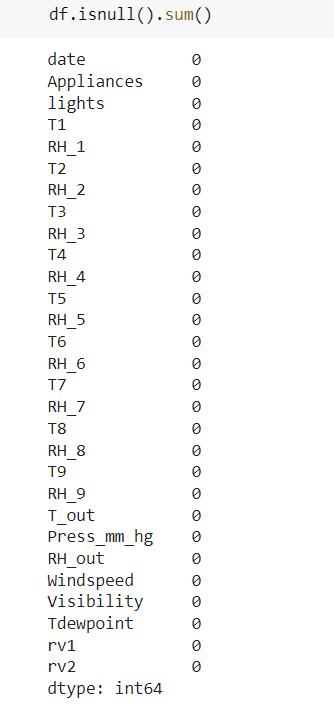
*“The more data we have, the more likely we are to drown in it.” — Nassim Taleb*

Not only are we interested in raw data but in the data from which valuable insights can be drawn. To do so, let us take a glimpse at another proverb.

*“More data beats clever algorithms, but better data beats more data.” — Peter Norvig*

After loading the dataset, we can start the exploration but before that, we need to check and see that the dataset is ready for performing several exploration operations or not.

To know if there is any missing value or Nan value in the dataset, we can use the isnull () function.



So, we will need to prepare the dataset before performing exploratory data analysis on it.

**2.4 Data Preparation and Cleaning**

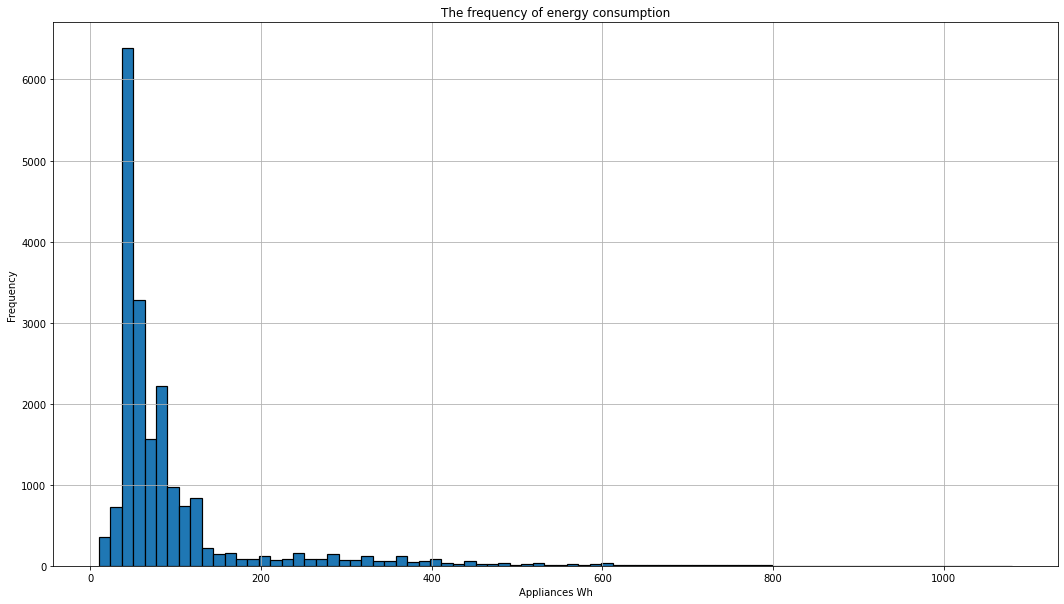
Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and the combining of data sets to enrich data. Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

● A regression problem is when the output variable is a real or continuous value. Regression problems require consideration of more than one predictor, and it is required to understand how the response *y* depends simultaneously on the predictors x1, x2,…,xp.

● The Appliances energy prediction dataset contains 29 features & 19,735 instances.

● The Appliances energy prediction dataset contains features with respect to temperatures inside & outside, humidity inside & outside, wind speed, visibility, Tdewpoint.

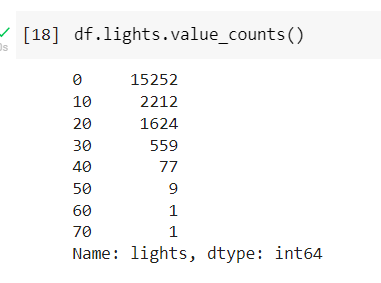
● The Appliances energy prediction dataset contains no Null or missing values. The dataset does not contain any duplicate values as well.



Energy consumption of appliances ranges from 10 Wh to 1080 Wh, about 75 % of energy values lie below 100 Wh, and about 93 % of them lie below 300 Wh.

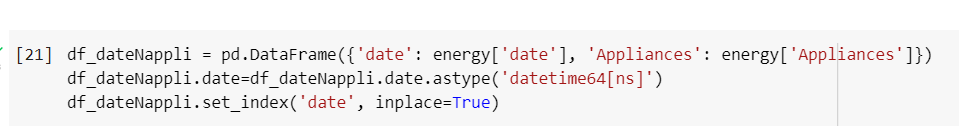
Our target variable seems to be highly skewed, and our task is to predict the usual as well as the large surges in energy in the building**.**

The column ‘lights’ has been dropped as maximum value in lights attribute is 0, it won’t be playing much role in our model. Hence we are dropping the lights attribute from our dataset.



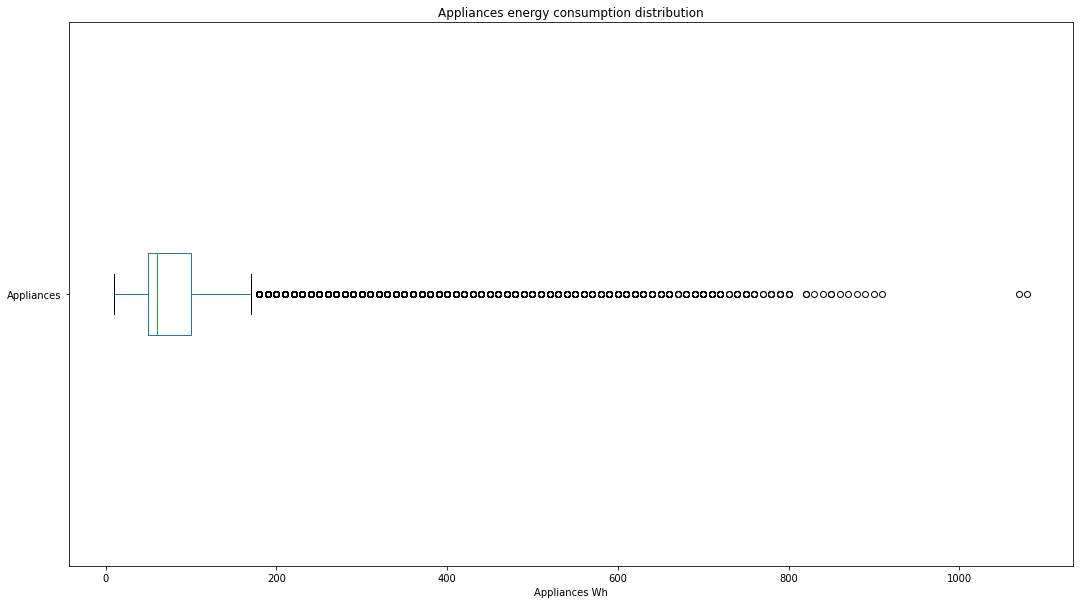
We also formed a new DataFrame “df\_dateNappli” with right attributes to plot graphs.

We took “date” and “Appliances” attributes into the new data frame. When we created the new data frame, date attribute's type got changed to objects. Then we converted it back to datetime64[ns] type. This set the new data frame’s index as date.



# ****3. Outlier Detection****

After observing the dataset we checked for outliers and found out that Appliances energy prediction has some high value but we did not considered them as outlier because those high values were of importance.

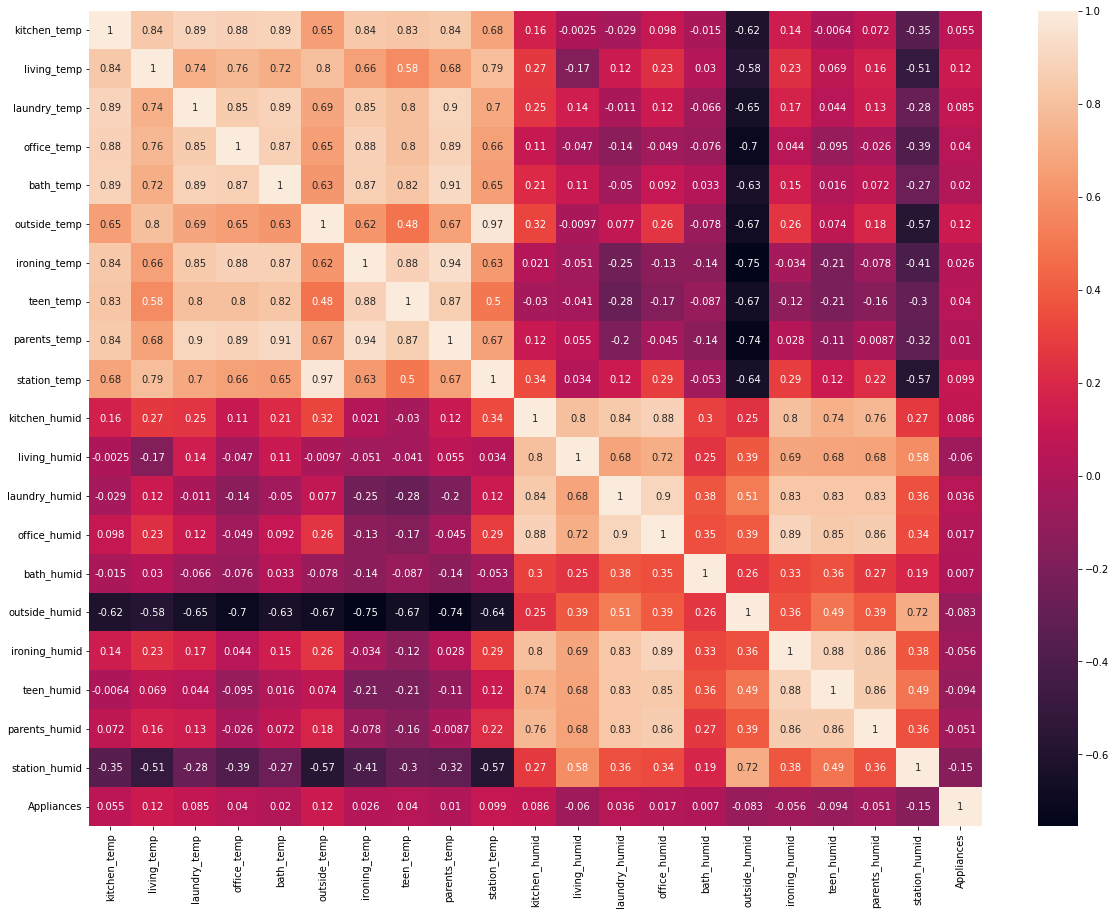


# ****4. Feature Engineering****

The features in your data will directly influence the predictive models we use and the results we can achieve. We can say that: the better the features that we prepare and choose, the better the results we will achieve. It is true, but it is also misleading. The results we achieve are a factor of the model we choose, the data we have available and the features we prepared. Even framing of the problem and objective measures we’re using to estimate accuracy play a part. We need great features that describe the structures inherent in our data.

● The column ‘lights’ has been dropped as maximum value in lights attribute is 0, it won’t be playing much role in our model. Hence we are dropping the lights attribute from our dataset.

● Plotted the correlation matrix for temperature and humidity levels.

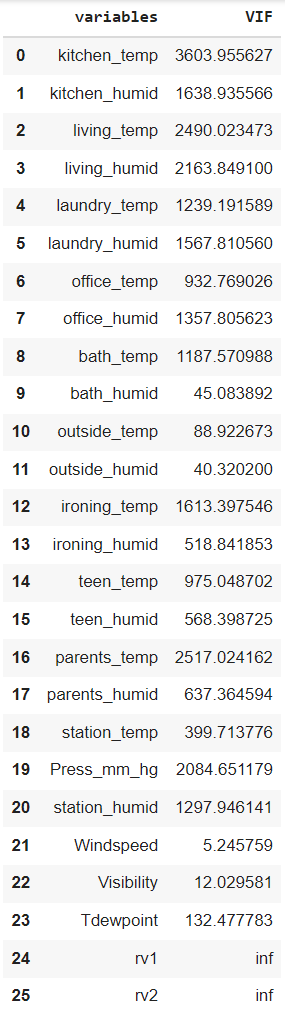


The below correlation shows

1. From the correlation graph we clearly observe that the features related to temperature and features related to humidity have positive correlation within themselves whereas have a very low or negative correlation with each other.
2. Humidity outside have a strong negative correlation with temperature levels.
3. Apart from that we observe that a couple features such as humidity at station, temperature outside the building and temperature in the living room have a comparatively high absolute correlation (above 0.12) with Appliances energy consumption.

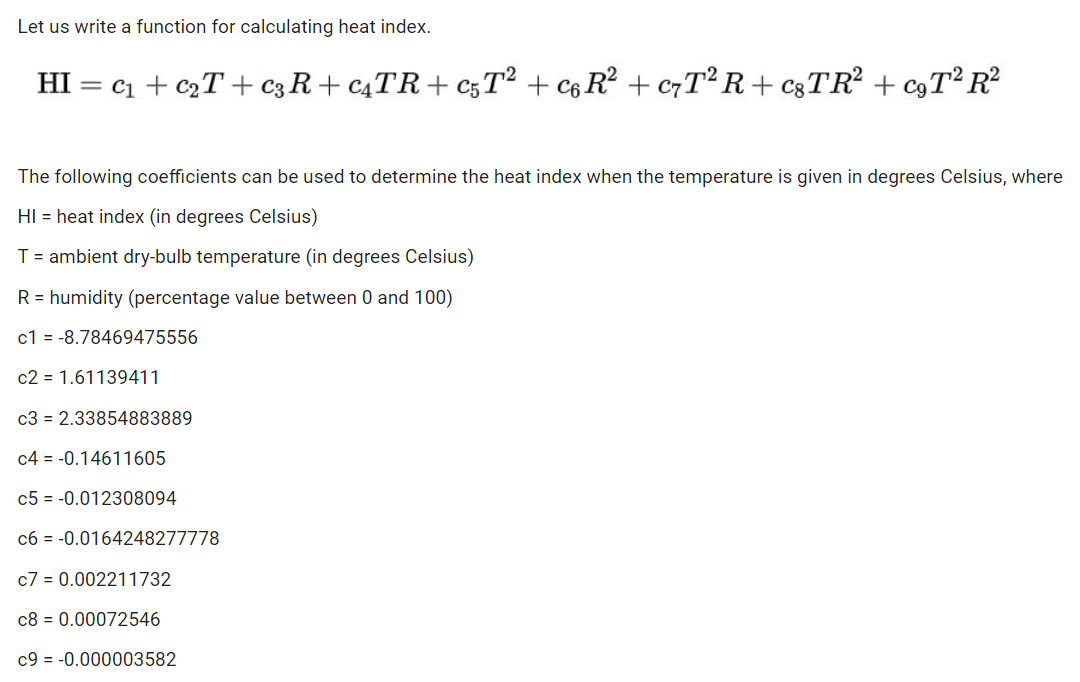
**4.1 Reducing Multi-collinearity**

We worked upon reducing the multi-collinearity using Variation Inflation Factor (VIF) and got the below table.



There is very high multi-collinearity among the independent variables as seen.

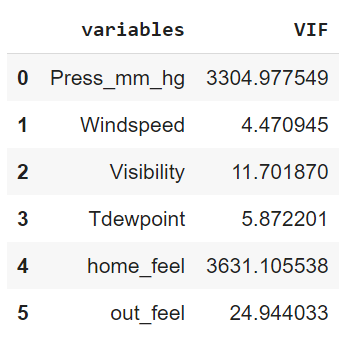
We tried on working to eliminate the temperature and humidity by converting them into heat index given by the below formula.



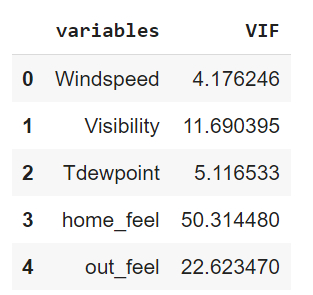
After performing this, we then checked the VIF of the new variables and removed the rv1 & rv2 due to infinite variance.



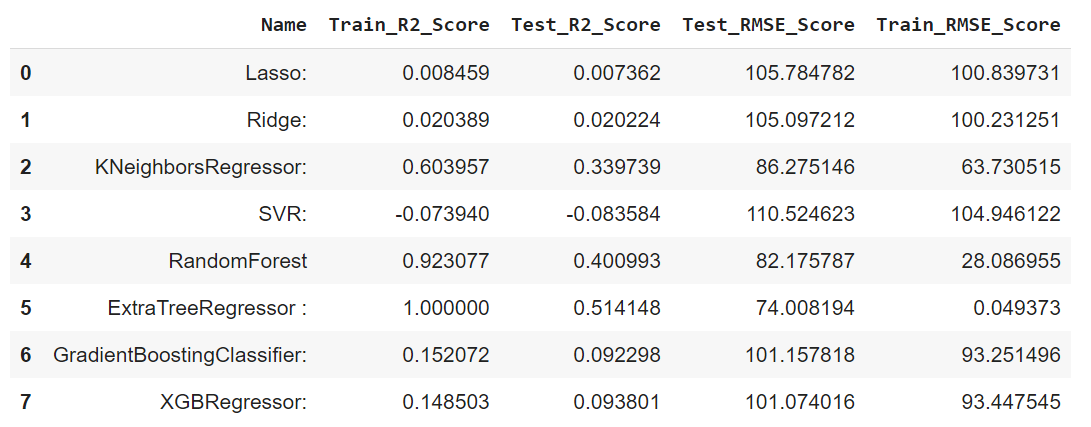
We can still see high multi-collinearity. Then we worked on reducing the features to inside and outside by taking an average of the inside feel and the outside feel and then eliminating variables used to calculate the new features.

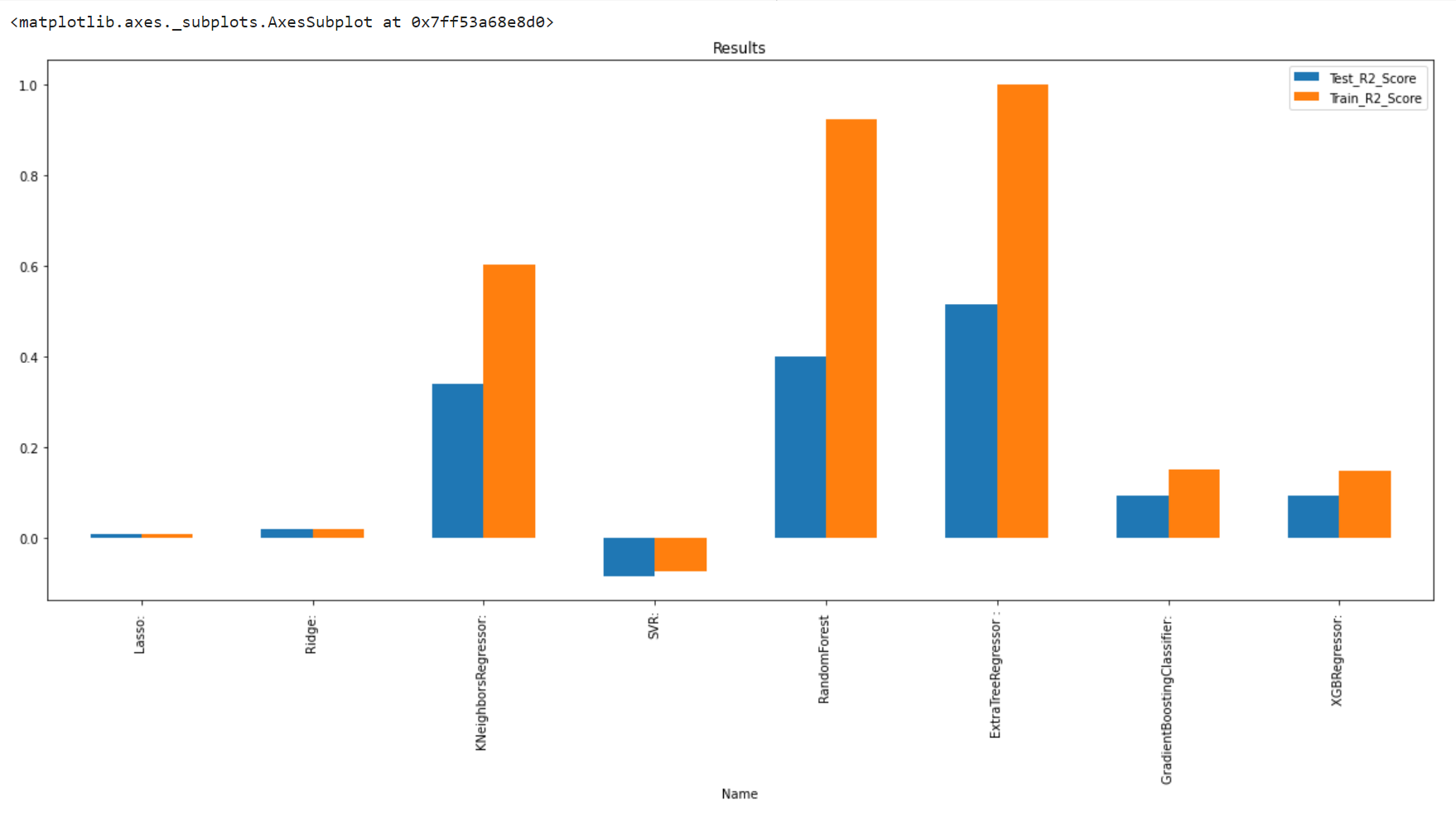


We still see a high multi-collinearity among press\_mm\_hg, home\_feel & out\_feel. We then decided to drop press\_mm\_hg and got the below VIF among the available features.



These features were then worked upon and then we got the following results:





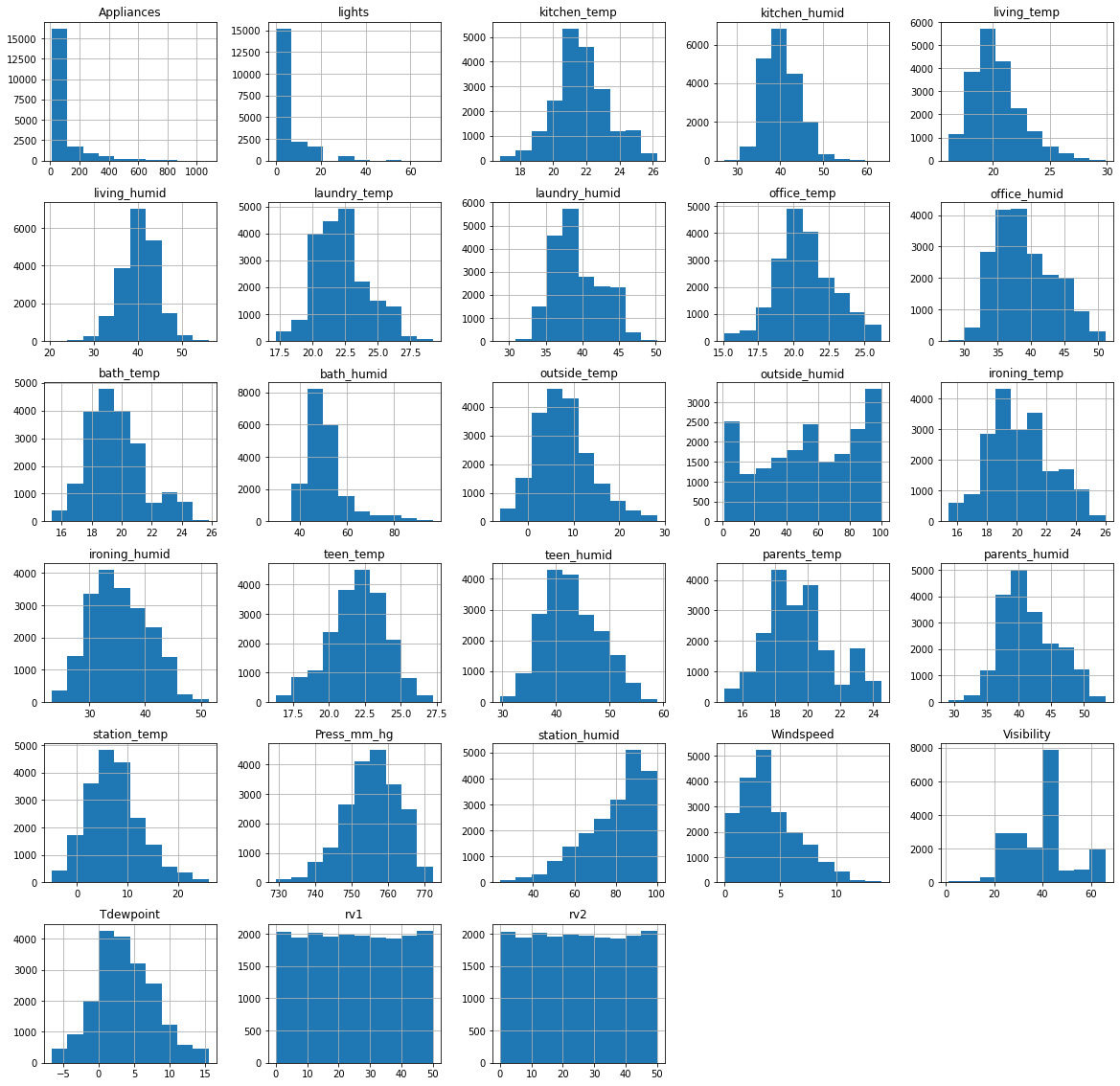
The below were observed:

1. We can see an improvement in the Lasso regression, obvious due to removal of several features.
2. Ridge underperformed by 15% when compared with its performance by using all features.
3. K-Nearest underperformed by 31% when compared with its performance by using all features.
4. SVR has performed the worst.
5. Random Forest underperformed by 31% when compared with its performance by using all features, and the hyper-parameter tuned model.

**5. Feature Distribution and Selection**

The feature distribution helps in understanding what kind of feature we are dealing with, and what values you can expect this feature to have. We’ll see if the values are centered or scattered. This distribution is important because models learn from data give those incorrect data and they will learn incorrectly. A model is only as good as the data we feed it. If a feature can be seen as a random variable, and enough data is used and the bins are narrow enough, the look of the distribution could be bell-shaped. This is called a normal distribution (or standard/Gaussian). The normal distribution is a very important probability distribution that arises in many situations. Generally speaking, when you have a large number of independent samples from a naturally occurring phenomenon, the data will follow a normal distribution.

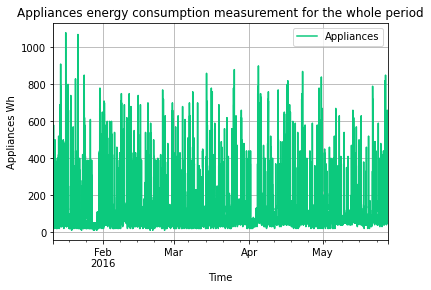
The data distribution plots plotted below gives us the following inferences



1. Temperature columns T1 (Temperature in kitchen area), T2 (Temperature in living room area), T3 (Temperature in laundry room area), T4 (Temperature in office room), T5 (Temperature in bathroom), T6 (Temperature outside the building), T7 (Temperature in ironing room), T8 (Temperature in teenager room) follow normal distribution.
2. Temperature column T9 (Temperature in parent’s room) does not follow normal distribution.
3. Humidity columns follow normal distribution RH\_1 (Humidity in kitchen area), RH\_2 (Humidity in living room area) RH\_3 (Humidity in laundry room area), RH\_4 (Humidity in office room), RH\_5 (Humidity in bathroom) RH\_7 (Humidity in ironing room), RH\_8 (Temperature in teenager room) follow normal distribution.
4. Humidity column RH\_6 (Humidity outside the building) and RH\_out (Humidity outside)does not follow normal distribution, primarily because these sensors are outside the house.
5. Visibility column is negatively skewed.
6. Wind speed column is positively skewed.

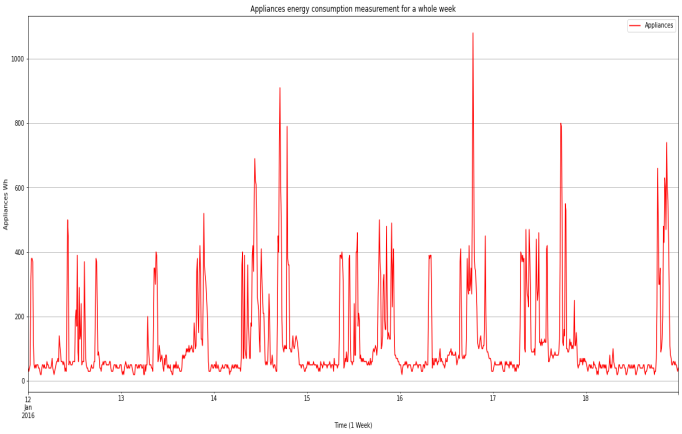
**6. EDA on Features**

* 1. **Appliance energy consumption measurement for whole period**



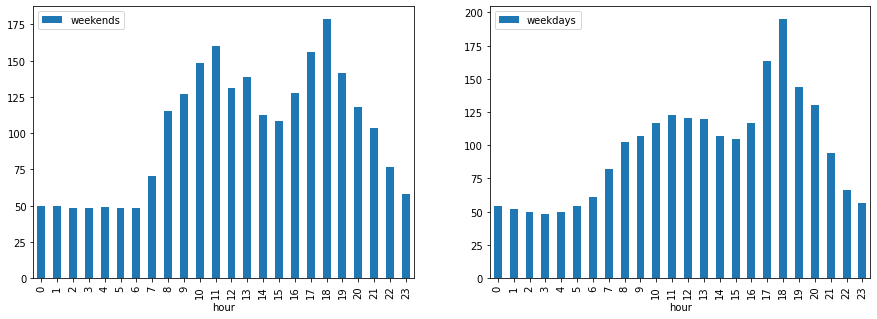
The above plot shows the energy consumption profile for the period. The energy consumption profile shows a high variability. We can see from here that the highest Appliances Wh is around 1100 and it was in January month. Also, we can see that, at the end of January, February and March, there were big fall down in Appliances usage.

* 1. **Appliances energy consumption measurement for a whole week.**

****

Frome the above **Energy Consumption pattern of the household at different times of the day over different time periods. This** is a representation of average energy consumption of appliances at different time of the day over a period of 4.5 months. We observe two peak hours. One at 11 am in the morning and other at 6 PM in the evening. While the peak at 11 am is shallow and low, peak at 6 PM is comparatively higher and sharper. We observe that over the sleeping hours (10 PM - 6 AM) the energy consumption of appliances is around 50 Wh. After about 6 AM, energy consumption starts to rise gradually up until 11 AM (probably due to morning chores). And then gradually decreases to around 100 Wh at about 3 PM. After which the energy consumption drastically shoots up up until 6 PM in the evening (probably due to requirement lights in rooms). However energy consumption of appliances reverts back to 50 Wh, as night approaches and people in the house go to bed at around 10 PM.

* 1. **What about weekdays and weekends?**

****

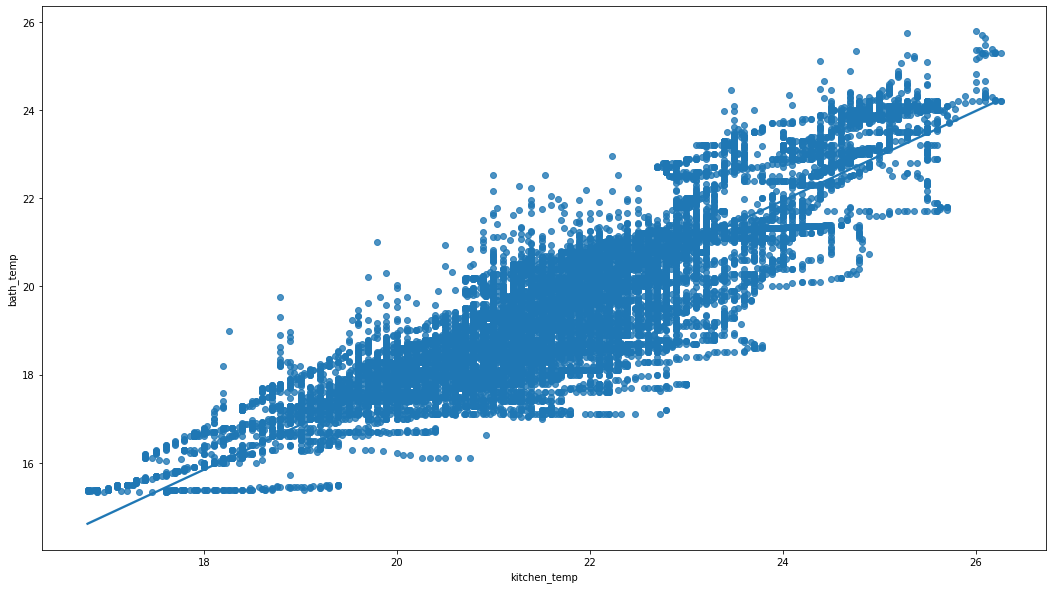
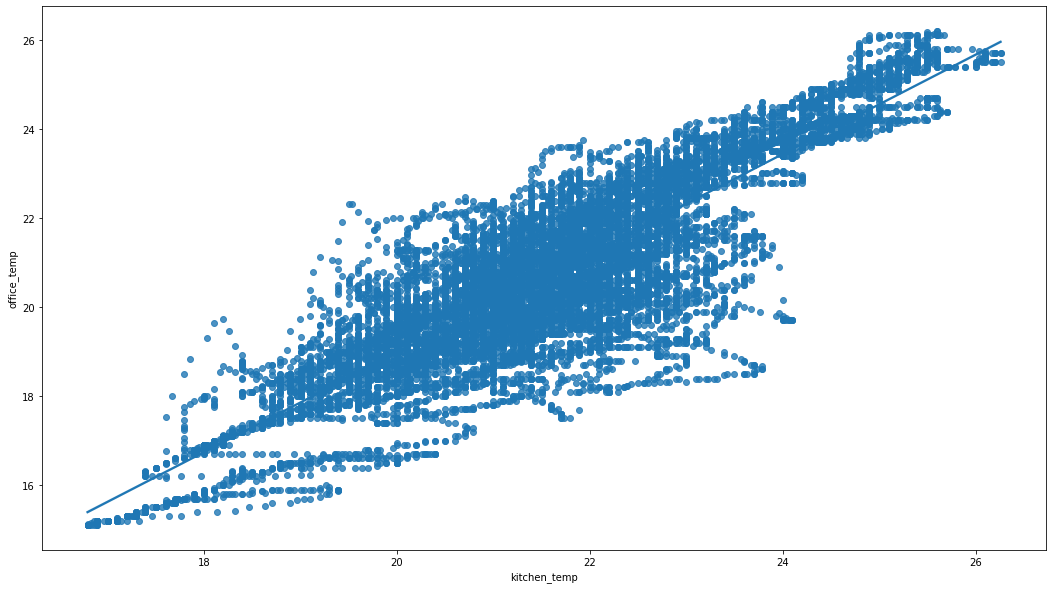
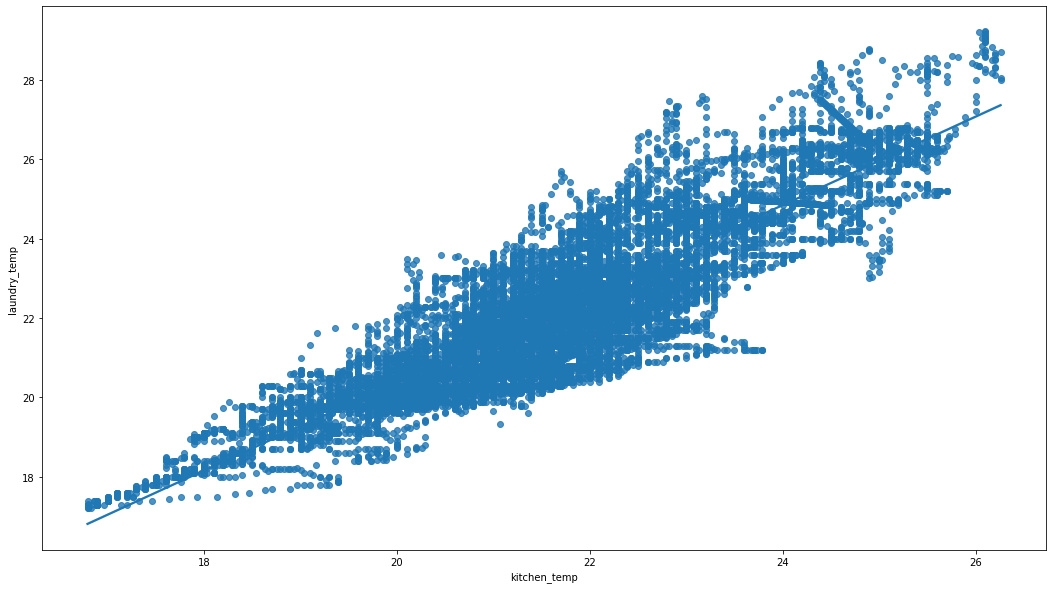
While observing energy consumption in weekdays and weekends We observe that the energy consumption of appliances during the 8 AM - 4 PM is higher in weekends compared to the weekdays. Also, average overall consumption in weekends is pretty high.

We also observed that-

* Outside Average temperature over a period of 4.5 months is around 7.5 degrees and ranges from -6(min) to 28(max) degrees.
* Inside the building average temperature has been around 20 degrees for all the rooms and ranges from 14(min) to 30(max) degrees.

**Note**: These points imply that warming appliances have been used to keep the insides of the building warm. There must be some sort of direct correlation b/w temperature and consumption of energy inside the house.

1. Visualization on Temperature data



**7. Machine Learning**

**Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed. As it is evident from the name, it gives the computer that makes it more similar to humans that is  the ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect. Machine Learning AI is an upward innovation which permits PCs to gain naturally from past information.

Assume we have a diverse issue, where we need to accomplish a few forecasts, so as opposed to composing a code for it, we simply need to provide information to generic algorithms, with the assistance for these calculations, machine assembles rationale according to information, foresee the yield. AI has changed our perspective about this issue.

**7.1 Classification of Machine Learning**

Supervised Learning is regularly characterized as learning with correct regulator; else you can say that learning inside the presence of educators. The algorithm learns on a labeled dataset with an answer key and does the training and evaluation. Administered learning is anticipated on "train me" idea. Supervised learning has next measures:

• Classification

• Random forest

• Decision tree

• Regression

There are following machine AI algorithms:

• Linear Regression

• Logistical Regression

• Support Vector Machines (SVM)

• Neural Networks

• Random Forest

• Gradient Boosted Trees

• Decision Trees

• Naive Bayes

From the above algorithm we trained our model using:

1. Linear Regression
   1. Lasso Regression
   2. Ridge Regression
2. K Neighbor Regression
3. Support Vector Machine
4. Random Forest Regression
5. Extra Tree Regression
6. Gradient Boosting Regression

**7.1.1 Linear Regression**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression

**Lasso Regression**

Lasso regression is a type of [linear regression](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/) that uses [shrinkage](https://www.statisticshowto.com/shrinkage-estimator/). Shrinkage is where data values are shrunk towards a central point, like the [mean](https://www.statisticshowto.com/mean/). Thelasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of [muticollinearity](https://www.statisticshowto.com/multicollinearity/) or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

**Ridge Regression**

Ridge [regression](https://www.mygreatlearning.com/blog/what-is-regression/) is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

**7.1.2 K- Neighbor Regression**

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood. The size of the neighborhood can be chosen using cross-validation select the size that minimizes the mean-squared error.

While the method is quite good but it becomes erroneous when there are many independent variables.

**7.1.3 Ensemble Models**

**Random Forest Regression**

Random Forest Regression is a supervised learning algorithm that uses ensemble **learning** method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, over fitting may easily occur, we must choose the number of trees to include in the model.

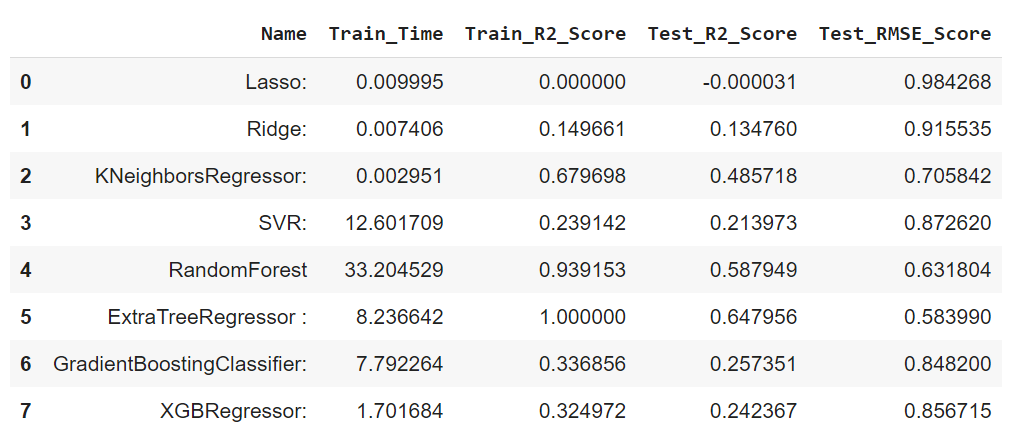
**Gradient Boosting Regression**

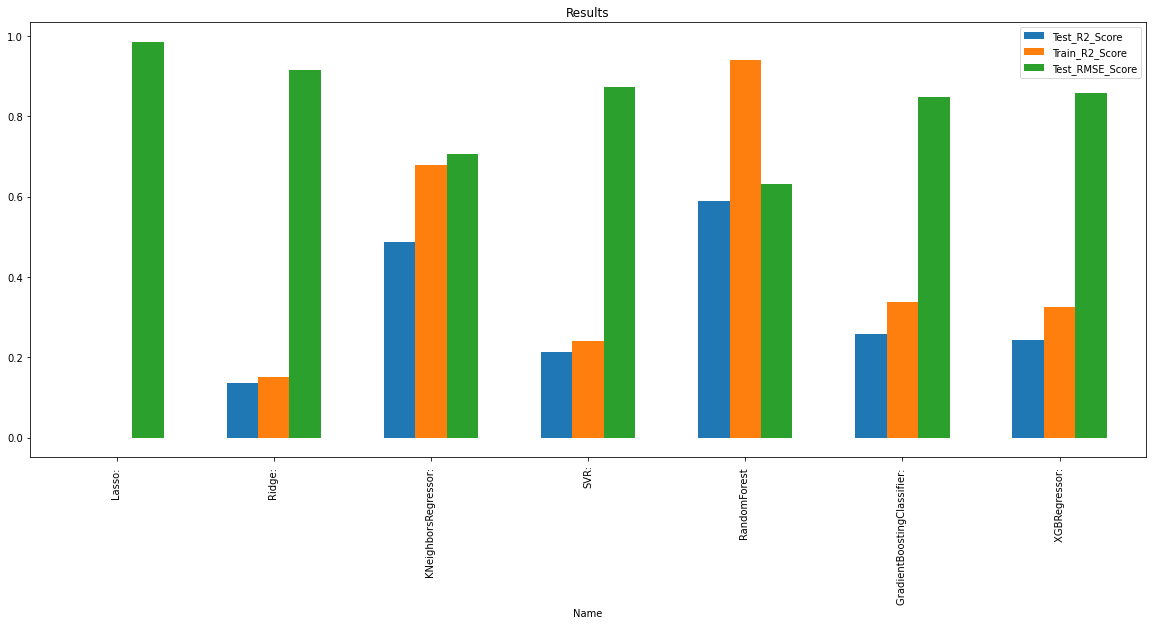
Gradient boosting is a Machine learning technique. It gives a prediction model in the form of an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning) of weak prediction models, which are typically [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning). When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms [random forest](https://en.wikipedia.org/wiki/Random_forest). A gradient-boosted trees model is built in a stage-wise fashion as compared to other [boosting](https://en.wikipedia.org/wiki/Boosting_(machine_learning)) methods, but it generalizes the other methods by allowing optimization of an arbitrary [differentiable](https://en.wikipedia.org/wiki/Differentiable_function) [loss function](https://en.wikipedia.org/wiki/Loss_function).

**7.1.4 Support Vector Regression**

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points. Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value. The threshold value is the distance between the hyperplane and boundary line.

**8. Comparison of the performance of all the Models**

****



Observation

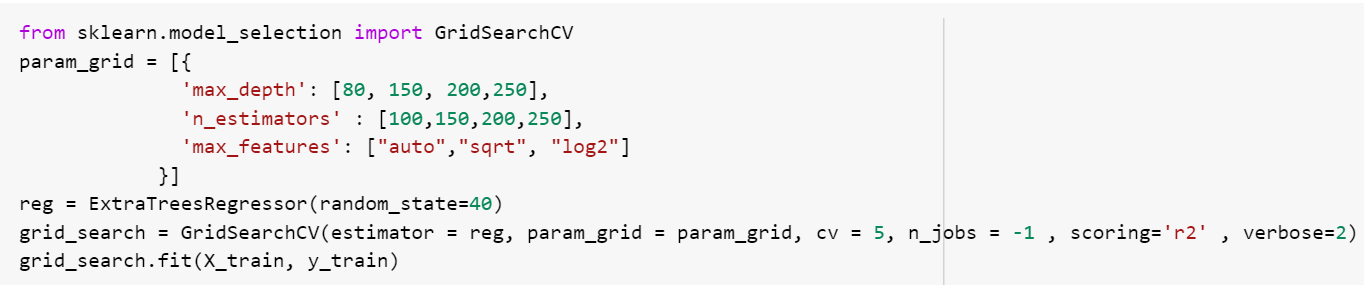
1. Best results over test set are given by Extra Tree Regression with R2 score of 0.64
2. Least RMSE score is also by Random Forest 0.58.
3. Lasso regularization over Linear regression was worst performing model.

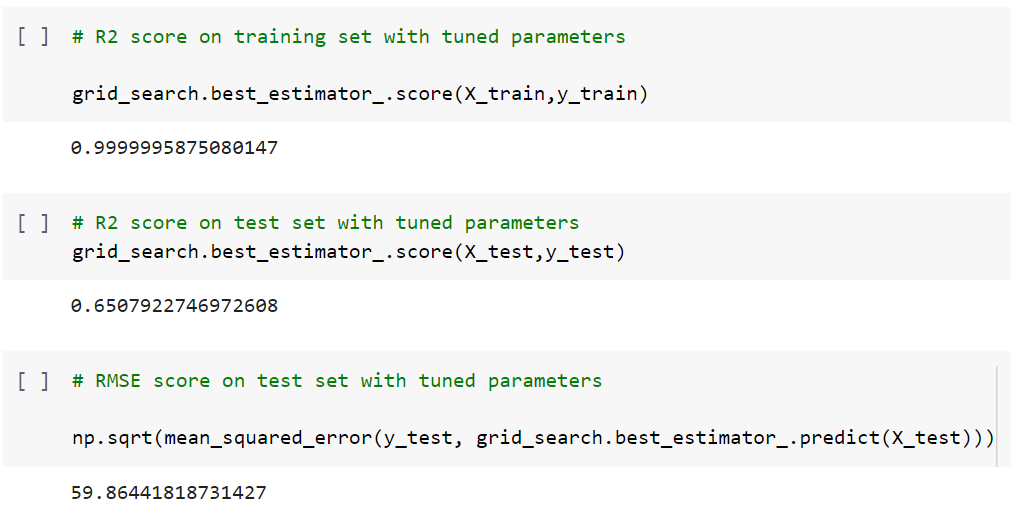
As we can see Extra Tree Regression it is showing R2 score for train set is 100% but for test it is 64% which shows that our model is overfitting on training data so to overcome this we did hyper parameter tuning for Extra Tree Regression.

**9. Hyper-Parameter Tuning**

Whenever a machine learning algorithm is implemented on a specific dataset, the performance is judged based on how well it generalizes i.e. how it reacts to new, never-before-seen data. In case the performance of the learning algorithm is not satisfactory or there is room for improvement, certain parameters in the algorithm need to be changed/tuned/tweaked. These parameters are known as ‘hyperparameters’ and the process of varying these hyperparameters to better the learning algorithm’s performance is known as ‘hyperparameter tuning’.

These hyperparameters are not learnt directly through the training of algorithms. These values are fixed before the training of the data begins. They deal with parameters such as learning rate, i.e how quickly the model should be able to learn, how complicated the model is, and so on There can be a wide variety of hyperparameters for every learning algorithm. Selecting the right set of hyperparameters so as to gain good performance is an important aspect of machine learning.

****

****

Based on parameter tuning step we can see that the best possible parameter combination are - 'max\_depth': 80, 'max\_features': 'sqrt', 'n\_estimators': 200

Training set R2 score of 1.0 may be signal of overfitting on training set

Test set R2 score is 0.63 improvements over 0.57 achieved using untuned model

Test set RMSE score is 0.60 improvements over 0.65 achieved using untuned model

**Conclusion:**

The goal of this research was to find out whether we can predict energy consumption of home appliances based on humidity and temperature which can got from sensor network in Belgium, and also to predict the energy consuming behaviour of people effectively and accurately. This study compares the performance of various algorithms like Lasso Regression, Ridge Regression, K-Neighbour Regressor, Support Vector Regression, Random Forest Regression, Gradient Boosting Regression for predicting appliance energy usage with the help of given dataset. The result of this study indicates that the Extra Tree Regression algorithm is the most efficient algorithm for prediction of energy usage. Some more inferences that we can draw are as follows. The time zone of the day plays an important role in deciding power consumption of appliances. After studying average energy consumption of appliances at different time of the day over a period of 4.5 months. We observe two peak hours one at 11 am in the morning and other at 6 PM in the evening. While the peak at 11 am is shallow and low, peak at 6 PM is comparatively higher and sharper.

We observe that over the sleeping hours (10 PM - 6 AM) the energy consumption of appliances is around 50 Wh. After about 6 AM, energy consumption starts to rise gradually up until 11 AM (probably due to morning chores). And then gradually decreases to around 100 Wh at about 3 PM. After which the energy consumption drastically shoots up until 6 PM in the evening (probably due to requirement lights in rooms). However energy consumption of appliances reverts back to 50 Wh, as night approaches and people in the house go to bed at around 10 PM.

The highest Appliances Wh is around 1100 and it was on 16th of January. Also, we can see that, it follows a bit of a pattern though it is not a strong one. In comparison between weekends and weekdays we noticed that the energy consumption of appliances during the 8 AM - 4 PM is higher in weekends compared to the weekdays. Also, average overall consumption in weekends is pretty high.

**Future Scope:**

Future work includes adding more predictive parameters such as number of family members, the area of the house, day to day activities performed by the family members, more indoor and outdoor activities performed by the family members, by using different datasets can also improve the prediction. Training other machine models such as ANN (Artificial Neural Network) can further boost the predictive capacity; energy consumption is a vast domain and has lot of reach in future also we can work on the day/week feature to explore more on the model performance, try various hyper-parameter tuning methods.

**References:**

AlmaBetter Tutorials

<https://www.python.org/>

<https://pandas.pydata.org/>

<https://matplotlib.org/>

<https://numpy.org/>

<https://www.w3schools.com/python/pandas/default.asp>