|  |
| --- |
| **APPLIANCES ENERGY PREDICTION**  **Manoj Patil M**  **Gulzar**  **Bindu Kovvada**  **Saksham Tripathi  Deepak Kumar Gautam**  **Data science trainees,**  **Alma Better, Bangalore** |

**Abstract:**

Energy consumption is quickly growing in today's globe. We are experiencing a scarcity of energy as a result of increased energy use in some regions of the world, which is causing environmental damage. We have observed excessive energy consumption in home appliances in some locations, so the main goal of this project is to analyses what factors are influencing the increasing energy consumption of home appliances, how we can reduce energy consumption of home appliances, and predict energy consumption of appliances using regression models.

The energy prediction challenge will fall under supervised machine learning, with the goal of predicting appliance energy usage for a house based on characteristics such as temperature, humidity, and pressure. Many approaches, such as the linear regression and many other algorithms have been used to predict the energy consumption.

**1.Problem Statement**

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).

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.

**1.1 Data Set**

Below is the info that is available in given dataset-

lights - Energy use of light fixtures in the house

* lights - Energy use of light fixtures in the house
* T1 - Temperature in kitchen area
* RH\_1 - Humidity in kitchen area.
* T2- Temperature in living room area.
* RH\_2 - Humidity in living room area
* T3 - Temperature in laundry room area
* RH\_3 - Humidity in laundry room area
* T4 - Temperature in office room
* RH\_4 - Humidity in office room
* T5 - Temperature in bathroom
* RH\_5 - Humidity in bathroom
* T6 - Temperature outside the building
* RH\_6 - Humidity outside the building
* T7 - Temperature in ironing room
* RH\_7 - Humidity in ironing room
* T8 - Temperature in teenager room 2
* RH\_8 - Humidity in teenager room 2
* T9 - Temperature in parents’ room
* RH\_9 - Humidity in parents room
* T\_out - Temperature outside (from Chievres weather station)
* Press\_mm\_hg - Pressure (from Chievres weather station)
* RH\_out - Humidity outside (from Chievres weather station)
* Windspeed - Wind speed (from Chievres weather station)
* Visibility - Visibility (from Chievres weather station)
* Tdewpoint - Tdewpoint (from Chievres weather station)
* rv1 - Random variable 1
* rv2 - Random variable 2
* Date - Date and time format
* Appliances - Energy used by appliances (Target Feature)

**2. Introduction**

Understanding the consumption of electricity in daily home appliances such as washing machines, televisions, microwaves etc., which constitutes the major part of electricity demand of a low energy household, could provide major insights in utilization of electricity. Studying this data is important in finding the key factors that could influence the electricity consumption in appliances and thereby work on those factors to decrease the consumption of electricity by appliances. The problem predicts the appliances usage of electricity based on various factors that could influence the consumption of electricity.

The electricity consumption in low energy houses is determined by two main factors, the number of electrical appliances in the house and the usage of appliances by the occupants of the house. There are many factors that could influence the usage by appliances, some of the factors that could influence are the indoor environmental factors near the vicinity of the appliances such as temperature, humidity, light, vibrations etc. The occupancy level of house in different locations could also help in determining the usage of appliances. Developing prediction models for this problem can be useful for many applications such as detecting abnormal energy usage patterns, determining energy demand, to use in building performance simulation etc.

The goal of the project is to predict the energy of the appliances using various features. Regression model is used in this problem for prediction of energy use of appliances. The problem is non-trivial as it has non-linear behavior, that is, it is difficult to predict the energy based on the given features as it is on dependent on all of them. Also, the data had time-series feature which needed some specific pre-processing but as for now time-series did not teach in class yet so as of now, I am not proceeding it with time-series approach instead of that I will be solving our problem using ML techniques. But if given more time, I would give a chance to time series also and would check how our data behaves on time-series models. Even if I am not proceeding with time series approach, we can extract useful information from the date-time feature.

**3. EDA**

* **Observing and Exploring Dataset**

After loading the dataset, we performed some basic functions and methods for knowing the data type. This process helped us figuring out various aspects and relationships among the appliance’s energy consumption and the data factors. It gave us a better idea of which factor behaves in which manner compared to the appliance’s energy consumption.

After observing the data, we would say that there are 29 columns and 19735 rows. According to appliances column there are range of energy consumption which is vary between 10 to 1080 Wh. May be there can be some outliers also we will also handle that.

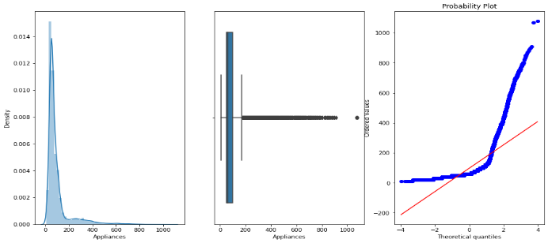
* **Null, Missing and Duplicate Values Treatment**

It is an important aspect of Data Cleaning because there can be some null, missing and duplicate values in our dataset. But our dataset doesn't contains a null or missing values which might tend to disturb our accuracy, if it has null or missing values then we have to drop them at the beginning of our project in order to get a better results.

* **Dropping unneccessary Column**

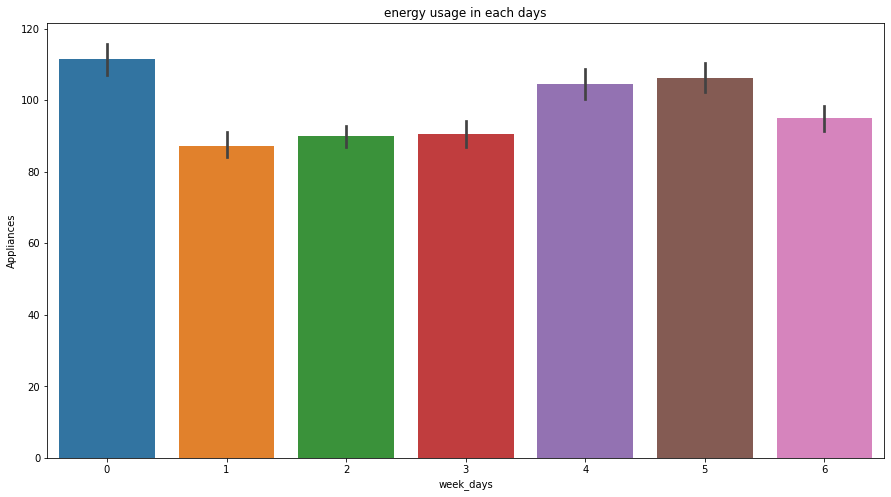
Our dataset does contains some unnecessary column like Lights, we didn't need it for analysis so we can drop them out from dataset(as per our problem statements).

* **Outliers handling**

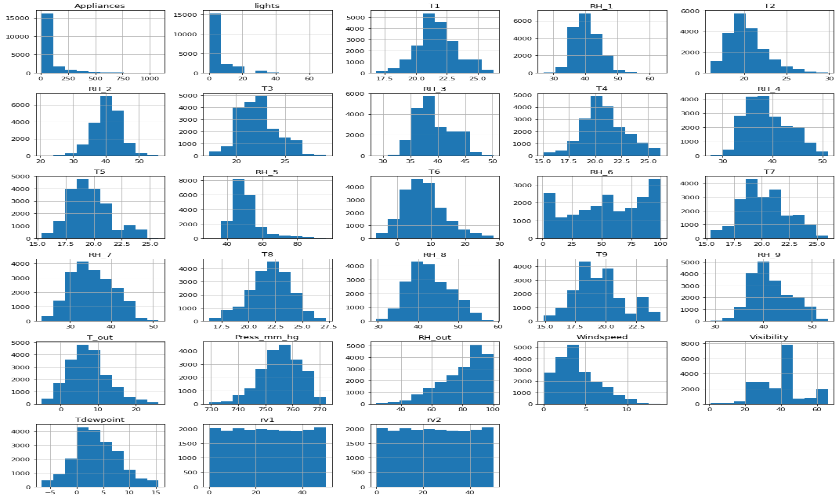
We have a lot of outliers in our dataset and in our target column. According to IGR method we have 2138 rows with outliers. We have done both method removal and non-removal of outliers but removal of outliers done great with our accuracy. 

* **Analyzing and Visualization of data(Including Distribution)**

In these steps we used plots like box plot to check the outliers and we used histogram for observing the distribution of our features as well as probability plot. We also used heatmap for finding correlation between our target and other features. 

Next we used bar plot for finding top hours and weekday in which energy consumption is higher than other hours and weekdays respectively after extracting some information from date column. We are using rel plot and kde plot also for our target feature.

After observation we can say that most of our feature are normally distributed except appliances, RH\_5, RH\_out, T6, T5,T\_out, rv1 and rv2 . We also tried Transformation like log,sqrt etc. but didn't get the good accuracy so we removed or commented those transformation.



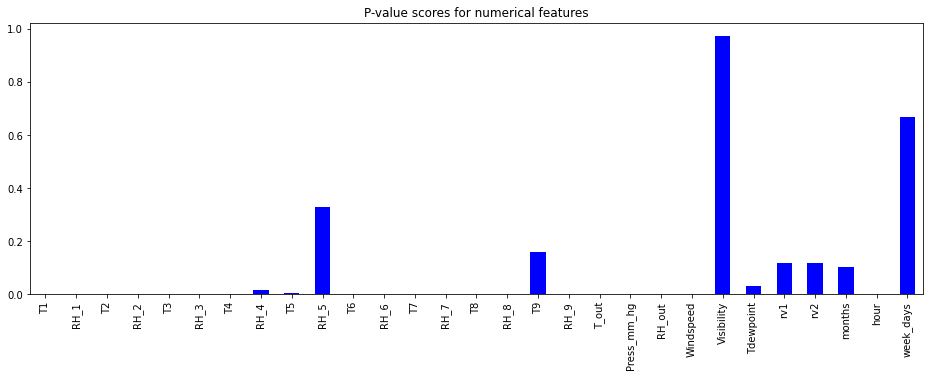
* **Using different visualization methods**

For visualization we tried various plots like:

* **Prob Plot**
* **Bar Plot**
* **Rel Plot**
* **Histogram**
* **Box plot**
* **Kde Plot**
* **Heat map**

**4. Feature Engineering**

For feature selection we used two method -

1. **VIF Method-** In this method we didn't get good score so we skip(commented) it.
2. **F-regression-**This method gives us promising results so we move ahead with F-regression method and removed some unnecessary features like T9, RH\_4, RH\_5, months, Tdewpoint, Visibility, rv1, rv2 and weekdays. 

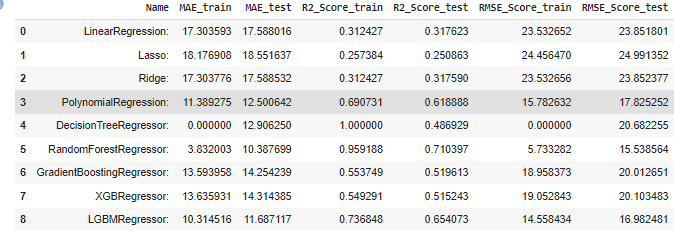
**5. Preparation for Model Making**

* **Splitting** - In this particular step we splitted our data to train and test data with 20% test data.
* **Scaling** - We used two scaler Min-max and Standard for our data but Standard scaler gives good results so we proceed further with it.
* **PCA** - we also applied pca on our data and play with but none of components gives good accuracy so we removed or commented it as well.
* **F-regression and Select best** - We also tried these for all and top features but none of them done great with our results.

**6. Making Models**

I used lot of regression processes to train my model in the beginning such as

* Linear Regression
* Ridge Regression
* Lasso Regression
* Decision Tree Regressor
* Lgbm Regression
* Random Forest Regressor
* Gradient Boosting Regressor
* XGB Regressor



Top 2 models were selected from these above using pre-train set on the basis of good r2 score and rmse. And they were used for hyperparameter tuning for further improvement.

Note- The lower the value of the RMSE and the higher the R2 the better the model.

**7. Hyperparameter Tuning**

We only did tune for top 2 models:

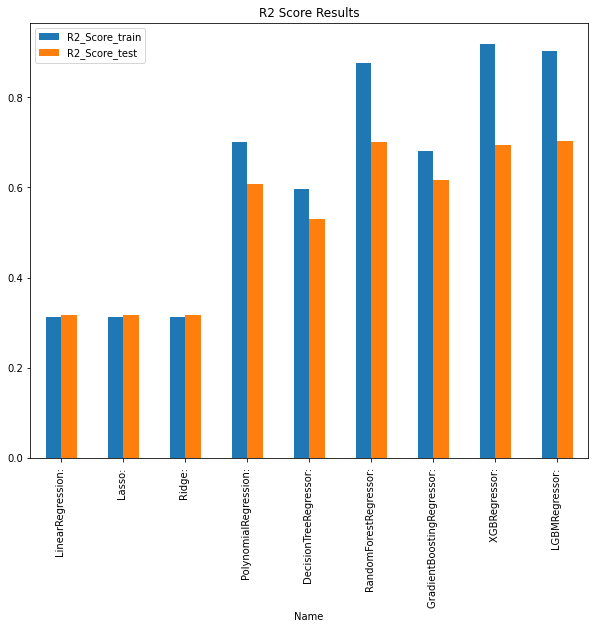
* **Random Forest Regressor**: Random Forest for regression operates by constructing a multitude decision trees at training time. It uses number of estimators that fit a number of decision trees on various sub-samples of dataset. It uses averaging to improve the predicted results and controls over-fitting. The main advantage of random forest regression is that prevents over-fitting. Random Forest is good at handling numerical data and it is also able to capture the non-linear interaction between the features and the target.

I used an inbuilt sklearn. ensemble.RandomForestRegressor() with the help of Grid search(commented) and bayesian search min\_samples\_split and max\_features as my hyperparameter to tune the model.

After playing a lot with these parameters we got our best parameter which were improving our r2 score

**Lgbm Regressor**:

We also gone for Lgbm Regressor for tuning with the help of Grid search(commented) and bayesian search with n\_estimators, max\_depth, min\_samples\_leaf and min\_samples\_split as my hyperparameter to tune the model.

After playing a lot with these parameters also we didn't got our best parameter which can improving our r2 score as well as rmse

**8. Final Result**

For test set, the best model obtained after validation was used and the best parameter as well. The model performance was now tested on the test set and according to results Random Forest is doing good from all of the above model so our final model for this dataset is Random Forest Regressor. It was observed that the random forest gives best results as it prevents overfitting and optimizes the data. It is also seen that the error is not that low and this might be due to the dataset available. The main reason for this is that appliances’ consumption profile is highly variable. The features which were present were not all dependent on the target as observed from the correlation results and that resulted into high error. Whereas, elimination of these features could have helped a little but they play different roles for different regressors and that’s why it was little difficult to determine which one to consider and which one to eliminate for a regressor. Therefore, I just found which features were more important for the model performance and arranged them in an order where top one is the best feature and last one is the least important feature. The result found is as follows:

1. T2

2. RH\_8

3. RH\_1

4.T8

5. RH\_5

6. RH\_6

7. T\_out

8.RH\_out

11. Tdewpoint

12. T4

13. T7

14. T5

15. T3

16. Windspeed

17. T1

**9. Conclusion**

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).

1) Main aim of the project is to predict energy consumption of Appliances. First, we analysed the data but the information from the data set is collected in regular interval of time so it's time series data. We are not implementing time series technique on the model because of less knowledge on time series.

2) Then we used the matplotlib and seaborn to do Exploratory Data Analysis on data by plotting different graphs like scatter-plot, bar plot, boxplot, subplot and heat map. From this we got useful insights like:

* Many columns in the dataset are not normally distributed and target column is also right skewed.
* Dataset has many outliers and no null values.
* We have hours column which is highly correlated with dependent variable and there are lot features that have lesser than 0.1 correlation with dependent variable and it is a non-linear dataset.
* Energy consumption in month of March is high and low in January and the increase in temp leads to more energy consumption.
* Decrease in Humidity leads to increase in power consumption. Humidity is inversely proportional to dependent variable i.e Energy Consumption.
* Hour of the Day is the most important influencing parameter for Energy consumption.
* High Electricity consumption of >140Wh is observed during evening hours 16:00 to 20:00. Weekends (Saturdays and Sundays) also observed high consumption of electricity. (> 25% than Weekdays)
* lights have very low importance as a feature.

3) In feature selection we used variance threshold, F\_Regression and Pearson correlation matrix and using them we removed features that are not important for predicting dependent variable.

4) In feature engineering technique we removed outliers in our model.

5) Algorithms like Linear regression, Polynomial regression, Decision tree, Random Forest, Gradient boosting, XGBM and LGBM regression are used and cross validation hyperparameter tuning was done on the all models. By comparing all models, we found that random forest regressor performs good having high r2 score and MSE, RMSE value is also low for random forest. Some overfitting is happening because dataset is time series and we are not implementing time series concept.

6) Finally, Model explain ability Shap technique is used to know which features are important for predicting output and understanding model. Hour feature is the most important feature.

.**10. Summary and Conclusion**

* **Problem:** Regression Problem
* **Best Model:** Random Forest Regressor
* **R2 score on train:**
* **R2 score on test:**

**References-**

* Stack overflow
* GeeksforGeeks
* Kaggle
* Analytics Vidhya
* Machinelearningmastery
* Stack exchange