Appliance Energy Prediction

V Rupesh Kumar Patro
Shashank Maindola

Abstract:

In this century, the world is being driven towards more cleaner means of energy. The electricity generated using the renewable sources plays a major role. But simply generating and utilizing the electrical energy is not viable, any extra generation will lead to non utilization and less generation will lead to power outages. Hence, we need demand side and supply side management so we can flatten the generation distribution curve. The residential load contributes to 27% of the total energy consumption (**Source:**:Eurostat). Residential household consumption is generally governed by weather conditions. Apart from that occupancy, time of day, time of the year and day of week also plays a major role in the energy consumption

Introduction:

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. Here we have a data of one of e residential building in Belgium and using machine learning techniques we have to use various regression models and predict the appliance energy consumption using the best model.

Data Set

Below is the info that is available in given dataset-

Note:- 'lights' feature 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

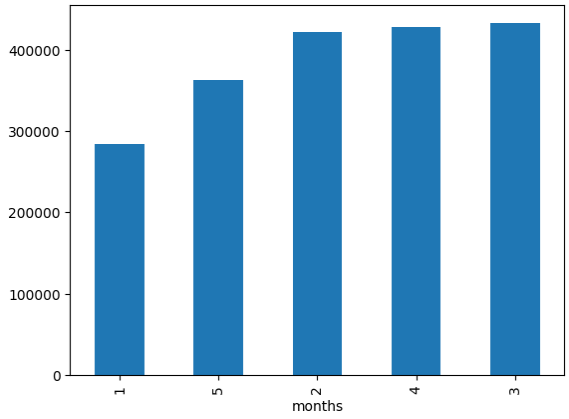
- **RH_5** - Humidity in bathroom
- **T6** - Temperature outside the building
- **RH_6** - Humidity outside the building
- **T7** - Temperature in ironing room
- **RH_7** - Humidity in ironing room
- **T8** - Temperature in teenager room 2
- **RH_8** - Humidity in teenager room 2
- **T9** - Temperature 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)

Steps for appliance energy prediction using machine learning

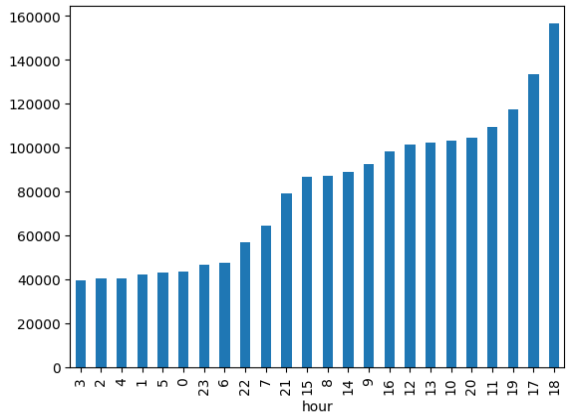
1. **EDA**
2. **Clean up**
3. **Feature Engineering**
4. **Pre Processing**
5. **Model Implementation**
6. **Model Explanation**

1. Exploratory Data Analysis

Seasonal and Hourly consumption Pattern



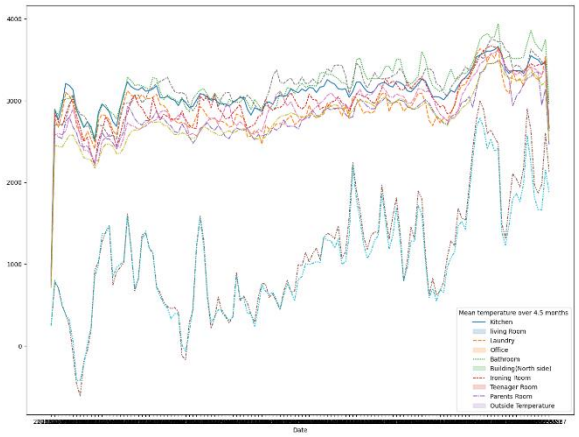
Total monthly energy consumption

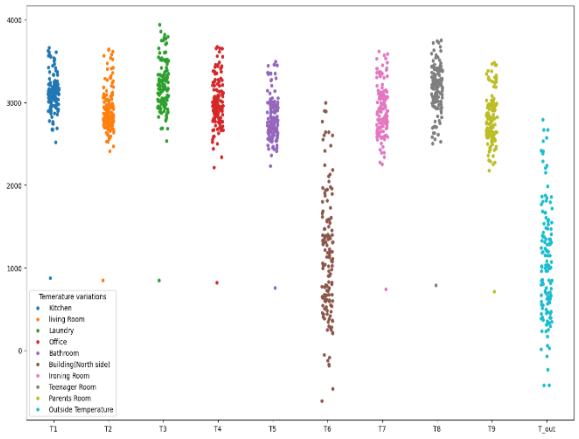


Total hourly energy consumption

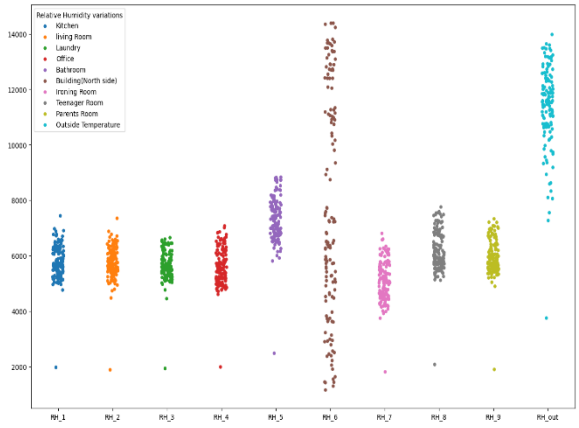
- It is clearly understood that march has the maximum energy consumption.
- The maximum power consumption is around evening while the least consumption is around the early morning hours.

Temperature and Humidity



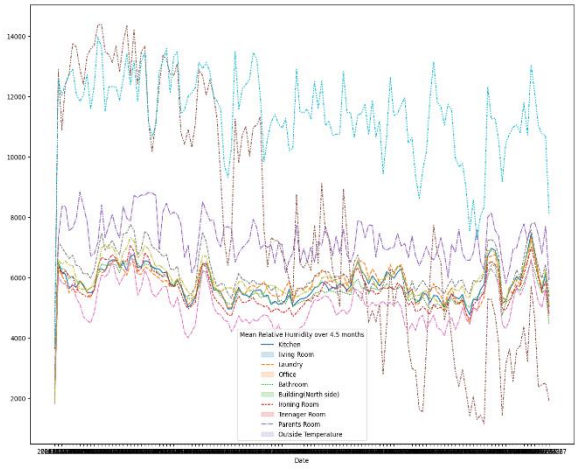


Household and outside temperature

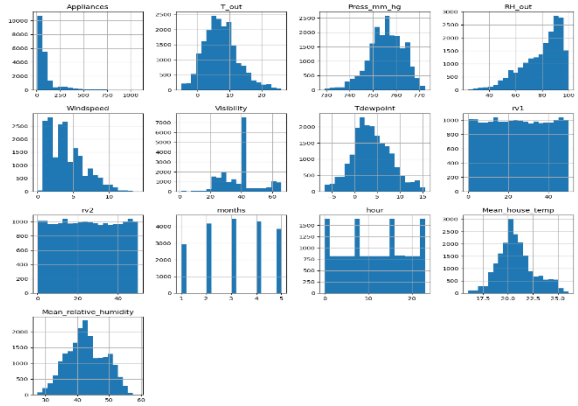


Household and outside Humidity

- The northside temperature as it is similar to north side temperature, hence it verifies that sensor was working perfectly and the data gathered is valid. The temperature inside is more or less same.
- There is a variation in relative humidity of Building (North side) and Outside. The outside humidity is from airport weather sensor so the humidity can be different while the temperature outside the building is different due to the neighborhood factors like landscaping etc., hence we will be ignoring the relative humidity data from the airport**.**

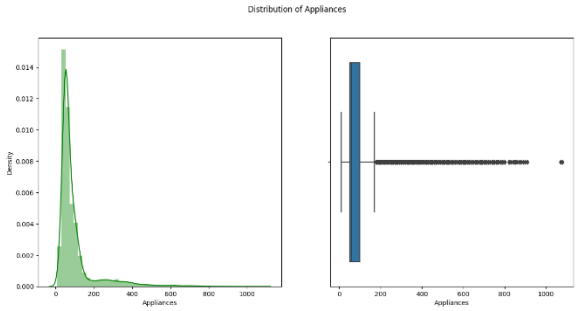


Distribution

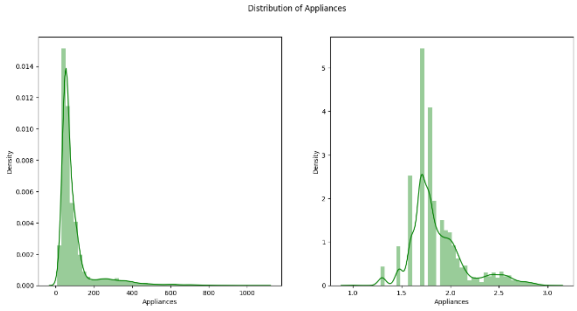


- Press_mm_hg, Visibility, TDewpoint, rv1, rv2, Mean_house_temp, Mean_relative_humidity,months, hour are normal distributed data.
- Positively skewed(>1):- Appliances.
- Moderately Positively skewed(0.5 to 1):- T_out, Windspeed.
- Normal Distributed(-0.5 to +0.5):- Press_mm_hg, Visibility, Tdewpoint, rv1, rv2, Mean_house_temp, Mean_relative_humidity, months,hour.
- Negative skewed(-0.5 to -1):- RH_6.

Target Variable



Distribution of appliance energy consumption



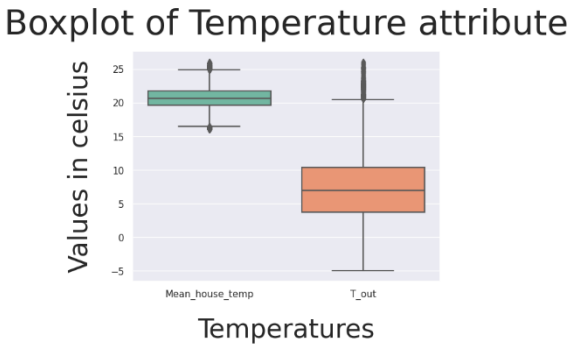
Our graph is moving towards to y axis as it is positively skewed and we couldn't get any better visualization with these type of graph. Hence we took Log or Square Root or Exponential of the dependent variable and draw the graph.

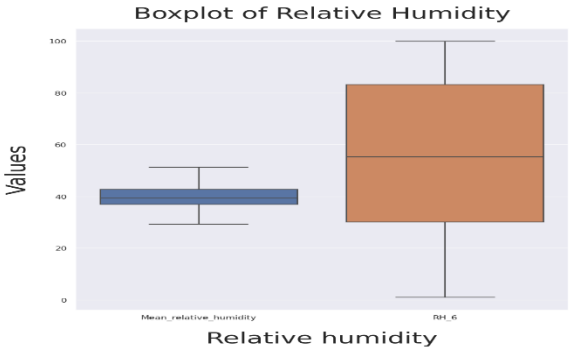
- Dependent variable is right skewed and lot of outliers present in our data set but they are not ignored at there are sometimes situations when the consumption increases here we see there are lot of such instances.
- There is positive correlation between temperature inside and outside with the appliance energy consumption.
- There is low correlation of appliance energy consumption with other variables.

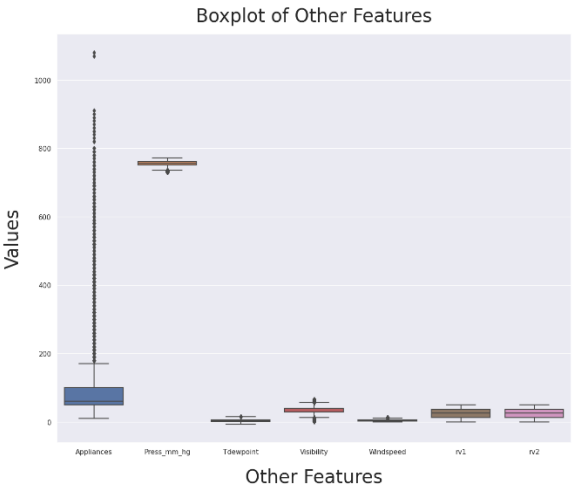
2. Clean Up

Data cleaning is one of the most essential subtasks of any data science project.
Our dataset has no missing values but we found a number of outliers

Appliances	2138
T_out	436
Press_mm_hg	219
RH_6	0
Windspeed	214
Visibility	2522
Tdewpoint	10
rv1	0
rv2	0
months	0
hour	0
Mean_house_temp	512
Mean_relative_humidity	0







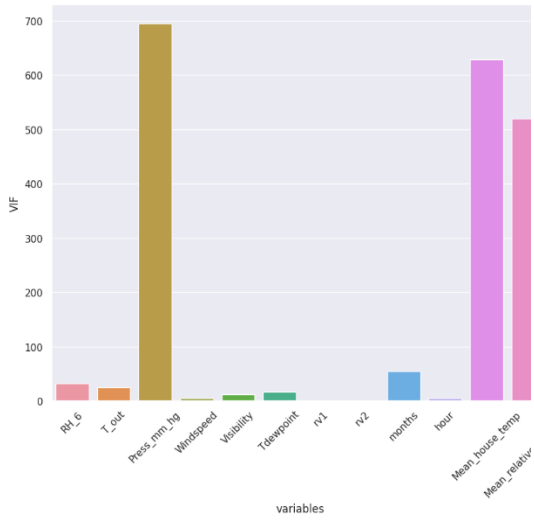
Firstly, we removed certain features like temperature in individual rooms, humidity in individual rooms and replaced them with the mean values for faster model training. We ignored the humidity in the bathroom and humidity at the airport as humidity in bathroom will always be higher than rest of the house and humidity of the airport doesn't represent the humidity of the neighborhood. Temperature outside is almost same as temperature of the northside of the building hence we will ignored temperature out side building. We removed r1 and r2 features as they have infinite VIF. We also removed light feature from our data set.

Secondly, we observe that there are a lot of outliers in visibility and appliance but we will not remove them as there are conditions when there is spike in demand and they are realistic hence we will not ignore them. Rest of the feature have little outliers so we will not remove them.

3. Feature Engineering

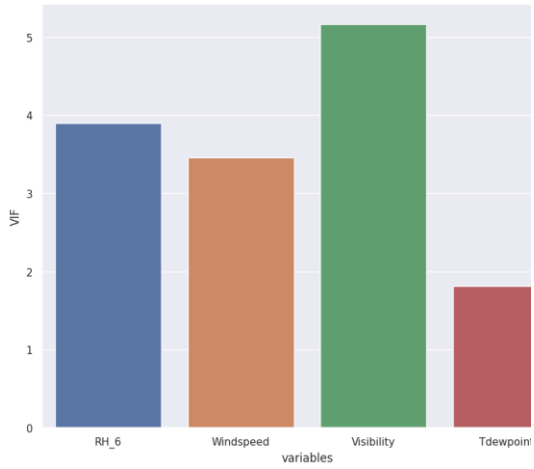
We calculate Variance Inflation Factor for the all the feature to measure the amount of multicollinearity.

VIF	
RH_6	32.20155615335796
T_out	24.95561180861631
Press_mm_hg	694.2964804113375
Windspeed	5.005980999514861
Visibility	11.771630692843596
Tdewpoint	7.168244694540604
rv1	Infinity
rv2	Infinity
Months	53.95771288882704
Hour	4.7905065440715555
Mean_house_temp	27.8826782455517
Mean_relative_humidity	519.4920680561612



rv1 and rv2 has infinite Variance Inflation Factor hence we will remove them.

We again calculated the VIF after removing the independent features



After Removing Multicollinearity of Independent Variables

1. RH_6(**3.892468**), Windspeed (**3.458756**), Tdewpoint(**1.815175**) :-Moderate Multicollinearity
2. Visibility (**5.159890**): - High Multicollinearity

4. Pre Processing

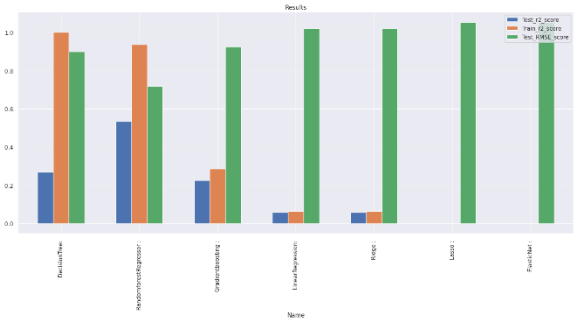
	count	mean	std	min	25%	50%	75%	max
T_out	19735.0	7.412580	5.318464	-5.000000	3.670000	6.920000	10.400000	26.100000
Press_mm_hg	19735.0	755.522602	7.399441	729.300000	750.933333	756.100000	760.933333	772.300000
RH_6	19735.0	54.609083	31.149806	1.000000	30.025000	55.290000	83.226667	99.900000
Windspeed	19735.0	4.039752	2.451221	0.000000	2.000000	3.666667	5.500000	14.000000
Visibility	19735.0	38.330834	11.794719	1.000000	29.000000	40.000000	40.000000	66.000000
Tdewpoint	19735.0	3.760995	4.195248	-6.600000	0.900000	3.430000	6.570000	15.500000
rv1	19735.0	24.988033	14.496634	0.005322	12.497889	24.897653	37.583769	49.996530
rv2	19735.0	24.988033	14.496634	0.005322	12.497889	24.897653	37.583769	49.996530
months	19735.0	3.101647	1.339200	1.000000	2.000000	3.000000	4.000000	5.000000
hour	19735.0	11.502002	6.921953	0.000000	6.000000	12.000000	17.000000	23.000000
Mean_house_temp	19735.0	20.815611	1.812567	16.012708	19.663000	20.597500	21.764375	26.061940
Mean_relative_humidity	19735.0	39.832333	3.929901	29.264857	36.826714	39.224490	42.698075	51.238571

Our final dataset for model training is split into two one which containing the features for model training(X) and the target variable(y). The two data set are also divided using 20:80 split for training and test the model

For feature scaling we used standardization using standard scalar.

5. Model Implementation

We used Decision Tree, Random Forest Regressor, Linear Regression, Gradient Boosting , Lasso, Ridge , Elastic Net algorithms for model training.

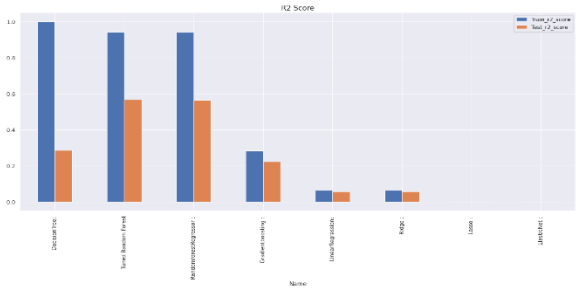


So after training data Random Forest Regressor has the least RMSE score and a decent R2 score.

Before proceeding to try next models, let us try to tune some hyperparameters and see if the performance of our model improves.

Hyperparameter tuning is the process of choosing a set of optimal hyperparameters for a learning algorithm.

Name	Train_r2_score	Test_r2_score	Train_MSE_score	Test_MSE_score	Train_RMSE_score	Test_RMS
DecisionTree:	1.000000	0.286427	4.063988e-34	0.790025	2.015933e-17	
Tuned Random Forest	0.942690	0.568343	5.730530e-02	0.477904	2.393853e-01	
RandomforestRegressor :	0.941369	0.565458	5.863053e-02	0.481098	2.421374e-01	
Gradientboosting :	0.283830	0.224162	7.161700e-01	0.858960	8.462683e-01	
LinearRegression:	0.063877	0.058172	9.361230e-01	1.042735	9.675345e-01	
Ridge :	0.063877	0.058173	9.361230e-01	1.042734	9.675345e-01	
Lasso :	0.000000	-0.000371	1.000000e+00	1.107550	1.000000e+00	
ElasticNet :	0.000000	-0.000371	1.000000e+00	1.107550	1.000000e+00	

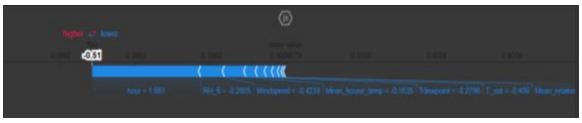


Result

The Dataset does not contains null values ,but there is very less correlation between features and target variables.
* By fitting all the model get best score in Random Forest regressor , after tuning the hyper parameter using GridsearchCV, GET Train r2 score 0.94 and test r2 score 0.5622 because of improper dataset and less correlation between feature and target variable.
* The performance is low due to like:- no proper pattern of data, less correlation , not enough relevant features.

6. Model Explaination

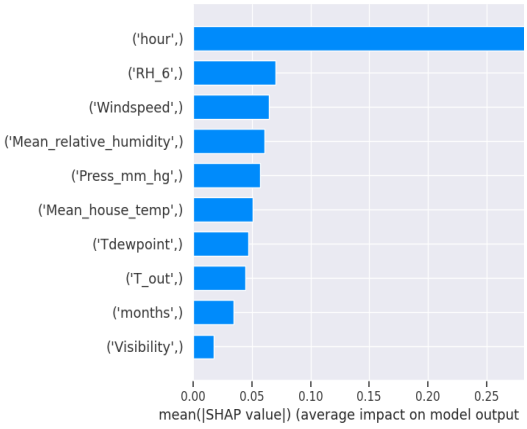
We used SHAP to explain our model to explain our machine learning model.



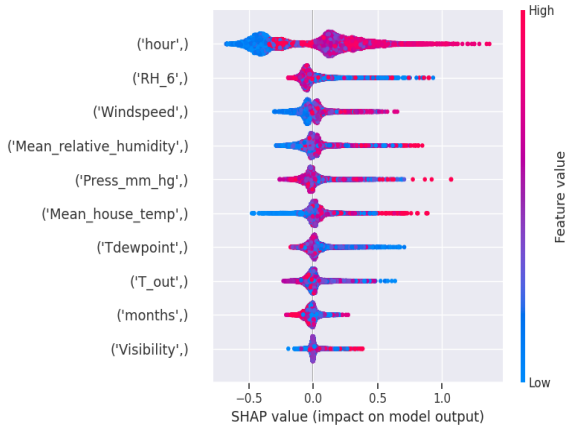
We used Random Forest Regressor model as it has the most accurate prediction. This SHAP plot gives us the explain ability of a single model prediction. Force plot can be used for error analysis, finding the explanation to specific instance

prediction. The model is explained as follow:-

- The model output value: -0.51
- The base value: this is the value that would be predicted if we didn't know any features for the current instance. The base value is the average of the model output over the training dataset
- The numbers on the plot arrows are the value of the feature for this instance.
- Red represents features that pushed the model score higher, and blue representing features that pushed the score lower.
- The bigger the arrow, the bigger the impact of the feature on the output.
- The amount of decrease or increase in the impact can be seen on the x-axis.



Bar Summary Plot



Dot Summary Plot

Conclusion

- The most important determining factor for energy consumption is the hour of day.
- Energy consumption is high in March and low in January, and a rise in temperature results in higher energy consumption.
- As a feature, lights are extremely undervalued.
- Decreased humidity leads to an increase in electricity consumption. Humidity is proportional to the dependent variable.
- We have a high correlation with the dependent variable in the hours column, and many features have less than a 0.1 correlation with the dependent variable in the non linear dataset.
- During the evening hours of 16:00 to 20:00, there is a high usage of electricity of more than 140Wh. Electricity use is highest on weekends (Saturdays and Sundays).

(more than 25% higher than on weekdays)

- The dataset has many outliers and no null values.
- Many columns in the dataset are not normally distributed, and the target column is also right skewed.

not only helps the developer to know the shortcoming in the app so that the developer can provide a great user experience.

References-

1. Investopedia.com
2. GeeksforGeeks
3. Stackoverflow
4. https://shap.readthedocs.io/
5. Kaggle
6. Stackoverflow