

**ADVANCED DATA SCIENCE**

Assignment 2

Machine Learning with Energy Datasets

Prof. Sri Krishnamurthy



Team Members: Sayali Borse Rishi Rajani Komal Ambekar

* Content:

|  |  |
| --- | --- |
| **Sr.No.** | **Content** |
| 1 | Research papers summary |
| 2 | Exploratory Data Analysis |
| 3 | Feature Engineering |
| 4 | Prediction Algorithms |
| 5 | Feature Selection |
| 6 | Model Validation and Selection |
| 7 | Final Pipeline |

* Summary:

As interest in IOT and sensors pick up steam, companies are trying to build algorithms and systems to understand consumer behavior to help them make better decisions. One such application is energy modeling. Though, most consumers are aware of their aggregate consumption of energy, few are aware of how and where energy is consumed. With increasing sensors in equipment, it is becoming easier to find out which equipment/instruments consume the most power. AdaptiveAlgo Systems Inc. works on solutions to build algorithms and platforms to address energy modeling challenges. The company is putting together a solution for energy modeling and is interested in understanding consumer energy usage and the attributes that contribute to appliance energy usage. The data scientists there came across a recent paper and dataset and are interested in building various machine learning models that could contribute to understanding energy usage by appliances and the attributes that contribute to aggregate energy usage. With the knowledge of energy consumed by various equipment, seasonality and attributes like temperature and humidity, a machine learning model could be used to predict aggregate energy use.

* Research Paper Summary:

We were given three research papers for review and analysis. The study is related to appliances and energy consumption and prediction of energy consumption. Each document had to offer different features. The three papers were:

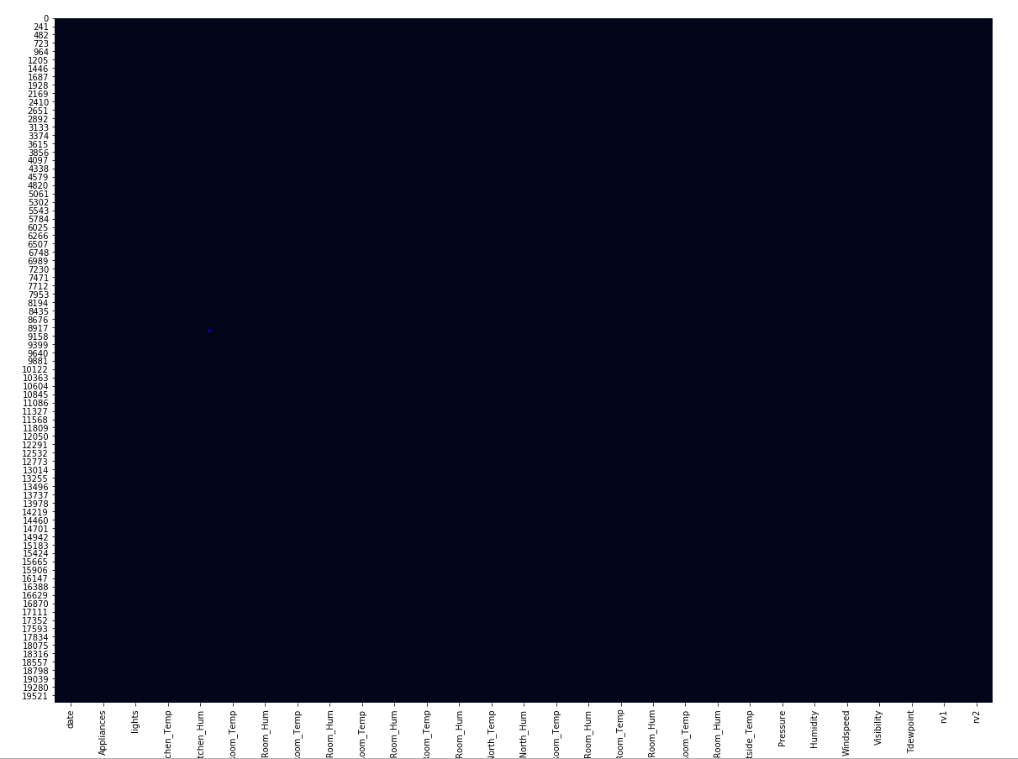
1. Data driven prediction models of energy use of appliances in a low-energy house.
2. A review of artificial intelligence-based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models
3. Prediction of appliances energy use in smart homes

* Exploratory Data Analysis:

It is an approach to analyze data sets to summarize their main characteristics, often with visual methods. EDA is for seeing what the data can tell us beyond the formal modeling. It is typically the first step of analysis.

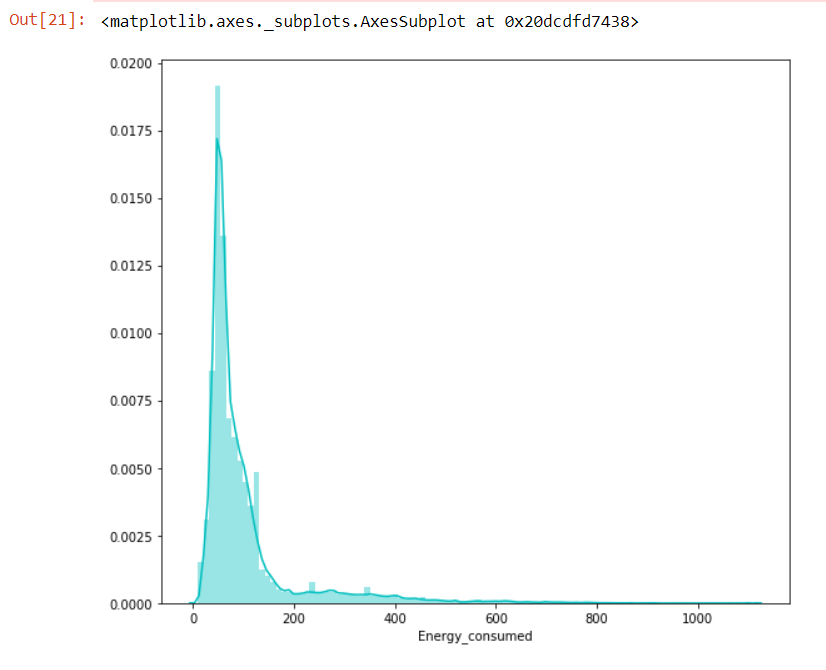
To check all null values in the dataset:



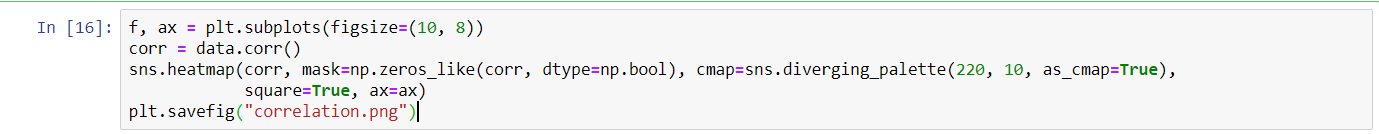


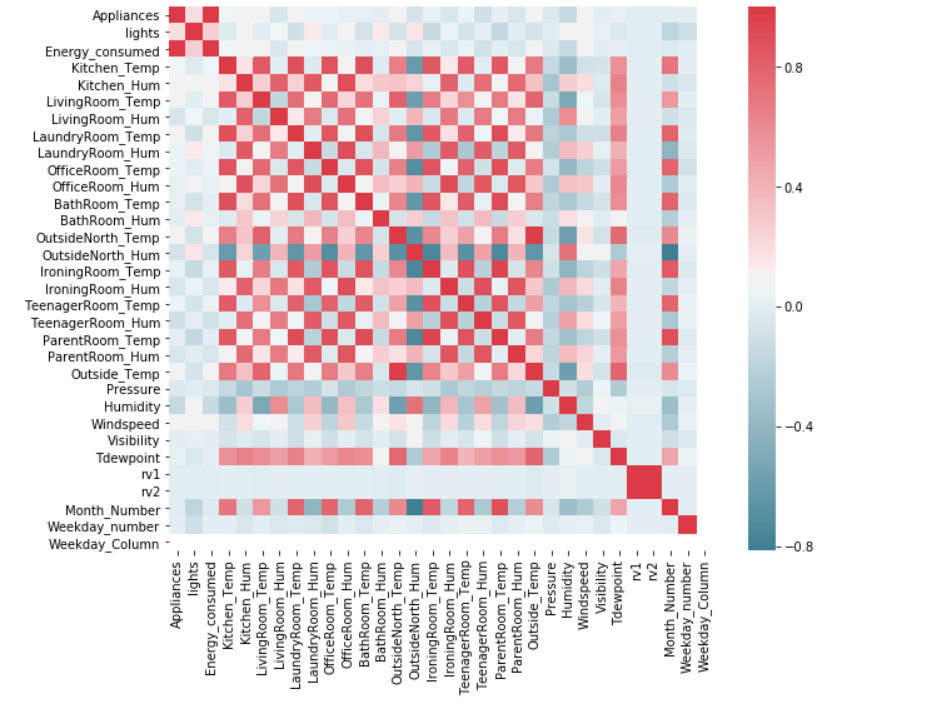
* Energy Consumed:



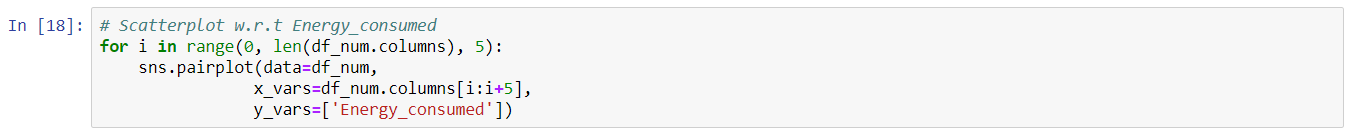


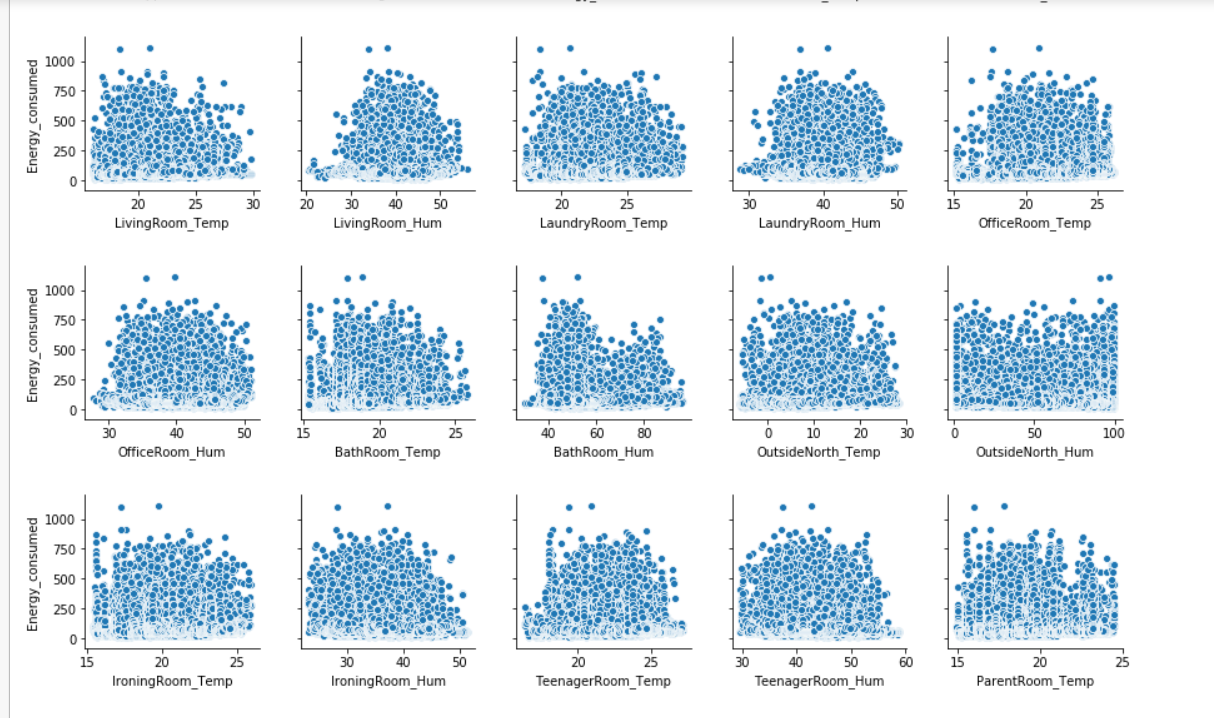
* Correlation Matrix



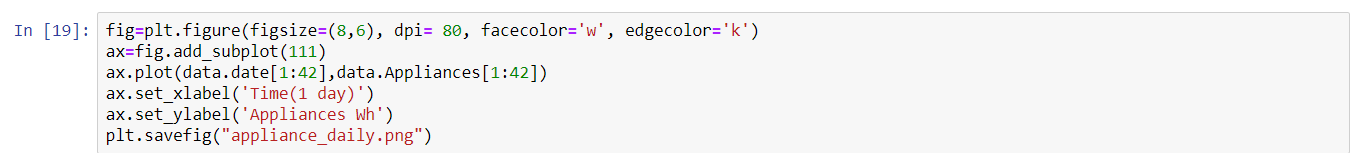


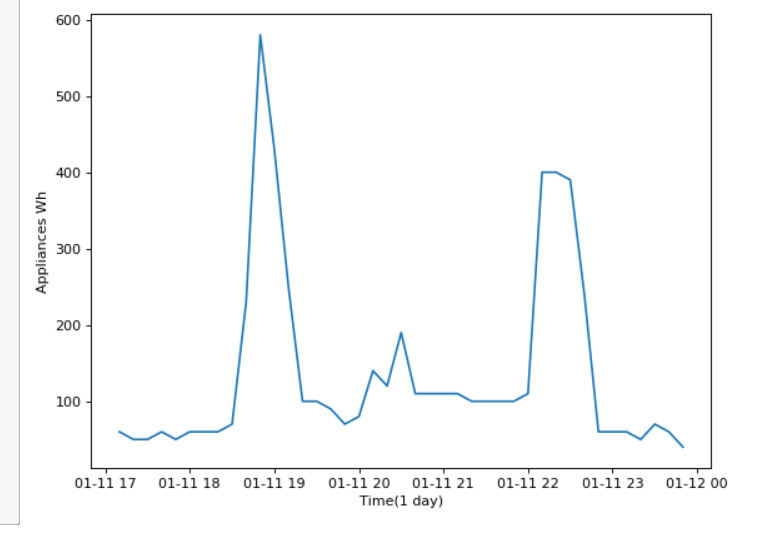
* Scatter plot wrt Energy Consumed



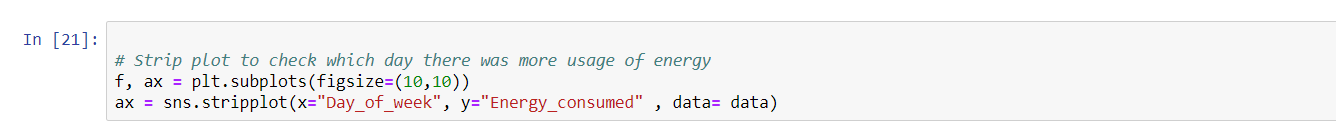


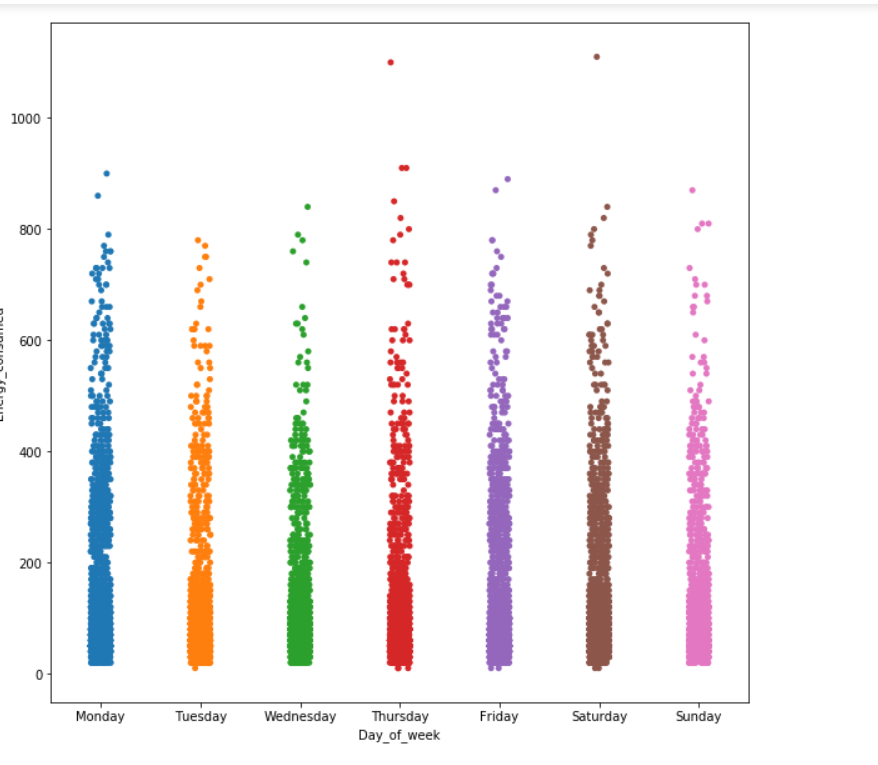
* Daily Energy Consumption





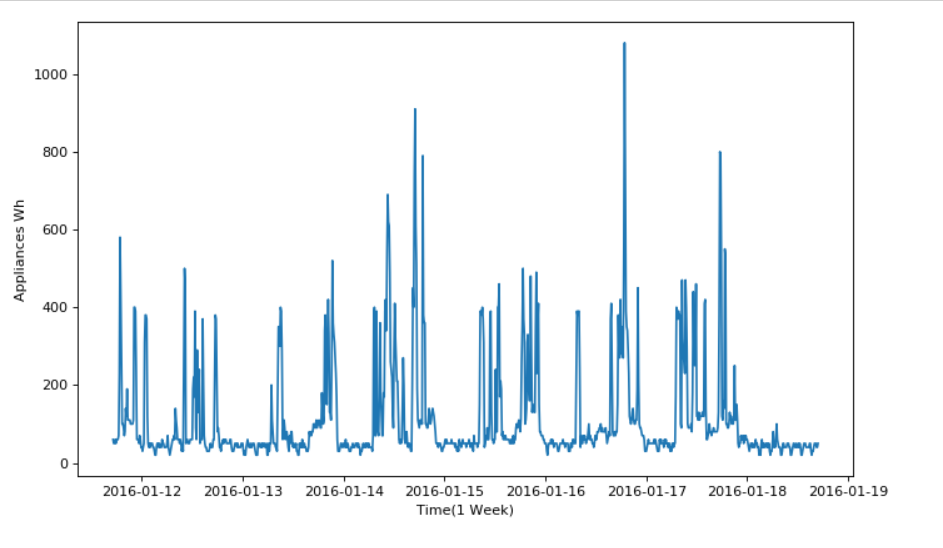
* Weekday wise energy consumption





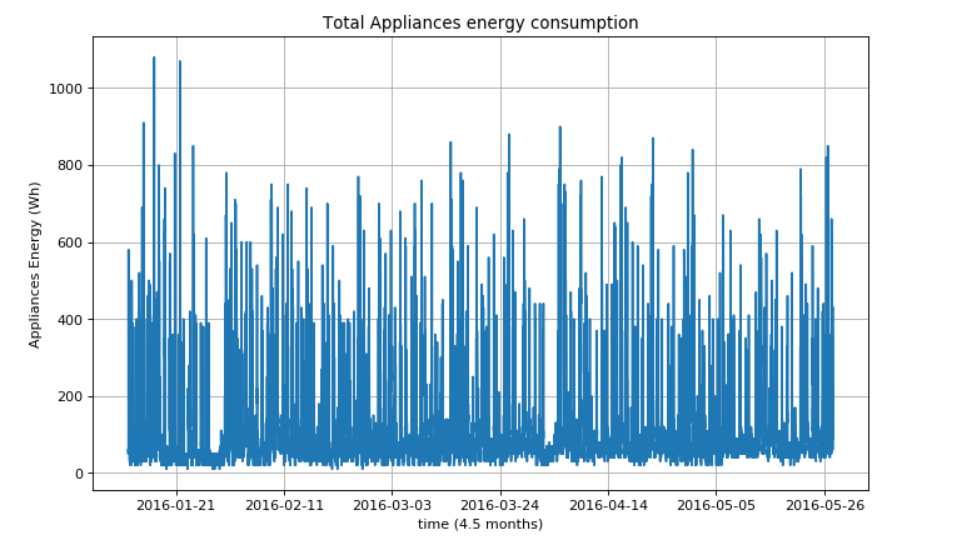
* Weekwise Energy Consumption



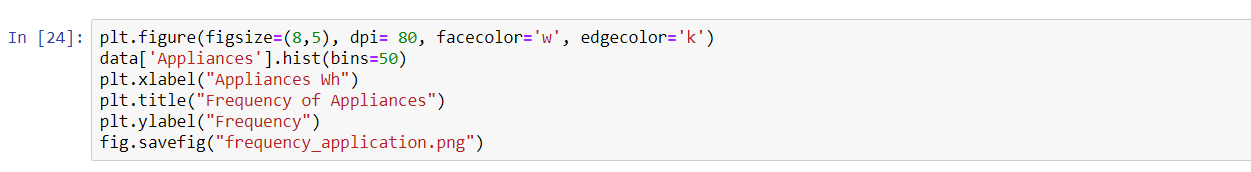


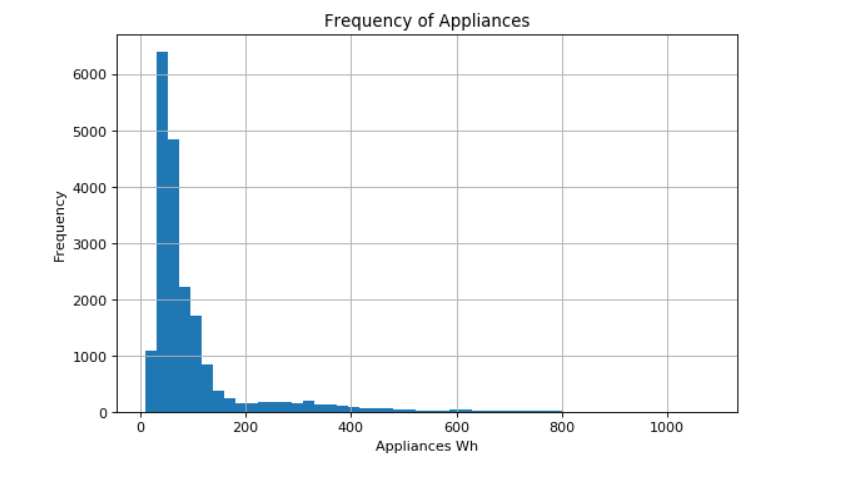
* Total Appliances energy consumption



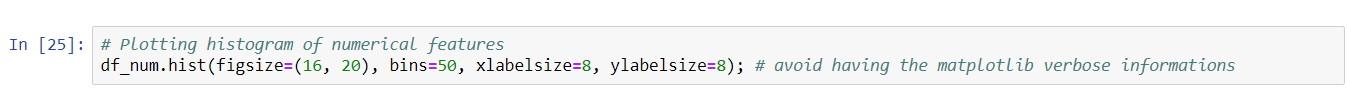


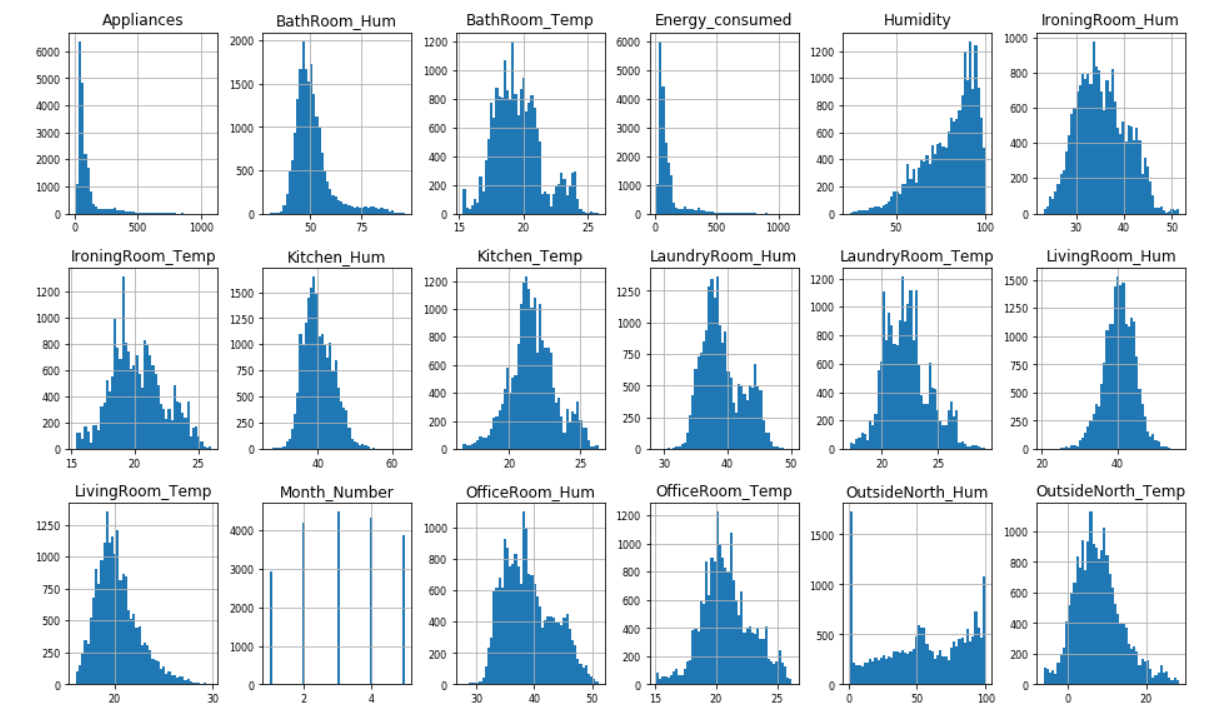
* Frequency of Appliances





* Plotting Histogram of Numerical Features





* Feature Engineering:

The dataset is numeric as we don’t have any categorical features in it.

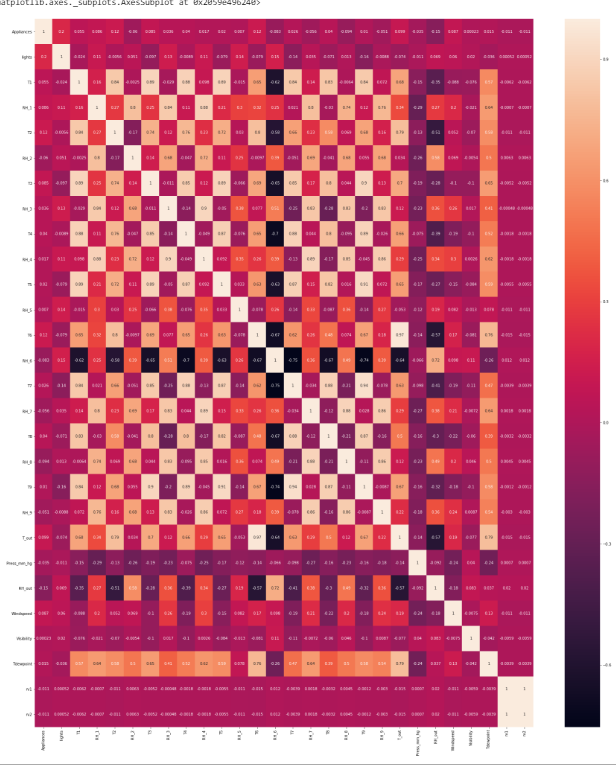
We also do not have any missing values

While studying the data we inferred that the total amount of energy consumed was derived by energy consumed by appliances and the light. The problem statement required us to predict the energy consumed basis the attributes that contribute to the energy consumption.

We also noticed the date time object is not helping us and hence decided to further divide it into attributes like time date, day of the month, week of the month, month number etc.

We also found the correlations between all variables and noticed that rv1 and rv2 were highly correlated and hence dropped rv2.

Lastly, we also renamed our columns for a better understanding.



* Predictive Models:

In this section we have explored numerous machine learning models which can be used in our study to help getting best results in prediction.

We implemented the predictive models - Linear Regression, Random Forest and Neural Networks on our data sets and derived the accuracy score for all of them. We divided the fragment in 66% and 33% basis the sample training and testing data given in the problem statement.

We noticed that the accuracy score for random forest was the best among the lot where as neural networks was bad. Kindly find the below table indicating the matrix.

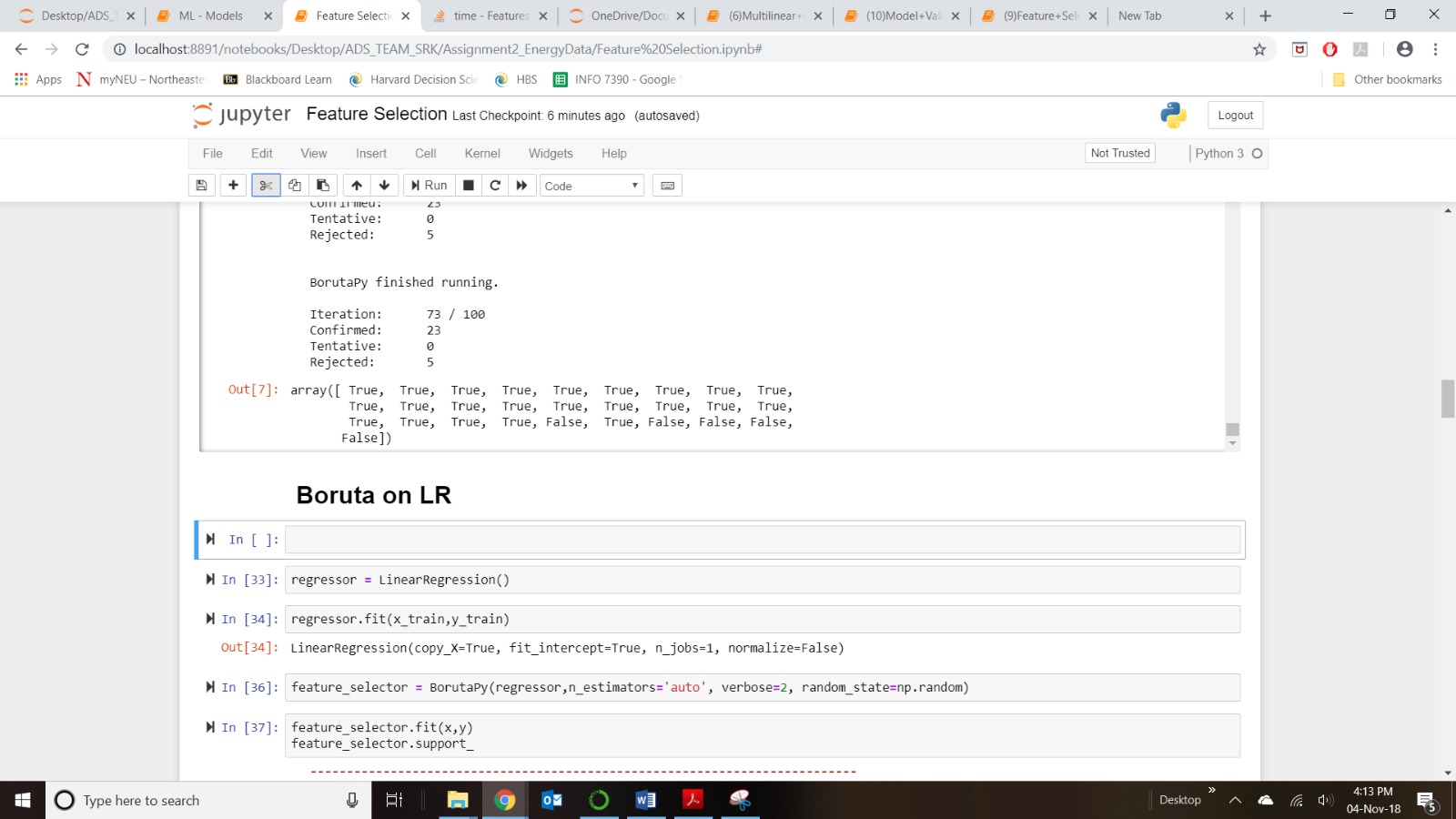
* Find the analysis of the models below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name of Model | Train / Test | Score | MAE | RMSE | R2 | MAPE |
| Linear Regression | Train | 0.1643 | 55.08 | 95.64 | 0.16 | 121.12 |
| Test | 0.1411 | 54.05 | 95.35 | 0.14 | 120.06 |
| Random Forest | Train | 0.9420 | 11.96 | 25.18 | 0.94 | 103.20 |
| Test | 0.5763 | 31.51 | 67.38 | 0.5763 | 108.53 |
| Neural Network | Train | 0.077 | 60.87 | 100.54 | 0.077 | 128.52 |
| Test | 0.074 | 59.28 | 100.54 | 0.074 | 128.15 |

Later in the report you will also come across Extra Tree Regressor which we selected basis a technique we came across while performing feature selection.

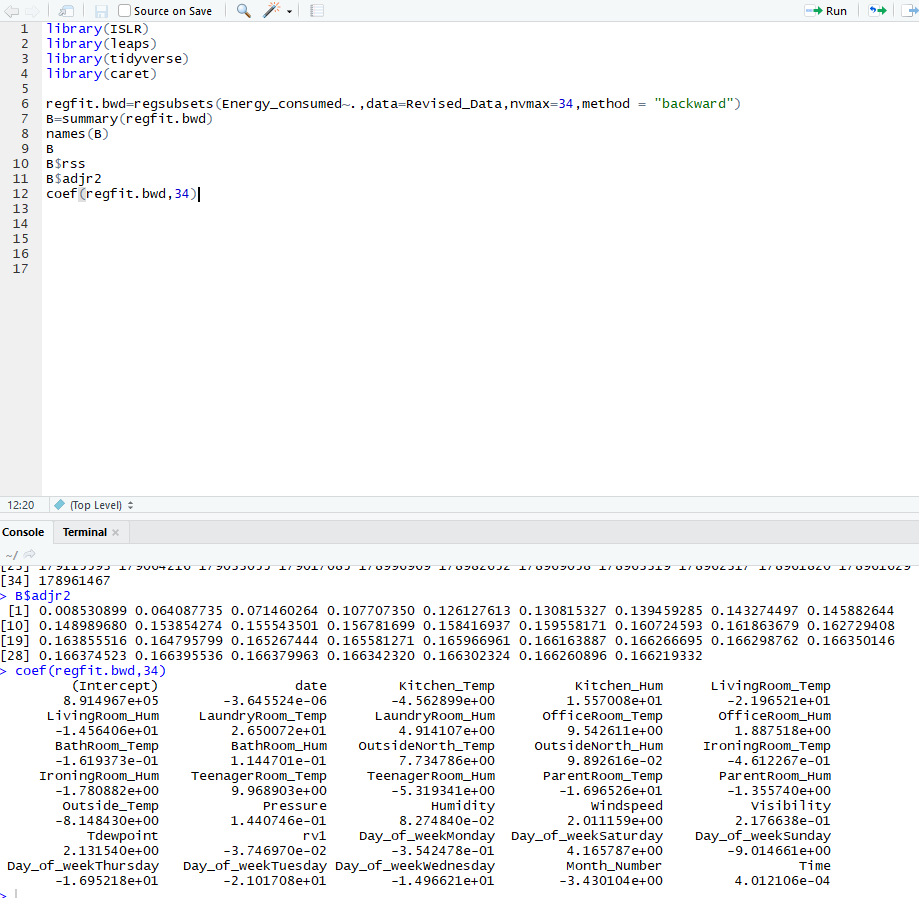
* Feature Selection
* Feature Selection Using Boruta:

1. Feature selection deals with analyzing the feature and ranking them with different performance metrics and using them in the further model to get the best results.
2. We have explored several feature selection techniques like:
3. Boruta Package

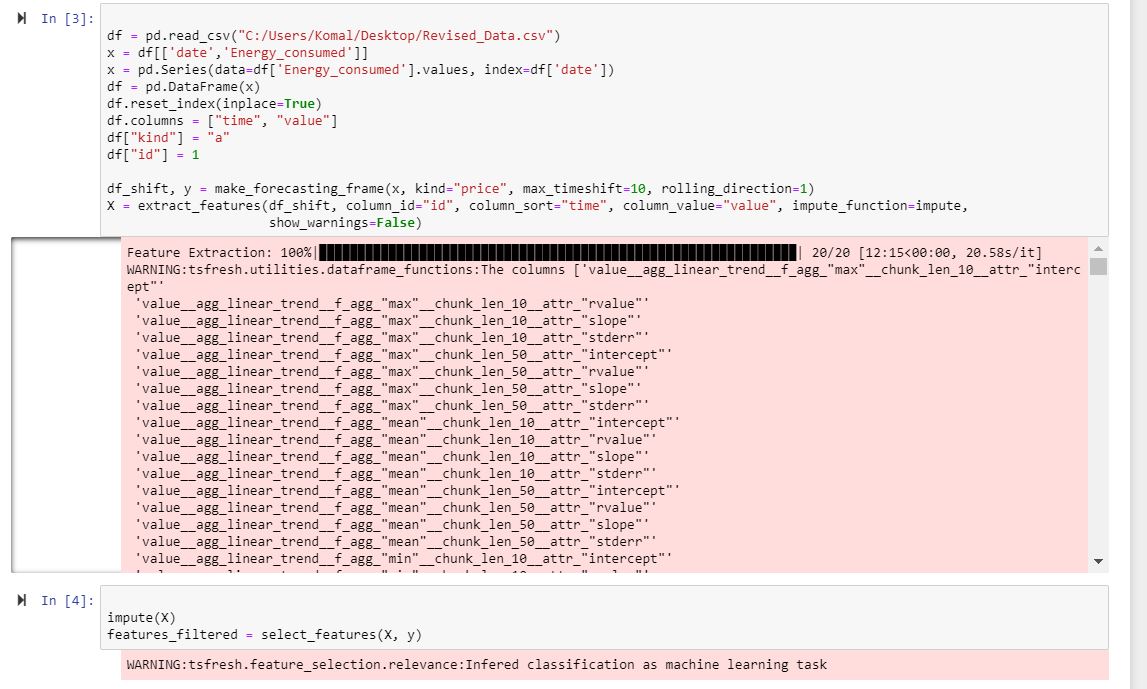


1. Forward and Backward Selection

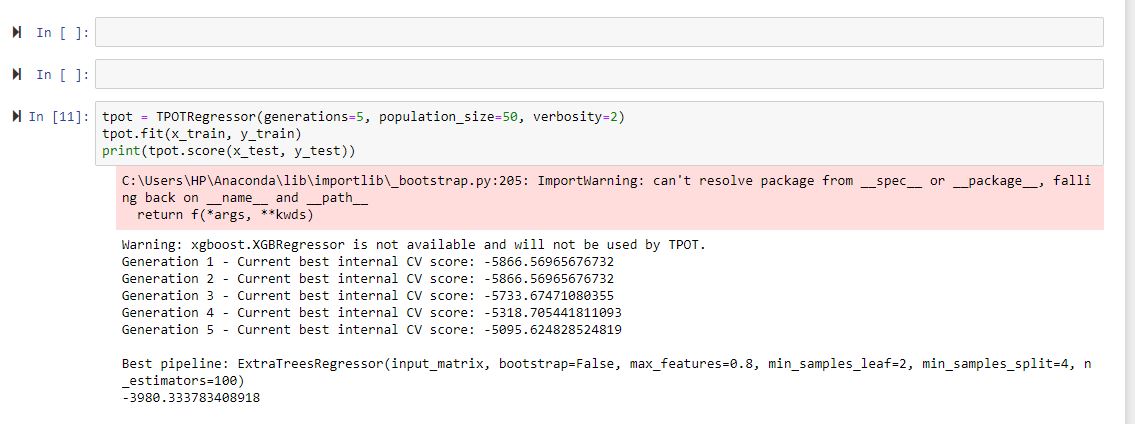




1. Tsfresh



1. TPOT



1. Considering the results obtained in this 5 techniques some of them were not suitable for the data-set and hence we chose the best one to refine our model and get best prediction accuracy through it.

* Model Validation & Selection

1. We have used different model validation techniques to choose the best model for our dataset to predict values:

a) Cross Validation

b) Grid Search

c) Bias Variance Trade-off

d) Regularization

1. After training the models, we have calculated RMSE and R2. The best models are the ones that provide the lower RMSE and highest R2 values.

* Final Pipeline

We have created a pipeline to automate the entire model from data ingestion to final model prediction.