11/4/2018

**Machine learning with Energy**

**datasets**

**Assignment 2- Report**

Team

Ankit yadav

PREETAM JAIN

RITU AGARWAL

**Content**

1. Research
2. Exploratory Data Analysis
3. Feature Engineering
4. Prediction Algorithms
5. Feature Selection
6. Model Validation and Selection
7. Final Pipeline
8. Summary

**1. Research Paper Analysis**

**RESEARCH PAPER 1:**

**RESEARCH PAPER 2:**

**RESEARCH PAPER 3:** Prediction of appliances energy use in smart home

Link - <https://www.sciencedirect.com/science/article/pii/S0360544212002903>

This paper has been written with an aim to predict the energy consumption in household for the next day. To achieve this, they have collected data of homes of France and have analyzed and have come up with certain predictors. It has been studied that residential sector is the biggest sector in electricity consumption. And it is required that we understand the pattern of electricity consumption in household so that Industries can generate and transfer only that amount of energy to the household area to better circulate power. The energy market is divided into distinct categories, but the Day Ahead Market or Spot Market is of great interest. This type of energy market involves bidding the energy consumption of the next day. It is a very complex mechanism, which requires a very good knowledge of the demand for the power suppliers There were lots of theories which were proposed but it is important to understand the each and every criterion like number of appliances, usage of these appliances, day of week, etc. So, this paper concentrates more over discrimination of usage of electricity on appliance level which would make things easier to understand the pattern of usage of electricity over the course of time. In order to get a better load control, the energy prediction has to go down from total household energy consumption to electrical device consumption. The concept of smart grid has been introduced to tackle power system challenges. Smart grid initiatives seek to improve operations, maintenance and planning using modern technology to better manage energy use and costs. This would help industries to smartly circulate the generated electricity to different industry which would be much more efficient than present method. There have been lot of expectations which was not getting met as the usage of appliance differ over period of time on daily basis. A reliable model was required as usage of appliances on peak time was different as compared to other times. Thus, they came up with a concept called demand dispatch which is ability to control individual loads in precise manner at all the times and not only during peak times. This load management id of two types.

• **Direct Control**: This method refers to classical method of load control which involves  
increasing the energy production in case of higher load demand.

• **Control by cost**: This method refers to change the load curve shape in such a way that energy consumption peak decreases, even though the total energy consumption for the specific house stays the same. When it has been understood that we have to consider the usage of appliances to get a better picture of electricity consumption and load balancing, there are 4 different type of predictors which can be considered to calculate the same.  
8  
• **The “will always consume” predictor**: According to this predictor, we assume that an appliance is always running and consuming electricity.

• **The “will never consume” predictor**: According to this predictor, we assume that an appliance is not at all being used and is not consuming electricity.

• **The ARMA predictor**: ARMA stands for Autoregressive Moving Average. According to  
this method current value of a time variable is assumed to be a function of its past values  
and it is expressed as a weighted sum (moving average).

• **The proposed predictor**: According to this model, an inhabitant in the house interacts  
with various electrical devices as part of his routine activities. Thus, energy consumption  
can be modeled as a process which is having a random probability distribution or pattern  
that may be analyzed statistically but may not be predicted precisely.

Improving the precision of prediction is highly necessary. It is important for us to understand the pattern of usage of electricity. The segmentation of data can be made considering various aspects such as the season, month, period of the day (day/night), type of day (weekday/weekend). The objective of this operation is to reduce the average dispersion to improve the prediction. In such conditions, k-means clustering method can be precise to cluster similar data together. At last, I would like to conclude by saying forecasting the energy consumption in homes is an important aspect in the power management of the grid, as the consumption in the residential sector represents a significant percentage in the total electricity demand. The development of the smart grid is not possible without a good prediction of energy consumption. The trend nowadays is to get the prediction of energy consumption not only at house level, but at household appliance level. The prediction of energy consumption in housing is very dependent on inhabitants’ behavior, so a stochastic method for prediction has been presented in this paper.

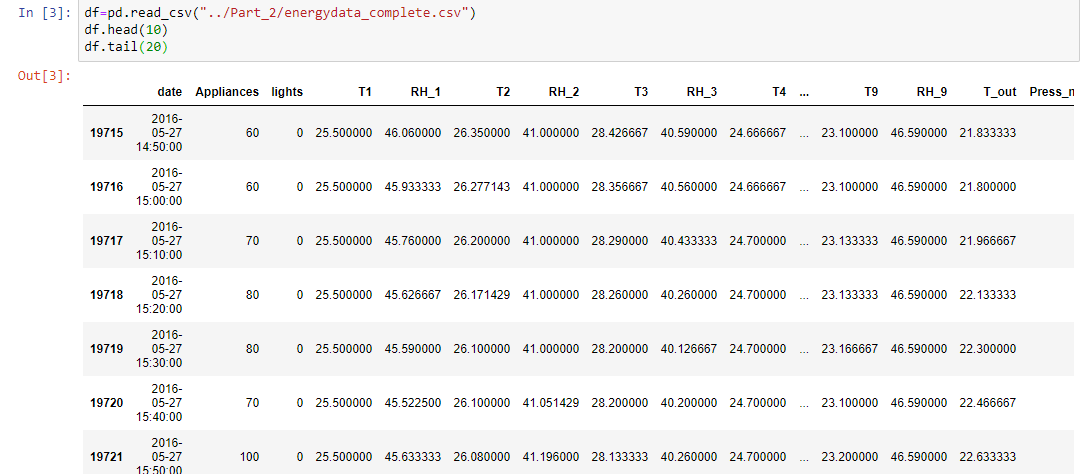
**2. Exploratory Data Analysis**

We imported following libraries

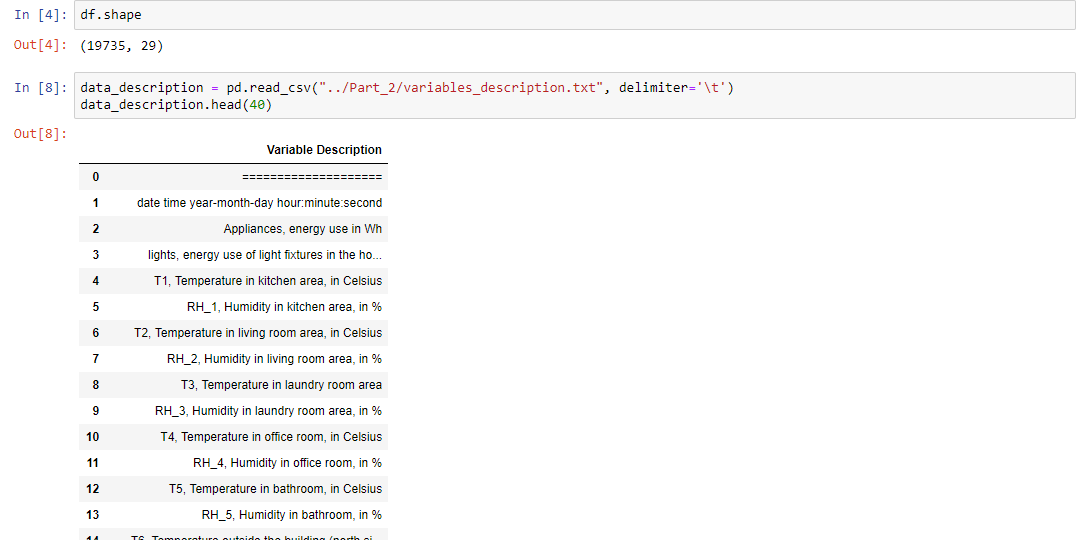


Then we read the data and viewed its head.

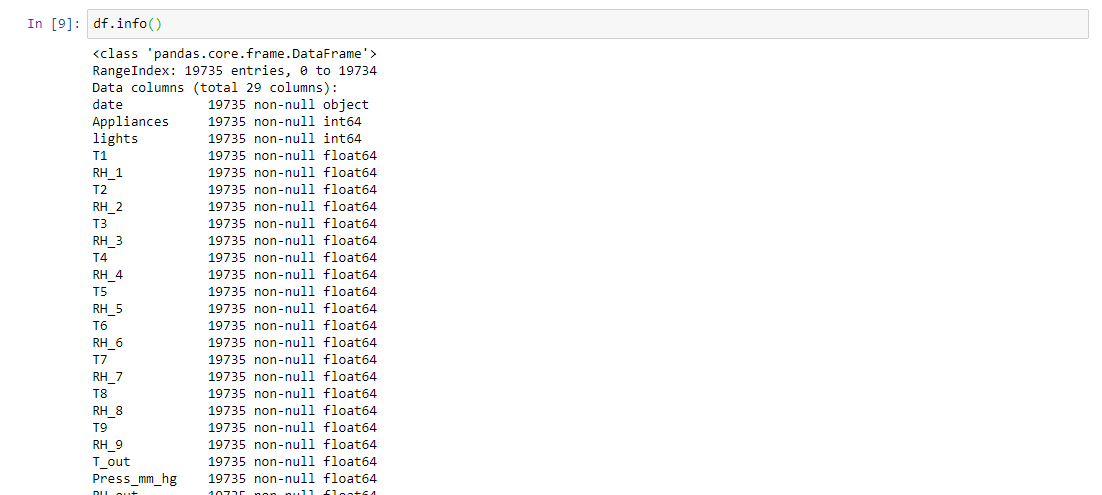
There were mainly variables for temperature and humidity at various parts of house.



We also checked its shape along with the variable information.



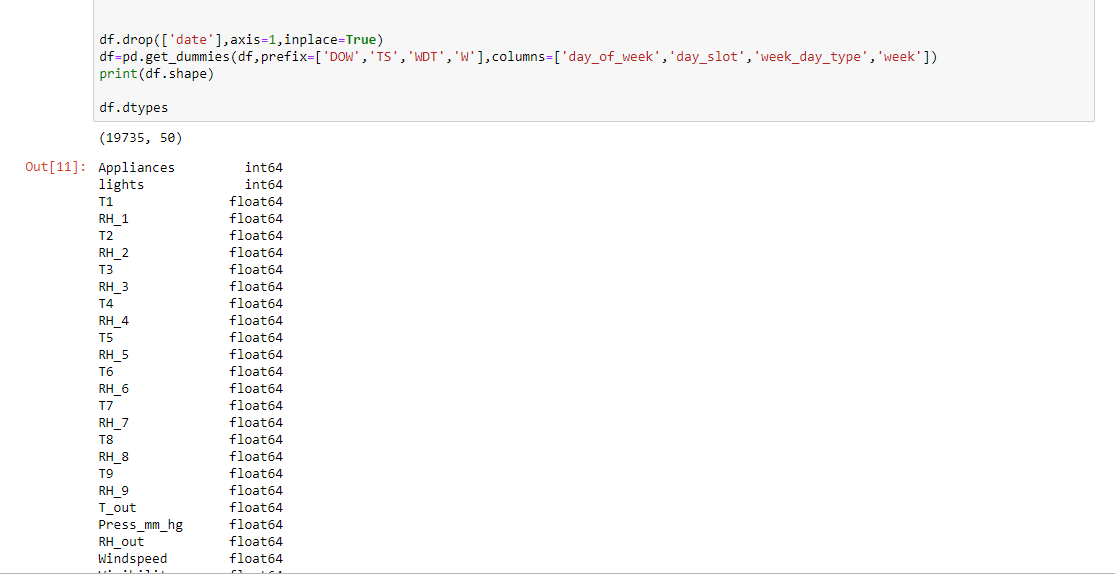
We also needed to check the variable information for further use.



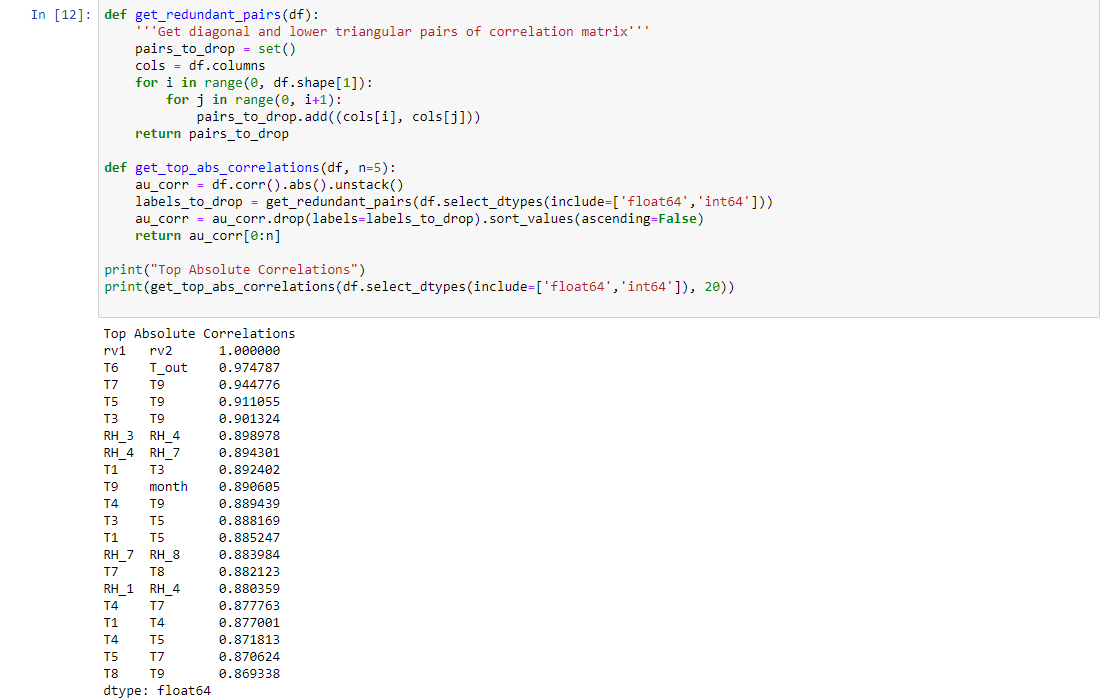
Then we change the date time according to month, week and time of day.



According to that we added new columns into our data frame



After this we checked the correlation check for redundant variable .



We found that rv1 and rv2, t6 and t\_out etc are highly correlated .

So, we removed these variables from our data.

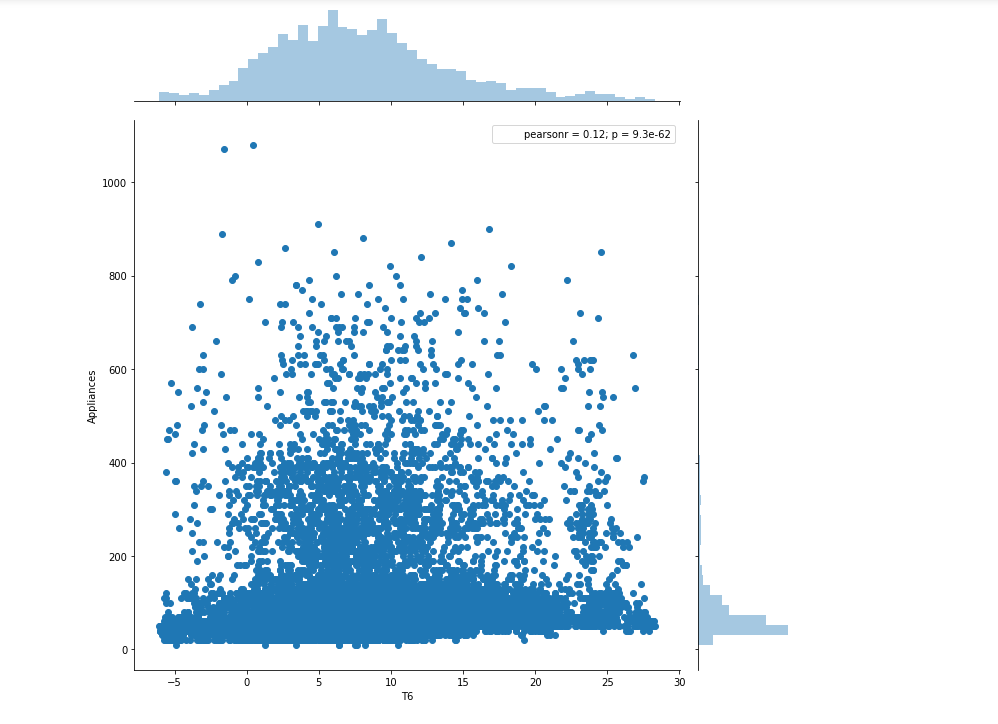


We also checked by using heatmap of variables.

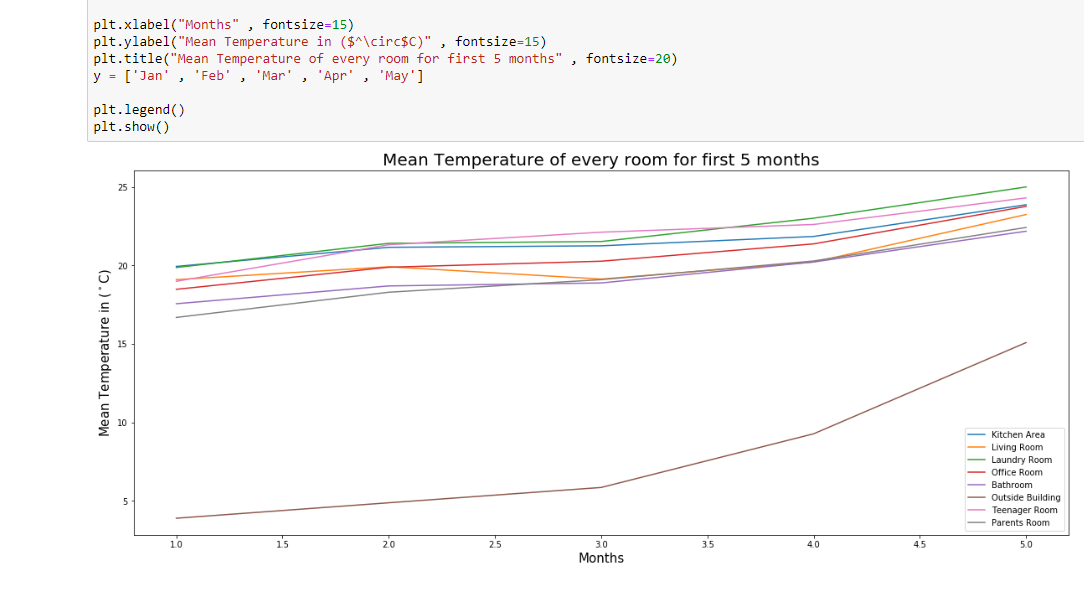


We also made following plot.

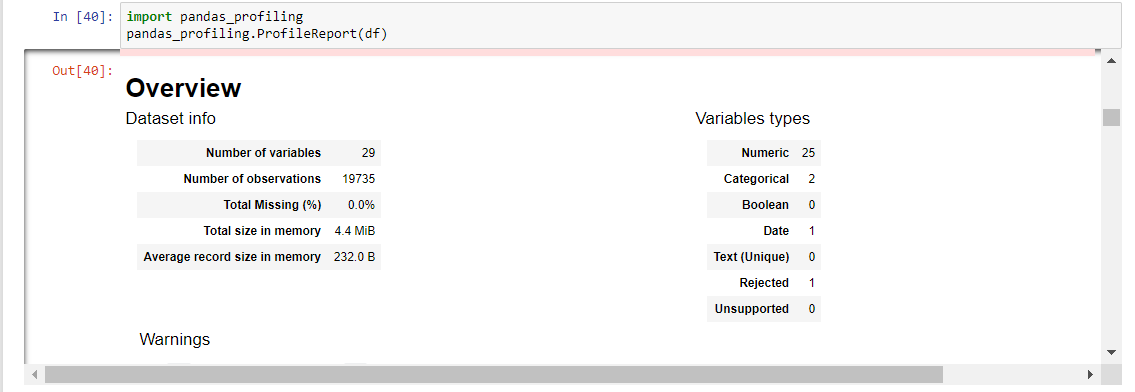


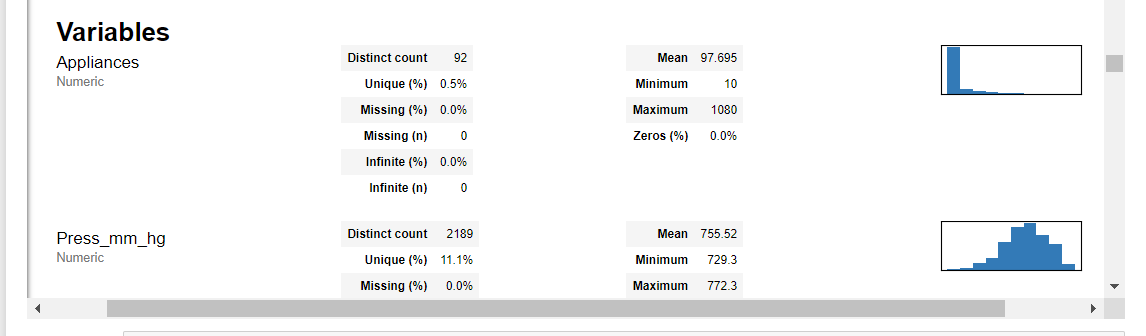


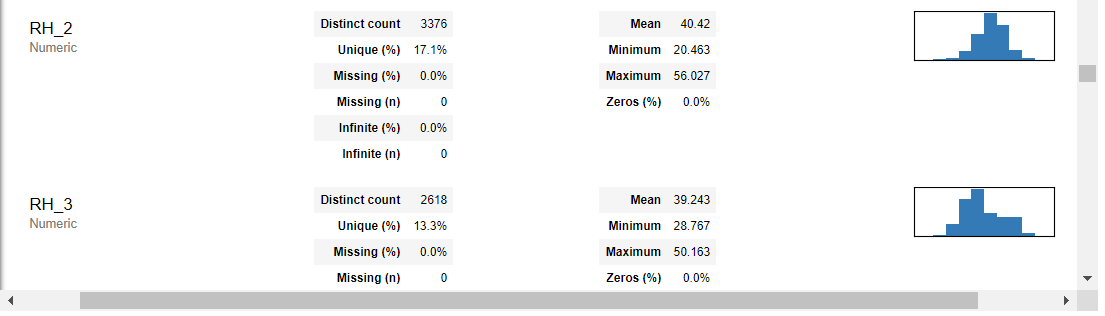
We also plotted the graph to look at the changes in temperature over time for various appliances

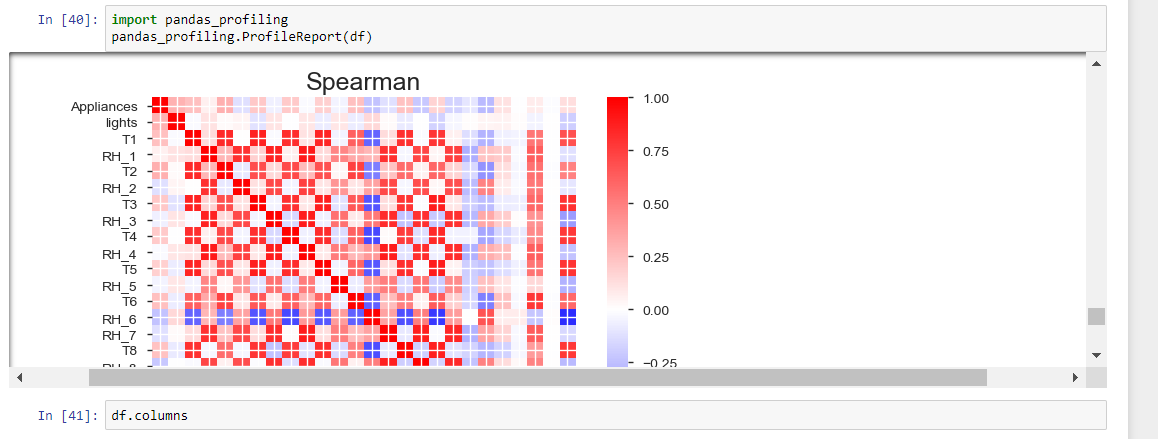


We also used pandas profiling to check for various variables.



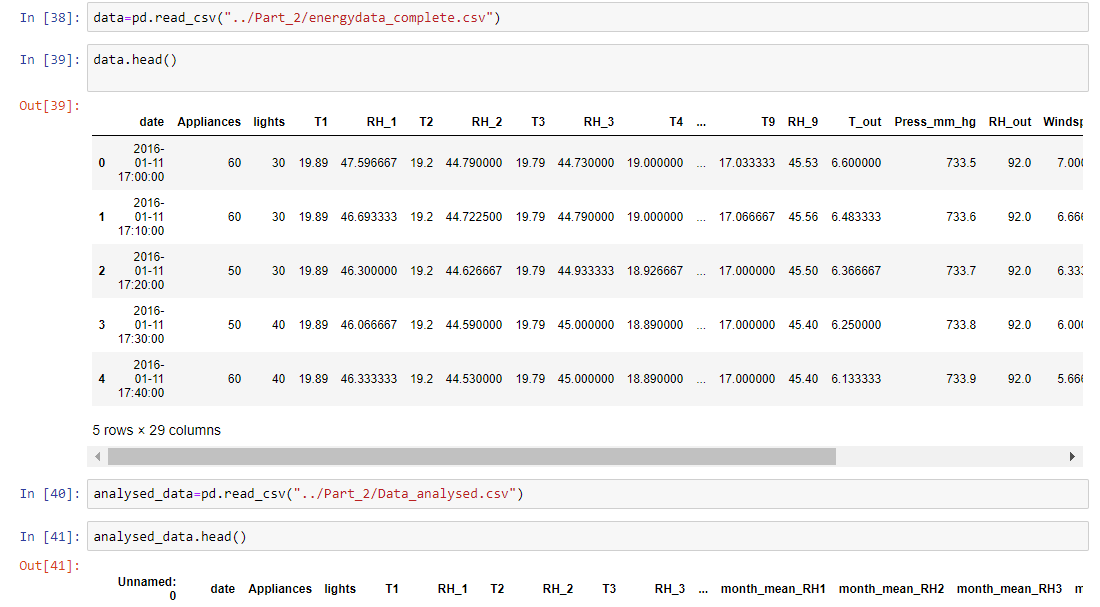




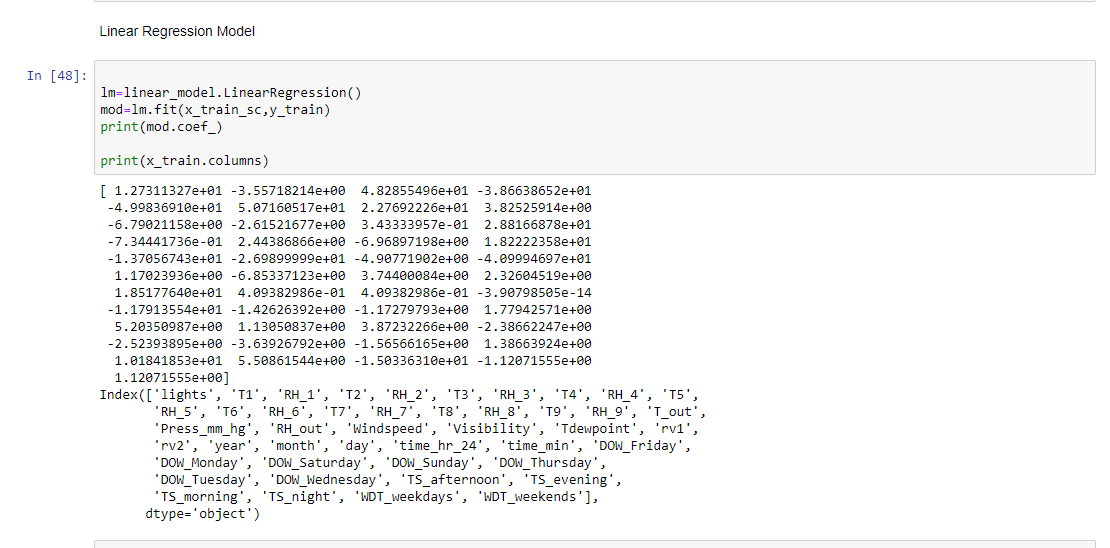


**3. Feature Engineering:**

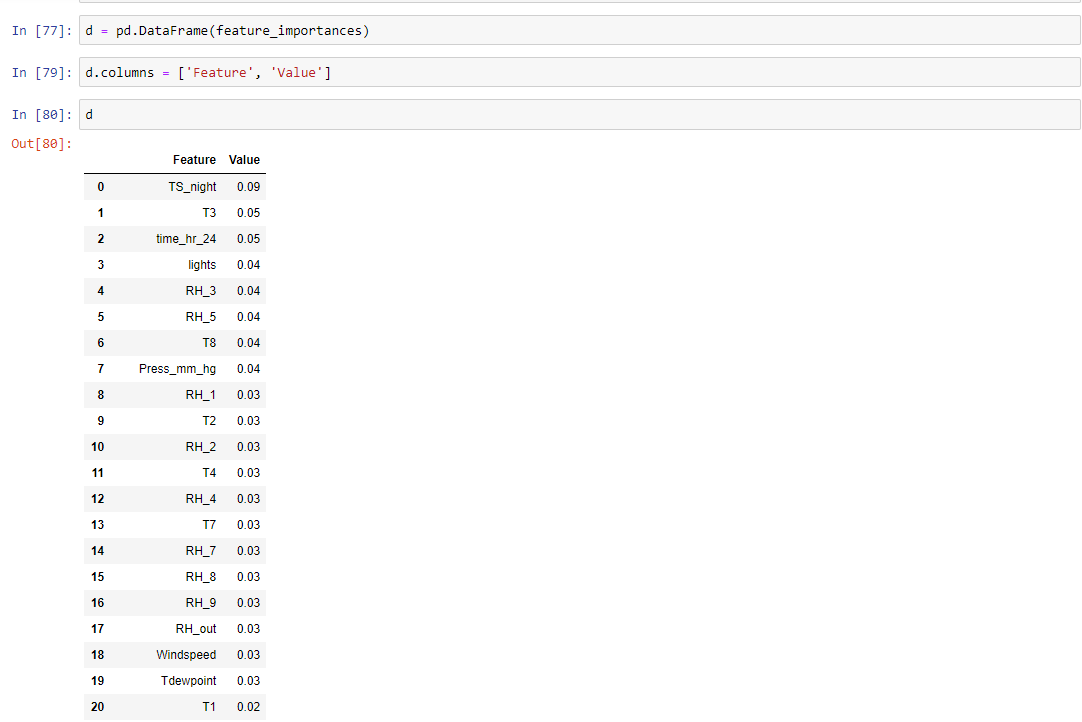
There are lots of features in our dataset. Temperature of 8 different rooms are recorded in degree  
Celsius. Humidity of 8 different rooms are recorded in percentage. Outside temperature is also  
recorded in degree Celsius and humidity in percentage. Along with it pressure has been recorded  
in millimeter scale in mercury, visibility in kilometer, dew point in degree Celsius and windspeed  
in m/s.  
Before performing any kind of test, we have to analyses the data and look into it.  
We also have to understand how much data does the file contains.  
It is important for us to analyze whether there are null values in the dataset and how are values  
distributed.



We also did linear regression on the data.



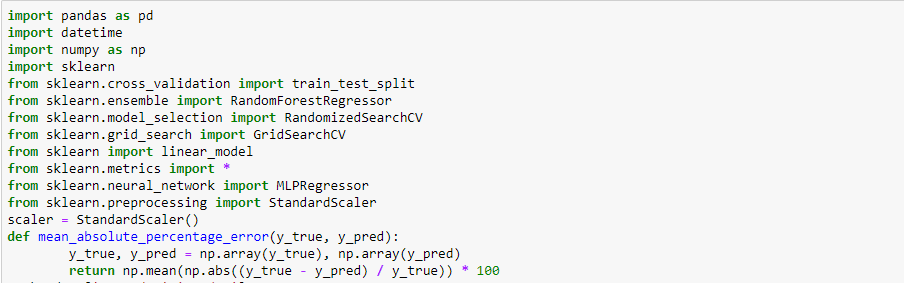
Finally, We found the importance of various features in our dataset.



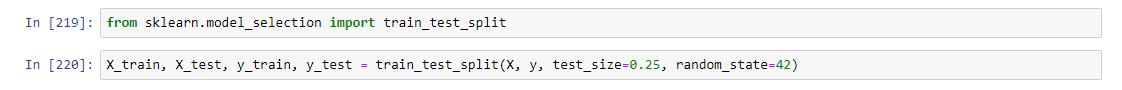
We also plotted a graph to check the same.



**4. Prediction Algorithms**

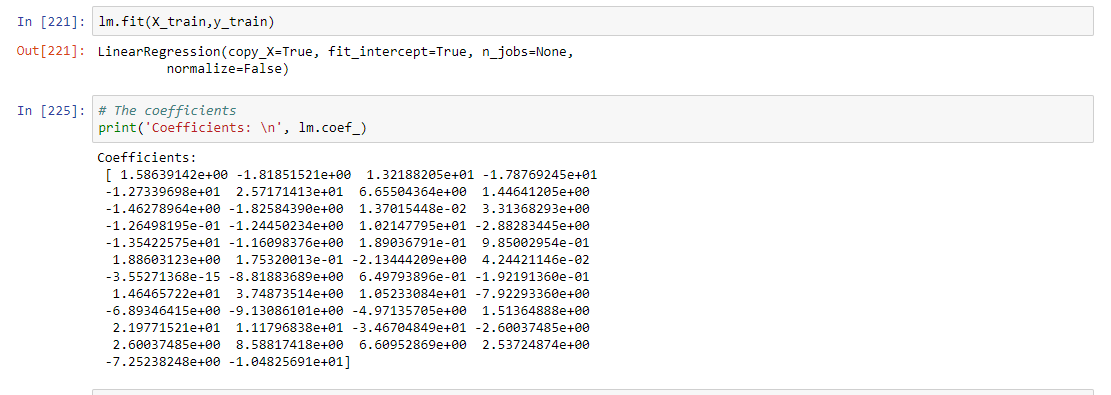


Splitting data

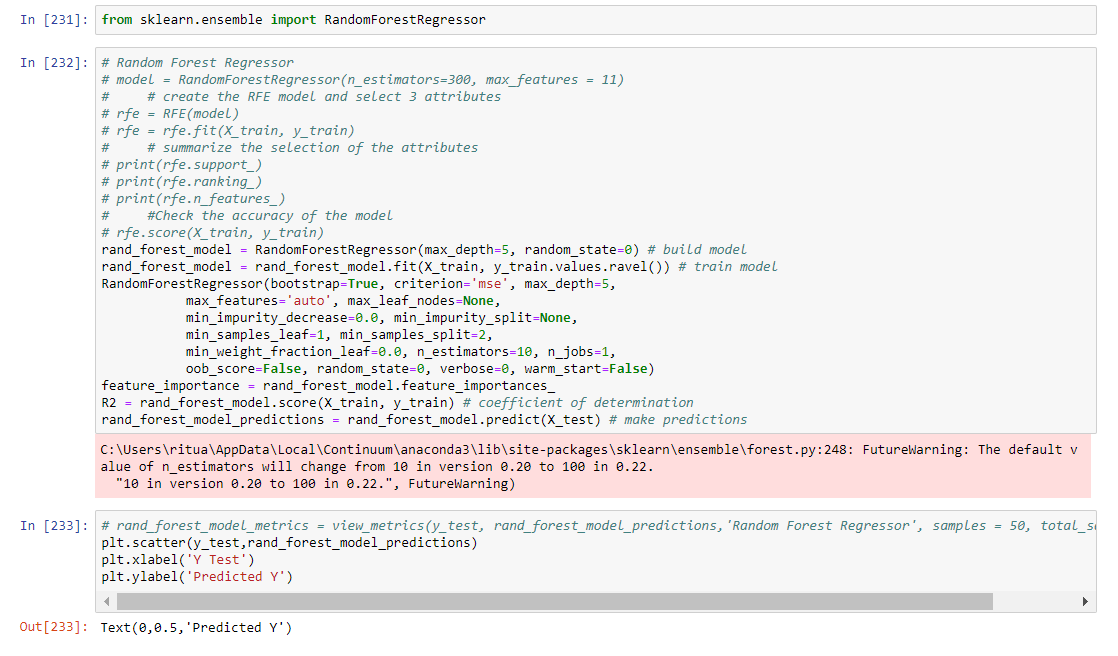


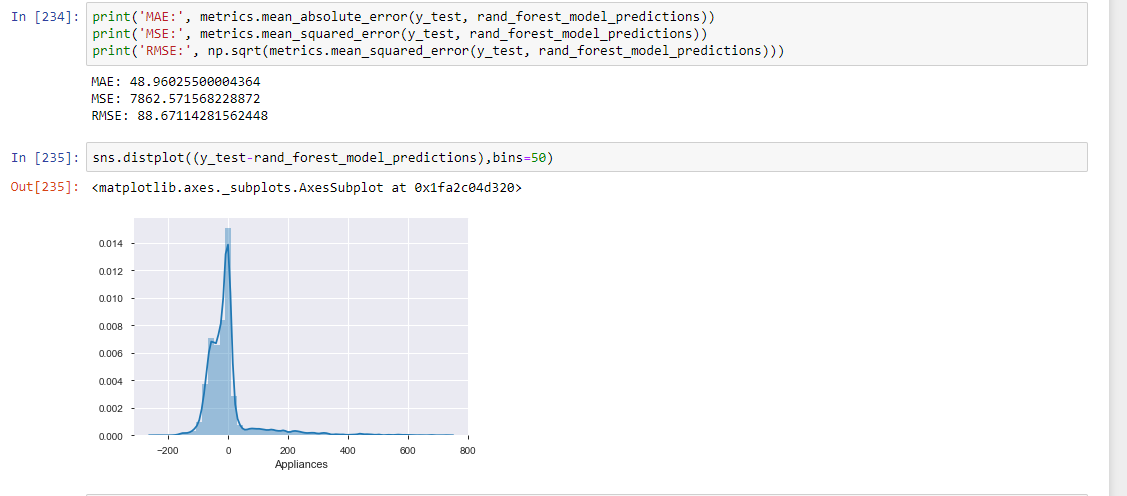
Linear Regression





Using Random Forest





Neural Network Models



**5. Feature Selection**

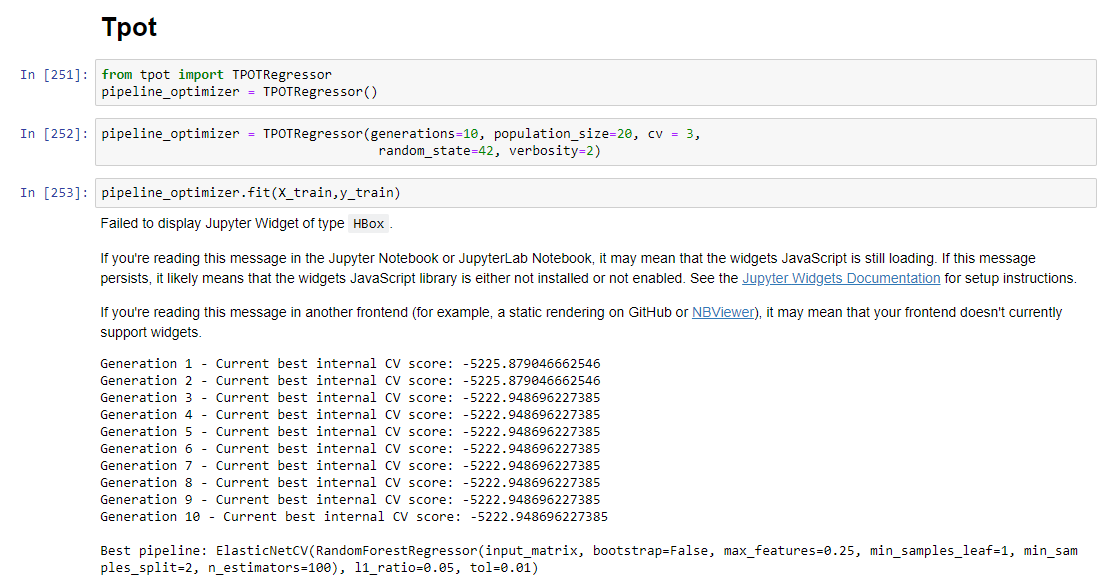
In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features variables, predictors) for use in model construction. Feature selection techniques are used for four reasons:

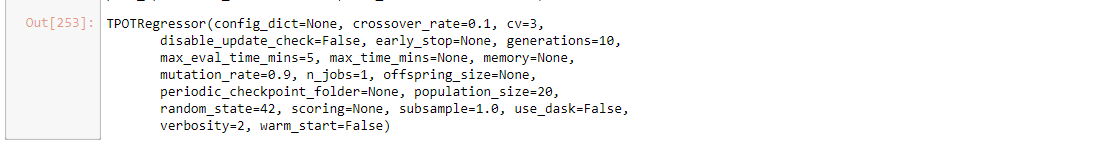
• simplification of models to make them easier to interpret by researchers/users,  
• shorter training times,  
• enhanced generalization by reducing overfitting (formally, reduction of variance)

The central premise when using a feature selection technique is that the data contains many  
features that are either redundant or irrelevant, and can thus be removed without incurring  
much loss of information. Redundant or irrelevant features are two distinct notions, since one  
relevant feature may be redundant in the presence of another relevant feature with which it is  
strongly correlated.

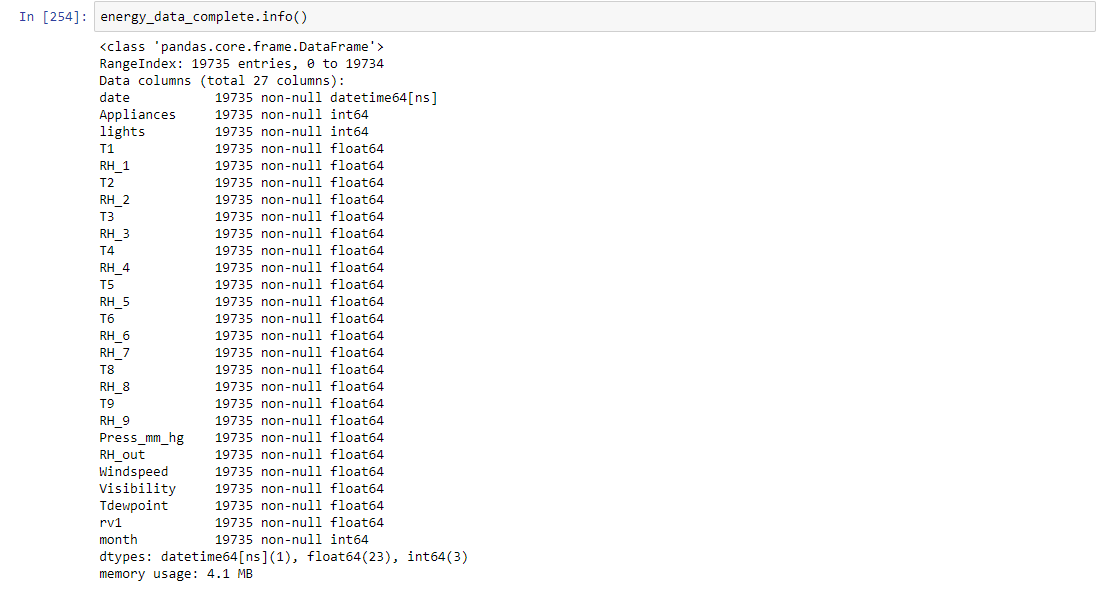
**TPOT:**

The Tree-Based Pipeline Optimization Tool (TPOT) was one of the very first AutoML methods  
and open-source software packages developed for the data science community. The goal of TPO   
is to automate the building of ML pipelines by combining a flexible expression tree representation of pipelines with stochastic search algorithms such as genetic programming. TPOT makes use of the Python-based scikit-learn library as its ML menu





Info of model



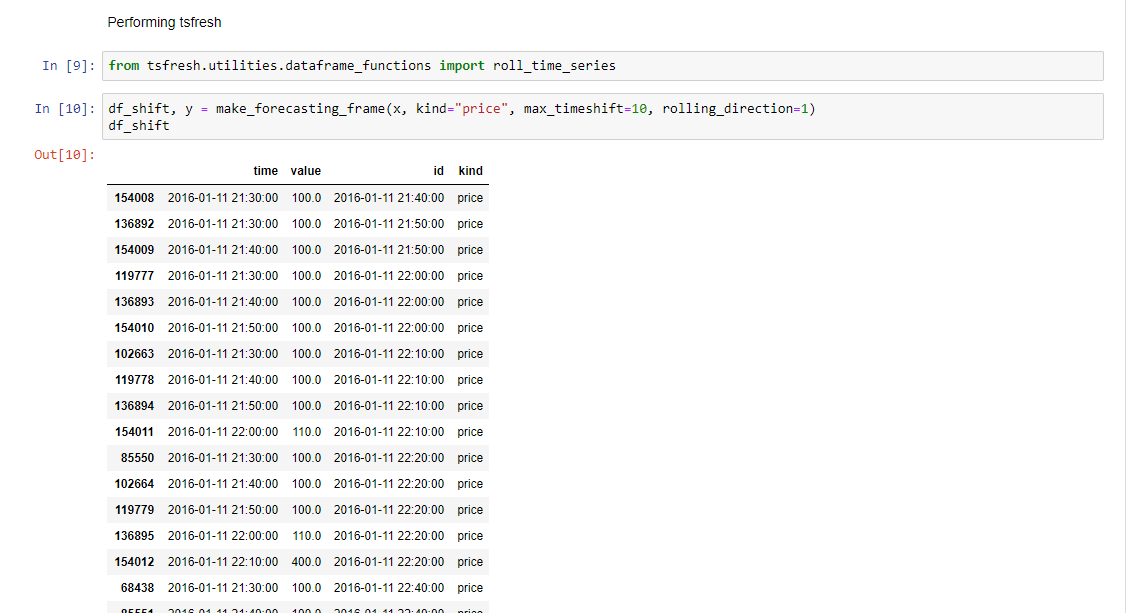
When we get the output, we check the accuracy of score.

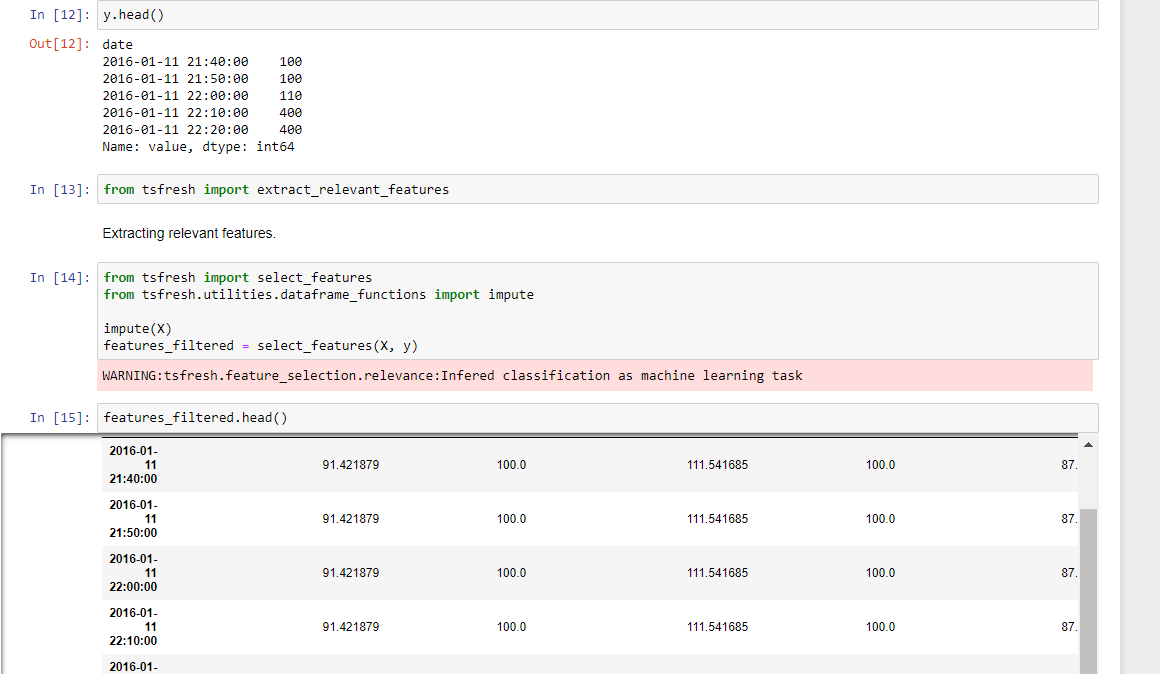


**TSFresh**

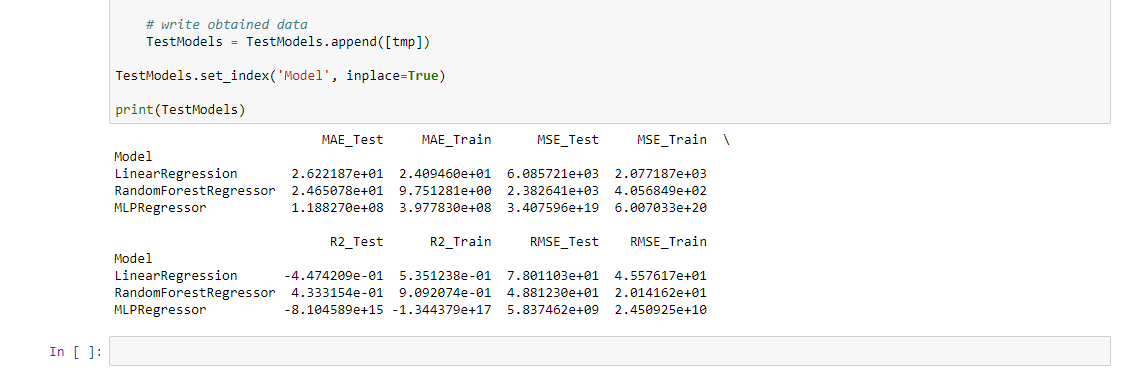
tsfresh is a python package. It automatically calculates a large number of time series characteristics, the so called features. Further the package contains methods to evaluate the explaining power and importance of such characteristics for regression or classification tasks.

We start by creating the model for testing.





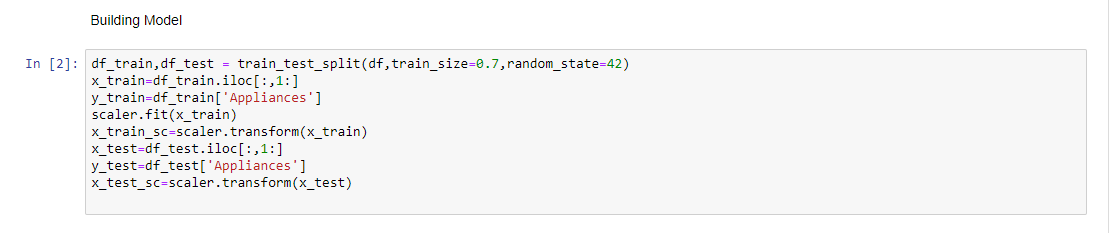
Getting final accuracy.



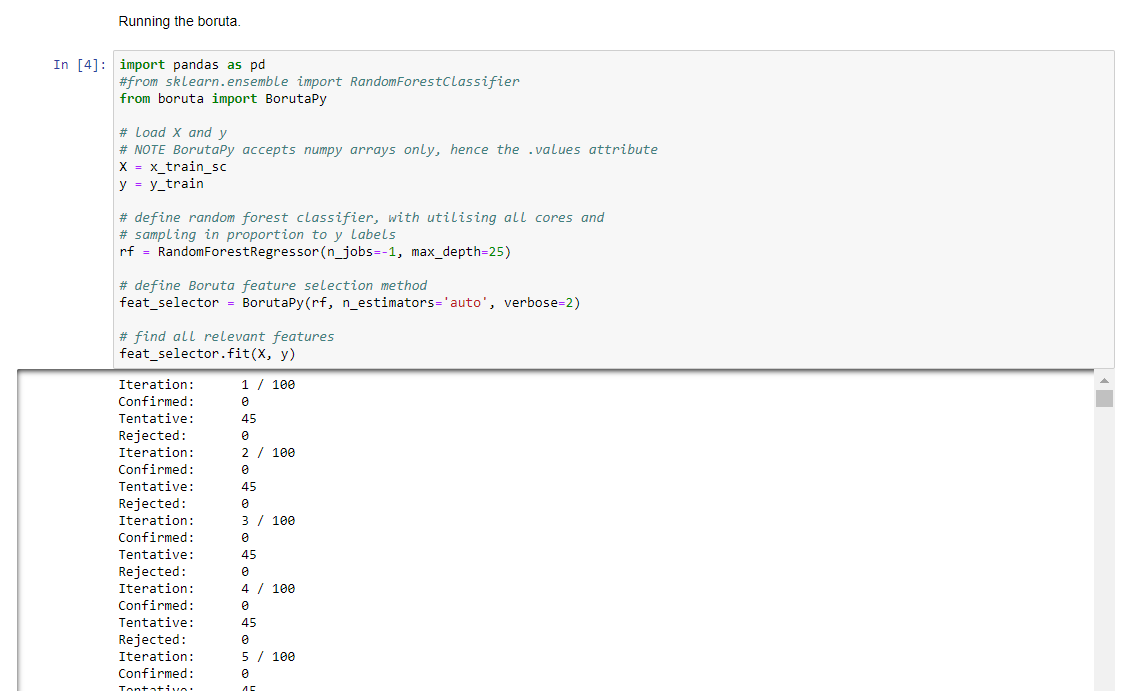
**Boruta**

An all relevant feature selection wrapper algorithm. It finds relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies (shadows).

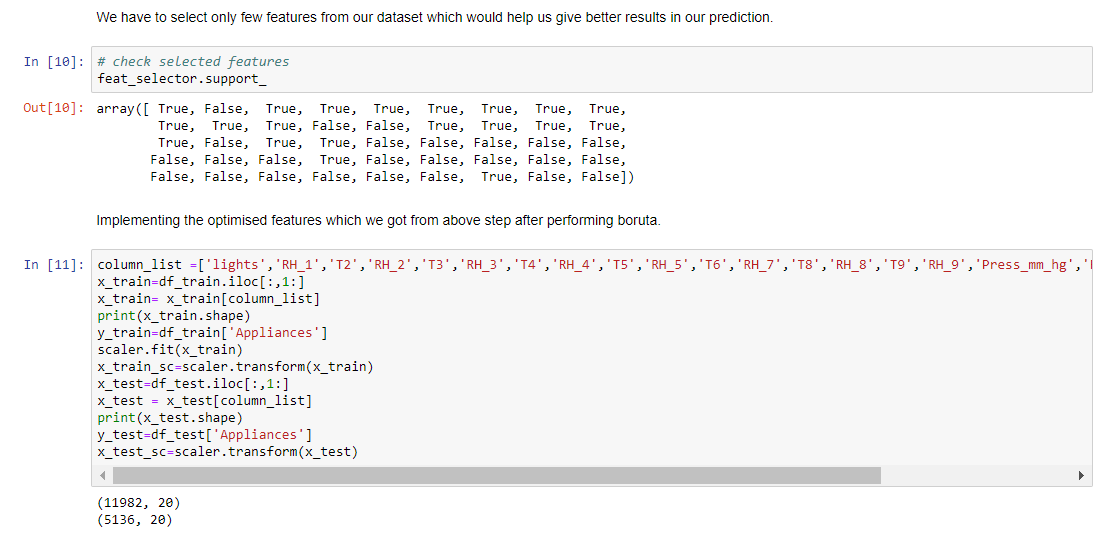
First, build the model.



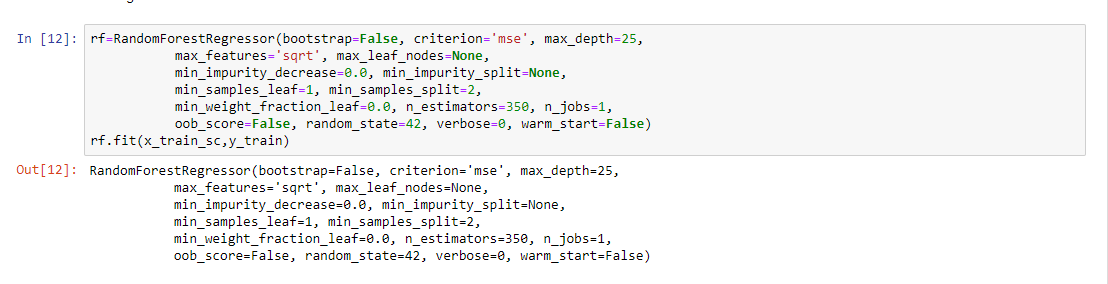
Running Boruta on this model



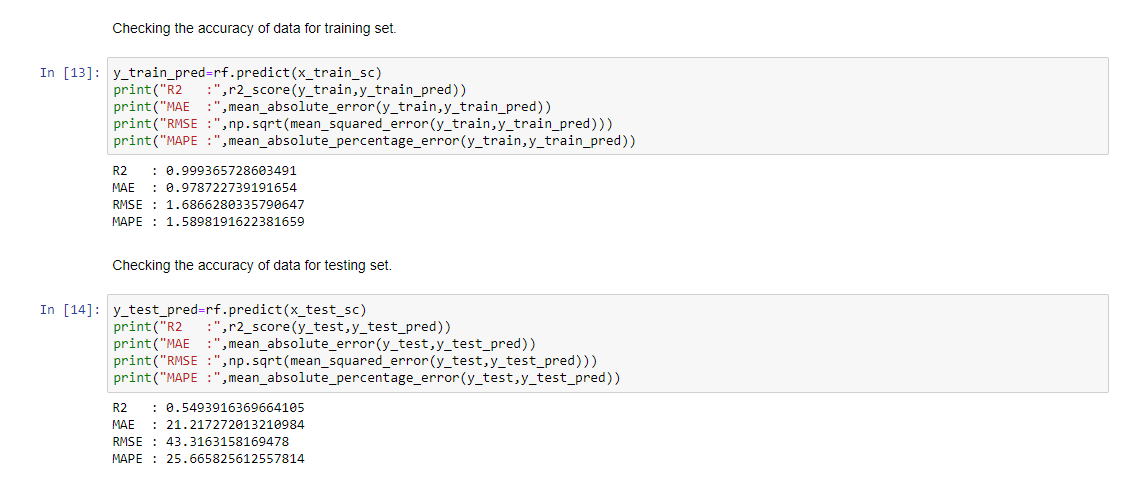
We need to select only helping features from our model.



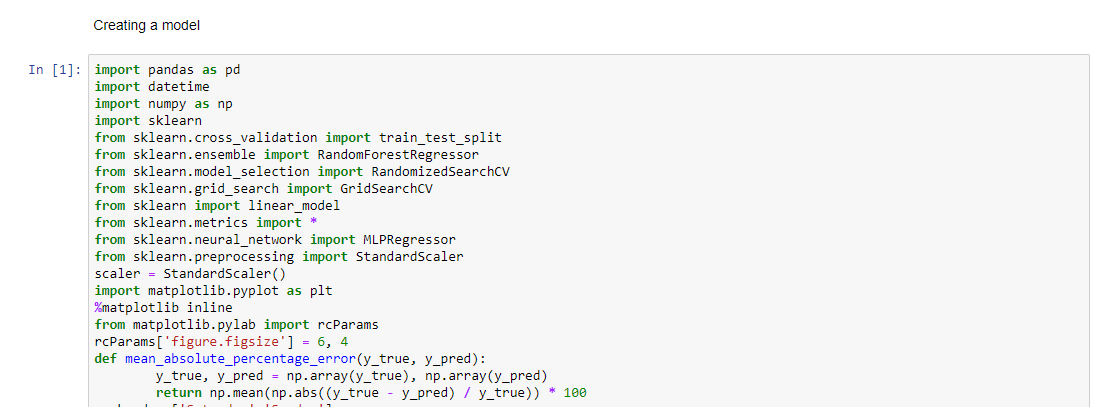
Making a model for random forest

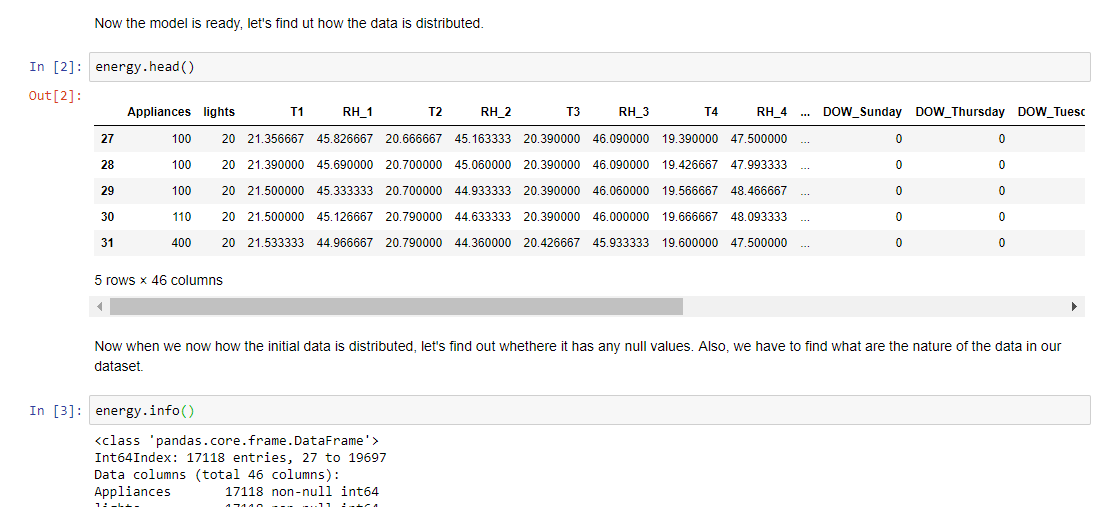


Checking the accuracy of training and testing data



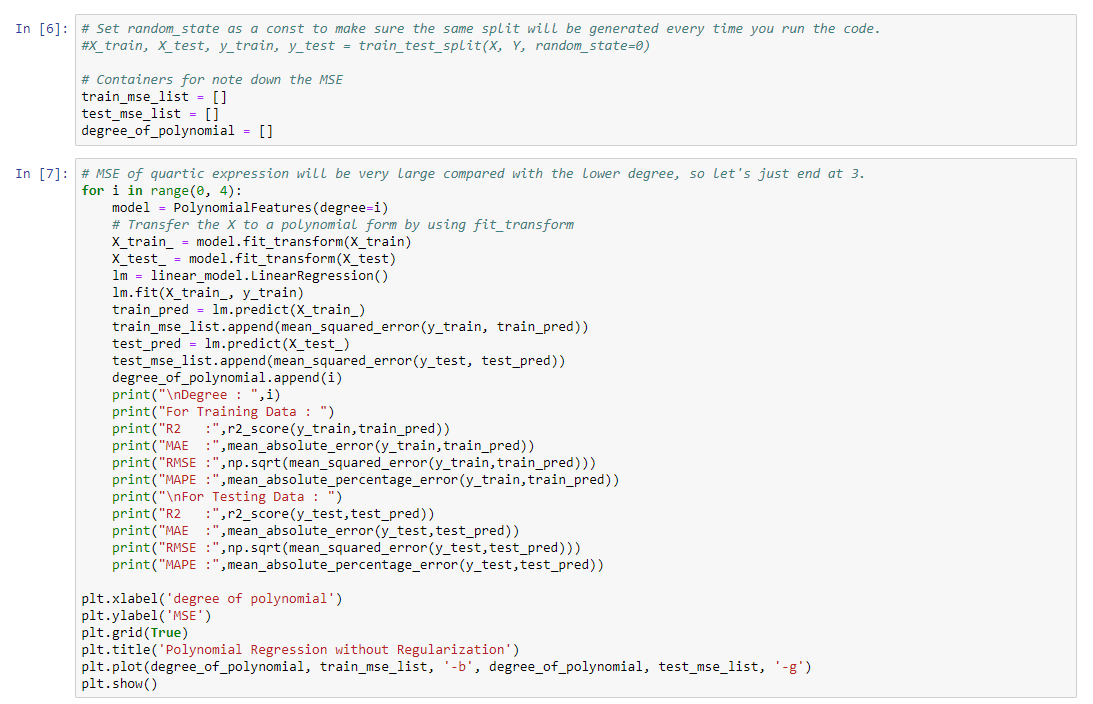
**6. Model Validation and Selection**

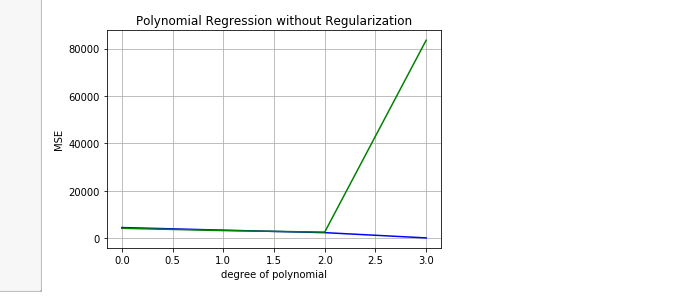






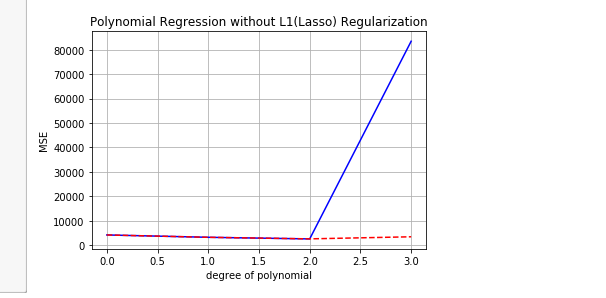
Regularization





Performing Lasso





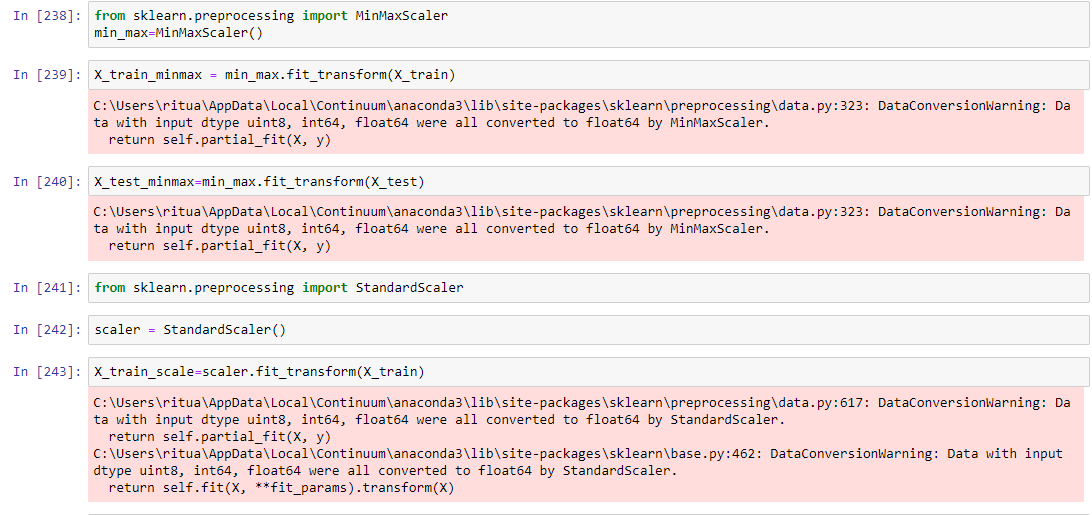
Cross Validation Technique



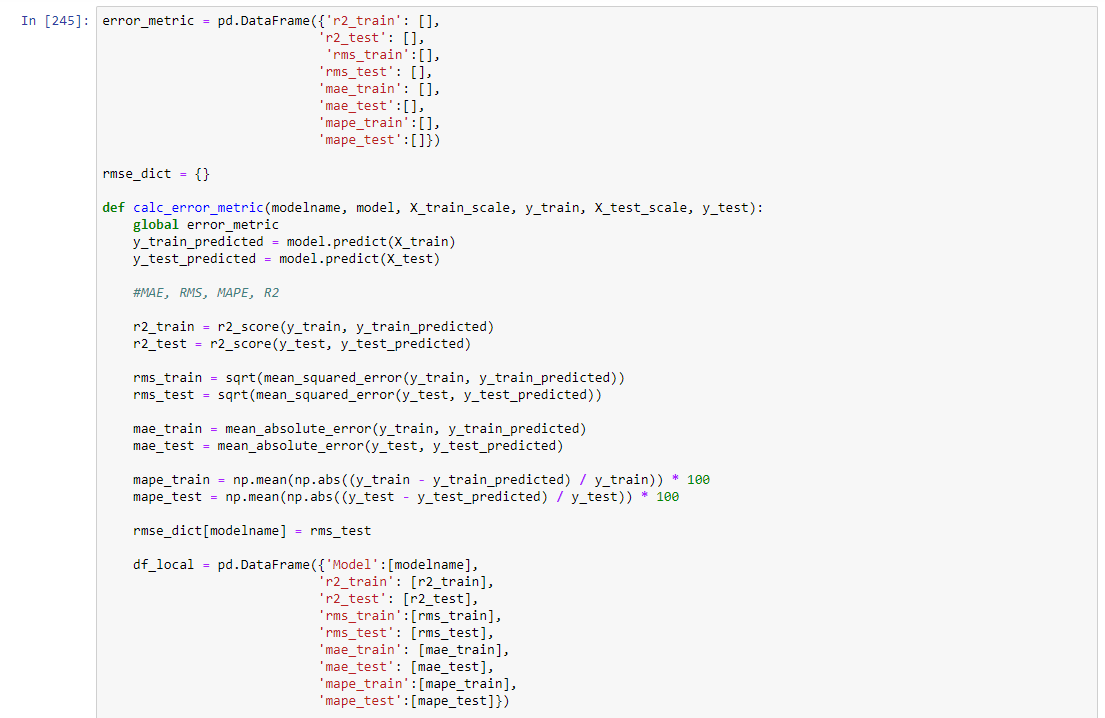
Getting Accuracy of model



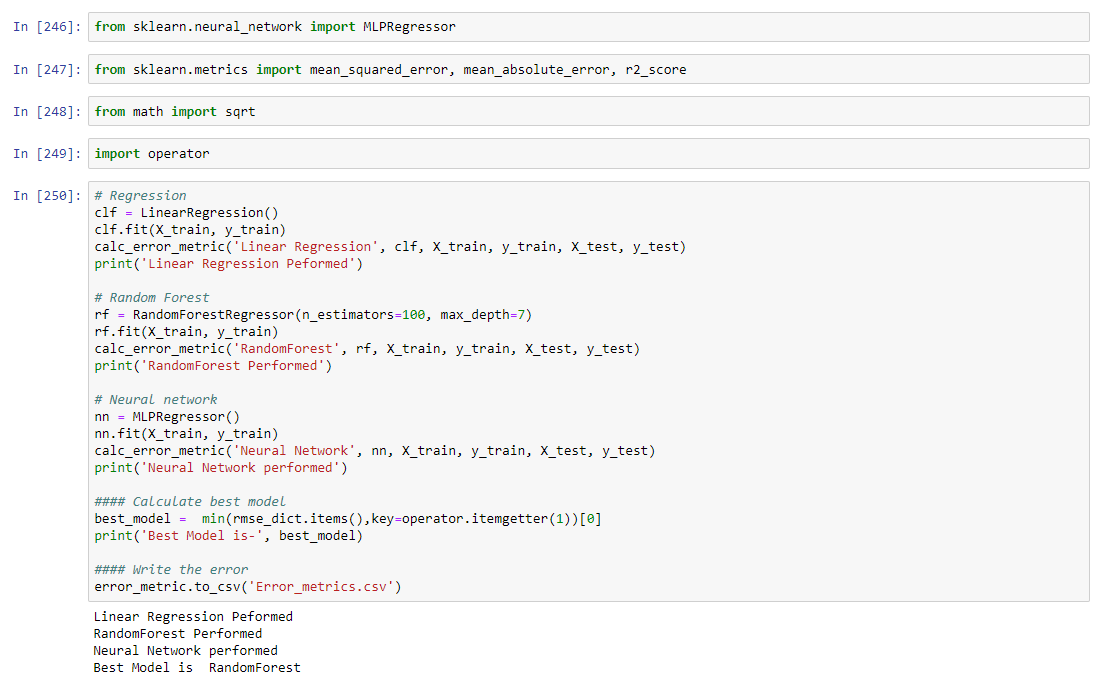
**7 Final Pipeline**

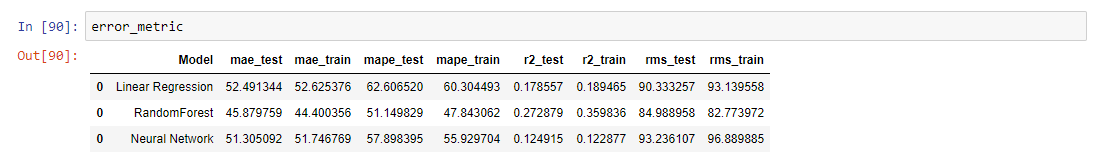


Error metric

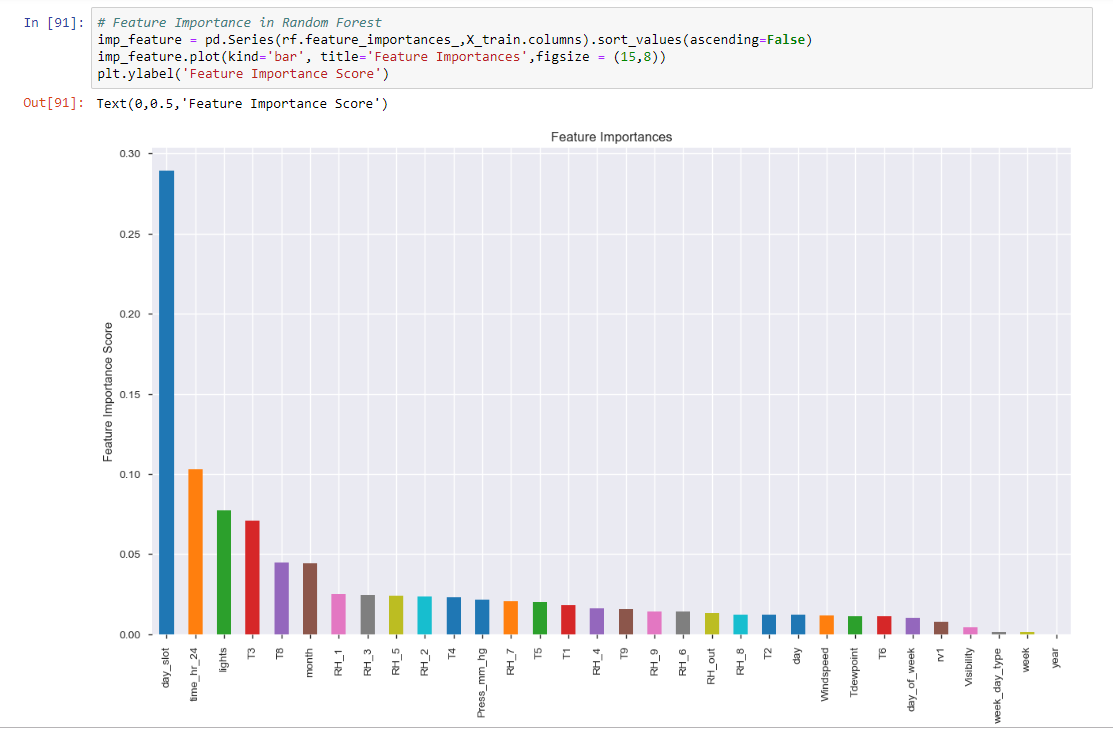


Getting best model.

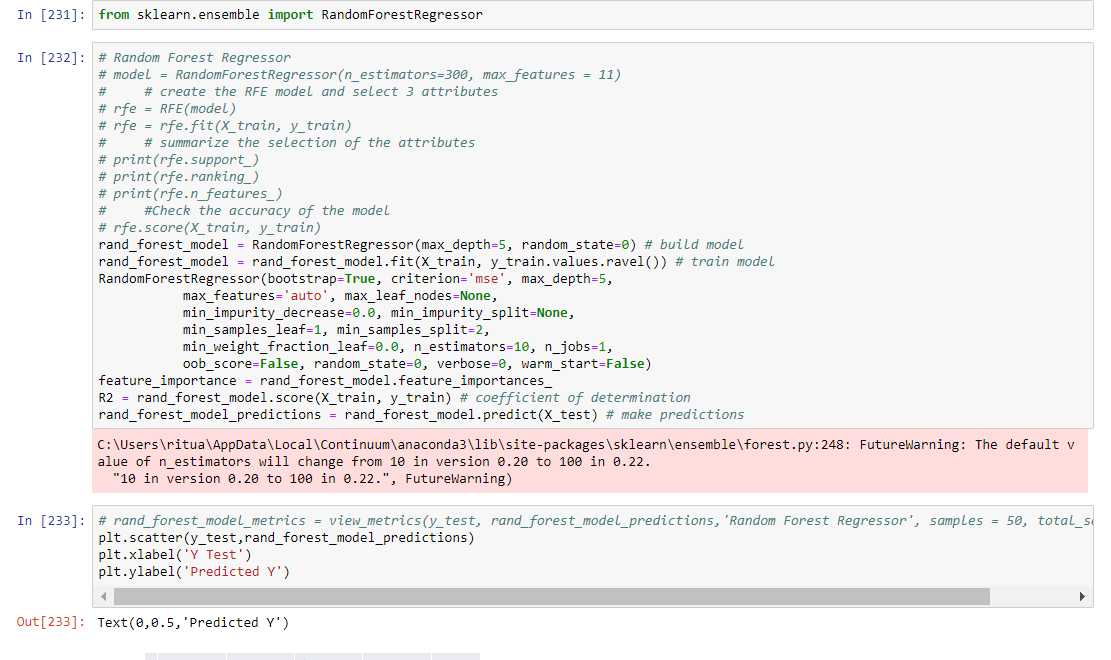


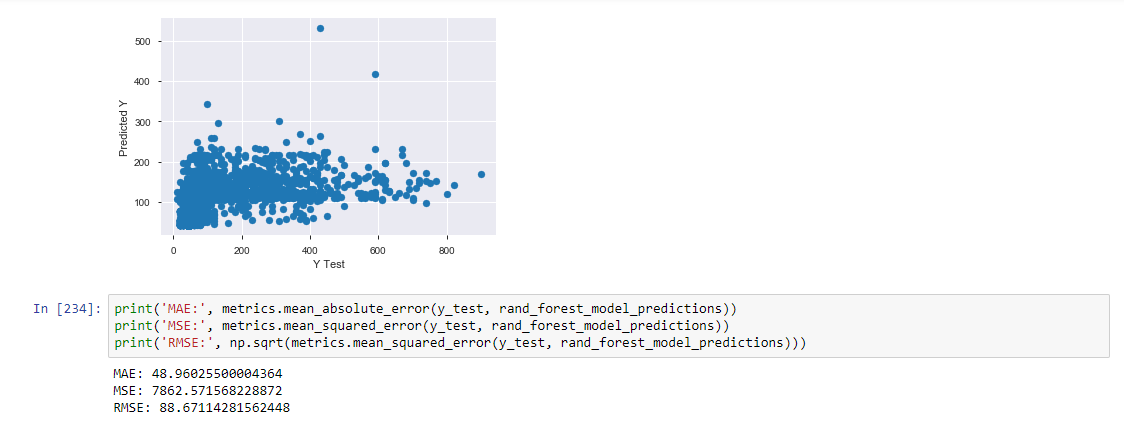


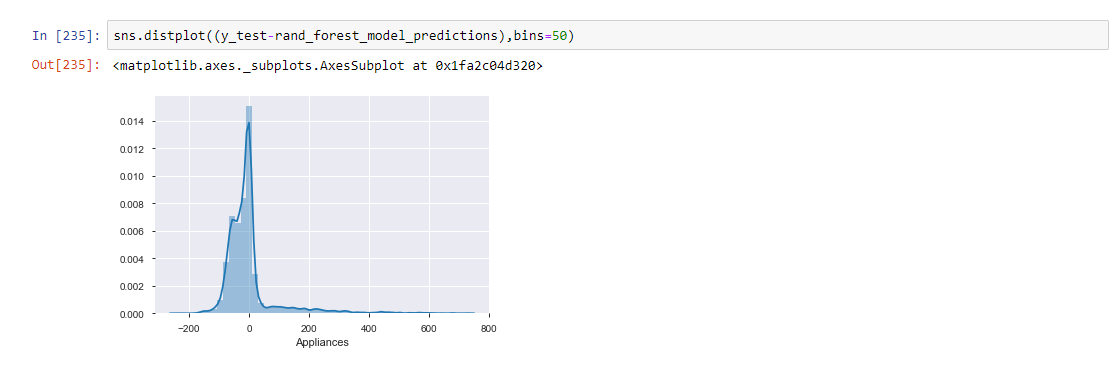
Getting feature importance of random forest



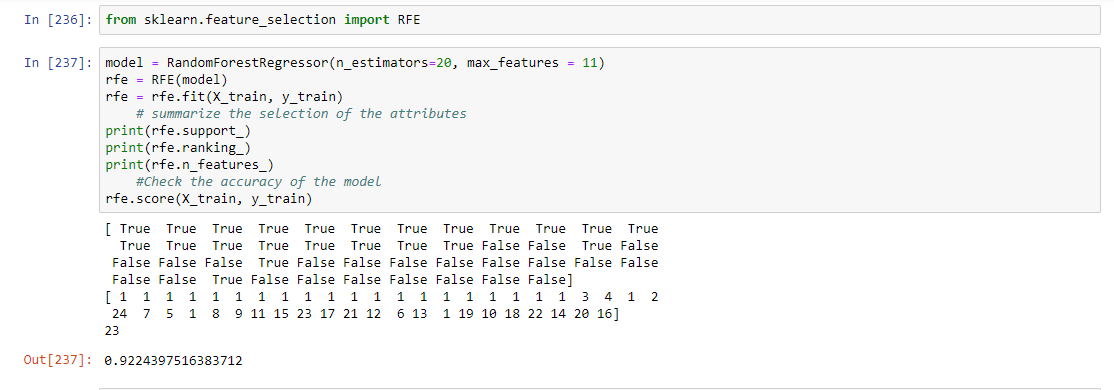
Doing prediction by random forest.



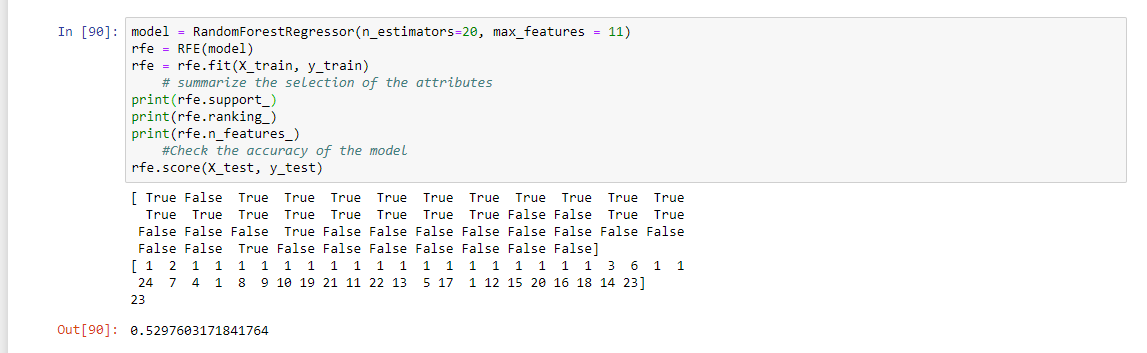




Predicting using Random forest algorithms on train data



Now, using the same on our testing data.



**So, our model gives 52.9 % accuracy.**

**8. Summary**

By the above analysis we have concluded following points regarding the data given to us.

* Best model to analyze and predict is Random Forest.
* There are many columns which are highly correlated, and they need to be removed in order to get good prediction
* The data have almost none outliers and no NULL valued column. So, the data is almost clean.