11/4/2018

**Modeling of Energy Data**

**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:**

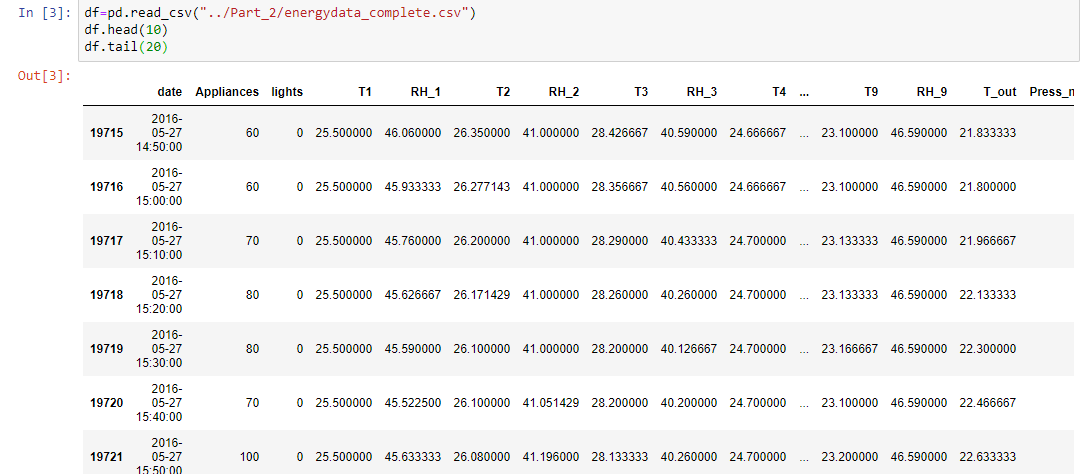
**RESEARCH PAPER 3:**

**2. Exploratory Data Analysis**

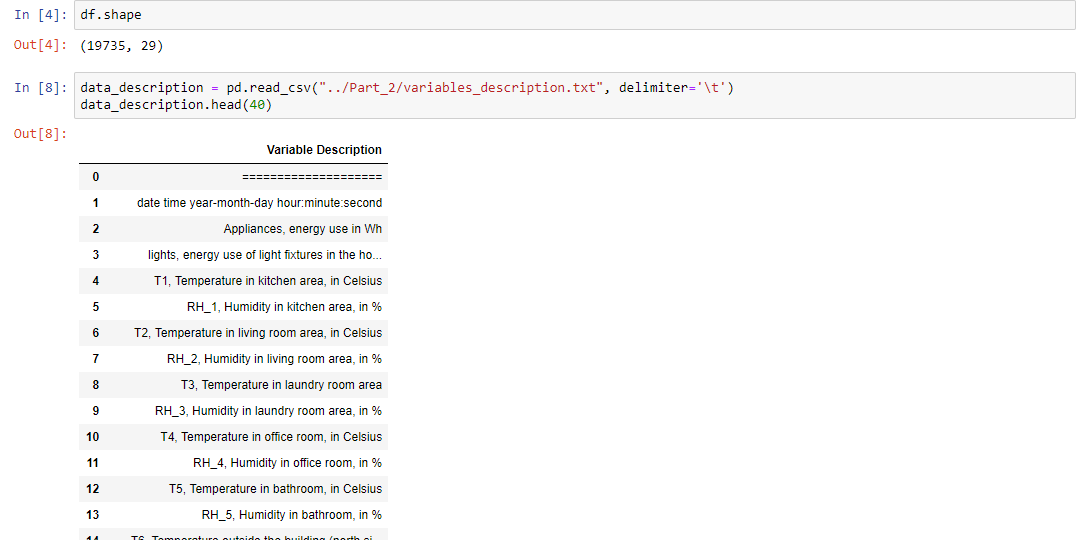
We imported following libraries



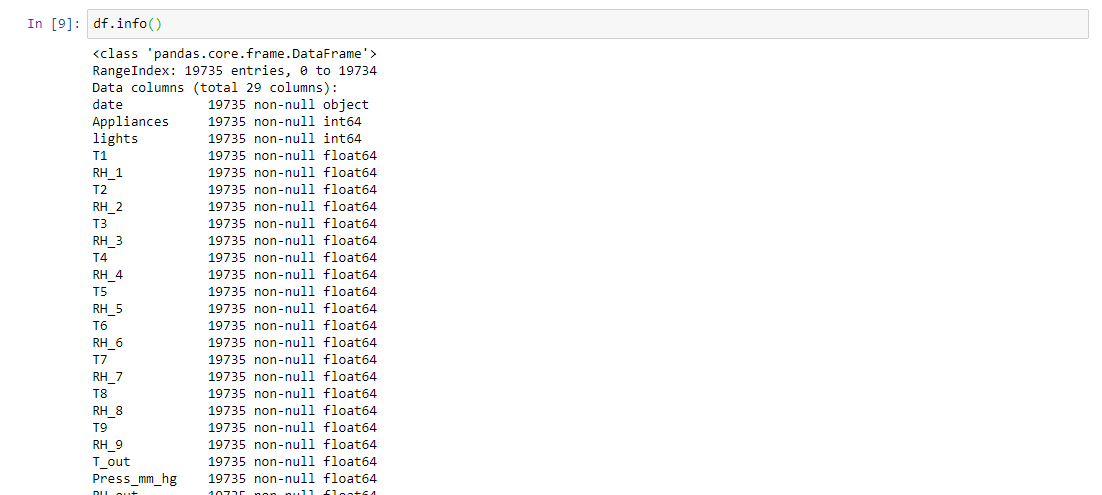
Then we read the data and viewed its head.



We also checked its shape along with the variable information.



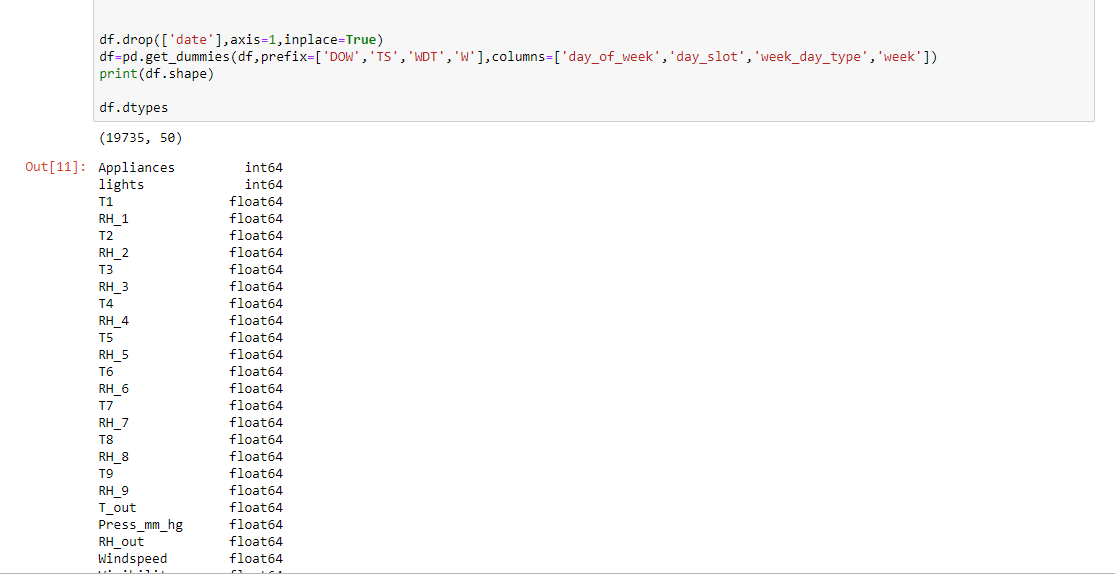
We also needed to check the variable information for further use.



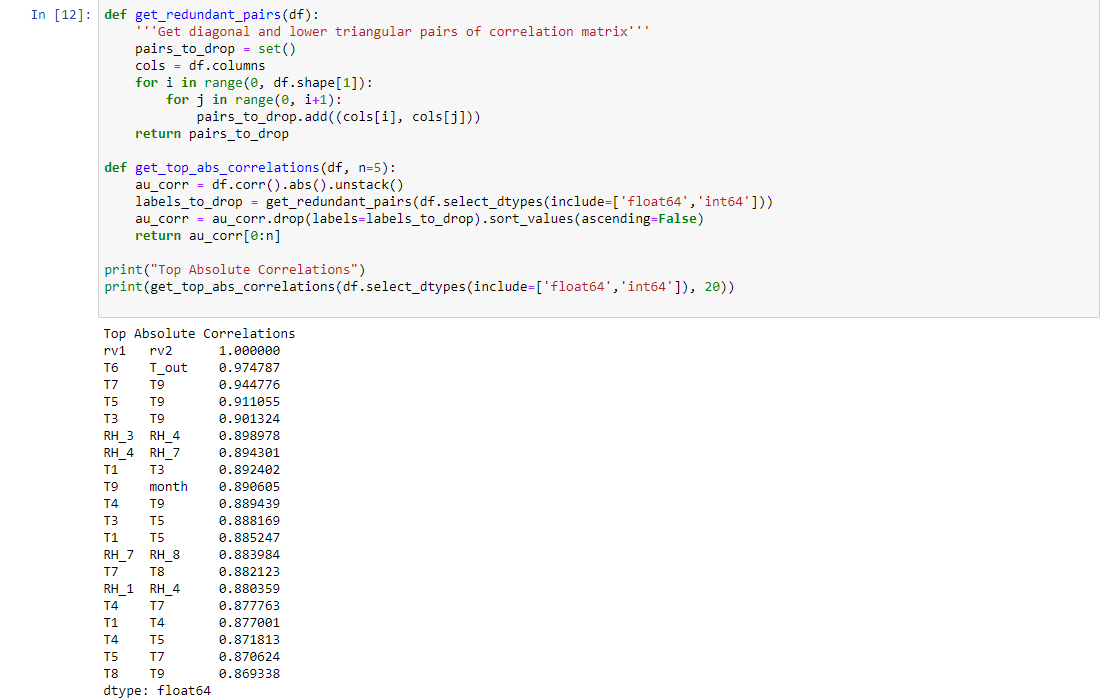
Then we change the date time according to month, week and time of day.



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 hihly correlated .

So, we removed these variables from our data.

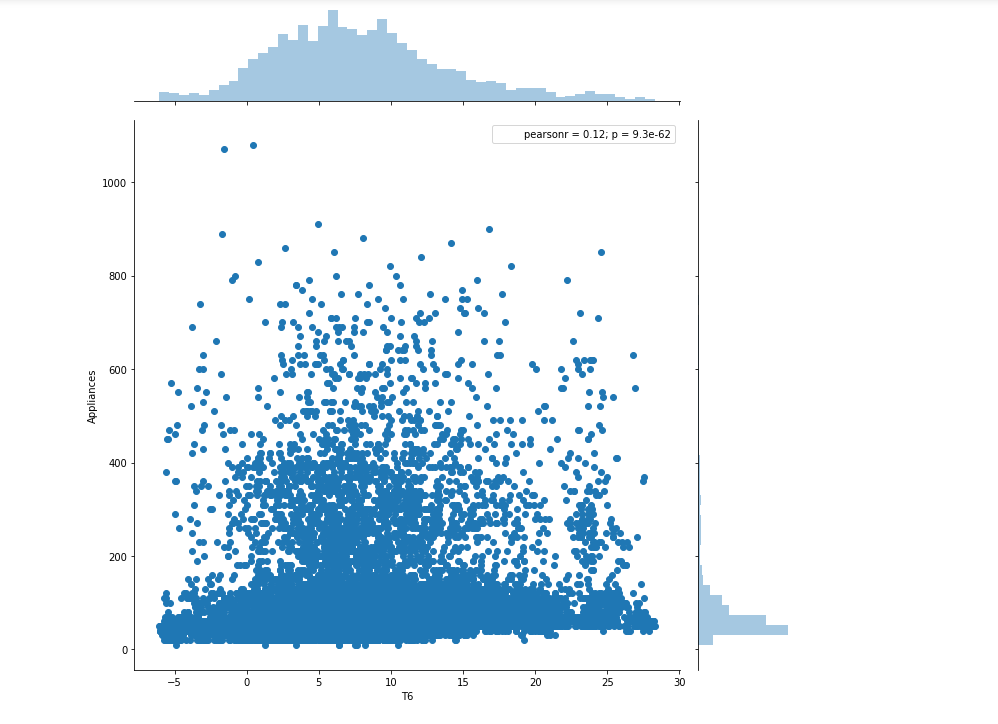


We also checked by using heatmap of variables.

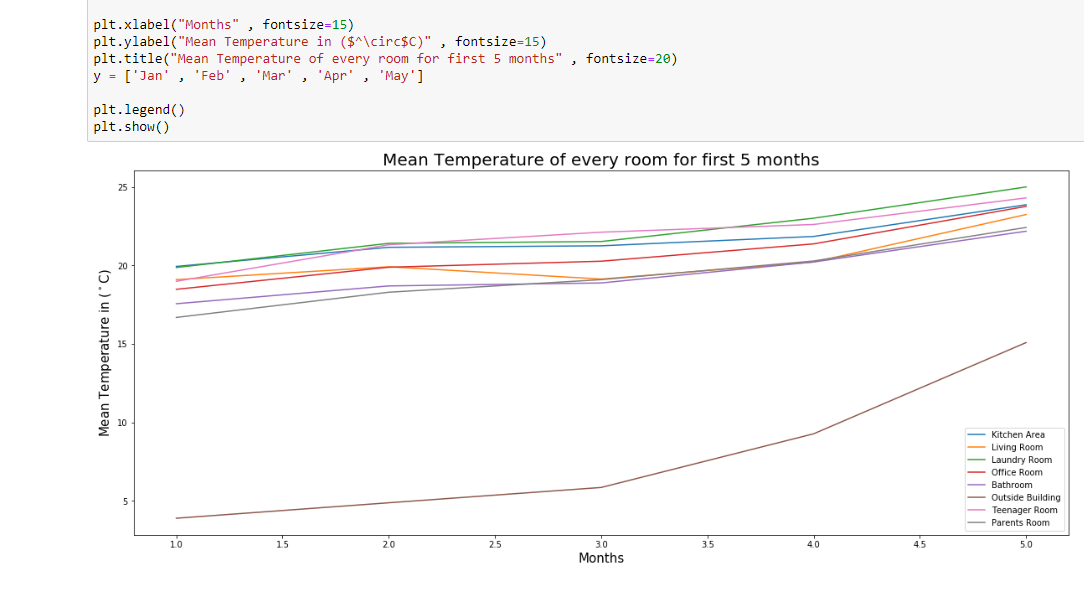


We also made following plot.

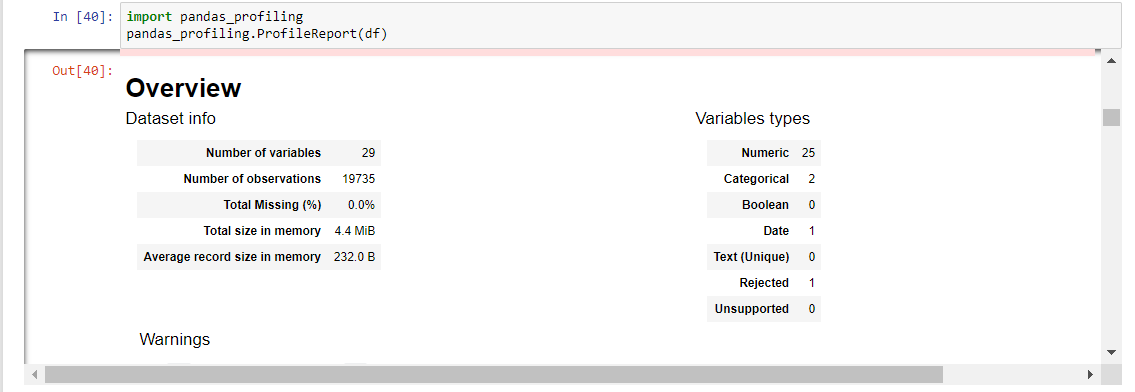


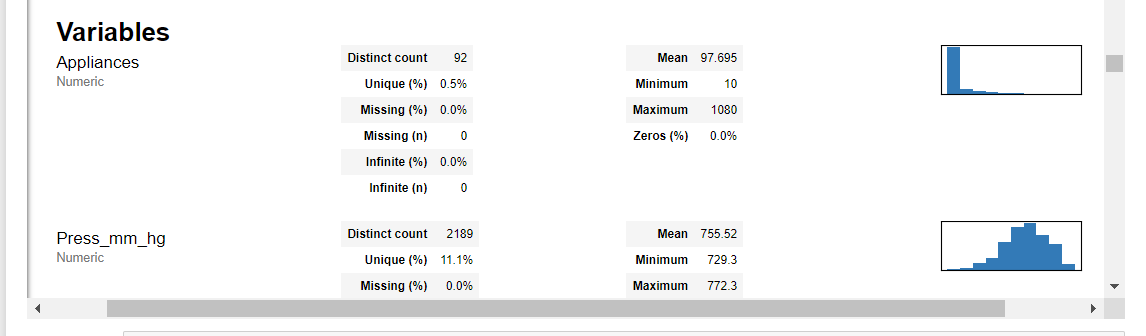


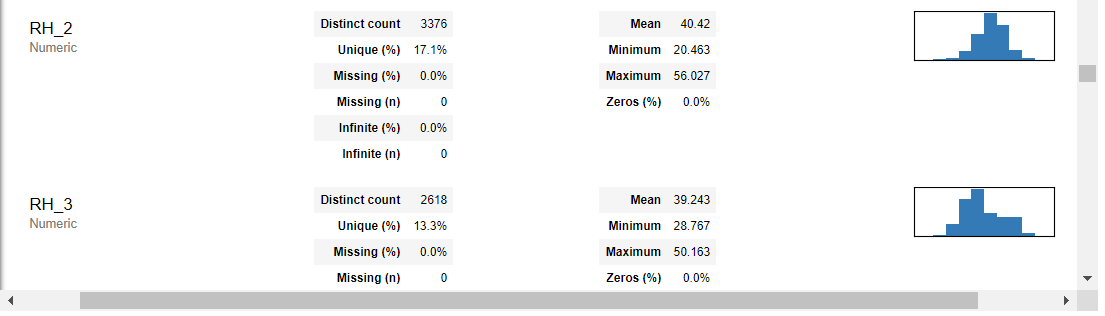
We also plotted the graph to look at the changes in temperature over time for various appliances

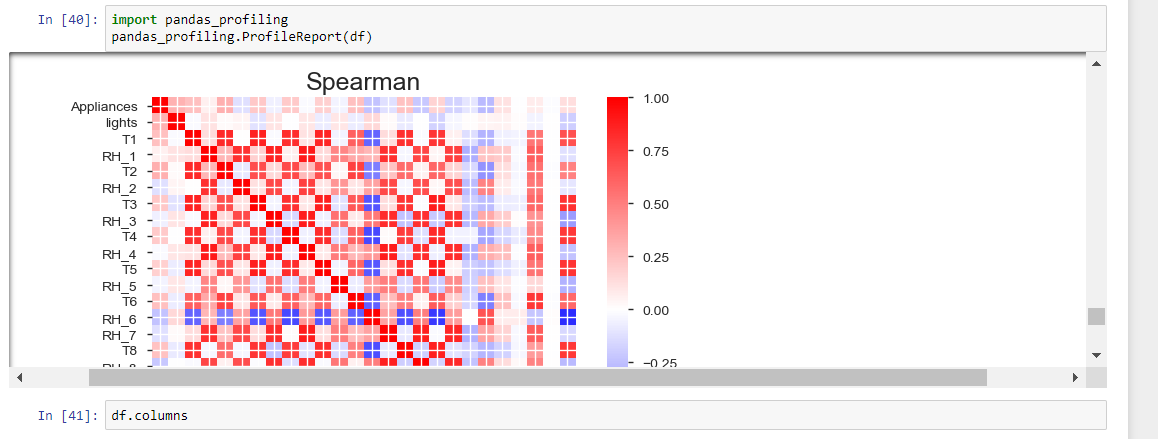


We also used pandas profiling to check for various variables.



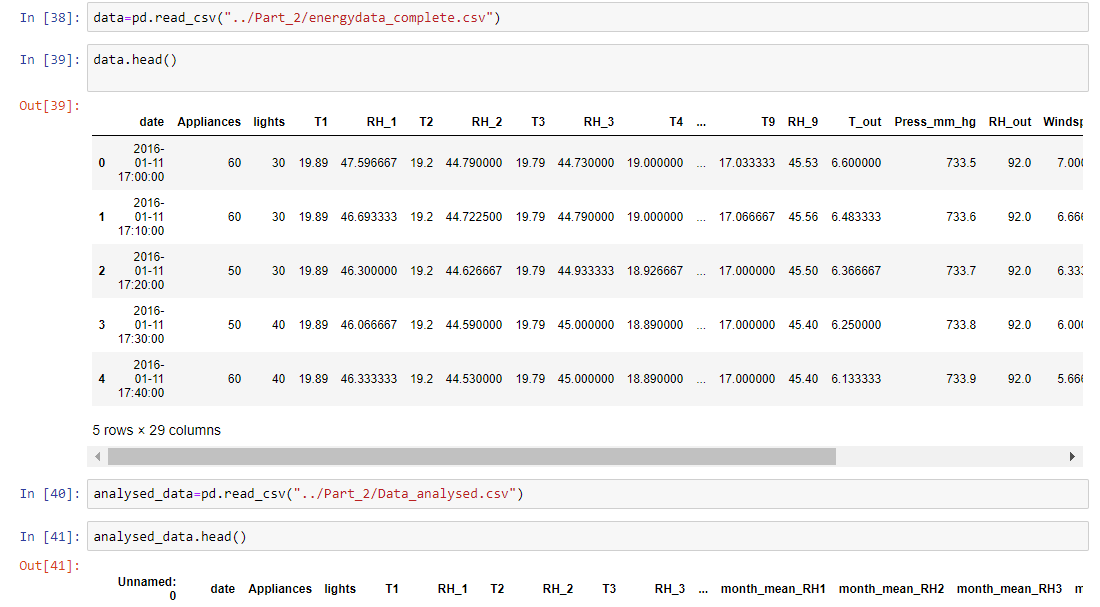




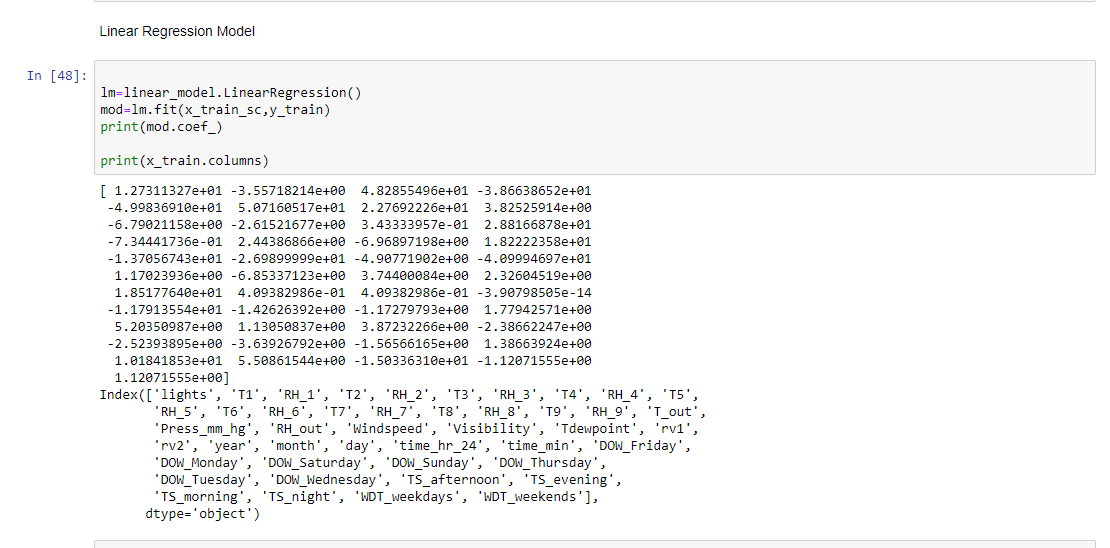


**3. Feature Engineering:**

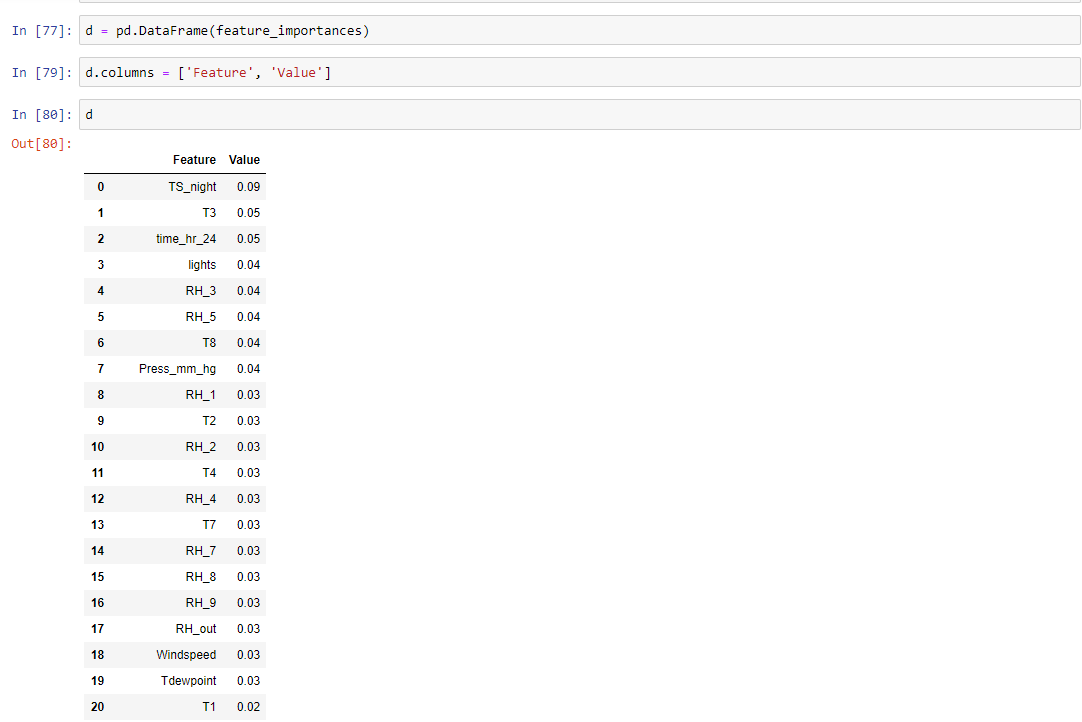
There are lots of features in our dataset. Temperature of 8 different rooms are recorded in degree  
Celsius. Humidity of 8 different rooms are recorded in percentage. Outside temperature is also  
recorded in degree Celsius and humidity in percentage. Along with it pressure has been recorded  
in millimeter 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.



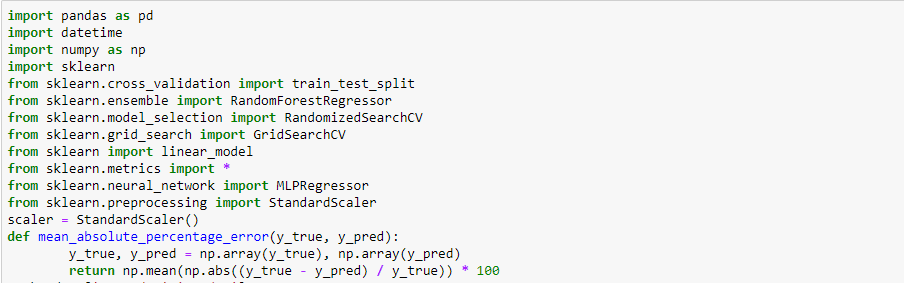
Finally, We found the importance of various features in our dataset.



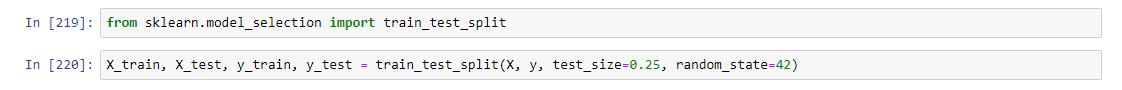
We also plotted a graph to check the same.



**4. Prediction Algorithms**

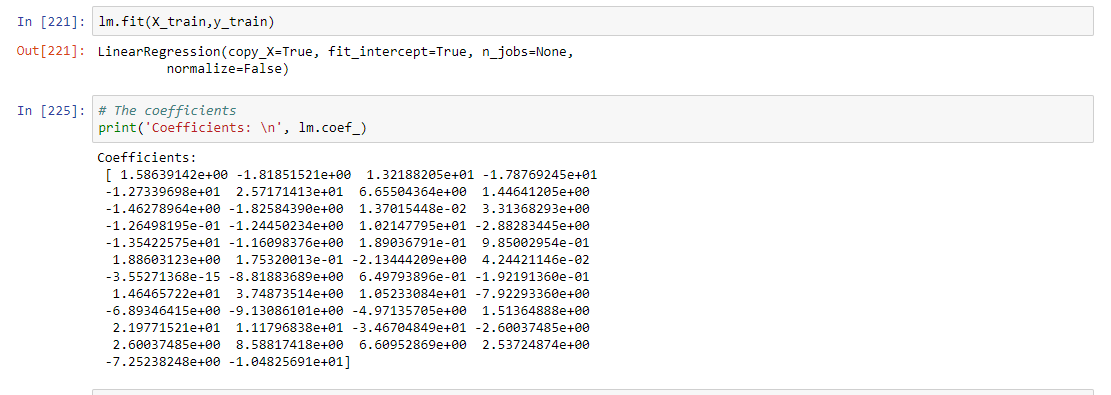


Splitting data

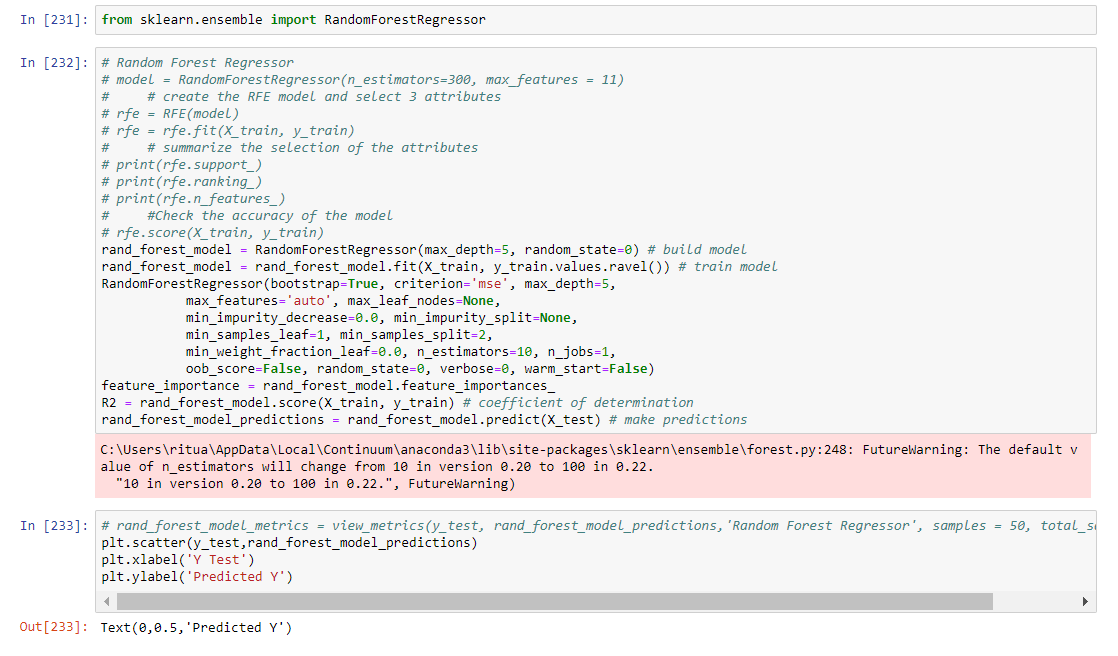


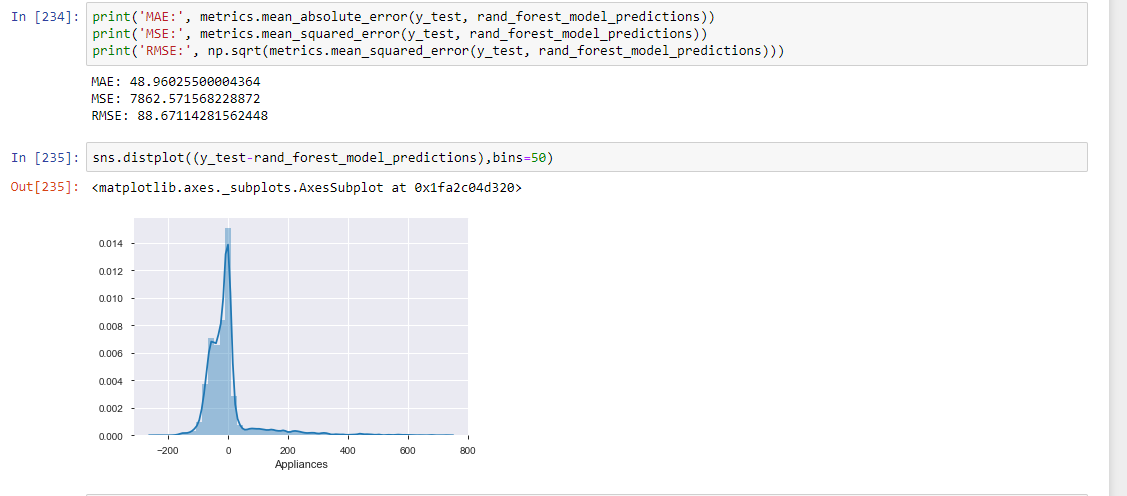
Linear Regression





Using Random Forest



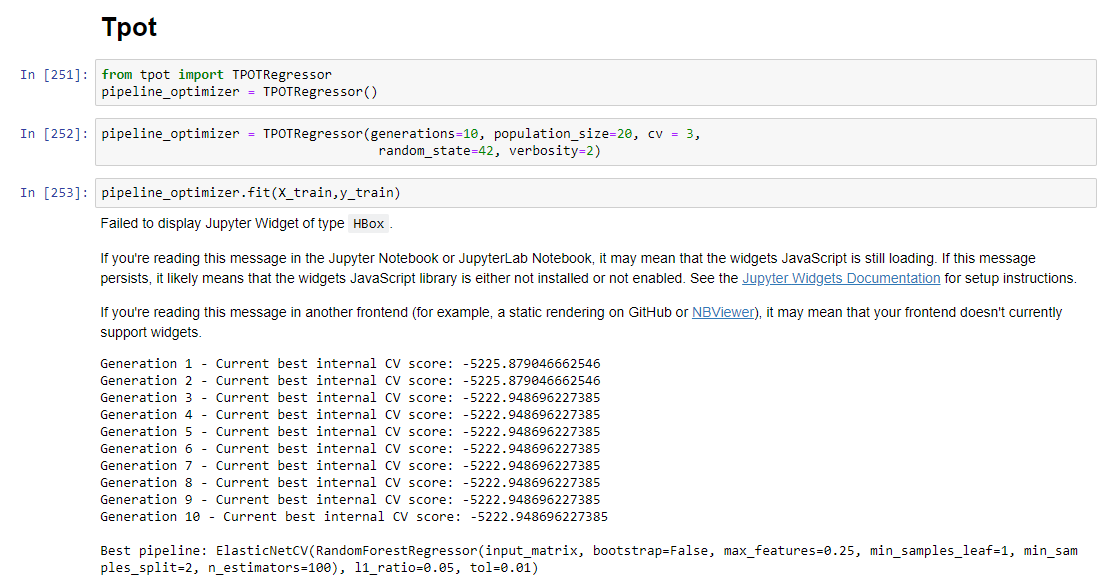


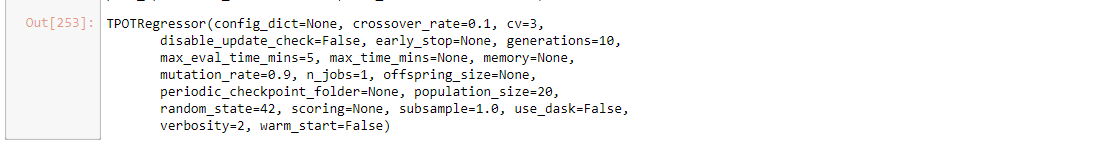
Neural Network Models



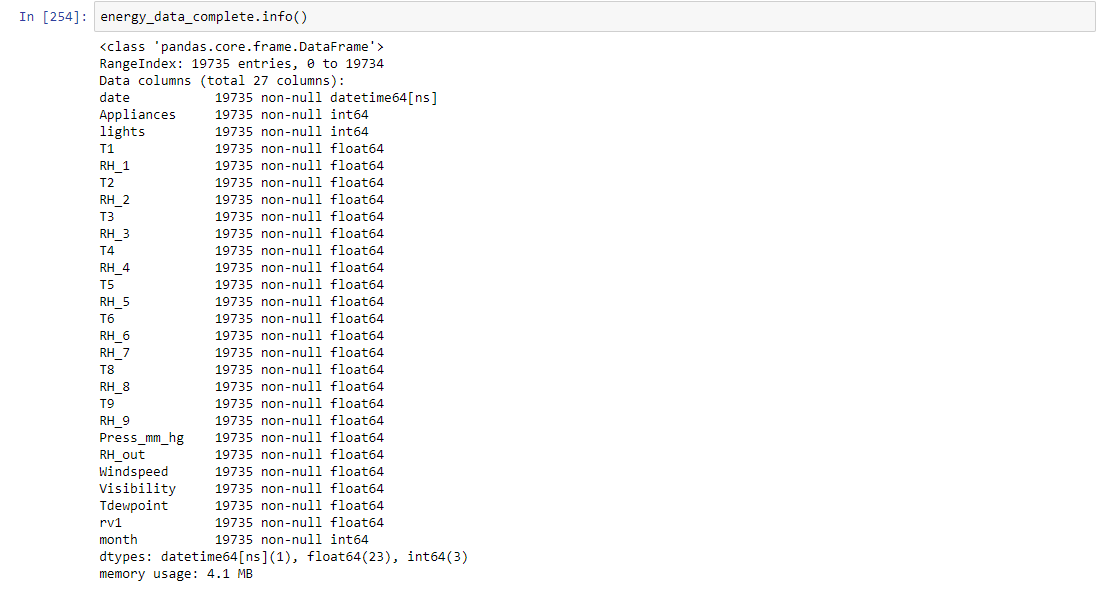
**5. Feature Selection**

**TPOT:**





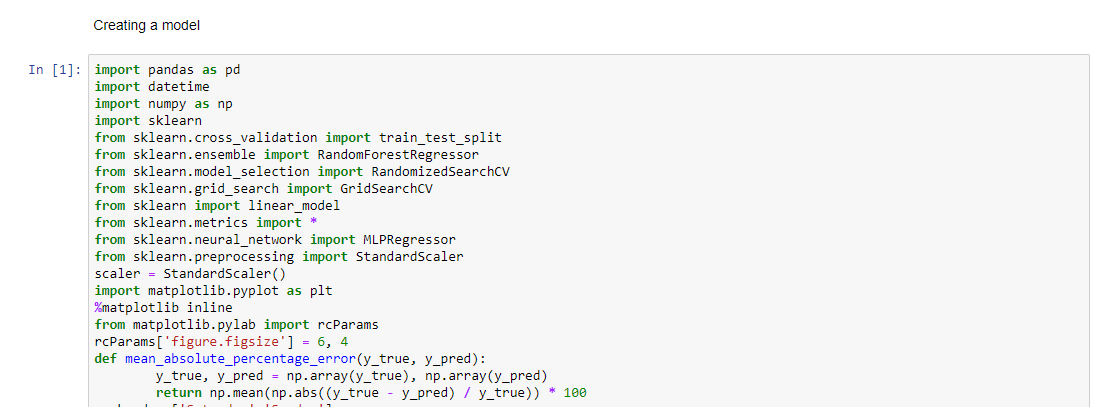
Info of model

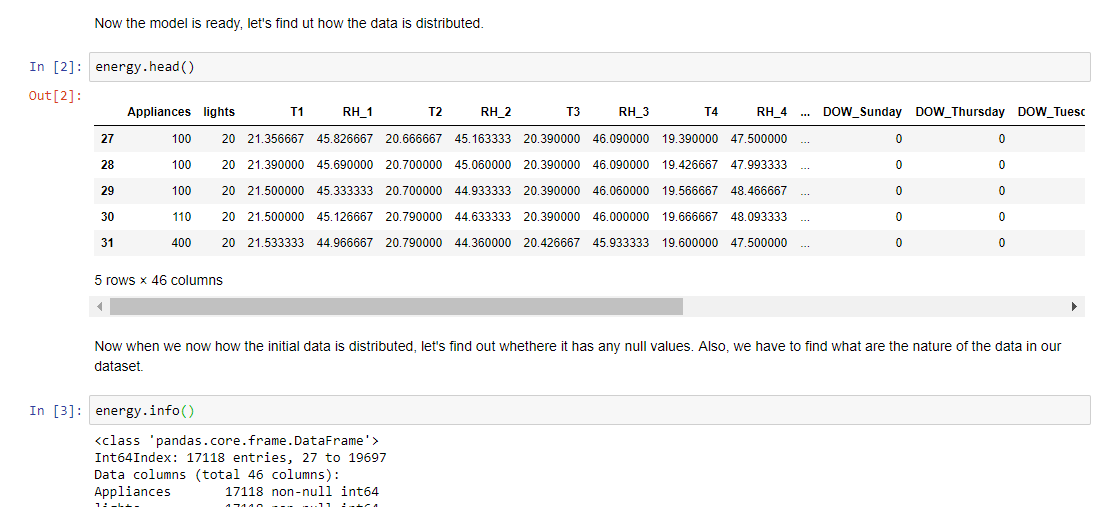


Final pipeline score



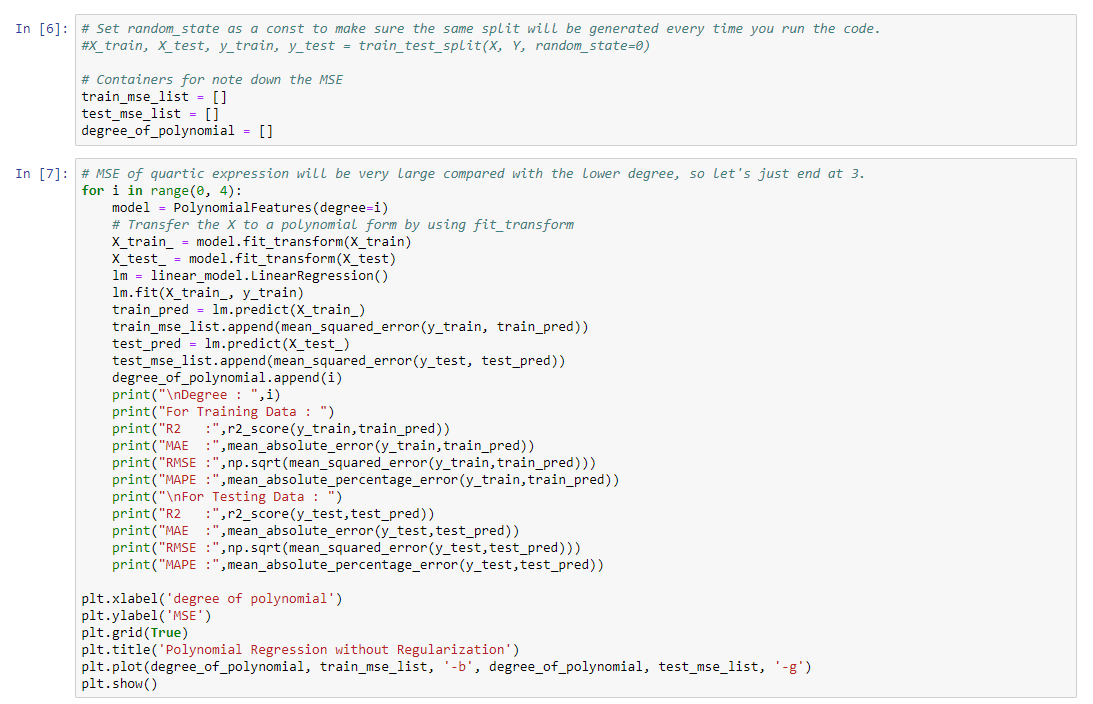
**6. Model Validation and Selection**

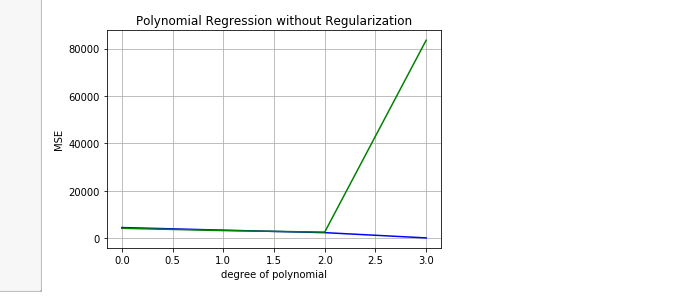






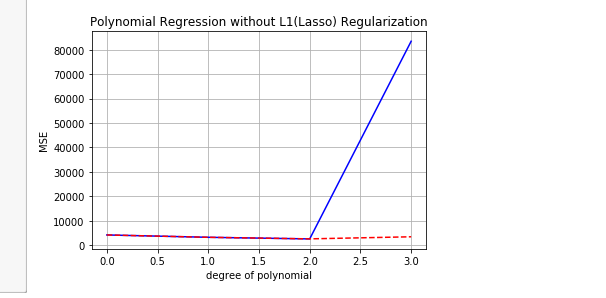
Regularization





Performing Lasso





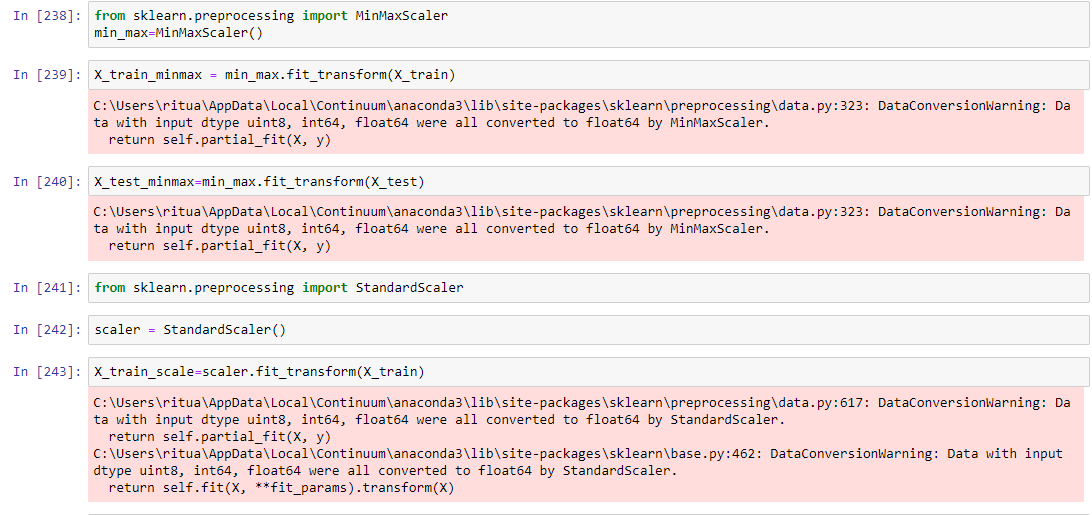
Cross Validation Technique



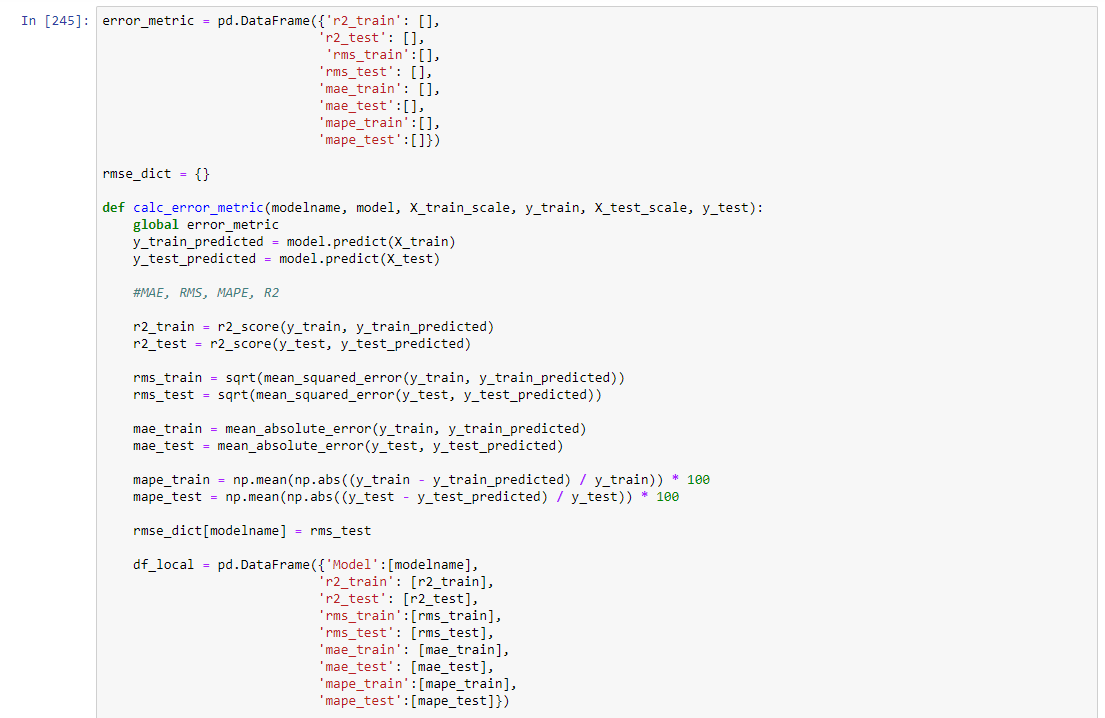
Getting Accuracy of model



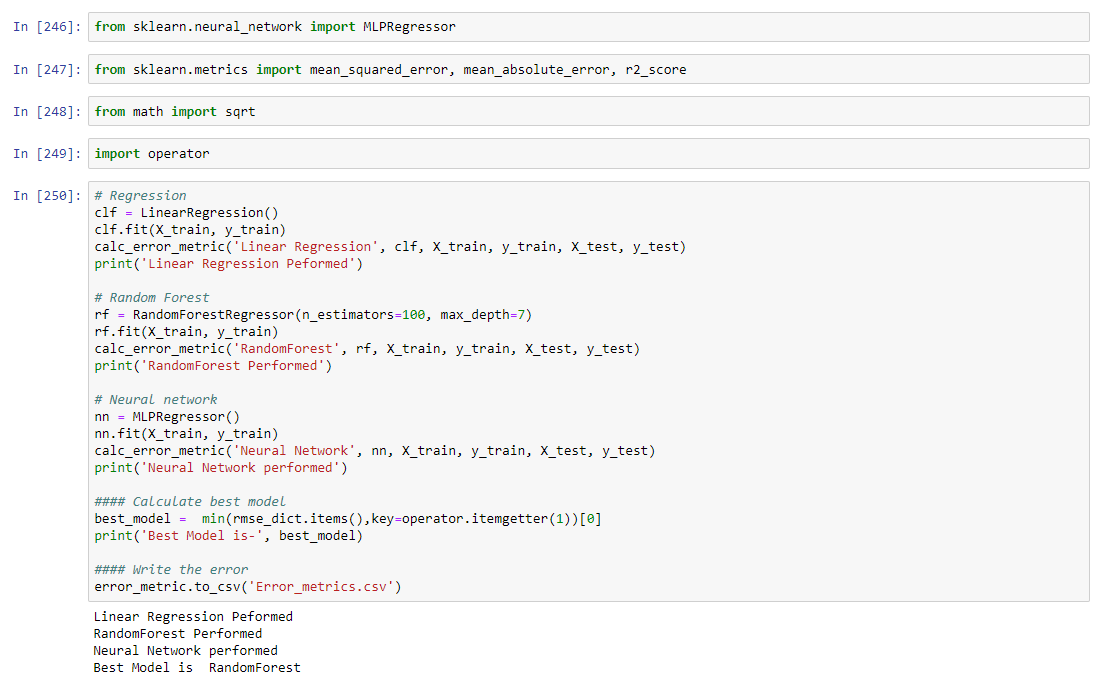
**7 Final Pipeline**

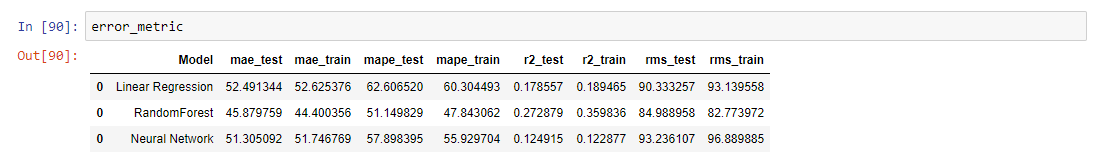


Error metric

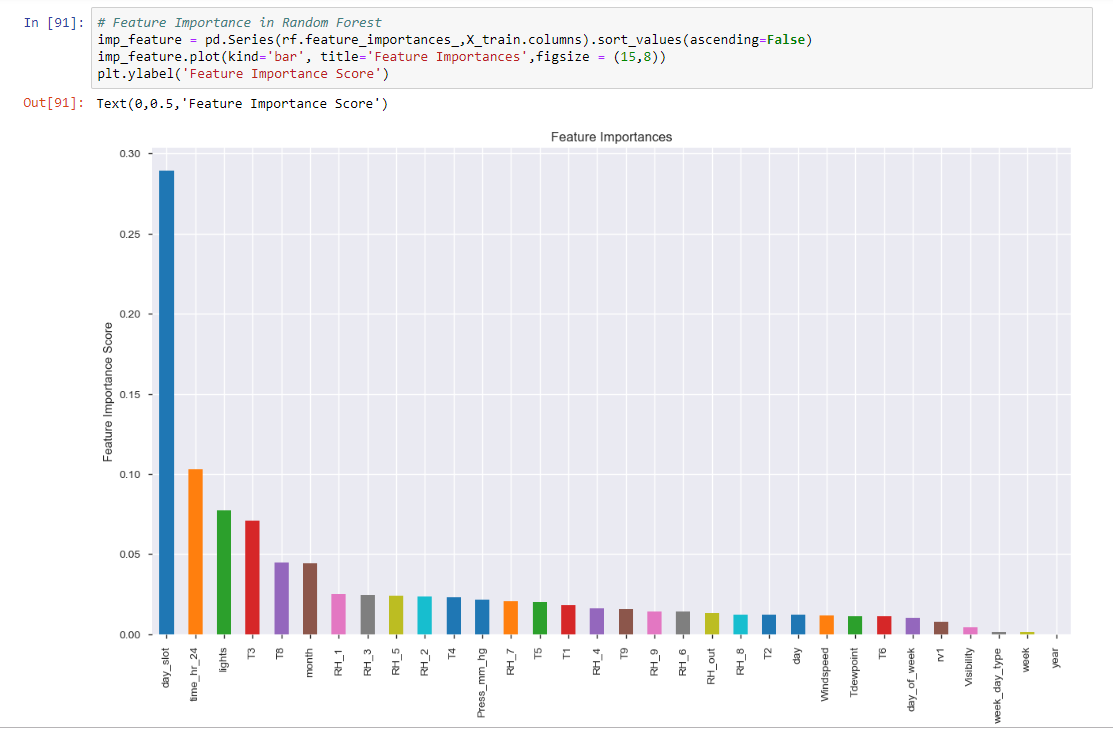


Getting best model.

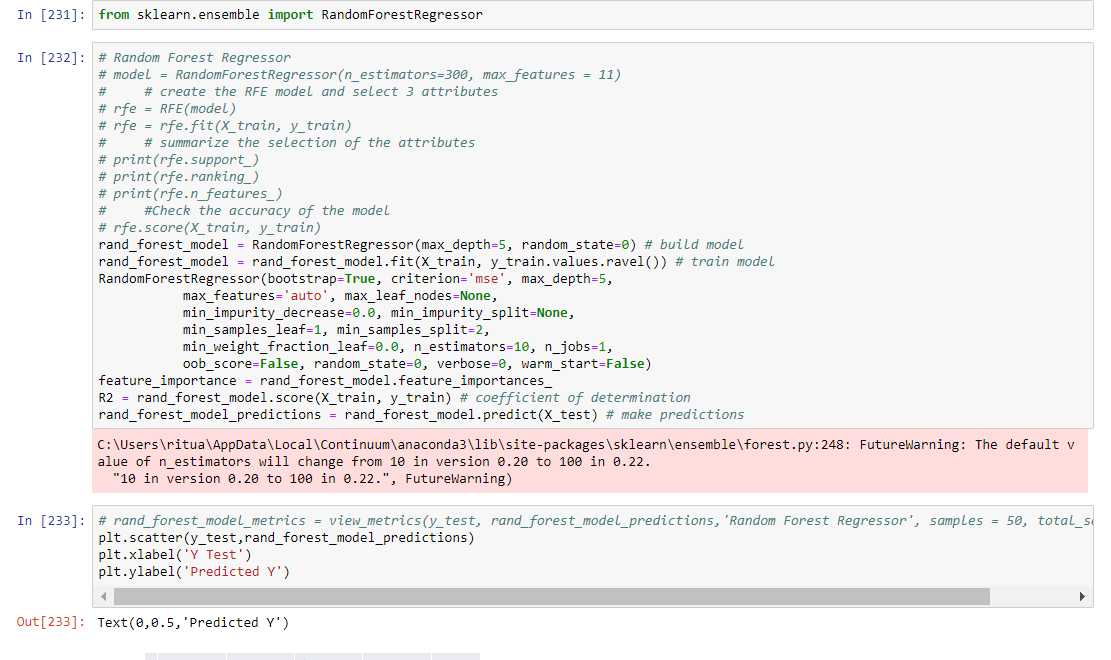


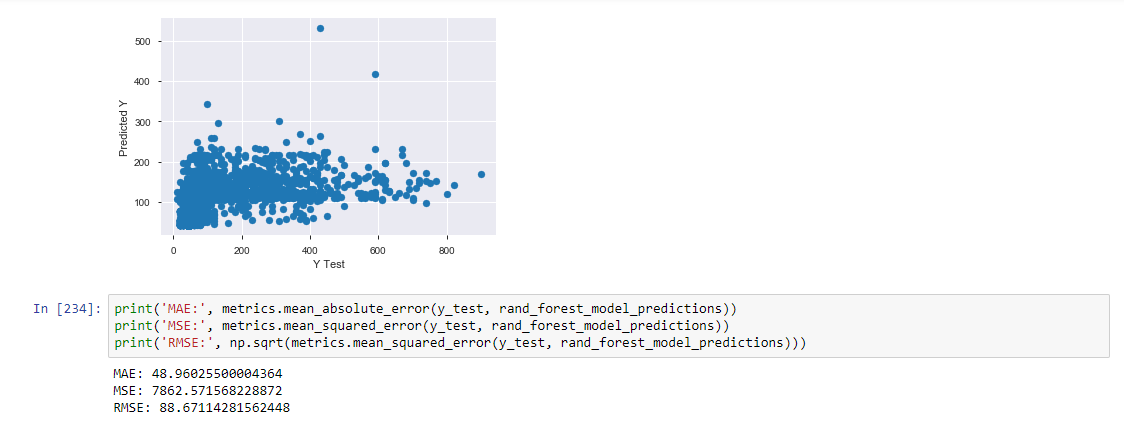


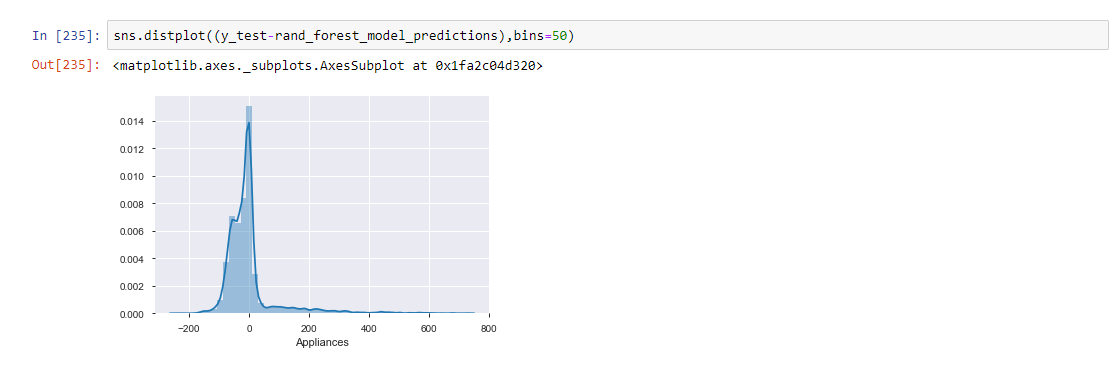
Getting feature importance of random forest



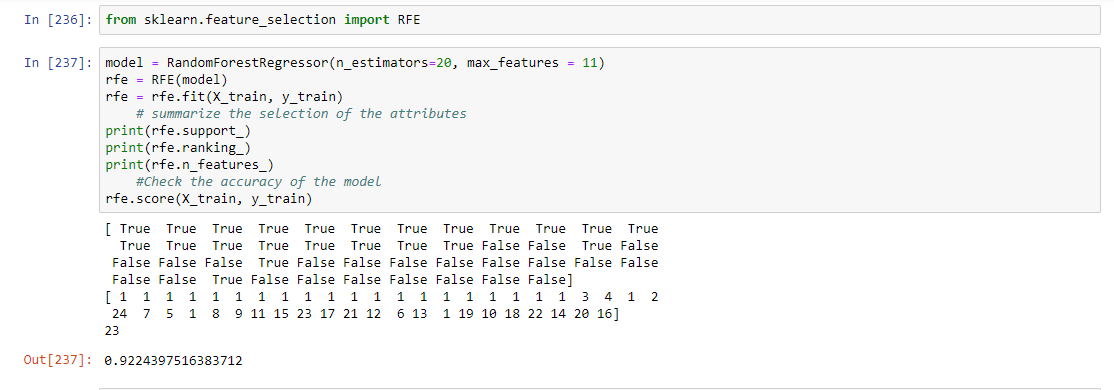
Doing prediction by random forest.







Predicting using Random forest algorithms



**So, our model gives 92 % 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.