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:** Data driven prediction models of energy use of appliances in a low-energy house

Link -  <https://www.sciencedirect.com/science/article/pii/S0378778816308970?via%3Dihub>

**Abstract**

This paper presents and discusses data-driven predictive models for the energy use of appliances. Data used include measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station and recorded energy use of lighting fixtures. The paper discusses data filtering to remove non-predictive parameters and feature ranking.

1. **Introduction**

The electricity consumption in domestic buildings is explained by two main factors: the type and number of electrical appliances and the use of the appliances by the occupants. The occupancy level of the building in different locations could also help to determine the use of the appliances. Specifically, data from a nearby airport weather station, temperature and humidity in different rooms in the house from a wireless sensor network and one sub-metered electrical energy consumption have been used to predict the energy use by appliances. Four regression models have been tested, namely (a) multiple linear regression model(lm), (b) support vector machine with radial basis function kernel(SVM-radial), (c) random forest(RF) and (d) gradient boosting machines(GBM) with different combinations of predictors.

**- 1.1 Literature review**

**1.1.1 Appliances' loads in buildings and numerical modeling of their consumption**

This section reviews some articles regarding modeling of appliances and other socio economic factors that help to understand the different data and methodologies that have been used in the past to understand appliances’ energy use.

**1.1.2 Electricity load prediction**

This section presents and discusses research addressing electricity load prediction to identify the parameters, models and other methods that have been useful for energy prediction.

Models such as multiple regression, neural networks, forecasting methods, engineering methods, support vector machines, time series techniques and forecasting methods have been used to predict the electricity demand. The models usually have considered parameters such as the time of day, outdoor temperature, month, weekend, holidays, yesterday’s consumption, rainfall index, global solar radiation, wind speed and occupancy.

From the data available to the researchers, an extensive list of variables was studied: weather, location (ZIP code), age of building, ownership, presence of double pane windows, energy efficient light fixtures, floor area, pet ownership, number of refrigerators and entertainment devices, number of occupants and income level were studied. The researchers concluded that the most important variables were weather, location, and floor area. Also, the number of refrigerators and entertainment appliances are among the most important determinants of daily minimum consumption. Another study found that being at home during the day correlated with lower appliance efficiency. The provided explanation by the researchers is the lower efficiency was likely due to the increased use of appliances when the house is occupied more often.

A prediction system for the problem of individual appliance prediction was presented in. The system used information such as past consumption, hour, day, season and month. The system is capable of learning from past data. One of the main conclusions was that the last 24 h are the most relevant for prediction.

**- 1.2 Research objectives and methodology outline**

The purpose of this work is to understand the relationships between appliances energy consumption and different predictors.

1. **House description**

A house located in Stambruges is taken into consideration for analyzing the aggregated electric energy consumption per month. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network.

1. **Recorded data and description**

The energy (Wh) data logged every 10 min for the appliances is the focus of this analysis. The 10 min reporting interval was chosen to be able to capture quick changes in energy consumption.

**- 3.1 Data sets and exploratory analysis**

The combined data set is split in training and test validation using CARET’S (Classification and Regression Training Package) create data partition function. 75% of the data is used for the training of the models and the rest is used for testing.

**- 3.2 Data features filtering and importance**

The Boruta package is used here to select all the relevant variables. Several researchers have used this package for variable filtering. To test the Boruta algorithm, two random variables were introduced in the data sets. Moreover, this feature or variable selection helps in model interpretability and reduces complexity of the model.

The Boruta package compares importance of attributes with importance of shadow attributes that are created by shuffling original ones. The Boruta algorithm is capable of detecting the two random variables that have no predicting power for the appliances’ energy consumption.

To test how many variables are optimal to minimize the RMSE the recursive feature elimination (RFE) is used to select the optimal inputs.

**- 3.3 The performance of regression models**

In order to compare the performance of each of the regression models, different performance evaluation indices are used here: the root mean squared error (RMSE), the coefficient of determination or R-squared/R2, the mean absolute error (MAE) and the mean absolute percentage error (MAPE).

**- 3.4 Model selection**

The best models are the ones that provide the lower RMSE and highest R-square values.

1. **Conclusion**

For all the models, the time information was ranked as the most important to predict the appliances’ consumption.

When using all the predictors the light consumption was ranked highly. However, when studying different predictor subsets, removing the light consumption appeared not to have a significant impact. This may be an indication that other features are correlated well with the light energy consumption.

This study has found curious relationships between variables. Future work could include considering weather data such as solar radiation and precipitation.

**RESEARCH PAPER 2:** A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models

Link - <https://www.sciencedirect.com/science/article/pii/S1364032116307420>

Building energy prediction can be broadly classified into engineering, Artificial Intelligence (AI) based, and hybrid approaches. While engineering and hybrid approaches use thermodynamic equations to estimate energy use, the AI-based approach uses historical data to predict future energy use under constraints. Tall buildings and skyscrapers accounts for 30% of total electricity consumption. This gave wide spectrum to data analyst to analyze this scenario which can be exercised through engineering methods. There are 2 major methods which are as follows: (1) AI based method also called as the black box predicts energy use without knowing the internal relationship of the building and the individual components. (2) Hybrid is called as grey box which requires detailed building information for simulation for model development.

There are 2 prediction methods which are widely used for energy systems. They are as follows: (1) Single Prediction – utilize one learning algorithm (2) Ensemble Prediction – integrate some of the single prediction to improve accuracy of prediction.

There are many parameters which can be used for predicting energy usage by buildings like building type which can be commercial, residential, educational or research. We can assume them to be one among educational, research or commercial due to availability of data. Residential building has privacy issues. The predicting model used for energy usage varied from one prediction algorithm to ensemble. Ensemble model is preferred due to its demonstrated superiority over one prediction after much research and development into it over the years. Artificial Neural Network (ANN) is used because of ease of implementation and reliable prediction performance besides regression, SVR, ARMAX, CHAID for building energy use prediction. The energy type was divided into 5 categories i.e. whole building energy/electricity (35%), heating and cooling energy (11%), heating energy (11%), cooling energy (13%) and all others (8%). Other considerable parameter was Prediction time scale which represents the time resolution of the prediction which is often impacted by the sampling interval of sensors and the research. Preferred time scale was Hour (49%), Day (19%), Year (8%) and Other (24% - min-by-min, week, month). Based on knowledge researchers collect data. Meteorology (60%), Occupancy (29%), Other (54%). Different patterns could be identified and analyzed.

The various AI-based prediction models involve methods like Data collection, data preprocessing, model training and model testing. In the single prediction method, I use multiple linear regression where inputs consist of shape factor, envelope U-value, window-to-floor area ratio, building time constant and climate which is defined as a function of sol-air temperature and heating set-point. These models were easy and efficient forecast tools for calculating heating demand of residential buildings. Catalina et al. simplified the MLR model by introducing only three inputs namely, building global heat loss coefficient, south equivalent surface, and the difference between the indoor set point temperature and sol-air temperature. Their results indicated that the proposed method closely predicted future building heat demand. Jacob et al. improved the performance of regression model by introducing the rate of change of the indoor air temperature as an independent variable. Their study indicated that the performance of MLR could be improved by introducing appropriate independent variables. In ANN, a nonlinear statistical technique which consists of Input, output and hidden layers interconnected. Ben-Nakhi and Mahmoud applied General Regression Neural Network (GRNN) to predict the cooling load for commercial buildings. Multiple results show ANNs could perform better than regression method for short-term forecasting. To detect complex nonlinear relationships between inputs and outputs implicitly which is good for real time monitoring. However, ANN method fails to establish any interconnection relationship between building physical parameters and building energy use, which limits the model's fitting ability when changes are made to building components or systems. In Support Vector Regression where input data is mapped via a nonlinear function. It finds the most deviation from the obtained target. Selection of kernel function is important as it affects the learning ability of SVR. SVR demonstrates its ability to predict hourly cooling load in the building. SVR helps in optimization

In the Ensemble prediction method, the aim is to provide best possible prediction performance by automatically managing the strengths and weaknesses of each base model. It has multiple learning models depending on base model resampling, manipulation or randomization of training data, learning algorithm and parameters. Ensemble models can be Homogenous and Heterogenous models. Homogenous model uses bagging and boosting. The various ways to implement it is firstly via Input feature identification, Data monitoring and preprocessing, Learning algorithm selection, Base model generation, Model integration.

Proposed by Hansen and Salamon 1990 to solve classification problems. They state that the collective decision produced by the ensemble model is less likely to be in error than the decision made by any of the individual models. Multiple classification algorithms and integration schemes were used to develop ensemble models. heterogeneous ensemble classifier that consisted of five different classification algorithms to solve health-related short text classification problem. A parameter sensitivity analysis was carried out to obtain the best possible features. Multiple model integration schemes such as multi-staging, reverse multi-staging, majority voting, and weighted probability averaging were used to combine the classification results of the base classifiers. The result indicated that the proposed ensemble classifier performs better than the single SVM classifier in the studied problem. Three combination techniques such as the majority voting, the LSE-based weighting, and the double-layer hierarchical combining were used to aggregate the individual SVMs. Three typical classification problems: data classification, handwritten digit recognition, and fraud detection were used to test the efficacy of the proposed ensemble model.

Their results indicated that the ensemble model outperforms single SVM model in terms of classification accuracy which is performed on buildings around the globe.

**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.

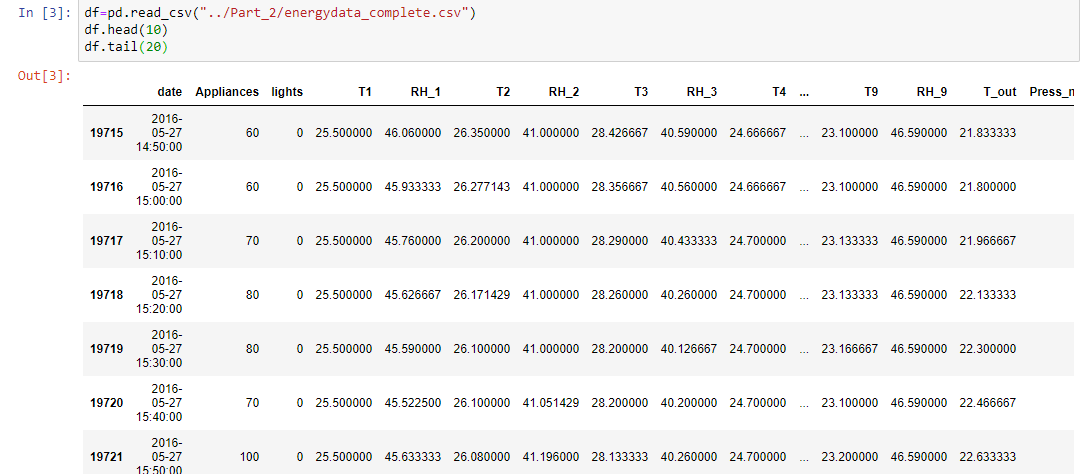
**2. Exploratory Data Analysis**

We imported following libraries

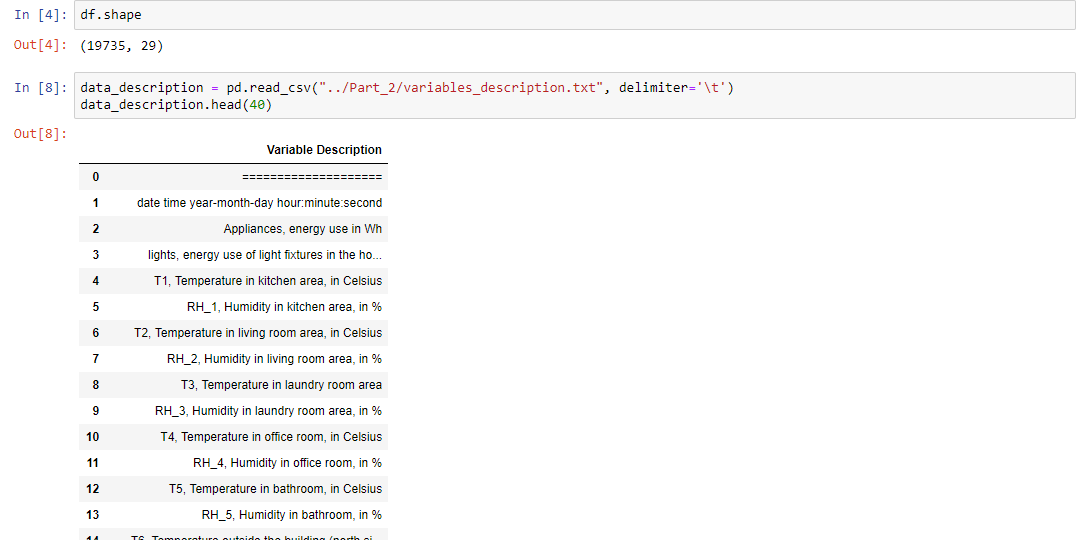


Then we read the data and viewed its head.

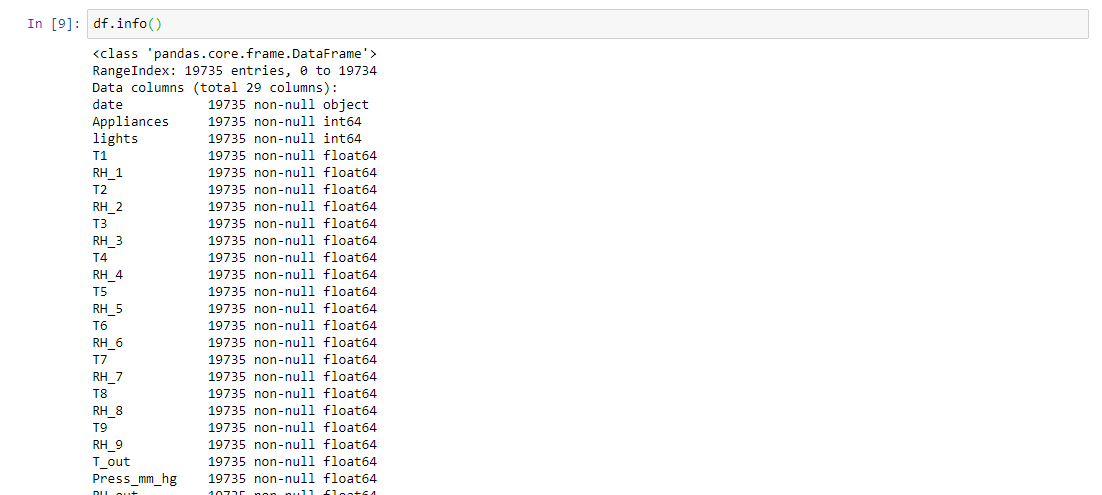
There were mainly variables for temperature and humidity at various parts of house.



We also checked its shape along with the variable information.



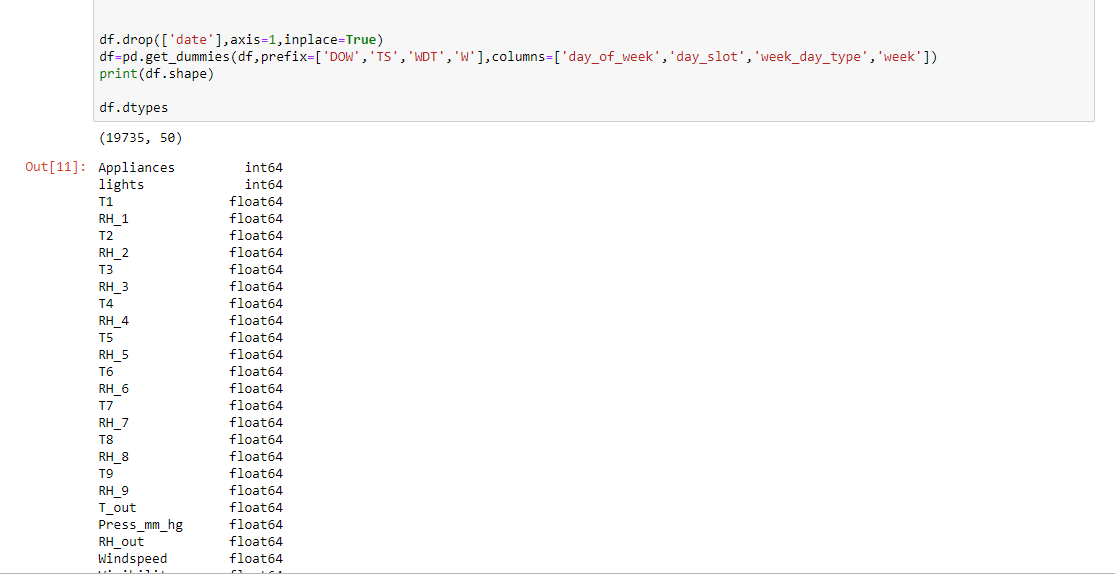
We also needed to check the variable information for further use.



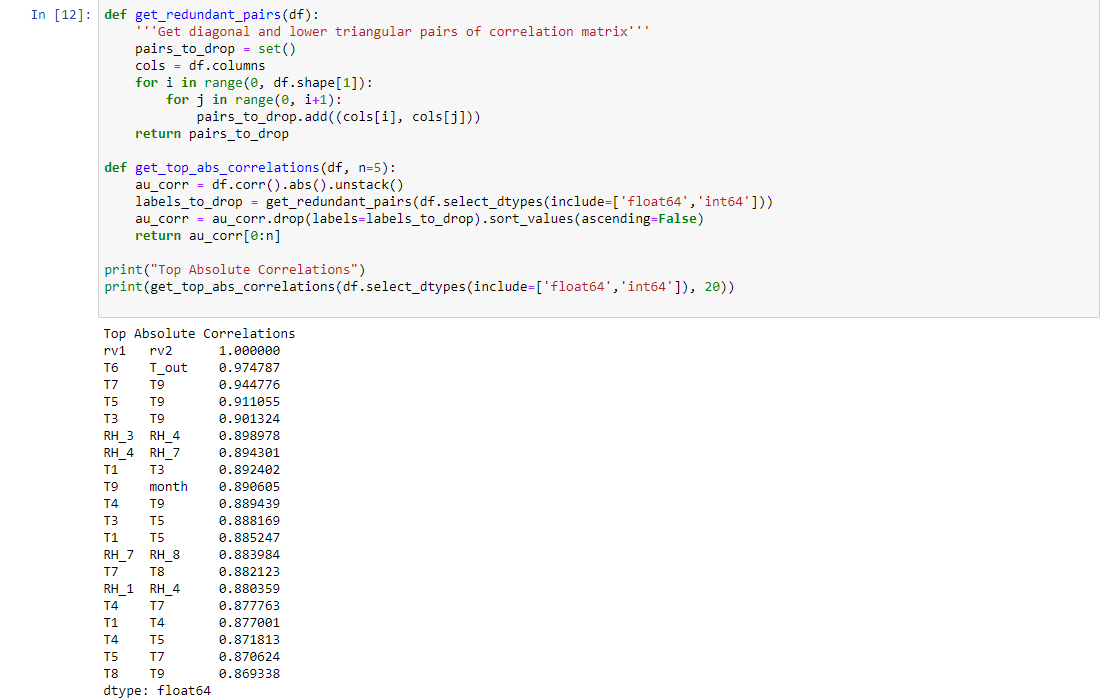
Then we change the date time according to month, week and time of day.



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.

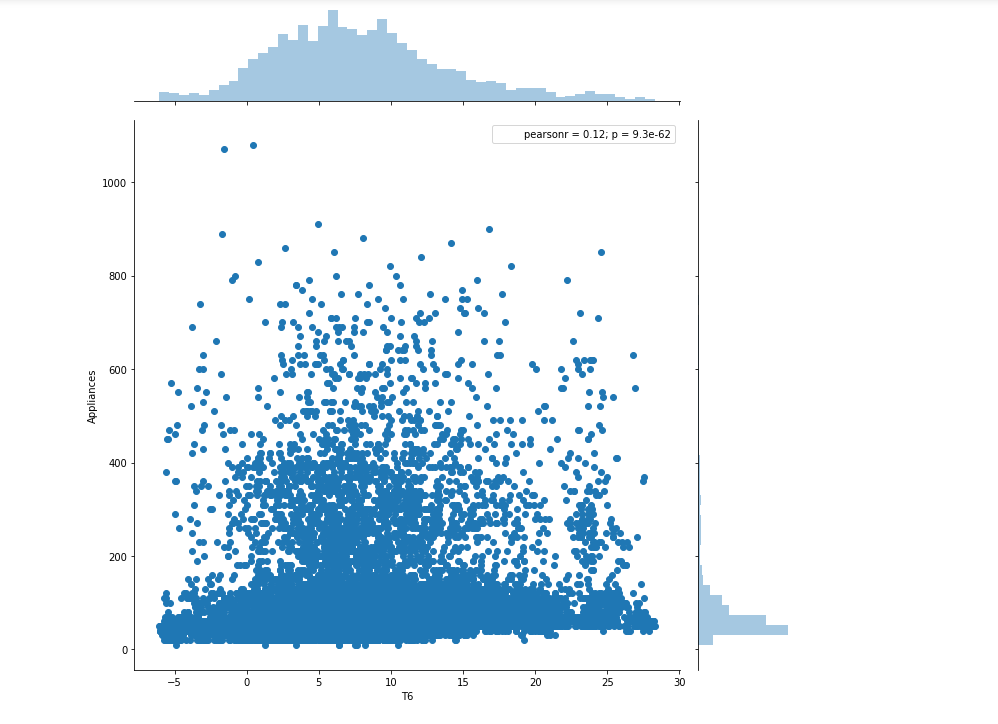


We also checked by using heatmap of variables.

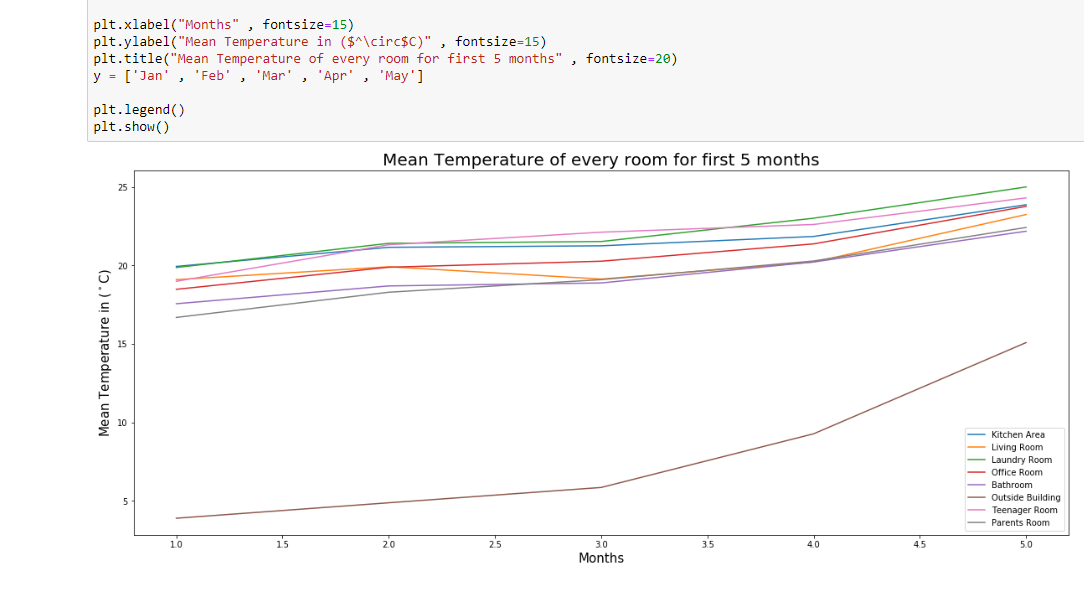


We also made following plot.

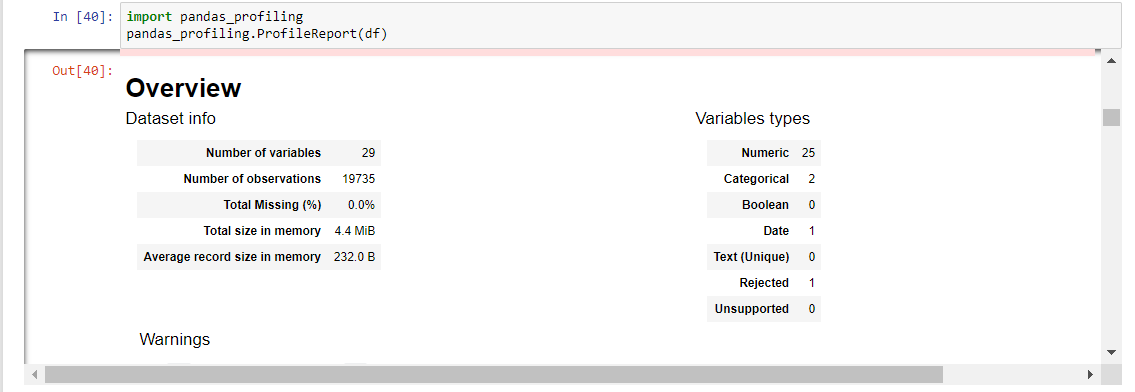


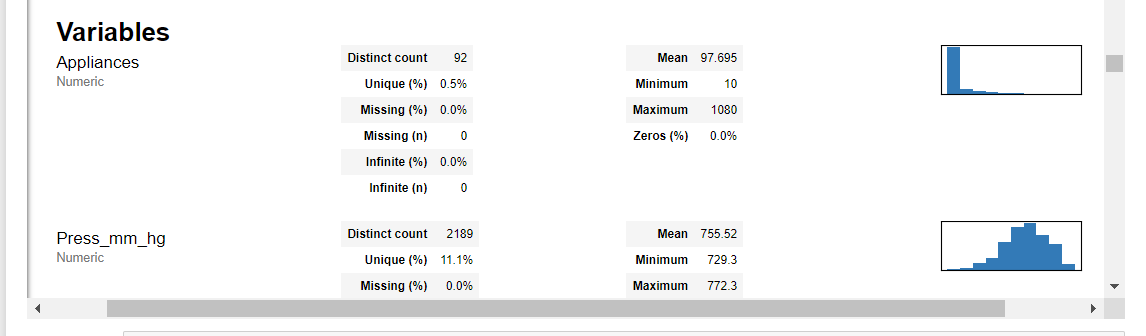


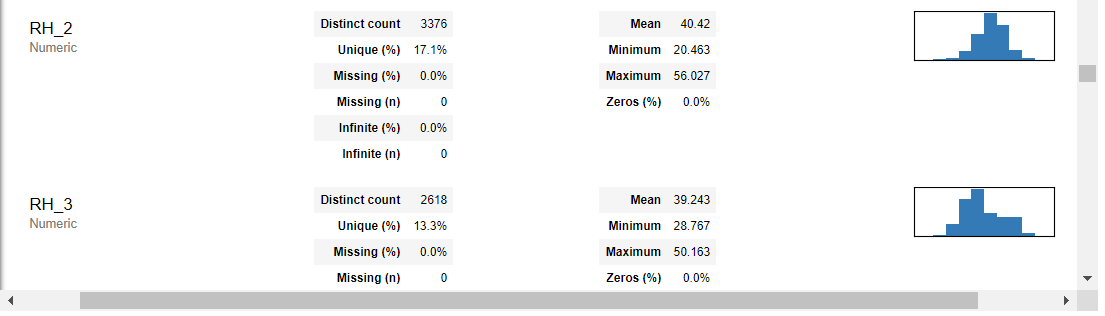
We also plotted the graph to look at the changes in temperature over time for various appliances

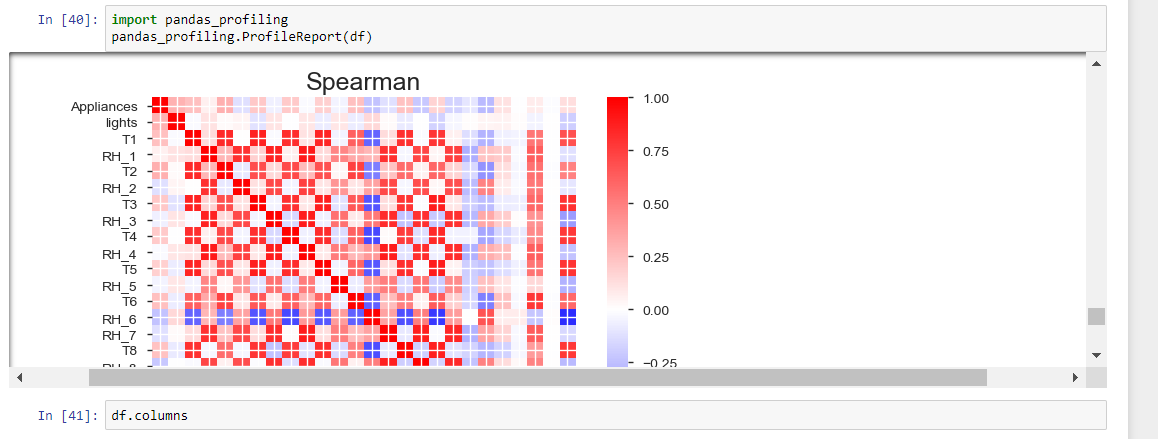


We also used pandas profiling to check for various variables.



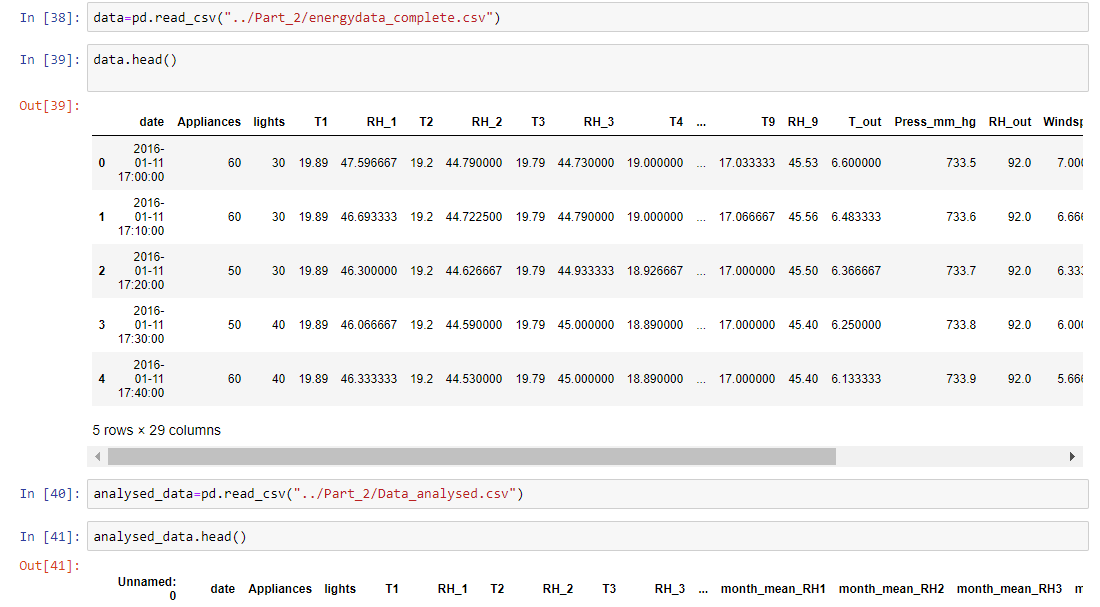




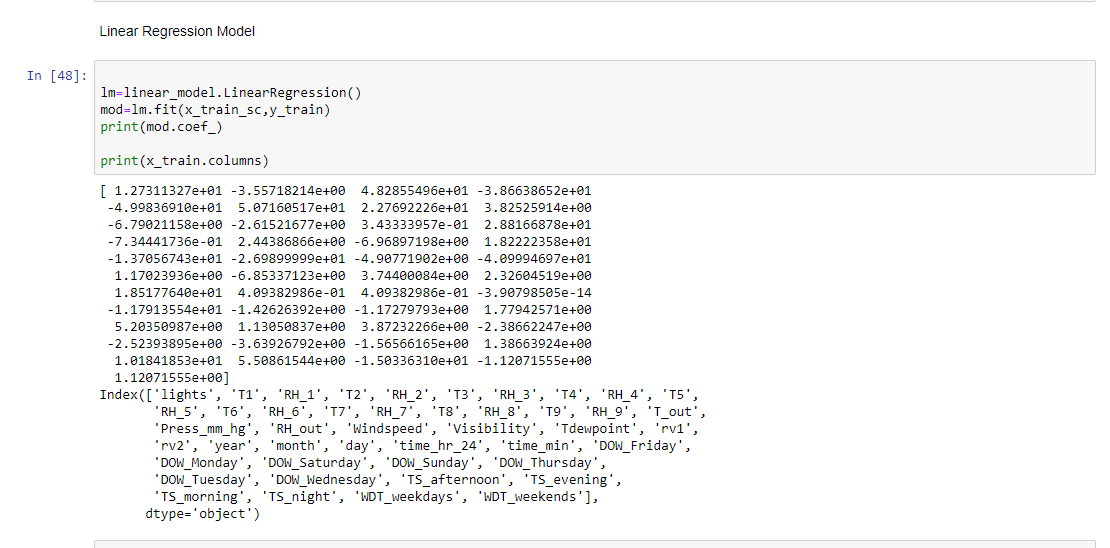


**3. Feature Engineering:**

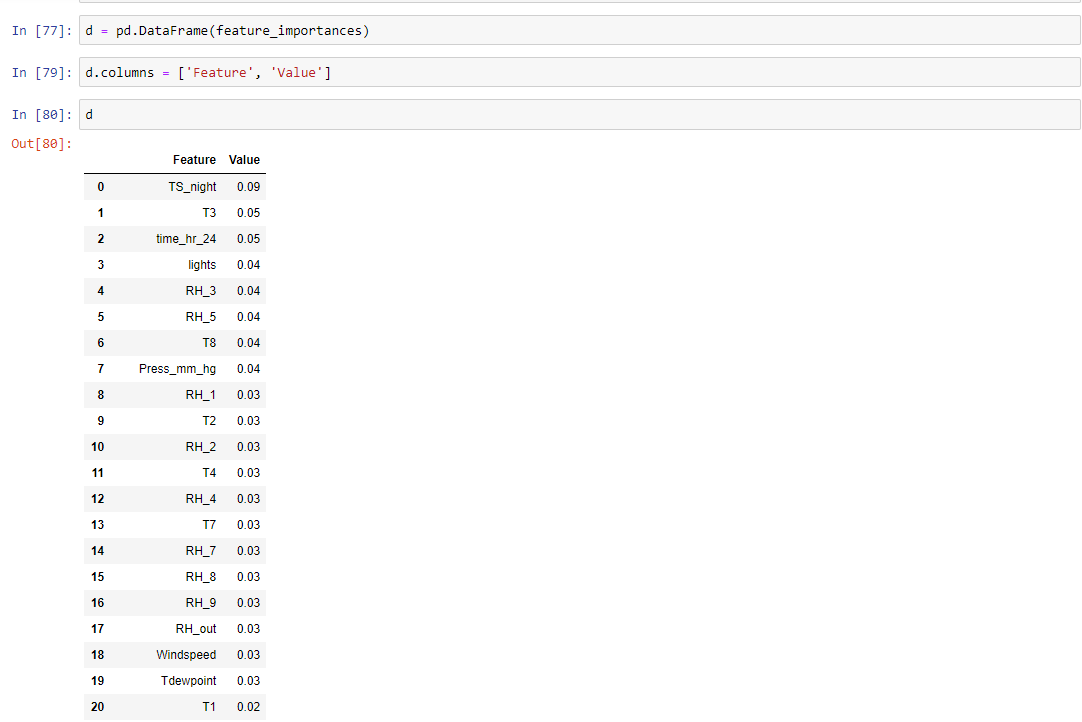
There are lots of features in our dataset. Temperature of 8 different rooms are recorded in degree  
Celsius. Humidity of 8 different rooms are recorded in percentage. Outside temperature is also  
recorded in degree Celsius and humidity in percentage. Along with it pressure has been recorded  
in millimeter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



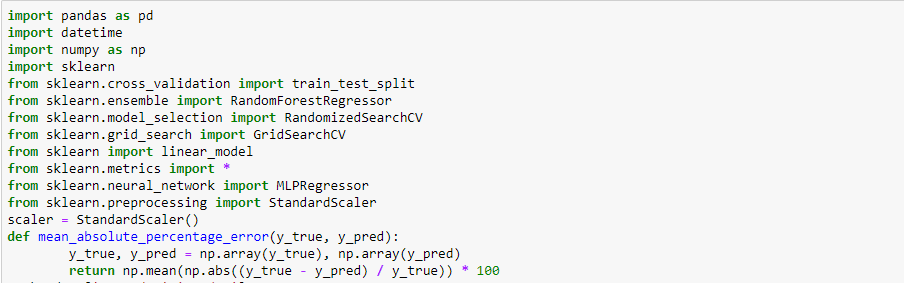
Finally, We found the importance of various features in our dataset.



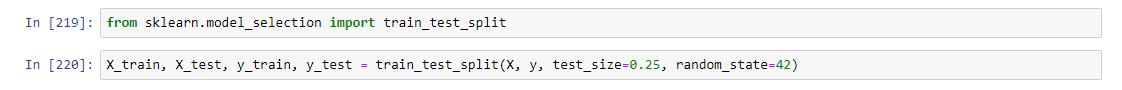
We also plotted a graph to check the same.



**4. Prediction Algorithms**

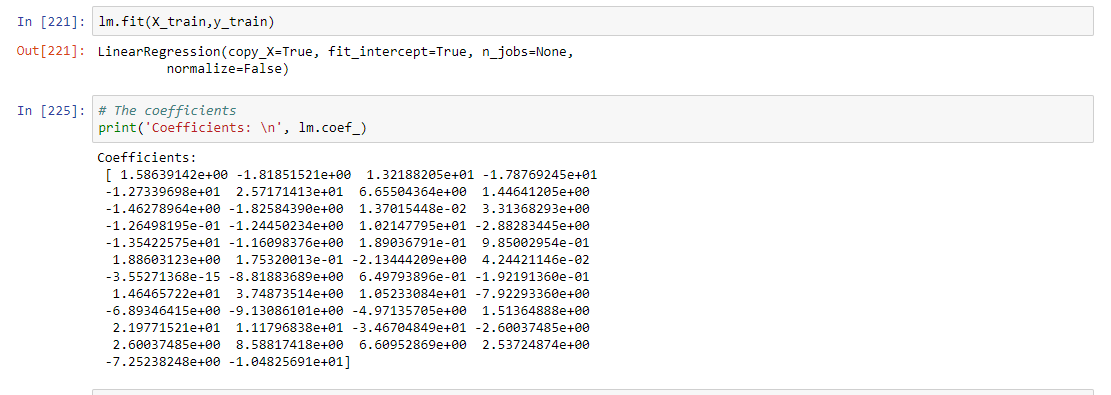


Splitting data

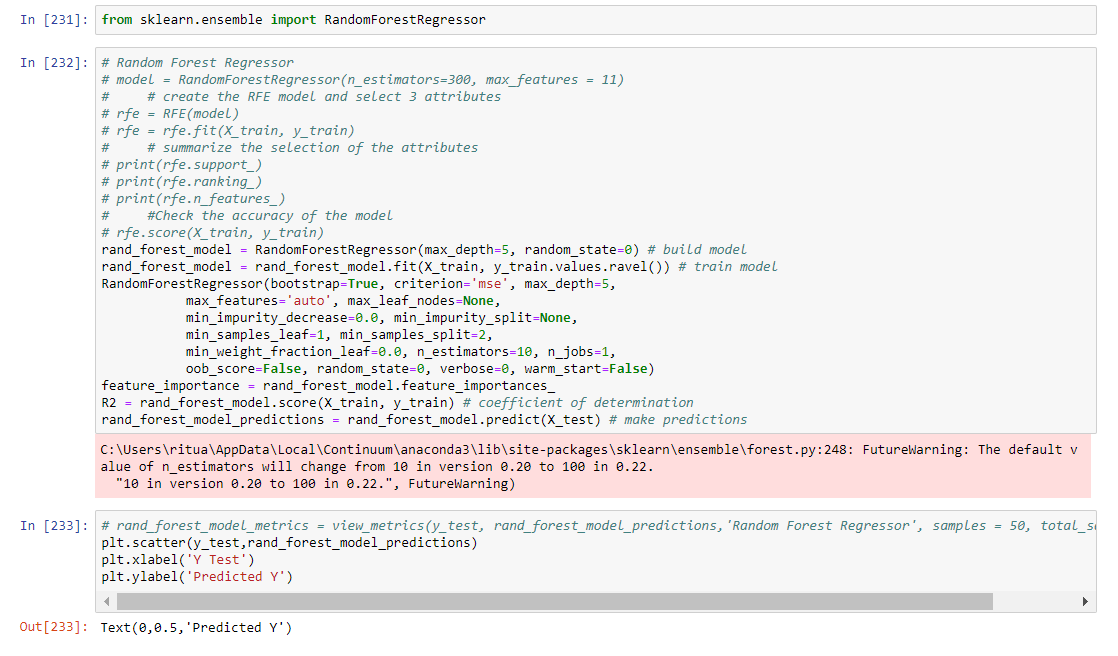


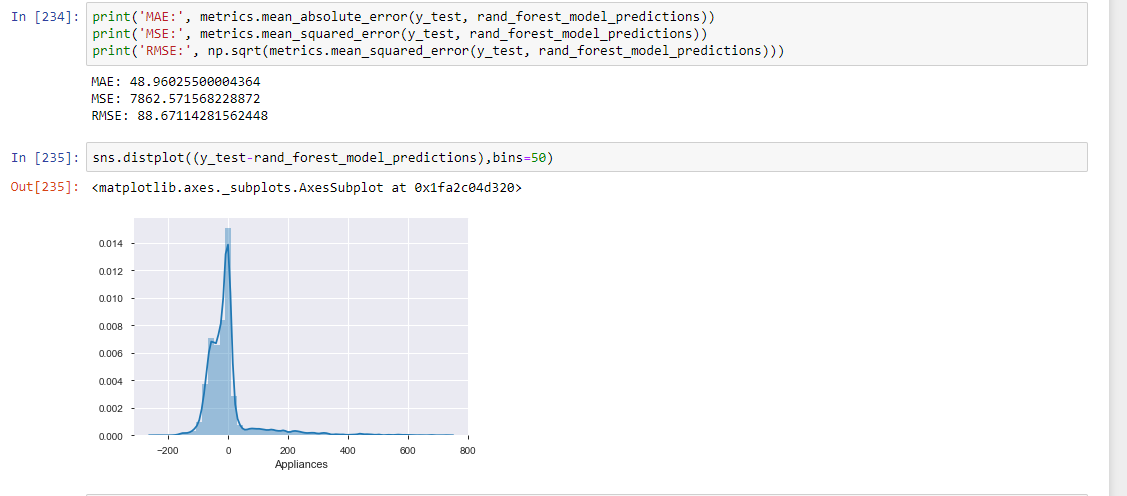
Linear Regression





Using Random Forest





Neural Network Models



**5. Feature Selection**

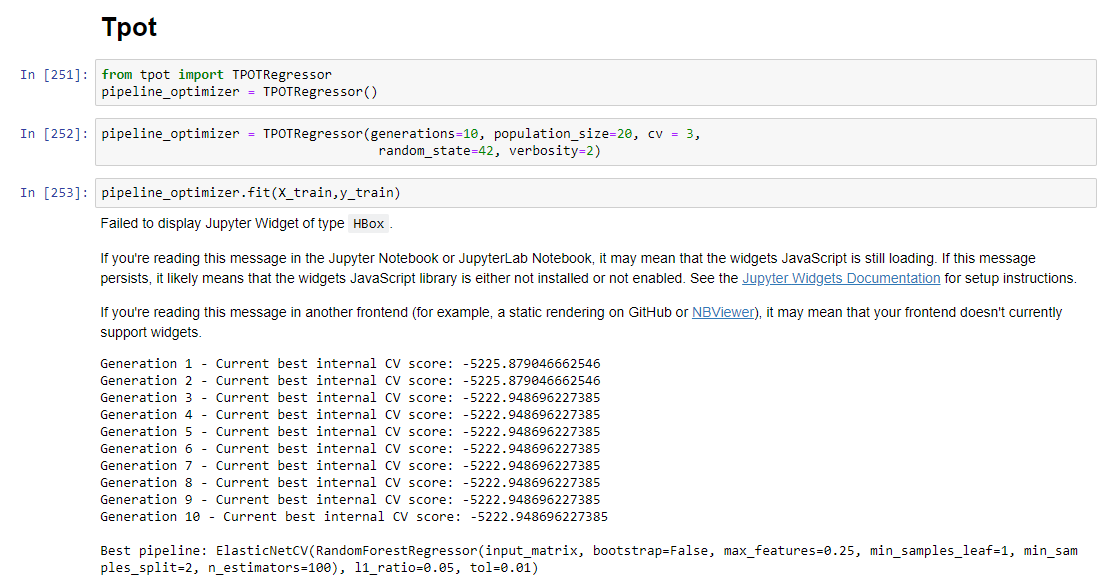
In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features variables, predictors) for use in model construction. Feature selection techniques are used for four reasons:

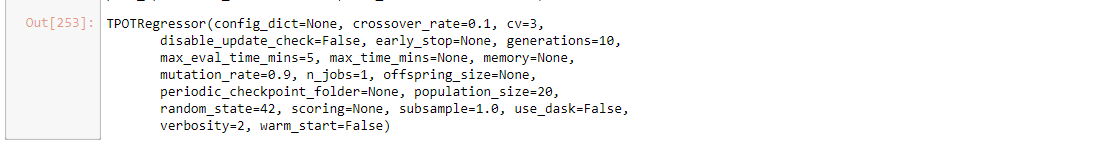
• simplification of models to make them easier to interpret by researchers/users,  
• shorter training times,  
• enhanced generalization by reducing overfitting (formally, reduction of variance)

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.

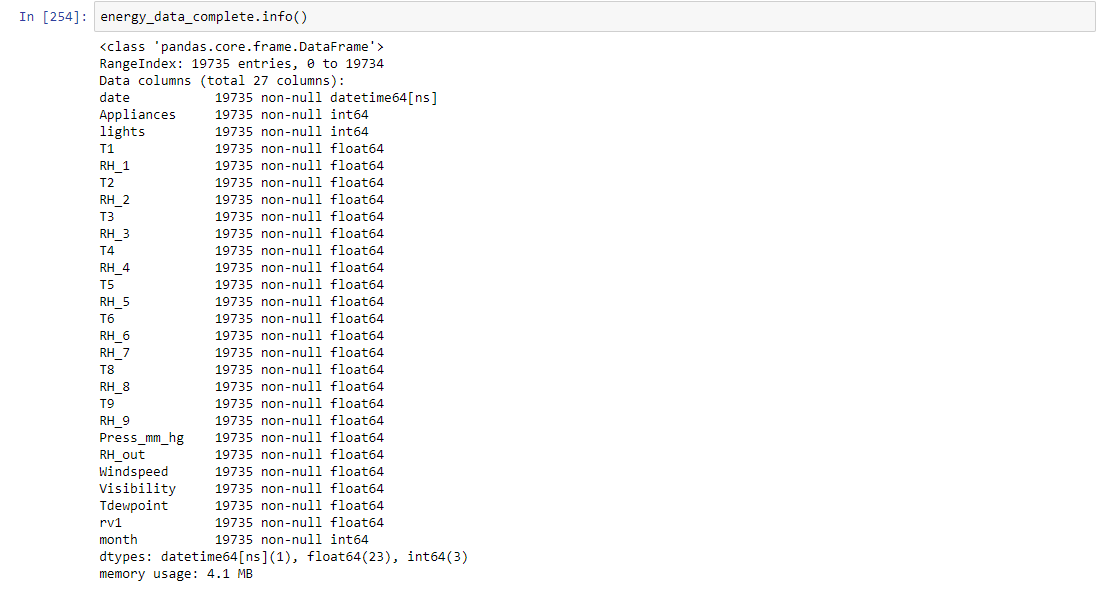
**TPOT:**

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



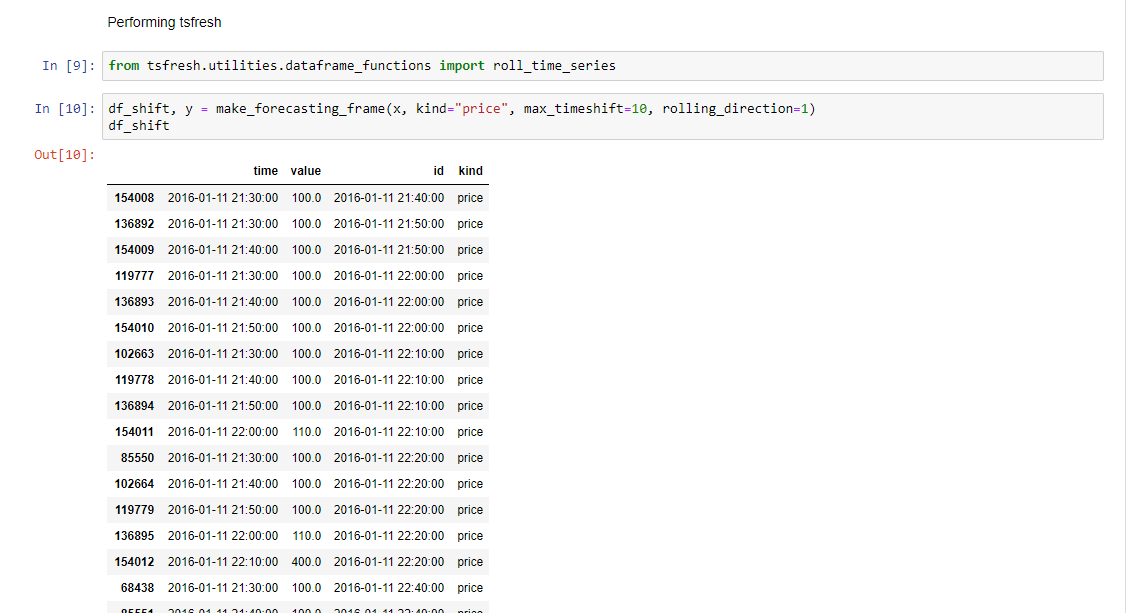
When we get the output, we check the accuracy of score.

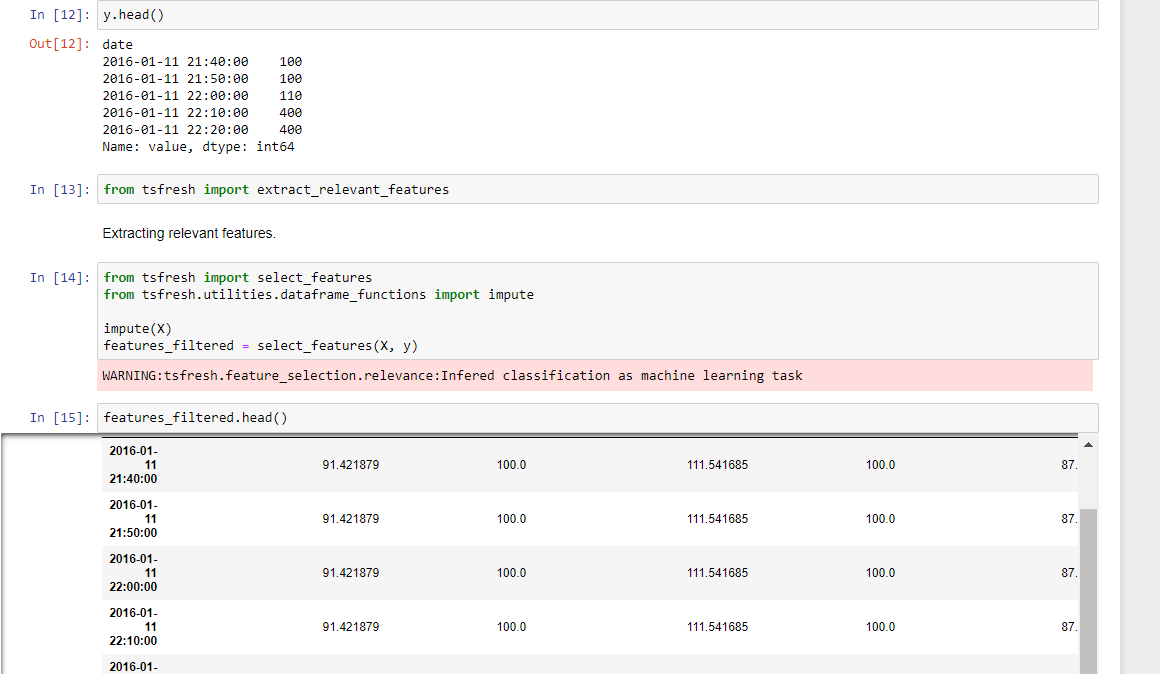


**TSFresh**

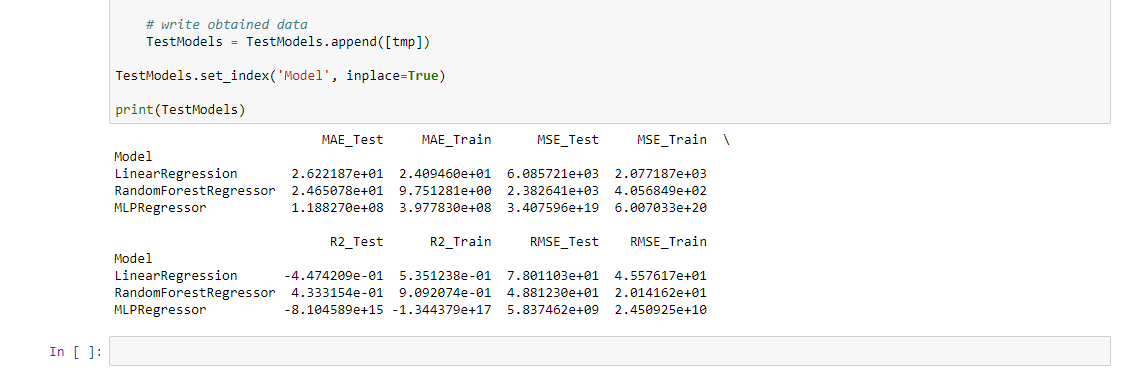
tsfresh is a python package. It automatically calculates a large number of time series characteristics, the so called features. Further the package contains methods to evaluate the explaining power and importance of such characteristics for regression or classification tasks.

We start by creating the model for testing.





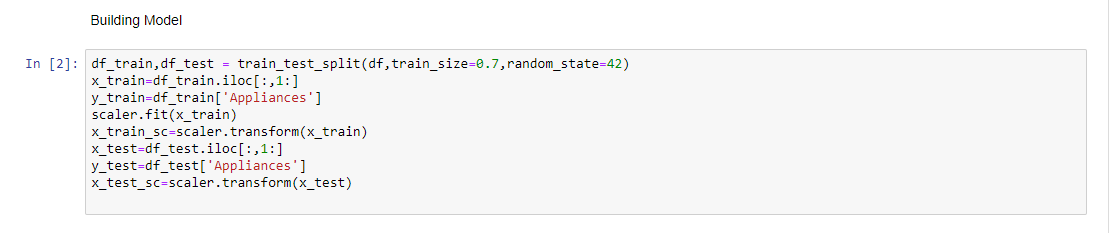
Getting final accuracy.



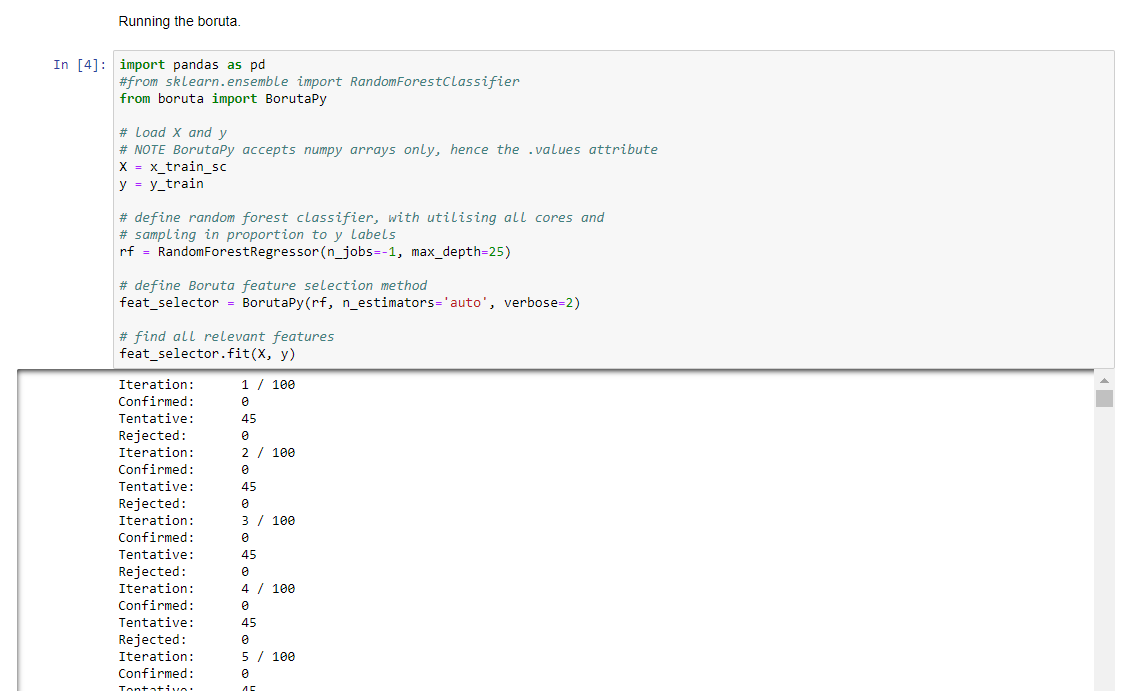
**Boruta**

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).

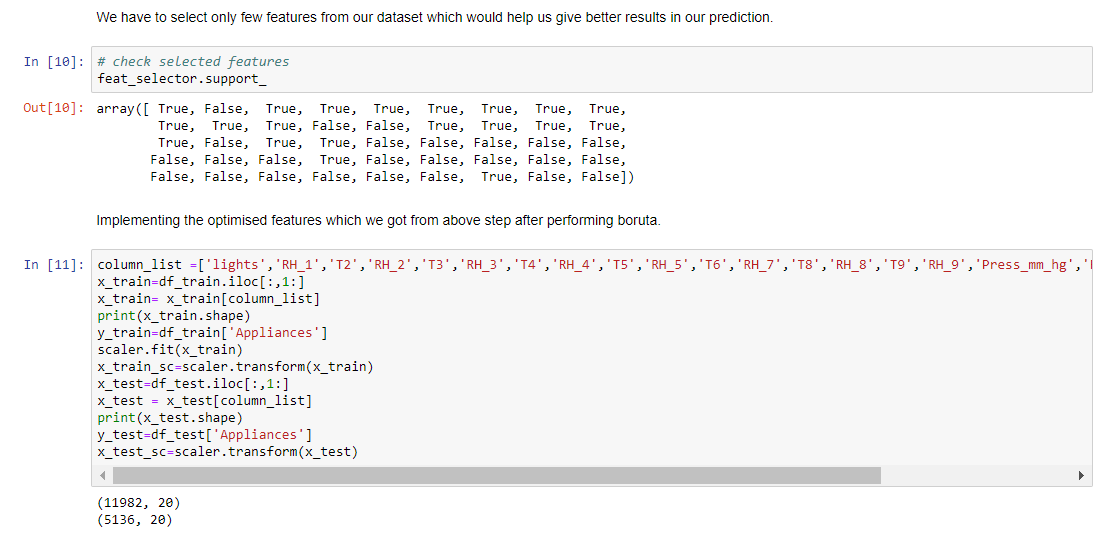
First, build the model.



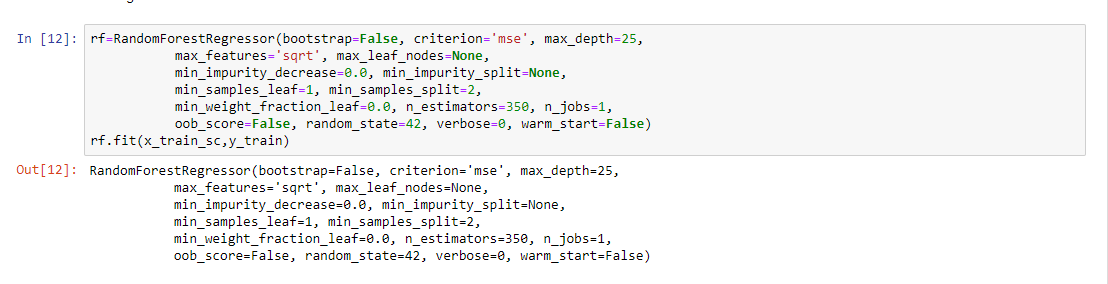
Running Boruta on this model



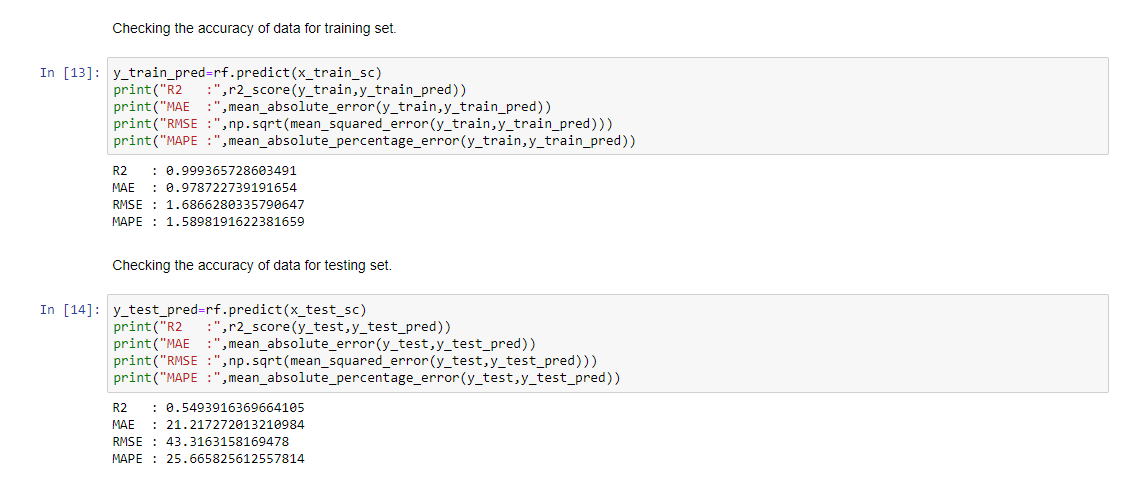
We need to select only helping features from our model.



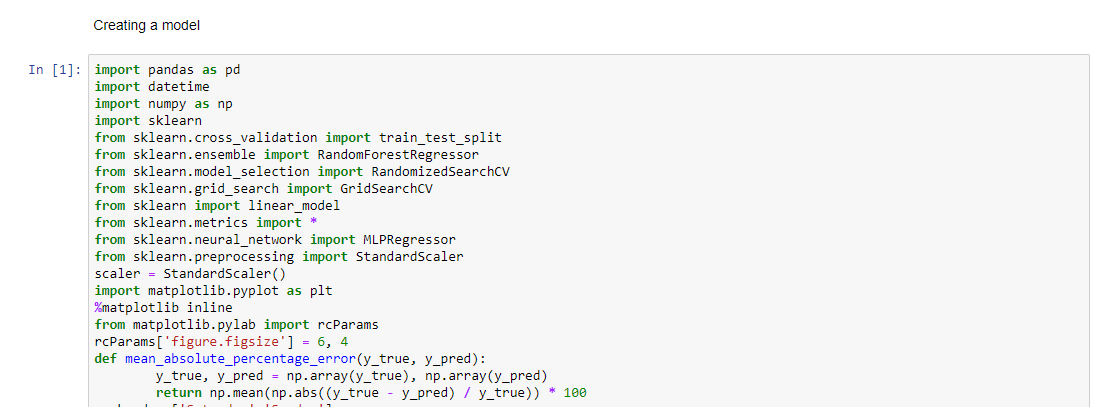
Making a model for random forest

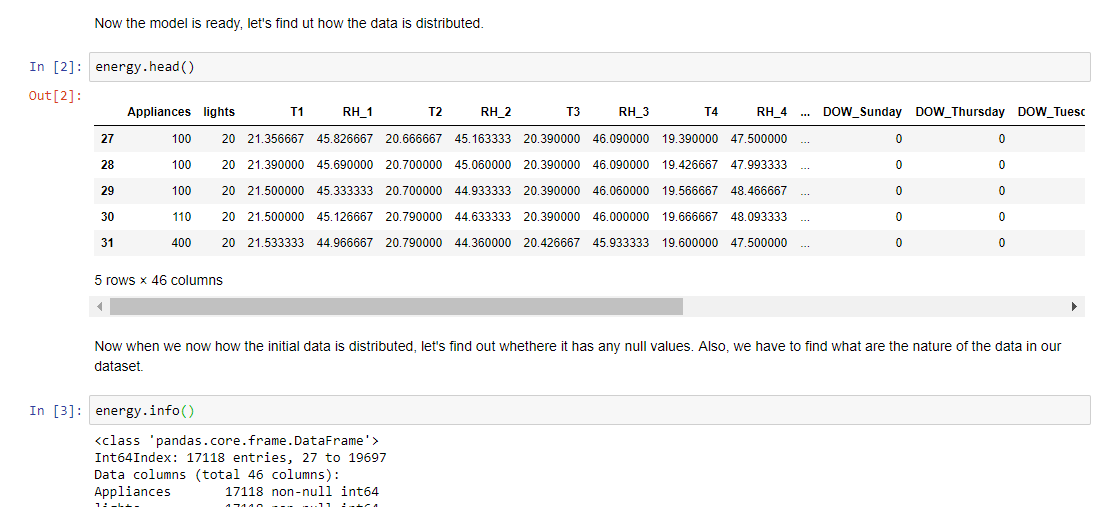


Checking the accuracy of training and testing data



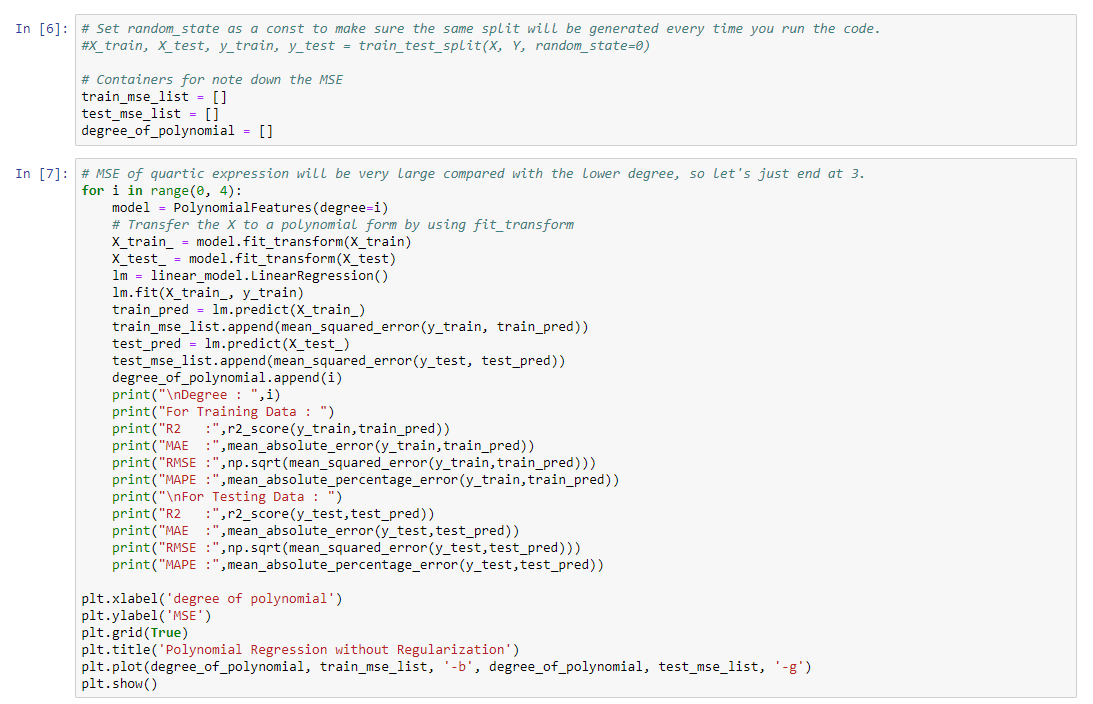
**6. Model Validation and Selection**

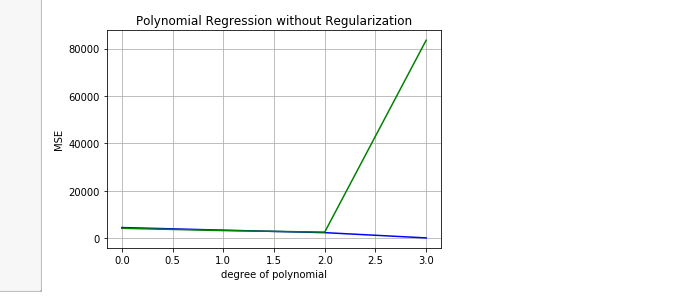






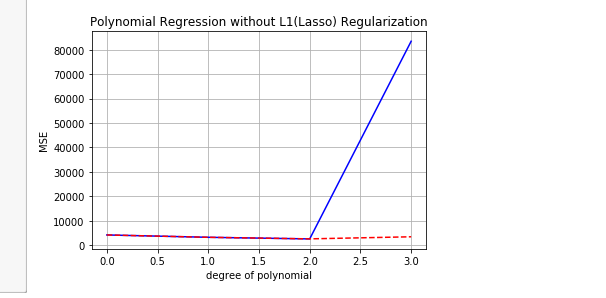
Regularization





Performing Lasso





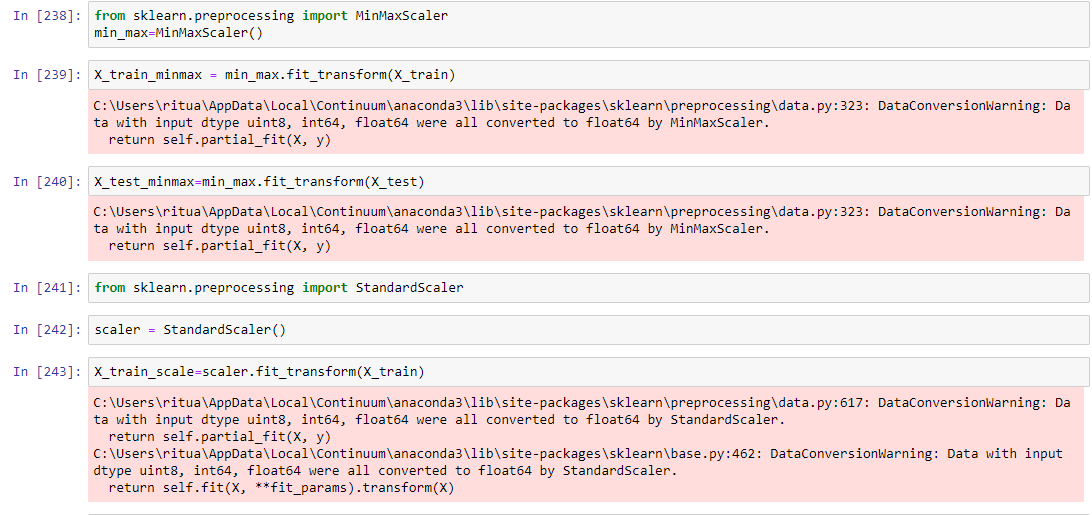
Cross Validation Technique



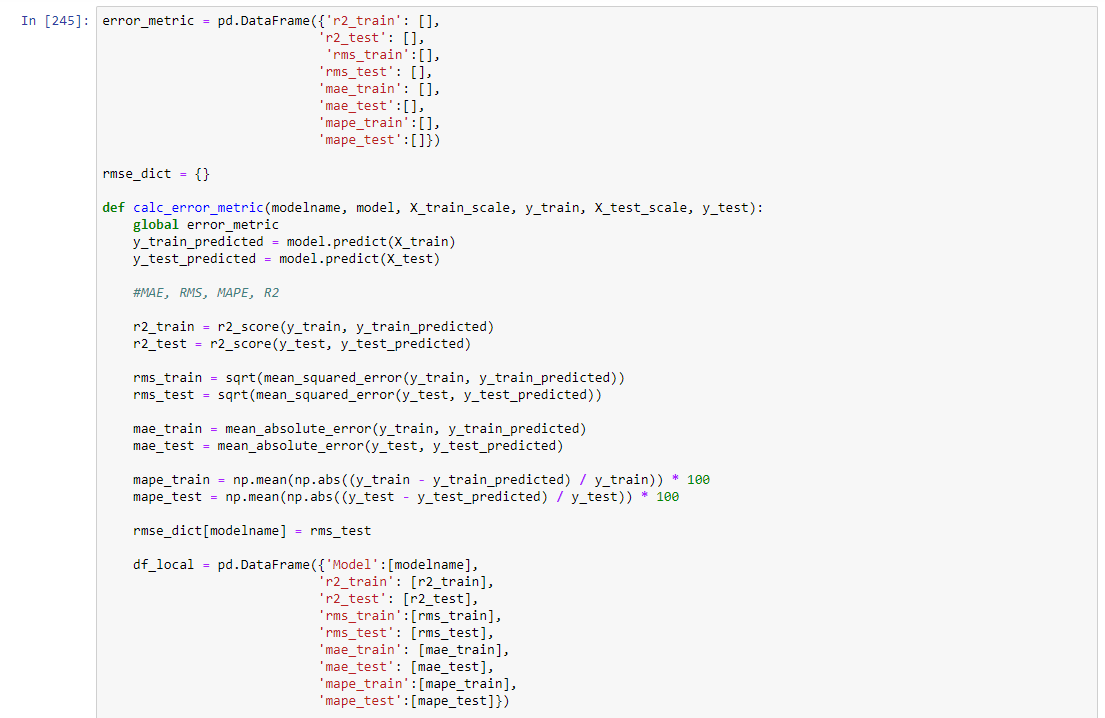
Getting Accuracy of model



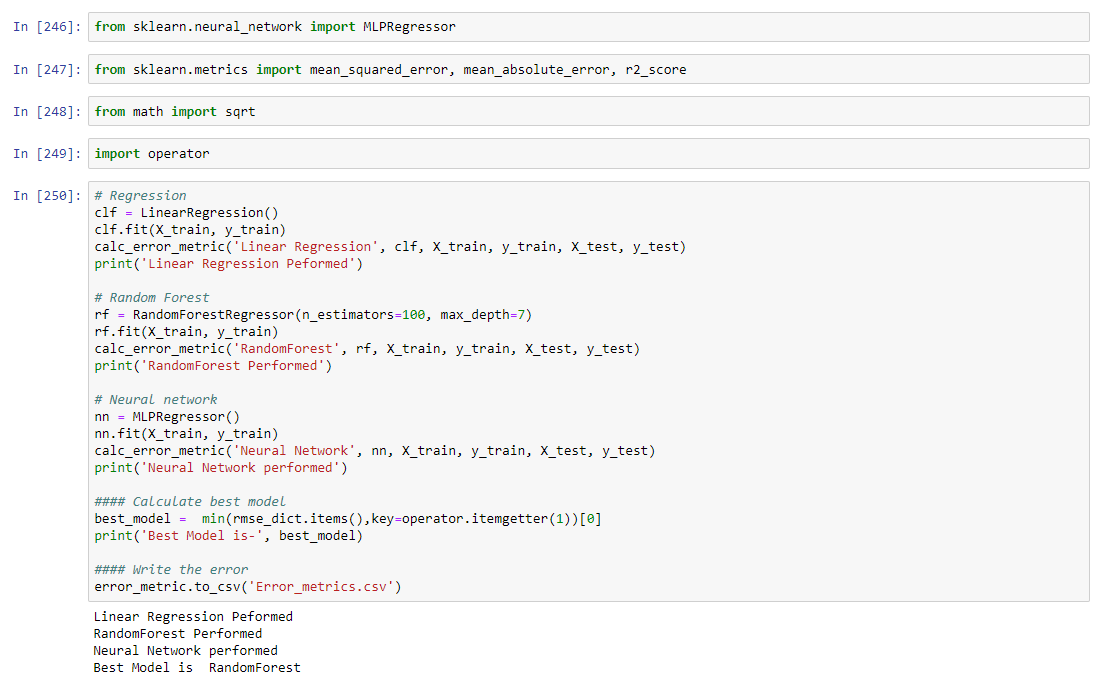
**7 Final Pipeline**

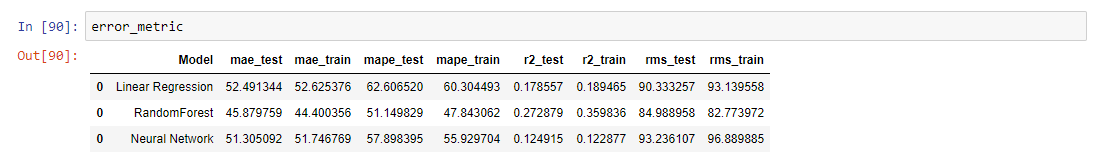


Error metric

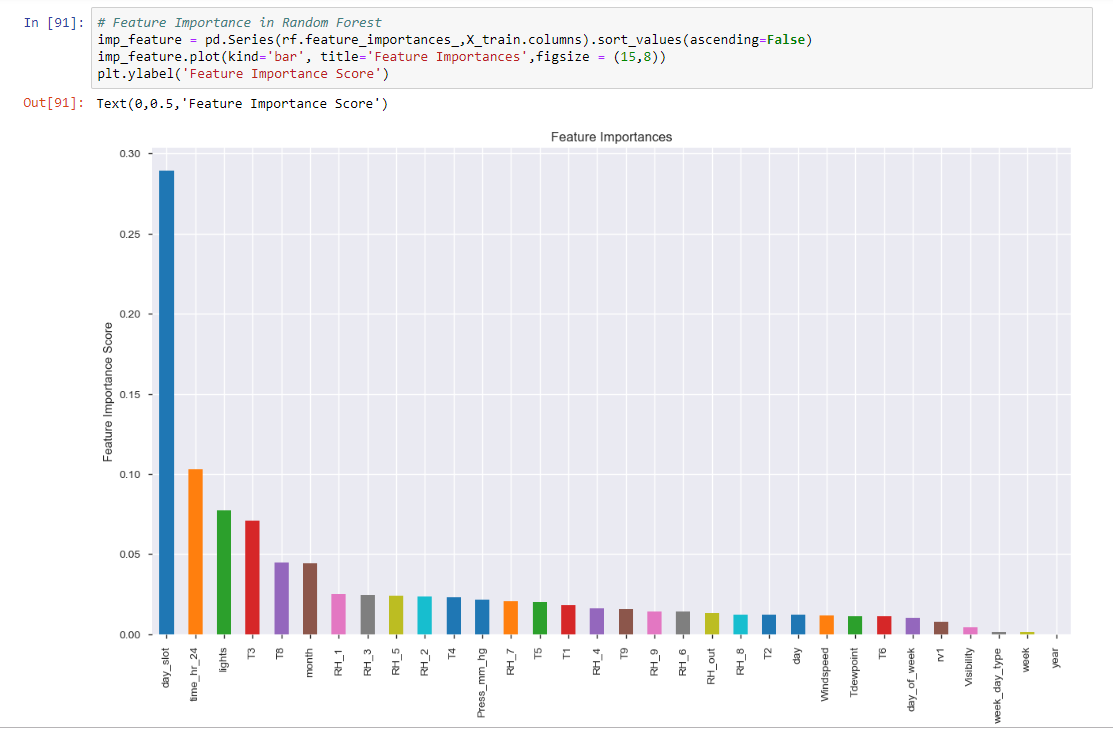


Getting best model.

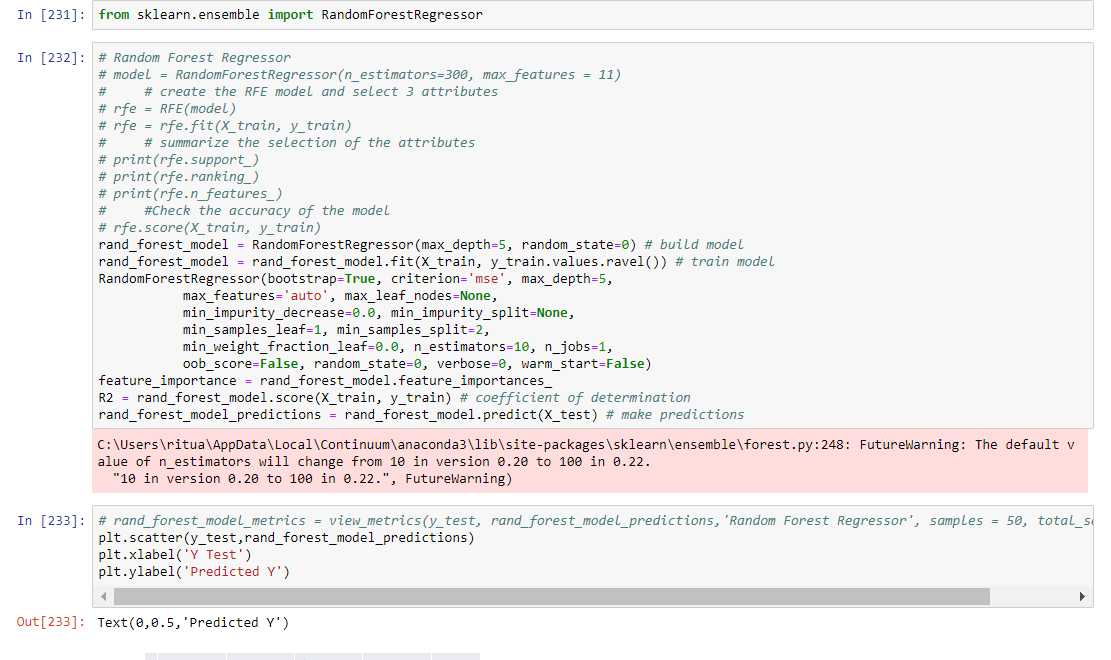


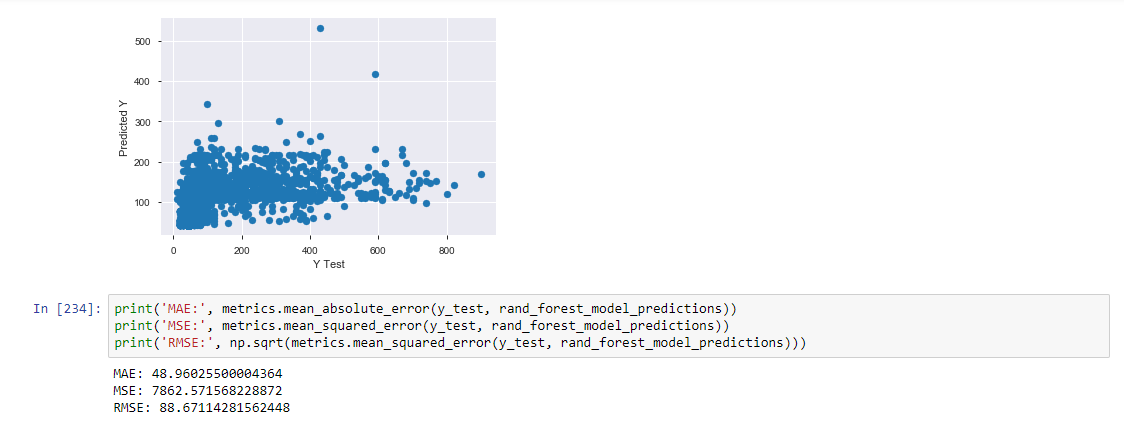


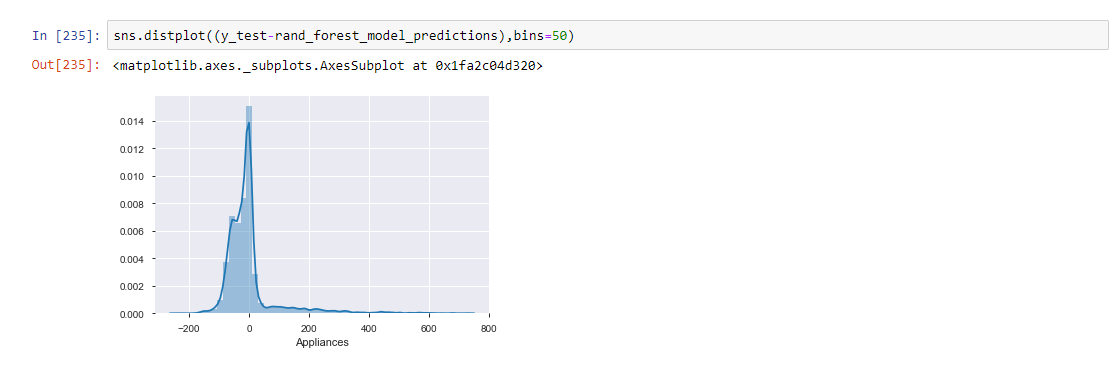
Getting feature importance of random forest



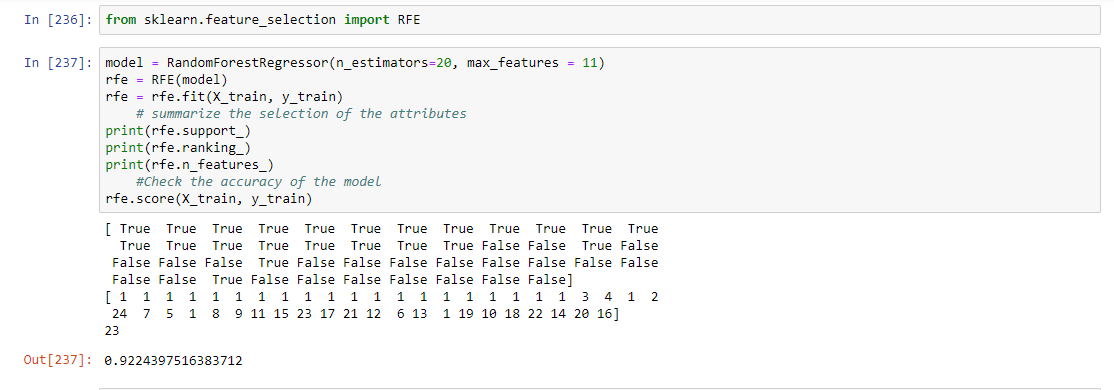
Doing prediction by random forest.







Predicting using Random forest algorithms



**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.