**Capstone Project -1**

**Appliances Energy Prediction**

Business Problem Description – Dataset contains the house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. As per the description on UCI website, 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. Combining this data with the weather data based on the date time columns

To find the key feature from the dataset which contributes to the most and the

1. Predict the appliance energy consumption.
   1. With collected data of temperature and humidity (indoor and outdoor) sensors.
   2. Weather data collected
   3. Fuel price over the period of time.
2. Best prediction model with best parameter for future prediction of the appliance Energy.

**Dataset Details –**

We have two data sets - **energydata\_complete.csv** and **CrudeOilPrice.csv**. We have

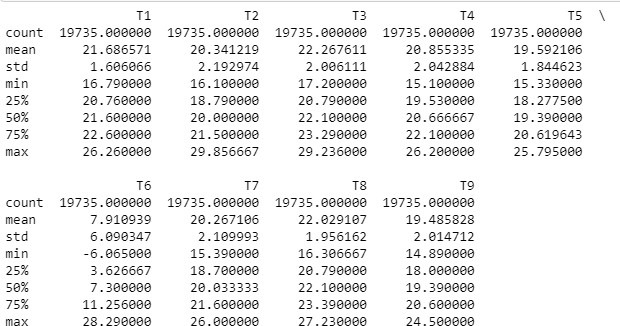
taken two different dataset to get better prediction with analyzing the engorge consumed and how was the fuel price during the particular date.

We do not have any missing values in energydata\_complete.csv; it has 19735 observation with 29 attributes pertaining to temperature, humidity, light, wind speed, dew, and visibility from local weather channel.

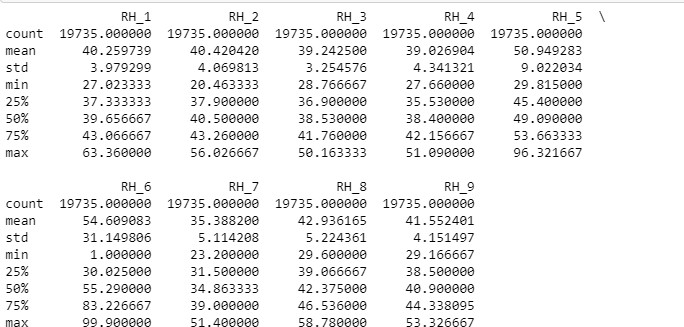
We do not have any missing value in CrudeOilPrice.csv, which has the fuel price for respective months and dates. This dataset has 2519 observation and 2 attributes of date and fuel price.

Few Key observation are as below –

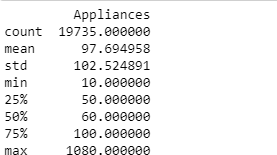
1. The dataset is from 2016-01-11 and 2016-05-27; have data starting JAN to MAY of 2016.
2. These are the temperature reading captured inside and outside the house. From the explored reading of each sensor is between 14.89 and 29.85 but ‘**T6’** is between -6 and 28.29. The possible reason can be its reading are for outside.



1. There are Humidity related information as well in the dataset, from the explored reading of each sensor is between 20.46 to 58.79 but ‘**RH\_5’** and ‘**RH\_6’** has max of 96.32 and 99.9.



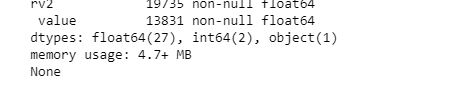
1. The max value is 1080wh, whereas 75% of usage is under 100wh. Some of the appliances has high consumption. These can be outliers but, currently keeping them as part of the dataset and not dropping them from the dataset.
2. If we see the statistics for Appliance Attributes, the minimum value is 10 and max value is 1080, and the mean is 97.69 and 75% of records are below 100 KWH. This column has outliers and we will keep them and check during our modeling.



When merging the two datasets, in energydata dataset, date is a timestamp and

in crudeoilprice dataset, date is a date datatype, so we have to normalize the date, in order for us to merge the two datasets.

1. After the merge, we observe that “ values" columns is merged on the dataset, but it doesn’t have all the dates values and 5904 records has null values.



To solve these null values, we used the “**forward fill” method** and value column was populated with previous day values for the records, which were null and renamed the column to "oilprice".

The total number of observation is 19735 and 30 Attributes.

There are 19735 and 30 attributes. Key features from the dataset are

|  |  |
| --- | --- |
| **Columns** | **Description** |
| date time | year-month-day hour:minute:second |
| Appliances | energy use in Wh |
| lights | energy use of light fixtures in the house in Wh |
| T1 | Temperature in kitchen area in Celsius |
| RH\_1 | Humidity in kitchen area in % |
| T2 | Temperature in living room area in Celsius |
| RH\_2 | Humidity in living room area in % |
| T3 | Temperature in laundry room area |
| RH\_3 | Humidity in laundry room area in % |
| T4 | Temperature in office room in Celsius |
| RH\_4 | Humidity in office room in % |
| T5 | Temperature in bathroom in Celsius |
| RH\_5 | Humidity in bathroom in % |
| T6 | Temperature outside the building (north side) in Celsius |
| RH\_6 | Humidity outside the building (north side) in % |
| T7 | Temperature in ironing room in Celsius |
| RH\_7 | Humidity in ironing room in % |
| T8 | Temperature in teenager room 2 in Celsius |
| RH\_8 | Humidity in teenager room 2 in % |
| T9 | Temperature in parents room in Celsius |
| RH\_9 | Humidity in parents room in % |
| To | Temperature outside (from Chievres weather station) in Celsius |
| Pressure (from Chievres weather station) | in mm Hg |
| RH\_out | Humidity outside (from Chievres weather station) in % |
| Wind speed (from Chievres weather station) | in m/s |
| Visibility (from Chievres weather station) | in km |
| Tdewpoint (from Chievres weather station) | Â°C |
| rv1 | Random variable 1, nondimensional |
| rv2 | Random variable 2, nondimensional |

Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru). There are 2 random variables, which has be explored more.

Another dataset used is to about the fuel price on the year of 2016, integrating this with overall dataset will help understand the fuel price impact the appliance energy consumption.

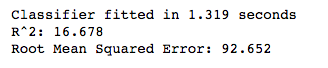
Reference data source –

* <https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction>
* <https://www.macrotrends.net/2516/wti-crude-oil-prices-10-year-daily-chart>

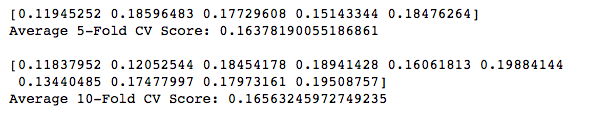
**Approach –**

**Model Building and Implementation –**

1. As identified earlier – we are dropping the below field –
   1. 'date\_x', 'appliances','rv1', 'rv2','t6','t9' from the dataset.
   2. Created the X will all the features and Y with the target feature.
   3. Using train\_test\_split method, we have done the split of the dataset in 70% Training data and 30% test data.
   4. Upon running the Linear regression model- we get the below score for R^2 and RMSE(Root Mean Square Error)

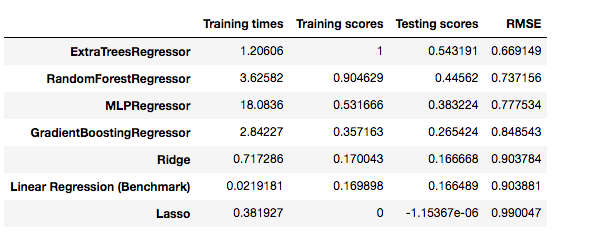


1. Also, with Cross validation, we didn’t see the improvement in the performance of the benchmark algorithms –

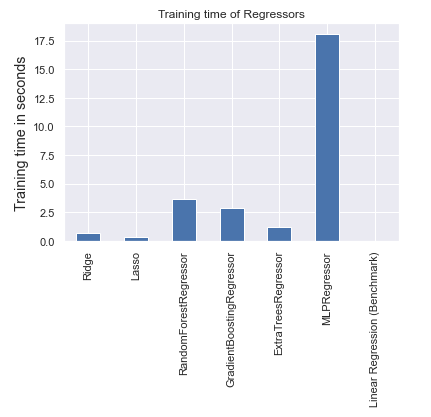


1. Now, we will try to scale the data and find best performing model –
   1. Dropped the x\_date feature from the dataset, and using the StandardScaler method, scaled the dfactual dataframe.
   2. From the scaled dataset, dropped the 'appliances','rv1', 'rv2','t6','t9'.
   3. Created the Training and Test Dataset with 70-30% ratio.
2. Create the following models with key important features –

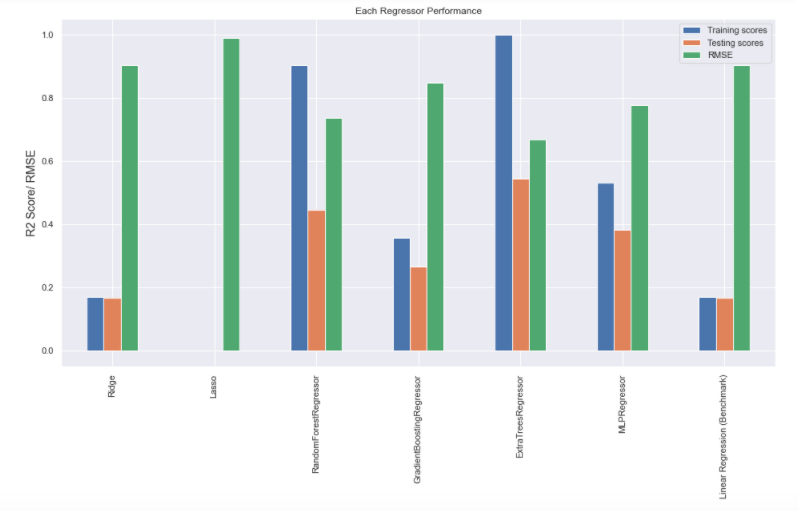
* Regularized linear models as an improvement over Linear Regression.
  + Ridge Regression
  + Lasso Regression
* Ensemble based Tree Regression models, which deal with number of features and outlier data.
  + Random Forests
  + Gradient Boosting
  + Extra Trees
* Neural networks for non-linear relationships target feature and predictors.
  + Multi-Layer Perceptron
  1. Implemented them in a iterative manner with creating different functions
     1. Created function to capture the fit and predict the models and capture the score and accuracy for the models
     2. Created the pipeline and passed all the algorithms to be executed in the above function.
     3. Created a function to display and store the results / outcome.
     4. Below is the results displaying the R^2 and RMSE and time it took to predict.



* 1. Comparing the Training time –



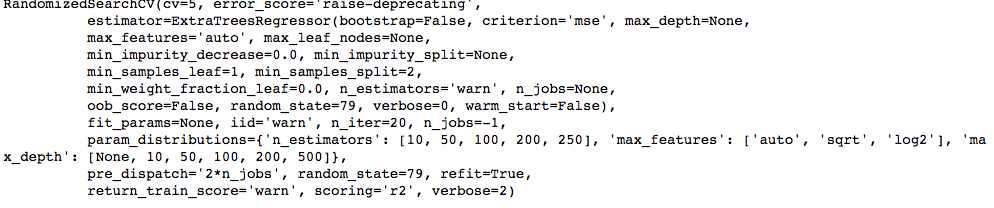
* 1. Plot to compare the performance of the algorithms on datasets



Interpretation –

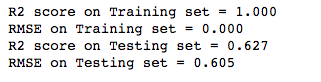
Least performing Repressor - Lasso Repressor and best performing Repressor - Extra Trees Repressor. Even though Extra Trees Repressor has a R2 score of 1.0 on training set, which might suggest over-fitting but, it has the highest score on test set and also, it's RMSE value is also the lowest. Clearly, ExtraTreesRegressor is the best model out of given models.

* 1. Hyper-parameter tuning the best Model – “ExtraTreesRegressor” observed from above step – Using the RandomizedSearchCV, we will find the best estimators and using those estimators we will perform the prediction.

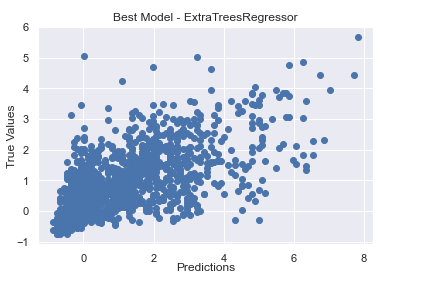




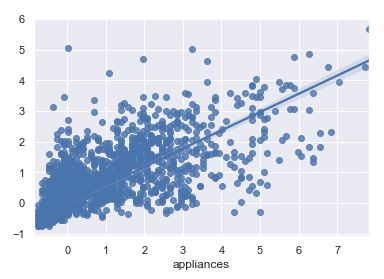
* 1. Using the best parameter we will fit and predict the training data and predict on test data.

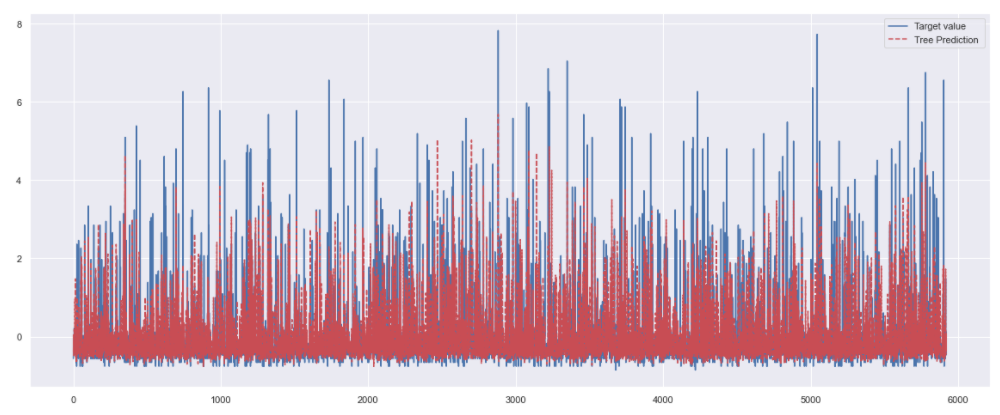


* 1. Plotting the data for y\_test\_s and predicted data –



Using the seaborn plot using regplot function

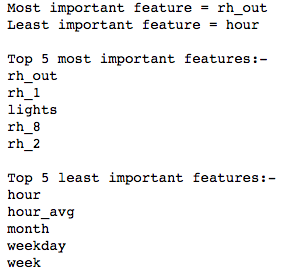


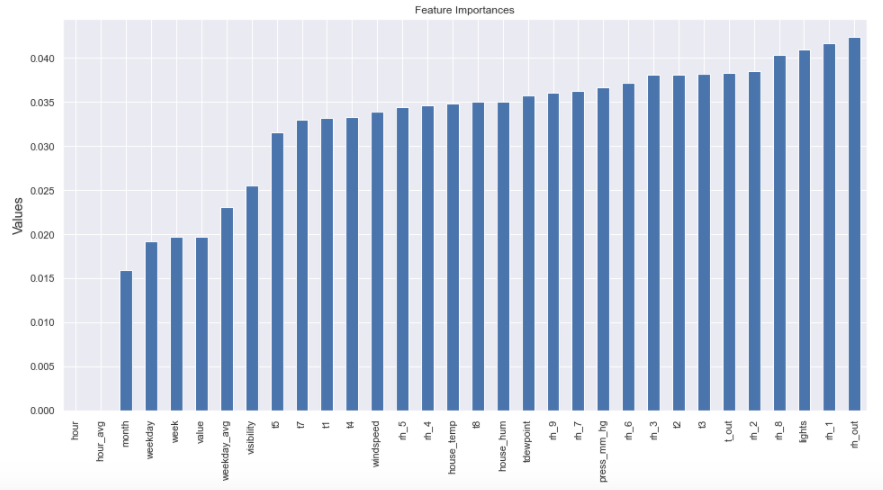


Overlaying the test data and predicted data, we can see that the prediction on not so accurate.

### Interpretation from Implementation -

* R2 score improvement compared to Benchmark model = 0.463.
* RMSE improvement compared to Benchmark model = 0.301.
* R2 score improvement compared to without tuned model = 0.086.
* RMSE improvement compared to without tuned model = 0.066.
  1. Important features contributing from the data set are as below –

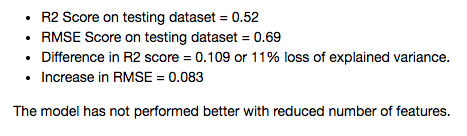




1. Feature and Model Evaluation-
   * + 1. Clone the above best model clone with the 'rh\_out', 'rh\_1', 'lights', 'rh\_8', 'rh\_2', 't\_out', 't3', 't2' and do a prediction only with the most important feature, to verify if there is any improvement in the model accuracy.



* + - 1. Comparing these results with above best performing algorithms – Extratreeregressor



1. **Conclusion** -
   1. Best Algorithm = Extra Trees Regressor
   2. Variance explained on test set = 63%.
   3. RMSE error = 60.3%