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.
    a. With collected data of temperature and humidity (indoor and outdoor) sensors.
    b. Weather data collected
    c. 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.

	T1	T2	T3	T4	T5	\
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	
mean	21.686571	20.341219	22.267611	20.855335	19.592106	
std	1.606066	2.192974	2.006111	2.042884	1.844623	
min	16.790000	16.100000	17.200000	15.100000	15.330000	
25%	20.760000	18.790000	20.790000	19.530000	18.277500	
50%	21.600000	20.000000	22.100000	20.666667	19.390000	
75%	22.600000	21.500000	23.290000	22.100000	20.619643	
max	26.260000	29.856667	29.236000	26.200000	25.795000	

	T6	T7	T8	T9
count	19735.000000	19735.000000	19735.000000	19735.000000
mean	7.910939	20.267106	22.029107	19.485828
std	6.090347	2.109993	1.956162	2.014712
min	-6.065000	15.390000	16.306667	14.890000
25%	3.626667	18.700000	20.790000	18.000000
50%	7.300000	20.033333	22.100000	19.390000
75%	11.256000	21.600000	23.390000	20.600000
max	28.290000	26.000000	27.230000	24.500000

3. 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.

	RH_1	RH_2	RH_3	RH_4	RH_5	\
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	
mean	40.259739	40.420420	39.242500	39.026904	50.949283	
std	3.979299	4.069813	3.254576	4.341321	9.022034	
min	27.023333	20.463333	28.766667	27.660000	29.815000	
25%	37.333333	37.900000	36.900000	35.530000	45.400000	
50%	39.656667	40.500000	38.530000	38.400000	49.090000	
75%	43.066667	43.260000	41.760000	42.156667	53.663333	
max	63.360000	56.026667	50.163333	51.090000	96.321667	

	RH_6	RH_7	RH_8	RH_9
count	19735.000000	19735.000000	19735.000000	19735.000000
mean	54.609083	35.388200	42.936165	41.552401
std	31.149806	5.114208	5.224361	4.151497
min	1.000000	23.200000	29.600000	29.166667
25%	30.025000	31.500000	39.066667	38.500000
50%	55.290000	34.863333	42.375000	40.900000
75%	83.226667	39.000000	46.536000	44.338095
max	99.900000	51.400000	58.780000	53.326667

4. 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.

5. 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.

	Appliances
count	19735.000000
mean	97.694958
std	102.524891
min	10.000000
25%	50.000000
50%	60.000000
75%	100.000000
max	1080.000000

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.

```
rv2             19735 non-null float64
 value          13831 non-null float64
dtypes: float64(27), int64(2), object(1)
memory usage: 4.7+ MB
None
```

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)

```
Classifier fitted in 1.319 seconds
R^2: 16.678
Root Mean Squared Error: 92.652
```

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

```
[0.11945252 0.18596483 0.17729608 0.15143344 0.18476264]
Average 5-Fold CV Score: 0.16378190055186861

[0.11837952 0.12052544 0.18454178 0.18941428 0.16061813 0.19884144
 0.13440485 0.17477997 0.17973161 0.19508757]
Average 10-Fold CV Score: 0.16563245972749235
```

3. 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.
4. Create the following models with key important features –
    - Regularized linear models as an improvement over Linear Regression.

    - Ridge Regression
    - Lasso Regression

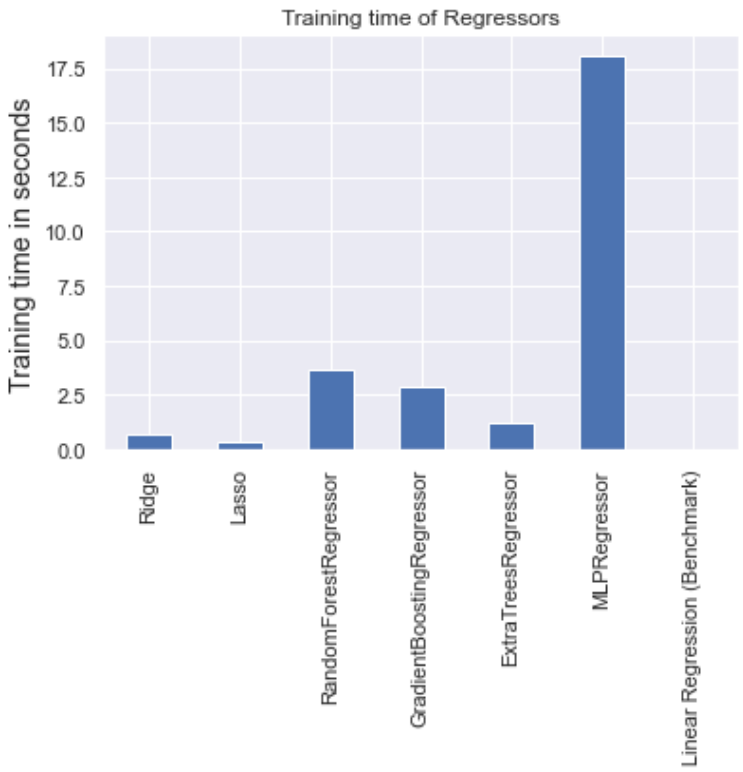
- Ensemble based Tree Regression models, which deal with number of features and outlier data.
    - Random Forests
    - Gradient Boosting
    - Extra Trees

- Neural networks for non-linear relationships target feature and predictors.
    - Multi-Layer Perceptron

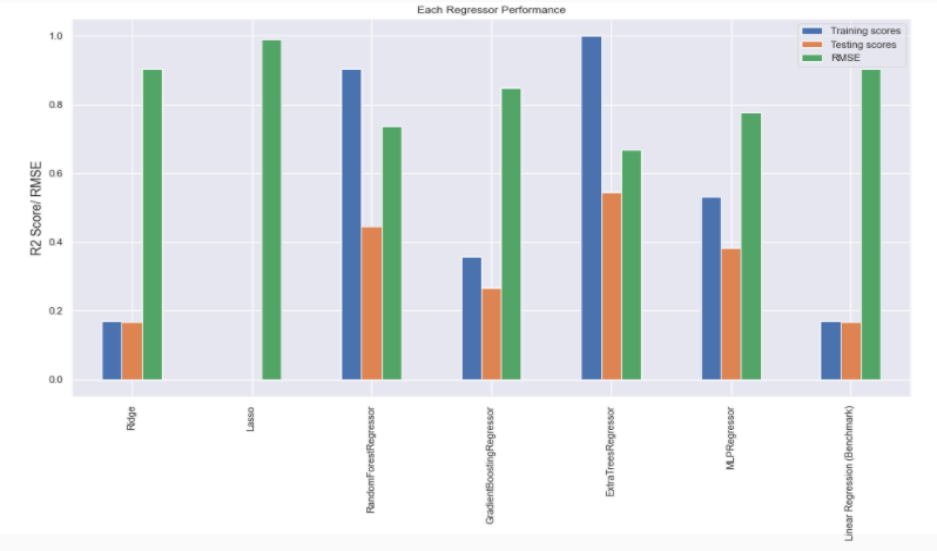
1. Implemented them in a iterative manner with creating different functions
    1. Created function to capture the fit and predict the models and capture the score and accuracy for the models
    2. Created the pipeline and passed all the algorithms to be executed in the above function.
    3. Created 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.

	Training times	Training scores	Testing scores	RMSE
ExtraTreesRegressor	1.20606	1	0.543191	0.669149
RandomForestRegressor	3.62582	0.904629	0.44562	0.737156
MLPRegressor	18.0836	0.531666	0.383224	0.777534
GradientBoostingRegressor	2.84227	0.357163	0.265424	0.848543
Ridge	0.717286	0.170043	0.166668	0.903784
Linear Regression (Benchmark)	0.0219181	0.169898	0.166489	0.903881
Lasso	0.381927	0	-1.15367e-06	0.990047

2. Comparing the Training time –



3. 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.

4. 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.

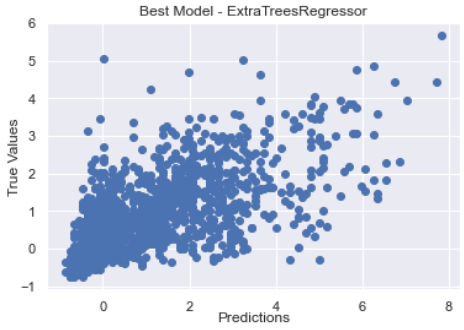
```
RandomizedSearchCV(cv=5, error_score='raise-deprecating',
          estimator=ExtraTreesRegressor(bootstrap=False, criterion='mse', max_depth=None,
          max_features='auto', max_leaf_nodes=None,
          min_impurity_decrease=0.0, min_impurity_split=None,
          min_samples_leaf=1, min_samples_split=2,
          min_weight_fraction_leaf=0.0, n_estimators='warn', n_jobs=None,
          oob_score=False, random_state=79, verbose=0, warm_start=False),
          fit_params=None, iid='warn', n_iter=20, n_jobs=-1,
          param_distributions={'n_estimators': [10, 50, 100, 200, 250], 'max_features': ['auto', 'sqrt', 'log2'], 'ma
x_depth': [None, 10, 50, 100, 200, 500]},
          pre_dispatch='2*n_jobs', random_state=79, refit=True,
          return_train_score='warn', scoring='r2', verbose=2)


 Parameters of best Regressor : {'n_estimators': 250, 'max_features': 'log2', 'max_depth': None}
```

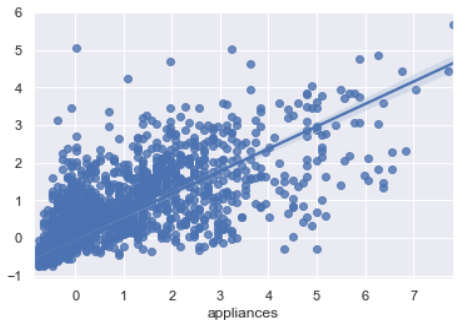
5. Using the best parameter we will fit and predict the training data and predict on test data.

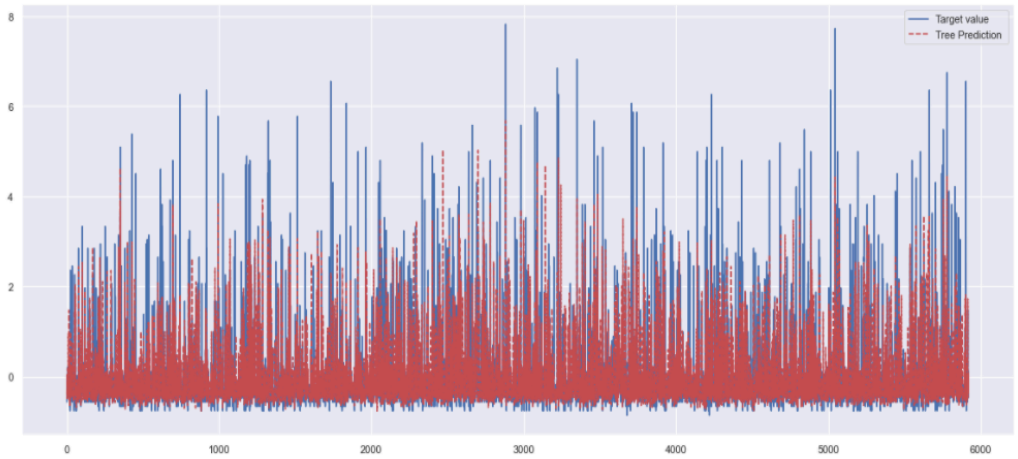
```
R2 score on Training set = 1.000
RMSE on Training set = 0.000
R2 score on Testing set = 0.627
RMSE on Testing set = 0.605
```

6. 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 -

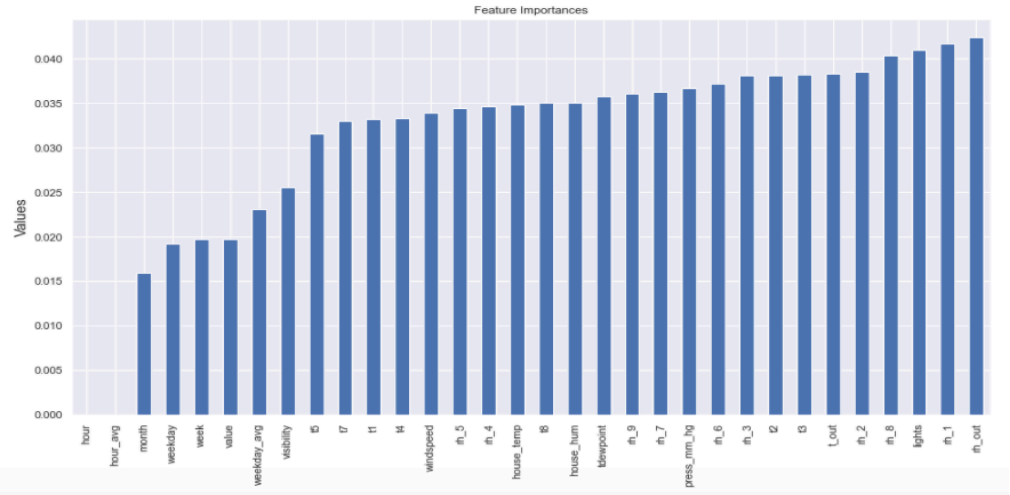
- R2 score improvement compared to Benchmark model = 0.463.
- RMSE improvement compared to Benchmark model = 0.301.
- R2 score improvement compared to without tuned model = 0.086.
- RMSE improvement compared to without tuned model = 0.066.

7. Important features contributing from the data set are as below –

```
Most important feature = rh_out
Least important feature = hour

Top 5 most important features:-
rh_out
rh_1
lights
rh_8
rh_2

Top 5 least important features:-
hour
hour_avg
month
weekday
week
```



5. 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.

```
R2 Score on testing dataset = 0.519
RMSE Score on testing dataset = 0.686
```

2. Comparing these results with above best performing algorithms – Extratreeregressor

   - R2 Score on testing dataset = 0.52
   - RMSE Score on testing dataset = 0.69
   - Difference in R2 score = 0.109 or 11% loss of explained variance.
   - Increase in RMSE = 0.083

   The model has not performed better with reduced number of features.

5. **Conclusion** -

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