REPORT ANALYSIS OF THE UCI_ELECTRICITY DATA MODELLING

Name: Subhasree Mohapatra
Student ID: s3779215

Objective:

The objective of the study is to understand the time series dataset for UCI energy use of Appliances along with various factors of impact like Temperature and Relative Humidity of internal environment, Pressure, Windspeed, Visibility of external environment and develop models for prediction for energy use. The problem comes under regression analysis and hence regression models will be constructed.

Data Preparation and Exploration:

The data was studied, and it was noted that the data was recorded daily for every 10 minutes from January to May 2016 with 27 features and 19735 instances of data.

date	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4	T5	RH_5	...	RH_9	T_out	Press_mm_hg
2016-04-19 20:30:00	22.200000	39.500000	20.566667	37.656667	22.230000	37.030000	22.318571	36.610000	20.633333	62.166667	...	33.90	9.70	766.100000
2016-03-05 04:40:00	20.356667	37.126667	17.566667	40.230000	20.890000	37.663333	18.700000	36.260000	18.463333	43.560000	...	41.09	0.30	740.333333
2016-03-14 12:40:00	20.926667	38.790000	21.100000	35.526667	21.600000	36.290000	21.000000	34.826667	18.100000	46.126667	...	38.76	4.40	768.466667
2016-01-22 15:30:00	18.290000	38.900000	17.290000	39.260000	18.390000	39.326667	16.100000	38.790000	16.100000	47.700000	...	39.20	3.35	760.600000
2016-02-10 00:40:00	22.290000	42.333333	21.600000	40.433333	22.666667	43.363333	19.100000	40.900000	19.290000	50.745000	...	43.73	3.20	738.900000

Figure 1: Dataset View

Based on DateTime indexing that was performed on the dataset for time series analysis, more features like Hours, Minutes, Day of the Year, Day of the Week, Year, and Month were constructed from data and added to it, in order to understand the behaviour of the energy use as per the time and date. The new dataset contained 33 features.

The data were seen to be varied with the newly constructed features like Hours and Weekday, which were plotted later on. While understanding the features it was found that some features tended to be distributed normally in consideration to Central Limit Theorem, while some do show bimodal distribution and skewness hence to attend a uniformness, normalization through MinMaxScaler was carried out, so that all features do fall under the bell-shaped curve. Not much of outliers were noted. However, we did notice some outliers in the target variable considering an hourly and daily basis consumption distribution . The energy consumption does show a variation with an hour and Weekday as plotted in the report. (Refer to the attached Jupyter notebooks in the zip file for boxplot and histogram plotting of features for outlier detection and normalization study).

The data was studied based on the correlation heat map as below(**Appendix A**) which described the strength and direction of the relationship and it was noted that the Temperatures bear a strong correlation between themselves similar to the Relative Humidity which bears a moderate correlation. A strong correlation can be seen between variable T_out and T_6 of 0.97, while T_9 bears a strong correlation T3, T5, T7, T8 along with this it was seen that many features do share less correlated relation with Target variable except the 'Hours' of the day which shared a comparatively higher relation but still these relations do not show strong linearity with the target variable. Moreover, Visibility share 0.00 correlation with Target variable indicating that feature selection can play a significant role in reducing the dimensionality of the dataset.

The behaviour of the target variable and other features were plotted and found that they do tend towards non-linearity. The features do not show a strong Linear Relationship. We also found some specifics as for the Hours of the

day, the most energy was to be used at 8 AM and then in-between 3 PM to 9 PM and then gradually slows down through the night. Likewise, the energy use varies with the month of the year, most likely due to season fluctuation. (**Appendix C**)To study the dataset we separated the dataset initially into two parts as for training, validation set for training and validation purpose and 10% of data has been kept aside for test purpose for the selected model. After this, the data was resampled on an hourly basis by performing down sampling over the dataset.

Feature Selection:

With our basic knowledge we know that the factors affecting the energy consumption can be Temperature, Humidity, Pressure, Time of the day. To be more consistent and accurate in feature selection we did the feature selection with the SelectKBest Model which prioritizes features on the p-value and correlation of the feature as per Pearson's test. Here the Pearson's correlation has been used to identify 20 variables that pose a greater efficiency in predicting the target variable, which reduced nearly 40% of the total variables which can be quite significant in handling the dimensionality of the dataset. Some of the highly correlated variables were removed and the features which bore no relationship were also eliminated. Features with p-values with a value greater than 0.05 are considered candidates for elimination. The features did include temperatures, humidity, pressure, hours, in common along with other selections.

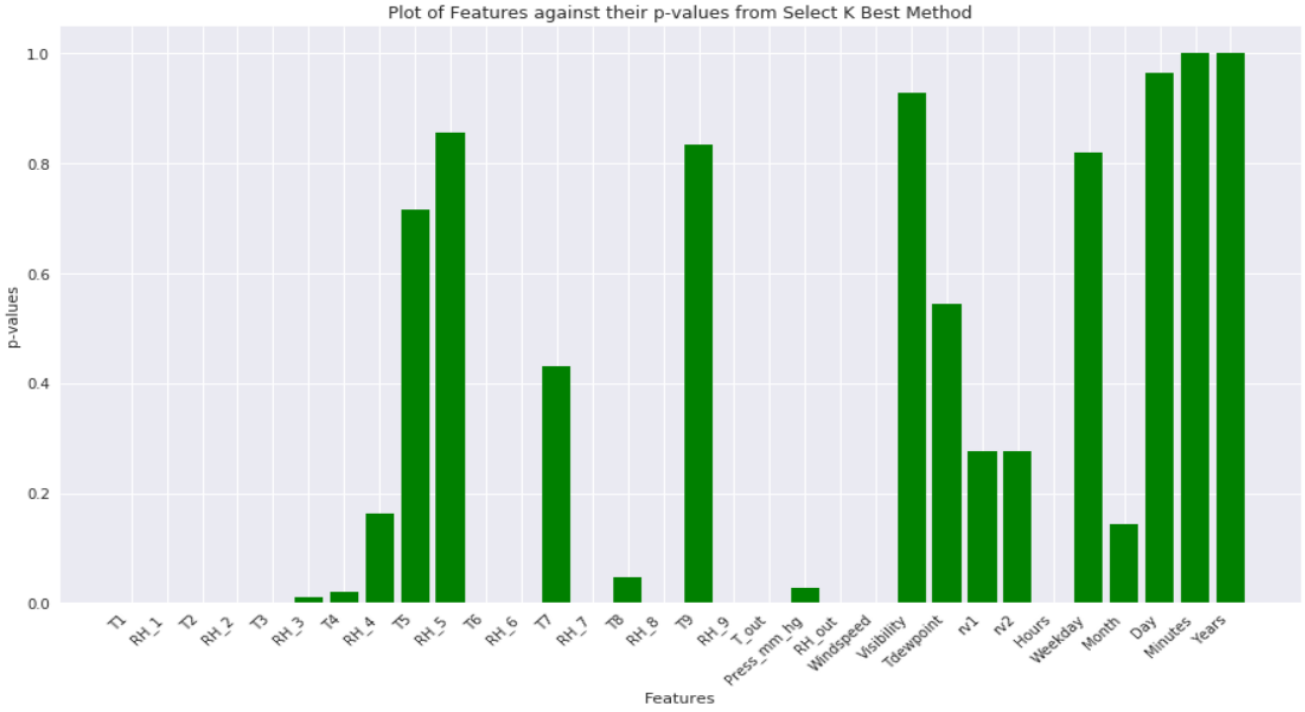


Figure 2: Mapping of p-values for Pearson's correlation.

It can be seen as the value from both datasets with all features and with selected features has been plotted against in the Linear Lasso Regression model and it does not show much of notable variation, hence estimating the fact that feature selection does not change the model prediction drastically.(Refer to **Appendix B**)

Model Development and Evaluation :

As the target variable does not bear a linear relation with the features as plotted in the scatter plots in the Jupyter Note and the Lasso Linear Regression Model does show a large variation estimating a large error in Linear Model Prediction(**Appendix B**) hence, we will be choosing a non-linear model prediction approach. To initiate with, the non-linear model like KNN, Gradient Boost, and Random Forest algorithms were chosen to build the Regressor Model. Alongside we also explored the ANN and LSTM network to study and train a model for prediction and analyse the performance. The dataset has been cross-validated with 80 % of training data and 20 % of validating data, the data

has been prepared and passed to each model in the GridSearchCV with cv=5. The GridSearchCV iterations ensure that no data has been left out for the training set/validating set. The models along with the set of parameters are passed to the Grid and the optimal model with the set of optimum parameters are fetched from the search as shown below:

Result for Random Forest Regressor:
RMSE score for RF 69.47818097524039
Train data score is 0.8974013868732335
Validation data score is 0.30421992987076707
The best parameter for the model is
{'criterion': 'mse', 'max_features': 'auto'}

Result for Gradient Boost Regressor:
RMSE score for GBM 67.75323303736272
Train data score is 0.5898627051421654
Validation data score is 0.33833958522306085
The best parameter for the model is
{'criterion': 'friedman_mse', 'learning_rate': 0.1, 'loss': 'ls', 'max_features': 'auto'}

Result for KNN Regressor:
RMSE score 62.086214624035925
Train data score is 1.0
Validation data score is 0.4443958649309778
The best parameter for the model is "{'algorithm': 'auto', 'n_neighbors': 6, 'p': 2, 'weights': 'distance'}

To begin with, the KNN model showed the highest overfitting of 1.0 as compared to other models and indicating this may not be a good choice for unseen data in terms of flexibility. While in the Random forest model, getting a score of nearly 0.89 on training data, while it scoring 0.30 on validation data shows that this model does tend to overfit hence it may not be an adaptable model if a new set of data are introduced in future. The Gradient boosting Model did show a lower train score of 0.58 and a higher validation score of 0.34 as compared to Random Forest indicating that the Gradient Boosting Regressor does demonstrate low overfitting in comparison to other constructed models as the difference between the training score and the test score(validation set) is not huge, which does imply that the Gradient Boosting Model can be more flexible than other models for prediction. Moreover, the learning rate in the Gradient Boosting helps in the regularization of the model such that it learns from the past error and updates the next choice for modelling which makes it a more robust than Random Forest.

Apart from the classical models, neural models for ANN and LSTM models to analyse model development. To begin with, the neural models were constructed based on the fact that models used here are likely to be befitting for non-linear timeseries information during the model development for sequential approach. The vanishing gradients issues due to backpropagation of neurons data has been attempted to overcome by Multilevel hierarchy and Long-term short memory for appropriate modelling strategy. After modelling we do see that the neural models tend to recognize more data points than the classical models as they show a better score if compared between the validation and the training scores(data provided below). The neural models were found to behave differently in tuning the epoch number, the number of neurons which was done in a permutation and combination approach, and the better performing model are being recorded in demonstrated in notebooks and reports. The introduction of Dropout in the neural models introduces the regularization in model development and thus it aids into lowers the overfitting and does avoid a complex structure learning of the neural network. This enables the model to be more fitting to predict unseen data in the future and restricting overfitting. The model does behave similarly in recognizing the data points of the validation dataset and train dataset as seen with the below figures from the model development. The difference in the R-squared score for validation and train dataset is nominal indicating that the performance of the model over the train and validation sets are equivalent.

Result for ANN model with a single layer

Validation data score
RMSE for the model 75.31099706541939
R2 score for the model 0.42719106824047914

Train data score
RMSE for the model 75.6701227229069
R2 score for the model 0.4201915695423384

Result for ANN model with multiple layers
Validation data score
RMSE for the model 70.20089993617786
R2 score for the model 0.5382095286060785

Train data score
RMSE for the model 68.55594104800367
R2 score for the model 0.5693110355819272

Result for Stacked LSTM approach
Validation data Scores
RMSE for the model: 70.1197937079971
R2 score for the model: 0.5397313234291728

Train data Scores
RMSE for the model: 67.52086230936057
R2 score for the model: 0.586830827305189
The above scores indicate that as how the neural network performed on the same analysis. The R2 scores are pretty good. The model selection depending upon the behaviour of the model on unpredicted data is more explained in the below Model Prediction Section.

Model Prediction :

The model performance was predicted over the 10% of data which has been set aside for prediction purposes from the beginning and we did find that the below behaviour between the predicted and actual values(*Figure* 3). For the Random Forest Regressor, there was a high bias that gradually decreases as we move from GBM to neural models. While selecting the model we did consider its tendency of the model to overfit which we find is gradually decreasing from Random Forest to Multilayered Neural Models, along with that we do find the RMSE, R2 scores for judgment. For this reason, we have dropped the idea of KNN as it showed the highest overfitting. The neural models with multiple layers in particular showed less overfitting and less bias as compared to classical models as demonstrated in Figure 3. The R2 scores for the neural do indicate that those models performed well in data coverage than classical models and can demonstrate more flexibility in future predictions by the classical model. The R2 score indicates that around 48.2 % of variations in data can be explained by the neural models, while it is 26.18 % in the case of the Gradient Boost Model and 20.29% for the Random Forest Model. Among the neural models, the double-layered neural model was performing well with metrics than the single-layered. But this does not prove that more layers should be added to increase efficiency, as it can enhance complexity and decreased performance. A clear study needs to be done with combinations to analyse the model.

Result for Gradient Boosting model with multiple layers for predicting unseen data:

RMSE score for the model:70.79430945971536
R2 score for the model: 0.2617923152899184
Result for Random Forest model with multiple layers for predicting unseen data:

RMSE score for the model: 73.56073301755752
R2 score for the model:0.20297130289024612

Result for ANN model with multiple layers for predicting unseen data:

RMSE for the model: 72.20360651150375
R2 score for the model: 0.481776798140798

Result for Stacked LSTM approach for predicting unseen data:

RMSE for the model:67.52086230936057
R2 score for the model: 0.48128842495609037

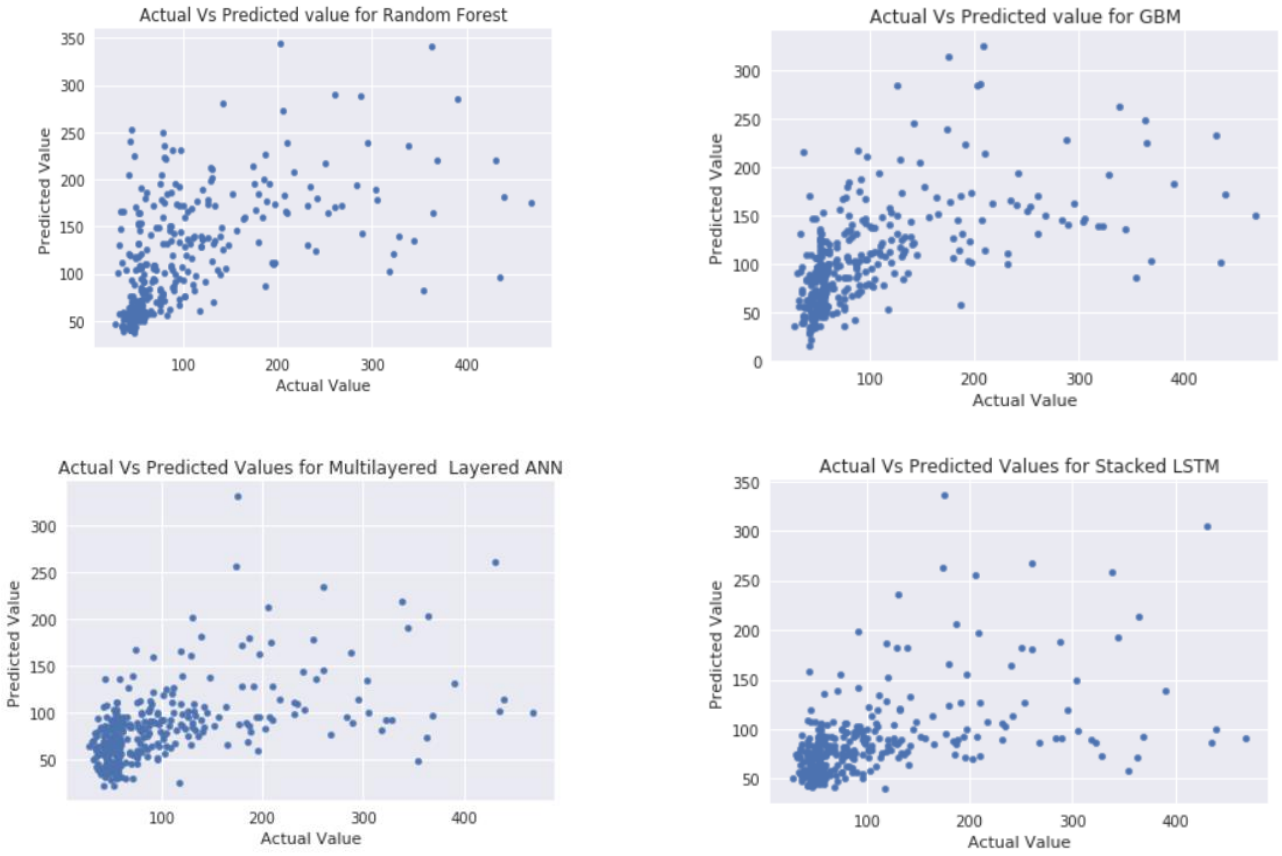


Figure 3: Prediction for various implemented models

Independent Evaluation:

Energy Consumption has been very used to understand the power consumption in fields like domestic buildings, commercial institutions, and buildings, ships and ports, etc. Some papers which were referred to during this study were very much helpful in understanding how to approach the problem.(Refer to number in brackets for details of paper)

1.*Data-driven prediction models of energy use of appliances in a low-energy house, Energy, and Buildings*.**[1]**

This has been the base for the study. Here the analysis has been done over the various classical models. The paper does demonstrate various models while in this study we have been taken the Gradient Boost and Random Forest Model similar to the paper, we also had opted for KNN Model for the non-linearity approach in solving the problem but it shows that the model has been highly overfitting in comparison to prediction. We also opted for Select K Best feature selection. Although the Gradient boost model did not demonstrate such a high accuracy of 97 % it did show an optimal response when both overfitting and prediction of train and validation sets were considered. While the Random Forest did show overfitting by giving a train score as 0.9 and test score 0.30 for our result. But in both reports, the GBM performed better than other models.

2. *Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings*.**[2]**

The paper does demonstrate the approach to the ANN model development method with the Lagrange interpolation to adjust the data for the time-series which is different from our study where we have taken the approach of resampling of the time series data. The paper had considered the multiple dense layers unless the least mean squared error has been found. In contrast, we have demonstrated single and double layers of ANN models and found out that introducing more hidden layers does increase the efficiency of the model but it does show a disadvantage that unnecessary introduction of layers increases the complexity of the model and thus decreases the performance. We did plot models on changing the factors like neurons, epochs, batch_sizes , and included the models which performed better than others in the report. Although the score has been different we did find a similarity that ANN behaved well in solving the problem statement than classical models.

3. *An End-to-End Adaptive Input Selection with Dynamic Weights for Forecasting Multivariate Time Series, Tsendsuren Munkhdalai, Tsatsral Amarbayasgalan, Lkhagvadorj Munkhdalai***[3]**

This paper demonstrated the ARIMA model for Multivariate Time Series Analysis. The paper used models like Decision Trees, LSTM, AIS-RNN with softmax layer, AdaBoost. The paper does say the implementation of PCA(Principal Component Analysis) with Recursive Feature Elimination for feature selection before model development while we took the SelectKBest approach for classical models. It explains how the RNN ability for retaining past information is held better than the MLPs and the basic functionality of Elman RNN. The LSTM and GRU have been analyzed for performance and the LSTM model did perform well in regards to other models. But the focus of the paper is quite difficult to judge in regards to consumption as there isn't any gradual order of model selection as in this report we have mentioned the linear regression approach and gradually moved towards the Neural Approach. The current report also demonstrates LSTM as a good model in consideration of fitting over data points and errors. But this paper does help us understand how to do the weights of the neurons to help in neural network learning.

Challenges:

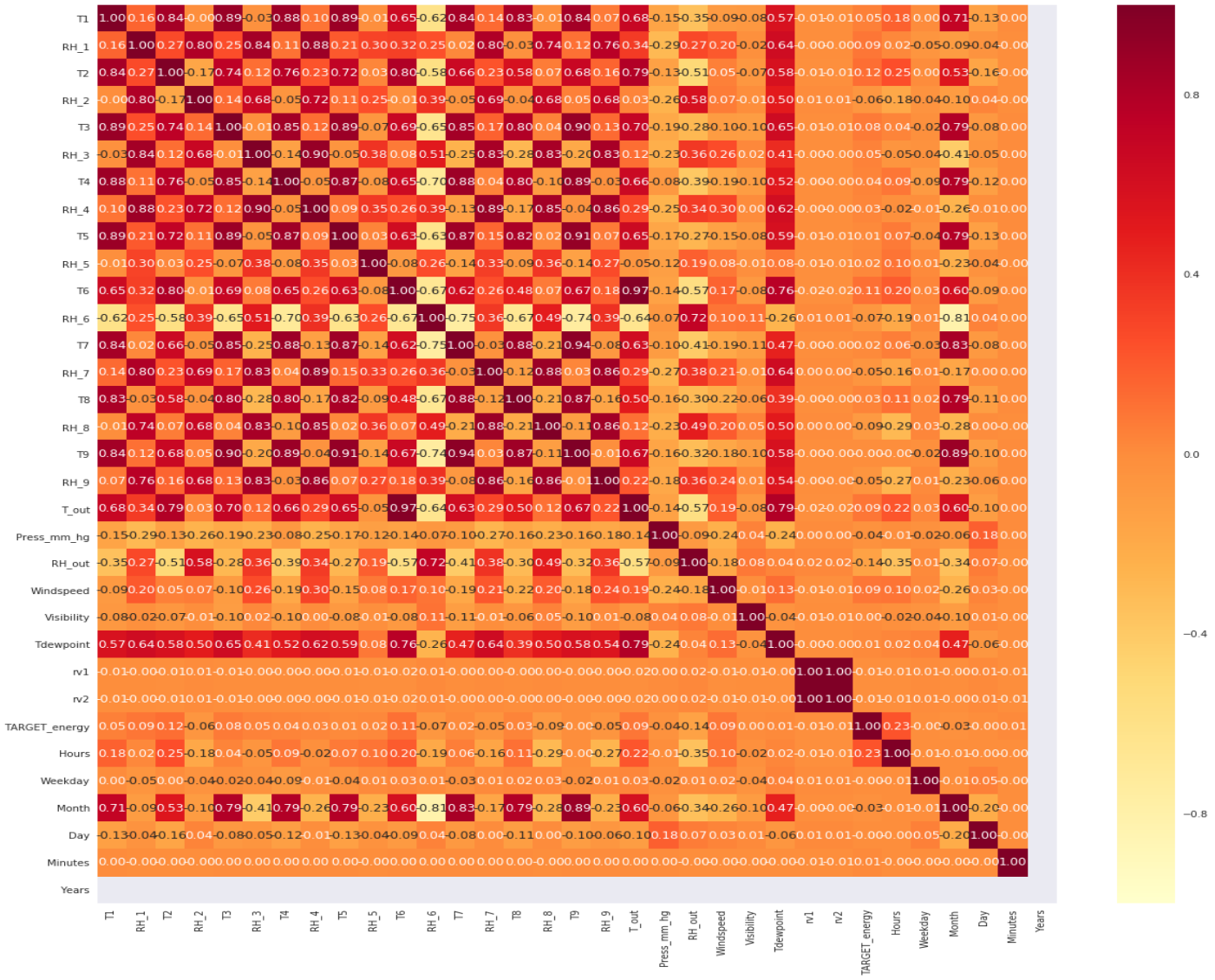
Below challenges were recognized on an overall basis to address the problem and compare it for related published papers:

1. During the process, I did face technical issues in implementation as the algorithms were taking much time to get implemented and running. The algorithms were getting stuck most of the time hence we have to interrupt the kernel and rerun the step.
2. Another challenge was to synchronize the basic steps of studies to understand various implementation functions of them.
3. Although the dataset is the same, the various papers had addressed the process of cross-validation, normalization in various ways do have included the more statically implemented functionalities. In context with the parameters taken, there was not a specific mentioning of selected model parameters specification which makes it quite difficult to compare the difference in the outcomes of this report and the papers.

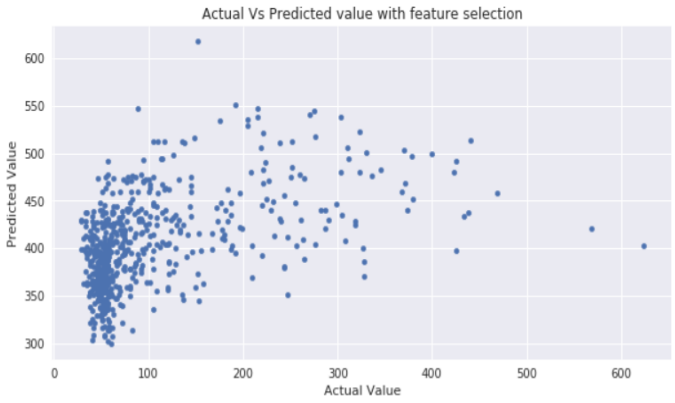
Conclusion:

Thus, the model which performed well was Neural Models since they did show an equivalent consistency in recognizing and predicting the data for the Test, Train, and Validation datasets and in support with the above-stated facts. The GBM Model did perform well in classical models since in Random Forest model showed a large training score and test score supporting the chances of overfitting while this is not seen in the neural model and GBM. But GBM did show a low variation understanding of data as compared to that of Neural Models. The independent evaluation shows how different approaches have been taken to solve the same or relatively similar problems and helped to provide the judgment.

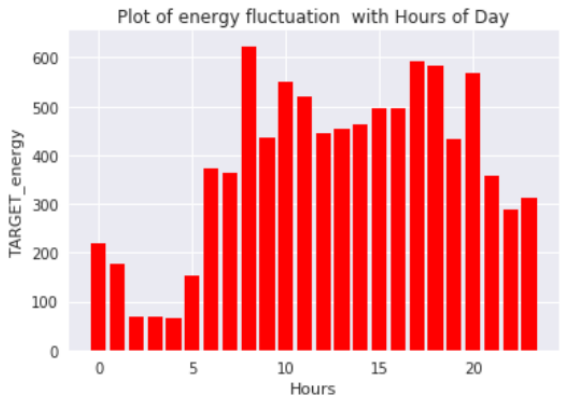
Appendix A:

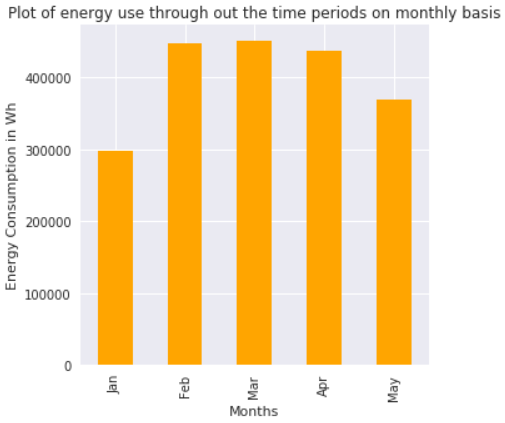


Appendix B:



Appendix C:





Independent Evaluation List:

[1] Luis M. Candanedo, Veronique Feldheim, Dominique Deramaix, Data-driven prediction models of energy use of appliances in a low-energy house, Energy and Buildings, Volume 140, 1 April 2017, Pages 81-97, ISSN 0378-7788

[2] Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. Shalika Walker, Waqas Khan, Katarina Katic, Wim Maassen, Wim Zeiler

[3] An End-to-End Adaptive Input Selection with Dynamic Weights for Forecasting Multivariate Time Series, Tsendsuren Munkhdalai, Tsatsral Amarbayasgalan, Lkhagvadorj Munkhdalai

References:

[1]https://towardsdatascience.com

[2] https://www.researchgate.net/

[3]An efficient data model for energy model and prediction using wireless sensors Michael Chammas,Abdallah Makhoul https://www.semanticscholar.org/