SUMMARY REPORT

PART – 1 SUMMARIZING RESEARCH PAPERS

1) Part 1 includes summarizing 3 research papers, the study is related to Appliances energy consumption and building models to efficiently predict the consumption of energy.
2) The First paper focuses on following things:
   a) Focuses on energy use of appliances in a low energy house.
   b) Features on Data Removal to remove non-predictive parameters and feature ranking.
   c) 4 models which are used to evaluate the results obtained.
3) The second paper focusses on following things:
   a) Prediction of appliances Energy use in smart homes.
   b) Building efficient predictive models.
4) The third paper focusses on following things:
   a) Contrasting the capabilities of single and ensemble prediction models using artificial intelligence techniques.
   b) Study of AI based prediction model which includes Data Collection -> Data Pre-processing -> Model Training -> Model Testing

PART -2 EXPLORATORY DATA ANALYSIS

1) This part includes exploratory data analysis of the energy data set ad finding relationships and patterns present between different features in it.
2) The problems associate with dataset were:
   1) We have a date feature with Object type, which won't be suitable for the analysis.
   2) So, we will change the data type for that feature to datetime.
   3) Also, rest all other features are numeric.
   4) Thus are dataset is numeric and don't have any categorical features.

3) It is found that the data is time series and we don’t have any categorical features present in the dataset.
4) We have given the Appliances energy utilization reading for the 10 min interval with the temperature, humidity, visibility and windspeed reading for that instance.
5) rv1 and rv2 shows perfect correlation. And as derived in above observations, rv2 is redundant feature.
6) Feature T9 shows high correlation with T3, T4, T5, T7.
Thus T9 can be consider as a overhead in the dataset.
7) Similarly, T5 has a high correlation with the T9, T3 and T1 which thus acts added redundancy.
8) Also, we know T6 and T_out are exterior temperatue features, and shows high degree of correlations.
9) We can get rid of either of the above two fetures to lower the data redundancy.
10) From the month January to February the appliance energy consumption increases and then decreases in the month of March.
11) Again the rise in the appliance energy increase in April and further drop in the month of May.
12) Similarly, there was a rise in the light usage for the month February and then gradual decrease in March, April and May.
13) Few of the observation which can he highlighted in this Data Analysis are:
   a) (Average) Appliance energy utilizations is low when lights are OFF.
   b) (Average) Appliance energy utilizations is high when lights are ON.

c) Also, Appliance energy utilizations follows a trend with respect to time.
d) For the night time (00:00:00 to 06:00:00) the energy consumption is lowest.
e) From the time 06:00:00 energy consumption started rising, and got steady till time 17:20:00
f) Then the peak hours started till 19:00:00
h) After that, the energy utilization started decreasing.

14) Feature rv1 does not have any specific pattern in the Appliances energy consumption over time.

PART 3 FEATURE ENGINEERING

1. We have appliance energy consumption time series dataset.
2. Also, the dataset is numeric, as we don't have any categorical features in it.
3. The data type for date feature needs to be datetime.
4. The date scope of the records is from 11th January 2016 to 27th May 2016 with 10 min interval in each record.
5. Total Number of records are 19735.
6. We do not have any missing vaules in energy dataset.
7. We have rv1 and rv2 features with perfect correlation and found to be same. Hence we can remove one of the feature from the dataset.
8. 90.29% of data for the appliance energy consumption is between 0-200 Wh.
9. 77.28% of data have lights off in the house, i.e. no energy consumption were recorded for the lights.
10. We have different temperature and humidity features. Also, features like windspeed, visibility and pressure has been recorded.
11. All temperature features are significantly correlated to each other.
12. All humidity featured has good correlation with each other except for RH_5, RH_6 and RH_out, due to the surrounding were it measured.
13. Feature T9 shows high correlation with T3, T4, T5, T7 and feature T5 has high correlation with T9, T3, T1. Hence can be consider as a redundant feature, giving us scope to eliminate those features.
14. Also, T6 and T_out are exterior temperatue features, and shows high degree of correlations. Thus we can get rid of one of these featurs.
15. To get more insight on the time series data for energy consumption we need to some more derived feature.
16. Features like month, time, DOY (Day of year), Only_Date and Date of week has been derived from date feature and added to the main energy dataset.
17. From the month January to February the appliance energy consumption increases and then decreases in the month of March.
18. Again the rise in the appliance energy increase in April and further drop in the month of May.
19. Similarly, there was a rise in the light usage for the month February and then gradual decrease in March, April and May.
20. (Average) Appliance energy utilizations is low when lights are OFF.
21. (Average) Appliance energy utilizations is high when lights are ON.

22. Also, Appliance energy utilizations follows a trand with respect to time.
23. For the night time (00:00:00 to 06:00:00) the energy consumption is lowest.
24. From the time 06:00:00 energy consumption started rising, and got steady till time 17:20:00
25. Then the peak hours started till 19:00:00
26. After that, the energy utilization started decreasing.
27. for normal hours 07:00:00 to 17:00:00 the energy utilization on weekend increases as compared to weekdays.
28. But for the peak hours ie. 17:00:00 to 19:00:00 the energy utilization decreases.
29. Feature rv1 does not have any specific pattern in the Appliances energy consumption over time.
30. Thus, giving us scope to think on eliminating the rv1.

PART 4 PREDICTION MODELS

1) In this section we have explored numerous machine learning models which can be used in our study to help getting best results in prediction.
2) Our study primarily included several models like:
    a) Ridge Regression
    b) Lasso Regression
    c) Elastic Net Regression
    d) Random Forests
    e) Gradient Boosting
    f) Extra Trees
3) Least performing Regressor - Lasso Regressor,ElasticNet
4) Best performing Regressor - Extra Trees Regressor
5) Even though Extra Trees Regressor had a R2 score of 1.0 on traininig set, which might suggest overfitting 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

¶

PART 5 FEATURE SELECTION

1) Feature selection deals with analyzing the feature and ranking them with different performance metrics and using them in the further model to get the best results.
2) We have explored several feature selection techniques like:
    a) Boruta Package
    b) Forward and Backward Selection
    c) Tsfresh
    d) Regressive feature elimination
    e) TPOT
3) Considering the results obtained in this 5 techniques some of them were not suitable for the data-set and hence we chose the best one to refine our model and get best prediction accuracy through it.

PART 6 MODEL VALIDATION AND SELECTION

1) We have used different model validation techniques to choose the best model for our dataset to predict values:

a) Cross Validation
b) Hyper Parameter Tuning
c) Bias Variance Trade-off
d) Regularization

PART 7 PIPELINING

We have automated the whole process from data ingestion to implementing best model and getting best results.