**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. Rutvina Papadkar   Email- [papadkarrutvina@gmail.com](mailto:papadkarrutvina@gmail.com)   * Data inspection * Exploratory data analysis   + - * Checking linear relation of all features using scatter plot       * Analyzing which features causing power consumption       * Checking which day in a week have high power consumption       * Checking distribution of features * Feature selection * Variance Threshold * F\_Regression * Heat map for finding a correlation between the features with the target column * Feature Engineering * Checking null values * Removing outliers * Model training cross validation and hyper parameter tuning * Model explain ability * conclusion with improvements points * Technical documentation * Fitting multiple models * Training & Testing * PowerPoint Presentation * Conclusion.  1. Sagar Khairkar   Email-[sagarkhairkar15@gmail.com](mailto:sagarkhairkar15@gmail.com)   * Data inspection---- * Head, tail, describe, null values, duplicates etc * Exploratory data analysis---- * Checking distribution of features * Checking outliers * Heatmap * Feature selection---- * F\_Regression * Feature Engineering----- * Standard scaler * Min max scaler * Log transformation * Sqrt transformation * Model training -----   1.Train test split  2.Used different algorithms   * Model training cross validation and hyper parameter tuning * Model explain ability * Conclusion.  1. Nayan Rewatkar   [Email-rewatkarnayan@gmail.com](mailto:Email-rewatkarnayan@gmail.com)   * Data inspection * Data Cleaning * Checking distribution of target feature by skewness * Checking skewness of features * Applying transformation for normal distribution * Fetching information from date column * Consumption visualization * Hourly, daily and monthly for target feature * Outliers handling * Checking Multicollinearity * 1Features Selection via Correlation and VIF method * Checking skewness of features * . Also used select best with F\_Regression * Preparation and Model making * . Used standard and minmax both scaler * . Used PCA but didn't give good results * . Try out 9 models and comparing their results with barplot * . Hyperparameter tuning for top 2 models * For Random Forest Regressor * For Lgbm Regressor * setting up best parameters after a lot of playing * Model Explainability * Using Shap * draw summary and force plot * making explanation for these above * plot to support our model * conclusion with improvements points * Technical documentation  |  |  | | --- | --- | |  |  | |
| **Please paste the GitHub Repo link.** |
| Github Link:- https://github.com/RutvinaPapadkar/Data\_appliances\_energy/upload/main |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| **Problem statement:** The data set is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. 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. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru) and merged together with the experimental data sets using the date and time column. Two random variables have been included in the data set for testing the regression models and to filter out non-predictive attributes (parameters). **Approach:**   * First, we load data set into Panda’s frame and initialize all the library which are required for doing EDA. * Then we did inspection of data on a basic level. * Then we did data cleaning by removing null values, duplicate values and outliers. * Then we used the matplotlib and seaborn to do Exploratory Data Analysis on sample data by plotting different graphs like count plot, pie chart, lmplot, bar plot, boxplot, subplot and heat map from this we got useful insights and correlation between target column and other features * In feature selection use variance threshold and F\_Regression to select best features * In feature engineering check null values, removed outliers in the data set. * use multiple models to predict power consumption did hyper parameters tuning Random Forest regression is best model. * With model explain ability technique for knowing which features is important   **Conclusion:**   * Many columns in the dataset have not normally distributed and target column is also right skewed * Dataset has many outliers and no null values * We have hours column is high correlation with dependent variable and there are lot features have lesser than 0.1 correlation with dependent variable and its non-linear dataset. * Energy consumption in month of march is high and low in Jan month and increase in temp leads to more energy consumption * Decrease in Humidity leads increase in power consumption. Humidity is inversely proportional to dependent variable. * Hour of the Day is the most important influencing parameter for Energy consumption * High Electricity consumption of >140Wh is observed during evening hours 16:00 to 20:00. Weekends (Saturdays and Sundays) are observed to have high consumption of electricity. (> 25% than Weekdays) * lights are having very low importance as a feature * Random forest is the best model its performance is good compare to others have high r2 on test * Through model explain ability found which feature has high importance. |