Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) Bindu Kovvada
   Email- bindukovvada187@gmail.com
   - Data Inspection
   - Exploratory Data Analysis:
     Which month the Energy Consumption is high?
     Which day in a week the Energy Consumption is high?
     Checking distribution of Dependent and Independent Features
   - Feature Selection:
     F Regression
     Heatmap to check correlation
     Checking multicollinearity using VIF
   - Feature Engineering:
     Outlier Analysis
     Removing Outliers using IQR method
   - Fitting multiple models
   - Training & Testing
   - PowerPoint Presentation

2) Manoj Patil M
   Email- smmanoj208@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.

3) Saksham Tripathi
   Email-saksham757474@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

4) Gulzar
Email-gulzarkhan9980@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 bar plot
- 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

5) Deepak Kumar Gautam
Email- deepakpracheta@gmail.com

- Data inspection
- Exploratory data analysis
  Checking linear relation of all features using scatter plot
  Analyzing which features causing power consumption
- Feature selection
  - Variance Threshold

- F_Regression
- Heat map for finding a correlation between all the features with the target column (Appliances)

- Feature Engineering
  - Checking null values
  - Removing outliers
- Model training cross validation and hyper parameter tuning
- Model explain ability
  Conclusion.

Please paste the GitHub Repo link.

GitHub Link: - https://github.com/bindukovvada/Appliances_Energy_Prediction

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:

Importing the Relevant Libraries
• Data Inspection
• Exploratory Data Analysis
• Feature selection
• Feature engineering
• Train and Test split
• Model training
• Cross validation
• Model Explainability
• Conclusion

Conclusion:

• Many columns in the dataset are not normally distributed, and the target column is also right skewed.
• The dataset has many outliers and no null values.
• We have a high correlation with the dependent variable in the hour's column, and many features have less

than a 0.1 correlation with the dependent variable in the nonlinear dataset.
• Energy consumption is high in March and low in January, and a rise in temperature results in higher
energy consumption.
• Decreased humidity leads to an increase in electricity consumption. Humidity is proportional to the
dependent variable.
• The most important determining factor for energy consumption is the hour of day.
• During the evening hours of 16:00 to 20:00, there is a high usage of electricity of more than 140Wh.
Electricity use is highest on weekends (Saturdays and Sundays). (more than 25% higher than on weekdays)
• As a feature, lights are extremely undervalued.
We excluded features that were not important for predicting the dependent variable using a variance
threshold, f regression, and the Pearson correlation matrix. We removed outliers from our model using
feature engineering.
• Implementing the XGBM and LGBM regression algorithms was done along with cross validation and
hyperparameter adjustment on all models. Decision tree, Random forest, Gradient Boosting, and Linear Regression were also used. In a comparison of all models, the Random Forest regressor is the best, having a high r2 score, a low MSE, and a low RMSE value. Due to the time series nature of the dataset and the lack of time series concept implementation, some overfitting is occurring
The model explainability Shap approach is used to determine which attributes are crucial for predicting output and understanding the model. The most significant feature is the hour feature.