Job-a-Thon Analytic Vidhya Nov'22

Problem Statement

Welcome to Green - A country well known for its greenery and natural resources. Green is working towards the betterment of the environment, natural resources, and health of citizens. Over the past few years, Green has improvised its natural resources by enabling the technologies for a safer future. The major investment happened to be in renewable energy. As of today, renewable energy contributes to 60% of the total energy capacity in the country. By 2030, the target is to generate 95% of the total energy through renewable energy. In order to achieve this objective, the government of Green Energy would like to use Data Science to understand the total energy demand of the country in the near future. This will help the government to build the infrastructure and technologies to achieve 95% of the total energy capacity via renewable energy. They have captured the estimated total energy demand from the past 12 years on an hourly basis. Now, the government of Green Energy is looking for a data scientist to understand the data and forecast the total energy demand for the next 3 years based on past trends. Help Green! Save Nature! Stay Healthy!

Objective

Task at hand is to build a machine learning/deep learning approach to forecast the total energy demand on an hourly basis for the next 3 years based on past trends.

About the Dataset

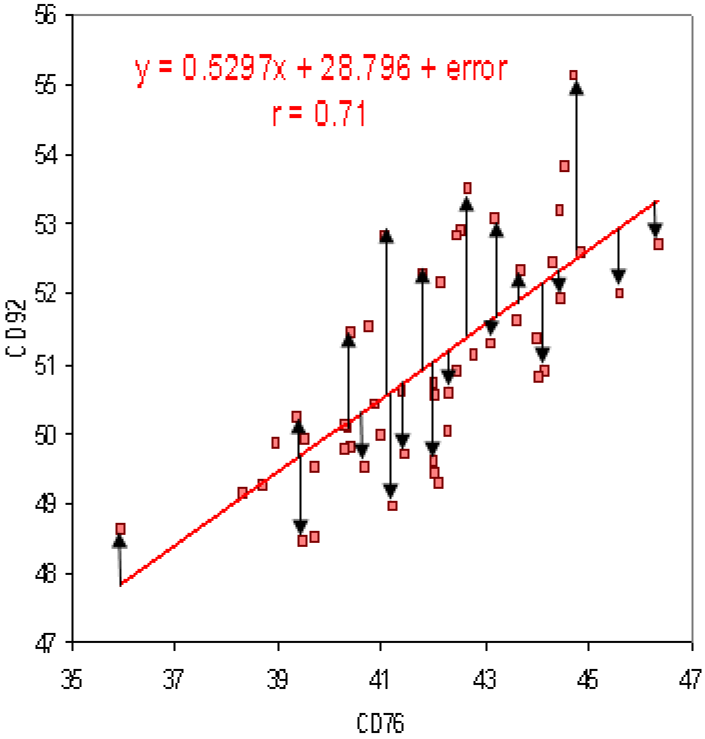
We are provided with total energy demand on an hourly basis for the past 9 years from March 2008 to Dec 2018 in the training set. You need to forecast the total energy demand on an hourly basis for the next 3 years from 2019 to 2021 in the test set.

Evaluation metric

The evaluation metric for this hackathon would be RMSE.

Root Mean Square Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.



Approach & Concept

Basic Operation Prior Anything

- Installing all the Basic Libraries
- Checking the Default Path
- Checking the Files in that Path

Read training and Test data using pandas

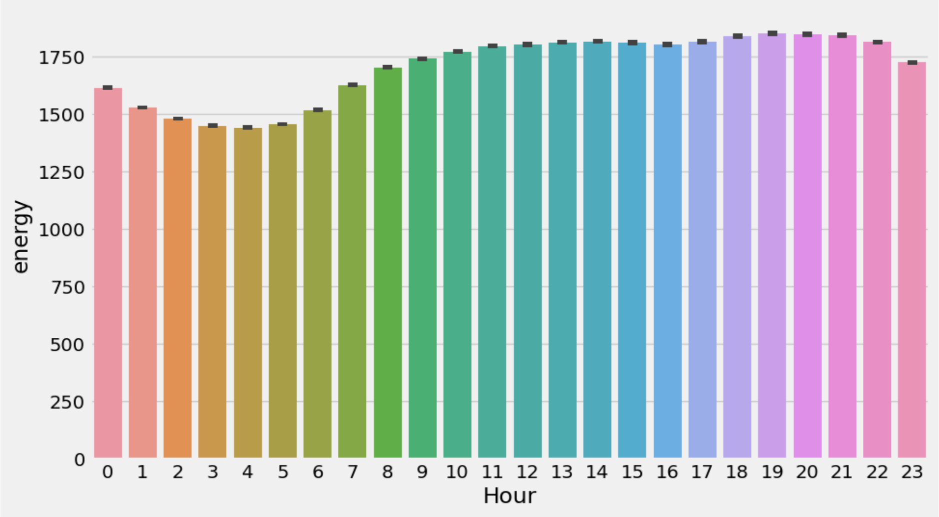
Data Cleaning and Wrangling

- Handling of all the Missing Values using Various operation -In this Dataset There where Missing Values in the Target Variable ['Energy']
- Handling of infinite, NaN , Null Values
- Handling of Missing Rows & Columns
- Using Dropna

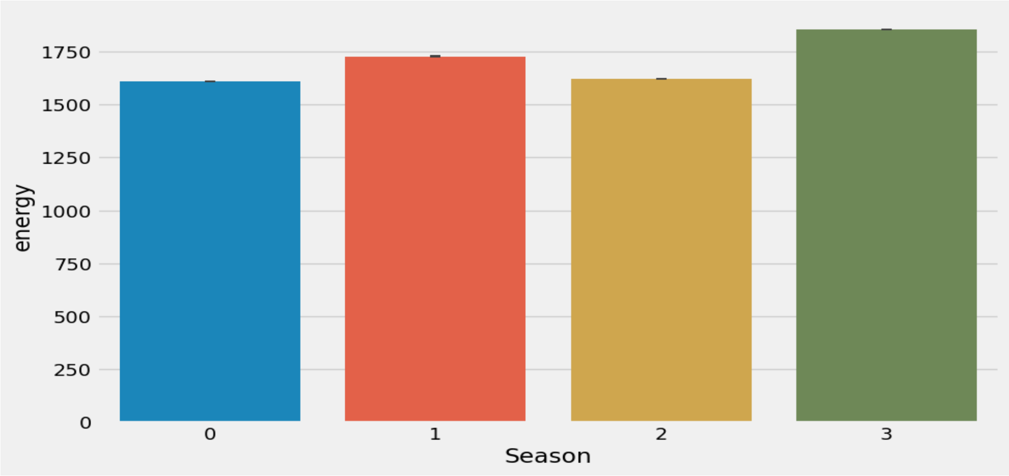
Note :- TimeStamp in this Dataset is a String Object not a datatime object so we need to Change the dtype to datetime[64ns]

Explanatory Data Analysis [EDA]

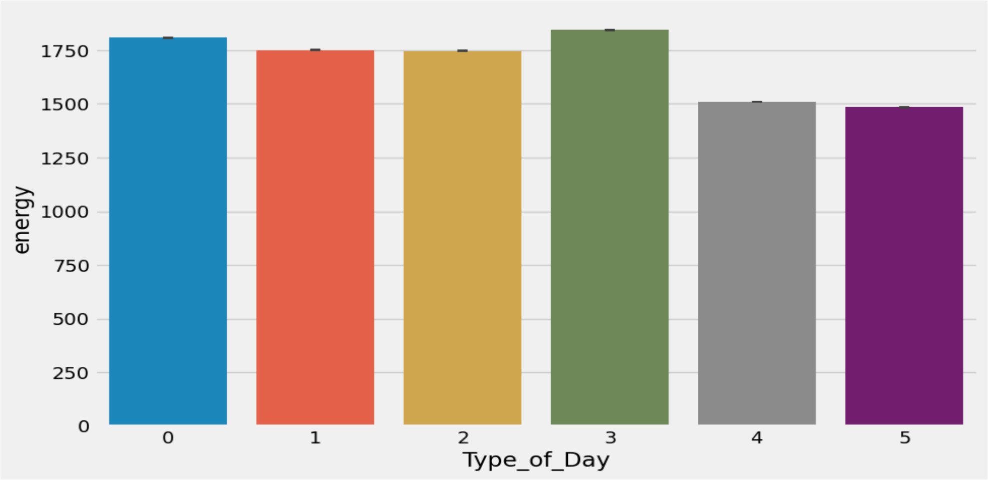
- **Deriving Data & Time Features from datetime column [Derived Features]**
  - Hour
  - year
  - month
  - day
  - day_of_week
  - is_quarter_date
  - is_weekend
  - Type_of_Day
  - Season
- **Data Visualization**
  - **Hour vs Energy**



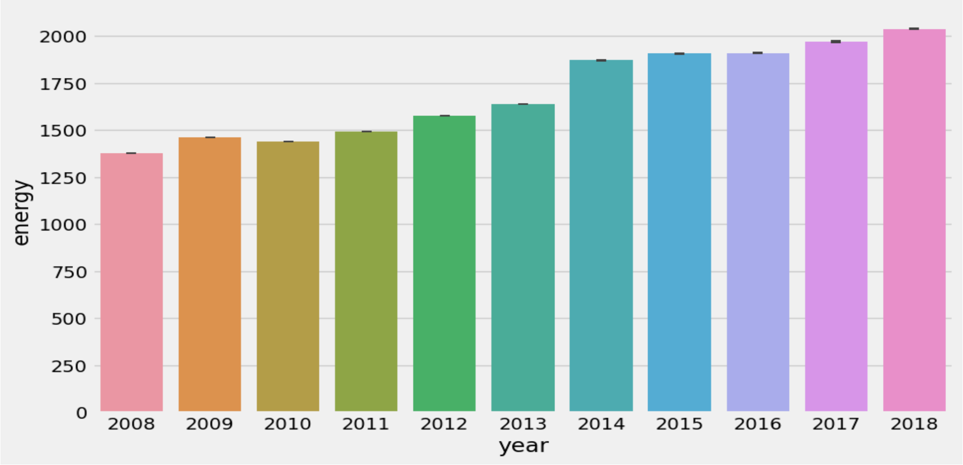
- **Season vs Energy**



- **Type_of_Day vs Energy**



- **Year vs Energy**



Variance Inflation Factor

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results.

- **According Removing Features where VIF is very high i.e year, is_quarter_date**
- **Standardization Using Z Score**
- **Feature Engineering**
- **Categorical encoding using Label-Encoding**

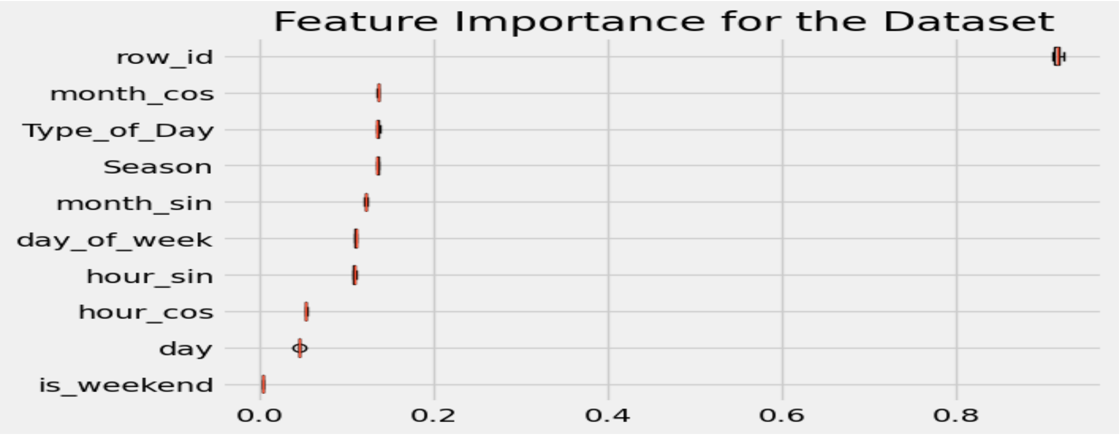
We Need to Round Function in the Pandas to decimal values to 2 so that Computional Power will be reduced while Model Training.

Few Important Points :--

. -> Keep remind of the Shape
-> Energy Demand for the past 9 years is on rise
-> Found to be their is some trend in Month, Weekday, Season and Others.

Model & Evaluation

- **Machine Learning Algorithms - Linear Regression(Ridge) - Neural Network - Random Forest - Catboost - XGBoost Regressor - XGBoost RF Regressor - Linear Regression(lasso) - Linear Regression - Light Gradient Boosting**
- **Train Test Split using Scikit-Learn**
- **Feature Importance**



- **Predicting Test.Values**

Best Model Result :--

Got the Best Result in XGBoost RF Regressor with a Score of 267 in Analytic Vidhya Solution Checker.

Submission

```
SS.to_csv('Submission_JOTNovXGBRF.csv', index=False)
```

Submitting the files in .csv Format