Electricity Consumption Prediction

AMS Assignment

Mohith Kurakula
SBU ID: 112504214
mohith.kurakula@stonybrook.edu

Data Preprocessing

Electricity and Weather Data is loaded by using **Pandas**. Convert the **time** format in both the datas to EST, Rename **Date & Time** in weather data to **time** and **merge** the two data sets on the column **time**.

Data Analysis

Daily Consumption, **Average Daily Consumption** and **Monthly Consumption** were calculated for Analyzing the Data. From **Average Daily Consumption** and **Monthly Consumption** we can find on which **days** and which **month** the Power consumption is high respectively and prevent the **high consumption** on those days and months in the future.

Features Generation

I Generated the following features using the **time** column. These features values are added to the data to help us get more accurate predictions: We use this feature set for **Linear Regression**, **Random Forest**, and **XGBoost** Regressor models.

1. **Weekend**
   During the weekdays the Power consumption is **uniform** whereas during the weekends the consumption **fluctuates** because on weekends people don't go to their jobs and they may spend their entire weekend in the home **increasing** the Power Consumption or may go on a trip **decreasing** the Consumption.

2. **Season**
   Power consumption will vary Season to Season. For example, If we take **Winter** Power Consumption is **high** due to **heating requirements**, If we take **Summer** Power Consumption is high due to the **high** usage of Air **Conditioners** and If we compare the above with the **Spring** then the Power Consumption is less as it does not require the above requirements.

3. **Sleep**
   As the Human Activity is less to none during the sleeping hours the **Power Consumption** is **low** and lesser than that of the day. During the day we use **light bulbs** and other appliances so Power Consumption is High.

Models

Naive Approach:

In the naive approach, the Predictions are equal to the **last observed value** without adjusting them. It is used mainly for comparing with the predictions generated by the better techniques.It is highly used in the **timeseries**, which have patterns that are difficult to **accurately predict**.

Linear Regression:

Linear Regression attempts to model by fitting the observed data to the **Linear equation**. It works fine with the **Linear** data, if the data is in **non-linear** shape then it doesn't work well as it only captures the **linear features** and **non-linear features** are lost.

From the data we can notice that there is a change in the Consumption when the **seasons** change. To analyze if there is a linear change or not we used a **linear model**.

Random Forest:

A **Random Forest** is a technique capable of performing both **Regression** and **Classification** tasks with the use of multiple decision trees and a technique called **Bootstrap Aggregation**, also known as **Bagging**. In **Bagging** we generate n samples and predict for each sample and average all the predictions(combining all the decisions and average them). Random forests work well if the data contains **Categorical features**. As the Generated Categorical features (Season, weekend, sleep) are added to the data, we use the **Random Forest Regressor** to Predict better results with less absolute error.

XGBoost:

In **XGBoost** Algorithm we use the concept of **Boosted Trees**.In XGBoost, the random sample that we take for training each minitree is picked with replacement over weighted data. Due to this, we can make each sample to be a **weak learner**. Based on their accuracy of classification, weights are assigned to each of these weak learners. Average of these **weak learners Prediction** is taken for the final predictions. In simple terms we can explain it as **many weak** can make one **strong**.

Results:

Home/model	Naive	Linear Regression	Random Forest	XGBoost
Home B	1.53	1.16437	1.16533	1.16436
Home C	1.54	1.11119	1.11050	1.11118
Home F	5.07	7.22297	7.22260	7.22292