**Synopsis**

**Introduction**

We are provided with load demand data of a northern Indian state, our object is to predict the load for a particular date. The dataset consists of data of three years. In which we have the date, time, load and other features which are relevant for the prediction. The data is a time series data, with an interval of 15 minutes. So, there will be ninety six entry for one day.

**Procedure & Insights**

From the dataset we can understand that it’s a supervised regression problem. First, we need to clean the data, during cleaning we have to check whether any data is missing, and if there is any do the needful. Here the load data for last two days are missing (which we want to predict). Also, it had a column with date and time values, from which I extracted the date, time, day, year. Hence those columns got added to the dataset. By using label encoding method, I have changed the data in time, day, month accordingly.

Using seaborn and matplotlib libraries I had plotted some graphs to identify the relation between some of the features. And based on that I had some inferences which I noted down in the notebook. I had used the correlation method to find the relation between the features. It is used when we want to remove some features from the dataset. If independent features are highly correlated with the dependent feature, then need not remove those kinds of features. If independent features are highly correlated among independent variable by 80% or 90% then drop those kinds of features and train the model with remaining feature. Here, even though correlation is checked none of the features are removed since the dataset is not that huge.

After cleaning the data, I split the data into train and test. Last two days in the dataset is not having the load data so for splitting I took the dataset without last two days values. Since, it’s a time series data I didn’t use the train-test split method. Instead, I split the data such that the last one month as test data and the remaining data as test data.

I used Linear Regression, Random Forest, XGBoost and Gradient Boost algorithms for training the model.

**Conclusion**

This problem was a good one to practice the Machine Learning. I had identified this problem as a supervised regression problem and based on that I did the modelling. I used Linear Regression, Random Forest, XGBoost and Gradient Boost algorithms. For each model I calculated the R-square value, and based on that value I selected the model to forecast load value for the desired date.

Here, XGBoost model gives the best R-square value when compared to other models so I used that model to predict the load value.