**Dataset Description**

The data was collected from the Delhi Electricity Board. In this research study, there are three significant datasets are collected such as electrical load data, weather data, and calendar data. The key reason for considering Delhi Electricity Board data is extremely fluctuating weather. The time-series data is transformed into suitable for the machine model building. Eight distinct electrical load/demand datasets are modeled using various characteristics such as hourly, daily, weather, and calendar data.

* Electrical load/demand data consists of two basic components: a timestamp in hours and a distribution of electrical load in Mega Watts.
* The maximum temperature in Celcius, the minimum temperature in Celcius, and the relative humidity at two different time points comprise the weather data.
* Calendar data is represented in the following manner using the label encoding technique.

Weekdays (Monday to Friday) are labeled by the number 0, whereas weekends (Saturday/Sunday) are labeled by the number 1, and public holidays are labeled by the number 2.

The downsampling technique is used to transform hourly data into daily data. The available datasets are grouped into two categories: training and testing. The training phase involves the use of data from the calendar years 2017 and 2018. The test phase includes the use of data from the calendar year 2019.

**The short-term electrical load forecasting is carried out using eight different data sets.**

1. Hourly Prediction with the electrical load data.

2. Day-wise prediction with the electrical load data.

3. Hourly Prediction with the electrical load data and weather data.

4. Day Wise prediction with the electrical load data and weather data.

5. Hourly Prediction with the electrical load data and the calendar dates inclusive of a holiday or not

6. Day Wise prediction the electrical load data and the calendar dates inclusive of a holiday or not

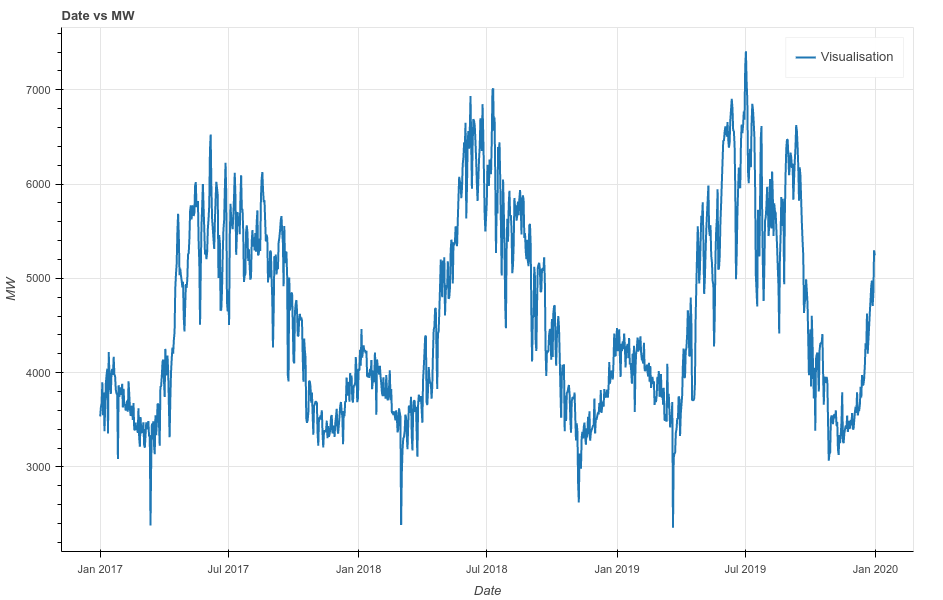
7. Hourly Prediction with the electrical load data, weather data, and the calendar dates inclusive of a holiday or not

8. Day Wise prediction with the electrical load data, weather data, and the calendar dates inclusive of a holiday or not

The primary level analysis is done with the time series analysis like AR, MR, ARMA, ARIMA, and machine learning models like linear regression, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN), and Random Forest. There are few ensemble models are considered for the prediction such as XG Boost, Gradient Boosting, and AdaBoost algorithms.

The panel data is analyzed using neural network models like RNN and LSTM.

**Exploratory Data Analysis**



It is observed that weather proves to be an important factor for election load consumption. As clearly visible in the above diagram, load consumption is more in summer as compared to winter. It reaches its peak value in July.

Another observation is for weekdays and weekends. The dataset shows that electric load consumption is maximum on weekdays as compared to weekends.

**Statistical Models**

The purpose of this research is to forecast future electrical load demand in Mega Watt based on previous electrical load usage that has been observed (and data collected) at regular time intervals. On the first three datasets, four univariate time series models are trained: Autoregressive(AR), Moving Average(AR), Auto-Regressive Integrated Moving Average(ARMA), and SARIMAX(Seasonal Auto-Regressive Integrated Moving Average with exogenous factors).

| **Name of the model** | **p** | **d** | **q** | **s** |
| --- | --- | --- | --- | --- |
| Dataset 1 | 5 | 1 | 1 | 12 |
| Dataset 2 | 1 | 0 | 3 | 12 |

**Notations:**

p: Number of autoregressive lags,

d: Order of differencing required for stationarity

q: Number of lagged forecast errors in the prediction equation

s: Measure of seasonality

Augmented Dickey-Fuller test (ADfuller()) is performed to know the stationarity of data. It returns a value of p. Lower value of p is considered good.

**Final Inference**

Considering the statistical univariate models, which consider only one parameter to predict the electrical load, so we are not getting desirable results. We are not able to consider all the parameters that are affecting electrical load consumption.

.

**Machine Learning**

The purpose of this research is to estimate the electrical load demand in Mega Watt. As a result, it is critical to employ numerical prediction models. On each of the eight datasets, four machine learning models are trained: Multiple Linear Regression, K-Nearest Neighborhood, Support Vector Machine (SVM), and Random Forest. The findings of the best machine learning models are summarized. Loss functions such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE) are calculated for each of the four models. The R-Square value quantifies the predictive power of a model in terms of percentage. R-Square is the proportion of dependent variable variation that is predicted by the independent variable.

| **Name of the model** | **Loss Function** | | | **R\_Square Value (%)** | **Dataset** |
| --- | --- | --- | --- | --- | --- |
| **MAE** | **MSE** | **RMSE** |
| Linear Regression | 665.83 | 694152.05 | 833.15 | 56.78 | 7 |
| KNN | 415.24 | 286639.813 | 535.38 | 76.23 | 8 |
| SVM | 429.49 | 303490.63 | 550.89 | 74.53 | 4 |

**Ensemble Models**

| **Name of the model** | **MSE** | **RMSE** | **explained\_variance\_score** | **r2\_score** | **Dataset** |
| --- | --- | --- | --- | --- | --- |
| XGBoost | 184579.72 | 429.63 | 88.86 | 84.50 | 8 |
| Gradient Tree Boosting | 183584.47 | 428.47 | 89.09 | 84.59 | 8 |
| AdaBoost | 202500.90 | 450.00 | 86.88 | 83.00 | 8 |
| Random Forest | 336.54 | 187942.76 | 433.52 | 84.22 | 8 |

**Conclusion**

The data itself is challenging

Data inadequate

Training Data 2017 & 2018 peaceful years

2019 is pandemic

**Ravi**

**Option1:**

1. LSTM with attention
2. GRU with attention
3. RNN with the attention

I start with RNN with attention..