POWER PLANT POWER GENERATION PREDICTION

Case Study:

We're provided with power and weather data for about 2 years. We're expected to predict power generated over the next 1month given weather data for that particular duration.

I handled the problem as a regression problem due to the following reasons:

- ➢ The target variable (power generated) is continuous.
- ➢ Weather data for the prediction month is provided, thus we have a test set (this rules out the idea pure time series forecasting).

Methodology:

- *Data understanding and overall exploration:*
  - Loaded the datasets (power_actual, weather_actuals, weather_forecast) into the working environment, checked their summary and dropped unnecessary columns.

- *Exploratory Data Analysis and Time Series Analysis:*
  Merged the power and actual weather datasets then performed the following analysis:
  - Univariate non graphical analysis - calculation of spread, measures of central tendency, and percentiles among other summaries for all the variables.
  - Univariate graphical analysis - plotted distribution plots (histograms and qqplots) for numerical variables and count plots for categorical variables.
  - Bivariate graphical analysis - plotted line plots and bar charts to explore relationships between variable pairs (more focus on how features related with **power**).
  - Multivariate graphical analysis - plotted correlation heatmap, pairplot and boxplots to explore relationships among more than 2 variables.

- *Preprocessing:*
  Concatenated the 2 weather datasets (actual == train features, forecast == test features) for uniform preprocessing. Performed the following preprocessing steps:
  - Null values: Dropped features with null values > 50%, imputed some with weekly means and others through regression imputation.

    - Outliers: Checked if outliers existed in the dataset, but did not treat them. This knowledge was used to select the best feature scaling approach.
    - Encoding categorical variables: used one hot encoder technique to remap categorical variables into numerical form.
    - Feature Scaling: Did a log transformation on the target variable, then used Robust Scaler on the predictor variables.
    - Dimensionality reduction: Applied Principle Component Analysis for dimensionality reduction and used components that explained upto 99% variation. This helped sort out problems of multicollinearity.

- *Modeling:*
    - Reconstructed the preprocessed data into train and test sets.
    - Used KFold cross-validation for model training and evaluation. Models used: **LinearRegression**, **XGBRegressor** and **LGBMRegeressor**. Their performance based on mean absolute error was as follows: **.64, .58. 56** respectively.
    - Methods considered for improving model performance:
        - GridsearchCV for hyperparameters tuning.
    - Methods and techniques to be considered to improve model performance:
        - Treatment of outliers.
        - Use of deep learning models like Neural Networks.
        - Models stacking and blending.
    - Averaged the mean of xgb and lgb predictions to get the predicted power generated.