1. INTRODUCTION
   1. **Problem Statement**

The objective of this Case is to predict the generation of power in the duration: October 1st, 2019 to October 27th, 2019

* 1. **Data**

Data provided with the problem is **weather\_actuals.csv**, **power\_actual.csv,**

**Test data : weather\_forecast.csv**

Overview of dataset: Independent variable should be a feature, “ power”, I have merged two datasets **weather\_actuals.csv**, **power\_actual.csv,**  by taking “datetime\_local” common column.

I have observed that most of the power generation is happening in between 6AM to 6PM , which makes sense, because sun will be visible only for that time. So, eliminating remaining timestamps.

Removing summary and icon features as, they explain same things as explained by numerical features like, cloud covering, dew point, humidity.

As I have to forecast the generated power on hourly basis, so, I have extracted hourly data of generated power from dataset

1. **PROJECT IMPLEMENTATION**

Below are the different phases of the implementation of project

* 1. **Rough sketch of project**

Here is the rough plan of the project in different phases, using CRISP-DM process. The whole project is divided in 7 phases (and further sub-phases). Below are the phases defined:

* Understanding the problem statement
* Gather the data
* Prepare data for consumption
* Perform Exploratory Data Analysis
* Modelling
* Evaluate and compare Model performances and choose the best model
* Produce sample output with selected model
  1. **Actual Implementation of project**

This is labelled dataset, having “power” as the independent feature,

So, clearly this problem is of category – **Supervised Machine Learning Regression Problem.**

There were some features, having missing values in them. I have removed those features. Below are such features.

["precip\_type","wind\_chill", "heat\_index", "qpf","snow", "pop","fctcode","precip\_accumulation”]

* + 1. **Perform EDA**
       1. **Outlier Analysis using Boxplots**

**There were some outliers, i have tried to replace and then impute them with mean, and for some of the rest features there were 5000+ outliers, So, I considered it as due to weather conditions and kept as it is, and then normalised the numerical features.**

* + - 1. **Analysis of Numerical Features**

1. **Correlation Analysis**

**Below features are correlated to each other.**

**['ozone',"wind\_speed","apparent\_temperature","precip\_intensity","precip\_probability","uv\_index"]**

* + - 1. **Choosing the Performance Measures for the Models**

We are working on regression problem, so I think the best performance matrix could be

* + - MSE

Going to measure above error metrics and compare them.

* + - 1. **Building the ML predictive Models**

Building linear regression by using sm.OLS()

And SVM by using SVR() with kernel rbf

Basic things while building the model, I have split the data between train and test then used the training set data features to train with training set target variable and then used that model to test on testing data set variable to produce results and then compared those results with target variable of test data set.

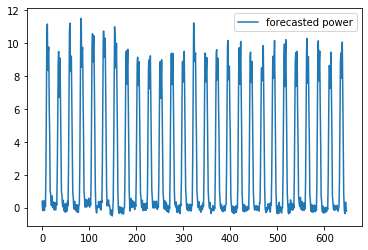
After that step i have performed cross validation so that 5 folds of the datasets should be formed and there will 5 different results of the model and average of those 5 datasets will be counted.

* + 1. **Model Evaluation and Comparision**

After taking mean scores after testing model on cross validation : I Have concluded that, SVM was best fit model

CONCLUSION

Forecasted generated power

****