1. Introduction

Electrical utilities need to diligently plan ahead of time to match their regional energy demand (MW), because if the demand is higher than the generation it can cause several blackouts resulting in a huge loss to the economy; on the other hand if the generation is higher than the demand the extra electricity will be wasted and it can also create an unnecessary load on the transmission lines.

So, it is very important for the energy companies to have a forecast of the energy consumption and the price category to be able to allocate appropriate resources to meet their demand. A daily forecast based on the weather reports, can help the utilities plan for a larger time scale but for smoother operations, daily forecast can prove very useful.

This project will involve analysing past 8 months of half hourly energy consumption data to find trends in energy consumption on a daily basis and also to check if factors like weather conditions in the region affect the energy consumption. That is, a model can be built to predict the energy consumption given parameters like temperature, rainfall, sunshine, wind direction, wind speed, etc.

➢ **What wrangling and aggregation methods have you applied? Why have you chosen these methods over other alternatives?**

Data wrangling also called data cleaning refers to a variety of processes designed to transform raw data into more readily used formats, ensuring data is in a reliable state before it’s analysed and leveraged.

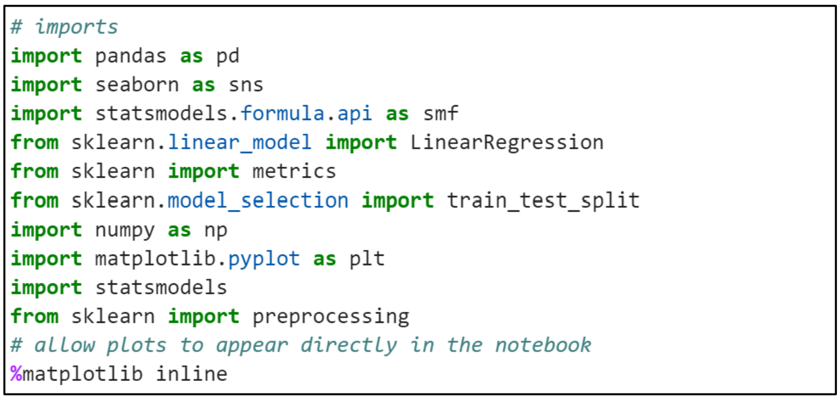
The main objectives of data wrangling in our assignment are the following:

- Merging multiple data sources into a single dataset for analysis
- Identifying blanks or missing data and either replacing or deleting them
- Deleting data that is either unnecessary or irrelevant to the project
- Identifying extreme outliers in data and either explaining the discrepancies or removing them so that analysis can take place

Following are the steps that we have used in our project as part of data wrangling and aggregation:

2. Importing libraries and input files

First, we imported various libraries those will be useful in our project. All these libraries have minimized our manual work thus making data modelling and analysis simple and fast.



```
# imports
import pandas as pd
import seaborn as sns
import statsmodels.formula.api as smf
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.model_selection import train_test_split
import numpy as np
import matplotlib.pyplot as plt
import statsmodels
from sklearn import preprocessing
# allow plots to appear directly in the notebook
%matplotlib inline
```

Figure 1: List of imported libraries

3. DATA

3.1. Importing Data

The next step was to **Input files** those contain raw data related to price demand and weather demand which we'll be cleaning, pre-processing and modelling as part of our data analysis

3.2. Evaluate usability of data

The second step of data wrangling is to gain a better understanding of the raw data. We have been provided with two raw files in this project in .csv format with information on price and weather demand. With commands like data.info(), price_data.head(), weather_data.tail() etc

In our case, we have two different data files, which consists of different sets of information on price and weather demand. Price demand file consists of half-hourly price demand while weather demand contains daily weather demand information. At this stage, dimension of our price_demand array is (11664 x 4) while that of weather demand is (243 x 11).

3.3. Cleansing Data

Once we understood the structure of our data, next step is to clean and format them. At this step, we:

1. Renamed Settlement date in price_demand file to Date
2. Converted half-hourly data to daily data using .groupby()

Further, we inferred the following:

- Grouping our data has reduced the price categories to predominantly two categories-Low/Mediums as against Low/Medium/High/Extreme.
- Plotting a graph of Daily energy consumption against date, showed that our energy consumption data is quite scattered
- Mean price demand data in low categoryis approx. 5000 while that in Medium category is approx. 6500.
- Removing Outlier – Since we had only one value for the month of September, this was considered as an outlier and eliminated it from the dataset using .drop() method.

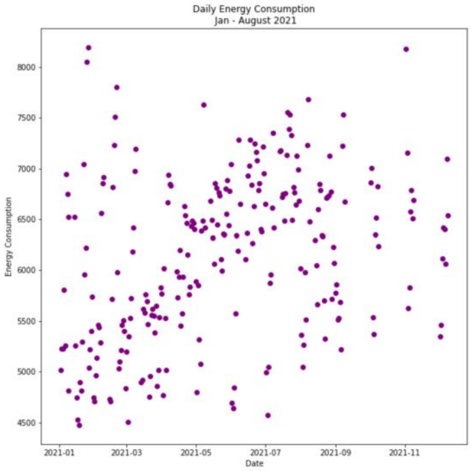


Figure 2: Plot of energy consumption over time versus Total Demand

3.4. Merging Data

In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified. In our case, taking Date as a common parameter and using **.join** method, we merged our two dataframes and now we have a single dataset. From the histogram below, we concluded majority of our total demand lies between 5200-7200.

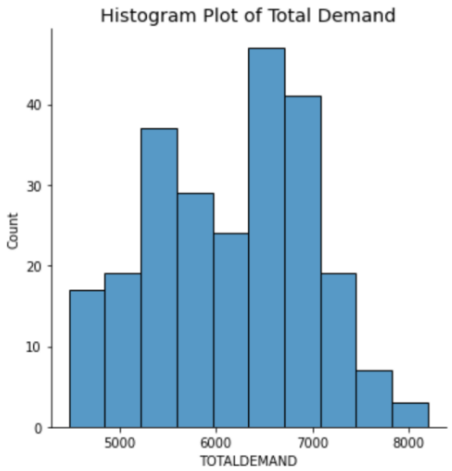


Figure 3: Histogram plot between count and total demand

4. Data Pre-processing

In order to increase the readability of our data frame and improve quality and consistency and making those appropriate to use in data modelling, following are the steps performed as a part of pre-processing:

- Removed Null values using .dropna() method or replacing with most frequent value
- Encoded categorical data in price_demand to numerical data using.LabelEncoder() method
- Found an outlier string called "Calm" in a column which should have integers and replaced with most frequent value.
- Converted wind direction data (datatype-string) to degrees (datatype-float) using windrose axis reference.
- Converted date from dataframe datatype to integer as well
- Rearranged the columns and renamed them to remove special characters
- And finally, we exported this final data to a new csv file.

- ➢ **How have you gone about building your models and how do your models work?**
- ➢ **How effective are your models? How have you evaluated this?**
- ➢ **What insights can you draw from your analysis?**

5. Exploring the data & Feature Selection

Feature selection is the procedure of selecting a subset of the input variables that are most relevant to the target variable. Target variable here refers to the variable that we wish to predict. In this project, it is Total Demand for problem 1 and Price Category for Problem 2.

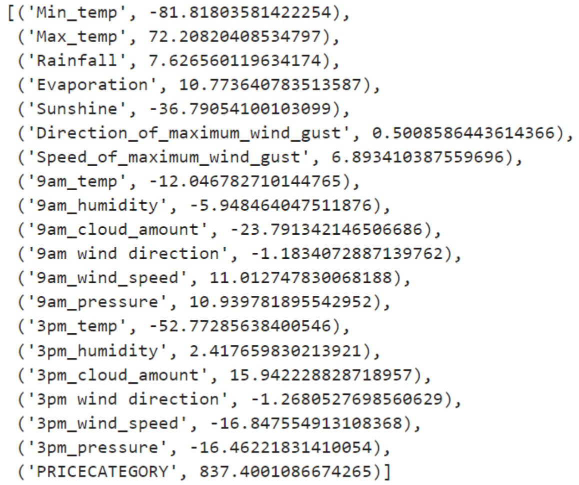
For this report we assumed that we only have numerical input variables and a numerical target for regression predictive modeling. The 2 feature selection techniques that have been used for Problem 1 are the following:

- Pearson's Correlation
- Mutual Information (MI)

5.1. Feature selection using Pearson's Correlation

We tried to understand the correlation between our Target Feature (Total Demand) with the rest of the features. This has been done using sns.pairplot().

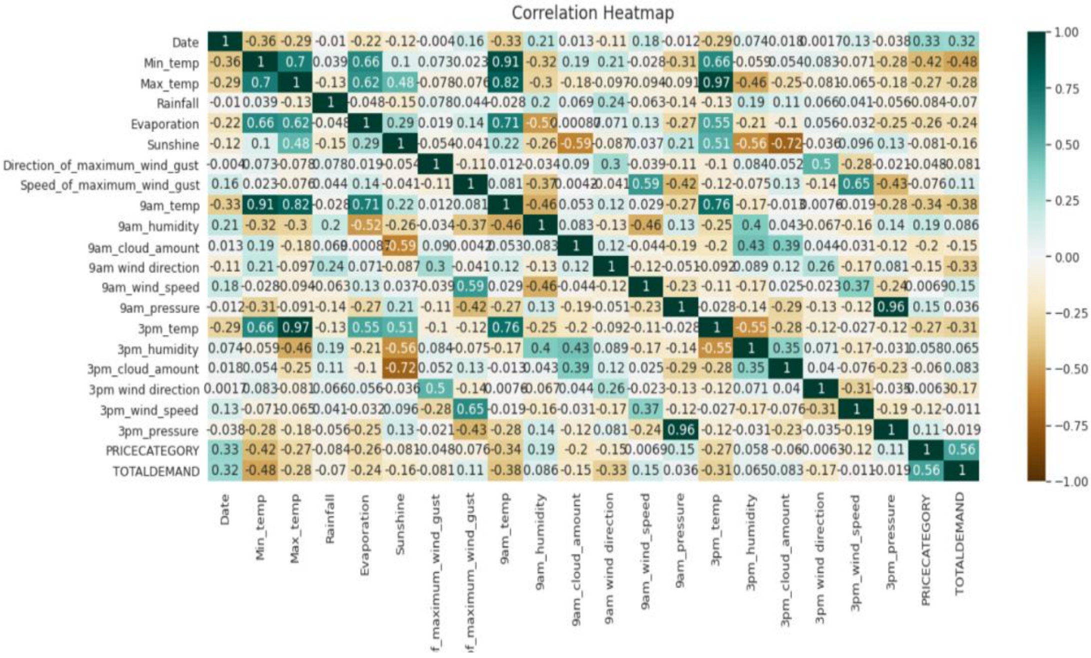
Further we visualize the relationship between the features and the response using scatterplots and estimated the model coefficients for all features. This helped us understand which features had linear correlation.



```
[('Min_temp', -81.81803581422254),
 ('Max_temp', 72.20820408534797),
 ('Rainfall', 7.626560119634174),
 ('Evaporation', 10.773640783513587),
 ('Sunshine', -36.79054100103099),
 ('Direction_of_maximum_wind_gust', 0.5008586443614366),
 ('Speed_of_maximum_wind_gust', 6.893410387559696),
 ('9am_temp', -12.046782710144765),
 ('9am_humidity', -5.948464047511876),
 ('9am_cloud_amount', -23.791342146506686),
 ('9am wind direction', -1.1834072887139762),
 ('9am_wind_speed', 11.012747830068188),
 ('9am_pressure', 10.939781895542952),
 ('3pm_temp', -52.77285638400546),
 ('3pm_humidity', 2.417659830213921),
 ('3pm_cloud_amount', 15.942228828718957),
 ('3pm wind direction', -1.2680527698560629),
 ('3pm_wind_speed', -16.847554913108368),
 ('3pm_pressure', -16.46221831410054),
 ('PRICECATEGORY', 837.4001086674265)]
```

Figure 4: Coefficient values for each feature

The Pearson's coefficient was also plotted using a Heatmap and this was then used for determining the feature of importance.



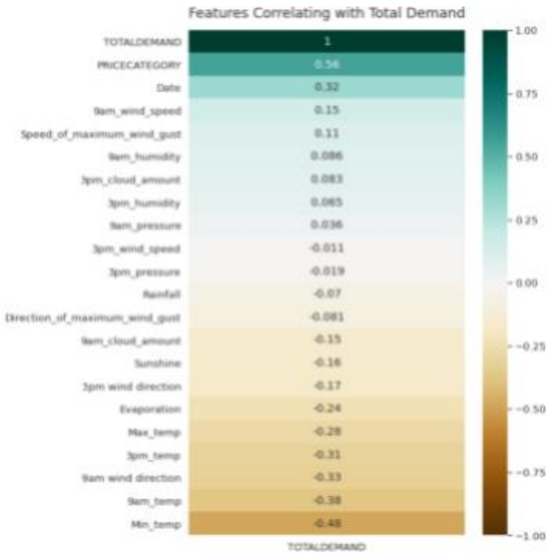


Figure 5: Heatmap / Correlation coefficient of one column: Total Demand

5.2. Feature selection using Mutual Information Metric:

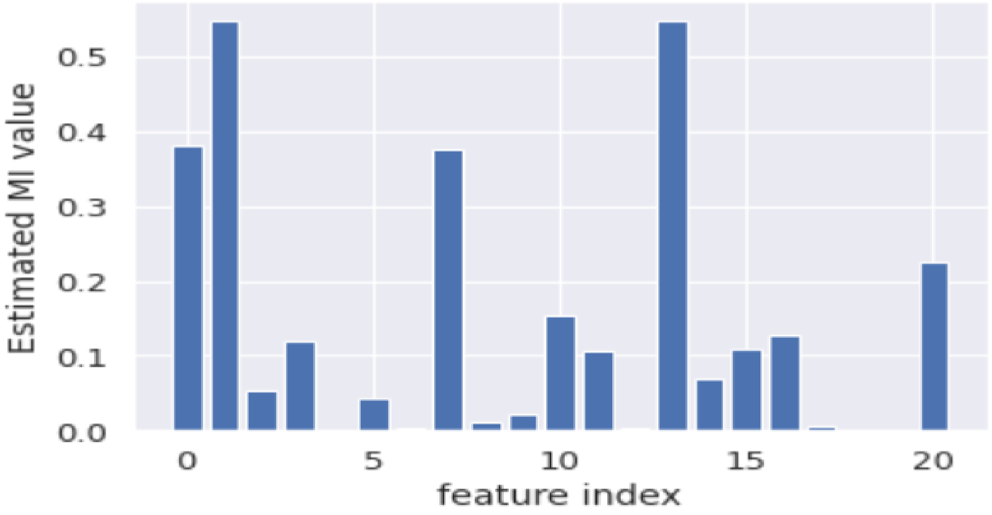


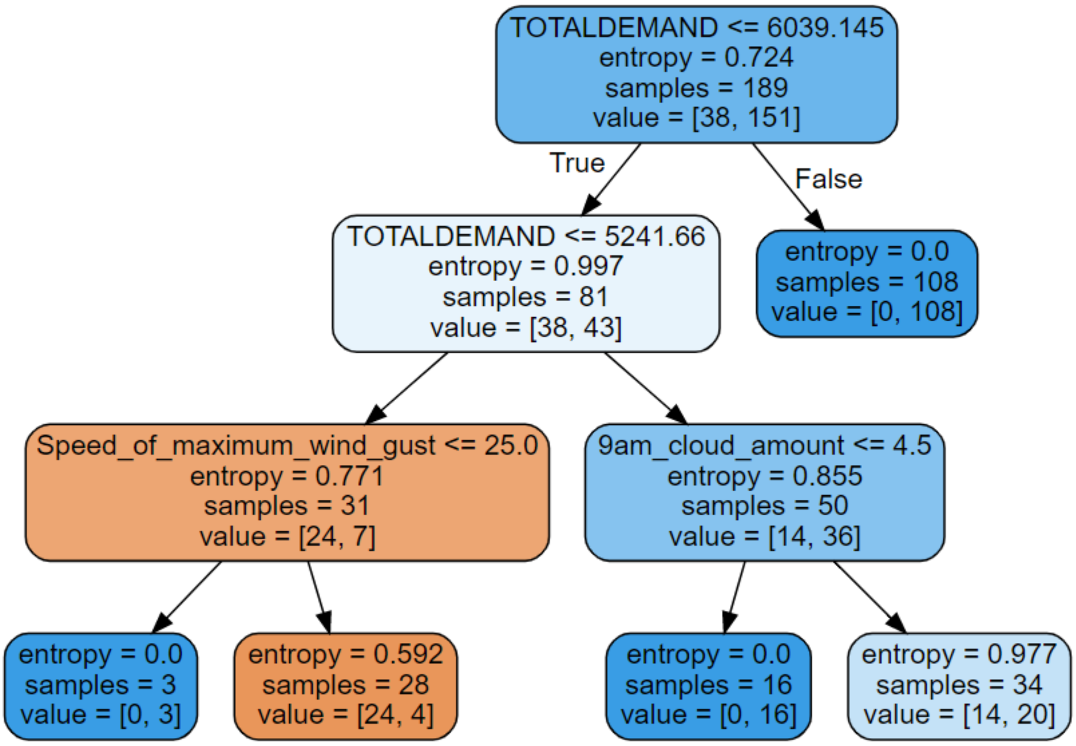
Figure 6: Plot of Estimated MI versus features

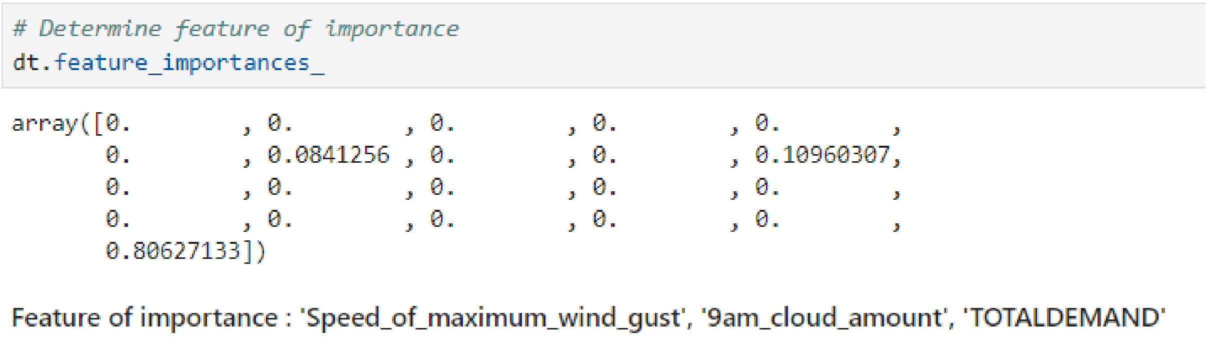
- We considered MI values above 0.2 as acceptable.
- The plot above shows that feature 1,2,8,14,20 are more important than the other features as these have MI above 0.2

5.3. Problem 2 : Feature selection using Decision Tree Classifier

For problem 2, DT classifier method for feature selection was used for predicting categorical feature

[143]:





```
# Determine feature of importance
dt.feature_importances_
```

```
array([0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.0841256 , 0.        , 0.        , 0.10960307,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.80627133])
```

Feature of importance : 'Speed_of_maximum_wind_gust', '9am_cloud_amount', 'TOTALDEMAND'

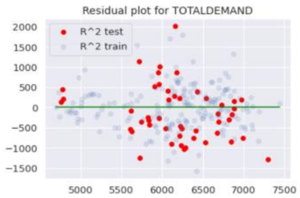
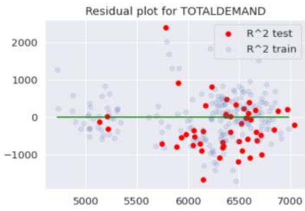
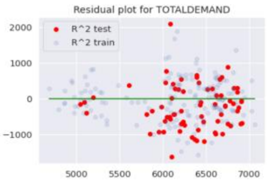
6. PREDICTIVE MODEL DEVELOPMENT

6.1. Model Development for Problem 1:

For problem 1, using 2 methods for feature selection, we created 3 regression models:

- MODEL 1 - Using all the features
- MODEL 2 - Using features from Mutual Information Metric
- MODEL 3 - Using feature selection from Pearson’s correlation

The table below summarizes the R2 score for each of these models. We have used residual plots to visually confirm the validity of our model.

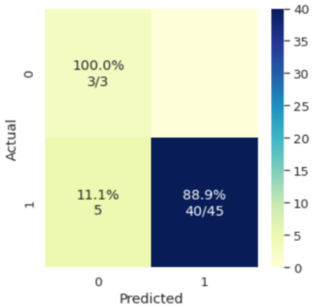
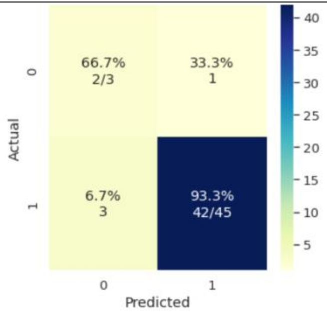
	Model 1	Model 2	Model 3
Feature Selection	All Features	'Min_temp', 'Max_temp', '9am_temp', '3pm_temp', 'PRICECATEGORY'	'PRICECATEGORY', '9am wind direction', 'Evaporation', 'Min_temp', 'Max_temp', '9am_temp', '3pm_temp', '9am_wind_speed'
R2_train	0.60	0.49	0.54
R2_test	0.17	0.10	0.24
Residual Plots			

6.2. Model Development for Problem 2

Problem 2, using the DT classifier method for feature selection, we have created 2 logistic models here:

- MODEL 1 - Using all the features
- MODEL 2 - Using feature selection from DT classifier

The table below summarizes the R2 score for each of these models. We have used confusion matrix to visually confirm the validity of our model.

	Model 1	Model 2
Feature Selection	All Features	'Speed_of_maximum_wind_gust', '9am_cloud_amount', 'TOTALDEMAND'
R2_train	0.895	0.92
R2_test	0.90	0.93
Confusion Matrix		

7. Inference and conclusion

- ➢ **Why are your results significant and valuable?**
- ➢ **What are the limitations of your results and how can the project be improved for future?**

7.1. Significant results:

Problem 1;

Three regression models were built to predict the total demand of electricity.

Model	**Feature of Importance**	**Accuracy rate of Training data**	**Accuracy rate of Testing data**
MODEL 1	All features	59%	16%
MODEL 2	Min_temp, Max_temp, 9am_temp, 3pm_temp, PRICECATEGORY	49/%	10%
MODEL 3	PRICECATEGORY, 9am wind direction, Evaporation, Min_temp, Max_temp, 9am_temp, 3pm_temp, 9am_wind_speed	54%	24%

The table above shows the results we obtained from the models. The accuracy rate of the training data ranges from 49% to 59%, while the accuracy rate of testing data ranges from 10% to 24%. Meanwhile, the residual for model 3 is the closet value to 0 as compared to the other two models. We also identify that there is no bias or trend in the residuals. Also there is no pattern or trend in the residuals. Thus, we conclude model 3 fits the dataset and thus can provide decently accurate prediction for future demand of electricity.

Problem 2:

In order to predict the maximum daily price category, we built a logistic regression model

MODEL 1: Logistic Regression using the feature of importance
MODEL 2: Logistic Regression using all features

Model	Feature of Importance	Accuracy rate of Training data	Accuracy rate of Testing data
MODEL 1	Speed_of_maximum_wind_gust, 9am_cloud_amount, TOTALDEMAND	92%	93%
MODEL 2	All features	89.5%	90%

The table above reflects the results we obtained from the models. The accuracy rate of MODEL 2 is 92%, while the MODEL 1 returns the score of 85%.We found when all features were employed, the accuracy score for the model dropped. To sum up, feature selection using Decision Tree classifier could be a better approach for this model.

7.2. **Limitations:**

The energy consumption is highly dependent on the outside temperature and has strong multiple seasonality — daily, weekly, and yearly , specific regional data. Although the linear regression models lead to a reliable prediction, we found the accuracy results in our models hasn't reached the maximum value.

The first limitation for the current project is lack of enough data. Currently, we only have a dataset in 8 months, and this length of dataset will probably limit our understanding of the reality. The best way to capture the trend, which is a combination of all the above factors and maybe more, is to make the model learn the trend over a long period of time. The seasonality is an important part of predicting the energy consumption of a region, so getting that part right is also very crucial for improving the model's performance. Additionally, more data type should be considered when building models. For example, the electricity was transported by retailers or generated by separated solar panels may conclude in different relationship between weather conditions.

When we grouped our data, we ended up with only 2 price category – Low and medium. We believe more data points will be supportive to classify more categories. Also a better understanding of the price category will be required.

Furthermore, instead of generic data about region, specific geographical and weather information of sites will be helpful to increase the accuracy rate of the model.

The results of our regression models also reflect the relationship between weather records and the electricity data may not be a simple linear relation. Other regression models such as XGBOOST are recommended for better understanding.