EXPLORATORY DATA ANALYSIS

Aditya Vinayak Sawant
Ashay Arun Singh
Manish Satish Aradwad
V.J.T.I. College

Contents

0.1 Introduction

0.1.1 Overview:

The influence of temperature and monthly or seasonwise trends on the annual electricity consumption in IIT-Bombay has been investigated in order to develop a simple and data light electricity consumption forecasting model, to be used as part of more complex planning tools. The time period considered for the historical data is from January 2017 to December 2017. Multivariable Linear regression models are developed using historical electricity consumption, day, month, and weekday. Annual electricity consumption was strongly related to the selected variables.

0.1.2 Objective and Scope of the study:

1. To study household electricity consumption so as to identify opportunities to optimize the consumption of electrical energy.

2. To predict electricity consumption using the available data for subsequent time periods.

Scope of study is focused on electricity consumption from a residential building in IIT-B campus.

0.1.3 Approach:

The dataset for electricity consumption from campus building is pre-processed to convert it into suitable format. This converted dataset is used for Analysis and Prediction.

0.1.4 Importance:

Electric power is one of the major input factors in economic development. To support economic growth and meet power requirements continually in the future, electricity consumption forecasting has become a very important task for electric utilities. Electric consumption will continue to increase in the power system. Electricity forecasting has become one of the most important aspects of electricity utility planning. Moreover, an accurate forecasting is helpful in developing a power supply strategy, financing planning,

marketing research and, of importance today, planning to use alternative energy or renewable energy of a country in the near future. This study focused on the prediction of electricity consumption for a residential building in IIT-Bombay.
Multivariable Linear Regression method was used to train and test the developed model.Climate data and other factors like weekday, month and day recorded for the year 2017 was used as input for model development and verification.Predictions were done for target year 2018.

0.2 Software Used

1. *Anaconda Module:* It comes with all the modules required for data analysis using Python. Along with that it installs latest version of Python.

2. *Jupyter Notebook:* This is preferred over IDEs because Jupyter allows us to run parts of code separately in different cells which saves a lot of time.

3. *Pandas Module:* It contains all the functions required for data separation and formatting which is used while Pre-processing.

4. *MatplotLib Module:* It contains all the functions required for data visualisation using Line Graphs, Bar Graphs, Scatter plots, etc.

5. *Scikit Learn Module:* It has various machine learning models which are used for prediction of data.

6. *NumPy Module:* It performs Mathematical and Logical Operations on multidimensional arrays, which are used during Analysis and Prediction.

0.3 Block Diagram and Flowcharts

Figure 1: Flowchart for Approach

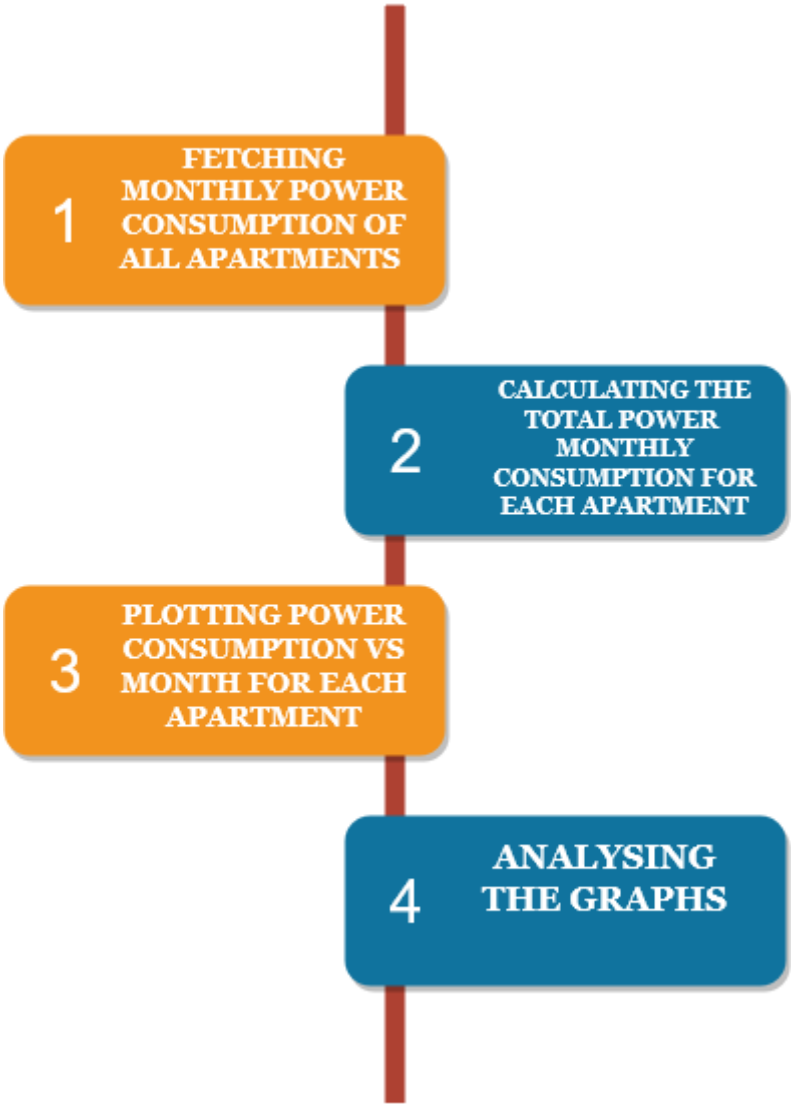

Figure 2: Monthly Power Consumption Analysis

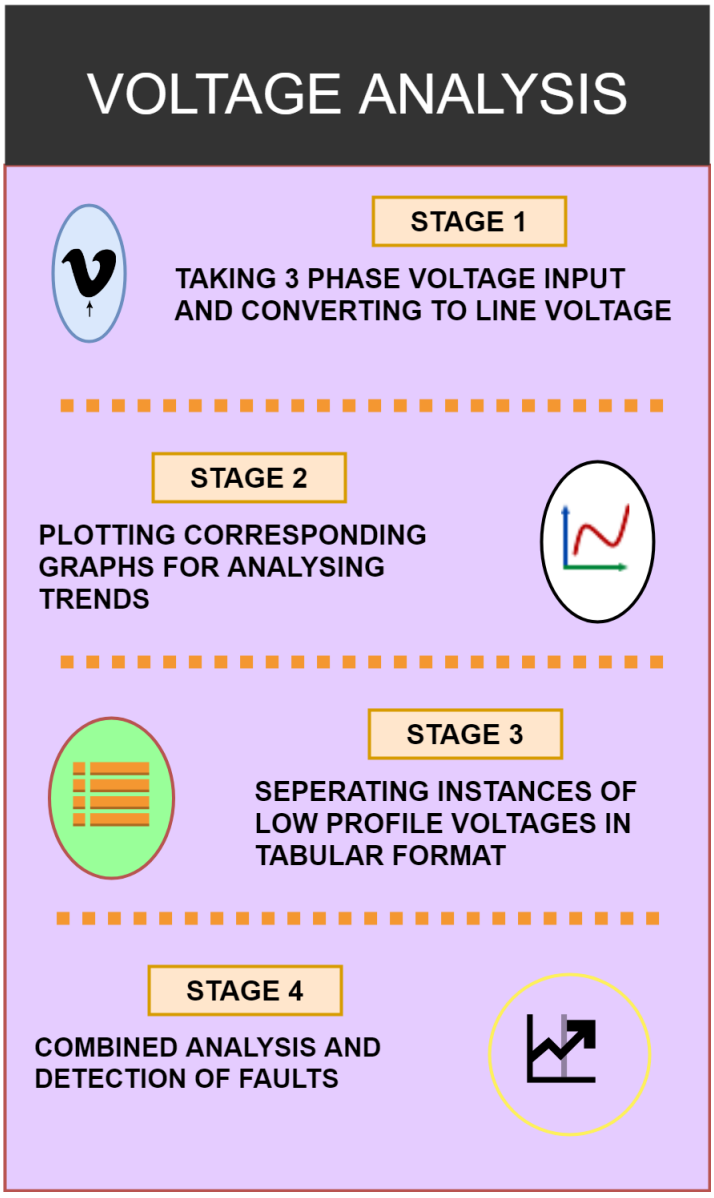

Figure 3: Voltage Profile Analysis

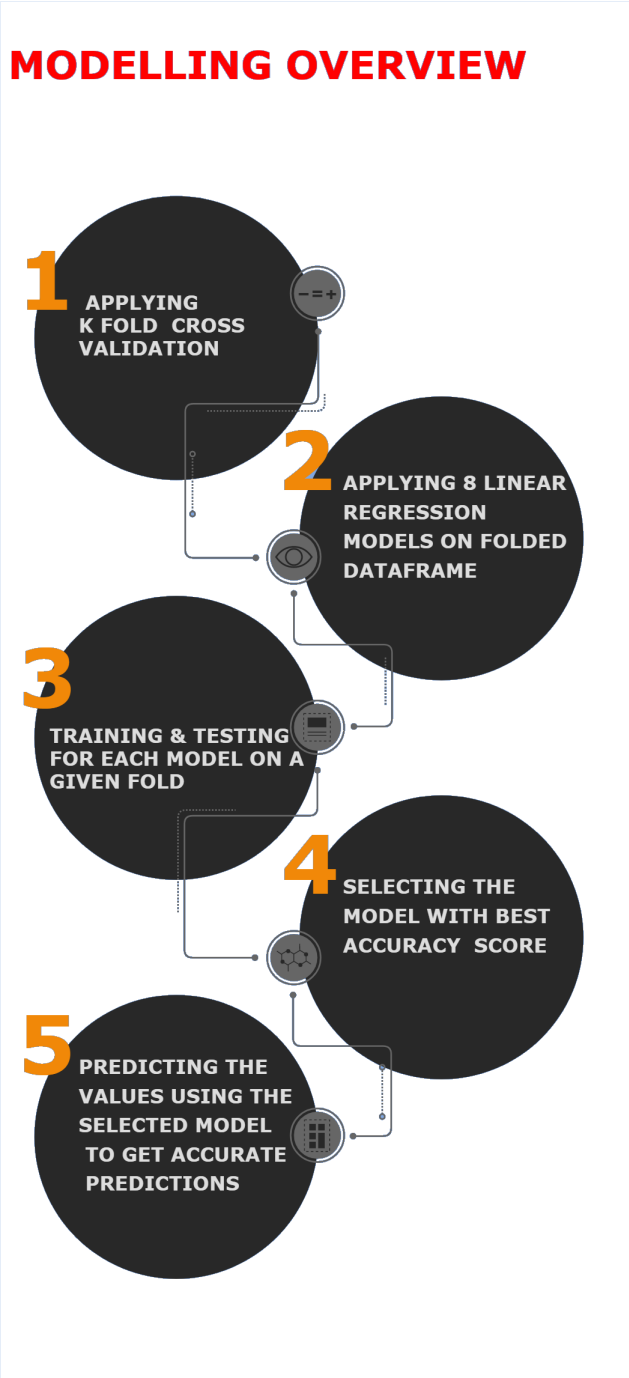

Figure 4: Flowchart for Prediction

0.4 Implementation

0.4.1 Data Pre-processing

The dataset consists of electricity consumption data (December 2016 to Jan 2018) from a high-rise residential building inside the IIT Bombay campus. The building consists of 60 3BHK(3 Bedrooms, I Hall and a Kitchen) apartments, each instrumented with a smart-meter, logging data at a sampling period of 5-8s. The data shared in the link in References is downsampled at 1-hour granularity. All the timestamps mentioned in the dataset is of Indian Standard Time(GMT+5.30). For privacy reasons, apartments are kept anonymous. The folder consists of 39 CSV files each representing an apartment.

Apartments having significant data loss are removed from the list.
The headers in the CSV files are as follows:

1. TS - Unix Time stamp (epochs)
2. V1 - Voltage of phase 1 (V)
3. V2 - Voltage of phase 2 (V)
4. V3 - Voltage of phase 3 (V)
5. W1 - Electricity consumption of phase 1 (Wh)
6. W2 - Electricity consumption of phase 2 (Wh)
7. W3 - Electricity consumption of phase 3 (Wh)

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

The Apartments with more than 5000 data entries were selected for Analysis and Prediction, while remaining apartments were discarded. The selected apartments are shown in tabular format on the next page:

This dataset was then Normalized for the entire year 2017.

Apartment No.	No. of Data Entries
1	7508
2	6042
3	7425
4	7425
5	7422
6	7480
7	7462
8	7508
9	7508
10	7508
11	7501
12	6928
13	36
14	799
15	5334
16	6408
17	799
18	799
19	799
20	650

Apartment No.	No. of Data Entries
21	7315
22	799
23	799
24	799
25	799
26	799
27	765
28	7315
29	7114
30	7315
31	7089
32	7315
33	7315
34	7158
35	7315
36	7315
37	7509
38	7478
39	799

Figure 5: No. of Entries in Raw Dataset

Apartment No.	Data Entries in 2017
1	6971
2	5721
3	6894
4	6894
5	6891
6	6943
7	6925
8	6971
9	6971
10	6971
11	6964
12	6452
15	5101

Apartment No.	Data Entries in 2017
16	5872
21	6838
28	6839
29	6713
30	6839
31	6552
32	6839
33	6839
34	6713
35	6839
36	6839
37	6972
38	6941

Figure 6: No. of Entries in Processed Dataset for the year 2017

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Figure 7: Normalization formula

The Normalized dataset is then arranged in following format Monthwise and interpolated Linearly. The data was then arranged as below with hours as the rows and days as columns:

	0	1	2	3	4	5
0	0.34425	0.34038	0.39175	0.38165		0.38118
1	0.34199	0.35054	0.33931	0.35276		0.36285
2	0.38585	0.35847	0.37743	0.38665		0.37465

Figure 8: Non Interpolated dataset

	0	1	2	3	4	5
0	0.34425	0.34038	0.39175	0.38165	0.38141	0.38118
1	0.34199	0.35054	0.33931	0.35276	0.35781	0.36285
2	0.38585	0.35847	0.37743	0.38665	0.38065	0.37465

Figure 9: Interpolated dataset

A similar method was adopted for the Temperature Dataset as well. These datasets obtained were used for prediction and consumption analysis. The original dataset was directly used for Voltage Profile Analsysis.

0.4.2 Data Analysis

- Power Consumption Analysis:

  The dataset normalised for entire year of 2017 was used for this analysis. The power consumption data for each apartment is fetched according to month and stored in a dictionary with two-dimensional keys. This dictionary was then used to calculate the Total Power Consumption of each apartment for every month. This data obtained was then plotted for Graphical Visualization(X-Month,Y-Power).

  The apartments were then separated according to their Monthly Power Consumption. The criteria adopted for this separation was as follows:

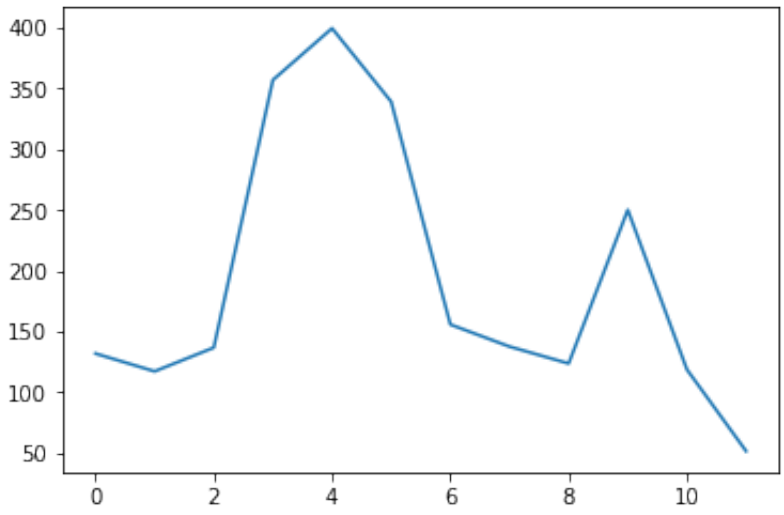

Figure 10: Example Graph: Apt7

1. Low Consumption Household is the one which consumed lesser than the average consumption of all apartments for that particular month.
2. High Consumption Household is the one which consumed greater than the average consumption of all apartments for that particular month.

- Voltage Profile Analysis:

  The voltage data of all the three phases was used for graphical and inferential analysis. The interpolation methods used in the other two parts was not carried out here, as the analysis is directly based only on the available data.

  The three phase supply to each apartment is a balanced three-phase power, where all three phase voltages are equal in magnitude and 120 degrees apart in phase.

  The line voltage was then calculated using the formula:

$$\text{Line Voltage} = (\text{V1} + \text{V2} + \text{V3})/\ \sqrt{3}$$

  Graphs of all the four obtained Voltages were plotted for Graphical Visualization.

  Similarly instances of low profile voltage were separated and saved for determining faults in the supply line. This analysis helps us to utilize the data from smart meter to prevent overheating or derating of appliances in near future. The analysis can also help user to keep a check whether the utilities are complying with the norms and regulations set up by the government.

0.4.3 Data Prediction

Multiple Linear Regressions Method:

This model assumes that the load at a particular period can be estimated by a linear combination of some independent variables. Generally, the longer the data set, the better the result in terms of accuracy . A larger computational time for parameter identification is required. The multiple linear regression trend models are generally expressed as:

$$y = b0 + b1*x1 + b2*x2 + ... + bn*xn$$

y=Observation of Dependent Variable
x=Independent variables
b=Slope Coefficients for each of the Independent Variables
b0=Error Term

PERFORMANCE MEASURES

This study uses Mean absolute percentage error (MAPE) and measures the accuracy of the model predicting electricity consumption. It expresses accuracy as a percentage. The MAPE is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}.100\%$$

TRAINING AND TESTING MODEL

The Regression Models used are:

1. Linear regression
2. Thiel-Sen Regression
3. PassiveAgressiveRegressor
4. Lasso-Lars
5. BayesianRidge
6. ARDRegression

7. SVR Regressor

Independent Variables

Electricity consumption was found to depend on the day, month, weekday and climate. In this study, the following factors are carefully selected for the prediction of electricity consumption:

1. Month: Helps determine seasonal trends of Electricity consumption.
2. Day: Gives trends between the day of a month and Electricity consumption.
3. Weekday: Gives trends between day of the week and Electricity consumption.
4. Climate : Gives relation between Temperature and Electricity consumption.

Dependent Variables

1. Electricity consumption.

	coef	std err
Month	-0.0011	0.001
Day	0.0004	0.000
weekday	0.0023	0.001
Temperature	0.0115	0.000

Figure 11: Evaluated coefficients using Stats Model

	weekday	Month	Day	Power	Temperature
weekday	1.000000	0.012860	-0.012152	-0.053332	-0.003486
Month	0.012860	1.000000	0.011893	-0.190052	-0.167732
Day	-0.012152	0.011893	1.000000	-0.018602	0.093002
Power	-0.053332	-0.190052	-0.018602	1.000000	0.234584
Temperature	-0.003486	-0.167732	0.093002	0.234584	1.000000

Figure 12: Correlation between various parameters

K-fold cross validation was applied on the dataset.The dataset was divided into training and testing data in the ratio of 4:1. This training and testing data was then evaluated on each of the 7 regression models for each fold. The accuracy of each model on different folds was calculated using Mean Absolute Percentage Error(MAPE) formula and stored. Finally, the model with best accuracy score was selected. Using the selected model the

future Power Consumption values were predicted based on data entered by user.

The Predicted values of Power Consumption were further used for monthly Bill Prediction.

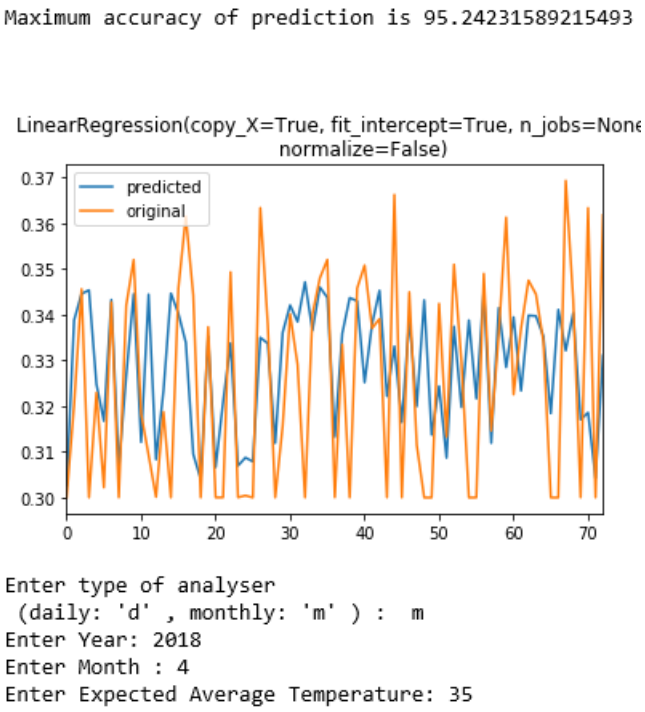

Figure 13: Plot of Predicted and Testing dataset

0.5 Results

Predictor:

```
Enter Apartment number: 1


Maximum accuracy of prediction is 95.24231589215493


Enter type of analyser
 (daily: 'd' , monthly: 'm' ) :  d
Enter Year: 2018
Enter Month : 5
Enter Day : 31
Enter Expected Average Temperature: 32.275




Expected units:
8.141553768039588


Actual units:
8.841538544415
```

Figure 14: Power Consumption Values for 31st May, 2018

The given dataset was trained and tested only for year 2017 due to lack of enough datapoints for the year 2018. The model has scope for further development once more datapoints are made available.

Power Consumption Analysis:

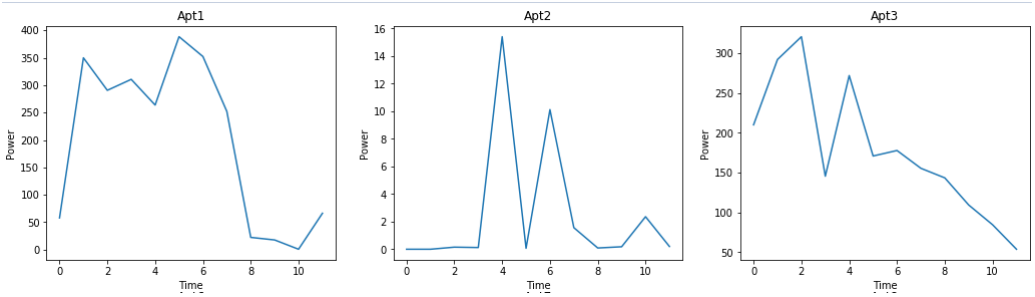

Figure 15: Graphs showing Yearly Consumption of Apts 1,2,3

	High	Low
0	Apt1	Apt2
1	Apt3	Apt6
2	Apt4	Apt7
3	Apt5	Apt9
4	Apt8	Apt12
5	Apt10	Apt15
6	Apt11	Apt16
7	Apt21	Apt30
8	Apt28	Apt31
9	Apt29	NaN
10	Apt32	NaN
11	Apt33	NaN
12	Apt34	NaN
13	Apt35	NaN
14	Apt36	NaN
15	Apt37	NaN
16	Apt38	NaN

Figure 16: Apts separated as per Power consumption

From this analysis, we can observe that the general trend for Power Consumption during Summer months is higher than any other months. Two apartments(Apt2, Apt31) showed very less power consumption (almost negligible) which signifies that these apartments were unoccupied during the given period.

Voltage Profile Analysis:

For Line Voltage :

	date_time	tot_vol
2017-04-27	04:30:00	387.072984
2017-05-07	05:30:00	389.792516
2017-09-21	02:30:00	385.972621
2017-10-20	12:30:00	388.646271
2017-11-01	12:30:00	389.615416
2017-11-29	01:30:00	388.900331

Figure 17: Table showing instances of Low Line Voltage

Figure 18: Scatter plot showing Voltage Profile of 3 phase supply line

The analysis indicated that the Phase 1 supply line was indeed faulty, the reason being unknown, and can be further analysed manually. The other two phases lines didn't show much low profile voltage instances. Further Line Voltage, being dependent on all three phase lines, was not found to be affected by the frequent low voltages occurring at phase 1 supply line.

0.6 Conclusion

Hence, using Exploratory Data Analysis, the apartments have been analyzed individually using Graphical and Descriptive methods of Analysis.

The apartments have been filtered on the basis of Monthly Power Consumption. The analyser can also assist the governmental bodies to study consumption characteristics of various apartments and plan their future policies accordingly.

By using Voltage Profile Analysis, a system was developed to check whether the utilities are complying with the norms and regulations set up by the government. Apartment wise analysis helped to inspect faults in supply line of the particular apartments , while building wise analysis helped to inspect faults in distribution line. The analysis results can then be used by the consumers to lodge complaints to their supplier, thus preventing any possible harm to the connected household appliances.

The various Linear Regression Models under Scikit Learn were implemented on the given dataset and model giving the optimum results was chosen at runtime. Based on the selected model, the future values were predicted based on parameters provided. The predicted values were compared with actual values successfully and showed negligible variance. Thus, training and testing were performed successfully.

References

- Link for Raw Dataset: *link*
- Data Filtering:
    1. Pandas: *link*
    2. Normalization (statistics): *link*
- Data Analysis:
    1. Pyplot from MatplotLib Module: *link*
    2. Voltage Profile Analysis: *link*
    3. Voltage Regulations as per MAHAVITARAN: *link*
- Data Prediction:
    1. Forecasting Electricity Consumption: *link*
    2. Using regression analysis to predict the future energy consumption of a supermarket in the UK: *link*
    3. Regression analysis for prediction of residential energy consumption: *link*
    4. Predicting Energy Usage of School Buildings Daniel Sambor, Rohith Desikan, Vikhyat Chaudhry, CS 229: Machine Learning, Stanford University, Fall 2016: *link*
    5. Regression analysis for prediction of residential energy consumption: *link*
    6. Scikit Learn: *link*