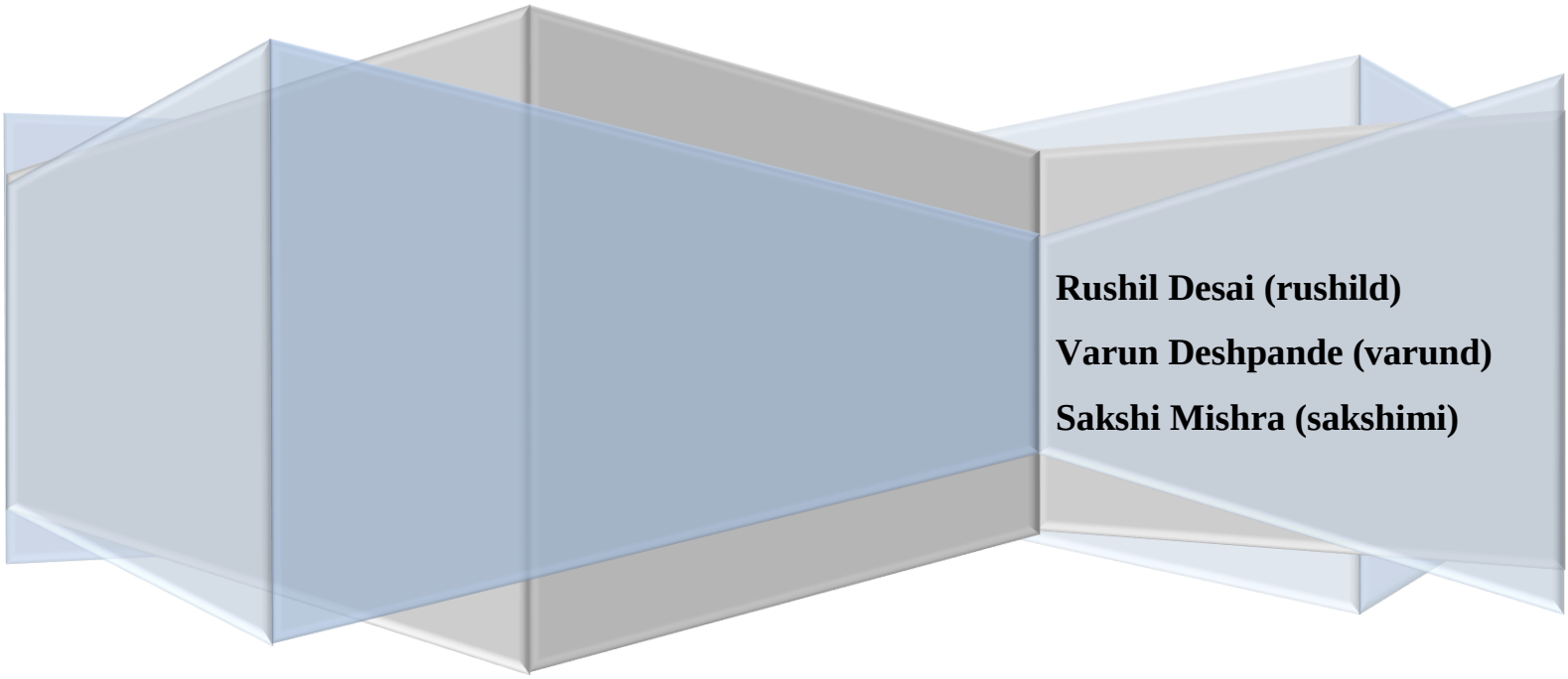
Carnegie Mellon University

Comparison of the accuracy of line-ar regression model based on the granularity of data: Scaife Hall

12-752 Data-driven Building Energy Management



Rushil Desai (rushild)

Varun Deshpande (varund)

Sakshi Mishra (sakshimi)

Abstract

The objective of this project is to analyze the dependence of power consumption in Scaife Hall, Carnegie Mellon University on outside air temperature and time of week. A linear regression model is developed to predict power consumption. The model has been trained on a part of the dataset and tested on the remainder. The accuracy of the model is determined by analyzing the statistical properties of the coefficients of the regression model. Additionally, the advantage of availability of granular data has been assessed by testing the model on the load profiles of individual components/panels of Scaife Hall.

Contents

Chapter 1

1. Introduction

"Sound energy management" strategies can lead to significant improvements in the energy consumption of a building. A variety of strategies are available for the same, from behavioral programs to retrofits. But it is important to establish a baseline and to be aware of it. Even the simplest of strategies can lead to significant energy savings but only if one is aware of their energy consumption. As stated by Steve Hanawalt in "Energy Management as a Corporate Strategy", an "energy improvement opportunity" can be explored by asking 3 basic questions per energy asset [**1**]:

- What it is
- What it should be
- What it could be

"What it is" refers to identifying the present energy consumption i.e. setting up a baseline in order to evaluate against. "What it should be" refers to identifying design stage energy consumption of the current energy assets. "What it could be" refers to the best case scenario [**1**].

In this project, we are trying to answer the first part i.e. "What it is". The project aims at understanding the dependence of power consumption of various types of loads on time of week. Nowadays, it is much easier to get access to granular data as opposed to overall building-level power consumption with the increasing use of sensors. The sensors installed in Scaife Hall are being used to gather power consumption data for each of the components in the building such as lighting, fan coils, drinking water fountains, HVAC, and so on, for rooms, auditoriums, stairways etc. We would also like to take advantage of the availability of such a granular data to assess whether load prediction can be made more accurate if the power consumption of individual components/panels is modeled instead of that of the entire building.

Chapter 2

2. Dataset Description

Power Consumption data of Scaife Hall has been collected, with power consumption values measured at 1 minute intervals. The Scaife Hall data is classified on the basis of panel name, which is further divided into branches where each breaker corresponds to a branch. The dataset contains minimum and maximum current, minimum and maximum voltage, average active power, and average reactive/apparent power for each branch. There is significant inconsistency in the labeling methodology for the data. The data which we have specifically focused on is:

- Overall lighting power consumption
- Lighting power consumption of each panel

Chapter 3

3. Data Analysis Methodology

3.1 Data Selection

Though 1 minute interval power consumption data is available, we have scaled it up to data separated by 15-minute intervals because a 15-minute interval is sufficient to determine correlation between the dependent and the independent variables i.e. the power consumption and the time of week or outside air temperature. Also, the computer's RAM limitations are posed on having a massive amount of data with 1-minute interval loaded. Additionally, trying to predict load variations for 1-minute intervals may lead to unnecessary noise i.e. choosing data spaced out using 15-minute intervals effectively smoothens out the power consumption values. Thus, we chose to conduct the analysis on a 15-minute interval basis.

'AvgWatt' (power consumption), 'DateStamp', 'Panel Name' and 'Branch Name' are the four parameters which are being used throughout the analysis. For the purpose of evaluating the advantages of granular data v/s overall data, we have chosen to work on power consumption data corresponding to the spring semester (mid-January to mid-April).

3.2 Data Processing

Two years of raw data is available in csv (comma separated value) files, one for each day of the month. The total size of this data is 15 GB. We have selected data from months January to end of March for our analysis assuming this to be the active period of the spring semester at Carnegie Mellon. The three month data was roughly 2.5 GB and was loaded into the ipython notebook sequentially and consolidated into one data frame.

Each of the csv files had multiple columns of data including Branch ID, Branch Name, Breaker position, Panel name, Date stamp, Maximum and minimum current and voltage, phase ID, location ID, average active and reactive power.

Upon loading of the data into the data frame, only the relevant columns were extracted and saved which included Date stamp, Panel name, Branch name and average active power.

Further, based on the contents in the Branch name column, only the rows with the lighting loads were highlighted and extracted.

On this remaining data, date conversion from string to timestamp object was performed in order to be able to sub set the data later in the code based on the timestamp.

As a final process, only data at 15 minute intervals was chosen for the analysis by sub setting it based on the timestamp.

The data was segregated into two categories to test the robustness of the model - training data and testing data. The segregation was done using the isocalendar() function in the DateTime

module assigning alternate weeks to the two datasets. The dataset considered begins from a Thursday in mid-January. The first four data entries of the training dataset were not considered in order to initialize both the training and testing datasets from the same day of the week.

3.3 Analysis

To understand the dependence of power consumption of the specific panels on the time of week, we have divided our analysis in two parts where we compare the prediction accuracy of the regression model with and without granularity with respect to panel level data against the total building power consumption data. The analysis has been performed on one specific component of the building - "Lights".

The first step is to store the average power consumption values for all the lights in Scaife Hall as one variable and store the date and timestamps of the study horizon as another variable. The power consumption is the dependent variable and time of week is the independent variable for the regression model. The next step is to develop the model for predicting the power consumption. For model development, firstly the Design Matrix (DMX) is constructed with the number of columns equal to the number of 15-minute time intervals in a week and the number of rows in the DMX are equal to total number of data points in the duration of the dataset (about 6 weeks of training and 6 weeks of test data).

As the first step, the power consumption of the overall lighting load in Scaife Hall is predicted after segregation into training and testing datasets. In order to assess the performance of the linear regression model with panel level data (as opposed to the building level data), the second step is to construct linear regression model for individual panel's data, which are "Third Floor Ladies' Room", "Fourth Floor Men's Room", "Dean's Office 2", "Penthouse 277", "Scaife Hall 2F".

3.4 Model Validation

The model has been validated by calculating the R^2 values of the estimated beta coefficients to understand model's accuracy. The R^2 value for the overall building's lighting power consumption is 0.7779.

The R^2 values for panel level data are as follows.

"Third Floor Ladies' Room": 0.0002

"Fourth Floor Men's Room": 0.2259

"Dean's Office 2": 0.0758

"Penthouse 277": 0.7577

"Scaife Hall 2F": -0.1831

As can be observed, the model seems to be reasonably accurate in case of the total lighting power consumption as well as lighting consumption for the 'Penthouse 277' panel. Reasoning for the same and further analysis have been provided in the next section. Here, the panels

named Auditorium 1 and Auditorium 2 have not been considered because the data shows that lighting is not a component in those panels.

Chapter 4

4. Results and Inferences

4.1 Results and Discussion

Figure 4.1 shows the lighting power consumption of the building over the study period. Spring semester is defined from mid-January to the first week of April for the purpose of this project. Differences in the weekday and weekend power consumption can be visually identified from the plot.

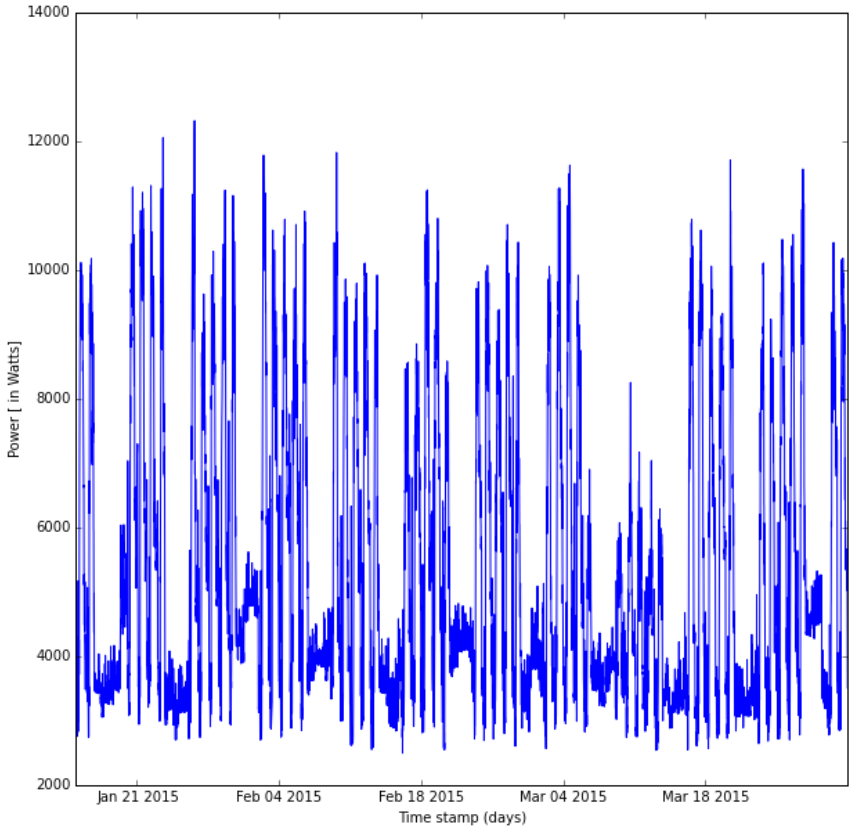


Figure 4.1 Total lighting load profile for Scaife Hall during the Spring semester

The plot of the actual power consumption and the predicted power consumption for the overall lighting data has been shown in Figure 4.2. The actual power consumption has been depicted in blue while the predicted power consumption in green.

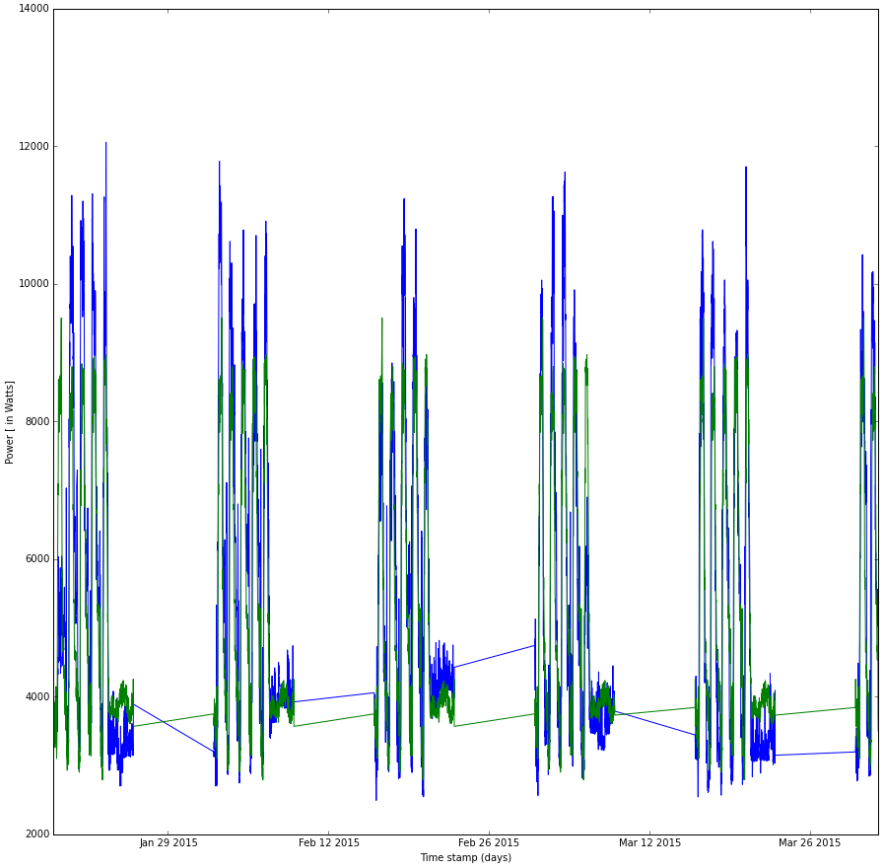


Figure 4.2 Actual (blue) and Predicted (green) power consumption for building-level lighting

The plot of the actual power consumption and the predicted power consumption for the "Third Floor Ladies' Room" panel has been shown in Figure 4.3. The actual power consumption has been depicted in blue while the predicted power consumption in green. A significant and consistent difference can be observed between the actual and predicted values.

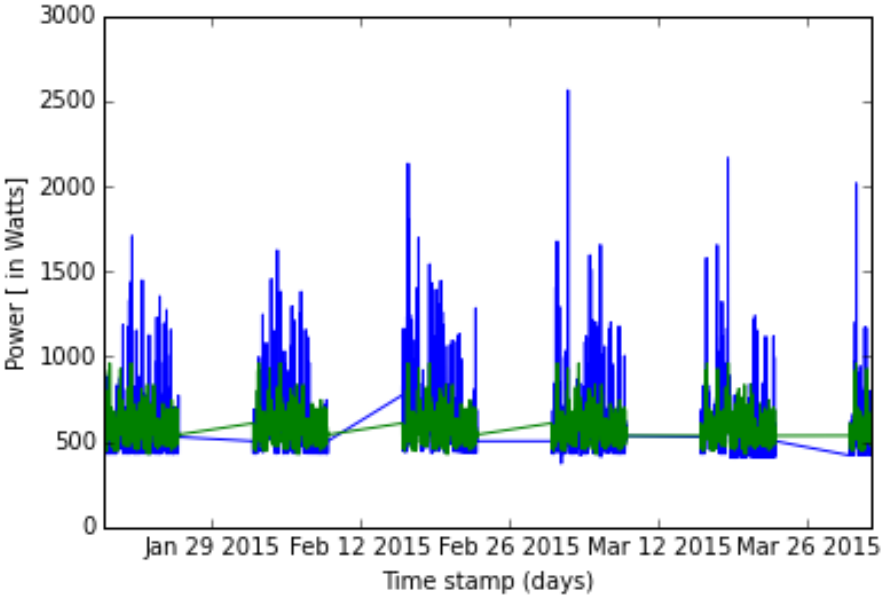


Figure 4.3 Actual (blue) and Predicted (green) power consumption for Panel 1

The plot of the actual power consumption and the predicted power consumption for the "Fourth Floor Men's Room" panel has been shown in Figure 4.4. Again, the difference between the actual and predicted values is large and consistent throughout the study period.

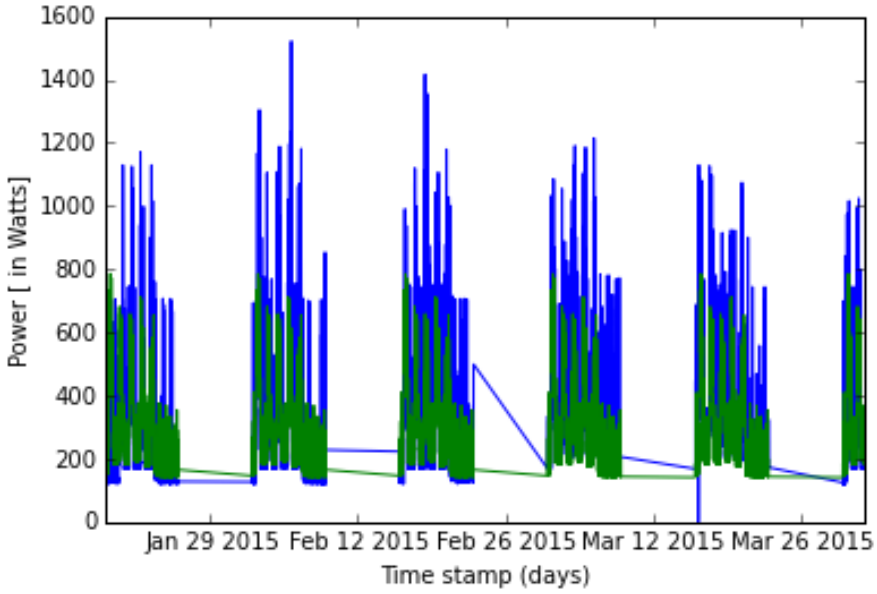


Figure 4.4 Actual (blue) and Predicted (green) power consumption for Panel 2

The plot of the actual power consumption and the predicted power consumption for the "Dean's Office 2" panel has been shown in Figure 4.5.

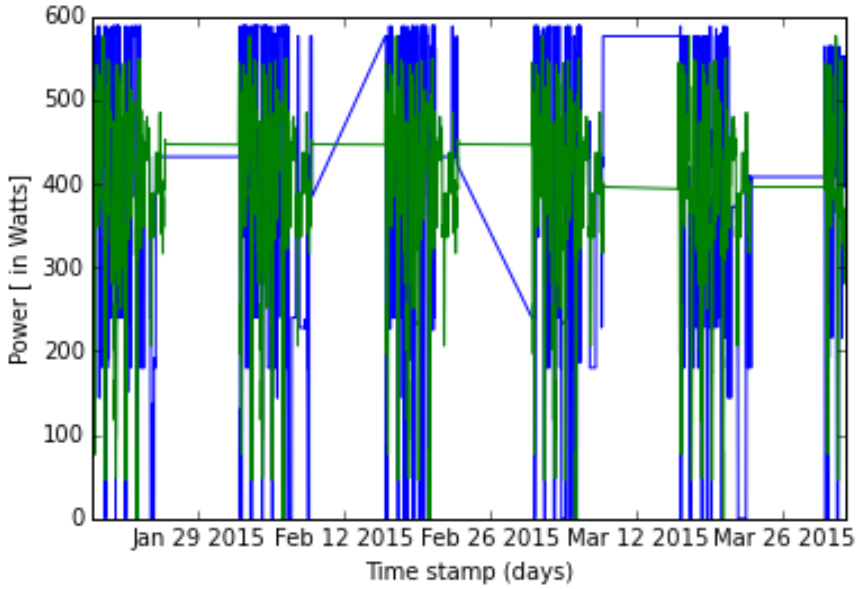


Figure 4.5 Actual (blue) and Predicted (green) power consumption for Panel 3

The plot of the actual power consumption and the predicted power consumption for the "Penthouse 277" panel has been shown in Figure 4.6.

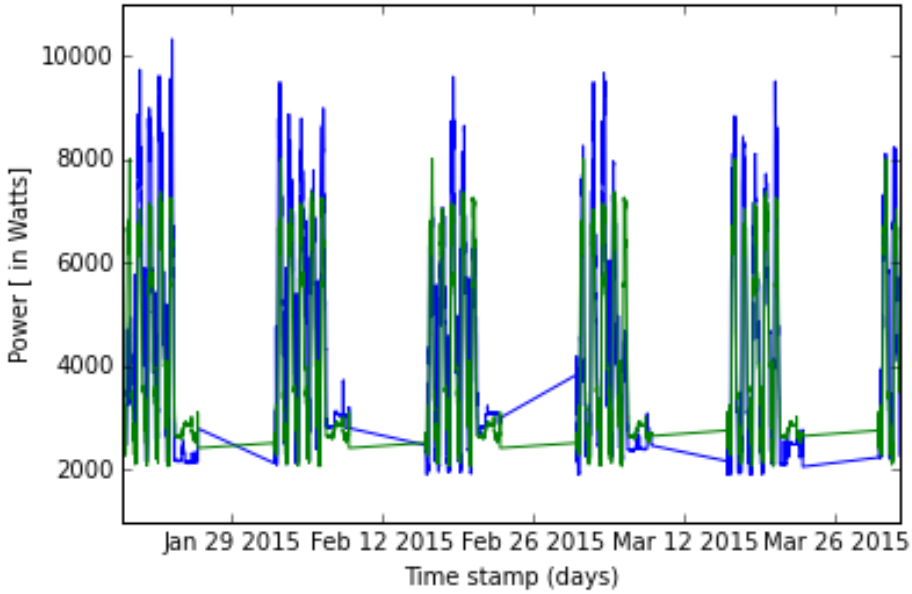


Figure 4.6 Actual (blue) and Predicted (green) power consumption for Panel 4

The plot of the actual power consumption and the predicted power consumption for the "Scaife Hall 2F" panel has been shown in Figure 4.7. The actual power consumption seems to be deviating significantly from the predicted values.

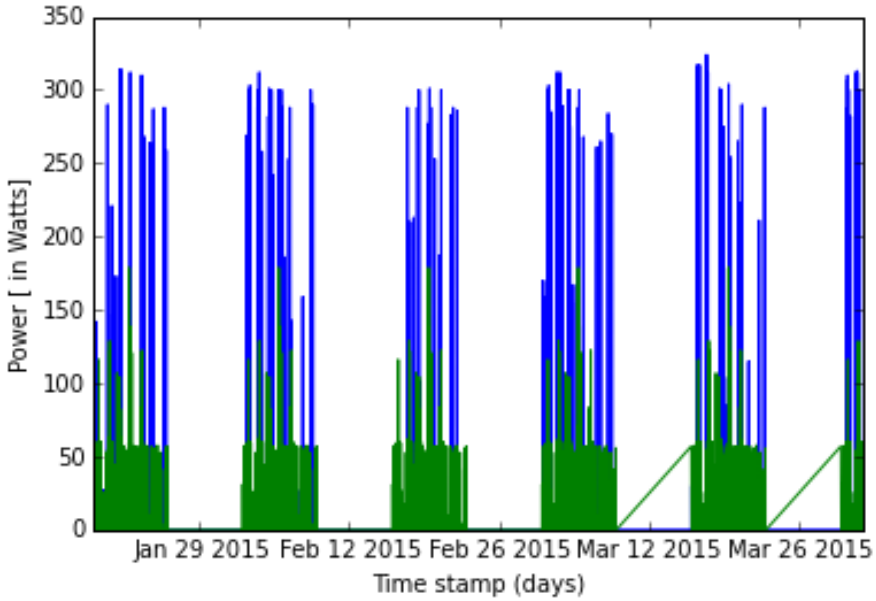


Figure 4.7 Actual (blue) and Predicted (green) power consumption for Panel 5

4.2 Inferences and Interpretation

It can be inferred from the plots that the predicted values are following the "weekday" and "weekend" trend pretty well. On the other hand, there is a noticeable difference between the magnitude of the actual and predicted power at any specific data point, as has been discussed. We believe that there exist some other variables such as cloud cover and occupancy schedule

(class timings, office hours etc.), influencing the data which have a consistent impact on the magnitude of the power consumption with respect to the actual power.

The R^2 value indicates how well data fits a statistical model. So, the R^2 values obtained clearly suggest that predicting lighting power consumption for the whole building has higher accuracy compared to the individual panel level power consumption predictions.

We believe, there can be two possible reasons for this difference in the prediction accuracy:

- The panel level consumption is not just based on the time of week. It is rather also dependent on the factors like class timings, events, office hours. On the other hand, for building level consumption, these factors do not play a significant role because they get normalized. Since building-level data is an aggregate of every individual panel, anomalies or fluctuations in individual panel data get absorbed/smoothened out while carrying out the summation
- For understanding panel level consumption, we need better information regarding transition of occupants between floors, which is not a significant attribute to consider for building level consumption since internal transitions do not affect overall occupancy of the building

This suggests that granulizing data is not a trivial job i.e. there is a certain grouping methodology that must be followed while collecting and grouping granular data in order to generate meaningful load prediction models.

Chapter 5

5. Future work

Although a reasonable correlation has been observed between lighting power consumption and time of week, the prediction might have been more accurate if we had taken other variables such as cloud cover and occupancy schedule (class timings, office hours etc.). Taking more variables into account would be more data-intensive but it would probably also lead to development of a more robust model. Additionally, with better labeling of panels and branches, deciphering labels and grouping data would become much smoother.

References

[1] Steve Hanawalt, "Energy Management as a Corporate Strategy," Uptime Institute, 2009.