**Predicting Solar Power Generation**

MIS 776 – Business Intelligence

November 30, 2020

Jungju Lim

Tracy Strycharski

Robert Thomas III

Aaron Wong

# Project Summary

This data analysis seeks to examine the power generation and weather data for two power plants in India to determine how that data can be used to predict power generation under a variety of conditions. In addition, the data also provides an opportunity to examine under what conditions an inverter may be performing suboptimally and to predict the factors that may lead to that suboptimal performance. Several methods were used to investigate the data including multiple linear regression and clustering techniques. The result of the analysis was a predictive model to forecast expected daily power generation at each plant and the discovery that one of the plants was significantly underperforming its predicted output.

# Introduction and Data Overview

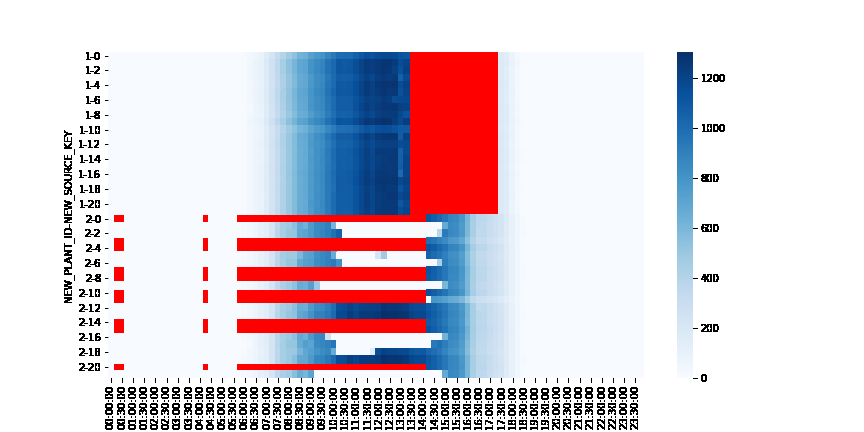
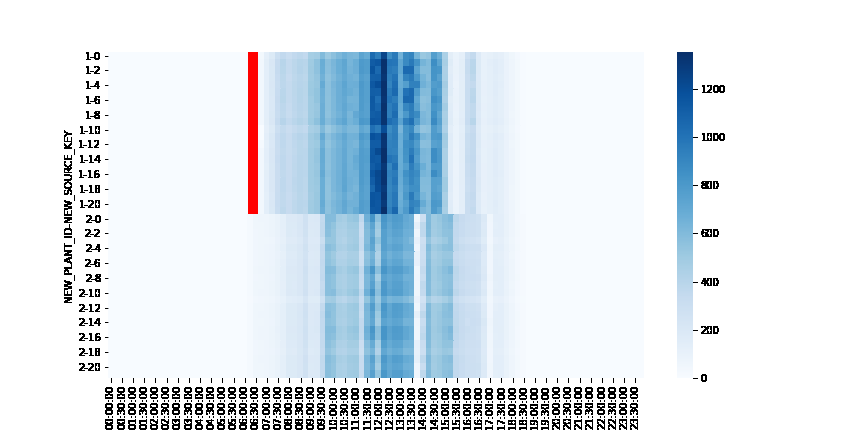
The dataset contains data from two solar power plants in India, gathered over a period of 34 days. The power generation data and corresponding weather sensor data about level of sunlight (irradiation) and the ambient temperature were recorded in 15-minute increments over the 34-day period. The data was stored in four distinct csv files: two contained data on AC, DC, and total power generated by both firms, and the other two contained corresponding weather information over the same timeframe.

The data in this dataset is compelling because it has the potential to generate analyses that can predict future power generation based on anticipated weather conditions, as well as to proactively identify underperforming or malfunctioning inverters in order to keep the plant running smoothly. This information is useful because it allows more effective management of the power grid and, when coupled with analyses that forecast demand, can enable the plant to ensure that it can meet that demand. Additionally, having data from two different power plants allows a comparison of plant output which allows a more effective analysis of plant performance overall.

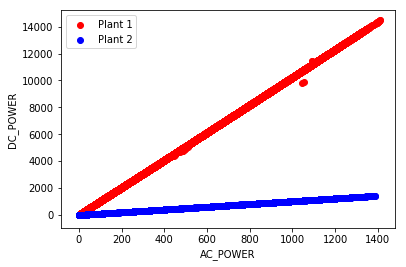
# Methods of Analysis

The first step of analyzing the data was to combine the power generation data for both plants into one file and then join the weather data files to create a single file for analysis. The combined date time field was separated into two separate fields, and the alphanumeric source keys for each plants’ 22 power inverters were replaced with a more manageable integer identifier.

We then began our initial review of the data by creating heatmaps of the AC power generation by inverter and time, which revealed several artifacts, illustrated by the below images. The inverters of the two plants are grouped together for visual clarity, and darker colors represent more energy production, while red spots indicate times when the inverter was offline. The image on the left is a day with effective power generation characterized by the inverters being offline while it was dark and the power generation roughly matches the solar cycle. The image on the right is a day when multiple inverters were not functioning, some inverters were generating no power during peak sunlight, and the entire first plant went offline for approximately a third of the daylight hours, though it was functioning well before that.

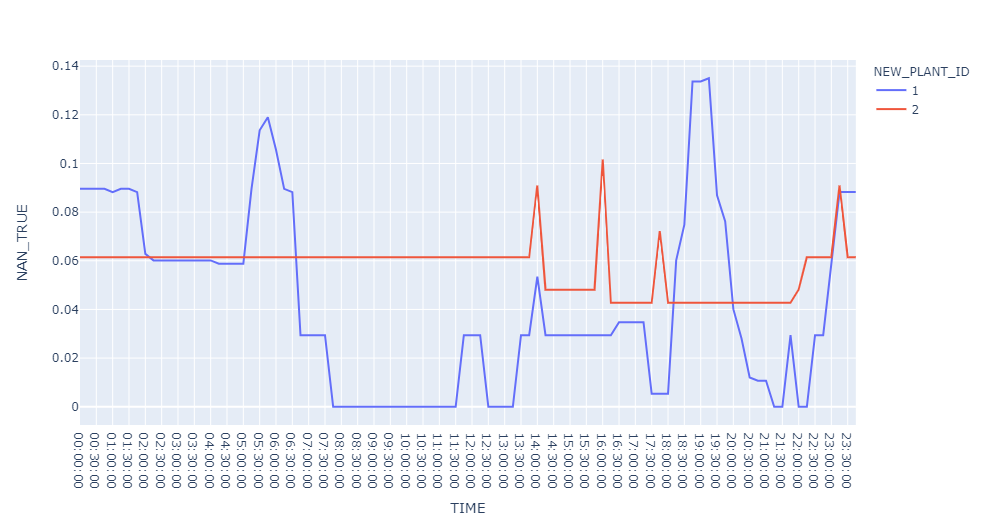


We also observed a significant difference between the two plants in their reporting of AC and DC power, as illustrated below. It is either the case that Plant 2’s AC\_POWER conversion is underperforming by a factor of 10 or they are reporting their data with the incorrect units.

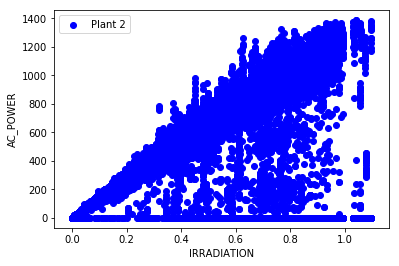
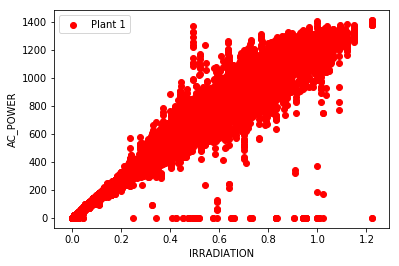


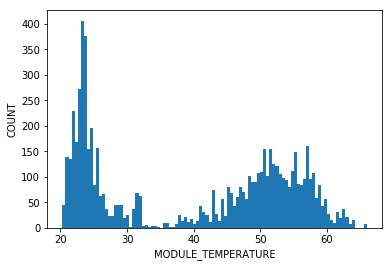
Hopefully, this is just a reporting issue and not a production issue, otherwise there are immediate serious deficiencies to address. Because the reported values of DC\_POWER appear to be off, we have used AC\_POWER production for the remainder of this report, as it provides a more reliable comparison between the two plants.

We then performed a downtime analysis of the two plants by counting the number of time intervals at each plant on each day that failed to report the current production. The following graph shows the total count of the inverters that were offline during each time interval over the entire observation period. It illustrates that Plant 1 is primarily taking its units offline just before first sunlight (which occurs at approximately 6 AM during this period) and just after darkness falls (which occurs at approximately 6 PM). These spikes show regular maintenance activities are being performed at specified times. Plant 2’s data reflects a period during which a number of inverters were offline for extended periods during a 9 day stretch and no periods that seem to reflect regular maintenance activities.

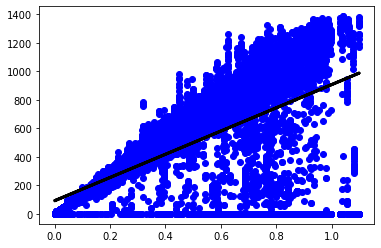
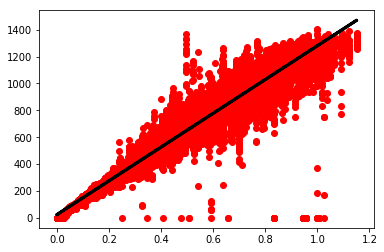


In order to come to an understanding of the consequences of differences in the management of the two plants, we plotted the level of sunlight compared to the energy production for each plant separately. Plant 1 shows very clearly that increased irradiation leads to increased power production. However, Plant 2 shows significant levels of underperformance, as demonstrated by the data points that fall below the main diagonal.



We want to bring particular attention to the horizontal line of data points from Plant 2 along the x-axis. These points correspond to times at which there is sunlight but zero energy production. When we isolated those points and plotted a histogram of the temperature of the module during those events, we discovered a clear bimodal pattern. The lower peak corresponds to very low light situations, such as early in the morning or late in the evening. The second peak appears to be periods where the equipment is overheating and is no longer capable of producing energy. This needs to be investigated on site to determine the root cause of the problem.

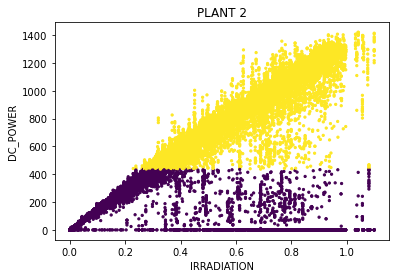
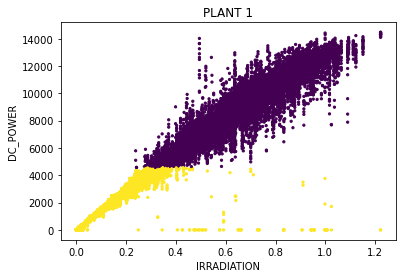
With this in mind, we set out to create a predictive model for energy production for each plant based using irradiation to predict AC power generation. We trained the model on the first several weeks of data and used that to calculate the expected output for the last week of data. The following graphs show the line of best fit for both models.



For Plant 1, when we gave it the irradiation levels for the final week, it predicted that the plant would generate 4,042,680 units of energy. The true energy generation was 4,027,380, which is accurate to within 0.5%. This is confirmation that the model works very effectively, and that Plant 1 continues to perform optimally.

For Plant 2, the prediction for the final week based on the measure of irradiation was that the plant should generate 3,480,916 units of energy. Instead, it only produced 2,924,922, which is a 16% underperformance by the plant. It is also worth noting that the model itself was created including the known periods of overheating, which lowered the expectation, and the plant still came up significantly short of its potential.

In an attempt to improve the predictive model, KMeans clustering was performed on the combined data from each plant with clusters constructed based on the irradiation, DC power, and module temperature. A characteristic challenge of KMeans clustering is in the selection of the optimal number of clusters. To determine this number, the KMeans algorithm was performed in a loop with the number of clusters incremented and the silhouette score, which is used to measure the separation distance between the resulting clusters, recorded for each number of clusters. For both plants, the optimal number of clusters was 2 with plant 1 having an average silhouette score 0.77 and plant 2 at 0.79. Once the clusters were identified, regression analysis was performed on each of the clusters to predict DC Power generation. However, the average completeness score, measuring the precision of the clustering, for each of the clusters was only 0.56, while the average homogeneity, a measure of clustering accuracy, was only 0.73. Each of these scoring measures range from 0.0 to 1.0, so the resulting clusters were suboptimal for prediction of the DC power generation. An additional requirement for KMeans clustering is that the clusters be globular in shape. Our datasets, as demonstrated from the scatterplots of the clusters below, did not fit this requirement.



# Operational and Business Insights

Our initial aim was to use the information contained in the dataset to build a predictive model to forecast expected power output based on weather data and identify underperforming inverters. However, based on our findings from the EDA, we concluded that while Plant 1 is operating within generally acceptable parameters, Plant 2 is highly deficient in a number of ways, and that this discrepancy may be a better focus for our data model. We set out to identify the drivers of Plant 2’s underperformance and theorized that we would be able to create a model that can quantify Plant 2’s underperformance in terms of actual energy production. The results of our analysis could then be used to provide advice and guidance to Plant 2’s management in order to bring their operation up to the same standard as Plant 1.

Our analysis generated several critical insights. This first, as described in the previous section, is the magnitude of Plant 2’s underperformance to its expected output. This 16% underachievement has serious implications for the management of the power grid in the local area and could impact the plant’s ability to meet demand. Additionally, we identified a lack of regular maintenance as a potential root cause of the underperformance, a recommendation that could be further investigated by management in an effort to rectify its lagging output.

# Conclusion

The dataset provided an opportunity to investigate the relationship between multiple weather-related factors and solar power generation and use this information to create meaningful insights to help maximize energy output and efficiency. After discovering the discrepancies between the two plants during our exploratory data analysis, we narrowed our analysis to focus on understanding the impact of Plant 2’s underperformance as well as to investigate the possible drivers of this deficiency. We applied both linear regression and clustering techniques to create predictive models for each plant, which confirmed our hypothesis that Plant 2’s data irregularities corresponded with a much lower total output than expected under optimal conditions. This type of analysis has multiple applications for the plant’s management team and can ultimately be used to reduce the costs of production by maximizing efficiency in both plants.