Capstone Two Project Proposal: Solar Power Generation

Solar power plants are becoming an ever-popular and prevalent source of energy production due to the decreasing costs of solar panels and increase in climate action by public and private institutions. A photovoltaic solar power plant is a facility that converts sunlight directly into electricity through the collection of solar energy through solar arrays and inverts the collected Direct Current (DC) energy into Alternating Current (AC) power. The solar arrays are made up of a series of modules (solar panels) that collect the energy from the sun and convert it into electricity which is then sent to an inverter to be converted into the usable format for the energy users. The major factors affecting the power generation at solar plants can be related to weather, cleanliness, efficiency, or maintenance. In this project, I will seek to identify equipment that is operating sub-optimally and predict the power generation over the next several days for the two power plants to better manage the plants and the expectations of those managing the energy grid.

The primary data for this project is the power generation information for two power plants in India. This is contained in two separate CSV files, one for each power plant, and has columns for the datetime, plant ID, source key (inverter ID), amount of DC power, amount of AC power, and daily yield, and total yield. Additionally, there is weather sensor data for each power plant, and contains columns for datetime, plant ID, source key (sensor panel ID), ambient temperature, module temperature, and irradiation. Both datasets are collected at 15-minute intervals over a 34-day period.

Using this data, I will investigate the reliability and energy production of the individual inverters at each plant. The data is likely going to peak during the day and be zero for each inverter over the nighttime period. One key step is going to be deciding how to deal with the zero values overnight and how to manipulate the data sampling to meet the different needs of the analysis. Identifying zero values for modules/inverters during the day will be an important part of finding the equipment that is operating sub-optimally or needs to be repaired. These zero values might cause a problem during model development for forecasting, however. Depending on the model methodology, it may be necessary to resample the data at the daily or hourly level or remove the zero values from overnight to create a time-series dataset.

The reliability of optimally operating inverters should be visible in the percentage of power converted from DC to AC. It should also be visible in the daily data at the individual inverter level: those that are underperforming are producing less DC power at a given time during the day and is likely happening over a multi-day period which should allow for detection. Identifying null values in day-to-day comparisons at the plant- or inverter-level can identify faulty equipment or where weather/temperature may be affecting the generation of power. Next, I will try to model the power generation data and predict the power generation for 3 days after the end of the data. My first thought is that the train and test datasets will need to exclude any faulty days and include the longest streak of non-faulty days. I haven’t worked with a lot of time-series data, I am only tangentially familiar with regression and ARIMA models, so. I will have to do some research on the best way to deal with daily data that spikes or behaves seasonally like the power generation data does. The final goal will be to determine if one of the plants is more reliable on average, which requires more maintenance, and have a power prediction model for each power plant. Additionally, it would be ideal to determine how similar the models are or if one can reliably produce more power given the trends modeled.

The deliverables for this project will be a GitHub repository containing the work for each stage of the project, a slide deck, and a project report. This will include the supporting code for data wrangling, exploratory data analysis, modeling, and visualizations that led to the project conclusions.