**Capstone 2: Solar Power Generation – Project Report**

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. The goal of this project will be to identify equipment that is operating sub-optimally and predict the power generation over the next several days for the two solar power plants to better manage the plants and the expectations of those managing the energy grid.

Diagram

Description automatically generated

**Data Identification and Data Wrangling**

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. The first step in the data wrangling phase was importing the 4 data files into pandas dataframes. The ‘DATE\_TIME’ column and ‘PLANT\_ID’ columns are common features between the dataframes and crucial for merging the appropriate power generation and weather sensor data for the associated plant.

Table

Description automatically generated

Table

Description automatically generated

The raw state of the respective ‘DATE\_TIME’ columns were inconsistent across the dataframes, which required reformatting and creating ‘datetime’ features using pandas datetime function. A routine check for data completeness showed that there were no missing values, which could be misleading due to the large number of zero values for power generation/weather data over the nighttime hours. This was an important item to remember for the Exploratory Data Analysis portion of the project.

**Exploratory Data Analysis**

After wrangling and cleaning the data, it is imperative to understand the individual features within the datasets as well as the scale and relationships between features. During the Exploratory Data Analysis phase of the project, the power generation properties of the two plants were plotted and compared in various ways. The first step was to identify the DC power and AC power ranges across any given day to identify if there were any clearly identifiable trends or issues (Plant 1 left, Plant 2 right.

Chart, line chart

Description automatically generated

Plant 1 has a relationship that suggests the conversion from DC to AC is a heavy-loss process, while Plant 2 has DC values overlapping the AC values with only a slight difference. Since the expectation that conversion of DC power to AC is not likely 100%, an assumption was made to suggest that the trend observed in Plant 1 was more likely than the trend in Plant 2. To correct this issue, the AC power column was divided by 10 to account for a possible transposition of the decimal point which resulted in the high conversion rate of DC power to AC power. After this conversion, it is clear that Plant 1 has a much more consistent distribution and higher DC power generation during the middle of the day. Additionally, the percentage of converted power between the two plants essential overlap, suggesting that the assumption is appropriate as long as it remains consistent (Plant 1 left, Plant 2 right).

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

The difference in the two DC power distributions is an important discover because it suggests that, while both plants are converting power at roughly the same rate, Plant 1 has a much cleaner distribution. This could be due to malfunctioning equipment, weather, or other unknown factors. Understanding this is crucial to determining whether a model can reliably predict future power at both plants or if only Plant 1 can be used. One way of ascertaining this information is by plotting the DC power for each individual solar module for each plant (Plant 1 is top, Plant 2 is bottom).

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

The two module-level plots show sharply different characteristics: the modules of Plant 1 largely overlap with the exception of 2 modules, whereas the modules for Plant 2 are rather chaotic and have a wide range of peak DC power. This suggests that some, if not many, of the modules for Plant 2 are sub-optimally performing and might represent a complication for modeling and forecasting the power generation of Plant 2.

Looking deeper at the relationships between the weather data and power generation, it is important to understand some additional terminology used. One piece of information gathered from the weather sensor is the ‘irradiation’, which is the power per unit area received from the sun. Since the modules are a fixed area in size, the irradiation records how much solar power is being gathered at each 15-minute interval. The module temperature records the temperature of the solar panel modules, and the ambient temperature records the air temperature. These features have been considered relative to the converted power to identify any possible trends in the data.

Chart, scatter chart

Description automatically generated

In the plot above, irradiation is plotted against converted power as well as the average converted power line for both plants. The irradiation and converted power have a very tight relationship, there is almost no spread in the converted power at any individual irradiation. At both very low and very high levels of irradiation converted power is below average, with the optimal range between irradiation of 0.1 and 0.6 W/m^2 providing above average power conversion.

Chart, scatter chart

Description automatically generated

Plotting module temperature versus converted power shows a much less tightly-constrained distribution of points between the two plants than was present in the irradiation plot above. At low module temperatures, there is a spread of converted power percentage. At module temperatures above 40 degrees, there is a tighter distribution of converted power percentage. As it was seen in the irradiation plot, there is a range of module temperature values that produce above average converted power, approximately between 25 degrees and 50 degrees. The spread in converted power data points at temperatures below 35 degrees is not clearly explained, and may be better explained by bringing in the air temperature data.

Chart, scatter chart

Description automatically generated

The delta temperature feature is the absolute value difference between the ambient temperature and the module temperature. This feature has been plotted above against irradiation with the above-average converted power irradiation range. There is a positive correlation between these two variables, suggesting that increasing irradiation correlates to increasing delta temperature with slight heteroskedasticity. At higher irradiation, there is a higher density of points between 15 and 30 degrees of delta temperature. One practical assumption to explain the below-average converted power at higher irradiation and delta temperature is possible equipment malfunctions at higher temperatures. The module temperatures are typically higher than the ambient temperatures, suggesting that during periods of higher irradiation the modules get much warmer and perform sub-optimally, effectively overheating.

**Modeling: Training and Testing**

With a clear understanding of the variables and features within the dataset, the next step is to create a model that can predict the daily production. Given some of the issues in the data for Plant 2, specifically the issues with the DC power and individual module performance, the choice has been made to only model Plant 1 at this time. For each of the models presented below, the original dataset was subset to include only the ‘DATE\_TIME’ and ‘DAILY\_YIELD’ columns. Some key considerations in the modeling step include: model type (seasonal ARIMA, Prophet), length of training set/cross-validation, and which error values to use. All of these considerations will aid in the process to judge and select the best model for power generation prediction.

Of the two model-types, the ARIMA model is much more interpretable and gives clear information about the model parameters. The ARIMA models were selected using the auto-ARIMA function, allowing for iteration through various seasonal ARIMA parameters to minimize the Akaike Information Criterion (AIC). This process is rather time-intensive but each iteration yields information about the parameters that were used and the AIC values, including when the intercept was reached. To determine the most effective length of series to use, the full daily yield series as well as a shorter set trimmed from the end of the time series will be modeled to determine fit and error metrics.

Chart, line chart

Description automatically generated

The above plot shows the training set, test set, and auto-ARIMA model for the end of the time series training split, resulting in a SARIMAX(4, 1, 0)x(0, 1, 1, 96) model. This is a simple test on a small subset of the data, so it should be expected that the prediction closely matches the test set with minimal error. This model, while an interesting proof-of-concept, will likely struggle with generalizing to new, longer sets of data. This would suggest that while the error values are quite low for this model, it will be hard to trust the predicted values beyond the end of the set. A longer training and test set would be more valuable to developing a model with the ability to generalize.

To create a more robust model, the full daily yield series for Plant 1 was used to generate training and test sets (70:30 split). The auto-ARIMA function results in a SARIMAX(0, 1, 0)x(1, 1, 0, 96) model (below). Forecasting using this model on the test set, the SARIMAX model predicts relatively constant daily yield values for each of the 10 days. The model is sensitive to the sharp, knife-like changes in the training dataset, evidenced by this same trend in each of the forecasted daily yield peaks. This model succeeds in a way that the simple model did not: none of the predicted values go below zero, which should be ideal for a power plant that cannot generate negative power or start below zero for the daily yield.

Chart

Description automatically generated

The Prophet model is part of an open-access package released by Facebook for seasonal time-series modeling that can produced automated forecasts and can be tuned by hand easily. It is based on an ‘additive model where non-linear trends are fit with yearly, weekly, and daily seasonality’. It is capable of handling missing data, shifts in trends, deals with outliers relatively well, and runs very quickly. On the surface, there are a lot of positives to this modeling approach: there is very little pre-processing needed, the process is automated, and it is a robust modeling program. Some disadvantages of this approach are the lack of interpretability of the model parameters and the process is inherently a ‘black-box’, wherein the inner-workings are not easily visible or interpretable like the auto-ARIMA model.

Chart

Description automatically generated with medium confidence

In the plot above, the original data is represented by the black dots and the Prophet model is represented by the dark blue lines with a light blue 95% confidence interval shading. There are a number of observations that can be made from this visualization of the model. First, the Prophet forecast stays within a rather tight range for peak daily yield, which results in poor fit for some of the more extreme days. Second, the Prophet predictions for nearly half of the days dip below zero, an artifact of the modeling process but ultimately a mis-representation of the data because the power plant cannot generate negative daily yield.

**Model Metrics**

The model metrics for the 3 separate models are summarized in the image below. From a practical standpoint, all 3 models have a very high R2 score suggesting they are all good at predicting on the test set based on their respective training sets. The other two error metrics, MAE and RMSE, represent a clear separation between the 3 models. In both cases, the Prophet model has the highest MAE and RMSE and the ‘end of series’ SARIMAX model has the lowest error. The EoS SARIMAX model had a much smaller training and test set, so this could be expected. The Prophet model has a more rigorous fitting process that accounts for missing data/outliers and generated negative forecast values which could all attribute to additional error in the model that is not present in the full-series SARIMAX model. The full-series SARIMAX model generalizes well to the test set and has a MAE and RMSE between the two other models.

Text

Description automatically generated

**Forecasting Power Generation**

To determine the best model for forecast power at Plant 1, it is important to consider the priorities of the model results. All 3 models are shown below with the forecast predictions for 3 days beyond the end of the series. As stated in the training and testing section above, each model comes with benefits and challenges.

Chart, line chart, scatter chart

Description automatically generated

The EoS SARIMAX model is not an advisable model for forecasting 3 days of power but represents a portion of what the cross-validation process might look like (graphical representation below).

Background pattern

Description automatically generated with low confidence

If the priority is speed, robustness of the model, and the ability to forecast longer intervals, the Prophet model is likely the best choice. If the priority is to obtain a highly interpretable model with identifiable parameters and the purpose is to forecast for only the next few days, the SARIMAX would produce less error based on the model produced here. Both of these models would benefit from a more robust process of cross-validation, using a rolling window. In conclusion, both the Prophet model and full-series SARIMAX model are suitable for predicting the daily yield of Plant 1 for 3 days.

Chart

Description automatically generated

Chart

Description automatically generated with medium confidence

**Summary**

The data for the two power plants explored and modeled in this report represent a unique opportunity to understand how solar power is generated and how grid operators might benefit from two different modeling methods to better understand power generation capabilities. Through exploring the data, it was clear that Plant 1 had a more complete and less complicated dataset. The equipment at Plant 2, particularly the solar modules which were underperforming, could be the cause of the lower and less-consistent DC and AC power trends. From the weather sensor data, the converted power in the two plants are directly related to the irradiation and difference in the module and ambient temperatures. At higher temperatures and irradiation equipment begins to generate below-average converted power, suggesting an optimal temperature window for power generation. By testing multiple different models on training/test sets of the ‘DAILY\_YIELD’ feature in the dataset, suitable models for forecasting power generation were able to be obtained. Each model has positives and challenges, but the full-series SARIMAX and Prophet models generalized the best to the test data and produced similar forecasts for the 3 days beyond the end of the series. Further work should include a more rigorous cross-validation during model training and testing and attempting on the more-complicated dataset for Plant 2.