Project Pitch 2: The Product

Introduction

To develop our data product, we used a solar power generation dataset from two solar power plants in India (link to data: https://www.kaggle.com/anikannal/solar-power-generation-data). In short, the dataset contains IoT data from two sources. On the one hand, we have data on the power generation of the 22 groups of panels present in each plant (each group shares a single power converter). On the other hand, we have data from weather sensors containing the ambient temperature and sun irradiation at the plants as well as the temperature of an optimally placed panel. Each data point was collected during a 34-day period at 15-minute intervals for both solar power plants.

Following the Analytics Canvas model, our main objective is to use the raw data and transform it to create value to the companies that operate solar power plants. This data product will create value in two ways. First, it will enable the plant operator company to quickly detect faulty equipment and other potential problems of their operations. This allows the operator to act fast in case of any malfunction and prevents serious financial losses. Secondly, it provides the operator with predictions on generated power for the near future. This allows the management to keep solar power supplies balanced with demand and to keep power systems operating within tightly constrained limits. It also helps with such issues as grid management, operational planning, etc..

Concretely, we believe that we can achieve this by establishing the relationship between the weather situation characteristics and the power generated through the solar panels. Specifically, we analyzed the level of sun irradiation as well as ambient temperature at the plants location. In this way, we can make predictions about the power generation based on local short-term weather forecasts which contain information about temperature and solar irradiation. These are accessible for example through the API of *Solcast (https://solcast.com/solar-radiation-data/)*. It is also possible to connect to the *Solcast* API for continuously updated real-time weather forecasts which can be found on numerous websites, and we can identify possible suboptimal performing equipment when we see that a piece of equipment is significantly underperforming compared to the level of performance, we predict it should have based on the weather data.

We will use the measure “AC power generated” as the prediction target of our model, since this is the type of electricity the plant sells to its customers. Additionally, in order to compare power generation with the different weather conditions, we decided to use ‘solar irradiation’ and ‘ambient temperature’, since they seem to have a clear impact on generated power. As we only have the generation data until the 17 of June (end of the available excel file data), we will consider that at the moment of the analysis and the implementation of our solution we are at the end of the 17 of June 2020. Thus, we will be making predictions for the following two weeks based on weather data we found.

Before setting up our final data product for deployment, we created a first dashboard with the purpose of gathering insights about the performance of the two plants between May 15th and June 17th 2020. This first dashboard is a high-level analysis of both plants that reveals a significant underperformance of plant B in comparison to plant A. Despite being exposed to similar levels of solar irradiation and ambient temperature, power generation of plant B limps behind plant A throughout the observation period. Additionally, it is noteworthy that the level of generation at plant A better corresponds to sun irradiation, in contrast to the weaker relation found in plant B. Moreover, looking at the total amount of AC power generated by each solar panel group, we see that several groups generate considerately less power than other groups. On the contrary,

at plant A only two panel groups have considerably underperformed during the analyzed period. Thus, we conclude that there seem to be general structural problems at plant B that should be inspected. Further, the panels connected to the two underperforming inverters of plant A should be inspected as well.

Data Preparation

After revising the data, we identified and solved two main issues. First, the DC power data for Plant A had to be divided by a factor of 10 due to an error in the dataset. Second, the unit in which irradiation was measured was not explained. Further research on the matter did not bring clarification. As a result, we decided to standardize this measure of irradiation to have a clear indicator without scale, that can be easily understood and compared with data from other sources found online. Additionally, we merged the datasets of each plant in order to have just one dataset for weather and one for generation. Finally, we applied some stylistic corrections on some data features in order to facilitate the connection of both datasets on Tableau and to make it more understandable for the audience. After the initial cleaning and preparation, the data was eventually loaded to Tableau.

Prediction Model

Based on the available data, we create regression models of the produced energy based on two parameters: the irradiation level and the ambient temperature. After analyzing the different regression models (one based on all the available data, one based on the plant A data only, and one on plant B data), we decided to use the regression model based on plant A data due to the evident presence of numerous failures of the material of plant B. The regression model based on plant A data has a R-Square of 98% and its two causal variables are statistically significant. This indicating that our model correctly predicts the energy generation based on the ambient temperature and irradiation level. Among the factors, the most important predictor is the irradiation level. *(For further details, you can look at the excel file with the different regressions.)*

To identify the faulty equipment, we calculated a field called the “Forecasted Ac Power” that represents the amount of energy that a fully functioning panel should be generating. To calculate it, we use the real time weather data available at the moment of the production time. Then, we calculated the “Over-performance” measure that represents the difference between the actual energy production output per panel (“Ac Power”) and the forecasted output (“Forecasted Ac Power”). This measure helps identifying anomalies in the equipment.

In order to create the energy production forecast for the following two weeks, we collected weather data for each location from the *Solcast* API and introduce them into our regression model to estimate the production in the short term future for both locations.

Detecting Defect Equipment

The following dashboards show how the product looks like when deployed. If connected to generation data of a solar power plant, we can detect problems in real time.
The top graph of “Power Generation Analysis” displays the difference between predicted energy of each group of panels and the generated level of generation, given the current weather conditions. When our product is connected to generation data of a power plant, we can see the discrepancies in generated energy and predicted energy in real time. Given, that we have only limited data, the current dashboard displays the power plants at midnight, 17th of June 2020.
In addition, the graphs below compare generated power to predicted power as well as their difference on a daily basis over the past month.

Forecasting

Besides detecting faulty equipment, we can use the model to predict the amount of generated power in the near future. The top graph of the "Model Insights" dashboard show the historical and expected power generation at both plants given the current weather forecast of each location. It illustrates how each plant should be performing if working as expected. Since the data we worked with reaches up to the first half of June 2020, we selected the period from June 17th to July 3rd as forecast period. When comparing the forecasted generation to the historical generation for plant B, we notice that the forecast is significantly higher. The reason behind this discrepancy is that the forecast predicts what the generation would be if the inverters are working normally. However, as we discussed previously, plant B contains several systematically underperforming inverters.

Financial Impact of our Solution

Furthermore, this product also allows the company to know the impact of the faulty equipment in monetary terms. We estimated this impact by multiplying the underperformance of energy generation and the average price of electricity supplied in India, which equals 0.069 euros per kWh (March 2020 data). Additionally, we calculated the total estimated revenue using this price to understand the magnitude of the money lost. After doing both calculations, we found that the estimated monetary loss per month with our sample data is 295,914 euros and the total monthly estimated revenue is 2,587,782 euros.

Considering that our model can contain prediction errors, we correct for a potential error of 100 Ac power in the financial analysis. Therefore, all the negative values greater than -100 will be increased to 0. Also, values smaller than -100 will be increased by 100. For instance, an underperformance of -75 will be corrected to 0, while an underperformance of -120 will be corrected to -20. This decision is based on the standard errors value of our prediction model that is 57 Ac Power. Thus, an anomaly will only be considered given a difference higher than 100. This consideration enables us to keep a conservative estimate of the financial impact of our solution.