AASD 4016 Full Stack Data Science Systems.

Students: Diana C. Lopera ID 101495151, Zarina Dossayeva ID 101385370, Juan M. Henao ID 101488061,

Project: Solar Power Forecasting

Introduction

For the project of this subject, we approached the solar power forecasting problem. The scope of this problem involves being able to predict the amount of power that is going to be generated by a solar photovoltaic plant in the next couple of days. The solar power forecasting can be divided into "nowcasting" for prediction of some minutes and a couple hours, short-term predictions for the following hours and days and Long-term for predictions greater than a week. This problem is very significant because it allows both great and small photovoltaic power plants to schedule ahead and program the connections and disconnections to the electric grid, and also manage efficiently the maintenance of the power plant, this is crucial since these kinds of installations have very slow ROI (usually 15 – 30 years).

Data

The dataset used for this project was found on Kaggle https://www.kaggle.com/datasets/anikannal/solar-power-generation-data. This dataset contains observations about AC, DC, Irradiation, and temperature taken for 2 different power plants every 15 minutes for 34 days. This data set was used for the initial development and testing of our product.

Past Projects

Regarding the state of the art in this topic we found several points that are worth discussing. First, there are already some companies that offer some kind of service like the one developed on this project, for example AESO https://www.aeso.ca/aeso/ which is a Canadian company operating in North America. Also there are several paths to come up for a solution for the power forecasting, classic statistical models like ARIMA regression, monitoring by satellite and of course machine learning algorithm, so why do this if there are already persons in the market? simple because the biggest market for this is yet to be explored, countries that have a high solar photovoltaic potential like Colombia, Ecuador, and Brazil are still lacking solutions like the one we are proposing that can be easily implemented.

ML – Canvas:

Our creative and design process was highly influenced by the tool of ML-Canvas. The complete image of our ML-Canvas can be found at the end of this document.

Model Benchmark:

To develop our model, we trained 3 different models. 2 of them with the main purpose of benchmarking and a final one to be deployed on our MVP. The first model was a simple Artificial Neural Network, afterwards we moved on to the initial development of an LSTM model, and we further fine tuned it into a second version of LSTM model. In the end we tested 3 different models that improved performance for each iteration. It's worth to mention that ARIMA regression models are also very famous regarding this topic and are widely used, they could also be a very good point of comparison for benchmarking, a table

with a summary of the errors related to each of the models is presented below. Minimum value and Maximum value in the ranges of the predicted value are displayed for comparison.

	ANN	LSTM	LSTM_2	Min_Value	Max_Value
Average Error [W]	83,026.00	31,513.00	19,381.00	0.00	200,000.00

Table 1. Error for different models

Model Deployment:

Model deployment was executed using Flask, Docker and GCP. The initial notebook was turned into a python script and then turned into an API using Flask. Finally, the product was deployed on Google Cloud Platform. Although the product works as expected there are a couple of things that we should mention regarding the challenges that can be found for this kind of product. First, the product is prone to give wrong predictions when weather behaves unexpectedly, this problem can be addressed by correctly scheduling a re training action on the API of our product. Also, the product currently does not consider how drastically the efficiency of solar power cells drops when maintenance is needed. It's very well known that cleaning is a main characteristic that affects the efficiency of solar power cells, this aspect can lead to wrong predictions even if all the other variables behave as expected. In an ideal world also, we would be able to include more metrics that would enrich our feature space allowing the model to have an even better performance since during training it was noticeable how the model is hungry for data.

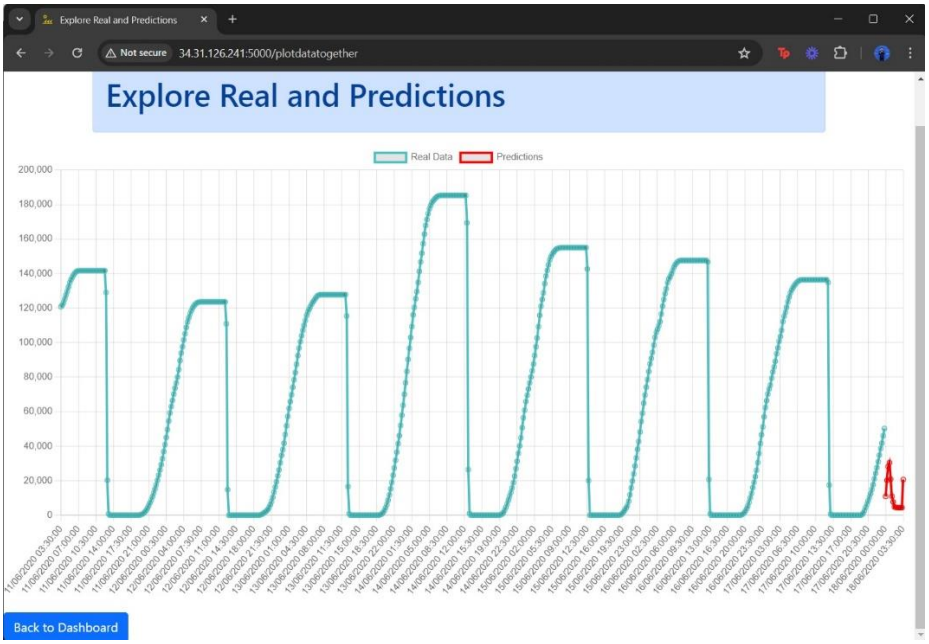

Figure 1. Historic data and predictions on our App

Conclusions:

The main conclusion of our product is how easy it is to implement and the great benefits that it can draw out of the solar power plants, specially helping maintain a good health of the hardware along time which is a crucial characteristic of solar power installations due to their long ROI. In future versions, automatic re training, and recommendations regarding actions that can be taken to prevent damage to the solar power plants can be included to the product.

Annex:

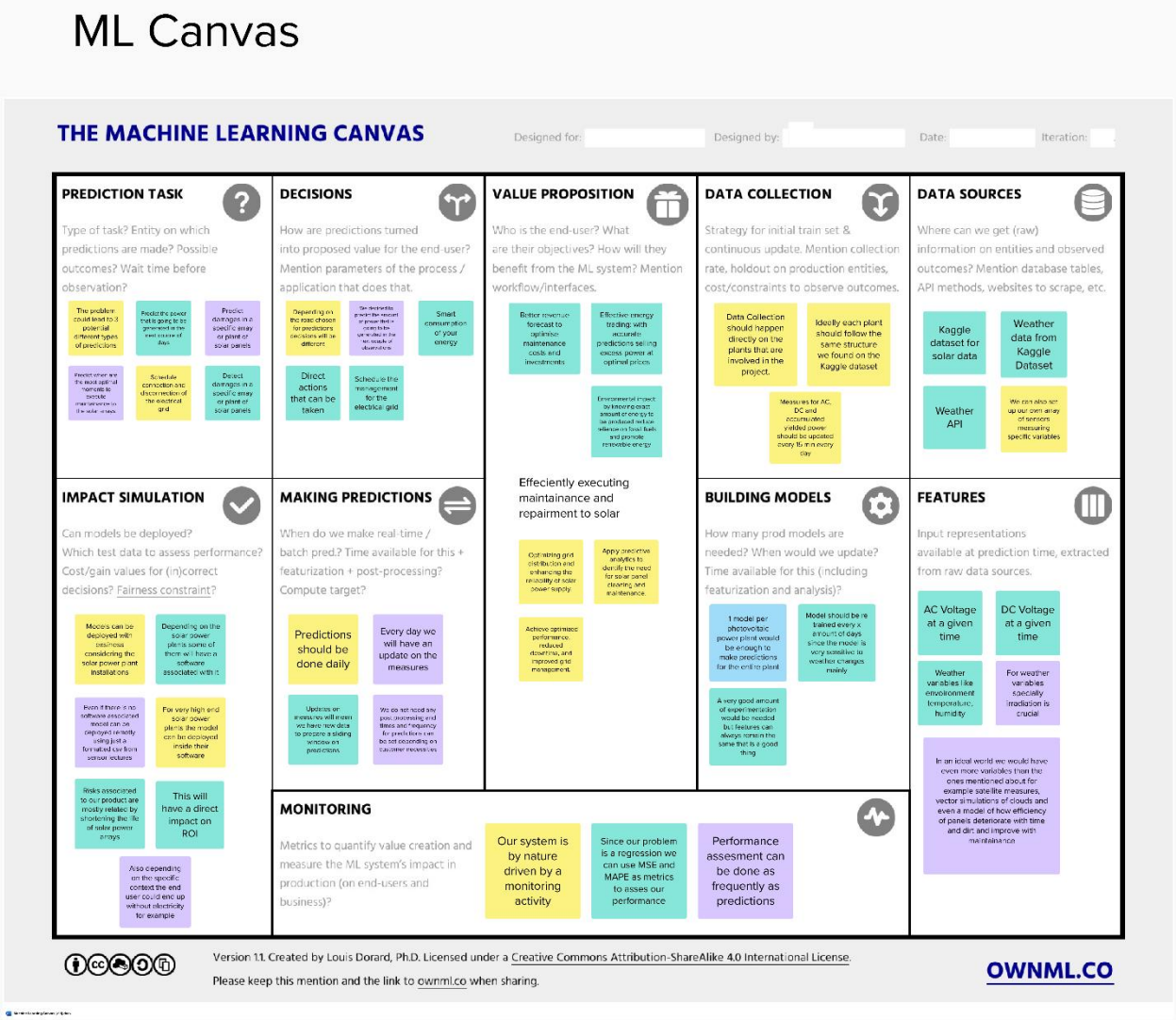

Figure 2. ML-Canvas

Links:

Visit our app: http://34.171.112.217:5000/

Visit our Github repository: https://github.com/catalinalopera/PowerForecasting/tree/main

ML Canvas

THE MACHINE LEARNING CANVAS

Designed for: Designed by: Date: Iteration:

PREDICTION TASK

Type of task? Entity on which predictions are made? Possible outcomes? Wait time before observation?

- The problem could lead to 3 potential different types of predictions
- Predict the power that is going to be generated in the next couple of days
- Predict damages in a specific array or plant of solar panels
- Predict when are the most optimal moments to execute maintainance to the solar arrays
- Schedule connection and disconnection of the electrical grid
- Detect damages in a specific array or plant of solar panels

DECISIONS

How are predictions turned into proposed value for the end-user? Mention parameters of the process / application that does that.

- Depending on the road chosen for predictions decisions will be different
- We decided to predict the amount of power that is going to be generated in the next couple of observations
- Smart consumption of your energy
- Direct actions that can be taken
- Schedule the management for the electrical grid

VALUE PROPOSITION

Who is the end-user? What are their objectives? How will they benefit from the ML system? Mention workflow/interfaces.

- Better revenue forecast to optimise maintenance costs and investments
- Effective energy trading: with accurate predictions selling excess power at optimal prices
- Environmental impact: by knowing exact amount of energy to be produced reduce relience on fossil fuels and promote renewable energy

Effeciently executing maintainance and repairment to solar

- Optimizing grid distribution and enhancing the reliability of solar power supply.
- Apply predictive analytics to identify the need for solar panel cleaning and maintenance.
- Achieve optimized performance, reduced downtime, and improved grid management.

DATA COLLECTION

Strategy for initial train set & continuous update. Mention collection rate, holdout on production entities, cost/constraints to observe outcomes.

- Data Collection should happen directly on the plants that are involved in the project.
- Ideally each plant should follow the same structure we found on the Kaggle dataset
- Measures for AC, DC and accumulated yielded power should be updated every 15 min every day

DATA SOURCES

Where can we get (raw) information on entities and observed outcomes? Mention database tables, API methods, websites to scrape, etc.

- Kaggle dataset for solar data
- Weather data from Kaggle Dataset
- Weather API
- We can also set up our own array of sensors measuring specific variables

IMPACT SIMULATION

Can models be deployed? Which test data to assess performance? Cost/gain values for (in)correct decisions? Fairness constraint?

- Models can be deployed with easiness considering the solar power plant installations
- Depending on the solar power plants some of them will have a software associated with it
- Even if there is no software associated model can be deployed remotly using just a formatted csv from sensor lectures
- For very high end solar power plants the model can be deployed inside their software
- Risks associated to our product are mostly related by shortening the life of solar power arrays
- This will have a direct impact on ROI
- Also depending on the specific context the end user could end up without electricity for example

MAKING PREDICTIONS

When do we make real-time / batch pred.? Time available for this + featurization + post-processing? Compute target?

- Predictions should be done daily
- Every day we will have an update on the measures
- Updates on measures will mean we have new data to prepare a sliding window on predictions
- We do not need any post processing and times and frequency for predictions can be set depending on customer necessities

BUILDING MODELS

How many prod models are needed? When would we update? Time available for this (including featurization and analysis)?

- 1 model per photovoltaic power plant would be enough to make predictions for the entire plant
- Model should be re trained every x amount of days since the model is very sensitive to weather changes mainly
- A very good amount of experimentation would be needed but features can always remain the same that is a good thing

FEATURES

Input representations available at prediction time, extracted from raw data sources.

- AC Voltage at a given time
- DC Voltage at a given time
- Weather variables like envoironment temperature, humidity
- For weather variables specially irradiation is crucial
- In an ideal world we would have even more variables than the ones mentioned about for example satellite measures, vector simulations of clouds and even a model of how efficiency of panels deteriorate with time and dirt and improve with maintainance

MONITORING

Metrics to quantify value creation and measure the ML system's impact in production (on end-users and business)?

- Our system is by nature driven by a monitoring activity
- Since our problem is a regression we can use MSE and MAPE as metrics to asses our performance
- Performance assesment can be done as frequently as predictions

OWNML.CO