Final Capstone Project Proposal
Solar Energy Anomaly Detection and Forecasting

Ruth Caswell Smith

January 29, 2021

Problem Statement

Can we predict future solar energy output from a solar array based historical data using time series forecasting techniques? Can we identify anomalies in production and develop a model with predictive power based on historical data and exogenous variables?

Context

Solar energy is radiant light and heat from the Sun that is harnessed using a range of ever-evolving technologies such as solar heating, photovoltaics, solar thermal energy, solar architecture, molten salt power plants and artificial photosynthesis.

It is an essential source of renewable energy, and its technologies are broadly characterized as either passive solar or active solar depending on how they capture and distribute solar energy or convert it into solar power. Active solar techniques include the use of large ground-based photovoltaic systems such as the one used for this project.

The large magnitude of solar energy available makes it a highly appealing source of electricity. The United Nations Development Programme in its 2000 World Energy Assessment found that the annual potential of solar energy was several times larger than the total world energy consumption[1].

Being able to accurately detect (and/or predict) anomalies as well as forecast energy output is an essential tool that can help solar installations operate effectively and return on capital investment.

Criteria for Success

A predictive model will be developed to accurately forecast energy output based on historical patterns. In addition, anomalies in production will be classified and a model will be developed with the aim of predicting outages of the solar inverters.

Scope of Solution Space

This forecast will be developed for one particular solar installation but could be applied to other solar installations as well.

Constraints

This is real-world data, and so there is a lot of missing data as well as periods of bad data. In addition, there is not a lot of information on outages, and so outages will need to be classified based on the data that is available.

Stakeholders

Operators or owners of large ground-based solar installations.

---

[1] Wikipedia

Data Sources
The data is from a 1MWatt installation of 5000 panels in the mid-Atlantic region. Data is available for six years at 15-minute increments, with each year being contained in a single CSV file.

The dataset consists of the following attributes:

- Timestamp
- Shark Meter, KWtotal Kilowatts
- AE 500kW 1, AC Power Kilowatts
- AE 500kW 2, AC Power Kilowatts
- (Offline) Weather Station - POA (POA)* Watts/meter²
- Weather Station (POA) (SO31456) (POA)* Watts/meter²
- RECx31 Weather Station, Module Temp Degrees Celsius
- RECx31 Weather Station, Ambient Temp Degrees Celsius
- Weather Station (POA) (SO31456), CabF Degrees Celsius
- (Offline) Weather Station - POA, CabF Degrees Celsius
- AE 500kW 1, PV current Amps
- AE 500kW 2, PV current Amps
- AE 500kW 1, PV voltage Volts
- AE 500kW 2, PV voltage Volts

Approach
Data wrangling will include a significant effort to deal with missing data. It is not immediately evident whether zeros indicate actual zero values or missing values, and so analysis will need to be performed to distinguish between the two cases. In addition, there are large sections of missing data.

Exploratory data analysis will develop an understanding of the correlation between the various input variables and the target variables, which is the power output of the two inverters. In addition, additional features based on historical data as well as calculated variables will be added. Seasonal decomposition will be performed to see trend and seasonality.

Anomaly detection will be challenging. This is because of the need to distinguish between real zero values, outliers, and missing data. We will attempt to classify anomalies through clustering techniques and statistical techniques.

A predictive model will be developed to see if patterns exist that can predict anomalies. Features for this will include historical data as well as exogenous variables such as irradiance and temperature.

Forecasting will be done using SARIMAX and FB Prophet using solar irradiance and temperatures as exogenous variables. Lastly, all code for this project will be written to be maintainable, taking advantage of best practices for data science.

Deliverables

Deliverables include python scripts and a Jupyter notebook, a project report, and a slide deck. A presentation of this capstone project will be given.