Springboard Capstone Project 1 – Project Proposal

Predicting Short Term Solar Energy Production


[6]

Connor McAnuff

June 3, 2019

1. Problem Description and Value to Client

1.1 Solar energy generation

Global solar energy generation capacity has been increasing exponentially since 2006 and is forecast to continue along the same trend through 2021 [1]. A form of solar energy generation is photovoltaic (PV) panels, which produce electricity through interaction with photons from the sun [2].

1.2 Predicting solar energy generation

The solar energy availability at a given location can be described by the level of solar irradiance (generally stated in W/m^2) and can be measured using a pyranometer [3]. Solar irradiance is used as a measure of the power available to enter a solar PV system. The system configuration and power losses through the PV system are well known, and thus if the solar irradiance is known, solar energy production can be accurately predicted. Thus, the short term solar energy production of a PV array can be predicted by proxy through the prediction of daily incoming solar energy (the sum of the solar irradiance throughout the day). Solar irradiance at the Earth's surface is in part determined by weather conditions, as cloud coverage and precipitation obscure the suns' rays from reaching the Earth's surface through reflection and refraction [4].

1.3 Value to client

Solar energy generation presents a unique challenge for electric utility companies as the energy generation varies in part depending on weather conditions. Utilities companies must be able to accurately forecast electricity production to prevent energy shortages and surpluses. Energy shortages can result in costly emergency purchases from neighbouring utility companies or blackouts, while energy surpluses can result in wasted energy as electricity cannot currently be feasibly stored in large amounts. Forecasting solar energy generation is a key component in several grid-balancing decisions such as reserve activation, short-term power trading (with other utility companies), peak load matching, and congestion management [4].

Accurately predicting short term solar energy production across many "solar farms" would provide value to utility companies by providing information for better grid-balancing decisions, resulting in a more efficient operation and reduced costs.

2. Dataset Description

2.1 Overview

The dataset to be used in this project has been made available for a former Kaggle competition that occurred in 2013 (https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/data). The training dataset includes data spanning from 1994-2007, while the testing dataset includes data spanning from 2008-2009. The testing dataset does not give the target variable values explicitly (testing for the Kaggle contest involves submitting predictions and receiving a single-value score only) so the training dataset will need to be segmented for use as the training and testing dataset for the purposes of this project.

Weather condition predictions are given at 144 evenly spread grid point locations surrounding and covering Oklahoma, and the "solar farm" locations are given as 98 Mesonet sites spread unevenly across Oklahoma. Locations of the weather condition prediction grid points and solar farms are shown in Figure 1.

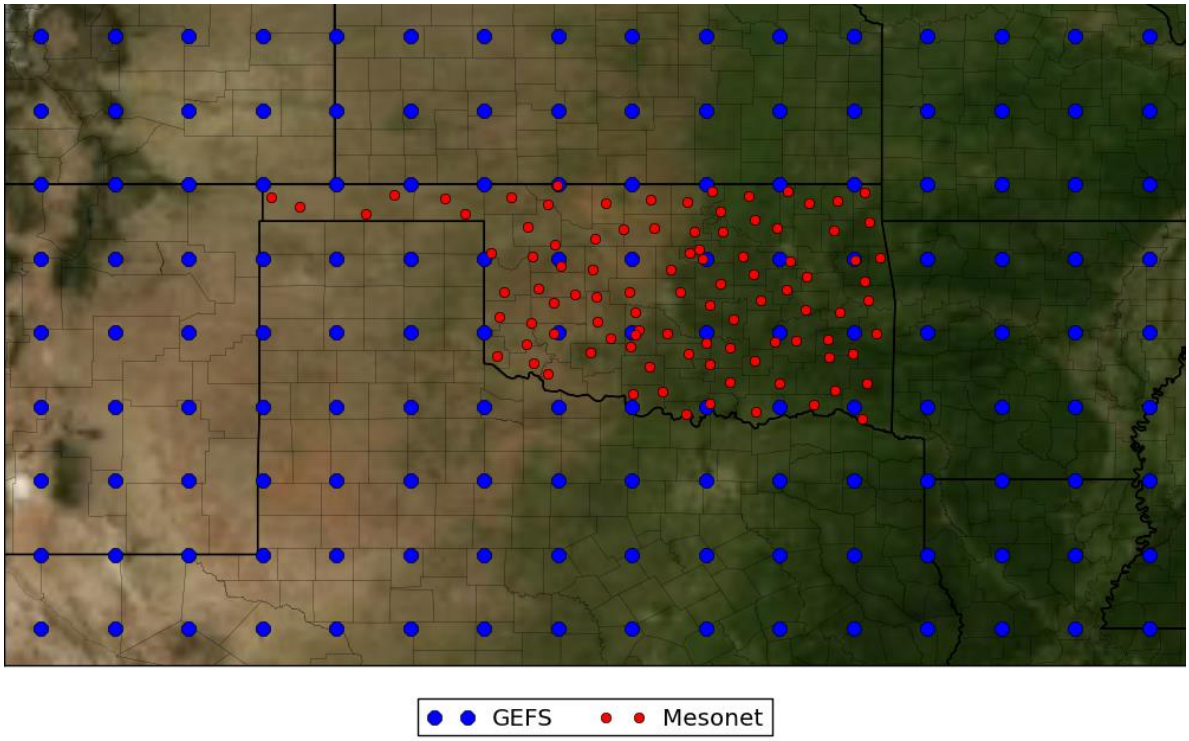

Figure 1: GEFS prediction grid points (blue) and Mesonet "solar farm" (red) locations.

2.2 Weather condition variable predictions

Weather condition data are provided as predictions from the NOAA/ESRL Global Ensemble Forecast System (GEFS). The predictions for each variable are given as 11 ensemble members, meaning there are 11 separate models producing 11 separate predictions for each variable at each timestamp for each forecast day. The 11 prediction models are used to account for the two usual sources of uncertainty in weather forecast models: (1) imperfect initial conditions, and (2) imperfections in the model formulation [5]. There are 15 weather condition variables. For each forecast day predictions are given at 5 timestamps at 3 hour intervals.

Weather prediction data are in netCDF4 files separated by variable. For each variable, the training dataset has 5 dimensions of size (5113, 11, 5, 9, 16), corresponding to:

- 5113 forecast days
- 11 prediction models
- 5 timestamps (at 3-hour intervals)
- 9 latitudes
- 16 longitudes

2.3 Incoming solar energy

The total daily incoming solar energy data provided has been directly measured using a pyranometer every 5 minutes and summed from sunrise to 23:55 UTC of each date at 98 Mesonet solar farms in Oklahoma. Solar energy data are in csv file format. The latitude, longitude, and elevation of each solar farm are also provided in the form of a csv file.

3. Proposed Solution Methodology

3.1 Machine learning model overview

- **Target variable:** Total daily incoming solar energy at 98 solar farm locations.
- **Features:** 15 weather condition prediction variables of 11 prediction models at 5 timestamps for each forecast day.
- **Model type:** Regression.

3.2 Data wrangling

- Target variable data imported from csv files into Pandas Dataframe and/or Python dictionary.
- Feature data imported in netCFD4 format.
- Split training data into training and testing data.

3.3 Exploratory data analysis

- Explore relationships between features and target variable.
- Explore statistics for the daily incoming solar energy at each location (e.g. mean, median, variance).
- Explore statistics for the 11 prediction models (variance likely the most important).

3.4 Use of GEFS predictions

- It will need to be determined how to make use of the GEFS prediction grid points according to their locations relative to the solar farms.
    - Likely a form of distance-weighted interpolation using x number of closest grid points.
- It will need to be determined how to make use of the 11 weather condition predictions resulting from the 11 prediction models.
    - A simple approach would be the median value.

4. Deliverables

- **Jupyter Notebook:** Annotated code with figures and short explanations for the data processing, exploratory data analysis, machine learning implementation, and results.
- **Technical Report:** In depth explanation and discussion of dataset, problem solving methodology, results, etc.
- **Slide deck:** High level overview of project, mostly non-technical.

5. References

[1] Forecast International - Powerweb , "Renewable Energy," 2016. [Online]. Available: http://www.fi-powerweb.com/Renewable-Energy.html.

[2] Live Science, "How Do Solar Panels Work?," Live Science, 2017. [Online]. Available: https://www.livescience.com/41995-how-do-solar-panels-work.html. [Accessed 2019].

[3] Hukseflux, "Pryanometers," Hukseflux, 2019. [Online]. Available: https://www.hukseflux.com/products/solar-radiation-sensors/pyranometers. [Accessed 2019].

[4] J. Lago, "Forecasting in the Electrical Grid," Incite, 2018. [Online]. Available: http://www.incite-itn.eu/blog/forecasting-in-the-electrical-grid/. [Accessed 2019].

[5] J. Slingo and T. Palmer, "Uncertainty in weather and climate prediction," *Philos Trans A Math Phys Eng Sci,* vol. 369, no. 1956, p. 4751–4767, 2011.

[6] http://news.mit.edu/2016/mit-neutralize-17-percent-carbon-emissions-through-purchase-solar-energy-1019, "MIT to neutralize 17 percent of carbon emissions through purchase of solar energy," MIT News, 2016. [Online]. Available: http://news.mit.edu/2016/mit-neutralize-17-percent-carbon-emissions-through-purchase-solar-energy-1019. [Accessed 2019].

6. Additional Resources

- https://bit.ly/2VJKlut: Kaggle competition winner GitHub repository.
- https://bit.ly/2HFg3Bm: Medium blog post on a similar project with GitHub / Jupyter notebook links.
- https://bit.ly/2HABi54: Resource for handling NetCDF files in Python.