Springboard Capstone Project 1 – Data Wrangling

Predicting Short Term Solar Energy Production

Connor McAnuff

July 5, 2019

1. Data Overview

This report outlines the process to translate the raw data into a form suitable for applying a machine learning model. The raw data has been provided in the following format (raw data source):

- station_info.csv:
  - Array of station IDs, latitude, longitude, and elevation (98 rows x 4 columns).
- train.csv:
  - Array of dates and the recorded daily available solar energy at each of the 98 Mesonet Solar Farms from 1994-01-01 to 2007-12-31 (5113 rows x 98 columns).
- Weather Variable Forecasts:
  - 15 NETCDF4 files (one file for each weather variable) listing the variable forecast value for each of the 11 predictive models, 5 forecast hours, 9 latitudes, and 16 longitudes for each of the 5113 forecast days (dimensions 11, 5, 9, 16, 5113).

2. Importing

Stations_info.csv and train.csv were imported directly into Pandas DataFrames named stations and energy respectively. The 15 weather variable forecast files were located using glob and the data were imported into a list of data using xarray. Next, the list of data was converted into a list of DataFrames.

3. Cleaning and Organization

3.1 Missing Values

There are no null values in the energy, station, and weather variable data. Null value checks were performed using isnull().

3.2 Outliers

The client stated that the pyranometers (sensors measuring solar energy availability) occasionally ceased functioning correctly. The client filled in the missing values with fictional values. Using value_counts(), it was determined that these fictional values end in non-zero numbers whereas the remaining (true) values end with zero. Figure 1 shows the 1999 solar energy availability for ACME and CLAY stations. From May-July, CLAY station shows a constant energy value of 12320768 J/m^2.

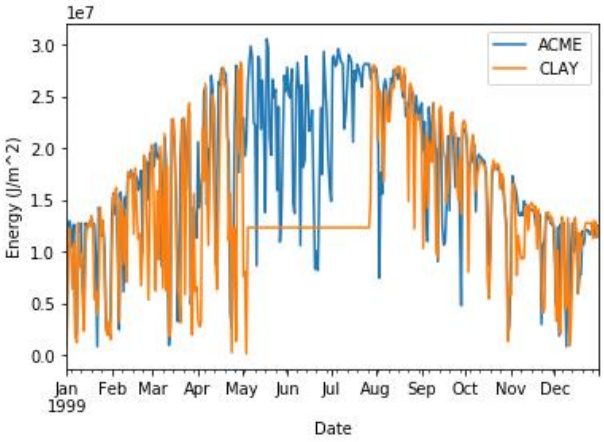

Figure 1: Daily available solar energy at two stations for the year 1999. CLAY has fictional data for May-July.

The fictional values were removed from the dataset after data formatting. They comprise 0.42% of the overall data.

3.3 Formatting

The goal of data formatting was to create a list of observations that include the date, station, available solar energy (target variable), and the machine learning features for that day and station. The process utilizes a nested for loop, iterating through each station, and for each station iterating through each weather variable. The steps are as follows:

1) Merge the list of energy data for a specific station with the list of stations.
2) Determine the closest weather forecast gridpoint (using longitude and latitude) to the station.
3) Get the weather variable forecasts for all dates, forecast hours, gridpoints, and 11 predictive models.
4) Use only the weather variable forecasts from the latitude and longitude of the closest gridpoint to the station – this process may be refined later.
5) Take the median value of the 11 different predictive models as a single weather variable forecast.
6) Pivot the forecast hour to be 5 different columns for each of the 5 forecast hours (and therefore 5 different features).
7) Merge the weather variable predictions/forecasts for each date with the total energy availability and add to the final DataFrame.

Additionally, year, month, and day were added as features.

4. Jupyter Notebook

Github Link: https://github.com/connormca12/Springboard-Projects/blob/master/Capstone-1/data_wrangling.ipynb