Analysis and Cost of Solar Power over Time

Data Wrangling Report

Springboard, DSCT

17Nov2017

Mark Sausville
saus@ieee.org

Data Wrangling Report on Analysis of Solar Installation over Time

Overview

In this project, we will analyze a large dataset characterizing over 1 million solar power installations, targeting the cost over time of the roof mounted, residential installation to support the residential solar customer by predicting the future cost of a solar installation.

This report describes the data cleaning and reformatting prior to analysis and modeling.

The Data Set

The project data were acquired from the National Renewable Energy Laboratory, at the Open PV Project. Several datasets are available as well as resources for downloading subsets of the full dataset by date, size and geography. The NREL put this data together from many sources including rebate program administrators, energy companies and volunteers. As such, it is to be expected that quality will vary.

NREL makes available the full dataset and a 'cleaned' dataset with some documentation. The initial notion was to take advantage of the cleaned dataset and utilize some of the documented data fields as modeling parameters.

Unfortunately, some potentially very useful documented fields are not actually present in the data. When contacted about this discrepancy, project personnel responded that the quality of these fields was very inconsistent and that they had made the decision to remove them from the public dataset.

While the lack of these parameters limits some modeling choices, we can still model with the most important variables.

We chose to work with the full dataset as potentially the most interesting.

The most important variables are the cost/watt and date of the installation. Other features (geography, installer, 3rd party owned, etc.) may be useful in cost modeling. Installation type (i.e. Residential, Commercial, etc.) is also significant because the study is restricted to the residential market.

Initial Findings

The full dataset contains information on over 1 million solar installations, with 82 columns. Only a relatively small set of columns appears to be populated and useful.

In the first pass, 42 completely null columns were discarded.

The data were indexed by the ***date_installed*** field (one of the few fully populated columns). During indexing, a small number of null values (less than 10) were discovered in this column and eliminated. These appear to have been introduced by data entry errors.

In a second pass, we evaluated all the remaining columns for utility and integrity. At this stage, 13 columns were identified as definitely pertinent to further investigation. These are summarized in the following table.

Time
```
date_installed    1020516 non-null object
```

Cost
```
size_kw           1020516 non-null float64
cost_per_watt     762941 non-null float64
cost              763102 non-null float64
```

Geography
```
state             1020521 non-null object
zipcode           1020516 non-null float64
city              798954 non-null object
county            998652 non-null object
```

Installation Features
```
install_type      977940 non-null object
installer         702466 non-null object
new_constr        27106 non-null float64
tracking          1930 non-null float64
3rdparty          306993 non-null float64
```

The time, cost and geography fields are all relevant to the study. The ***install_type*** variable is vital because it indicates whether the installation is residential, the main thrust of the investigation. The other installation features listed above may provide usable information.

The other features not described above have not been eliminated because it may be interesting to examine some of them during the EDA phase. The disadvantage of aggressively removing data is that some interesting stories in the data might be lost. Prior to modeling, we will be more stringent about restricting the data to PV installations that are most pertinent to the overall modeling goal.

Manipulation and Cleaning

Since we are particularly interested in time, we transformed the data into a time-series by parsing the ***install_date*** field and reindexing the data. This facilitates selection by date range in the analysis stage.

The ***install_type*** field was reworked for spelling and capitalization.

Zipcode was coerced to five character string format and invalid entries were flagged.

State was corrected to contain unique two character state codes.

Several fields were identified as boolean or possible categorical variables.

The 96% of the data is contained in the period from 2006 through 2015. In the next stage we will focus on these installations because they provide a strong basis for modeling.

Status of Reformatted and Cleaned Dataset

At this stage the reduced and clean dataset consists of 1M rows of 38 columns, of which 13 have been canonicalized to facilitate analysis. The information below summarizes the structure of the data.

Time

```
DatetimeIndex: 1020521 entries, 1909-07-07 to 2017-11-25
```

Size and Cost

```
cost_per_watt                762941 non-null float64
cost                         763102 non-null float64
size_kw                      1020516 non-null float64
```

Geography

```
zipcode                      1020516 non-null float64
state                        1020521 non-null object
city                         798954 non-null object
county                       998652 non-null object
```

Installation Features

```
new_constr                   27106 non-null float64
tracking                     1930 non-null float64
3rdparty                     306993 non-null float64
install_type                 977940 non-null object
installer                    702466 non-null object
```

Other Features

appraised	224036 non-null object
incentive_prog_names	797958 non-null object
type	1020516 non-null object
lbnl_tts_version_year	797958 non-null float64
lbnl_tts	797958 non-null object
utility_clean	792720 non-null object
tech_1	580919 non-null object
model1_clean	580919 non-null object
annual_PV_prod	780969 non-null float64
annual_insolation	780969 non-null float64
rebate	386698 non-null object
sales_tax_cost	355309 non-null float64
tilt1	383365 non-null float64
tracking_type	526058 non-null object
azimuth1	363281 non-null float64
manuf2_clean	231607 non-null object
manuf3_clean	209653 non-null object
manuf1_clean	201121 non-null object
inv_man_clean	49933 non-null object
reported_annual_energy_prod	204429 non-null float64
year	68 non-null object
pbi_length	5427 non-null float64
utility	2117 non-null object
bipv_3	5255 non-null float64
bipv_2	5255 non-null float64
bipv_1	5255 non-null float64