Case study: Detection of correct circuit definition and polarity

# Background

As you probably know already, Solar Analytics provides monitoring services for residential/commercial customers (usually with solar photovoltaics (PV) systems). One problem is that sometimes the circuits of the monitoring devices are miss-labelled, or the circuit polarity is wrong when our devices are first installed. Hence, we display PV data as load consumption data, or the load consumption data displays as negative due to wrong polarity.

We are hoping to build a machine-learning algorithm to detect and correct these circuits automatically over the first few days of the site being registered and the device being online. The metrology from these devices is reported at 5 minute intervals and the intent is to identify which circuits may need to be modified as soon as possible after the site is viewable by the customer. This means there is limited metrology to inform a decision (days, not weeks). Some work is done to label some circuits manually, and we have added some features which may help with the detection.

# Tasks

The dataset is provided, and the column information is mentioned in the section below.

Please note that this task is **not to build the model**. We are interested in how you would attempt the problem.

There are a few questions we would like you to answer for this case study:

* If you were to build an offline version of this algorithm, what steps would you take, and what types of algorithms do you prefer?
* We have noticed there is some missing data in the dataset. How would you deal with it?
* Let’s say you have built the best possible model using the provided dataset, what could we do to further improve the model?
* How would you monitor the performance of the model after being deployed in a production state?

# Dataset information

Each customer owns at least one site and each is assigned a unique site ID. A site could have multiple monitored circuits, each circuit is represented by a unique circuit id.

Here are the column definitions for the provided CSV:

|  |  |
| --- | --- |
| Column | Definition |
| site\_ac\_capacity | AC capacity of a site (kW) |
| num\_cons\_circ | Number of load consumption circuits of a site |
| num\_pv\_circ | Number of PV circuits of a site |
| energy\_hour(0-23) | Hourly energy of a day for that circuit (date is provided in the date column) |
| site\_id | site id |
| c\_id | circuit id |
| date | Date when the hourly data was pulled |
| Label (0-7) (definition and polarity): | Data labels, representing the ground truth.  label 0: sub-load; positive  label 1: sub-load; negative  label 2: PV; positive  label 3: PV; negative  label 4: gross load; positive  label 5: gross load; negative  label 6: net load; positive  label 7: net load; negative |