Device Failure Predict

Index

[Data Cleaning 1](#_Toc5567108)

[Data Exploration 1](#_Toc5567109)

[Data Cleaning 1](#_Toc5567110)

# Data Cleaning

## Data Exploration

The dataset is composed by 124494 samples and 12 columns: 9 columns are actually features, named attribute1, attribute2,…attribute9, one is device id and one is the date of the sample acquisition.

The objective column is named failure.

Overall, we have 1168 unique devices.

At the first glance over the dataset appears immediately that the features attribute7 and attribute8 have a similar distribution of value. With a more in dept view we have discovered that they are actually identical.

Also, it’s quite clear that some columns (attribute 1,5 and 6) have an even distribution of value between the max and min value, while the other features are mostly zeros with few very high values.

Moving from the features to the target column we have discovered two aspects:

* The target variable is unbalanced: over 1168 only 106 have had a failure
* No device has experienced more than one failure

Looking at the date we can also see that the failure is relative ad is the last datapoint collected.

As last point, we have observed that often a failure is associated with a rise of the values of some features in the previous days.

## Data Cleaning

Following the consideration above we have done the following operation over the dataset:

* Attribute7 it’ been removed
* Attribute 1, 5 and 6 are been normalized, while we have kept the others original
* Considering the full dataset as a unique time series, we have calculated the difference of each column’s values with the value at day-1, day-2 and day-3. Those values are been added to the dataset in specific columns named ‘\*\_lag1’, ‘\*\_lag2’ and ‘\*\_lag3’, where \* indicates a specific attribute.