A predicting model for correcting fault in billing readings

Paulo Benatti Alves[1], Olga Satomi Yoshida[1]
[1] Instituto de Pesquisas Tecnológicas do Estado de São Paulo – IPT

Abbreviated abstract: A model is proposed to predict the use of natural gas in three distinct commercial establishments. This model solves the problem of predicting the bill for collection in case of data loss, fraud, lack of in person or remote reading of gas meters. The results were evaluated using the Mean Absolute Error metric. It was concluded that the proposed models solves the problem of bill prediction with errors at levels currently acceptable by regulatory agencies.

Related publications:

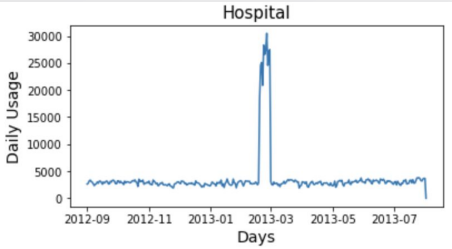
– T. Lane *et al*, Cybersecurity for Industry 4.0 , 01-33 (2017)
– SJ. Taylor et al, The American Statistician (1), 37-45 (2018)

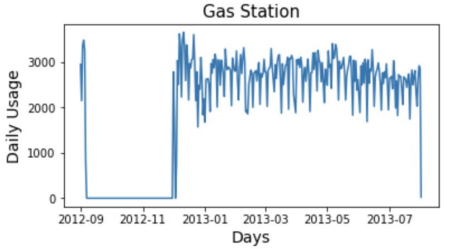
Problem, Data

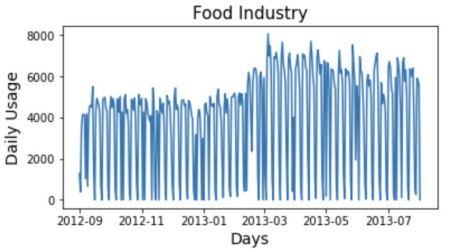
- With the digital transformations of the systems for measuring gas consumed and charging by utilities, problems such as data loss, fraud, lack of in-person or remote reading of gas meters, along with cyber-physical attacks and blackouts, are expected to occur more recurrently.

- The available data is the hourly gas consumption information between 1 September 2012 to 1 August 2013 for a food industry establishment, a gas station and a hospital. Data contains missing values, typo outliers and change of behaviour.

- To overcome these problems, machine learning algorithms were used at each facility to predict the gas consumption for short and medium-term forecasts.





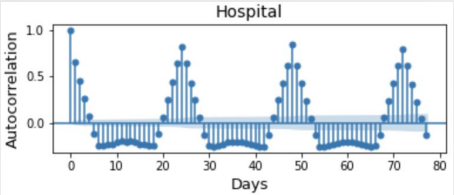


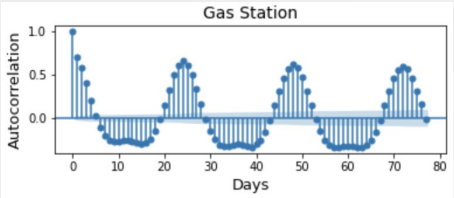
Methods

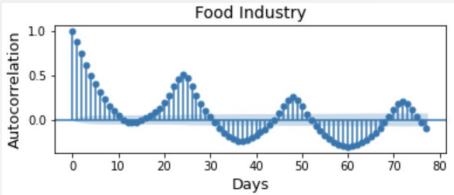
- Use of InterQuantile Range for outlier correction and median (by Day/Hour) for missing values
- For each establishment:
  - A sazonality study was done by Hour and Day of the Week.
  - Train between Sep 1, 2012 and Jun 30, 2013 and validation on July 2013.
  - 16 different Prophet (1) models with exogenous variables created considering Hour and Day of the Week for different data lags. A grid-search were performed considering different values for changepoint and seasonality prior scales, resulting a total of 256 models tested.

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (1)$$

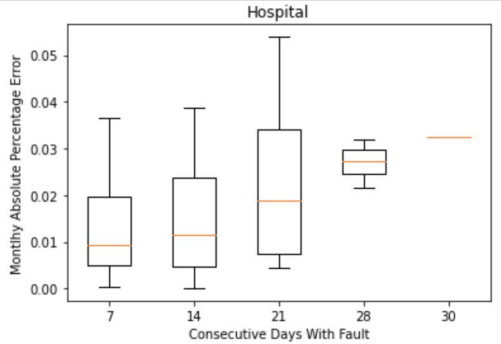
- To analyse the performance of each model a experiment was done considering different numbers of consecutive missing days (7, 14, 21, 28 and 30) to predict in one month.

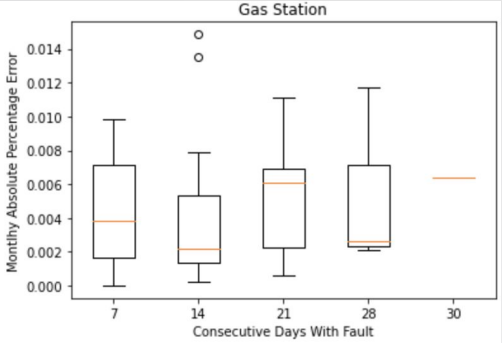






Results and Conclusions





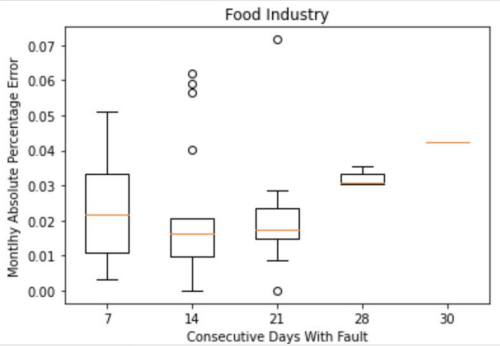


Table 1 - % Scenarios with MAPE < 5%

Days with Fault	Hospital	Gas Station	Food Industry
7	100%	100%	96%
14	100%	100%	82%
21	90%	100%	90%
28	100%	100%	100%
30	100%	100%	100%

- The models are fitting really well the data considering that at least 80% of the scenarios without data have a monthly error lower than 5%.
- Further studies will be done with more data and different models.