ANALYSIS AUTOMATION SYSTEM

PREDICT SIGNAL BEHAVIOR

BENAZIR ABIGAIL DE LA ROSA MUÑOZ

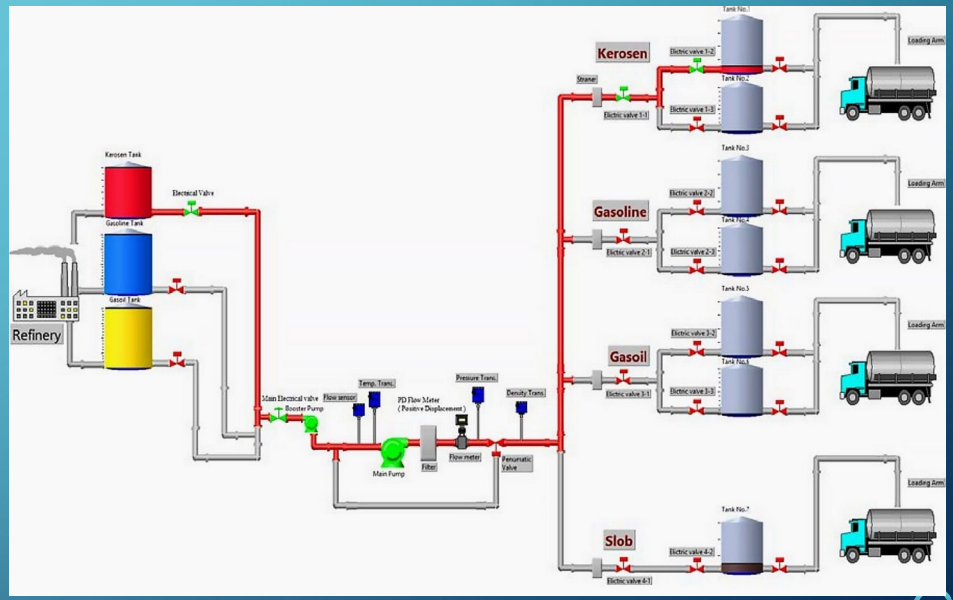
DATA SCIENTIST CONSULTANT

OVERVIEW

- Business Case Analysis
- Data Clening Analysis
- Descriptive Analytics
- ML/DL algorithm
- Evaluation method
- Results
- Next Steps

BUSINESS CASE

- The use case processed the data from a public website Anero energy. (https://anero.id/energy/data)
- The goal is to predict a signal from this dataset based on the other signals.
- The dataset represents the behavior of a SCADA system. It contains the signals from 318 different kind of actuators and sensors.
- The models were develop with Microsoft Azure platform
- Overall we have measures from different signals with Intervals of 5 minutes from 1 year period January 2018 to January 2019.
- The data is not enough to do feature engineering with seasonality, we need at least 2 years.

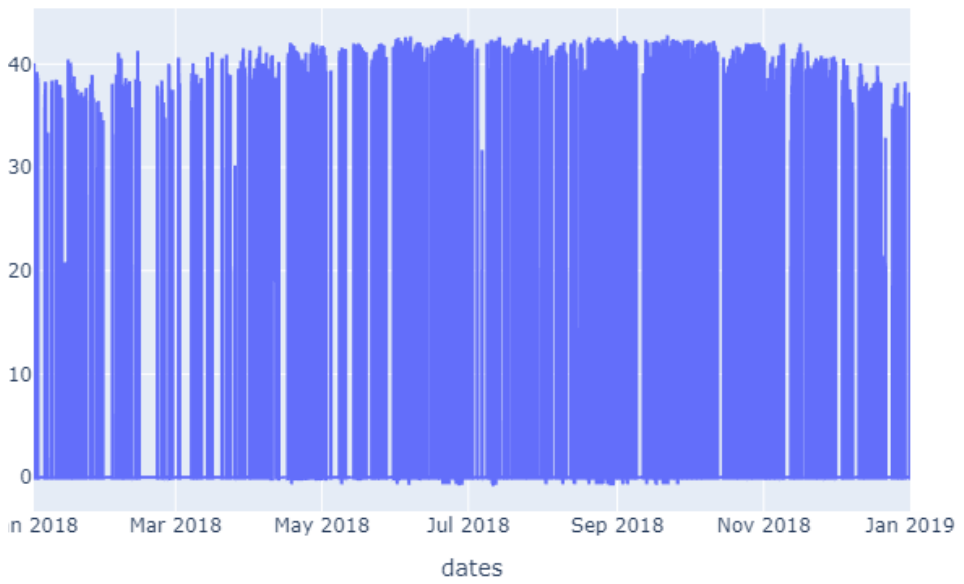


DATA CLEANING ANALYSIS: PCA

- Given that the dataset contains 318 signals it was necessary to use PCA, to find out how many of them represented a big impact over the dataset. As a result I found out that 30 features were useful for advance analytics with 70% of the total variance.
- From the selected features, I selected a dependent variable to predict based on the other variables. It was choosen because it seemed to me as an analog sensor which depended on other signals (it had always real values). The name of the variable BLD01.

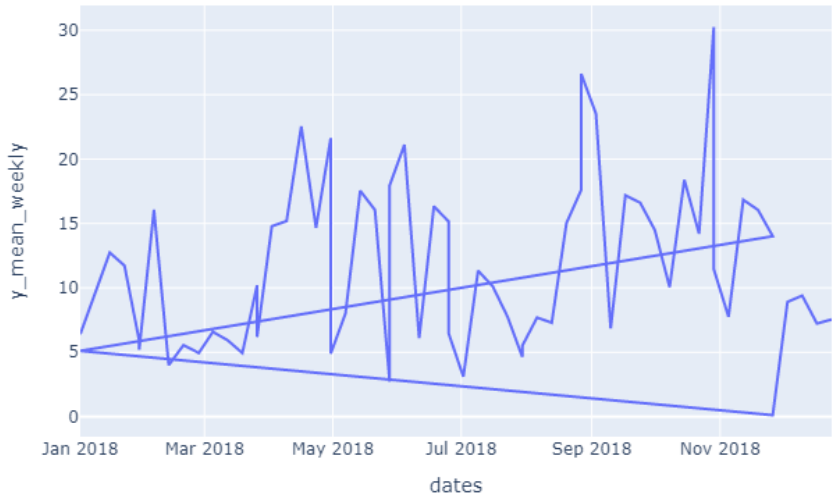
Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
variance	0.005859	0.036083	0.005717	0.072114	0.001273	0.001355	0.115245	0.024854	0.002438	0.134883	0.18213	0.181016	0.215934	0.004492	0.0042	0.002219	0.161269	0.141894	0.074948	0.103258	0.069845	0.010198	0.12322	0.021427	0.037199	0.089985	0.07079	0.047589	0.065251	0.035377

DESCRIPTIVE ANALYTICS: SIGNAL BEHAVIOR

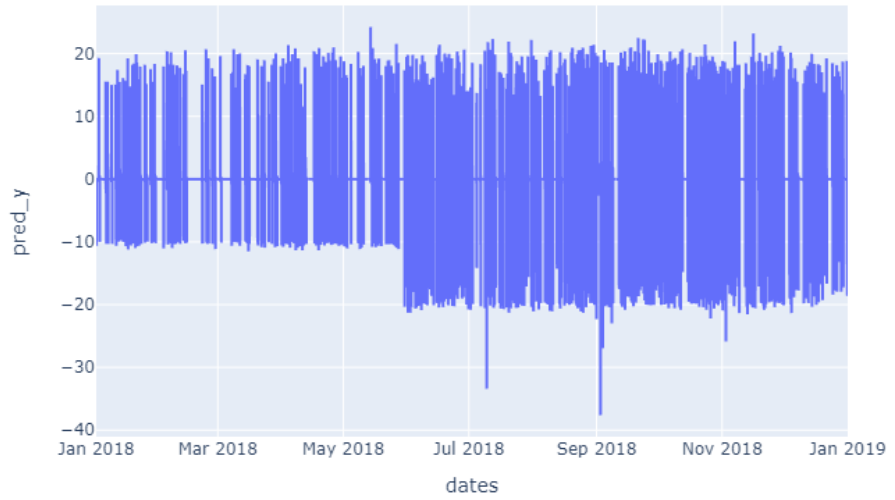


THE GRAPH REPRESENTS THE BEHAVIOR OF THE SIGNAL WITH ITS ORIGINAL 5 MINUTES INTERVAL.

DESCRIPTIVE ANALYTICS: SIGNAL BEHAVIOR



- The next graph represent the average aggregate values of the signal with a weekly period. It was necessary to be able to see that indeed it has a trend and a seasonallity.



DATA CLEANING: DEPENDENT SIGNAL NORMALIZATION

- To be able to predict better the signal it was normalized by obtaining the difference between the actual value and the previuos value. The graph shows the normalized signal.

MACHINE LEARNING / DEEP LEARNING

THE ANALYSIS WAS DONE WITH A XGBOOST REGRESSOR AND A LONG SHORT TERM MEMORY NEURAL NETWORK

MACHINE LEARNING: XGBOOST

- The input features from the XGBoost, were the 29 represented features, plus the previous value of the dependent signal and the timestamp decomposition by year, month, day, week and hour.
- Several models were trained (git repo for details), the selected one had the next parameters:
  - N_estimators 50
  - max_Depth 30
  - Learning rate 0.3

DEEP LEARNING: LSTM NEURAL NETWORK

- The input features from the LSTM, were the 29 represented features, plus the previous value of the dependent signal and the timestamp decomposition by year, month, day, week and hour.
- Several models were trained (git repo), the selected one had the next parameters:
  - Selected 128 neurons.
  - Droupout rate 0.2
  - Dense units 1
  - Epochs 12 (per month)
  - Batch_size 96 (8 hour period, intuition of behavior per morning, night and evening)

EVALUATION: MEAN SQUARE ERROR

- The selected metric to check the performance of the model was the MSE.
- The closer to 0, the better.
- Example of metric usage.

VALUE REAL	VALUE PREDICTED	
1	0.5	0.25
2	1.8	0.04
5	4	1
8	7.9	0.01
9	9.5	0.25
	Sum/n	0.31

RESULTS: MODELS COMPARISON

- The results of the XGBoost with a 15 k-fold cross validation were as follows:

MSE: 0.00014

- The results of the LSTM Neural Network with a15 k-fold cross validation were as follows:

MSE: 0.0000296

Both of them reach out a good score.

GIT REPO: https://github.com/nonameforpirate2/SCADAAnalysis

NEXT STEPS

- More feature engineering, the model is a baseline
- Model selection
- Deployment
- Raken Data Group Consultancy