Paris Session 2022

cigre
For power system expertise

Addressing anomalies with impact on technological and/or economic performance of the power system

D2, PS 1 – Question 1.1

“What experience on improving the performance of machine learning-based systems does exist in terms of addressing the anomalies (rare event with significant consequences), which may pose considerable impact on technological and/or economic performance of the power system? How should we distinct anomalies and data outliers between each other?”

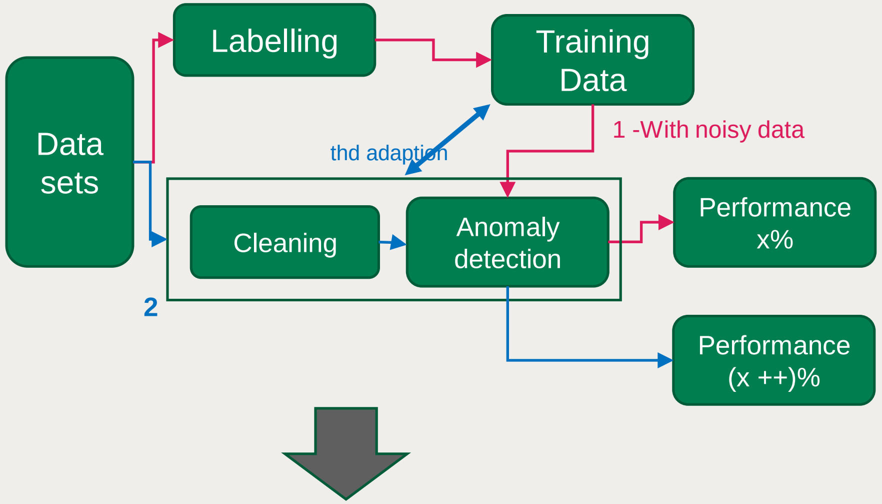
HITACHI
Inspire the Next

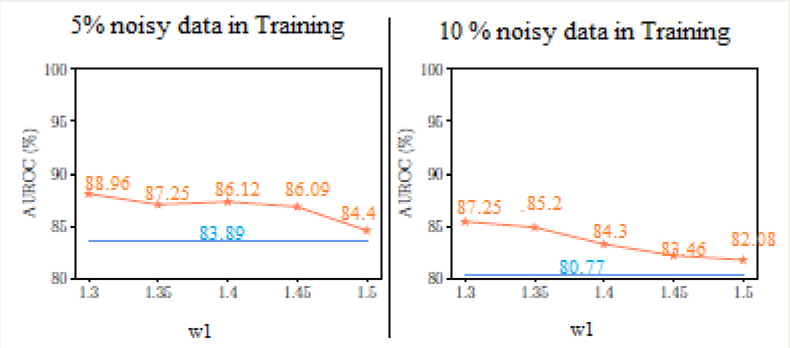
Sarala M Naidu, Sweden

Hitachi Energy

Improving the performance of machine learning-based systems in terms of addressing the anomalies..

- In the regular **supervised** approaches, the data labelling is often **challenging**, **time consuming and error-prone**
- **Leads to impurity in the datasets for training the models and** thus a limited accuracy say (AUROC metric) is obtained
- In the Power systems where large unlabeled datasets exists, such a learning process would need clean data sets for a reliable performance.
- Thus, **we acquire the following experience as crucial to improve the performance notably** :
  - Having a deeper understanding of the data- in collaboration with the domain experts (not just the AI experts):
    - To identify quality of the data – noise vs normal data
    - Outlier management strategy
    - Identification of false labelled data and removal of the same from Training data
    - Data pre-processing strategy
    - Information like the spatiotemporal context and slow drifts due to operational conditions of assets/systems effect the outlier removal, thus, such information should also be included in the training process
  - choosing the right model & its architecture , based on the data context
  - level of impurity in the datasets is unknown, different values of a weighting factor can be evaluated as a hyperparameter



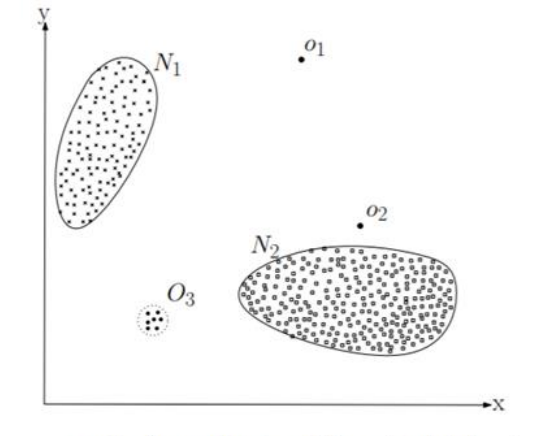


The area under the receiver operating characteristic (**AUROC**) is a performance metric that you can use to evaluate classification models

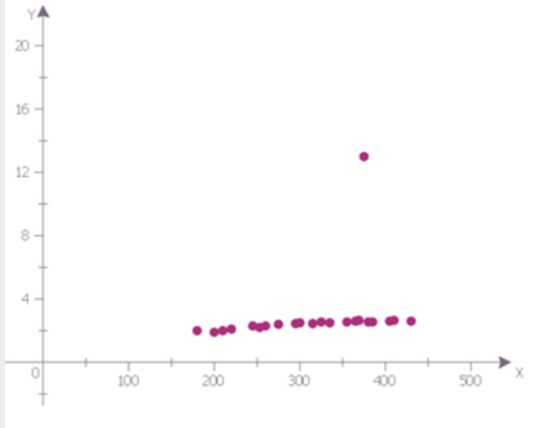
How should we distinct anomalies and data outliers between each other?

- Anomalies are patterns of different data within given data and detecting them leads to
  - New findings of patterns in the data or
  - new features that can be introduced to existing models

- Outliers would be merely extreme data points within data and detecting outliers leads to :
  - **improving the model accuracy through treatment of outliers.**
- Treating or altering the outlier/extreme values in genuine observations is not a standard operating procedure. Thus outlier management strategy is to be decided based on the use-case and this should be well documented or clearly stated on the subsequent report
- The outlier challenge is one of the earliest of statistical interests. And Statisticians have developed numerous algorithms for the detection and treatment of outliers.
  - A univariate outlier -extreme value on one variable. Some of the Univariate Outlier Detection Techniques popularly used are “The Box Plot Rule”, Grubbs Test
  - A multivariate outlier is a combination of unusual scores on at least two variables. Some of the Multivariate Outlier Detection Techniques popularly used are the Mahalanobis Distance, Cook’s Distance, etc.



example of anomalies in a 2-dimensional data set.



an Outlier

Thus, to improve performance in addressing anomalies, there is a large significance of the need for availability of spatiotemporal contextual information and also for clean datasets in training data