

Anomaly deTection

Problem Statement :

To identify *anomalous data points* that can potentially corrupt a dataset and *assign probabilities* of a given value to be an anomaly. Anomalies refer to **Outliers**, **change of unit** (for e.g. 85 bps reported as 0.85 bps) or **missing values** present in the data.

The data :

The given dataset contains quarter end data for 76 variables categorized into 8 Asset Classes (Equity, FX, Real Estate etc.) from **2007Q1** to **2018Q4** of a financial firm.

Data Preprocessing :

- A missing value can be present in any of the forms "n/a", "na", "--", "" or "NAN". Our task is to **identify & map** them to a single symbol.
- The identified missing values are estimated using **Linear Interpolation**.

Procedure Followed :

- **Binary Classification Models**

This type of model classifies each observation into 2 classes (anomaly or not), based on their underlying algorithm.

- **Probability Model**

Assigns probability of a given value to be an anomaly

- **Prediction**

Each model is assigned a weight based on its accuracy tested w.r.t their performance for the **CHECKED** columns/rows of data. The results are then combined w.r.t the weights and final prediction is made.

Binary Classification Models :

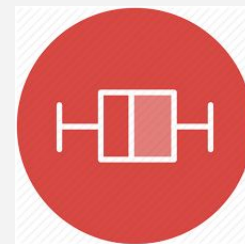
1. Local Outlier Factor (LOF)



2. Isolation Forest



3. Quantile Method (Using IQR)



Probability Models :

1. Empirical CDF Method



Results :

The final decision about a particular observation to be an anomaly is made upon the majority of the method's results.

The probability score is obtained by the ECDF method, where we used each asset class separately to identify their probability distribution.