# **Anomaly Detection – Unsupervised and Semi-Supervised Methods**

# **Problem Context and Description**

Anomaly detection refers to the problem of ﬁnding patterns in data that do not conform to expected behavior. Anomaly detection ﬁnds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities. The importance of anomaly detection is due to the fact that anomalies in data translate to signiﬁcant (and often critical) actionable information in a wide variety of application domains.

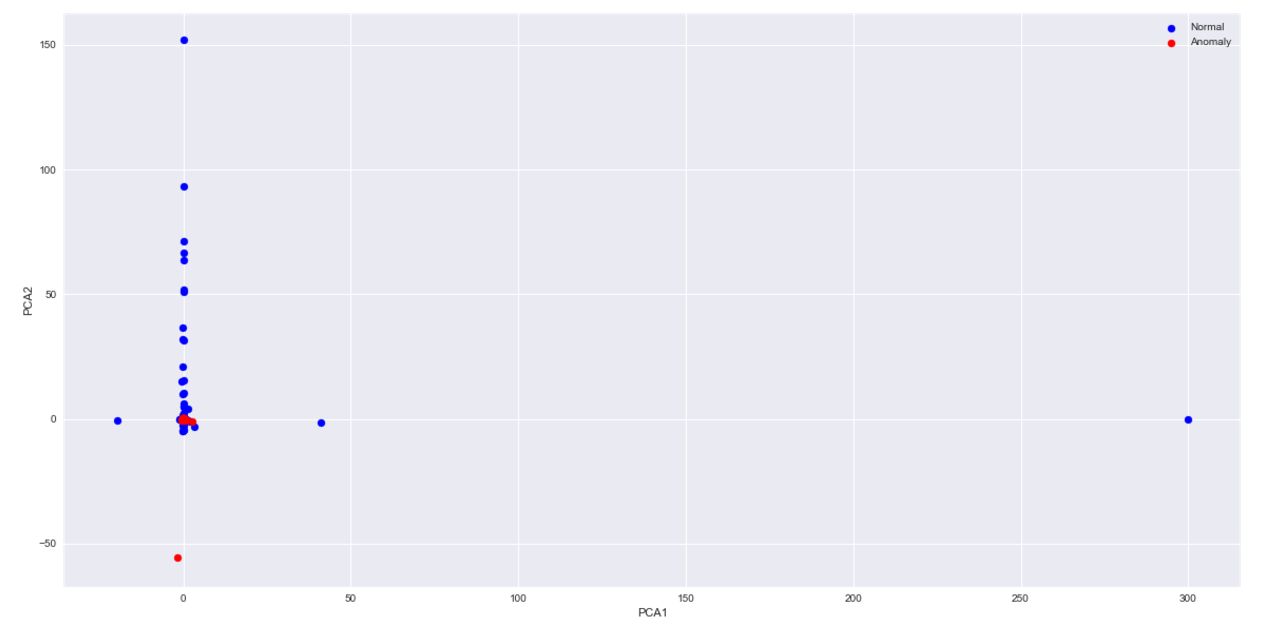
Although supervised techniques can be used to detect anomalies, availability of labeled data for training/validation of models is usually a major issue.

# **Dataset Description**

Data can be downloaded from: <https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data> (1-year data)

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service (EMIS, [[Web Link]](http://www.securities.com/)), which is a database containing information on emerging markets around the world. The data contains financial rates from 1st year of the forecasting period and corresponding class label that indicates bankruptcy status after 5 years. The data contains 7027 instances (financial statements), 271 represents bankrupted companies, 6756 firms that did not bankrupt in the forecasting period. Each observation contains 64 features e.g. net profit / total assets, total liabilities / total assets, working capital / total assets, current assets / short-term liabilities, etc.

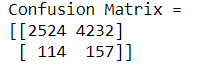
# **Exploratory Data Analysis**

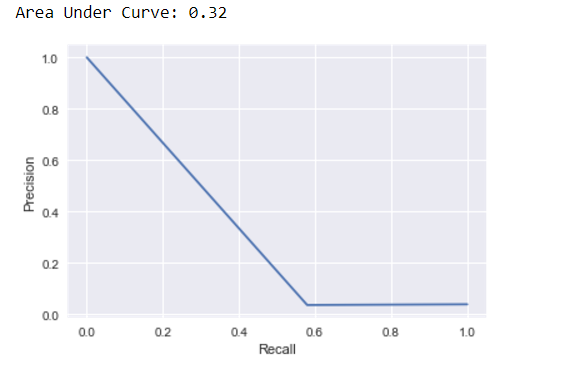


We can observe that

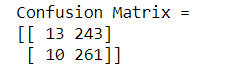
* Centroid-based clustering won't work as anomalous instances do not form a 'spherical' cluster
* Density-based clustering might work as anomalous instances form clusters of arbitrary shape. However, it might not be the best way of identifying outliers as densities of normal and anomalous instances overlap considerably.
* OneClassSVM, a semi-supervised technique, may be a better alternative, as it is immune to distribution of data

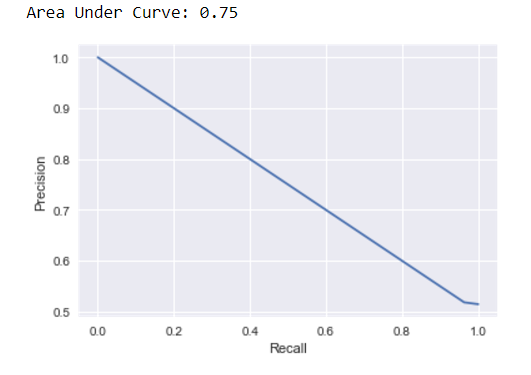
# **DBSCAN: Confusion Matrix and Precision-Recall Curve**





# **OneClassSVM: Confusion Matrix and Precision-Recall Curve**





We can observe from the true positive metric and precision-recall curve that oneClassSVM performs better than DBSCAN.