Anomaly detection

Here we present two data set for anomaly detection:

1. Simple synthetic dataset to evaluate 7 different algorithms;
2. Real dataset of trading volumes from the stock market to evaluate Isolation Forest Algorithm.

Synthetic case:

- **Input Data:**

The synthetic dataset has 300 data points with 10% outliers. We consider 2 features: feature 1 and feature 2. Figure 1 shows the cross-plot of Feature 1 vs Feature 2:

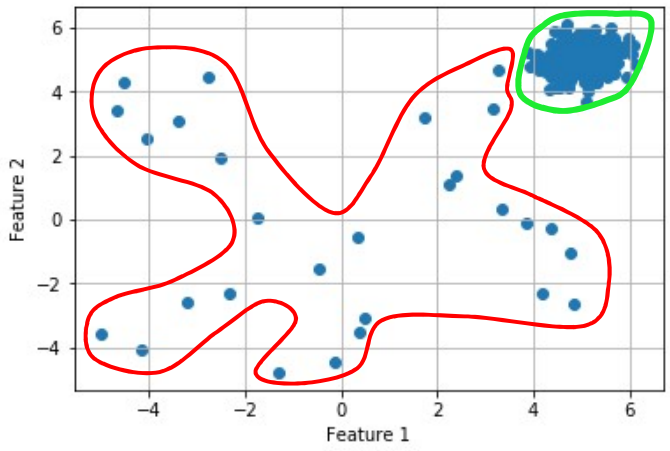

Figure 1: Feature 1 vs Feature 2 (300 synthetic data points). Inliers and outliers are highlighted with green and red contours respectively.

- **Anomaly detection:**

Here we use the library of PyOD for anomaly detection. In this case we investigate 7 different algorithms for anomaly detection:

Angle-based outlier detection (ABOD): it considers the relationship between each point and its neighbors. ABOD performs well on multi-dimensional data. There are two different versions of ABOD:
- FAST ABOD: uses k-nearest neighbors to approximate,
- Original ABOD: considers all training points with high-time complexity

k- Nearest Neighbors Detector: For any data point, the distance to its kth nearest neighbor is considered as the outlying score: PyOD support three kNN detectors:

- Largest: uses the distance of the kth neighbor as the outlier score.
- Mean: uses the average of all k neighbors as the outlier score
- Median: uses the median of the distance to k neighbors as the outlier score

Isolation Forest (IForest): it uses sklearn library. In this method, the data partitioning is done using a set of trees. The algorithm provides an anomaly score looking at how isolated the point is in the structure. The anomaly score is then used to identify outliers from normal observations. Isolation forest performs well on multi-dimensional data. (we will use this algorithm for the stock market).

Histogram-based outliers detection (HBOS): it is an effective unsupervised method which assumes the features independence and calculates the outlier score by building histogram. It is much faster than multivariate approaches, but with less precision.

Local Correlation Integral(LOCI): is very effective for detecting outliers and groups of outliers. It provides a LOCI plot for each point which summarizes a lot of the information about the data in the area around the point, determining clusters, micro-clusters, their diameters and their inter-cluster distances.

Feature Bagging: Fits several base detectors on various sub-samples of the dataset.

Clustering Based Local Outlier Factor: It classifies the data into small clusters and large clusters. the anomaly score is then calculated based on the size of the cluster the point belongs to, as well as the distance to the nearest large cluster

- **Results :** Anomaly detection

Table 1 shows a summary of the performance of all the 7 algorithms by highlighting the number of errors for each anomaly detector. Every algorithm, except ABOD, detected all the outliers. ABOD missed 3 outliers out of the 30 outliers in the data. Figure 2 provides a visual of inliers, outliers, and clustering using the 7 different algorithms.

Table 1: Number of errors for anomaly detection for different algorithms

Algorithms	**No of Errors**
Angle-based Outlier Detector (ABOD)	3
Cluster-based Local Outlier Factor (CBLOF)	0
Feature Bagging	0
Histogram-base Outlier Detection (HBOS)	0
Isolation Forest	0
K Nearest Neighbors (KNN)	0
Average KNN	0

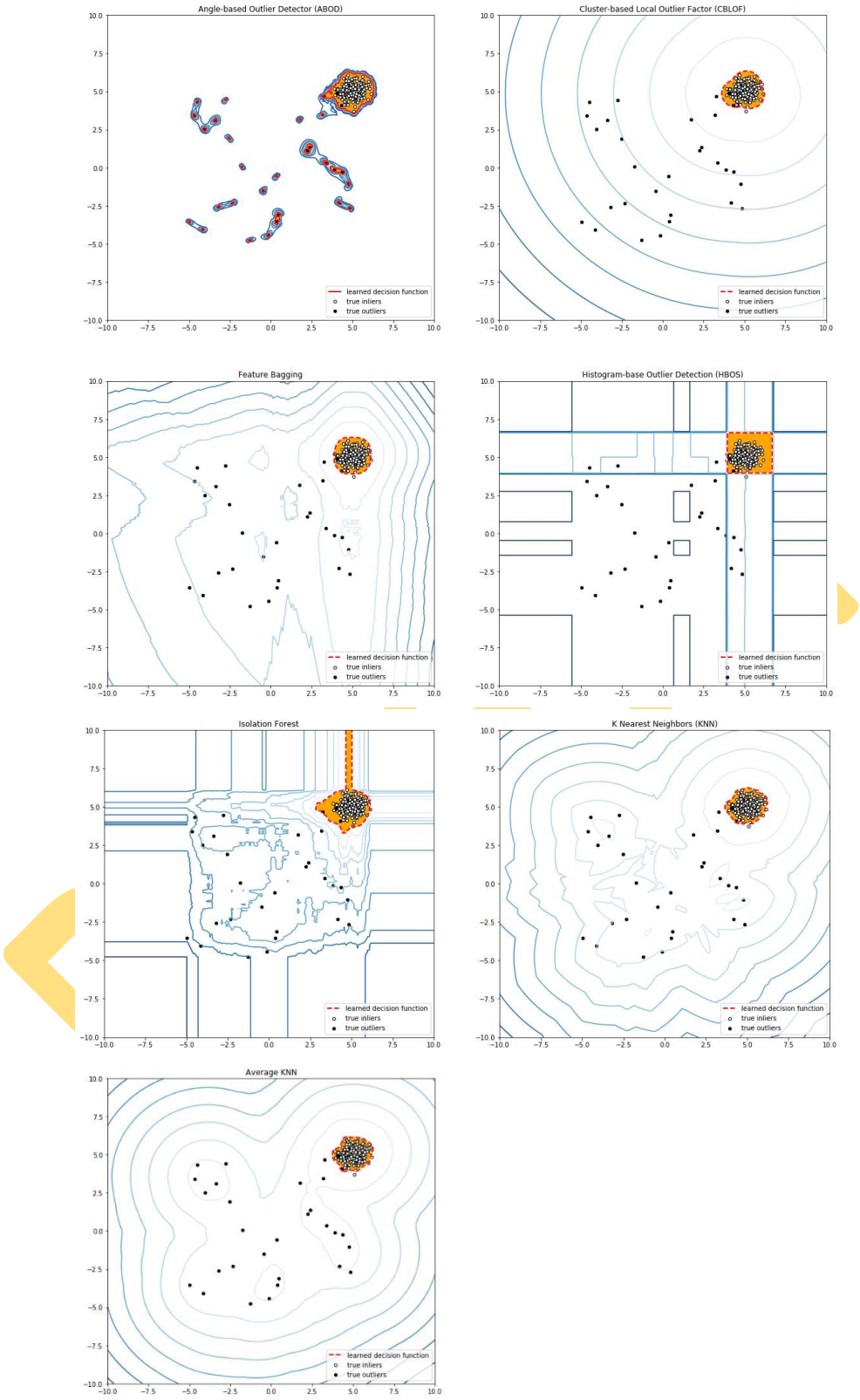

Figure 2: clustering of inliers and outliers by all the 7 algorithms

Real case dataset:

- **Input Data:**

As input data we use time series of trading volumes of ETFs from the US stock market. this application aims at evaluating when the trading volume for our list of symbols, as a whole, is in an anomalous state. this could mean, for example that we are detecting a spike in trading volume.

The symbols considered in this study are:

- 'SPY': SPDR S&P 500 ETF Trust, is used to track the S&P 500 stock market index
- 'IWM': iShares Russell 2000 Index, is used by day traders and investors alike to gain access to the small-cap segment of US stocks. It is highly liquid.
- 'DIA': SPDR Dow Jones Industrial Average ETF, a price-weighted index of 30 large-cap US stocks, selected by the editors of the Wall Street Journal.
- 'IEF': iShares Barclays 7-10 Year Trasry Bnd Fd, tracks a market-value-weighted index of debt issued by the US Treasury with 7-10 years to maturity remaining.
- 'TLT': iShares 20+ Year Treasury Bond ETF, Tracks the Barclays U.S. 20+ Year Treasury Bond Index.
- 'GLD': Tracks the gold spot price, less expenses and liabilities, using gold bars held in London vaults.
- 'SLV': SLV tracks the silver spot price, less expenses and liabilities, using silver bullion held in London.
- 'USD': us dollar

We are looking at trading volumes between 2012/01/01 and 2018/11/30. We download the historical data form yahoo Finance data API (see algorithm).

- **Data Visualization:**

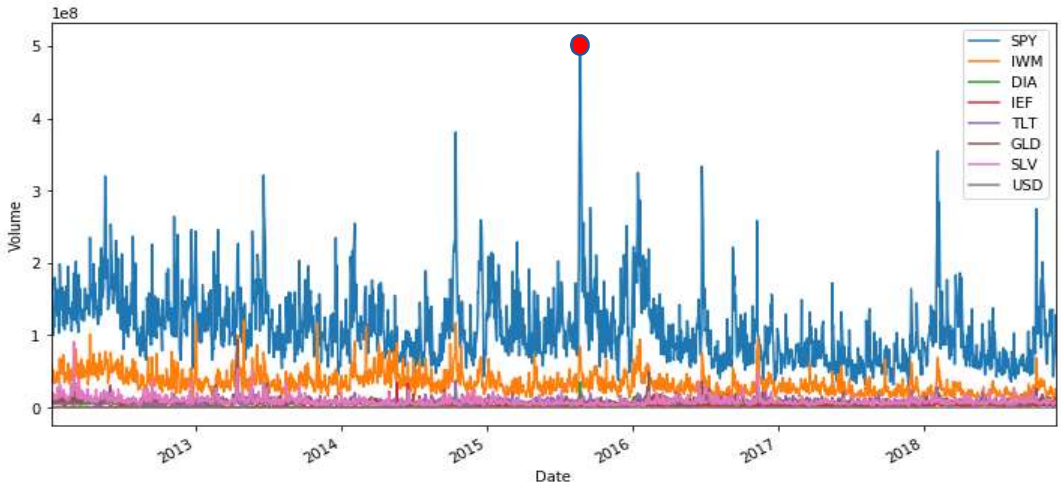

Figure 3: Trading Volumes for different ETFs between 2012-01-01 and 2018-11-30, the red dot isa a clear volume anomaly, it represents spike in volume the day Trump was elected.

Figure 3 shows the trading volumes variation with time. Visually there are clear spikes in volumes that represent anomalies in the market.

Histogram to visualize the trading volumes :

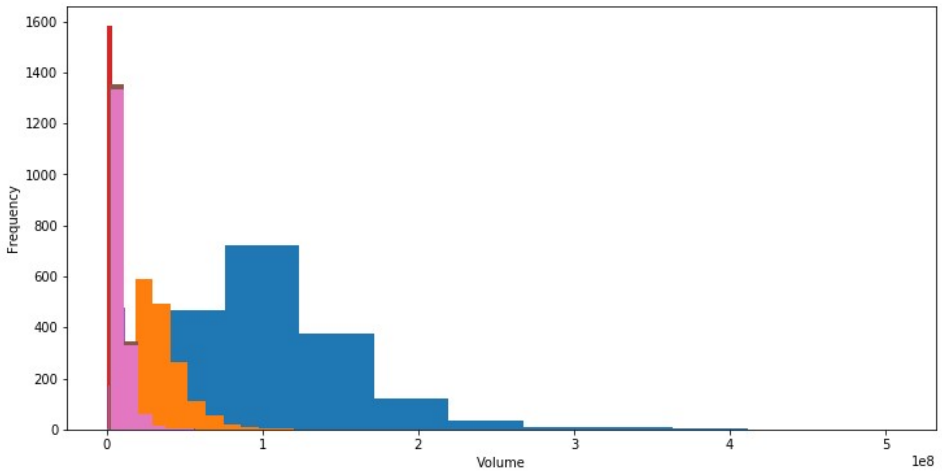

Figure 4: Histogram of trading volumes of the 8 different ETFs.

The histogram of the trading volumes of the 8 ETFs doesn't show clear separation between the inliers and the outliers. This was expected, since volume spikes only correlate with extraordinary events such as political events, events following a major Central bank/ Fed decision, sell-off triggered by technical trading.

- **Data clustering:**

To further explore the data we try to cluster the trading volumes and investigate the possibility of identifying outliers.

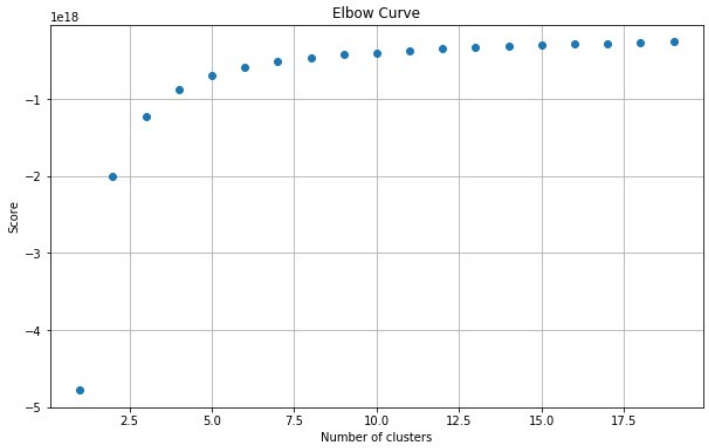

Figure 5: Elbow curve: search for optimum number of clusters base on the kmean score.

To find the optimum number of clusters for the dataset, we investigate the k-mean score for a range of number of clusters. Figure 5 is the Elbow curve of the kmean score for different number of clusters. The plot suggests that 10 clusters is the optimum number of clusters for this dataset. With 10 clusters the k-mean score start to plateau.

- **Visualize the clusters**

To be able to visualize the different clusters we apply Principal complement analysis PCA to reduced the dimensions to PCA1 and PCA2.

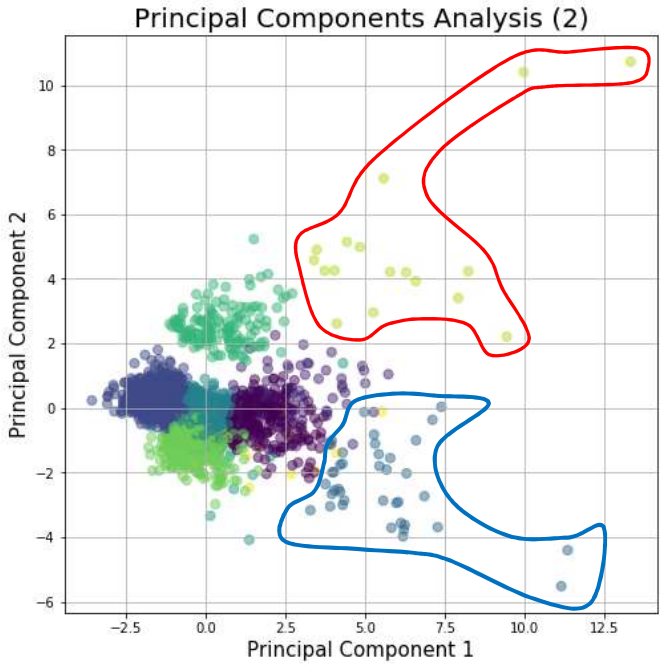

Figure 6: Data clustering - 10 clusters

Figure 6 shows the main 10 clusters, visually there 2 clusters, highlighted with red and blue contours, seem to represent the groups of outliers. This is not a quantitative approach to identify the outliers.

Isolation Forest: Anomaly detection

To detect anomalies we use the isolation forest to automatically detect the trading volume anomalies. Figure 7 shows the results from the Isolation Forest Algorithm. Red dots represent the anomalies detected. These anomalies correlate very well with major events that seems to affect market balance.

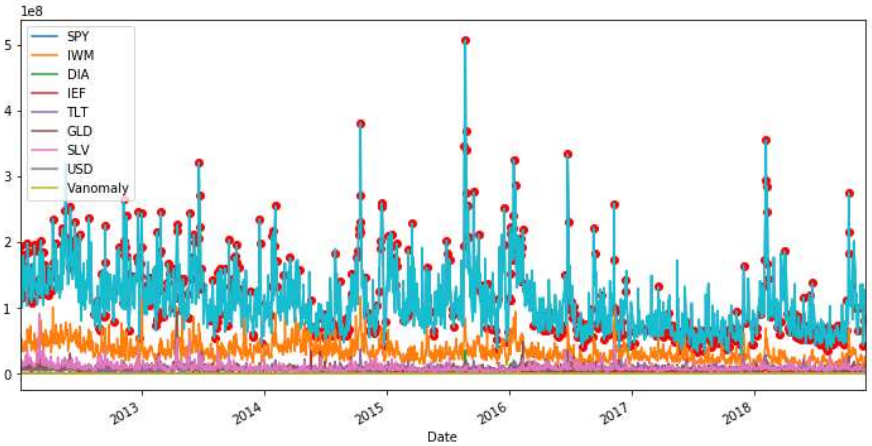

Figure 7: Trading Volumes over time. The red dots represent anomalies detected using Isolation Forest algorithm