Credit Suisse Hackathon 2019 Anomaly Detection

Purnendu Ghosh
Subhasish Basak
Trina De

Chennai Mathematical Institute

Anomaly detection - Approaches

- **LOF(Local Outlier Factor)**
- **Isolation Forest**
- **Density Based Approach**
- **Quantile Method**
- **Empirical CDF Method**

Thoughts on approaches :

All of the methods were applied using a different definition of outlier. Depending on what the definition of outlier is for us, all or any combination of methods can be used. Here we consider all such definitions important so we have used a combination of the methods for our identification.

(Details of thoughts on each approach in their respective slides.)

Alternate Approaches

- Since the data of different columns pertains to different portfolios which we do not know the relations of we cannot consider joint distributions among them. If we did we might be able to club columns together to find patterns in higher dimensions, that might increase accuracy.
- Extended Isolation Forest (Improves on the drawbacks of Isolation Forest).
- Mapper algorithm(Topological Data Analysis) can also be used for clustering the data that might reveal some abnormalities. This clustering can be done considering each column individually or by combining columns to form higher dimensional data points.

LOF-Local Outlier Factor

The local outlier factor is based on a concept of a local density, where locality is given by k nearest neighbors, whose distance is used to estimate the density. We fix a **k** and find the corresponding radius **r** such that the neighbourhood of a point p has at least **k** points. Corresponding to each point p we find the smallest **r** called the **local radius** of p. Then the LOF factor of a point is given by,

$$LOF(p) = local\_radius(p) / average\ local\_radius$$

This has a disadvantage because the value of this being less than 1 is not a clear indication of an outlier. We thus apply a second method where instead of considering only the k nearest neighbours, we consider the density or sparsity of the entire dataset as a whole.

Isolation Forest

This applies too, an algorithm based on the idea that an outlier point will be located in a region that is significantly different from a normal point. It exploits the idea of decision trees. It picks a feature at random and them chooses a random partition between the maximum and minimum values of the feature. It is believed that an outlier point would be identified at a shorter path length than usual. In this algorithm a score is assigned to each point based on the path length that is needed to identify it. This score is based on the path length to identify the point and the usual path length to categorize points in binary search.

Density Based Approach

In this method, we try to consider the variation of all points of a feature(i.e. A column) instead of considering only k nearest neighbours of it.
We take one column at a time. We then calculate the variation of points in that column. We have a function that computes the variation of values in that column omitting one value at a time. We then take the difference of each such value from the total variation of that column. Any values outside the **3-sigma neighbourhood of the mean of such difference** is believed to be an outlier and associated with the appropriate marker. The assumption is that something that causes more than usual reduction in the variation of the column is an outlier. And that the difference of the total variation with the variation leaving out a single value follows a normal distribution. We make this guess for the lack of a better more informed alternative.

Quantile Method

This is a nonparametric outlier detection method in a one dimensional feature space based on quantiles. Here outliers are calculated by means of the **IQR (InterQuartile Range)**.The first and the third quartile are (Q_1, Q_3) calculated.An outlier is then a data point x that lies outside the interquartile range.

$$X_i > Q_3 + K \cdot IQR \text{ or } X_i < Q_1 - k \cdot IQR$$

Where $IQR = Q_3 - Q_1$

Using the interquartile multiplier value k=1.5, the range limits are the typical upper and lower whiskers of a box plot.

Empirical CDF Method

All the above discussed method gives binary output,i.e. In other words it classifies each observation to either an outlier or not. In this method we estimate the **empirical cdf (F(t))** from different asset variable. We estimate the probability of an observation to be an outlier as :

$$p(\text{ x is an outlier }) = \min(F(x), 1-F(x)) / 0.5$$

The minimum takes into account both right & left tail of the empirical distribution and estimates only the extreme tail probability. The 0.5 normalization factor is introduced to scale the value in [0,1] interval so that it can be interpreted as probability.

Add Empirical Observations

- The data has 76 variables in all.
- Out of it 9 are checked and 67 are not checked.
- There are no patterns in the data over quarters.
- Since the data of different columns pertains to different portfolios which we do not know the relations of we cannot consider joint distributions among them.
- The following are the weightages of the different methods we have used(here the methods are LOF, IF, IQR, ECDF, VIF respectively) :

```
0.2488479262672811, 0.2488479262672811, 0.2488479262672811, 0.004608294930875576,
0.2488479262672811
```

- All methods have similar weightage except for ECDF which performs poorly on known checked data in comparison to others.

Combining The Results

The probabilities are converted to boolean using 0.5 as threshold and the classifiers are combined with appropriate weights. The weights are calculated based on the performance of the methods on the **checked** data.

Further Scopes - Statistical Testing

- **Dixon Q Test** : To apply a *Q* test for bad data, arrange the data in order of increasing values and calculate *Q* as defined :

$$Q = \text{gap/range}$$

Where gap is the absolute difference between the outlier in question and the closest number to it.If $Q > Q_{table}$, where Q_{table} is a reference value corresponding to the sample size and confidence level, then reject the questionable point.This statistical testing assumes normality of the underlying dataset.

Further Scopes - Statistical Testing

- **Grubbs’s Test :** Grubbs's test detects one outlier at a time. This outlier is expunged from the dataset and the test is iterated until no outliers are detected. However, multiple iterations change the probabilities of detection, and the test should not be used for sample sizes of six or fewer since it frequently tags most of the points as outliers.

  Grubbs’s test is then defined as :

  Null Hypothesis : There are no outliers in the dataset.

  Alternate Hypothesis : There is exactly one outlier in the data set.

  This test also requires the assumption of normality of the dataset.