Outlier Detection Methods

Ben Fuqua

Data Dojo

General Overview

An outlier is an instance of data that is different from the normal form of a dataset. In general, there are 4 types of data point in respect to outliers: True Positive, True Negative, False Positive and False Negative.

Types of outliers

In the graphic below I will describe these points in relation to our project:

- Blue points: True Positives or a real part number with the correct price/inventory/demand
- Grey points: True Negative or a fake part number with price/inventory/demand data that is outside the cluster
- Red points: False Positive or a fake part number with price/inventory/demand data that is within the cluster; or a real part number who's price/inventory/demand data is slightly outisde the cluster (this can only be determined by an algorithm)
- False Negatives rarely occur, and that is why they aren't represented but I will give a scenario
    - Let's say all of the manufacturers, except DK, ran out of HDMI cabels and they wouldn't be able to make more for 1 month. DK then raises their prices to $10,000/cabel for the next month.
    - This would be flagged as an outlier because the price is much higher than the norm.
    - This situation would be more correctly defined as an 'Anomaly'

Types of Outlier Detection

There are two different classification types, Global Detection/Local Detection and Univariate/Multivariate.

- Global and Local Detection
  - Global Detection is across the entire data with no respect to small clusters (Grey dots)
  - Local detection is with respect to the individual clusters (Red dot)
- Univariate and Multivariate
  - Univariate is with respect to 1 variable
    - ex: Does it make sense that person 1 ran 10 MPH and person 2 ran 100 MPH?
  - Multivariate is with respect to 2+ variables
    - ex: it makes sense that the RPM in a car can be 0 and MPH can be 0, but does it make sense when the MPH is 50 and the RPM is 0?
    - Very important to put causitive variables together in multivariate

The Process

Although there are many different styles of detection, the process remains the same.

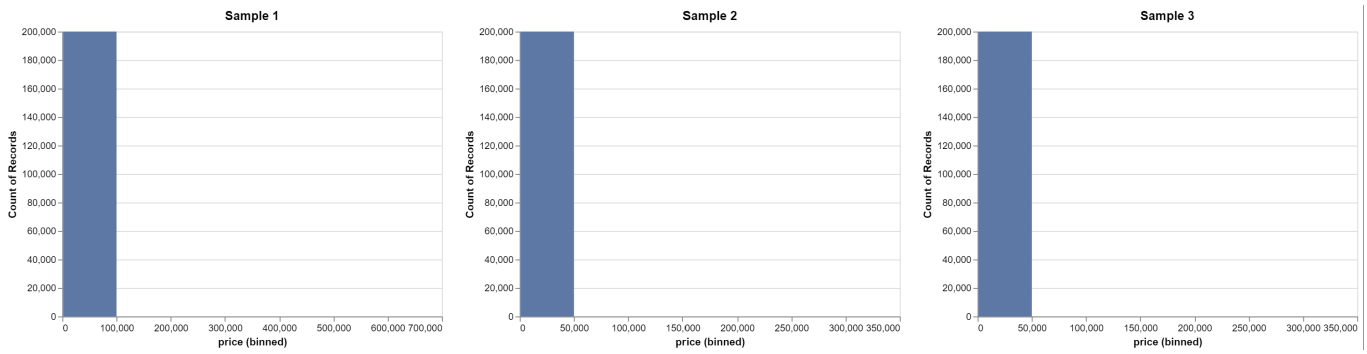
- First, you need to plot your distribution.
  - This normally looks like a box and whisker plot or a bar chart.
  - You do this to see if you need to run any global detection first.
- Run your global outlier detection
  - It is always good to run this because even though it may not look like an outlier, the algorithms can surprise you.
  - We do global outliers first becuase it allows the algorithms to look closer into the clusters of data.

- Second, run your local outlier detection methods
    - This will help better define the clusters and help train your ml model on 'normal' occurances.
- Lastly, plot your distribution again to make sure it looks the way you would expect it to look.
    - We do this to help us not only choose the best algorithm but to make sure it worked as expected.

Initial Distribution

All of the models I will be building will be working off of the same 3 samples. I have a population of 8,000,000 rows, and then I randomly sample 300,000 rows without replacement. Below you can see the distribution and how it is right skewed

```
index1 = list(swr(7999999,200000,random_state = 200))
index2 = list(swr(7999999,200000,random_state = 47))
index3 = list(swr(7999999,200000,random_state = 1999))
sample1 = data.iloc[index1].reset_index()
sample2 = data.iloc[index2].reset_index()
sample3 = data.iloc[index3].reset_index()
```



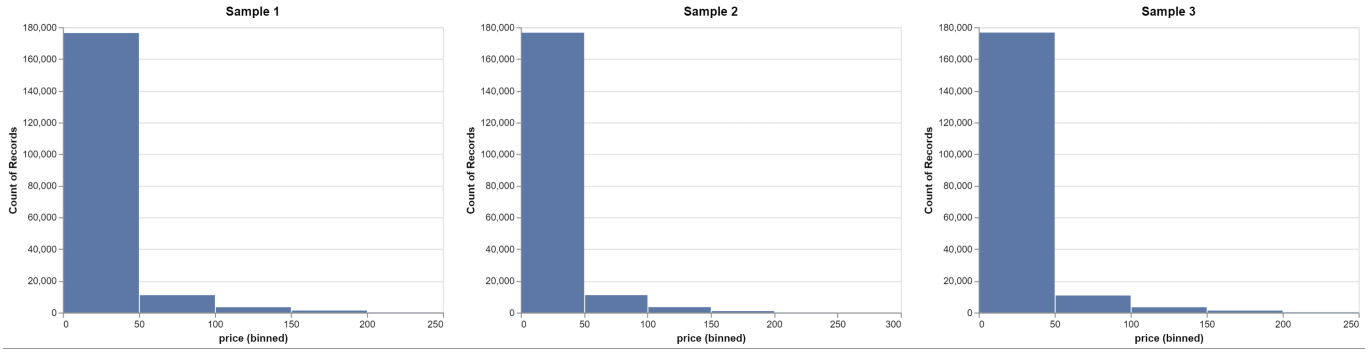
DBSCAN Detection Model

Global Outlier Detection Univariate

Below is the model set up, the important metric here is min_samples. We are saying there must be a minimum of 20 other points within 'eps'(epsilon) distance of it for it to not be considered an outlier

```
model = DBSCAN(
    eps = .2,
    metric = 'euclidean',
    min_samples = 20,
    n_jobs = -1
)
```

Here is the resulting distribution


Sample 1
Count of Records
price (binned)
Sample 2
Count of Records
price (binned)
Sample 3
Count of Records
price (binned)