Outlier Detection

DigitalSkola
Uncover The World Of Digital Skills With Us

Irony of the day

People wanted stand out and yet do not dare to do different.

DigitalSkola
Uncover The World Of Digital Skills With Us

Hello!
I am Agil Haykal

I am a Data expert with extensive experience in multiple industries such as marketplace, insurance, banking, general taxation, consulting, and training.

In total, I trained more than 300 data scientists, engineers, and analysts.

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

Story about Outliers

Cultures in each country has different methods to handle outliers.

Liberal countries tend to embrace and support outliers.

Conservative countries tend to remove outlier behaviors on their society.

As a rational people do we have to embrace it or remove it?

When to embrace or when to remove?

DigitalSkola
Uncover The World Of Digital Skills With Us

Story about Outliers

Liberal Quotes

Conservative Quotes

Really different, right?

What is an outliers

- Outliers are extreme values that deviate from other observations on data, they may indicate variability in a measurement, experimental errors or a novelty.
- Outlier is an observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism. — Hawkins (1980)
- To put it simply, outliers are observation that different from an overall pattern on a sample.

DigitalSkola
Uncover The World Of Digital Skills With Us

What causes outlier in data?

1. Data entry errors (human errors)
2. Measurement errors (instrument errors)
3. Experimental errors (data extraction by poor planning or execution)
4. Intentional (dummy outliers made to test the detection methods)
5. Data processing errors (data manipulation or data set unintended mutation)
6. Sampling errors (extracting/mixing data from wrong or various source)
7. Natural (not an error, novelties in data)

Types of Outliers

Global Outliers

Value is far outside the entirety of the data set in which it is found.

Contextual Outliers

Value significantly deviates from the rest of the data points in the same context.

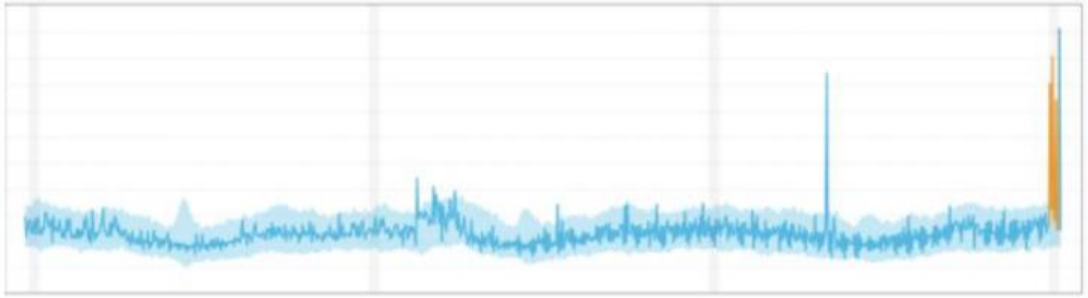
Collective Outliers

Values as a collection deviate significantly from the entire data set, but the values of the individual data points are not themselves anomalous in either a contextual or global sense.

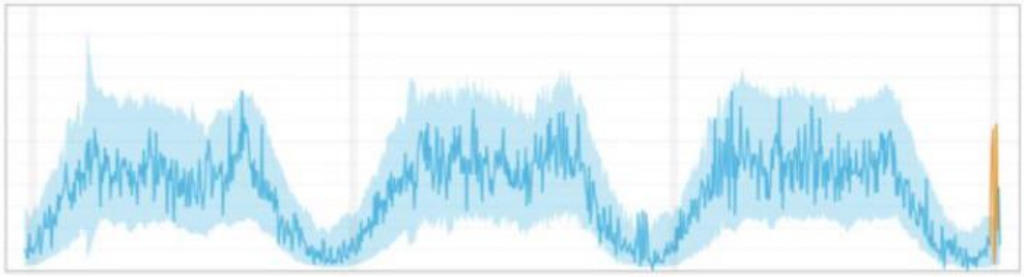
DigitalSkola
Uncover The World Of Digital Skills With Us

Types of Outliers

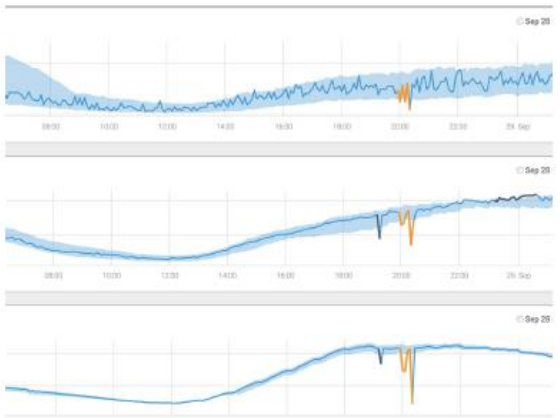
Global



Contextual



Collective



Why do we need to detect?

1. To prevent unwanted situation
2. To understand how extreme observation occurs
3. To get new direction in business

How to detect outliers?

1. Statistical Methods
2. Proximity-based Methods
3. Cluster-based Methods

Statistical Method

Z-score

The features are normally or approximately normally distributed.

IQR Based Filtering

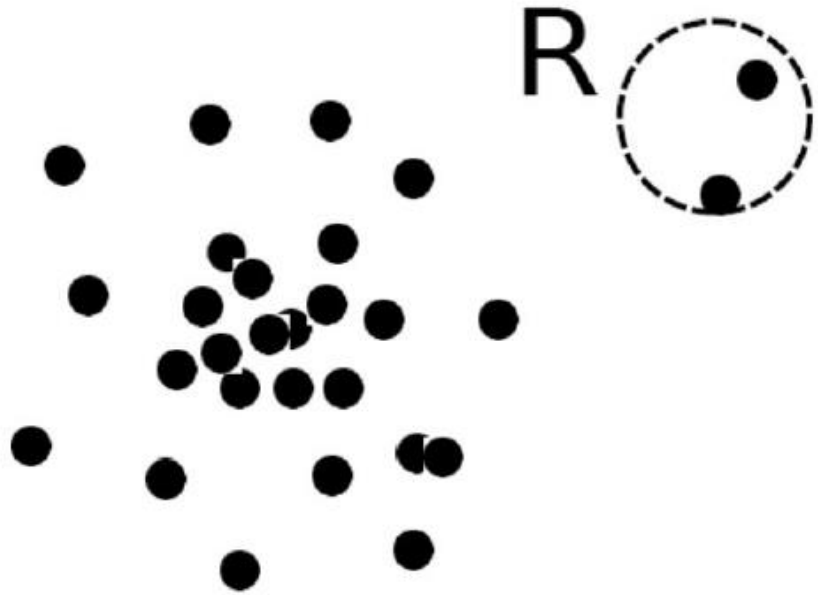
Used when our data distribution is skewed.

Percentile

Works by setting a particular threshold value, which decides based on our problem statement.

Proximity-based Method

An object is an outlier if the nearest neighbors of the object are far away. For example the distance of the object is significantly far away from the proximity of the most other objects.



Proximity-based Method

- This model rely on distance so normalization/standardization is important to do in advance.
- This model has disadvantage when unwanted outliers gather in a group.

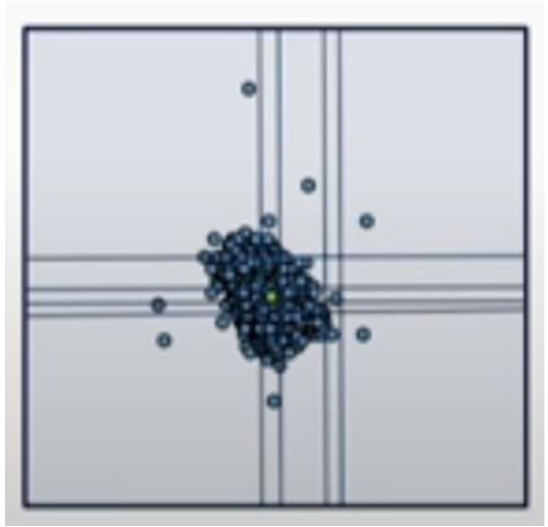
Cluster-based Method

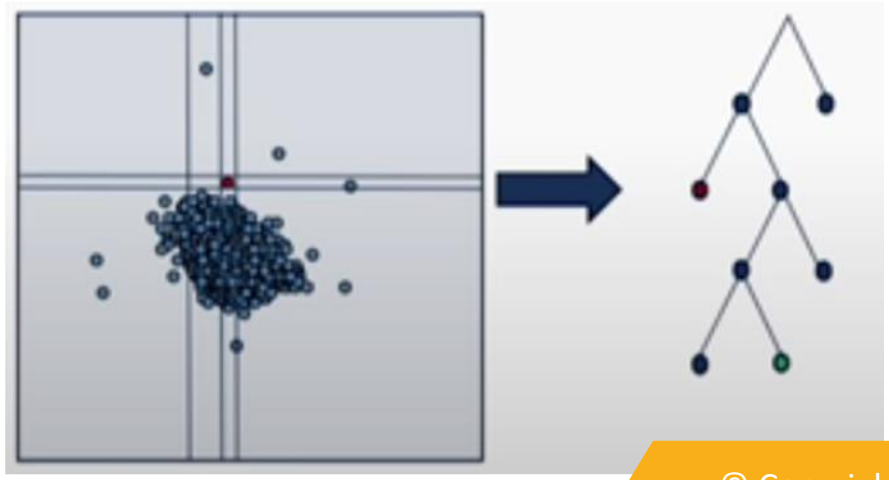
This method basically implement clustering algorithm to define the outliers. Each algorithm has their own method to define the outliers. So understanding the algorithm is essential before execute the model.

Isolation Forest

Isolation Forest is an outlier detection technique that identifies anomalies instead of normal observations.

Similarly to Random Forest, it is built on an ensemble of binary (isolation) trees. It can be scaled up to handle large, high-dimensional datasets.





How it works?

Isolation Forest is an ensemble regressor, and it uses the concept of isolation to explain/separate-away anomalies.

Isolation Forest builds an ensemble random trees for a given data set and anomalies are points with the shortest average path length.

Isolation forest can work as supervised and unsupervised classifier.

Why Isolation Forest?

It is a good method to detect outliers.

Since all cases anomaly data are imbalanced,

It can be used as anomaly data detection in a bank, such as money laundering, credit card fraud.

Outliers detection in Business

- **Fraud Detection** (The detection of fraudulent credit card transactions, insurance claims, expense reports or financial information.)

- **Quality Control** (The detection of production defects or product characteristics that do not fits the same standards as the other products that a company manufacturers.)

- **Intrusion Detection** (The detection of unauthorized access or attempts to computer networks or systems.)

Outliers detection in Business

- **4. Activity Monitoring** (The detection of (malicious) phone calls, messages and other forms of chatter which provides intelligence about people with bad intentions.)

- **5. Image analysis** (The detection of changed imagery. This can be applied to medical scans to detect certain types of diseases, or satellite imagery to detect abnormal or changed patterns.)

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU