What Are Outliers

Outliers are data points that deviate significantly from the majority of the data points in a dataset. They are observations that are either unusually high or low compared to the other data points, and they can have a significant impact on statistical analysis and machine learning models if not properly addressed.

Types of outliers

Global Outliers: Global outliers are data points that significantly deviate from the rest of the dataset across all variables. These outliers exhibit extreme values that are noticeable in the overall distribution of the data. They are typically identified using statistical methods that consider the entire dataset. Global outliers can be caused by measurement errors, data entry mistakes, or genuine rare events. For example, in a dataset representing the heights of individuals, a global outlier might be an exceptionally tall or short person compared to the majority.

Contextual (or Conditional) Outliers: Contextual outliers are data points that are considered outliers only within a specific context or subset of the data. They may appear normal when considered in the overall dataset, but when analyzed within a particular subgroup or condition, they exhibit unusual behavior. Contextual outliers are identified by taking into account additional contextual information or conditioning variables. For instance, in a dataset of housing prices, a property with a higher-than-average price in a specific neighborhood might be considered a contextual outlier if most other houses in that neighborhood have significantly lower prices.

Collective Outliers: Collective outliers, also known as structural outliers or group outliers, refer to groups or subsets of data points that collectively exhibit outlier behavior when considered together. These outliers are not necessarily extreme values on individual variables but instead demonstrate anomalous patterns or relationships among multiple variables. Identifying collective outliers often requires analyzing the interdependencies and relationships within the data. An example of collective outliers could be a group of customers who exhibit unusual purchasing patterns compared to the rest of the customer base.

Challenges of Outlier Detection:

Subjectivity: Determining what constitutes an outlier is often subjective and depends on the domain knowledge and context. Different analysts may have different opinions on what should be considered an outlier.

Data Quality: Outliers can arise due to data entry errors, measurement errors, or other data quality issues. It can be challenging to distinguish between genuine outliers and errors in the data.

Dimensionality: Outlier detection becomes more challenging as the dimensionality of the data increases. Visualizing and analyzing relationships between variables become difficult in high-dimensional datasets.

Influence on Analysis: Outliers can have a significant impact on statistical analysis, data mining, and machine learning models. They can affect parameter estimation, skew distributions, and distort results.

Identification Methods: There are various outlier detection techniques available, each with its own assumptions and limitations. It can be challenging to select the most appropriate method for a given dataset.

Trade-off with Data Size: Outlier detection becomes more difficult in large datasets due to computational constraints. Balancing computational efficiency and accuracy of detection methods can be a challenge.

Handling Outliers: Once outliers are detected, deciding how to handle them is another challenge. Should they be removed, transformed, or treated separately? The appropriate approach depends on the analysis objectives and the impact of outliers on the specific problem.

12.2 Outlier Detection Methods

Supervised Outlier Detection: These methods utilize labeled data with outliers and non-outliers for training a model. The model learns from the labeled data to classify new instances as outliers or non-outliers. Classification algorithms like SVM or Decision Trees can be used. Supervised methods require annotated labeled data, which can be time-consuming and expensive.

Semi-Supervised Outlier Detection: In this approach, a small portion of labeled data containing outliers or non-outliers is available, but most of the data is unlabeled. The goal is to use the labeled data to build a model that can detect outliers in the unlabeled data. Techniques like self-training or co-training can be employed. Semi-supervised methods aim to make the most of the limited labeled data available.

Unsupervised Outlier Detection: Unsupervised methods do not rely on labeled data. They identify outliers by finding patterns of anomalies or deviations from the normal behavior in unlabeled data. Clustering-based, density-based,

distance-based, and statistical methods are commonly used in unsupervised outlier detection. Unsupervised methods are versatile but may produce more false positives due to the absence of ground truth labels.

12.4 Proximity-Based Approaches

Proximity-based approaches are a category of outlier detection methods that identify outliers based on the proximity or distance of data points to their neighbors. These methods assume that outliers exhibit different proximity patterns compared to normal data points. Here are some commonly used proximity-based approaches:

Distance-Based Methods:

- Euclidean Distance: This method calculates the Euclidean distance between data points in the feature space. Data points that are far away from their neighbors are considered outliers.
- Mahalanobis Distance: The Mahalanobis distance accounts for correlations between variables by considering the covariance matrix. It measures the distance of a data point from the centroid or distribution of the data and identifies outliers accordingly.

Density-Based Methods:

- Local Outlier Factor (LOF): LOF measures the local density deviation of a data point compared to its neighbors. It assigns an outlier score based on how isolated a data point is from its local neighborhood. Points with significantly lower densities than their neighbors are considered outliers.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN groups data points into clusters based on density. Points that do not belong to any cluster or belong to small, sparse clusters are identified as outliers.

Subspace-Based Methods:

- Subspace Outlier Detection: These methods identify outliers based on the concept of subspaces, considering different subsets of features or dimensions. Outliers are detected when data points exhibit unusual patterns or behavior in specific subspaces.