OUTLIERS HANDLING

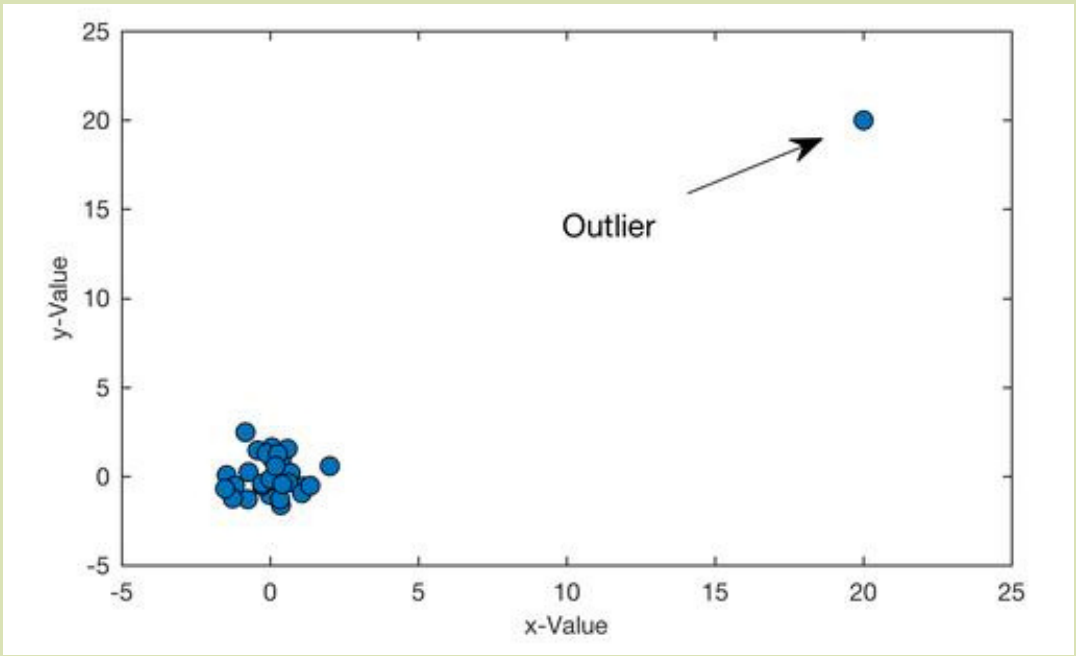
- Outliers and their types
- Impacts on ML model
- Detection of outliers
- Strategies for handling
- Key considerations

As data science enthusiasts, we know that data quality is a critical factor in achieving high-quality machine learning results.

Outliers are one common issue that can significantly affect the quality of our data and, therefore, our machine learning model's performance.

What are outliers?

Outliers are data points that deviate significantly from the rest of the data in a dataset.



Some of the common causes of outliers in a data are :

- **Measurement errors:** Outliers can be caused by errors in the measurement process, such as incorrect data entry or faulty measurement equipment.

- **Data entry errors:** Outliers can also occur due to errors in data entry, such as typos or missing values.

- **Sampling errors:** Outliers can be caused by errors in the sampling process, such as biased sampling or insufficient sample size.

- **Natural variation:** In some cases, outliers can occur naturally due to extreme values.

- **Anomalous data points:** Outliers can also occur due to anomalous data points, such as a sudden spike in a stock price or an unusual weather event.

What are the different types of outliers?

- **Point outliers:** A single data point that is significantly different from all other points in the dataset.

- **Contextual outliers:** A data point that is not an outlier in the global sense, but is an outlier in a specific context. For example, a price of ₹1,000 for a cup of coffee might be an outlier in a dataset of coffee prices, but not in a dataset of luxury item prices.

- **Collective outliers:** A group of data points that are collectively different from the rest of the dataset, but individually they may not be outliers.

- **Syntactical outliers:** Data points that are outliers due to errors in data entry or other syntactical issues.

What are the impacts of outliers on machine learning model?

- **Skew the distribution of our data:** Imagine you're analyzing the salaries of employees in a company. If the CEO's salary is much higher than other employees, it would skew the distribution and make it harder to draw meaningful insights.

- **Bias our model:** Outliers can influence the model's parameters and cause it to favor certain patterns or trends that may not generalize well to new data results in increase the models error rate.

- **Reduce accuracy:** Outliers can introduce noise and make it harder for the model to identify meaningful patterns and relationships in the data.

Now, How can we detect these outliers? Here are some popular strategies:

Z-score method :

- Significance **:** This method helps to identify and remove the data points that are too far from the mean, and hence are considered outliers.

- To detect univariate outliers.

Interquartile range (IQR) method :

- Significance : This method involves calculating the IQR of the data and then identifying and removing the data points that fall outside 1.5 times the IQR.

- To detect multivariate outliers.

Box plots

- Box plots provide a graphical representation of the IQR method.
- They display the median, quartiles, and outliers in a visual format. Any data point outside of the whiskers of the box plot is considered an outlier.

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that identifies outliers as points that are not part of any cluster.

- This method is useful in detecting outliers in high-dimensional data sets.

Local Outlier Factor (LOF)

- LOF is a density-based method that measures the local deviation of a data point with respect to its neighbors.
- Any data point with a significantly lower LOF score than its neighbors is considered an outlier.

These are some of the popular strategies used to detect outliers, and each has its own significance and use case depending on the dataset and problem at hand.

So how can we handle outliers? Here are some common strategies:

Removal

- This strategy involves identifying the outliers and simply removing them from the dataset.
- However, this approach should be used with caution as removing too many data points can result in a loss of valuable information.

Transformation

- This strategy involves transforming the data to reduce the impact of outliers.
- Common transformations include logarithmic, square root, and inverse transformations.
- These transformations can help to normalize the data and make it more suitable for analysis.

Capping

- This strategy involves setting a cap or limit on the values of the data. Values above or below the cap are then replaced with the cap value.
- This approach can help to prevent extreme values from distorting the data.

Binning

- This strategy involves grouping the data into bins or categories based on their values.
- This can help to reduce the impact of outliers by smoothing out the data.

Imputation

- This strategy involves replacing the outlier values with a new value based on other data points.
- For example, the mean, median, or mode of the data can be used to impute the missing or extreme values.

Machine Learning Models

- Some machine learning models are inherently robust to outliers.
- For example, decision trees and random forests can handle outliers well by partitioning the data based on its values.

It's important to note that the choice of outlier handling strategy depends on the specific use case and the nature of the data.

Here are some key considerations while handling outliers in data cleaning:

- **Understand the data:** It is important to understand the data and the domain before handling outliers. Different domains may have different definitions of what constitutes an outlier, so it is important to have domain knowledge.

- **Identify the cause of outliers:** Understanding the cause of outliers can help in choosing the appropriate strategy to handle them.

- **Determine the impact of outliers:** It is important to evaluate the impact of outliers on the data and the machine learning model. In some cases, removing outliers may be necessary, while in other cases, they may provide valuable information.

- **Choose an appropriate strategy:** There are various strategies to handle outliers, and the appropriate strategy depends on the cause of outliers and the impact they have on the data and the machine learning model.

- **Validate the strategy:** It is important to validate the chosen strategy to ensure that it is effective in handling outliers. This can be done by evaluating the performance of the machine learning model with and without the outliers.

- **Document the process:** Documenting the process of handling outliers is important for transparency and reproducibility. This includes documenting the chosen strategy, the rationale behind it, and the impact on the data and machine learning model.

In summary, managing outliers is essential for producing high-quality machine learning results.

By using best practices such as data visualization and statistical methods, and carefully considering how to handle outliers, we can avoid misleading insights and ensure that our model is accurate and reliable.

Happy learning!😊

THANK YOU

@ Gangadhar Neelam