Introduction to Data Analytics

AIGC5000
Lecture 6-Outliers

Instructor: Parisa Pouladzadeh
Email: parisa.pouladzadeh@humber.ca

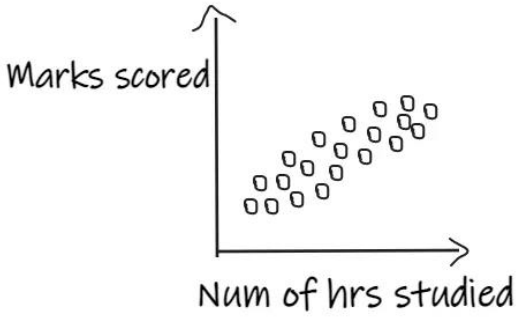
What are Outliers?

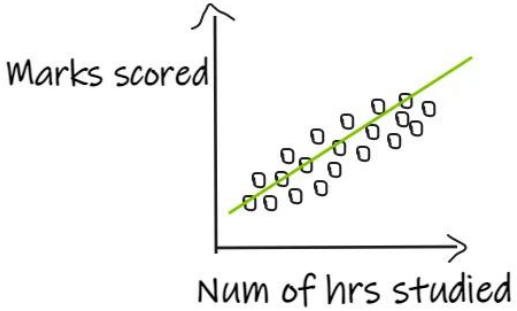
- **What are Outliers?**
- An outlier is that datapoint or observation which behaves very differently from the rest of the data.

- **Example:**
- If we are finding the average net worth of a group of people, and if we find Elon Musk in that group, then the complete analysis will go wrong because of just one outlier. This is a reason why outliers should be treated properly before building a machine learning model.

Outliers in linear Regression

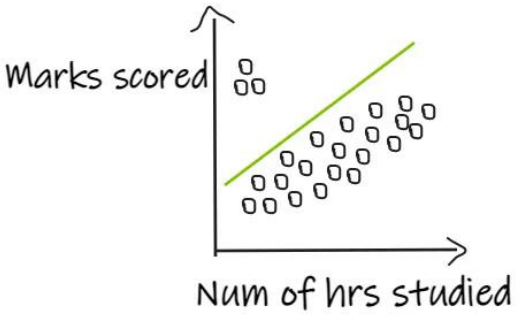
- If we are building a linear regression model, which has an independent feature, ‘Num of hours studied’, and the dependent feature, ‘marks scored’, and if the data is distributed as shown below, then the model will perform well.





Outliers in linear Regression

If we have 3 students who scored good marks even after studying for fewer hours, then the regression line shifts in order to fit the outlier points as shown below, resulting in giving bad results to the actual data.



Outliers in Machine learning algorithm

- Machine learning algorithms in which the calculation of weight is involved, like linear regression, logistic regression, Ada boost, and deep learning models, will get impacted by the outliers.
- Tree-based algorithms like Decision trees, Random Forest will get less impacted by outliers.

Outlier Treatment

- **Trimming**: Remove the outliers from the dataset before training a machine learning model. E.g., Remove the students from the dataset in the above example.

- **Capping**: Keep a maximum or minimum threshold and give values to the data points accordingly. E.g., if we are working on the age feature, we can keep the threshold of 85 and assign the value of 85 to all the people with age greater than 85.

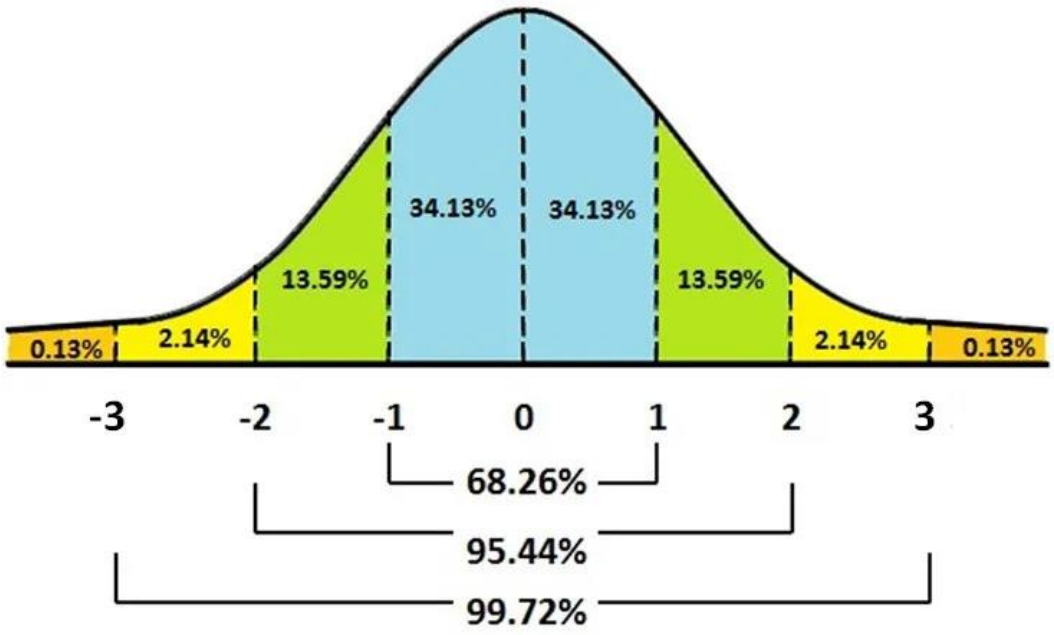
- **Discretization**: This is the method in which numerical features are converted to discrete using bins. E.g., if the age 80–90 is considered as a single bin, then all the ages between 80 and 90 will be treated equally.

Outlier Detection and Removal Techniques: 1. Z Score-based method

- **1. Z Score-based method**
- The main assumption in this technique is that the data should be normally distributed or close to normal distribution.
- If the data is normally distributed, the Empirical Rule says that 68.2% of the data points lie in between the 1st standard deviation, 95.4% of the data points lie in between the 2nd standard deviation, and 99.7% of the data points will be between the 3rd standard deviation.

Z Score-based method

- Data points that lie outside the 3rd standard deviation can be treated as outliers.



- As 99.7% of the data will lie within the 3 standard deviations, we can treat the rest of the data which lie outside the 3 standard deviations as outliers.

Standardization or Z-Score Normalization

Standardization or Z-Score Normalization is one of the feature scaling techniques, here, the transformation of features is done by subtracting from the mean and dividing by standard deviation.

This is often called Z-score normalization. The resulting data will have the mean as 0 and the standard deviation as 1.

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

Formula to calculate Z score

Treatment

Outlier Treatment:

Trimming: In this method, we can remove all the data points that are outside the 3 standard deviations.

Capping: In this method, the outlier data points are capped with the highest or lowest values

Let us look at the practical implementation of this technique.

2. IQR technique

- This method is used when the distribution of the data is skewed.

- The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first, find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

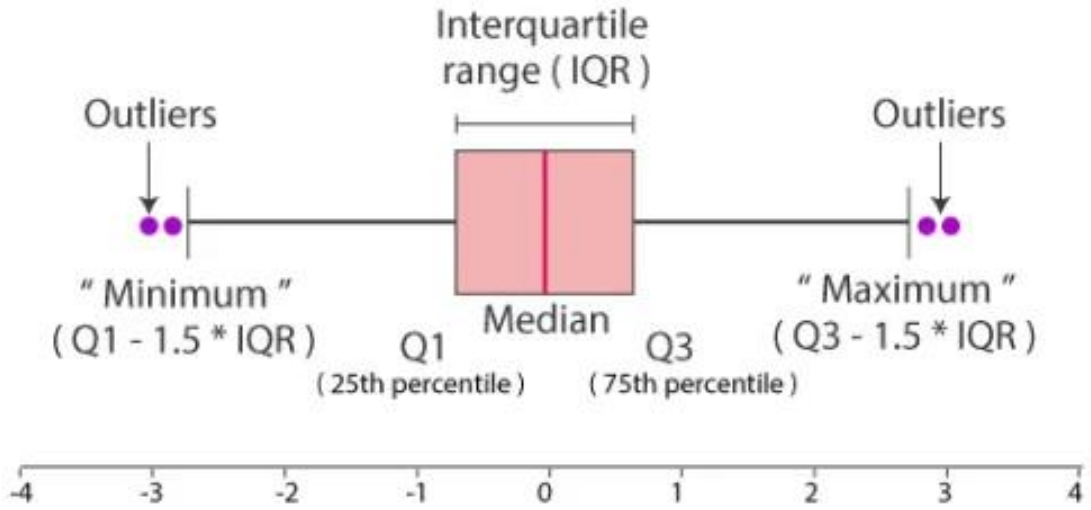
- IQR = Q3-Q1

- Minimum value = Q1–1.5 *IQR

- Maximum value = Q3+1.5*IQR

IQR technique

The data points which are lesser than the minimum value and the data points which are greater than the maximum value are treated as outliers.



Let us look at the practical implementation of this technique.

3. Percentile Method

In simple words, Percentile can be seen as the value below which a percentage of data falls. If my score is 90 percentile, then that means my score is better than 90 percent of the students who took the examination.

If you scored the maximum score, which happens to be 95, then it is 100 percentile, which means you scored more than the 100% of the students who took the examination.

Percentile Method

- This is one of the simplest techniques used to detect the outliers in a dataset. We need to just decide the threshold, Eg, if we decide 1 percentile, then it means we treat all the values above 99 percentile and below 1 percentile as outliers.

- Let us look at how we can implement this method practically.