FORMERLY Industries Helping Hands

THE iSCALE

Data Science | 30 Days of Machine Learning | Day - 17

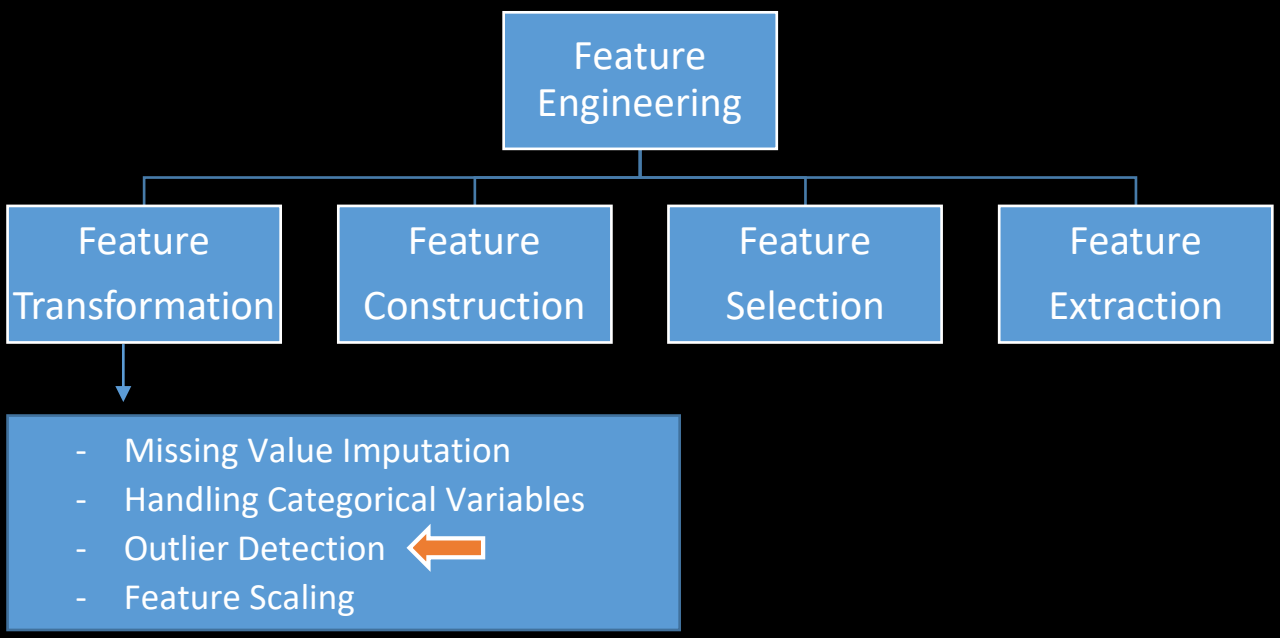
Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

----Today Topics | Day 17----

Feature Engineering (Outliers)

- **What are the outliers in machine learning?**
- **When is Outlier Unsafe?**
- **What role play anomaly detection algorithms in outliers?**
- **Effect of Outliers on ML Algorithms**
- **How to treat Outliers?**
- **How to detect Outliers?**
- **Techniques to detect & remove Outliers**

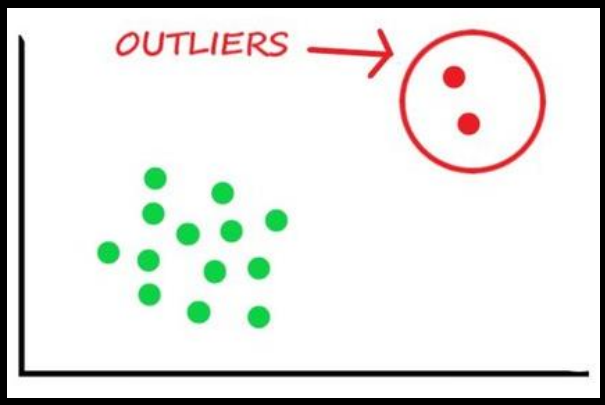
- Dataset Link GitHub: https://github.com/TheiScale/30_Days_Machine_Learning/



What are the outliers in machine learning?

In data analytics, outliers are values within a dataset that vary greatly from the others—they're either much larger, or significantly smaller.

In statistics, any observations or data points that deviate significantly and do not conform with the rest of the observation or data points in a dataset are called outliers. Outliers are extreme values in a feature or dataset.



For example, if you have a dataset with a feature height. The majority of the values in this feature range between 4.5−6.5 feet, but there is one value with 10 feet. This value would be considered an outlier, as it is not only an extreme value but an impossible height as well.

When is Outlier Unsafe?

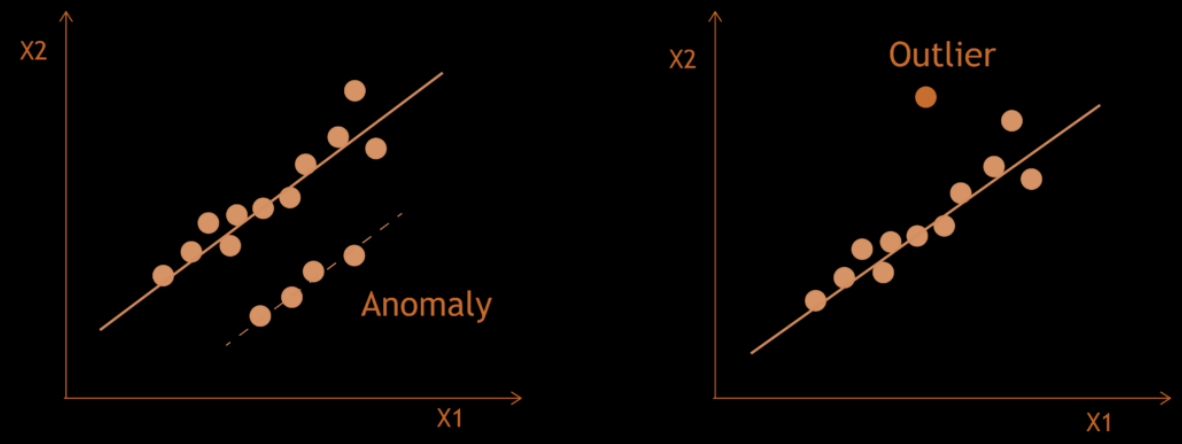
Outliers can be problematic/dangerous when they are caused by incorrect data entries. They can distort statistical measures and lead to misleading conclusions.

For example, let's say you're analysing a dataset of students' test scores in a class. Most of the scores fall within a reasonable range between 60 and 100, but there's one data entry showing a score of 1000. This outlier is clearly a mistake, as test scores typically don't exceed 100.

What role play anomaly detection algorithms in outliers?

Anomaly detection is a technique used to identify patterns in data that deviate significantly from the norm or expected behaviour. It involves detecting data points, events, or observations that are rare, unusual, or suspicious compared to the majority of the dataset.

Example: In credit card fraud detection, anomaly detection algorithms are used to identify transactions that deviate from a cardholder's typical spending behaviour. For instance, if a credit card is suddenly used for a large transaction in a foreign country where the cardholder has never travelled before, the transaction may be flagged as an anomaly and subjected to further investigation to determine if it's a fraudulent activity.



Effect of Outliers on ML Algorithms: -

Outliers disrupt weight-based algorithms like linear regression, logistic regression, and AdaBoost. They pull the model towards them, leading to inaccurate predictions (linear regression), misclassification (logistic regression), and overfitting (AdaBoost).

These algorithms aim to minimize the error or maximize the performance by assigning weights to each data point during the learning process.

How to treat Outliers?

Trimming or removing outliers completely :-One way to treat outliers is by removing them completely from the dataset. However, this approach should be used carefully because it may reduce the size of the dataset significantly if there are many outliers.

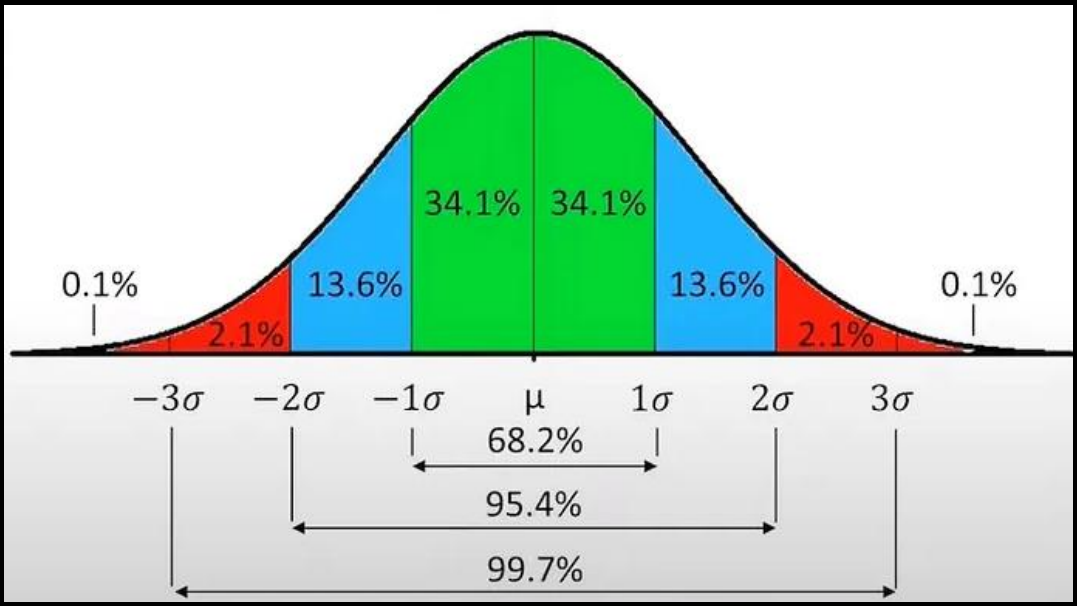
Capping: - This technique involves setting a limit or threshold on both ends of the data distribution and replacing outliers beyond those limits with the nearest non-outlier values. This approach helps in reducing the impact of extreme values without completely discarding them.

Treating outliers as missing values: - Another approach is to treat outliers as missing values and then handle them using appropriate missing data imputation techniques. This approach allows for a more flexible treatment of outliers based on the specific characteristics of the dataset.

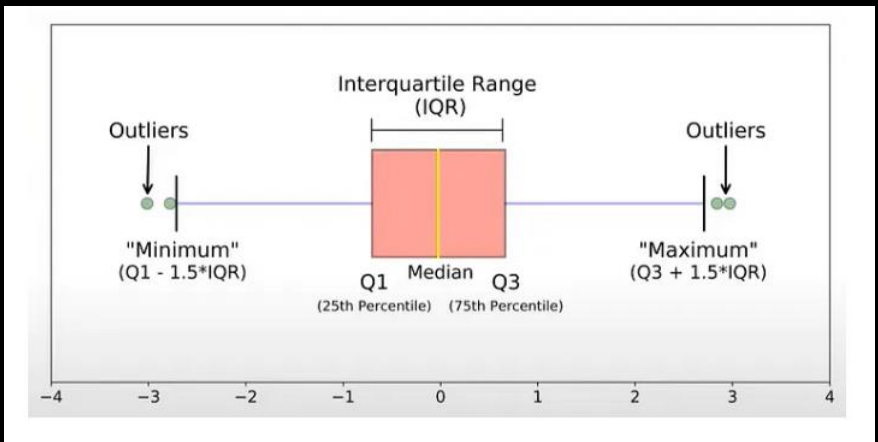
How to detect Outliers?

To detect outliers in a dataset, there are various approaches: -

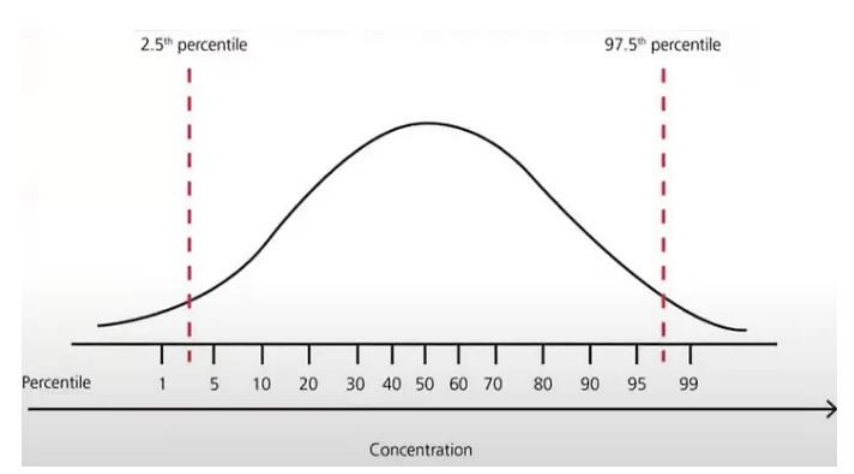
Normal distribution: - If a column follows a normal distribution, you can identify outliers by checking if a particular observation falls outside the range of mean plus or minus three standard deviations.

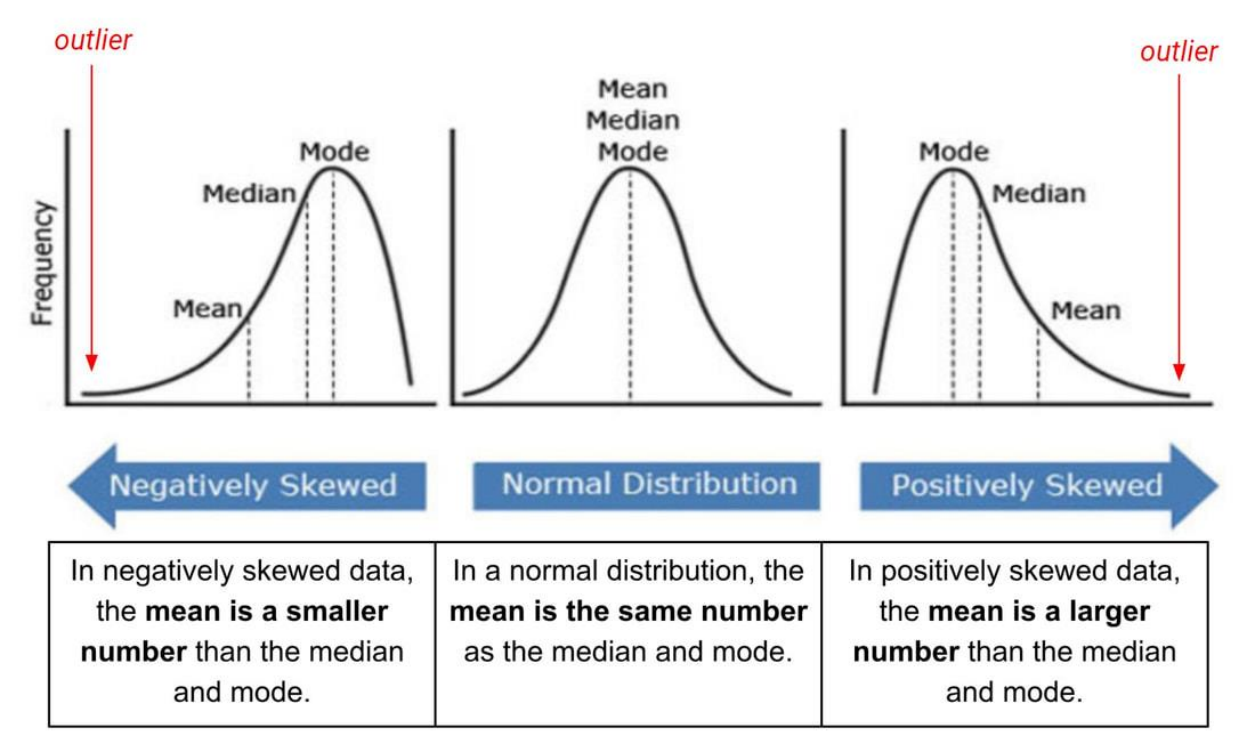


Skewed distribution: - For skewed distributions, you can use the interquartile range (IQR) proximity rule. Calculate the minimum value as (Q1–1.5 * IQR) and the maximum value as (Q3 + 1.5 * IQR). Any value that is lower than the minimum or higher than the maximum is considered an outlier.



Other distributions (percentile-based approach) :- In this approach, you can identify outliers by comparing the observations to specific percentiles. If an observation is greater than the 97.5th percentile or less than the 2.5th percentile, it is considered an outlier.





In negatively skewed data, the **mean is a smaller number** than the median and mode.	In a normal distribution, the **mean is the same number** as the median and mode.	In positively skewed data, the **mean is a larger number** than the median and mode.

Techniques to detect & remove Outliers: -

Z-score treatment: - This technique assumes that the column follows a normal distribution.

IQR (Interquartile Range) based filtering: - The IQR method involves calculating the range between the first quartile (Q1) and the third quartile (Q3).

Percentile method: - In this approach, a threshold is set based on percentiles. For example, if the threshold is set at 5%, any data point above the 95th percentile or below the 5th percentile is considered an outlier. These outliers can be removed or handled accordingly.

Winsorization:- Winsorization involves replacing outliers with values at a certain percentile, rather than removing them completely.

Day 17: Curious Data Minds

Art of Storytelling in Data Science:

- **Why Story telling is important for "Data Science" Interview?**

Storytelling is important for data science interviews because:

1. **Communication**: It helps explain complex data findings to non-technical people.
2. **Context**: It puts data analysis into a meaningful story, explaining why it's important.
3. **Engagement**: Stories capture attention and persuade people better than just data.
4. **Simplicity**: It simplifies complex data concepts, making them easier to understand.
5. **Memorability**: Stories are more memorable, ensuring that insights stick with people.

In interviews, storytelling shows you can explain data clearly, understand its context, and influence decisions—valuable skills in data science.

https://eightify.app/summary/miscellaneous/why-jeff-bezos-banned-powerpoint-at-amazon-lex-fridman-podcast-clips

- Focus point to remember in Interview

- **Tell interviewers what you know?**
- **Focus on Data Storytelling (About Projects)**

- How should be start the Data Story telling?

1. **Introduction**: Begin by introducing your project and why it's important.
2. **Hook**: Grab your interviewer's attention with an interesting fact or story related to your data.
3. **Problem Statement**: Clearly state the problem you're trying to solve or the question you're exploring.
4. **Objective**: Explain what you aim to achieve through your analysis.
5. **Preview of Insights**: Give a quick overview of the main findings or insights you'll be sharing.

- "Tell about your projects."
- "Discuss the architecture used."
- "Focus on the Life Cycle of Data Science Projects."
- "Focus on your roles in the team."
- "Tell about the challenges you faced and how you handled them."

https://towardsdatascience.com/fundamentals-of-data-architecture-to-help-data-scientists-understand-architectural-diagrams-better-7bd26de41c66