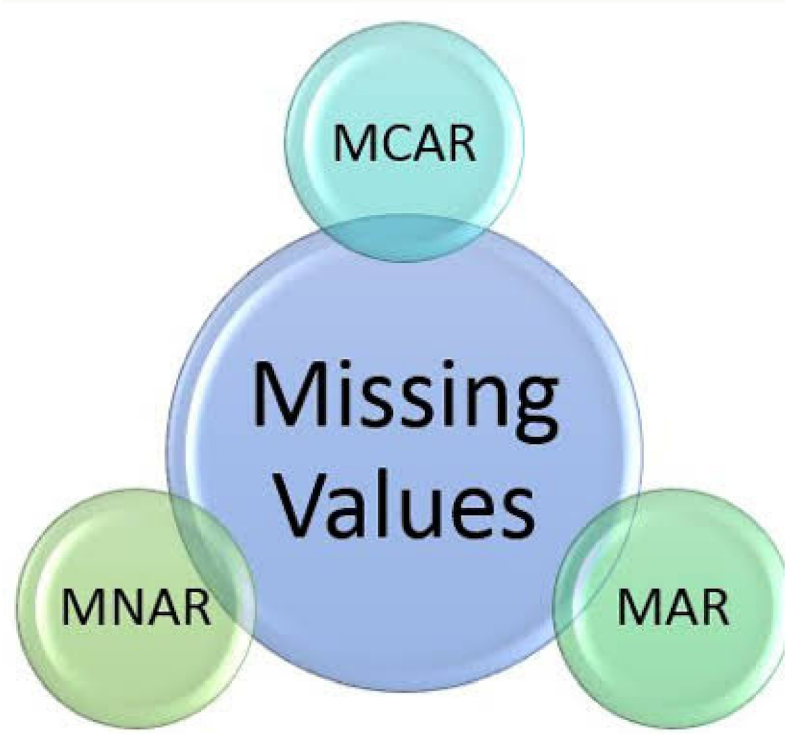
Day 2: Handling Missing Values



Author: Alisha Metkari

What Is Missing Value

- Missing value is a common problem in real world dataset. It is a datapoint that is missing(not known) in dataset.
- It arises due to incomplete data entry, error in data collection process, not answering a question due to privacy reasons, etc.
- It is represented as NaN in Pandas.
- Machine learning models don’t understand the missing/null values in data. They need to be treated before feeding data to the model.
- There are three types of missing values that we will see on next page.

Author: Alisha Metkari

Types of Missing Values

- Missing Completely At Random(MCAR): The probability of missing is same for all the observations. That is, there is no pattern and missing values are completely independent of the other data. The value can be missing due to human error, loss of data, etc.
- Missing At Random(MAR): There is some relationship between missing value and other data. The data is not missing for all observations. It is missing only within subsamples of the data and there is some pattern in the missing values.
- Missing Not At Random(MNAR): If the missing data does not fall under MCAR and MAR then it can be categorised as MNAR. Missing value depends on unobserved data.

Author: Alisha Metkari

Ways To Handle Missing Values

- Dropping features with majority of missing values.
- Dropping data points with missing values if data size is large enough.
- Imputing missing values with mean/median/mode or any other method like forward fill, backward fill.
- Predicting missing values with predictive model using rest of the data.
- Using cluster based models to identify the value of the missing data.
- Creating separate class for missing data if data is categorical.

Author: Alisha Metkari

End Notes

- Without handling missing data we cannot build ML models. So choosing the correct method is important.
- Selection of method depends on the data and the problem statement.
- Use multiple methods, evaluate the model on test data and choose the best performing method.

Author: Alisha Metkari