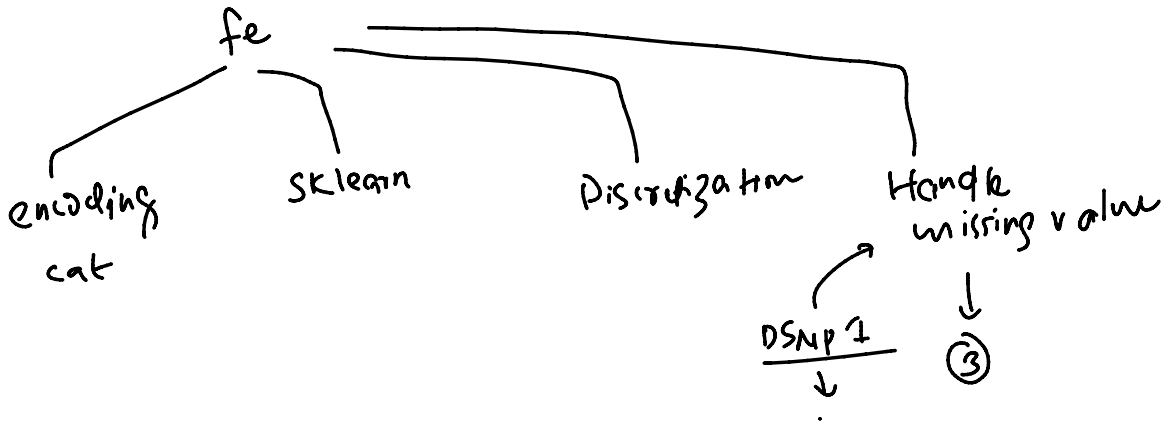
Recap

21 February 2024 19:47



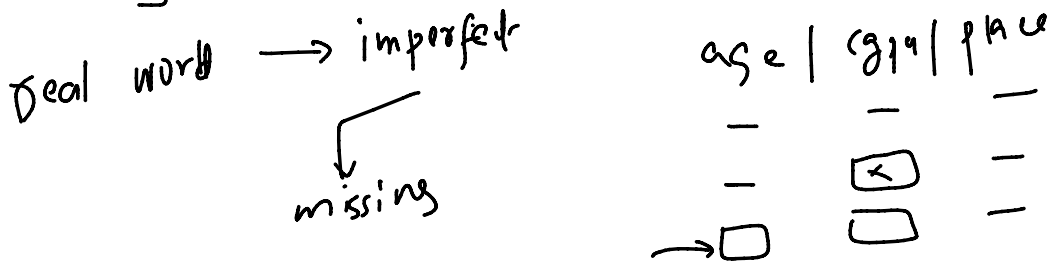
Intro

Missing Values

20 February 2024 10:42

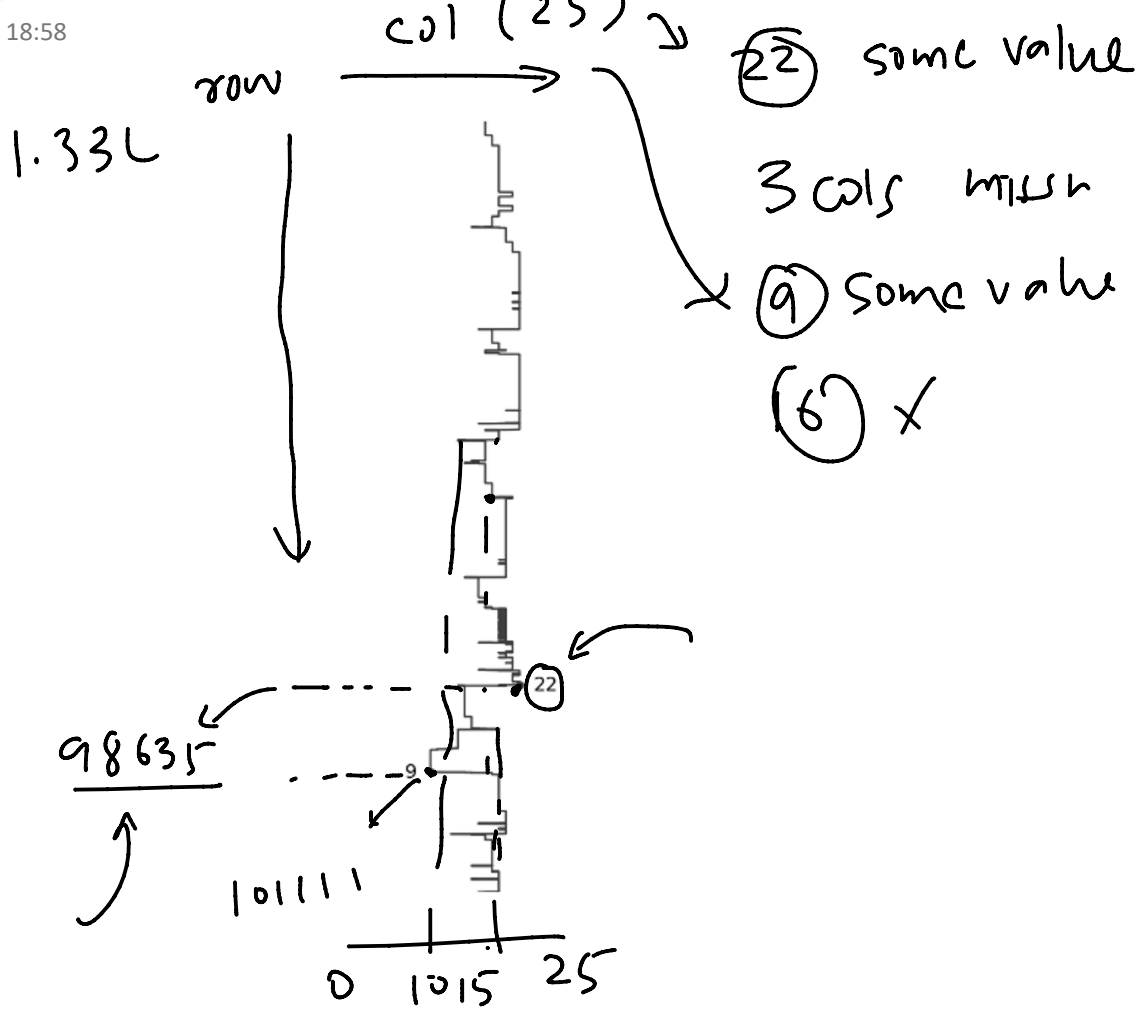
In the context of machine learning, a "missing value" refers to an instance where a data point is absent in the dataset.

How to find missing values in a dataset?



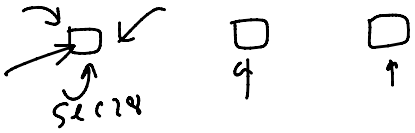
The missingo library

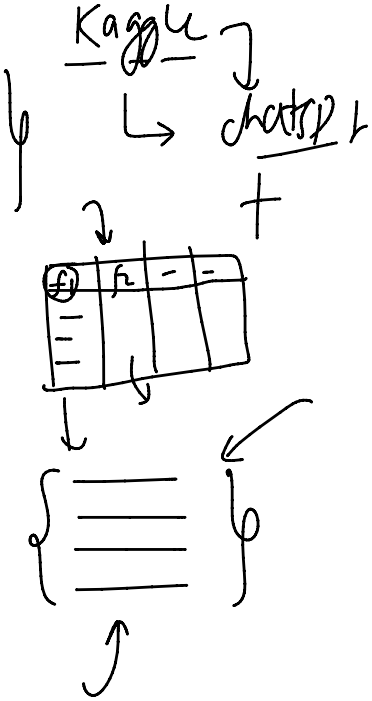
21 February 2024 18:58



Why missing values occur?

20 February 2024 11:00





1. Data Collection Issues

Example: A weather prediction model relies on data from various sensors across the country to predict weather patterns. If a sensor in a remote location malfunctions due to extreme weather conditions, temperature and humidity data from that site might be missing, creating gaps in the dataset used for prediction.

2. Data Entry Errors

Example: In a dataset for a machine learning model predicting patient health outcomes, a data entry clerk mistakenly types "NA" (not applicable) instead of "0" for "number of cigarettes smoked per day" by a non-smoker. This typo leads to missing values in a key predictive variable.

4. Privacy Concerns

Example: A dataset used for predicting credit scores is missing some values in the "previous loan amounts" field because individuals have opted out of sharing this information due to privacy concerns, leading to systematically missing data in this field.

5. Data Processing and Cleaning

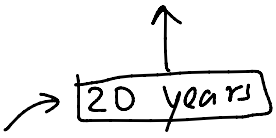
Example: During the cleaning of a dataset for a real estate price prediction model, rows containing extreme outlier values for house size are removed, assuming they are errors. If these outliers were actually correct but lacked accompanying price data, their removal could inadvertently result in missing information relevant to high-end property predictions.

6. Integration from Multiple Sources

Example: Combining hospital records from different health systems to predict patient readmission rates, you might find that some hospitals did not collect data on certain conditions or treatments. This leads to missing values in the combined dataset for those variables.

7. Evolution of Data Collection

Example: In a long-term study predicting the impact of lifestyle choices on health outcomes, the questionnaire was updated halfway through to include questions about electronic cigarette use—a source of data not available in earlier surveys, resulting in missing values for this variable in the first part of the study.

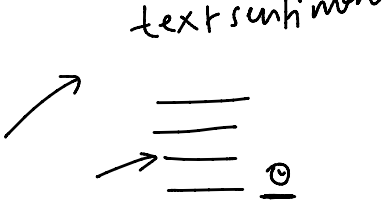


8. Censoring

Example: A model predicting time until a machine part fails is trained on data where some machines are still operational at the end of the observation period. For these machines, the "time until failure" data is censored (incomplete), as the failure event hasn't occurred yet.

9. Software Limitations

Example: A dataset being compiled from social media posts for sentiment analysis encounters issues because the scraping tool has a bug that fails to load and record posts containing certain emojis. This results in missing data specifically related to posts that could have been crucial for understanding sentiment nuances.

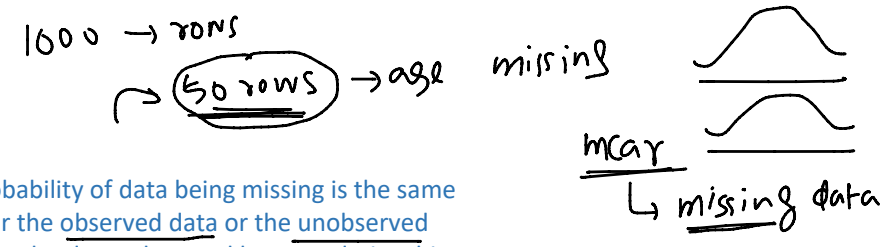


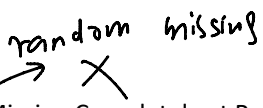
10. Resource Constraints

10. Resource Constraints

Example: In an effort to build a model predicting the spread of a rare disease, researchers could not afford to conduct genome sequencing on all available samples due to high costs. Consequently, the dataset lacks genetic information for a significant number of cases, limiting the model's predictive capabilities in genetic factors.

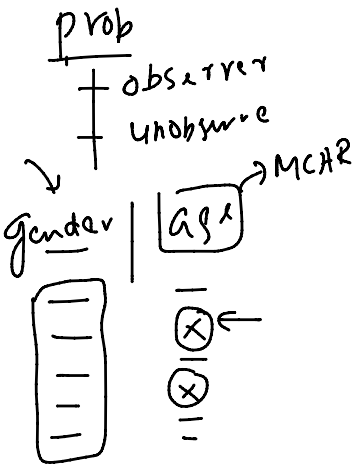
Types of missing values

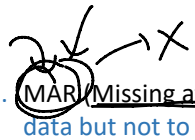




1. MCAR (Missing Completely at Random): The probability of data being missing is the same for all observations. It does not depend on either the observed data or the unobserved (missing) data. Essentially, the missingness is completely random and has no relationship with any other data, making MCAR the simplest case to handle since it introduces the least bias.

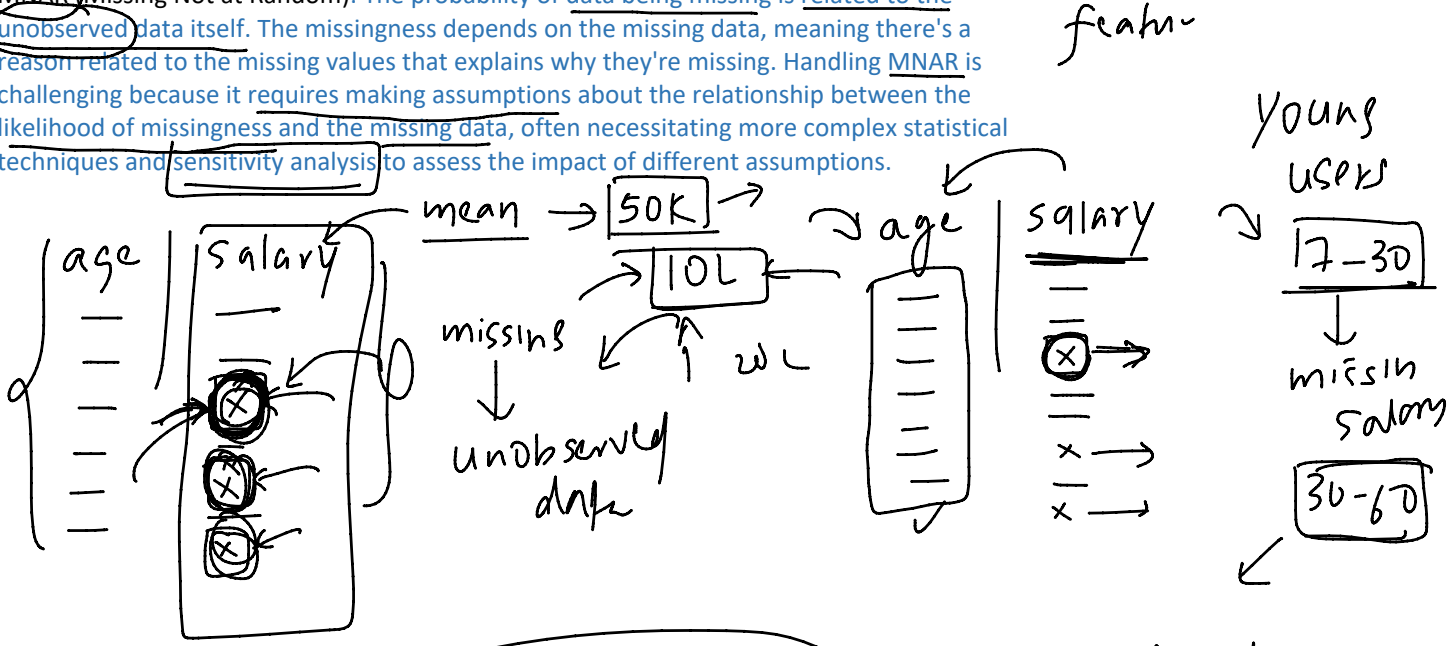
- Implications: When data are MCAR, the missing data can be considered a random subset of the dataset. Thus, analyses that exclude cases with missing data (like complete case analysis) do not introduce bias regarding the estimates of the model's parameters. However, MCAR can lead to a reduction in statistical power due to the smaller sample size.

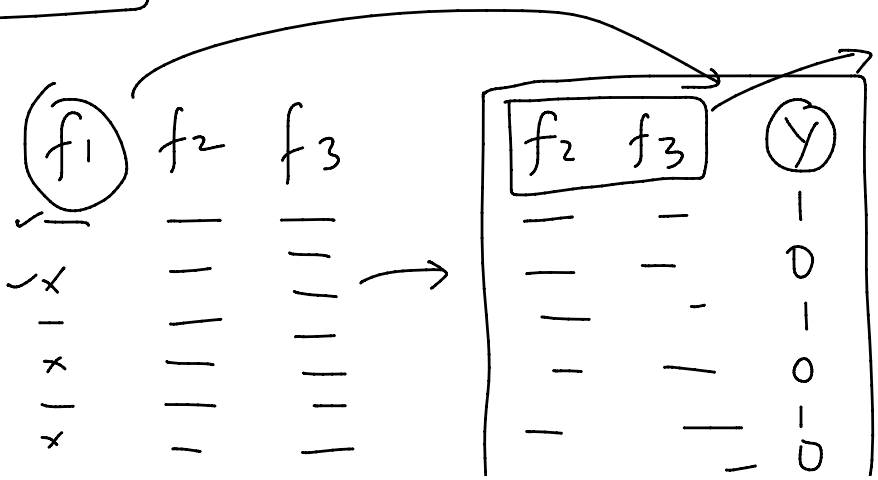


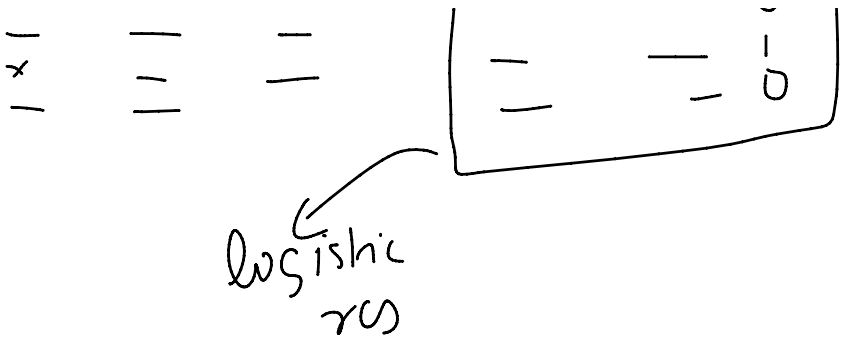


2. MAR (Missing at Random): The probability of data being missing is related to the observed data but not to the unobserved data. That is, the missingness can be explained by other variables in the dataset for which data is available. Under MAR, the missingness is systematic and can be modelled using the observed data, allowing for more sophisticated imputation methods that can reduce bias.

3. MNAR (Missing Not at Random): The probability of data being missing is related to the unobserved data itself. The missingness depends on the missing data, meaning there's a reason related to the missing values that explains why they're missing. Handling MNAR is challenging because it requires making assumptions about the relationship between the likelihood of missingness and the missing data, often necessitating more complex statistical techniques and sensitivity analysis to assess the impact of different assumptions.

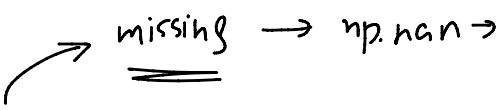





x
0
logistic
reg

How missing values impact ML models?

20 February 2024 11:00

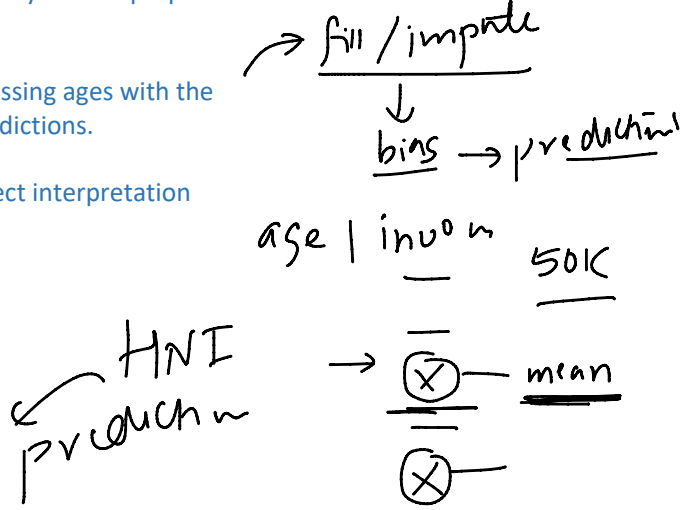


- Training Difficulties: Many ML algorithms require complete cases (rows of data without any missing values) to operate. Missing values can complicate the training process, requiring additional preprocessing steps like imputation or the use of models that can handle missing data explicitly.

- Increased Complexity: Handling missing data often requires additional coding and preprocessing steps (e.g., imputation strategies), increasing the complexity of data preparation and the risk of errors.

- Bias: Imputing missing values with incorrect assumptions (e.g., filling missing ages with the mean age) can introduce bias into the dataset, skewing the model's predictions.

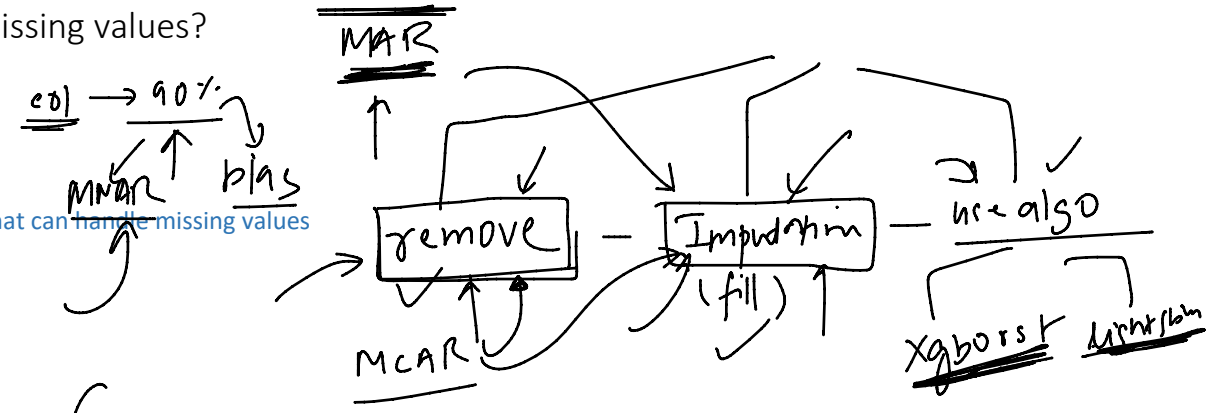
Interpretation: imputing with incorrect values might also lead to incorrect interpretation



How to handle missing values?

20 February 2024 11:00

1. Removing
2. Imputation
3. Using Algorithms that can handle missing values



How to decide?

- The Mechanism of Missingness: Understanding why data is missing (Missing Completely at Random, Missing at Random, Missing Not at Random) can guide the selection of the most appropriate imputation method.
- The Amount of Missing Data: If a feature has a very high percentage of missing values, it might be more reasonable to remove it entirely rather than impute.
- The Type of Data (Categorical vs. Continuous): Certain imputation methods are better suited to continuous data (e.g., mean imputation), while others work well with categorical data (e.g., mode imputation or using predictive models).
- The Model Being Used: Some machine learning models require complete datasets, while others have mechanisms to handle missing data directly.