Train In Data

Missing Data Imputation

Missing Data Imputation

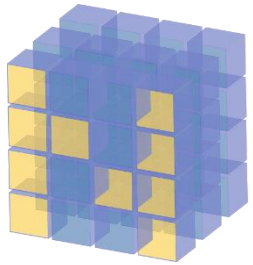
- Imputation is the act of replacing missing data with statistical estimates of the missing values.

- The goal of any imputation technique is to produce a **complete dataset** that can be used to train machine learning models.

Missing Data Imputation Techniques

Numerical Variables	Categorical Variables	Both
Mean / Median Imputation	Frequent category imputation	Complete Case Analysis
Arbitrary value imputation	Adding a “missing” category	Adding a “Missing” indicator
End of tail imputation		Random sample imputation

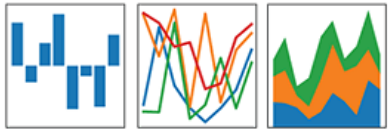
Missing Data Imputation Techniques



NumPy

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Feature-engine

Objectives

Understand the different techniques for missing data imputation.

- Learn multiple techniques
- Understand their impact on the variable and the machine learning model
- Learn how to implement it with pandas, Scikit-learn, and Feature-engine, within a machine learning pipeline

Section Structure

Three main sections:

- Learn multiple techniques (pandas and NumPy)
- Implement the technique with Scikit-learn
- Implement the technique with Feature-engine

Content

For each lecture:

- Presentation and video
- Accompanying Jupyter notebook
  - Examples of the variable characteristics in real datasets
  - Code to identify and the different variable characteristics

Final Summary

- Final article summarizing how the different variable characteristics affect the different machine learning models at the end of the section.

- Additional reading resources.

Train In Data

Train In Data

THANK YOU

www.trainindata.com