Train In Data

Frequent Category Imputation

Frequent Category imputation: definition

- **Mode** imputation consists of replacing all occurrences of missing values (NA) within a variable by the mode, or the **most frequent value**.
- Suitable numerical and categorical variables.
- In practice, we use this technique with <u>categorical</u> variables.

Mode imputation: example

Make
Ford
Ford
Fiat
BMW
Ford
Kia
Fiat
Ford
Kia

Mode = Ford

Price
Ford
Ford
Fiat
BMW
Ford
Kia
Ford
Fiat
Ford
Ford
Kia

Mode imputation: Assumptions

- Data is missing at random
- The missing observations, most likely look like the majority of the observations (aka, the mode)

Mode imputation: Advantages

- Easy to implement
- Fast way of obtaining complete datasets
- Can be integrated in production (during model deployment)

Mode imputation: Limitations

- Distortion the relation of the most frequent label with other variables within the dataset

- May lead to an over-representation of the most frequent label if there is a big number of NA

- **The higher the percentage of NA, the higher the distortions**

When to use Mode Imputation

- Data is missing completely at random
- No more than 5% of the variable contains missing data

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook
    - Frequent category imputation with pandas
    - Effect of the imputation on:
        - Variable distribution - proportions
        - Interaction with other variables - target

Train In Data

THANK YOU

www.trainindata.com