Train In Data

Missing Category Imputation

Missing Category imputation: definition

- This method consists in treating missing data as an additional label or category of the variable.
  - Missing observations are grouped in the newly created label **'Missing'**.

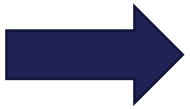
- This is the most widely used method of missing data imputation for categorical variables.

- Suitable for categorical variables

Missing Category imputation: example

Make
Ford
Ford
Fiat
BMW
Ford
Kia
Fiat
Ford
Kia

"Missing"



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

Missing Category imputation: Advantages

- Easy to implement
- Fast way of obtaining complete datasets
- Can be integrated in production (during model deployment)
- Captures the importance of "missingness" if there is one
- No assumption made on the data

Missing Category imputation: Limitations

- If the number of NA is small creating an additional category is in essence adding another rare label to the variable.

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook
  - Missing 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