Train In Data

Frequent Category Imputation

Frequent Category imputation

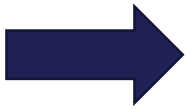
- **Mode** imputation consists of replacing all occurrences of missing values (NA) within a variable with the mode, or the **most frequent category**.

- We use this technique with categorical variables.

Example

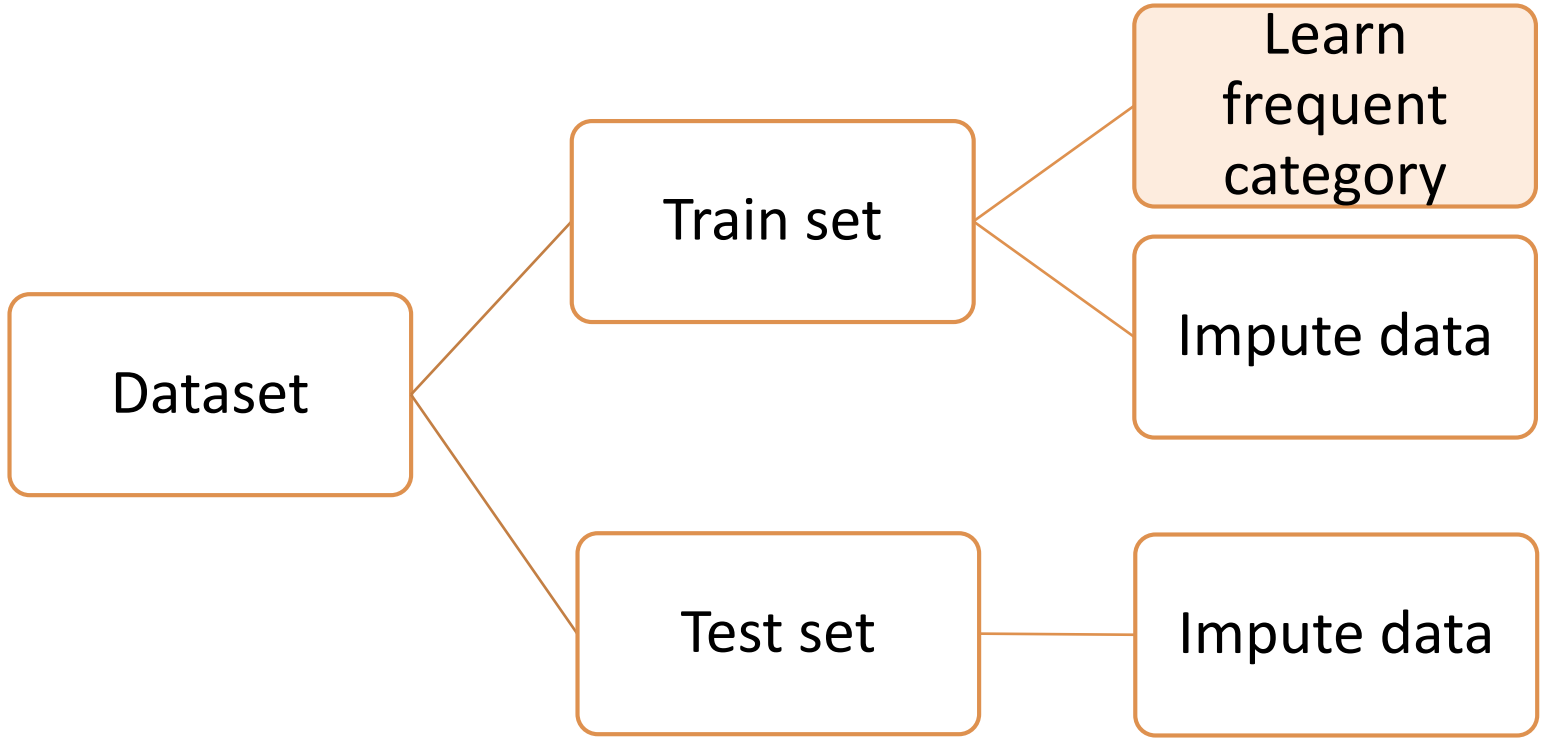
Make
Ford
Ford
Fiat
BMW
Ford
Kia
Fiat
Ford
Kia

Mode = Ford



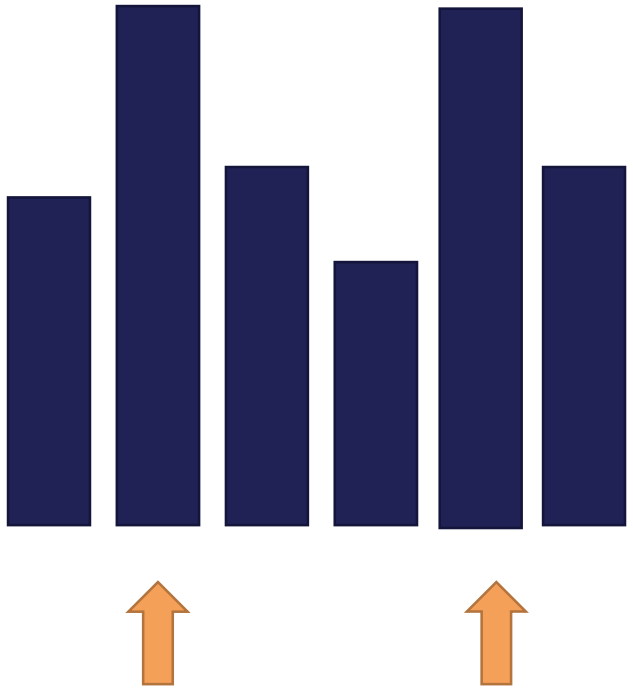
Price
Ford
Ford
Fiat
BMW
Ford
Kia
Ford
Fiat
Ford
Ford
Kia

Correct workflow



The frequent category is a "learned parameter", like the coefficients of a linear model, or the splits of a tree.

Caution



- Categorical variables can have **2 modes**.
  - (categories with equal number of observations)

- We either pick one category manually, or use arbitrary imputation.

Assumptions

- Data is missing at random
- The missing observations, most likely look like most observations
  - In categorical variables, the mode represents most observations
- Missing data are **blended** with the other values.

Good imputation strategy

Train In Data

THANK YOU

www.trainindata.com