Train In Data

Missing Indicator

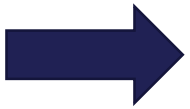
Missing indicator: definition

- A Missing Indicator is a binary variable that indicates if the data is missing for an observation (1) or not (0).
- Suitable for numerical and categorical variables

Missing indicator: example

Price
100
90
50
40
20
100
60
120
200

Missing Indicator

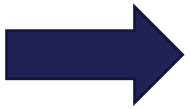


Price	MI
100	0
90	0
50	0
40	0
20	0
100	0
	1
60	0
120	0
	1
200	0

Missing indicator + Mean Imputation

Price
100
90
50
40
20
100
60
120
200

Mean = 86.66

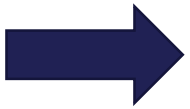


Price	MI
100	0
90	0
50	0
40	0
20	0
100	0
86.66	**1**
60	0
120	0
86.66	**1**
200	0

Missing indicator: example

Make
Ford
Ford
Fiat
BMW
Ford
Kia
Ford
BMW
Kia

Missing Indicator

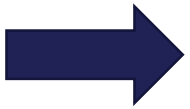


Make	MI
Ford	0
Ford	0
Fiat	0
BMW	0
Ford	0
	1
Kia	0
Ford	0
BMW	0
	1
Kia	0

Missing indicator + Frequent Category

Make
Ford
Ford
Fiat
BMW
Ford
Kia
Ford
BMW
Kia

Frequent category = Ford



Make	MI
Ford	0
Ford	0
Fiat	0
BMW	0
Ford	0
Ford	**1**
Kia	0
Ford	0
BMW	0
Ford	**1**
Kia	0

Missing indicator: use

- Missing Indicators are used together with methods that assume data is missing at random:
  - Mean, median, mode imputation
  - Random sample imputation (next section)

Missing indicator: Assumptions

- Data is NOT missing at random
- Missing data are predictive

Missing indicator - considerations

- Expands the feature space
- Original variable still needs to be imputed
- **Many missing indicators may end up being identical or very highly correlated**

Train In Data

THANK YOU

www.trainindata.com