Train In Data

Missing Indicator

Missing indicator: definition

- A Missing Indicator is an additional binary variable, which indicates whether the data was 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

- The Missing Indicator is used together with methods that assume data is missing at random:
    - Mean, median, mode imputation
    - Random sample imputation

Missing indicator: Assumptions

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

Missing indicator: Advantages

- Easy to implement
- Captures importance of missing data
- Can be integrated in production (during model deployment)

Missing indicator: Limitations

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

When to use a missing indicator

Typically, mean, median and mode imputation are done together with adding a binary "missing indicator" variable to capture those observations where the data was missing (see lecture "Missing Indicator"), thus covering 2 angles:

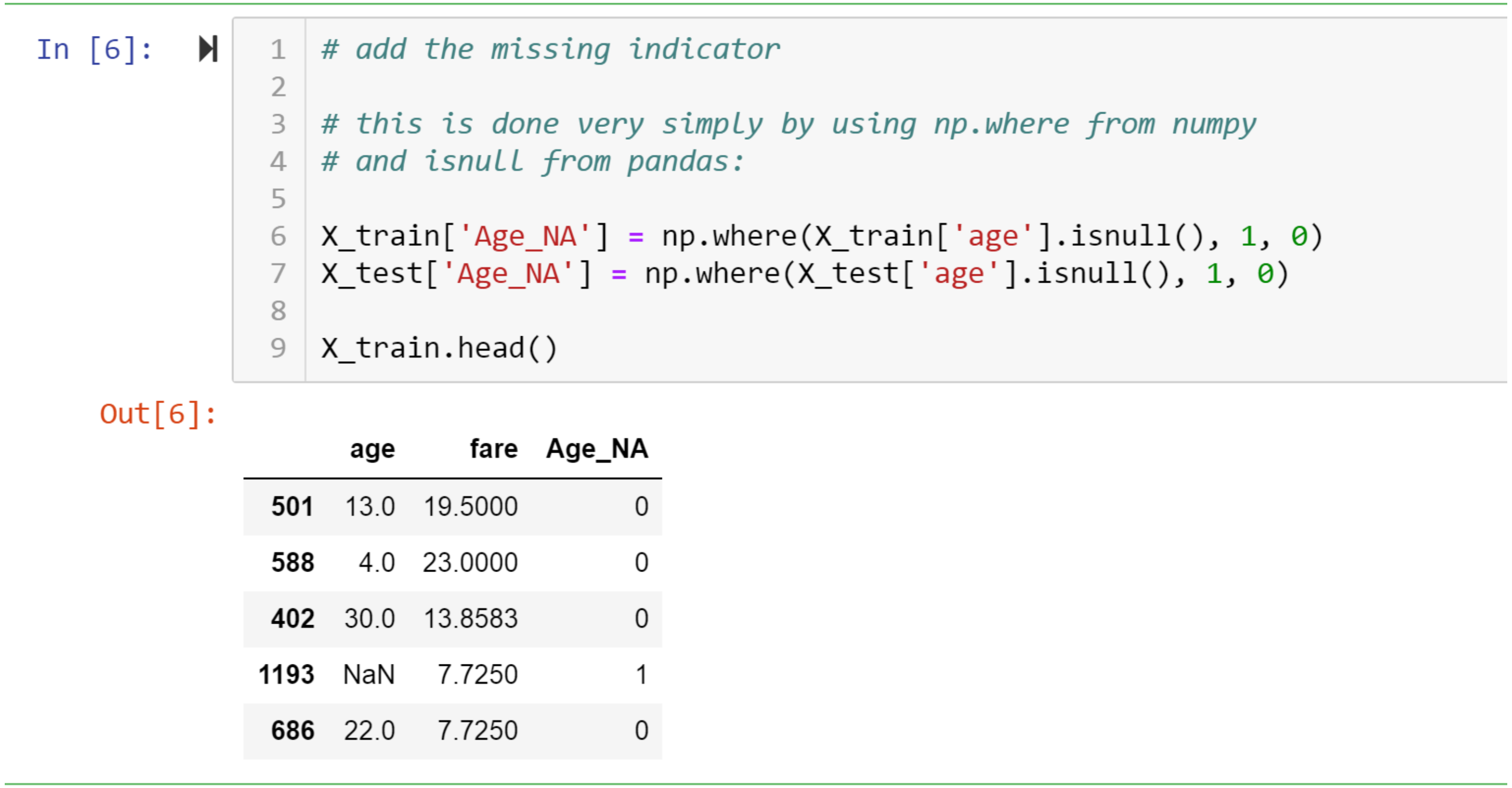
if the data was missing completely at random, this would be captured by the mean, median or mode imputation, and if it wasn't this would be captured by the additional "missing indicator" variable.

Both methods are extremely straight forward to implement, and therefore are a top choice in data science competitions.

Accompanying Jupyter Notebook

- Read the accompanying Jupyter Notebook
  - Missing indicator with pandas and NumPy
  - Followed by median imputation

Missing indicator with NumPy



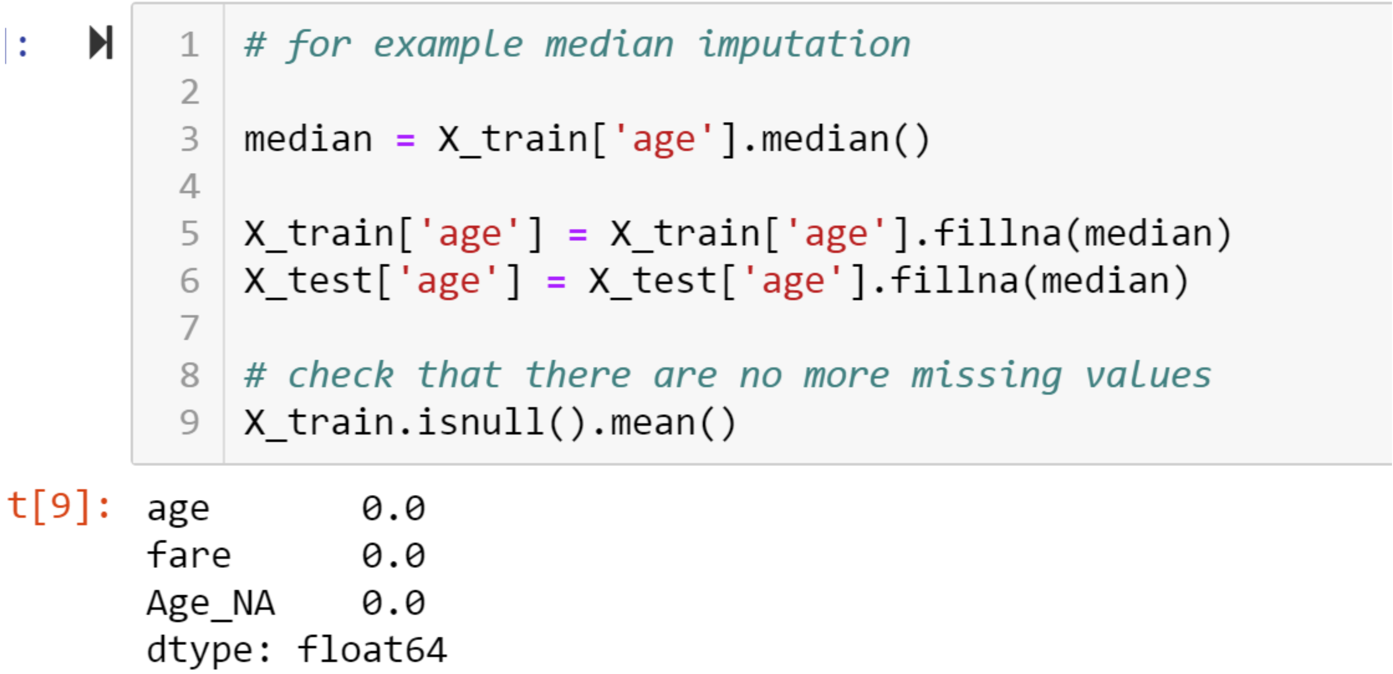
In [6]:

```python
# add the missing indicator

# this is done very simply by using np.where from numpy
# and isnull from pandas:

X_train['Age_NA'] = np.where(X_train['age'].isnull(), 1, 0)
X_test['Age_NA'] = np.where(X_test['age'].isnull(), 1, 0)

X_train.head()
```

Out[6]:

	age	fare	Age_NA
501	13.0	19.5000	0
588	4.0	23.0000	0
402	30.0	13.8583	0
1193	NaN	7.7250	1
686	22.0	7.7250	0

Missing indicator + Median imputation



```python
# for example median imputation

median = X_train['age'].median()

X_train['age'] = X_train['age'].fillna(median)
X_test['age'] = X_test['age'].fillna(median)

# check that there are no more missing values
X_train.isnull().mean()
```

```
age        0.0
fare       0.0
Age_NA     0.0
dtype: float64
```

Train In Data

THANK YOU

www.trainindata.com