Dealing with Missing Values

Preparing Data for Model Building

- Impute Missing Values
- Remove Categorical (String) Variables
- Treat Outliers
- Feature Scaling and Variable Transformation

Reasons for Missing Values

- Human Error
- Extraction Error
- Customer’s Privacy
- Other Factors

How to Identify Missing Values

gender	dependents	occupation	city	customer_nw_category
Male	0.0	self_employed	187.0	2
Male	0.0	self_employed	NaN	2
Male	0.0	salaried	146.0	2
NA	NaN	self_employed	1020.0	2
Male	2.0	self_employed	1494.0	3

How to Identify Missing Values

gender	dependents	occupation	city	customer_nw_category
Male	0.0	self_employed	187.0	2
Male	0.0	self_employed	NaN	2
Male	0.0	salaried	146.0	2
NA	NaN	self_employed	1020.0	2
Male	2.0	self_employed	1494.0	3

Methods of Treating Missing Values

- Deleting data points/rows with missing values
- Deleting features/columns with missing values
- Replacing with a new category/value
- Imputing Missing Values
    - Using central tendency
    - Using Relationship with other feature(s)
    - Using an ML model

Methods of Treating Missing Values

- **Deleting data points/rows with missing values**
- Deleting features/columns with missing values
- Replacing with a new category/value
- Imputing Missing Values
  - Using central tendency
  - Using Relationship with other feature(s)
  - Using an ML model

Delete Data points with Missing Values

Notebook

Delete Data points with Missing Values

gender	dependents	occupation	city	customer_nw_category
Male	0.0	self_employed	187.0	2
Male	0.0	self_employed	NaN	2
Male	0.0	salaried	146.0	2
NA	NaN	self_employed	NA	NA
Male	2.0	self_employed	1494.0	3

Do Not Delete Row

Can Delete Row

Analytics Vidhya
Learn everything about analytics

Methods of Treating Missing Values

- Deleting data points/rows with missing values
- **Deleting features/columns with missing values**
- Replacing with a new category/value
- Imputing Missing Values
  - Using central tendency
  - Using Relationship with other feature(s)
  - Using an ML model

Delete Columns with Missing Values

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Analytics Vidhya
Learn everything about analytics

Delete Columns with Missing Values

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

```
PassengerId    0.000000
Survived       0.000000
Pclass         0.000000
Name           0.000000
Sex            0.000000
Age            0.198653
SibSp          0.000000
Parch          0.000000
Ticket         0.000000
Fare           0.000000
Cabin          0.771044
Embarked       0.000000
dtype: float64
```

Do Not Delete Column

Can Delete Column

Don’t Delete Column

Delete Columns with Missing Values

Notebook

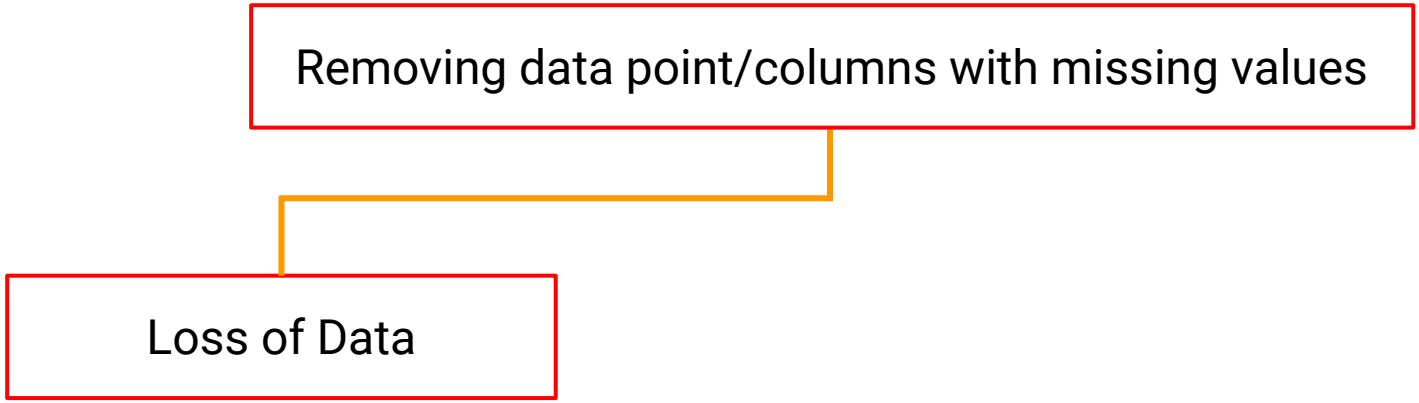
Methods of Treating Missing Values

- Deleting data points/rows with missing values
- Deleting features/columns with missing values

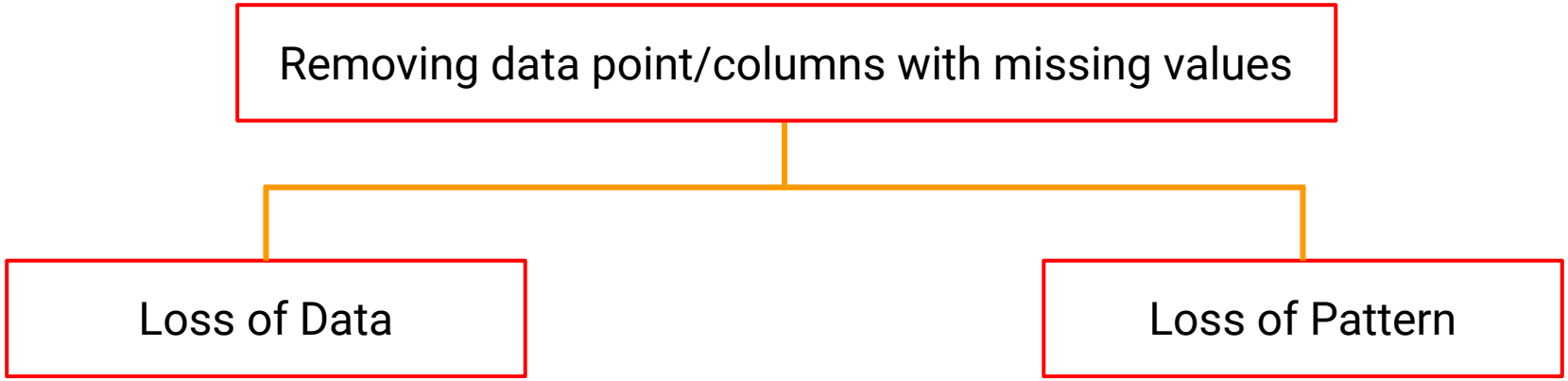
Not Recommended

- Replacing with a new category/value
- Imputing Missing Values
  - Using central tendency
  - Using Relationship with other feature(s)
  - Using an ML model

Problems with Deleting Missing Values



Problems with Deleting Missing Values



Pattern in Missing Values

Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN
Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85
Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN
Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123
Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN
Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN
McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46
Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN
Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN

Thank You

Dealing with Missing Values

Methods of Treating Missing Values

- Deleting data points/rows with missing values
- Deleting features/columns with missing values

Not Recommended

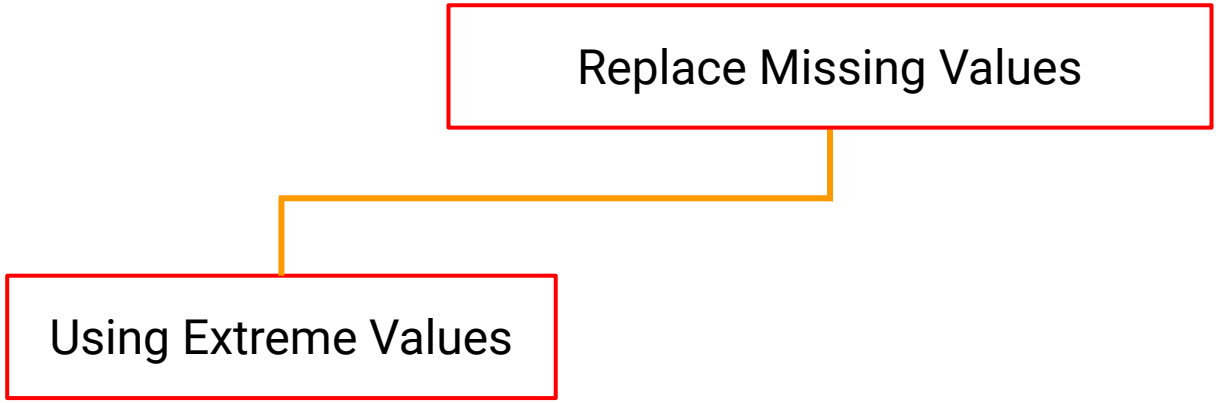
- Replacing with a new category/value
- Imputing Missing Values
  - Using central tendency
  - Using Relationship with other feature(s)
  - Using an ML model

Methods of Treating Missing Values

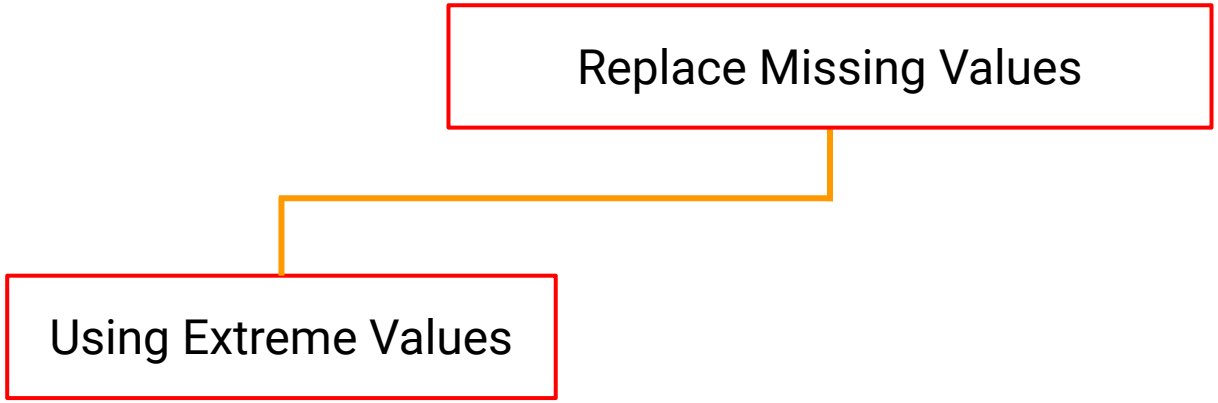
- Deleting data points/rows with missing values
- Deleting features/columns with missing values
- Replacing with a new category/value
- Imputing Missing Values
  - Using central tendency
  - Using Relationship with other feature(s)
  - Using an ML model

Replacing / Imputing Missing Values

Replace Missing Values

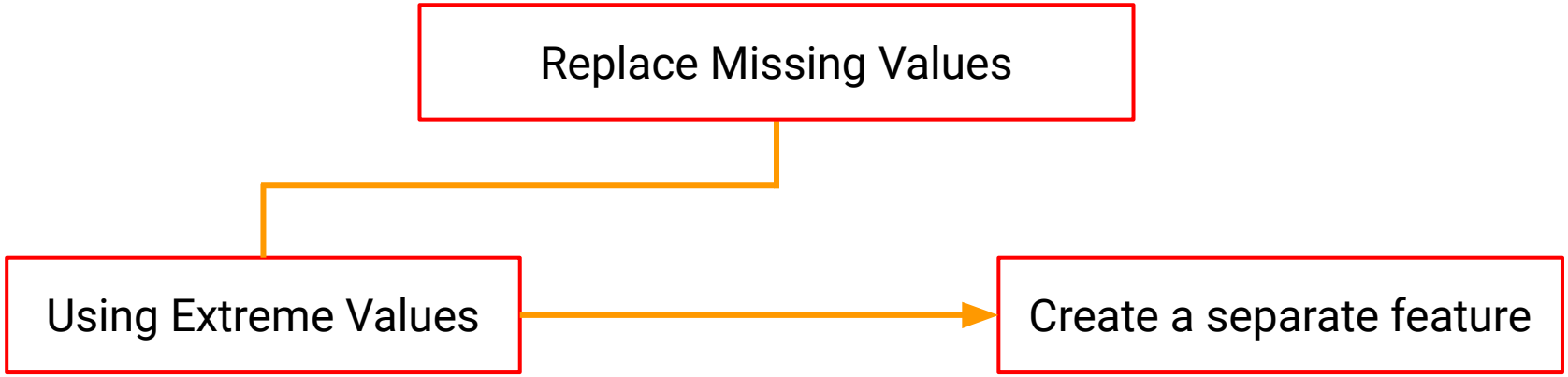


Replace Missing Values



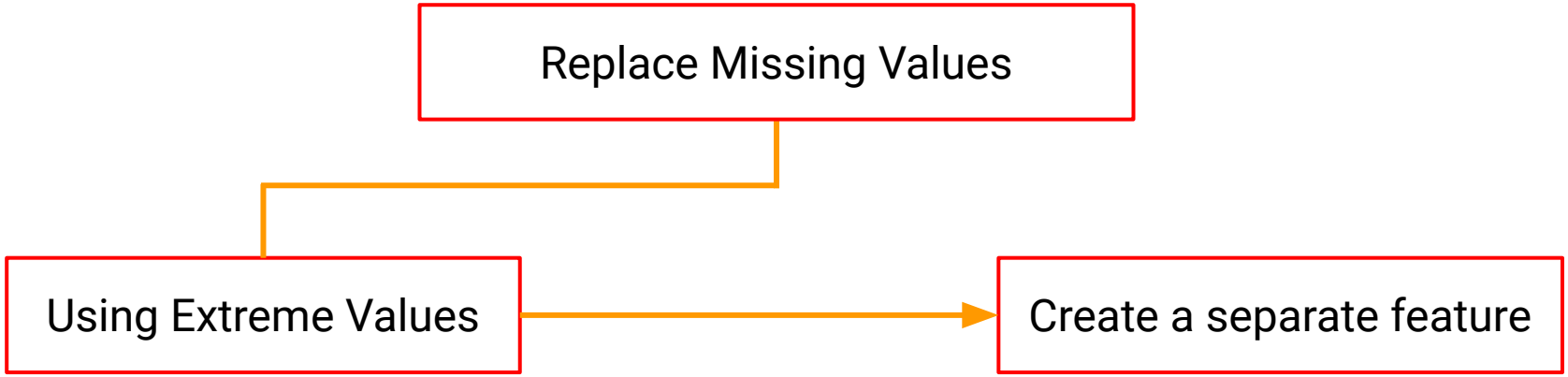
- Category as ‘Missing’ or ‘unknown’
- Numerical Values as 999

Replace Missing Values



- Category as ‘Missing’ or ‘unknown’
- Numerical Values as 999

Replace Missing Values



- Category as ‘Missing’ or ‘unknown’
- Numerical Values as 999

Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Cabin	Cabin_na
3	male	22	1	0	7.25	S	NaN	1
1	female	38	1	0	71.2833	C	C85	0
3	female	26	0	0	7.925	S	NaN	1
1	female	35	1	0	53.1	S	C123	0
3	male	35	0	0	8.05	S	NaN	1
3	male	NaN	0	0	8.458	Q	NaN	1
1	male	54	0	0	51.86	S	E46	0

Replace Missing Values with Extreme Value

Notebook

Dealing with Missing Values

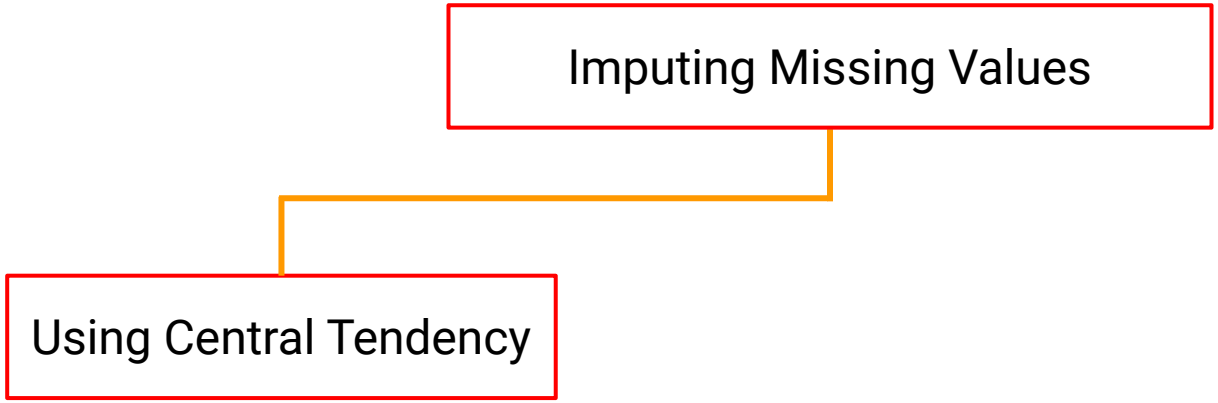
Methods of Treating Missing Values

- Deleting data points/rows with missing values
- Deleting features/columns with missing values
- Replacing with a new category/value
- Imputing Missing Values
  - Using central tendency
  - Using Relationship with other feature(s)
  - Using an ML model

Missing Values

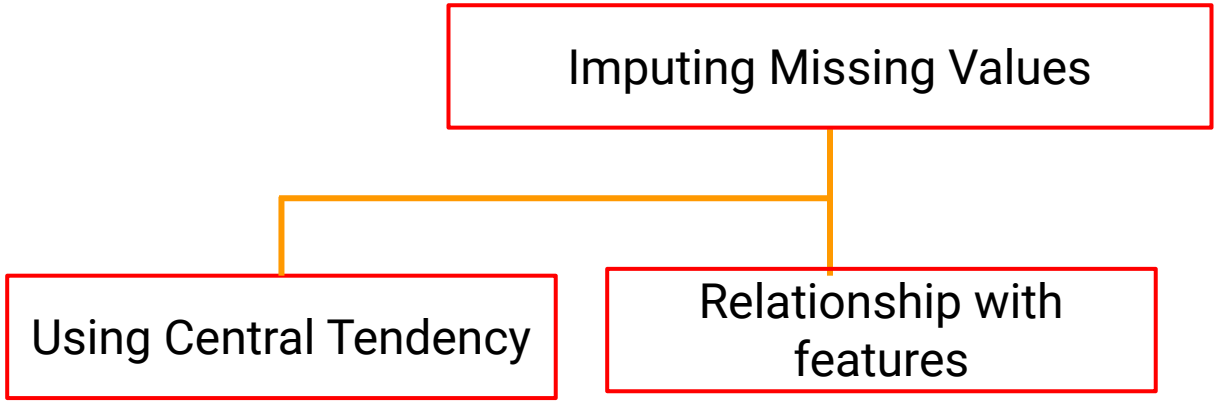
Imputing Missing Values

Missing Values



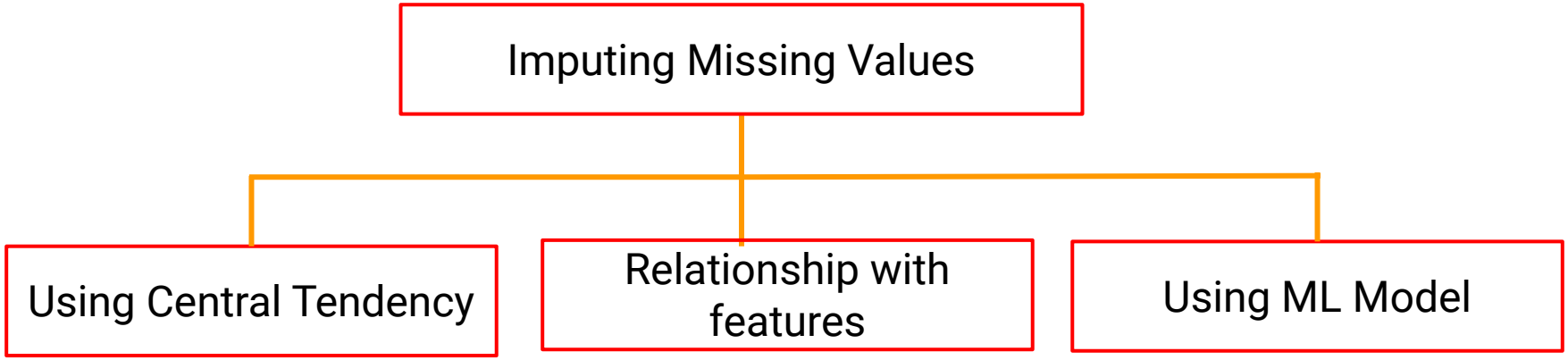
- Categorical - Mode
- Numerical - Mean, Median

Missing Values



- Using another variable with high correlation
- Pattern among features

Missing Values



- Target - Column with Nan
- Features - Other Attributes

Analytics Vidhya
Learn everything about analytics

Imputing Missing Values

Notebook

Thank You

Delete Data points with Missing Values

delete when very high number of missing

delete only columns with high missing, replacxe in others

recommended not to delete

Imputing Missing Values

ID	Age	Gender	Smoking_status
1	22	M	Never smoked
2	20	M	
3	24	F	Ocassiionally
4	32	F	Never smoked
5	21	F	
6	30	F	Smokes often
7	26	F	
8	29	M	
9	31	F	Never smoked
10	27	M	
11	25	M	