Feature engineering and selection

Feature Engineering Methods

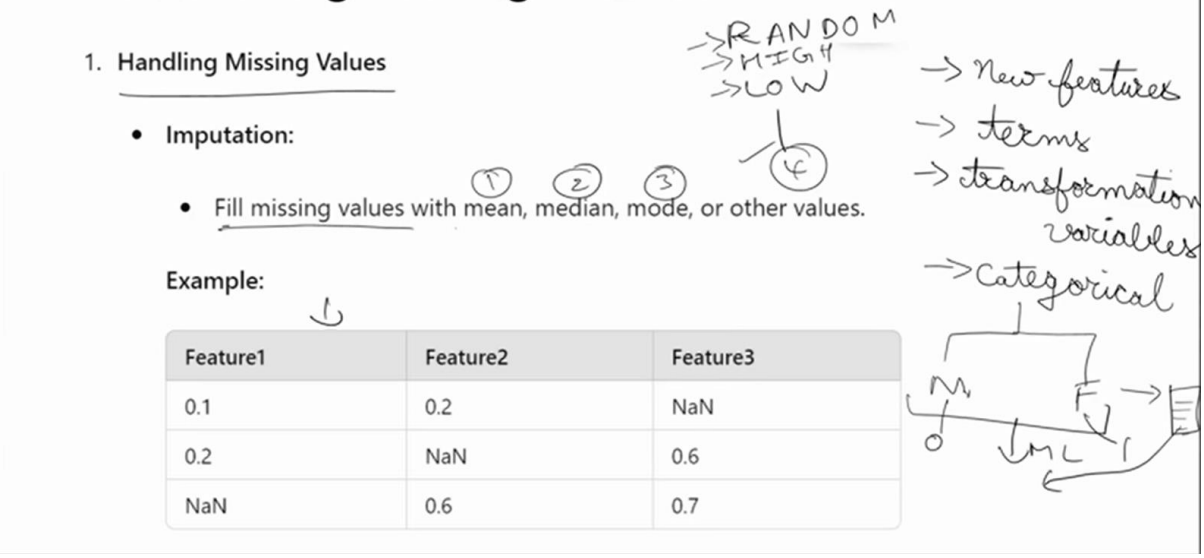
1. **Handling Missing Values**

   - **Imputation:**

     - Fill missing values with mean, median, mode, or other values.

   Example:

Feature1	Feature2	Feature3
0.1	0.2	NaN
0.2	NaN	0.6
NaN	0.6	0.7



After imputation:

Feature1	Feature2	Feature3
0.1	0.2	0.65
0.2	0.4	0.6
0.15	0.6	0.7

2. **Encoding Categorical Variables**

- **One-Hot Encoding:**
  - Convert categorical variables into a series of binary columns.

Example:

Color
Red
Blue
Green

After one-hot encoding:

Color_Red	Color_Blue	Color_Green
1	0	0
0	1	0
0	0	1

---

3. **Feature Scaling**

   - **Min-Max Scaling:**

     - Scale features to a fixed range, typically [0, 1].

     Example:

Feature1	Feature2
10	100
20	200
30	300

After min-max scaling:

Feature1	Feature2
0	0
0.5	0.5
1	1

4. **Feature Creation**

- **Polynomial Features:**
  - Create new features by taking polynomial combinations of existing features.

Example:

Feature1	Feature2
1	2
3	4
5	6

After creating polynomial features (degree=2):

Feature1	Feature2	Feature1^2	Feature2^2	Feature1*Feature2
1	2	1	4	2
3	4	9	16	12
5	6	25	36	30

Feature Selection Methods

1. Variance Thresholding

Explanation

Variance Thresholding is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e., features that have the same value in all samples.

Example Table Data

Feature1	Feature2	Feature3	Constant
1	2	3	1
1	3	4	1
1	4	5	1
1	5	6	1
1	6	7	1

In this table, 'Feature1' and 'Constant' have low or zero variance.

2. Correlation Matrix Filtering

Explanation

Correlation Matrix Filtering involves computing the correlation matrix for the features in the dataset and removing one of each pair of features with a high correlation. This helps to reduce redundancy in the data.

Example Table Data

Feature1	Feature2	Feature3	Feature4
1	2	2	5
2	4	4	6
3	6	6	7
4	8	8	8
5	10	10	9

In this table, 'Feature2' and 'Feature3' are highly correlated with 'Feature1'.

3. Domain Knowledge

Explanation

Domain knowledge involves using expertise from the specific field or industry to manually select the most relevant features. This method leverages human understanding of which features are likely to be important.

Example Table Data

Age	Salary	Height	Weight
25	50000	5.5	150
30	60000	6.0	160
35	70000	5.8	170
40	80000	5.9	180
45	90000	6.1	190

In this table, 'Age' and 'Salary' might be selected based on domain knowledge indicating their importance.

Summary

Here are three feature selection methods explained with examples:

1. **Variance Thresholding:**
    - Removes features with low variance.
    - Example removes 'Feature1' and 'Constant' due to low variance.
2. **Correlation Matrix Filtering:**
    - Computes the correlation matrix and removes highly correlated features.
    - Example removes 'Feature2' and 'Feature3' due to high correlation with 'Feature1'
    - Correlation matrix plot is provided.
3. **Domain Knowledge:**
    - Manually selects features based on expert knowledge.
    - Example selects 'Age' and 'Salary' as important features.