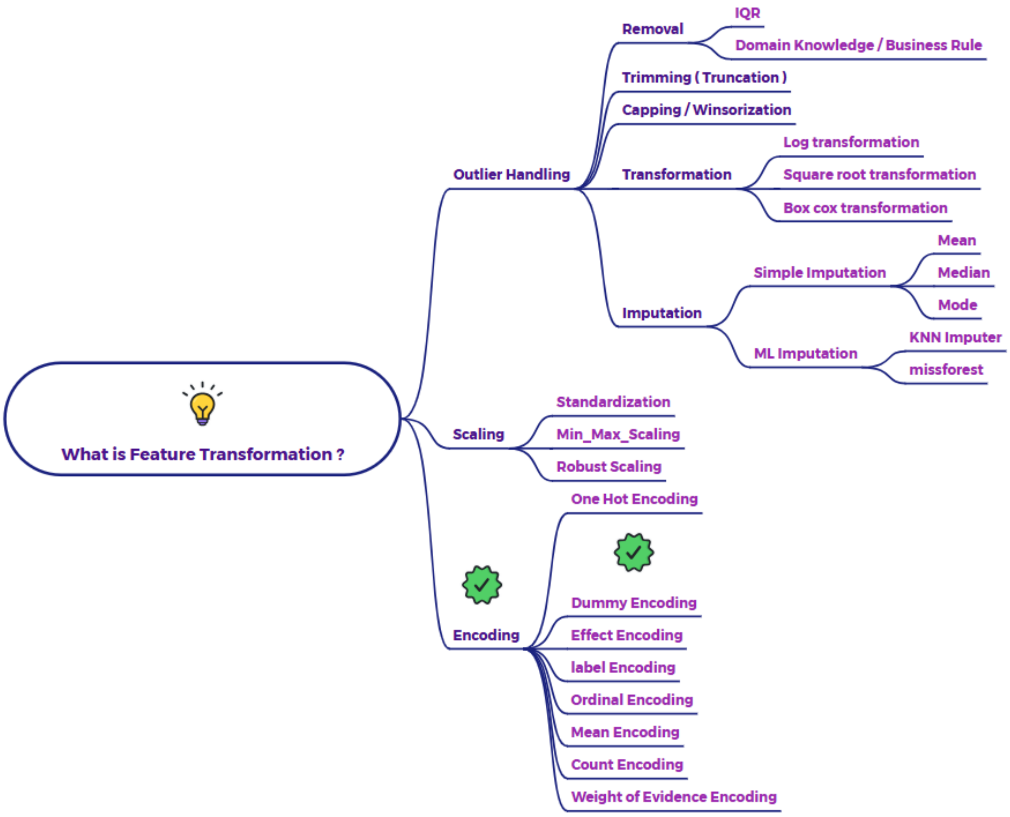
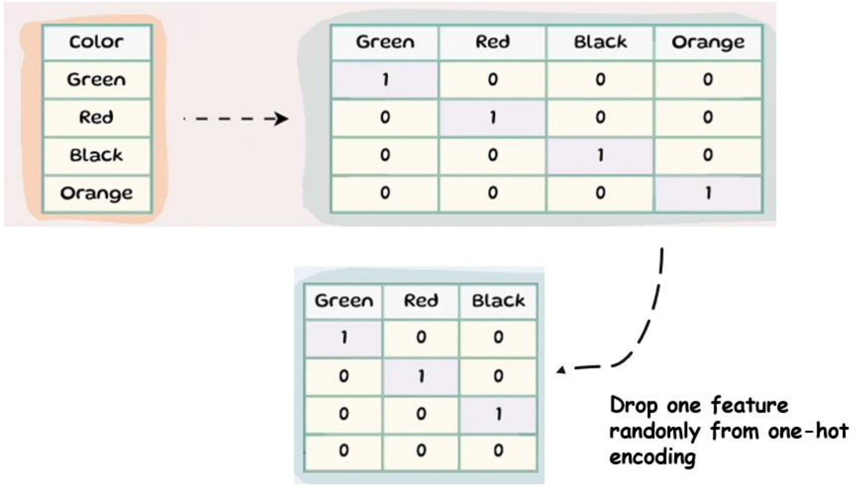
Explain Dummy Encoding with an example



1. Explanation of Dummy Encoding

Dummy encoding is a technique used to convert categorical variables into numerical values, similar to one-hot encoding. However, instead of creating a binary column for *every* category, it creates binary columns for *n-1* categories, where *n* is the total number of unique categories. This helps to avoid multicollinearity issues that can arise in some models.

Dummy Encoding



Color
Green
Red
Black
Orange

Green	Red	Black	Orange
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Green	Red	Black
1	0	0
0	1	0
0	0	1
0	0	0

Drop one feature randomly from one-hot encoding

2. How to Calculate Dummy Encoding

Here's a step-by-step explanation with an example:

Example:

Suppose we have a dataset with a "Color" column:

Color
Green
Red
Black
Orange

1. Identify Unique Categories: The unique categories in the "Color" column are "Green," "Red," "Black," and "Orange."

2. Create Binary Columns: Create binary columns for *n-1* of the categories. For instance, we might create columns for "Color_Green", "Color_Red", and "Color_Black" and drop "Color_Orange".

3. Populate Binary Columns: For each row, assign a value of 1 to the column corresponding to the row's color, and 0 to the other columns. If a row has the color that was dropped (Orange in this case), all the created binary columns will be 0 for that row.

The resulting dummy-encoded data looks like this:

Color	Color_Green	Color_Red	Color_Black
Green	1	0	0
Red	0	1	0
Black	0	0	1
Orange	0	0	0

3. When to Use Dummy Encoding

- When you want to convert categorical variables into a numerical format for use in machine learning models.
- Specifically used to avoid multicollinearity, which can be a problem in models like linear regression.

4. Strengths and Weaknesses of Dummy Encoding

- **Strengths:**
  - Avoids multicollinearity by reducing the number of created columns.
  - Provides a clear representation of categorical data.
  - Suitable for a wide range of machine learning algorithms.

- **Weaknesses:**
    - Slightly less interpretable than one-hot encoding (because the absence of all 1s across the dummy variables implies the dropped category).
    - Still increases the dimensionality of the data, though by one less column than one-hot encoding.