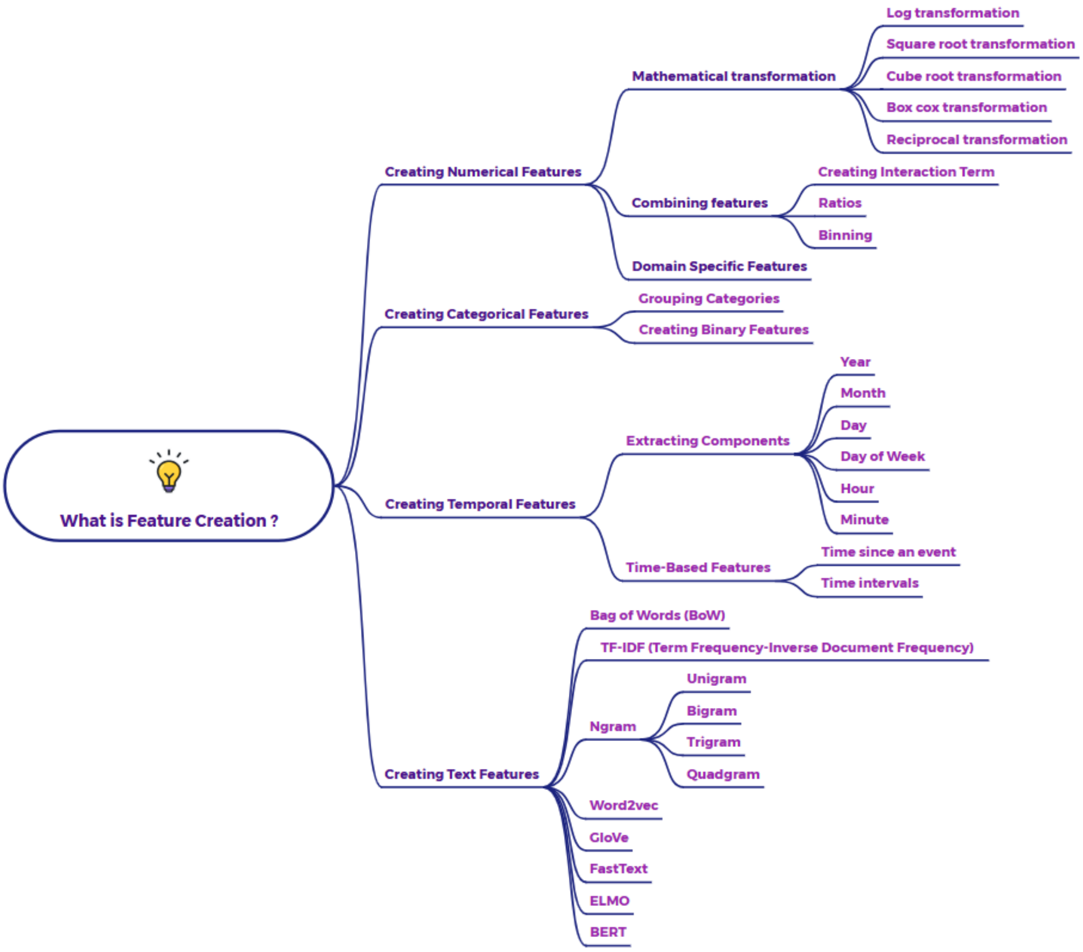
What is Feature Creation?



Feature Creation

In Feature Engineering, Feature Creation is the process of generating new features from your existing data. The goal is to derive features that can provide additional information or better represent the underlying patterns in the data, ultimately improving the performance of machine learning models.

Explanation

The illustration above highlights several ways in which new features can be created:

1. Creating Numerical Features

- **Combining Features**: This involves merging or transforming existing numerical features to create new ones.

    - **Ratios**: Creating new features by calculating the ratios between different numerical features.
    - **Creating Interaction Terms**: Generating new features by multiplying or combining two or more numerical features.
    - **Binning**: Discretizing numerical features into bins or intervals and representing them as categorical or numerical values.
- **Domain-Specific Features**: These are features created based on domain knowledge or specific understanding of the problem.

2. Creating Categorical Features

- **Grouping Categories**: Combining multiple categories of a categorical feature into fewer, broader categories.
- **Creating Binary Features**: Converting categorical features into binary (0 or 1) features, often using techniques like one-hot encoding.

3. Creating Temporal Features

- **Extracting Components**: Decomposing date and time variables into their individual components.
    - For example, from a date like "2023-11-16," you can extract "Year" (2023), "Month" (11), "Day" (16), and "Day of Week" (Thursday).
- **Time-Based Features**: Deriving new features based on time-related calculations.
    - "Time since an event": Calculating the time elapsed between a specific event and other data points.
    - "Time intervals": Creating features representing durations or time spans.

4. Creating Text Features

- Converting raw text data into numerical features that machine learning models can understand.
    - **Bag of Words (BoW)**: Representing text as the collection of its words, disregarding grammar and word order.

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: Weighing words based on their frequency in a document and their rarity across all documents.
- **N-gram**: Creating features from sequences of n words in a text.
- **Word Embeddings**: Using pre-trained models to convert words into dense numerical vectors.
    - Word2Vec, GloVe, FastText, ELMo, BERT