Feature Engineering - overview

Refining data for optimal model performance.

Bhupen

Define feature engineering

01

a process of creating new input variables (features) or transforming existing ones

02

aims to highlight relevant information, reduce noise, and address issues like missing data or outliers

03

requires a combination of domain knowledge, creativity, and a deep understanding of both the data and the machine learning algorithms being used.

Examples of feature engg

(standard numeric data)

Polynomial Features:

Generating polynomial features, like squared or cubed terms, to capture non-linear relationships in numerical data.

Log Transformation:

Applying a logarithmic transformation to skewed numeric features to make their distribution more Gaussian

Examples of feature engg (Text data)

TF-IDF Vectorization:

Converting text data into numerical vectors

Word Embeddings:

Representing words as dense vectors using techniques like Word2Vec or GloVe.

Examples of feature engg (Image data)

Color Histograms:

Extracting color distribution information from images using histograms

Edge Detection:

Applying edge detection algorithms, such as Sobel or Canny, to extract important structural information

Why should we perform feature engg?

Improved Model Performance:

capture relevant patterns and relationships within the data

Noise Reduction:

irrelevant or noisy information can be minimized, preventing the model from being influenced by irrelevant or less meaningful data.

Addressing Non-Linearity:

non-linear transformations help models handle non-linear relationships in the data

… more reasons

Handling Missing Data:

Feature engineering allows for the development of strategies to handle missing values

Enhanced Interpretability:

Thoughtful feature engineering can make models more interpretable

Optimizing Model Training:

Well-engineered features can expedite model training by providing a more focused and relevant set of input variables. This can lead to quicker convergence

two main categories

Feature Selection:

- **Definition:** Involves selecting a subset of the most relevant features from the original set.
- **Methods:** Techniques like filtering methods (e.g., correlation analysis), wrapper methods (e.g., recursive feature elimination), and embedded methods (e.g., LASSO regression).

Feature Extraction:

- **Definition:** Involves creating new features by transforming or combining existing ones.
- **Methods:** Techniques like principal component analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), autoencoders, and various transformations (e.g., polynomial features, log transformations).

Examples of feature selection methods

Relevance Filtering:

Idea: Keep only the features that seem directly related to the outcome.

Example: For predicting student success, focus on relevant features like attendance, study hours, and previous grades while excluding less relevant ones like favorite color.

Eliminating Redundancy:

Idea: Remove features that provide similar information to avoid redundancy.

Example: If both "hours spent studying" and "number of study sessions" convey similar information, consider keeping only one of them.

Keeping the Essentials:

Idea: Retain features crucial for making predictions, discarding those that contribute little.

Example: In a customer churn prediction, prioritize features such as customer tenure, satisfaction scores, and usage frequency, ignoring less influential factors.

… more examples of Feature Selection

Simplifying Complexity

- **Idea:** Reduce model complexity by focusing on a subset of features that still captures the essential patterns.
- **Example:** When predicting house prices, focus on key features like square footage, number of bedrooms, and location, avoiding overly detailed or complex variables.

Avoiding Overfitting:

- **Idea:** Remove features that may lead the model to memorize noise in the data rather than learning meaningful patterns.
- **Example:** If some features show random fluctuations and don't consistently relate to the target, excluding them helps prevent overfitting.

… more examples of Feature Selection

Enhancing Interpretability:

1. **Idea:** Keep features that are easy to understand and explain, promoting transparency.
2. **Example:** In a medical diagnosis model, focus on straightforward features like age, cholesterol levels, and blood pressure, while omitting complex biomarkers with unclear significance.

… more examples of Feature Selection

Examples of feature extraction methods

- Creating new features from existing ones to highlight important information or patterns in the data.
- It involves transforming or combining features to make them more informative or representative.
- For example, converting temperatures from Fahrenheit to Celsius or combining height and weight to create a body mass index (BMI) feature.
- Feature extraction doesn't necessarily reduce the number of features; it aims to provide a more meaningful representation.

Dimensionality reduction

- Simplifying the dataset by reducing the number of features while preserving essential information.
- It's like condensing a large book into a summary, keeping the key points intact.
- Dimensionality reduction is beneficial when dealing with too many features that might lead to computational challenges or overfitting.
- Methods like Principal Component Analysis (PCA) identify the most critical aspects and represent the data with fewer dimensions.

Popular Feature selection methods

Correlation Analysis:

Description: Measures the linear relationship between features and selects those with the highest correlation to the target variable.

Example: Pearson correlation coefficient, Spearman rank correlation coefficient.

Information Gain:

Description: Evaluates the information gain of each feature by assessing how well it predicts the target variable in a classification problem.

Example: Used in decision trees and ensemble methods.

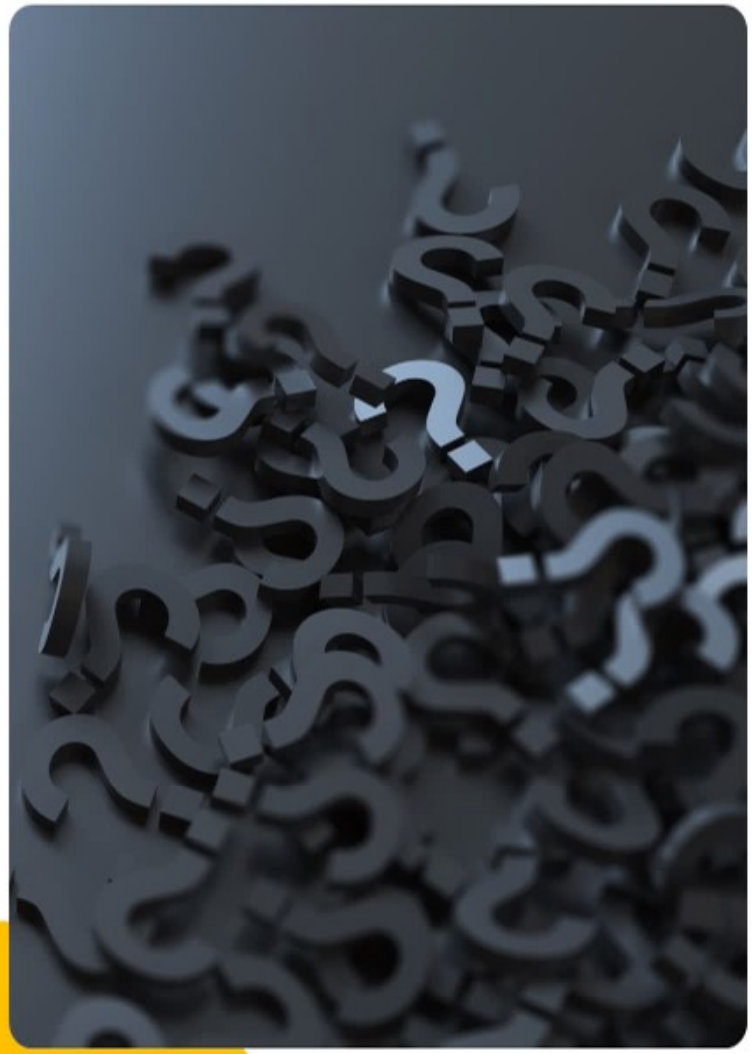
… more methods

Mutual Information:

- **Description:** Measures the dependence between two variables, providing insights into the relevance of a feature to the target variable.
- **Example:** Mutual Information feature selection.

VarianceThreshold:

- **Description:** Removes features with low variance, assuming they carry minimal information.
- **Example:** VarianceThreshold in scikit-learn.



... more methods

Univariate Feature Selection":

ANOVA F-Test:

1. **Description:** Measures the difference in means across multiple groups and selects features with significant variance.
2. **Example:** F-test for feature selection in ANOVA.

Chi-Square Test:

1. **Description:** Tests the independence of categorical variables by comparing the observed distribution with the expected distribution.
2. **Example:** Chi-square test for feature selection in categorical data.

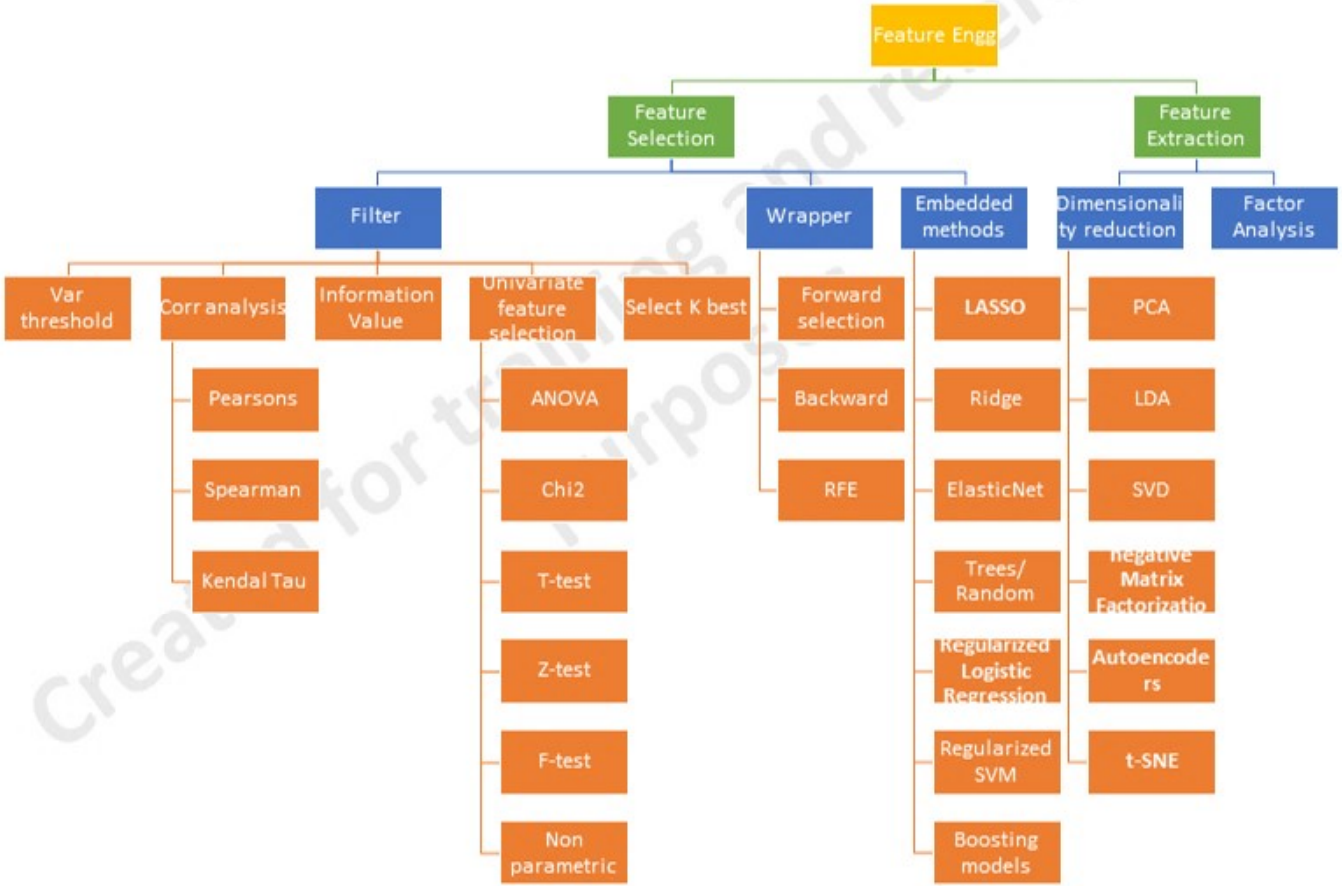
Feature engineering is challenging

Target encoding and other potent feature transformations have the potential to introduce leakage if not applied correctly.

Proficiency in domain knowledge is often crucial to understand how features interact, mitigating the risk of unintended consequences.

The process can be time-consuming, involving the execution of numerous experiments to fine-tune and validate transformations.

Taxonomy of FE methods



Thanks !!

- Next :
- Feature selection – Filter methods