Feature Engineering

Data & Variable Concepts

- **Types of Variables**: Classifying data based on its nature and mathematical properties (e.g., **numerical** like age, **categorical** like gender).
- **Variable Characteristics**: Describing the key properties of a variable, such as its **distribution**, number of unique values (**cardinality**), and presence of **missing data** or **outliers**.
- **Mixed Variables**: Variables that contain values belonging to **different data types** in the same column (e.g., a mix of numerical and text values).

Data Transformation & Preparation

- **Variable Transformation**: Applying a **mathematical function** (like a logarithm or square root) to a variable to change its **distribution** (e.g., to make it more Gaussian or reduce skew) or relationship with other variables.
- **Feature-creation**: The process of **generating new variables** (features) from existing ones to add more predictive power to a model (e.g., combining height and weight to create a BMI feature).
- **Feature-scaling**: Adjusting the **numerical range** of independent variables to a standard, comparable scale (e.g., from 0 to 1 or mean 0 and standard deviation 1). This prevents features with larger magnitudes from unfairly dominating the model.
- **Discretization-Basic**: Converting a **continuous numerical variable** into a **categorical/discrete variable** by grouping values into a limited number of ordered **bins** (e.g., turning age into "Child," "Teen," "Adult").
- **Discretization-Other**: Advanced methods for binning, often using techniques that consider the **target variable** to create optimal, non-uniform bin boundaries.
- **Outlier Engineering**: Techniques used to **detect** and **treat** extreme data points (**outliers**). Treatment can involve removing them, transforming them, or **capping/winsorizing** them to a less extreme value.
- **Date-Time Features**: Extracting and creating meaningful numerical or categorical **features** from **date and time stamps** (e.g., extracting the day of the week, month, or calculating the time elapsed since a specific event).

Missing Data Handling

- **Missing Data Imputation**: The process of **replacing missing values** (nulls or NaNs) with substituted values, such as the mean, median, or mode.
- **Imputation Alternative**: Methods for dealing with missing data that involve **not filling in** the missing values, such as deleting the rows or using models that can handle nulls natively.
- **Multivariate Imputation**: Imputation techniques that estimate missing values in a

variable by taking into account the **relationships** between the variable with missing data and **other variables** in the dataset.

Categorical Encoding

- **Categorical Encoding-Basic**: Converting **categorical variables** (text labels or strings) into a **numerical format** that machine learning models can understand, such as **One-Hot Encoding** (creating a new column for each category) or **Label Encoding** (assigning an integer to each category).
- **Categorical Encoding-Monotonic**: Encoding methods designed to convert categorical labels into numbers while maintaining a **monotonic relationship** (a consistent increasing or decreasing trend) with the **target variable**, such as Target/Mean Encoding.
- **Categorical Encoding-Rare Labels**: Methods for managing categorical variables with a large number of unique labels, many of which appear **infrequently** (rare labels), typically by grouping these low-frequency labels into a single, combined category.