Machine Learning

Feature Engineering

Missing Value Imputation

- Replace missing values with mean, median, mode, or a constant.
- Advanced methods like k-nearest neighbors (KNN) imputation or predictive models can be used.

Encoding Categorical Variables

- **Label Encoding:** Assigns an integer value to each unique category.
- **One-Hot Encoding:** Creates binary columns for each category in a variable.
- **Target Encoding:** Replaces categories with the mean of the target variable for each category

Binning

- **Discretization**: Converts continuous features into categorical by creating intervals (e.g., age groups).
- **Quantile Binning**: Divides the data into quantiles or percentiles.

Feature Scaling

- **Normalization**: Scales data to a [0,1] range, often used for algorithms like KNN.
- **Standardization**: Scales data to have a mean of 0 and a standard deviation of 1, often used for linear models.

Polynomial and Interaction Features

- **Polynomial Features**: Creates new features by raising existing features to a power (e.g., x2x^2x2, x3x^3x3).
- **Interaction Features**: Creates new features by combining two or more variables (e.g., product or sum of two features).

Log Transform and Power Transform

- **Log Transform**: Reduces skewness by applying a logarithmic transformation, often for right-skewed data.
- **Box-Cox & Yeo-Johnson**: Power transforms that reduce skewness and make data closer to normal.

Date and Time Extraction

- Extract components like year, month, day, hour, weekday, etc., from date/time features.
- Calculate elapsed time, duration, or seasonality indicators.

Text-Based Feature Engineering

- **TF-IDF** (Term Frequency-Inverse Document Frequency) and **Count Vectorization** for word frequencies.
- **Word Embeddings:** Uses pre-trained embeddings like Word2Vec or BERT to represent words as vectors.

Aggregations and Grouping

- Aggregating data by groups (e.g., average purchase amount per customer).
- Rolling and expanding functions for time-series data, like moving average or cumulative sum.

Dimensionality Reduction Techniques

- **PCA (Principal Component Analysis):** Reduces feature dimensions by finding principal components.
- **t-SNE and UMAP**: Non-linear methods often used for visualizations and clustering.

Feature Selection

- **Filter Methods**: Selects features based on statistical tests (e.g., Chi-square, ANOVA).
- **Wrapper Methods**: Uses algorithms like forward selection, backward elimination.
- **Embedded Methods**: Algorithms with built-in feature selection, like Lasso (L1 regularization).

Target Transformation

- Apply transformations to target variable (for regression problems) to stabilize variance or meet model assumptions.

Feature Importance

- Let's say you have a dataset with features like age, income, employment status, and education level to predict loan approval.
- Using feature importance methods, you find that:
  - Income and employment status are the top features, suggesting they heavily influence the loan approval out come.
  - Age and education level have lower scores, implying they are less relevant to the decision

Feature Importance

- Feature importance in machine learning refers to techniques used to assign a score to each feature (input variable) based on its usefulness in predicting the target variable.
- The higher the score, the more significant the feature is considered in making predictions.
- Methods:
  - In **Random Forest**, feature importance is often calculated as the average reduction in impurity brought by a feature across all trees in the forest.
  - In linear models (e.g., Linear Regression, Logistic Regression), feature importance can be interpreted from the magnitude of the coefficients.
  - SHAP values provide a detailed feature importance explanation by calculating the contribution of each feature to every prediction (python shap library, Explainer)