

Feature Engineering & Selection Notes

- Feature Engineering : Creating new features from existing ones to improve model performance.

* Techniques :-

- Binning : Grouping continuous values into discrete bins
- Scaling : normalizing features to a common scale
- Encoding : Converting categorical variables into numerical representations.
- Aggregating : Combining multiple features into a single feature.

* Benefits :-

- Improves model performance by capturing complex relationships
- Reduces dimensionality by removing redundant features
- Enhances interpretability by creating more meaningful features

- Computationally expensive, but guarantees ~~combinations~~ optimal solution
- Not feasible for high-dimensional data
- Recursive Feature Elimination [RFE]: recursively removes least important features.
- Faster than exhaustive search, but may not find optimal solution
- Can handle high-dimensional data, but may overfit or underfit
- Benefits :
 - Evaluates feature interactions & dependencies
 - Can handle high-dimensional data
 - Improves model performances by selecting optimal feature subset.

Exhaustive Feature Selection & Recursive Feature Elimination Notes

- Exhaustive Feature Selection: Evaluates all possible combinations of features
- Guaranteed optimal solution, but computationally

expensive.

- Not feasible for high-dimensional data
- Can be used for small datasets with few features
- Evaluates all possible feature combinations.
- Recursive Features Elimination [RFE]: Recursively removes least important feature
 - Faster than exhaustive search, but may not find optimal solution
 - Can handle high-dimensional data, but may overfit or underfit
 - Can be used for large datasets with many features
 - Evaluation features importance based on model performance.
- Key differences :
 - Computation Cost : Exhaustive feature selection is computationally expensive, while RFE is faster.
 - Optimality : Exhaustive feature selection guarantees optimal solution, while RFE may not.

not find optimal solution.

- **Feasibility**: Exhaustive Feature selection is not feasible for high-dimensional data, while RFE can handle high-dimensional data.

Titanic Dataset Features Notes

- **Family & Sibling**:
 - "Sibsp" feature represents number of siblings/spouses aboard.
 - "Parch" feature represents number of parents/children aboard.
 - Can be combined to create "Family Size" feature.
- **Ticket Price & Class**:
 - "fare" feature represents ticket price.
 - "Pclass" feature represents ~~social~~ socio-economic status [1st, 2nd, 3rd class].
 - Can be used to create new feature representing ticket price & class.

Variables in Dataset: Encoding, Embedding, Dimensionality Reduction Notes

- Reduces dimensionality by removing irrelevant features.
- Improves model interpretability by selecting most relevant features.
- Enhances model performance by reducing overfitting

Fisher Score & Wrapper Methods Notes

- Fisher Score : Evaluates feature ability to separate classes
 - Measures feature ability to distinguish between classes
 - Commonly used for classification problems
 - Handles missing values by ignoring or imputing them
 - Can be used for both binary and multiclass classification problems.
 - Evaluates feature relevance based on class separation.
- Wrapper Methods : Evaluate feature subsets using a machine learning algorithm
 - Exhaustive Feature Selection : Evaluates all possible combinations of features

- Feature Selection : selecting most relevant features to use in model training.
- * Filter-based Approaches :-
 - Information Gain : measures reduction in entropy or uncertainty
 - Evaluates feature relevance based on information gain
 - Commonly used in decision trees & random forests.
 - Chi-Square Test : Evaluates independence of feature & target
 - Tests whether feature is independent of target variable
 - Commonly used for categorical features
 - Fisher Score : Evaluates feature ability to separate classes
 - Measures feature ability to distinguish between classes
 - Commonly used for classification problems

* Benefits

- Linear Discriminant Analysis [LDA]: supervised dimensionality reduction
- Benefits: improves model performance, reduce overfitting, enhances interpretability

Principal Component Analysis, Linear Discriminant Analysis, Bagging, Boosting Notes

- Principal Components Analysis [PCA]:
 - linear transformation technique for dimensionality reduction.
 - Preserves variance in data, orthogonal components
 - Benefits: reduces dimensionality, removes multicollinearity
- Linear Discriminant Analysis [LDA]
 - Supervised dimensionality reduction ~~technique~~ technique
 - Maximizes class separation, minimizes within-class variance
 - Benefits: improves classification performances, reduces dimensionality.

• Encoding :

- Converting categorical variables into numerical representations

- Techniques : One-hot Encoding, label encoding, Binary encoding.

- Benefits : Enables machine learning algorithms to process categorical data

• Embedding :

- Representing high-dimensional data in lower-dimensional space.

- Techniques : Word2Vec, GloVe, Neural Network Embeddings

- Benefits : Captures complex relationships reduce dimensionality.

• Dimension Reduction :

- Reducing number of features in dataset while preserving information.

• Techniques :

- Principal Components Analysis [PCA] : linear transformation