An Introduction to Artificial Intelligence (AI) in Finance

Chapter 2:
Data Preprocessing

Universität
Münster

2. Data Preprocessing

1. Basic Definitions
2. Data Cleaning
3. Feature Engineering
4. Feature Selection

2.1 Basic Definitions

Data Preprocessing: Process of transforming raw data into a clean and usable format suitable for analysis and modeling in ML.

↓ Noise ↑ Model performance ↑ Generalization ability

Data cleaning: Specific steps within the data preprocessing process that focus on identifying and rectifying errors, inconsistencies, and inaccuracies in the dataset.

Feature Engineering: Specific steps within the data preprocessing process that use domain knowledge to create new input features or modify existing features from raw data.

Feature Selection: Specific steps within the data preprocessing that aim to reduce dimensionality by selecting a subset of relevant features for model construction to enhance model performance and interpretability

2.2 Data Cleaning
Addressing Data Inconsistencies

Standardizing Formats

- Numerical Variables
- Categorical Variables
- Date Variables

Removing Duplicates

Customer-ID	Customer Name	Order-ID
C-403-693-210	Jane Doe	O-821-290-299
C-403-693-210	Jane Doe	O-821-344-010
C-562-672-880	Jane Doe	O-821-555-502

2.2 Data Cleaning
Outlier Detection & Handling

Outliers: Data points that significantly differ from the rest of the dataset, often due to variability in measurement, experimental errors, or inherent variability in the data.

$$z_i = \frac{x_i - \mu}{\sigma}$$

Detection Methods:

- Visualization (e.g., box plots, scatter plots, histograms)
- Z-Score Method: How many standard deviations does a data point deviate from the mean?
- Interquartile Range (IQR) Method: Is the data point located between the first and third quartile?

Handling Methods:

- Removal
- Capping (Winsorization)
- Transformation
- Imputation

! Consider the domain context when determining whether to treat data points as outliers. Select strategies for handling outliers that align with analysis goals.

2.2 Data Cleaning
Data Transformation

Normalization: Rescaling the data to a specific range, usually [0, 1]

- ⊕ Retains relationships in the data.
- ⊖ Sensitive to outliers.

Applications: e.g., neural networks, k-nearest-neighbor (KNN) clustering

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

Standardization: Rescaling the data to have a mean of 0 and a standard deviation of 1.

- ⊕ Preserves the shape of the distribution; less affected by outliers compared to normalization.
- ⊖ May not be suitable for non-Gaussian distributions.

Applications: e.g., regression models, support-vector-machines (SVM), principal component analysis (PCA)

$$z_i = \frac{x_i - \mu}{\sigma}$$

2.2 Data Cleaning
Data Transformation

Log-Transformation: Applying the logarithm function to the data to reduce skewness and stabilize variance.

- ⊕ Reduces skewness and heteroskedasticity.
- ⊖ Only appropriate for data with exponential growth patterns or features that exhibit right-skewness.

Applications: e.g., financial data, population growth data, income data, sales data

$$z_i = log(x_i + c)$$

2.2 Data Cleaning

Missing Value Imputation: Types of Missing Data

Missing Completely At Random (MCAR): Data points are missing entirely by chance, unrelated to both the observed and unobserved data. Resulting data and model estimates are unbiased.

- Use basic or advanced imputation methods or drop observations with missing values.

Missing At Random (MAR): The probability of missing data is related to the observed data but not the missing data itself. Resulting model estimates are potentially biased.

- Use advanced imputation methods.

Missing Not At Random (MNAR): The missingness is related to the unobserved data itself. Resulting Data is inherently biased.

- Analyze the effect of the bias by conducting a sensitivity analysis, use auxiliary data to explain the missingness, and be cautious when using advanced imputation methods.

MCAR: Survey respondents might skip a question purely by accident.

MAR: Patients with higher levels of anxiety – an observed variable – may be less likely to report their medication adherence.

MNAR: People with high incomes may be less likely to report their income.

2.2 Data Cleaning

Missing Value Imputation: Basic Imputation Methods

Mean Imputation: Replaces missing values with the average of the observed values for that feature.

- \+ Simple and quick; maintains the mean of the dataset.
- − Can distort the variability of the data; not appropriate for skewed distributions.

Median Imputation: Replaces missing values with the median of the observed values.

- \+ More robust to outliers than mean imputation; preserves the median of the data.
- − Can distort the variability of the data.

Mode Imputation: Replaces missing values with the most frequently occurring value in the dataset.

- \+ Useful for categorical variables; preserves the mode.
- − May not be representative of the data distribution.

2.2 Data Cleaning

Missing Value Imputation: Advanced Imputation Methods

K-Nearest Neighbors (KNN) Imputation: Imputes missing values based on the values of k-nearest neighbors in the feature space.

- ⊕ Considers the local structure of the data; can provide more accurate imputations for complex data.
- ⊖ Computationally intensive; performance is sensitive to the choice of k.

Regression Imputation: Uses regression models to predict missing values based on other observed features.

- ⊕ Considers relationships among variables; can be tailored to data distributions.
- ⊖ Introduces additional assumptions; can lead to underestimation of variability.

! Consider the reason for missing values before choosing an imputation method. After imputing missing values, it is crucial to assess how the chosen method affects the model and its performance.

2.3 Feature Engineering
Creating New Features

Polynomial Features & Interaction Terms: Creating new features by raising existing features to a power or combining features

Applications: e.g., in regression modeling to capture non-linear relationships *(e.g., x_i^2, $x_i * x_j$)*

Lagged Variables: Creating features that represent previous time steps.

Applications: Time-series forecasting *(e.g., sales from the previous month)*

2.3 Feature Engineering
Encoding Categorical Variables

Encoding: Converting non-numerical variables into numerical formats.

One-Hot Encoding: Converts categorical variables into binary vectors.

Applications: Nominal variables.

(e.g., Color: “Red" becomes [1, 0, 0], “Blue” becomes [0, 1, 0], …)

Label Encoding: Assigns a unique integer to each category.

Applications: Ordinal variables.

(e.g., Education: “Bachelor" becomes [1], “Master" becomes [2], …)

2.3 Feature Engineering
Encoding Text Data

One-Hot-Encoding: Converting each unique word in the vocabulary into a binary vector.

Bag of Words: Representing text as an unordered collection of words.

Term Frequency-Inverse Document Frequency (TF-IDF): Measuring word importance by counting the occurrence of words within a document relative to its occurrences across documents.

$$TF - IDF = TF(t, d) * IDF(t, D)$$

$$IDF(t, D) = log\left(\frac{D}{1 + DF(d, t)}\right)$$

2.3 Feature Engineering
Encoding Text Data

Word Embeddings: Mapping words to a continuous vector space where similar words are located closer together.

"John is a data scientist. Sarah works as a data engineer." →

$$\begin{pmatrix} 0.2 \\ -0.4 \\ \dots \end{pmatrix}, \begin{pmatrix} 0.9 \\ 0.3 \\ \dots \end{pmatrix}, \begin{pmatrix} 0.1 \\ -0.1 \\ \dots \end{pmatrix}, \begin{pmatrix} -0.2 \\ -0.7 \\ \dots \end{pmatrix}, \begin{pmatrix} 0.1 \\ 0.5 \\ \dots \end{pmatrix},$$

$$\begin{pmatrix} -0.8 \\ 0.8 \\ \dots \end{pmatrix}, \begin{pmatrix} 0.7 \\ 0.4 \\ \dots \end{pmatrix}, \begin{pmatrix} 0.6 \\ 0.6 \\ \dots \end{pmatrix}, \begin{pmatrix} 0.1 \\ -0.1 \\ \dots \end{pmatrix}, \begin{pmatrix} -0.2 \\ -0.7 \\ \dots \end{pmatrix}, \begin{pmatrix} 0.2 \\ 0.4 \\ \dots \end{pmatrix}$$

Document Embeddings: Mapping entire documents to a continuous vector space where similar documents are located closer together.

2.3 Feature Engineering
Binning Continuous Variables

Binning: Converting continuous variables into categorical bins or ranges.

- \+ Reduces noise; makes models less sensitive to outliers; helps algorithms that perform better with categorical data.
- − Loss of information and granularity; choosing the right binning strategy is crucial for maintaining predictive power.

Equal-Width Binning: Dividing the range of the variable into equal-width intervals.

Equal-Frequency Binning: Dividing the data into bins such that each bin has approximately the same number of observations.

2.4 Feature Selection
Using Domain Knowledge

> **!** Understanding the context of the data should guide the selection of relevant features.

- Focus on features that are expected to have an impact based on theory, existing research, and/or expert knowledge.
- However, do not blindly rely on these hypotheses but let the data speak.

2.4 Feature Selection
Filter Methods

Filter Methods: Statistical methods used to evaluate the relevance of features (independent of any ML algorithm).

- ⊕ Fast and computationally efficient; can handle high-dimensional datasets; independent from ML model.
- ⊖ Ignore feature interactions; can lead to suboptimal feature sets; context dependent.

Common Techniques:

- Correlation Analysis (of features & target variable)
- Chi-Squared Tests (to assess independence of categorical features from target variable)

2.4 Feature Selection
Wrapper Methods

Wrapper Methods: Methods used to evaluate feature subsets based on the performance of a specific machine learning model.

- ⊕ Consider interactions between features; better performance than wrapper methods; applicable to any ML model.
- ⊖ Computationally intensive; risk of overfitting; low ability to generalize to other models.

Common Techniques:

- Forward Selection
- Backwards Elimination
- Exhaustive Feature Selection

2.4 Feature Selection
Estimating Feature Importance

Feature Importance: Quantifies the contribution of individual features when predicting the target variable in a ML model.

Common Model-Agnostic Methods:

- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley Additive Explanations (SHAP)
- Partial Dependence Plots (PDP)
- Permutation Importance

Common Model-Specific Methods:

- Regression Coefficients & Significance
- Tree-Based Feature Importance

Explainable Artificial Intelligence (XAI): *Methodologies designed to make the outputs of machine learning models more interpretable and understandable to humans (e.g., estimating feature importance, surrogate modeling, counterfactual explanations).*

2.4 Feature Selection
Embedded Methods

Embedded Methods: Methods that incorporate feature selection during the model training process.

- ⊕ Efficient and effective for high-dimensional datasets – combine advantages of both filter and wrapper methods.
- ⊖ High complexity; low ability to generalize to other models.

Common Techniques:

- Lasso Regression (L1 Regularization)
- ~~Ridge Regression (L2 Regularization)~~
- Elastic Net (L1 & L2 Regularization)
- Tree-Based Methods

$$L1 = \lambda_{L1} * \sum |\beta_j|$$

$$L2 = \lambda_{L2} * \sum \beta_j^2$$

E2 – Missing Values

Decide whether the missing values in the following scenarios are Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). Provide an explanation for your decision.

(1) A bank experiences a system failure that results in the loss of all transactions conducted on April 2, 2023, from the database. Fortunately, the account balances for all customers appear to remain accurate. The data science department is working on a project to model customer transactions using machine learning.

(2) The cyber security department discovers that the system failure was, in fact, orchestrated by a group of hackers. They suspect that the timing of the attack was intentionally chosen to conceal many suspicious and fraudulent transactions.

E2 – Missing Values

(3) A financial institute conducts a customer satisfaction survey via email. Unfortunately, only customers who opened their accounts after 2010 or who set up online banking were required to provide an email address. Furthermore, a significant number of the emails sent out have gone unanswered.

(4) Due to a system error, customers of an online bank were able to submit loan applications without providing their monthly income. The bank must now assess the creditworthiness of these applicants.