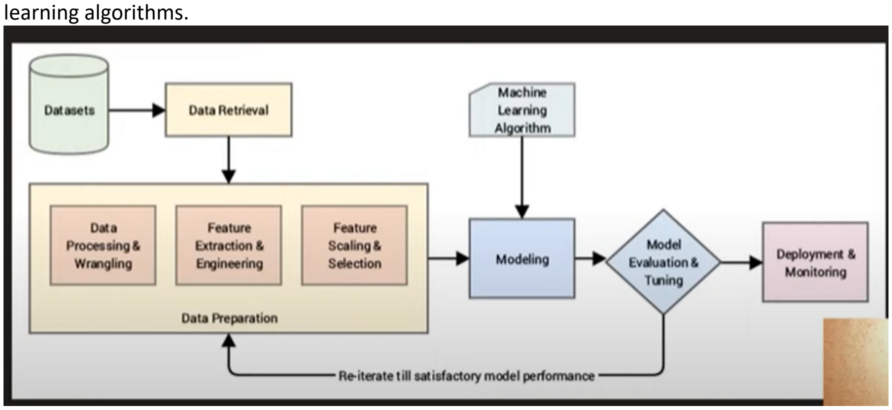
23 What is feature engineering Day 23

Introduction to Feature Engineering in Machine Learning

Understanding Feature Engineering

Definition

- **According to Wikipedia**:
  - Feature Engineering is the process of using domain knowledge to extract features from raw data that can improve the performance of machine learning algorithms.
  - 

- **Simplified Explanation**:
  - Preparing raw data into a format that machine learning algorithms can understand and learn from effectively.
  - Involves transforming, constructing, selecting, and extracting features to enhance model performance.

Importance

- **Key Points**:
  - **Critical Step**: Feature engineering is crucial in the machine learning pipeline.
  - **Impact on Models**: Better features can significantly improve model performance, even more than choosing a powerful algorithm.
  - **Analogy**: Feeding high-quality ingredients (features) into a simple recipe (algorithm) can produce better results than using poor ingredients with a complex recipe.
- **Quote**:
  - "A bad algorithm with good features can outperform a good algorithm with bad features."

The Machine Learning Pipeline

- **Overview**:
  1. **Data Gathering**: Collecting raw data.
  2. **Data Preprocessing**: Initial cleaning, handling missing values.
  3. **Exploratory Data Analysis (EDA)**: Understanding data distributions and relationships.
  4. **Feature Engineering**: Today's focus.
  5. **Model Training**: Building machine learning models.
  6. **Model Evaluation**: Assessing model performance.
  7. **Deployment**: Using the model in production.

Types of Feature Engineering Techniques

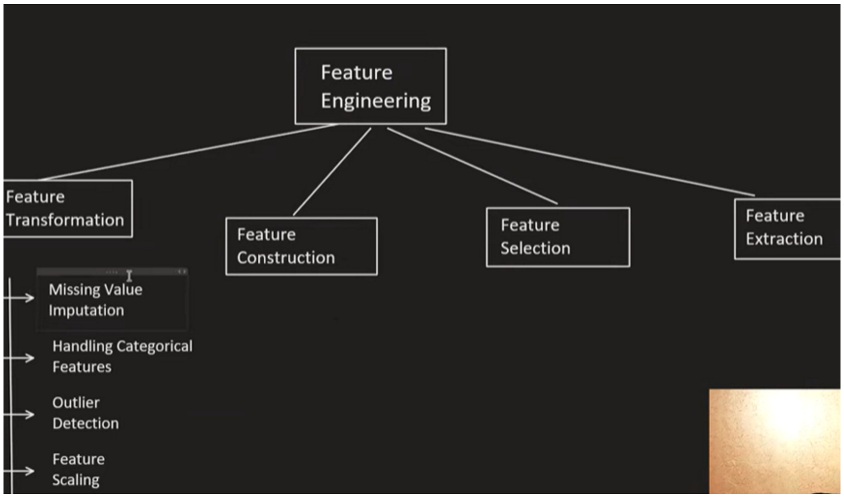
Overview

Feature engineering can be broadly classified into four categories:

1. **Feature Transformation**
   - Modifying existing features to make them suitable for modeling.
2. **Feature Construction**
   - Creating new features from existing data.
3. **Feature Selection**
   - Selecting the most relevant features for the model.
4. **Feature Extraction**
   - Reducing dimensionality by transforming data into a lower-dimensional space.

Visual Representation

- A flowchart illustrating the types and subtypes of feature engineering techniques.



1. Feature Transformation

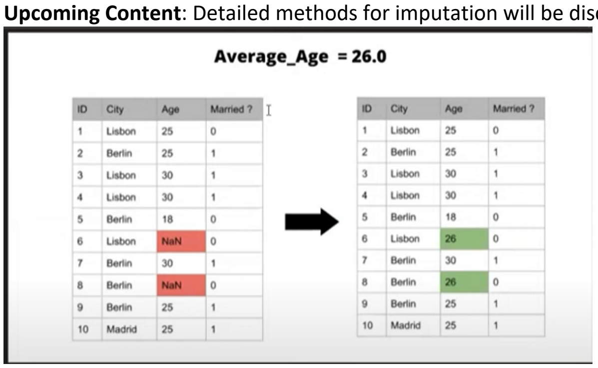
Objective

- Transform existing features into a suitable format for machine learning algorithms.

Techniques

a. Handling Missing Values (Imputation)

- **Problem**: Real-world datasets often contain missing values, which can cause issues during model training.
- **Solutions**:
  - **Removal**: If missing values are minimal, they can be removed.
  - **Imputation**: Filling missing values using:
    - **Mean**: For numerical data.
    - **Median**: When data is skewed.
    - **Mode**: For categorical data.
- **Upcoming Content**: Detailed methods for imputation will be discussed in future.

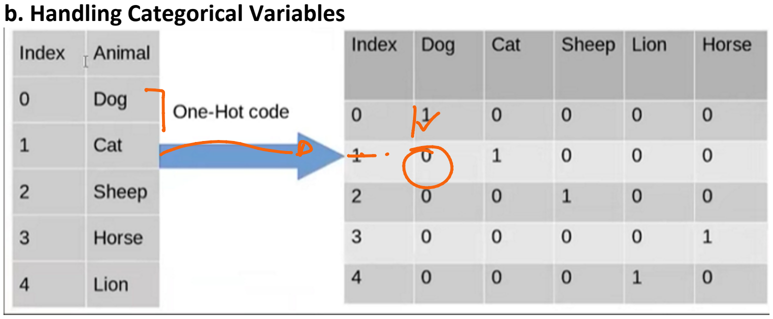


Average_Age = 26.0

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	NaN	0
7	Berlin	30	1
8	Berlin	NaN	0
9	Berlin	25	1
10	Madrid	25	1

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	26	0
7	Berlin	30	1
8	Berlin	26	0
9	Berlin	25	1
10	Madrid	25	1

b. Handling Categorical Variables



Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

- **Problem**: Machine learning algorithms require numerical input; categorical data must be converted.
- **Solutions**:
  - **Encoding Techniques**:
    - **One-Hot Encoding**:
      - Create binary columns for each category.
      - Example:
        - Categories: Dog, Cat, Sheep.
        - New columns: Is_Dog, Is_Cat, Is_Sheep.
    - **Label Encoding**:
      - Assign a unique integer to each category.
  - **Binning Numerical Variables**:
    - Convert numerical variables into categorical bins.

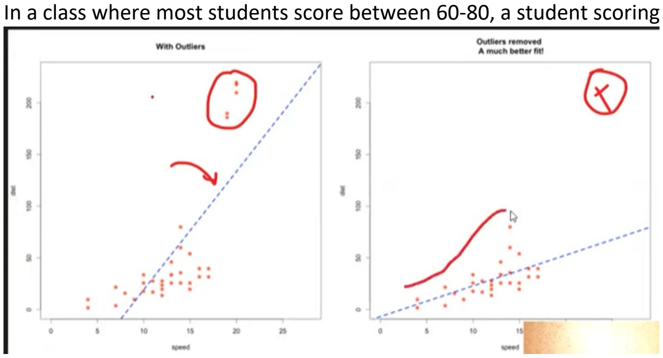
Label Encoding:

- ◻ Assign a unique integer to each category.

- ○ **Binning Numerical Variables**:
  - ▪ Convert numerical variables into categorical bins.
  - ▪ Example:
    - ◻ Age groups: 0-12 (Child), 13-19 (Teenager), 20-59 (Adult), 60+ (Senior).

c. Outlier Detection and Handling

- **Problem**: Outliers can skew data distributions and affect model performance.
- **Solutions**:
  - **Detection**: Identify outliers using statistical methods (e.g., z-scores, IQR).
  - **Removal**: Remove or cap outliers to minimize their impact.
- **Visualization**:
  - **Scatter Plots**: To visually detect outliers.
- **Analogy**:
  - In a class where most students score between 60-80, a student scoring 100 is an outlier.



d. Feature Scaling

- **Problem**: Features with different scales can bias models, especially those based on distance calculations (e.g., KNN).
- **Solutions**:
  - **Normalization**:
    - Scales data to a range of [0, 1].
    - Formula: (X - X_min) / (X_max - X_min).
  - **Standardization**:
    - Scales data to have a mean of 0 and standard deviation of 1.
    - Formula: (X - μ) / σ.
- **Importance**:
  - Ensures that no single feature dominates due to scale.
- **Analogy**:
  - Comparing the heights and weights of individuals; without scaling, weight may dominate due to larger numerical values.

	A	B	C	D
1	Country	Age	Salary	Purchased
2	France	44	72000	No
3	Spain	27	48000	Yes
4	Germany	30	54000	No
5	Spain	38	61000	No
6	Germany	40		Yes
7	France	35	58000	Yes
8	Spain		52000	No
9	France	48	79000	Yes
10	Germany	50	83000	No
11	France	37	67000	Yes

(44, 72000)

(38, 61000) $\Rightarrow (x_2 - x_1)^2 + (y_2 - y_1)^2$

this will dominate due to larger numerical values.

2. Feature Construction

Objective

- Create new features that can better capture the underlying patterns in the data.

Techniques

a. Combining Features

- **Example**:
-

P. Id	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Brigg	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmin	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

  - **Titanic Dataset**:

        - Original Features:
            - SibSp: Number of siblings/spouses aboard.
            - Parch: Number of parents/children aboard.
        - New Feature:
            - FamilySize = SibSp + Parch + 1 (including the passenger).
- **Benefit**:
    - Captures the total number of family members traveling together.

b. Creating Categorical Features from Numerical

- **Binning**:
    - Convert continuous variables into categorical bins.
    - **Example**:
        - Age bins:
            - 0-12: Child.
            - 13-19: Teenager.
            - 20-59: Adult.
            - 60+: Senior.
- **Why Use Binning**:
    - Simplifies models by reducing the effect of minor observation errors.
    - Can capture non-linear relationships.

c. Feature Interactions

- **Definition**:
    - Creating new features by combining existing ones.
- **Example**:
    - Multiplying Number of Rooms and House Age to create a new feature that captures the combined effect on housing prices.
- **Analogy**:
    - In cooking, combining ingredients in a specific way to create new flavors.

3. Feature Selection

Objective

- Identify and select the most important features that contribute to the predictive power of the model.

Importance

- **Reduces Overfitting**:
    - By eliminating irrelevant features, the model focuses on the most significant variables.
- **Improves Model Performance**:
    - Reduces computational complexity.
    - Enhances model interpretability.

Techniques

a. Filter Methods

- **Definition**:
    - Use statistical measures to select features.
- **Examples**:
    - **Correlation Threshold**:
        - Remove features with low correlation to the target variable.
    - **Chi-Squared Test**:
        - For categorical variables.

b. Wrapper Methods

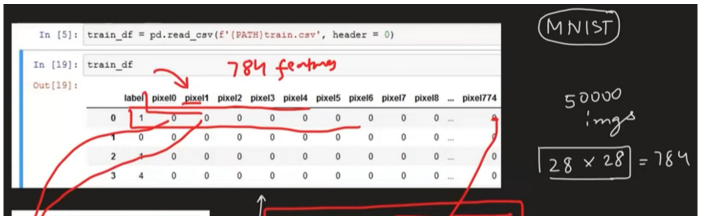
- **Definition**:
    - Use a predictive model to evaluate combinations of features.
- **Examples**:
    - **Recursive Feature Elimination (RFE)**:
        - Iteratively removes least important features.
    - **Forward/Backward Selection**:
        - Adds/removes features based on model performance.

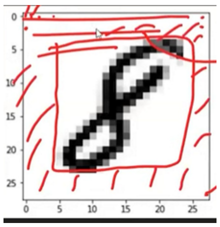
c. Embedded Methods

- **Definition**:
    - Feature selection occurs during model training.
- **Examples**:
    - **LASSO Regression**:
        - Uses L1 regularization to penalize less important features.
    - **Decision Trees**:
        - Inherently perform feature selection.

d. Example in Image Data

- **MNIST Dataset**:
    - Contains images of handwritten digits.

   - 

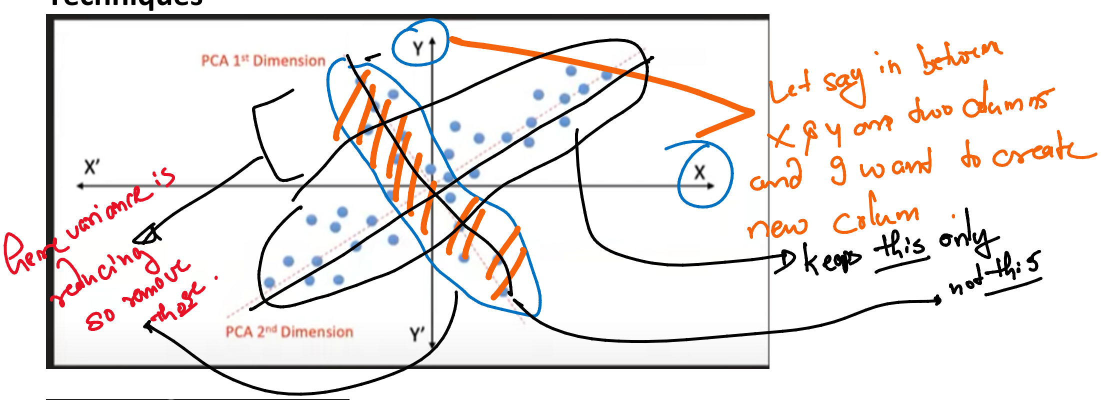
   - 
   - **Problem**:
     - Each image has 784 pixels (features), making the dataset high-dimensional.
   - **Solution**:
     - Use feature selection to identify the most informative pixels.
     - Focus on central pixels where the digit is likely to be, ignoring background pixels.

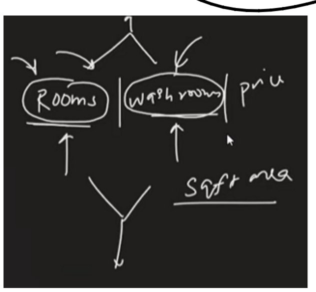
4. Feature Extraction

Objective

- Transform the data into a lower-dimensional space while retaining most of the information.

Techniques





a. Principal Component Analysis (PCA)

- **Definition**:
  - An unsupervised technique that transforms the data into a set of linearly uncorrelated variables called principal components.
- **Benefits**:
  - Reduces dimensionality.
  - Removes multicollinearity.
- **Analogy**:
  - Like summarizing a large book into key points.

b. Linear Discriminant Analysis (LDA)

- **Definition**:
  - A supervised method that projects data onto a lower-dimensional space maximizing class separability.
- **Use Case**:
  - Often used in classification tasks.

c. t-Distributed Stochastic Neighbor Embedding (t-SNE)

- **Definition**:
  - A non-linear technique primarily used for data visualization in 2D or 3D space.

- **Benefit**:
    - Captures complex relationships in data.

Importance

- **Reduces Computational Cost**:
    - Simplifies models by reducing the number of features.
- **Enhances Visualization**:
    - Easier to visualize and interpret data in lower dimensions.

The Art of Feature Engineering

Key Points

- **Creativity and Intuition**:
    - Requires understanding of the domain and the data.
- **No One-Size-Fits-All**:
    - Techniques vary depending on the dataset and problem.
- **Iterative Process**:
    - Often involves experimenting with different techniques and evaluating their impact.

Analogy

- **Cooking**:
    - Just as chefs experiment with ingredients to create a delicious dish, data scientists experiment with features to build effective models.

Conclusion

Recap

- **Feature Engineering**:
    - A crucial step in the machine learning pipeline.
    - Involves transforming, constructing, selecting, and extracting features.
- **Importance**:
    - Directly impacts model performance.
    - Helps in handling real-world data issues.

Next Steps

- **Upcoming Videos**:
    - We will explore each feature engineering technique in detail.
    - Practical implementations and examples will be provided.
- **Learning Outcome**:
    - By the end of this series, you will be proficient in feature engineering and able to apply these techniques to your own datasets.

Final Thought

- **Quote**:
    - "Data is the new oil, but it needs to be refined (through feature engineering) to unlock its true value."