22/01/26

Data Preprocessing and feature Engineering.

PR-1- Data Profiler.

PART-A Fundamentals

Q 1. Write a short note : What is Data Analysis?

→ Data Analysis is the process of collecting, cleaning, transforming, and modelling data to discover useful information, identify patterns, and support decision-making.

for ex :- a story store owner examine daily sales data to determine which product sell best on weekends versus weekdays, using simple totals and averages to decide what stock more of.

Q.2 Describe how to plan a Data Science Project, listing all steps.

→ Planning a data science project involves following a structured sequence of steps from problem defination to ongoing maintenance.

Core Planning steps:

- Problem Defination :- Clearly articulate the business objective, such as predicting customer churn, and define success metrics like achieving 85% accuracy. Identify stakeholders and constraints like budget or timeline.

2. Data collection :- Gather relevant data from sources including databases, APIs, CSV files, or web scraping while ensuring data quality, volume sufficiency, and compliance with privacy regulation.

3. Cleaning & Preprocessing :- Handle missing values through imputing or removal, remove duplicates, correct inconsistencies, and normalize data formats to prepare a ~~rea~~ reliable dataset.

4. Exploratory Data Analysis (EDA) :- Use visualization like histograms and correlation matrices alongside statistical summaries to uncover patterns, outliers, and relationships within data.

5. Feature Engineering & Selection :- Create new feature (e.g., customer age from birthdate), apply transformation like scaling, and select optimal features using techniques such as PCA or feature importance scores.

6. Model Selection :- Choose appropriate algorithms based on problem type - regression for continuous outcomes, classification for categories, or clustering for unsupervised tasks - and consider baseline models.

7. Model Training :- Split data into training / validation sets, fit models using libraries like scikit-learn, and apply techniques

such as cross-validation to ensure robust learning

8. Model Evaluation and Turning - Assess performance with metrics like precision, recall, or RMSE; use grid search or random search for hyperparameter optimization and address overfitting.

9. Model Deloyment :- Integrate the model into production via APIs, cloud services, or dashboards, ensuring scalability and real-time inference capabilities.

10. Model Monitoring & Maintenance - Track performance drift, data quality degraduation, and business metric alignment ; retrain prediodically and implement feedback loops for continuous improvement.

Q.3 Frame a Machine learning problem statement. Predict whether a customer will churn based on purchase behaviour.

→ Problem statement - Develop a machine learning classification model to predict customer churn (~~binary~~ using their historical purchase behavior data, achieving at least 85% accuracy on a holdout test set to enable targetted retention retention campaign.

Input features:

- Transaction frequency (last 6 M)
- Avg. order value (total spend ÷ no. of transaction)
- Recency (days since last purchase)
- Total spend (cumulative purchase amount)
- Product categories purchased (count of distinct category)
- Days since first purchase

Targetted Variable:

- 1 = Churned
- 0 = Retained

Ex - customer ID - 12345 has 5 transactions, avg order value $45, last purchase 2 days ago, total spend $225, bought from 3 categories, account age 180 days, Model predict churn = 0 (retain).

Q.4. Explain:

a) What are Tensors?

→ Tensors are multi-dimensional arrays of numerical data that generalized scalars (0D), vectors (1D), and matrices (2D) to higher dimensions, serving as the fundamental data

structure in Machine learning frameworks like TensorFlow and Pytorch.

b. In-depth Explanation of Tensor using ex. of Numpy.

→ Tensors enable efficient storage and computation for computation for complex data like images (h × w × channels) or video (time × height × width × channels), supporting vectorized operations and automatic differentiation for gradient-based learning.

```
import numpy as np

scalar = np.array(42)
print("scalar [0D]:", scalar.shape)

vector = np.array([1.5, 2.3, 0.8])
print("vector [1D]:", vector.shape)

matrix = np.array([[1.5, 2.3, 0.8],
                   [2.1, 1.9, 1.2],
                   [0.9, 3.1, 0.5]])
print("Matrix [2D]:", matrix.shape)

tensor_3D = np.array([[[1.5, 2.3, 0.8],
                       [2.1, 1.9, 1.2],
                       [0.9, 3.1, 0.5]],
                      [[1.8, 2.0, 0.9],
                       [1.9, 2.2, 1.0],
                       [1.1, 2.8, 0.6]]])
```

```
print ("3D tensor:" tensor_3D.shape)

customer_means = np.mean (tensor_3D, axis=0)
print ("customer avg features:", customer_means)
```

Output

Scalar (0D) - ()
Vector (1D) - (3,)
Matrix (2D) - (3, 3)
Tensor (3D) - (2, 3, 3)

customer avg feature : [[1.65, 2.15 0.85]
[2.0., 2.05 1.1]
[1.0, 2.95 0.55]]

This 3D tensor track purchase behaviour evolution, with axis=0 anging daily pattern per customer - directly applicable to g churn prediction.