Top Questions

1. What is the difference between supervised, unsupervised, and reinforcement learning?

Expected Answer:

- **Supervised:** Learn mapping from inputs to outputs using labeled data (e.g., regression, classification).
- **Unsupervised:** Find hidden patterns or structures in unlabeled data (e.g., clustering, dimensionality reduction).
- **Reinforcement:** Learn actions to maximize cumulative reward (e.g., game agents, robotics).

Follow-up: Give an example of each from your past project.

About the projects :

1) From where did you get the data ?

- Service Based : Client provided you the data
- Product Based : Internal Team - Data Engineering team.

2) What was your responsibility ?

- a) you analyzed the data , performed the EDA and finally built the ML model.
- b) you tested the hypothesis, that the data could be fitted with ML model.

3) Why did you chose the algorithm that you chose ?

- a) target variable : if continuous -> Regression | categorical -> Classification

4) What were the evaluation metrics you chose for this model and why ?

- a) Regression : Mean Absolute Error, Mean Squared Error, Root MSE and R2 score
- b) Classification : Accuracy, Precision , Recall , F1- score.

5) What were the steps you performed in feature engineering and why ?

- a) Handling Missing values

- b) StandardScaler -> standardize the data [(xi - mean) / std]
- C) One hot encoding : to convert cat -> numerical
- d) Target/label encoding : Gender (M - F) -> (1 - 0)
- E) Outlier Detection : IQR : Q3 - Q1 : Filter lower_fence and upper_fence -> To convert the skewed data into normal distribution data.

6) What was the train test split ratio ?

- 70 - 30 | 80 - 20 | 75 - 25 | 90 - 10

7) What strategies you used to tackle the missing values and why ?

- one column it is discount for the products, so I have 65% rows missing in it ?-> remove column
- -> discount : 10 % missing values and the column is of type float ? Is the data skewed , if the data skewed -> median() | if the data not skewed -> mean()

- a) histogram b) box plot -> outliers

8)

MSE | RMSE | MAE -> 0 the better

R2 score -> 1 the better

Model A : MSE - 100 | R2 -> 0.89

Model B : MSE - 40 | R2 -> - 0.9 (not at all understood the pattern)

Model C : MSE - 150 | R2 -> 0.95 (understand the data pattern)

9) Methodology Used ?

- CRISP - DM (Cross Industry Standard Process for Data Mining)

10) Store the Model ?

-> Pickle format -> .pkl

11) API Development -> FastAPI : used to build the API's to serve the trained model.

MlFlow -> Experimentation | Model serving ()

Airflow -> DAG -> weekly -> weekly -> workflow automation

trained_model.pkl -> V1

trained_model.pkl-> V2

12) Packages -

A) Scikit - learn -> ML models

B) Numpy -> numerical operations

C) Pandas -> data analysis

D) Matplotlib.pyplot -> to visualise

E) Seaborn -> to visualise