ALEENA MARY ABRAHAM

V01150203

SCMA - EXAM II

Q1. (a) This is a classification problem → the target is categorical; either the user will cancel or not cancel their subscription the next month.

(b) One real-world feature that can help in prediction is no. of hours watched this month. This feature allows to identify low engagement which often signals users dissatisfaction, which increases the likelihood of churn.

(c) Predicting how many minutes a user will watch next week is a regression problem. This is because you are predicting a continuous numeric value ie. the minutes watched next week.

Q2. (a) This challenge is due to class imbalance where most users haven't churned, so the model might learn to always predict "not churn" to achieve high accuracy.

(b) One way to handle this issue is by using resampling techniques or class weighting to give more importance to rare churning cases during model training.

Q3. (a) This is a time-series problem relating to forecasting ~~rates~~ of sales on time-dependent patterns like seasonality and trends.

(b) Patterns that the model should account for include -

(i) Seasonality (celebrations spike)

(ii) Trends (increasing / decreasing sales over time)

(iii) Holiday (like diwali ~~or~~ christmas etc) promotions

(c) It is incorrect to randomly split this data into training & test sets because random splits break temporal order, which leaks future data into training & gives highly optimistic results. It is suggested to use chronological train-test split to preserve time structure.

Q4. (a) To prepare for -

(i) 'Tenure' :- impute using median or predictive imputation

(ii) 'Plan type' :- convert to numerical type using one-hot encoding or label encoding.

(b) Accuracy is not a good evaluation metric for fraud detection as it is misleading in imbalanced datasets.

A better metric for fraud detection could be - (i) F1-score, or

(ii) ROC-AUC

(i) F1 - score : It balances precision & recall useful when false negatives & false positives matter

(ii) ROC - AUC : It measures the model's ability to distinguish between fraud & non-fraud in the dat asets.

X ——— X