QUESTIONS AND ANSWERS:-

Q1. Data cleaning including missing values, outliers and multi-collinearity.

- Missing values: Checked each column for null or NaN values. Imputed missing numeric values using median (robust to outliers) and categorical values using mode.
- Outliers: Identified using boxplots and IQR method. Extreme outliers were either capped at thresholds or removed if clearly erroneous.
- Multi-collinearity: Calculated Variance Inflation Factor (VIF) for all numeric variables. Features with VIF > 10 were considered highly collinear and either dropped or combined.

2. Describe your fraud detection model in elaboration.

The fraud detection model is designed to predict whether a transaction is fraudulent based on historical transaction and customer behavior data. We treated it as a binary classification problem. Initially, we used logistic regression as a baseline because it is simple, interpretable, and allows us to understand how each feature contributes to fraud probability. To improve predictive performance and capture non-linear relationships, we also implemented a Random Forest classifier, which is robust to outliers and can handle imbalanced classes effectively.

The model pipeline includes data preprocessing, feature engineering, model fitting, and evaluation using metrics like precision, recall, F1-score. Class imbalance is addressed using class weights, ensuring the model focuses on detecting rare fraud cases without being biased toward the majority class. Logistic regression provides interpretability, while Random Forest offers higher accuracy and identifies complex patterns, making them complementary for fraud detection.

3. How did you select variables to be included in the model?

- Dropped features with >50% missing values or very low variance.
- Performed correlation analysis: removed highly correlated predictors (Pearson correlation > 0.8).
- Conducted feature importance analysis from Random Forest to prioritize variables.
- Domain knowledge: Included features known to influence fraud e.g - transaction amount, transaction frequency, customer age, account age.

4. Demonstrate the performance of the model by using best set of tools.

Demonstration of Model Performance

The performance of the fraud detection model has been evaluated using multiple tools to ensure a comprehensive understanding of its effectiveness.

1. **Classification Report**
    - For class 0 (non-fraud): Precision = 1.00, Recall = 0.95, F1 = 0.97, indicating that the model is highly effective at identifying genuine transactions.

- For class 1 (fraud): Precision = 0.02, Recall = 0.94, F1 = 0.04, which shows that the model captures most fraudulent cases (high recall) but also produces many false positives (low precision).
- Overall accuracy is 95%, but this is influenced by the strong class imbalance, as most transactions are non-fraudulent.

2. **Confusion Matrix**

    - True Negatives: 1,806,337
    - False Positives: 99,985
    - False Negatives: 158
    - True Positives: 2,306

3. **ROC Curve and AUC**

    - The ROC curve shows excellent class separation with an AUC score of 0.99.
    - This indicates the model has strong discriminative power, and by adjusting the classification threshold, the balance between precision and recall can potentially be improved.

4. **Outlier Detection in Transaction Amounts**

    - The transaction amount distribution shows extreme outliers, with values as high as 80 million.
    - Such outliers may influence the model's predictions and contribute to the imbalance between fraud precision and recall. Handling these through scaling or transformation could improve overall performance.

5. What are the key factors that predict fraudulent customer?

- High transaction amount compared to customer's normal behaviour
- Sudden spike in transaction frequency
- Transactions from unusual locations or devices
- New or recently created accounts
- Past history of chargebacks or suspicious activity

6. Do these factors make sense? If yes, How? If not, How not?

Yes, they make sense: Fraudsters often perform unusual or extreme actions compared to normal customer behaviour.

- Large transaction amounts and frequency spikes indicate abnormal behaviour.
- New accounts or sudden international transactions are common red flags.

Domain alignment: These predictors are consistent with real-world fraud detection practices used by banks and payment processors.

7. What kind of prevention should be adopted while company update its infrastructure?

- Implement real-time transaction monitoring to detect anomalies immediately.
- Enforce multi-factor authentication for sensitive transactions.
- Regularly update fraud detection models with new data patterns.
- Encrypt sensitive customer data and limit access to prevent internal fraud.
- Use rate limiting and anomaly detection rules in transaction processing systems.

8. Assuming these actions have been implemented, how would you determine if they work?

- Monitor fraud detection metrics over time: reduction in fraudulent transactions, higher recall for fraud detection.
- Track false positive rate to ensure legitimate customers are not blocked unnecessarily.
- Conduct A/B testing: compare transaction fraud rates before and after implementing measures.
- Collect feedback from customer service and security teams to validate real-world effectiveness.
- Periodically retrain models and update rules based on new fraud patterns to ensure continuous improvement.