Fraud Detection Case Study – Answers

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

Missing Values:
The dataset does not contain missing or null values. Also I checked any duplicates values are present.This was verified using exploratory data analysis techniques. Even though no missing values were found, validation was necessary to ensure data reliability.

Outliers:
Transaction amounts and balances had very uneven (skewed) spreads. In fraud detection, these unusual values often signal real fraud, so we kept them instead of removing them. We used log transformations and sturdy models to handle them without losing important fraud clues.

Multicollinearity:
Some balance variables (like oldbalanceOrg and newbalanceOrig) were strongly linked to each other. Tree-based models like Random Forest don't mind this much, so we kept them all. This let us track transaction patterns without hurting the model's accuracy.

2. Describe your fraud detection model in elaboration.

The fraud detection problem was treated as a binary classification task, where each transaction is classified as fraudulent or non-fraudulent.

A **Random Forest Classifier** was selected due to its ability to:

- works well with complex data
- handles non-linear patterns
- performs better on imbalanced datasets like fraud data
- Reduce overfitting through ensemble learning

The model was trained using a train-test split and class weighting was applied to give more importance to fraudulent transactions since they are very few compared to normal ones.

Random Forest also helps identify important features, making the model easier to understand from a business perspective. The model produces probability scores, allowing the company to set fraud detection thresholds based on risk level.

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

- Variables were selected using domain knowledge and model results.
- Identifier columns like nameOrig and nameDest were removed because they do not help in predicting fraud.
- Important transaction features such as amount, transaction type and balance changes were kept as they strongly indicate fraud behavior.
- Feature importance from the Random Forest model was used to identify the most influential variables.

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

Since fraud detection involves highly imbalanced data, accuracy alone is insufficient to evaluate performance.

The following metrics were used:

- Confusion Matrix
- Precision
- Recall

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

- **Transaction Amount:** Fraud usually involves very high or unusual amounts.
- **Account Balances:** Sudden drops or zero balance after a transaction indicate fraud.
- **Transaction Type:** TRANSFER and CASH_OUT transactions are more risky.
- **Balance Mismatch:** When the balance change does not match the transaction amount, it looks suspicious.
- **Transaction Pattern:** Quickly moving money and then withdrawing it is a common fraud behavior.

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

Yes, these factors make sense in real life. Fraud usually happens when someone tries to move or withdraw a large amount of money in a short time. This kind of behavior is uncommon for normal customers.

Regular users rarely empty their accounts or show unusual balance changes. Therefore, the factors identified by the model match real fraud behavior and help in detecting suspicious transactions.

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

To better stop fraud, the company should use these steps:

- Real-time monitoring: Watch transactions as they happen.
- Smart spending limits: Set limits based on each user's risk level.
- Quick-transaction checks: Spot and flag too many transactions in a short time.
- Extra login security: Require multiple proofs (like codes or biometrics) for risky actions.
- Pause suspicious ones: Temporarily freeze transactions that look off.
- Regular model updates: Retrain fraud-detection AI often with new data.

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

You can check if the fraud measures are working by tracking these:

- Drop in fraud cases over time.
- Less money lost to fraud overall.
- Fewer wrong alerts (false positives) and missed frauds (false negatives).
- Fewer customer complaints or transaction arguments.
- Side-by-side tests: Compare new rules against old ones.
- Ongoing checks on how well the fraud model performs.