Fraud Detection Project Questions

1. Data cleaning:

- No missing values or duplicates in the dataset.
- Outliers exist in `amount` and balances but were **kept**, as high values may indicate fraud.
- Dropped `newbalanceOrig` and `newbalanceDest` due to **high correlation** with sender/receiver balances.

2. Fraud detection models:

- **Logistic Regression**: baseline with class weighting to handle imbalance.
- **Random Forest**: final model, handles nonlinearity, imbalanced data, and gives feature importance for interpretability.

3. Feature selection:

- Removed identifiers (`nameOrig`, `nameDest`) and highly correlated features.
- Encoded `type` transactions into one-hot variables.
- Kept balances, transaction amount, step, and transaction type as predictors.

4. Model performance:

Model	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.008	0.77	0.016	0.898
Random Forest	0.947	0.708	0.811	0.985

Random Forest clearly outperformed the baseline, especially for precision and F1-score.

5. Key factors predicting fraud:

- `oldbalanceOrg` (high sender balances)
- `amount` (large transactions)
- Transaction types: `CASH_OUT`, `TRANSFER`
- `step` (time patterns)

6. Do these factors make sense?

Yes. Fraudsters target high-balance accounts and transfer or cash out money. Time patterns capture bursts of suspicious activity. These insights align with real-world fraud behavior.

7. Recommended prevention measures:

- Real-time monitoring for high-value CASH_OUT and TRANSFER transactions
- Flag suspicious accounts based on balances and patterns
- Dynamic thresholds using model probabilities
- Combine rules + ML in a multi-layer pipeline
- Continuous model retraining and monitoring

8. How to check if these measures work:

- Track fraud detection rate and false positives
- Measure financial savings from prevented fraud
- Compare old system vs new (A/B testing)
- Monitor feature drift and model performance over time