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

In the data cleaning process, I found that there were no missing values, which minimized the need for extensive cleaning. Outliers were retained instead of removed, as tree-based algorithms like RandomForest are robust against outliers and removing them could lead to a loss of important information. When checking for multi-collinearity, I found that most data points were not correlated. However, there were correlations between oldbalanceOrig and newbalanceOrig, and between oldbalanceDest and newbalanceDest. These correlations are logical, given that they represent the addition and subtraction of amounts in transactions. Since these correlations are expected and do not suggest redundancy, no columns were removed.

2. Describe your fraud detection model in elaboration.

The fraud detection model follows these steps:

1. The process begins by importing the dataset and conducting an initial exploratory analysis to understand the data.
2. Correlations between data points are checked, especially focusing on removing any correlations between unrelated features.
3. Skewed columns undergo transformations to normalize their distribution, improving model accuracy.
4. : One-hot encoding is applied to categorical variables, such as nameOrig and nameDest, which are then preserved as additional features in the model.
5. The data is split into training and testing sets using an 80-20 ratio to ensure robust model training and evaluation.
6. Sampling techniques are applied to address class imbalance, a common issue in fraud detection datasets.
7. The model is trained using both RandomForestClassifier and Logistic Regression, allowing for comparison and selection of the best-performing algorithm.

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

Variables were selected based on their relevance and contribution to the predictive power of the model. The selection process involved:

- Only features directly related to fraud detection were included.
- Variables that did not exhibit significant correlation with each other were preferred to avoid multicollinearity.
- Skewed variables were transformed to normalize their distribution.
- New features were created when necessary to enhance the model's effectiveness.
- Data normalization was applied to improve the model's performance and accuracy.

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

The model's performance was evaluated using several metrics:

- While accuracy was measured, it was not solely relied upon due to the class imbalance common in fraud detection problems.
- The classification report provided a detailed view of precision, recall, and F1-score. RandomForestClassifier demonstrated higher recall but lower precision compared to Logistic Regression. Since recall is crucial in fraud detection, RandomForest was identified as the best-performing model.
- The confusion matrix was used to understand the model's prediction accuracy, particularly the number of true positives and false negatives. RandomForestClassifier outperformed Logistic Regression by producing significantly fewer false negatives.
- The ROC curve was plotted to visualize the trade-off between true positive and false positive rates, further validating the RandomForest model's effectiveness.

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

The key features identified as significant predictors of fraudulent activity were:

- 'isFlaggedFraud'
- 'oldbalanceOrg'
- 'type_CASH_OUT'
- 'type_TRANSFER'

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

Yes, these factors make sense:

- 'isFlaggedFraud': This feature is critical as it flags transactions based on predefined conditions in the fraud detection software, such as transaction amount limits. It serves as an essential precautionary measure.
- 'oldbalanceOrg': Customers with higher balances in their accounts are more susceptible to fraud, as fraudsters often target accounts with substantial funds.
- 'type_CASH_OUT': Cash-out transactions, where non-cash assets are converted to cash, are more prone to fraud, particularly in online scenarios involving high-value goods.
- 'type_TRANSFER': Fraud often occurs during transfers, especially when they are uninformed or poorly informed, making this feature a critical indicator of potential fraud.

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

To enhance fraud prevention, the company should consider the following measures:

- Implement multiple accounts or distribute funds for customers with large balances to reduce risk.
- Use separate accounts for online transactions to mitigate the likelihood of fraud.
- Introduce special protocols and security measures for cash-out transactions.
- Require additional verification steps for users frequently using transfer modes.

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

To determine the effectiveness of these measures, the company should collect and analyze data post-implementation. By comparing the rate of fraudulent activities before and after the changes, the company can assess whether the measures have successfully reduced the incidence of fraud. Continuous monitoring and periodic evaluation will also be crucial in ensuring the long-term success of these preventative strategies.