Case Study Questions

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

- There were no Missing values or NaN values.
- No need of data cleaning was needed.

The number of fraudulent transactions is very small compared to the total number of transactions. The fraudulent transaction can be considered outliers.

- I worked on outlier detection techniques as below.
    - I implemented the Isolation Forest technique for outlier detection. Thi technique is based on the decision tree and works on isolating the outliers.
    - I also implemented a local outlier factor algorithm for outlier detection.
    - The density of points in a region is a good differentiator between genuine and fraudulent transactions.
    - For a fraudulent transaction, the number of data points in the neighborhood is very small. For a genuine transaction, the number of data points in the neighborhood is large.
    - To understand the density, I implemented the DBSCAN technique.
- The results of the outlier detection techniques were not satisfactory. They could not identify the fraudulent transactions correctly.

Q2. Describe your fraud detection model in elaboration.

- As Fraud detection is a Binary type of classification task. The dataset provided is Labeled data. So we need ML models which support Supervised and binary types of data.
- After doing the EDA of the dataset I found that data is heavily imbalanced so I need performance evaluation metrics as: **a. Confusion Matrix:** To check TP,FP ,FN, TN **b. Precision:** Out of all data points predicted as Fraud, how many of them were actually a fraud? We want our model which will not miss any fraudulent transaction. **c. Recall:** Out of all Fraud transactions how much of them were predicted fraud? Because we want will not classify fraudulent transactions as Valid transactions. **d. ROC_AUC:** Tells us the discriminative power of the model between fraud and valid transactions.
- So Model selection criteria become **a. Model which does binary classification** b. The model which gives Probability values along with class labels
- Due to this reason, I selected Logistic Regression and the Random Forest model. a. Logistic Regression: Supports Binary classification by default and returns probability values. b. Random Forest: This is an ensemble type of model. The performance of these models is Superior because of a combination of multiple base models. Also, this model Supports Binary classification and returns probability values.

Q3.How did you select variables to be included in the model?

- The selection of variables was done after performing extensive EDA.

For example, dropping the 'isFlaggedFraud' variable was decided because it was violating the given criteria as per the dataset description and the number of data points was only 16.

- For the 'type' of transactions I included only CASH_OUT & TRANSFER transactions. Because fraud was only in that transaction.
- nameOrig and nameDest were dropped because of for fraud transactions, the account that received funds during a transfer was not used at all for cashing out.
- See EDA part for more details.

Q4. Demonstrate the performance of the model by using the best set of tools.

- As the data was heavily imbalanced, so selecting accuracy as a performance metric is not a good idea.Because even a dumb model will give good accuracy.
- So I selected the following performance evaluation metrics to evaluate the performance of model: a. Confusion Matrix: To check TP, FP ,FN, TN b. Precision: Out of all data points predicted as Fraud, how many of them were actually fraud? We want our model which will not miss any fraudulent transactions. c. Recall: Out of all Fraud transactions how much of them were predicted fraud? Because we want a model which will not classify fraud transactions as Valid transactions. d. ROC_AUC: Tells us the discriminative power of the model between fraud and valid transactions.

Q5. What are the key factors that predict fraudulent customers?

- CASH_OUT and TRANSFER type of transactions are responsible.
- ErrorbalanceOrig and ErrobalanceDest features.
- Hours_of_day feature.
- The account that received funds during a transfer was not used at all for cashing out.

Q6.Do these factors make sense? If yes, How? If not, How not?

- Yes, these factors make sense.

Based on domain knowledge and my own study of fraud detection, I believe we can detect fraudulent transactions with high accuracy by deploying my model.

- As per EDA, I separated the fraud and valid transactions. And found that fraud transaction happens only when the transaction type is CASH_OUT and TRANSFER.
- ErrorbalanceOrig and ErrobalanceDest feature: As per EDA, Most of the errorbalanceOrg of fraud transactions are -ve and up to 75% of transactions have an error value of less than 0. Hence this can be a significant point to categories of valid and fraud transactions. While in the case of valid transactions have a large errorbalanceorig.
- Hours_of_day feature: Valid transactions occur only within specific hours of the day.But Fraud transactions occur at any time of the day. This Hours_of_data can be an important reason to differentiate between fraud and valid transactions.

Q7. What kind of prevention should be adopted while the company updates its infrastructure?

- Improve the data collection of 'isFalggedFraud' feature, it's not working as per the criteria provided in the dataset.
- There is a lot of error in tallying the balance of (oldbalanceOrig,newbalanceOrig) and (oldbalanceDest,newbalanceDest)performing after transactions.
- while updating the company infrastructure, we should collect more features for the transactions. Some examples of additional features that should be collected are as follow.

    - Location of transaction: city, state, the country from IP address - if a user's primary address is Mumbai and if the same users make a large transaction in a North-East Indian state, such transaction is suspicious.
    - Device type of transaction: Android, iOS, Desktop, etc.
    - Check whether the transaction was done from a user-authorized device or not.
    - For each user, the average amount of transactions (if a user usually makes transactions of small amounts, and suddenly if a new transaction is of a large amount, it could be fraudulent.)
    - For each merchant, the average amount of transactions (if a merchant usually receives transactions of small amounts, and suddenly if a new transaction is of a large amount, it could be fraudulent.)
    - IP Address
    - User's last known location
- The amount of each Fraud transaction is higher compared to a valid transaction. So tracking transactions contain a higher amount with more attention.

Q8.Assuming these actions have been implemented, how would you determine if they work?

- We will deploy the fraud detection model. After the deployment, we will continuously monitor the performance of the fraud detection model.
- We will also periodically re-train the model by adding recent transaction data sets.
- We will also analyze data drift - the change in the type of data as compared to the original trained data. We will analyze the changes in the distribution of the data and make improvements to model parameters accordingly.