Data Scientist Takeaway Assessment

Objective: This assessment aims to evaluate your approach to handling real-world data challenges, your strategic vision for a data team, and your ability to translate technical details into business insights.

Task Overview

You are provided with a dataset. Perform the following:

- **Exploratory Data Analysis:** Identify patterns, outliers, and anomalies.
- **Predictive Modeling:** Build a predictive model that addresses a specific business outcome related to the dataset.
- **Insights and Recommendations:** Provide insights from your analysis and model. How would these insights influence business decisions?
- **Technical Documentation:** Document your methodology, code, and findings in a clear and concise manner.

Dataset Name: `financial_transactions.csv`

Dataset Overview: The dataset represents financial transactions for a hypothetical fintech company over the course of one year. These transactions involve payments processed for various services and products. The data includes normal transactions as well as anomalies such as fraud attempts, refunds, and chargebacks.

Columns in the Dataset:

- **Transaction_ID:** A unique identifier for each transaction.
- **Date:** The date of the transaction (format: YYYY-MM-DD).
- **Time:** The time the transaction was processed (HH:MM:SS).
- **Customer_ID:** A unique identifier for the customer.
- **Product_ID:** A unique identifier for the product or service involved.
- **Amount:** The amount of the transaction in USD.
- **Payment_Type:** The method of payment (e.g., credit card, debit card, PayPal, etc.).
- **Country:** The country from which the transaction was made.
- **Merchant_ID:** A unique identifier for the merchant.
- **Status:** The status of the transaction (Completed, Pending, Cancelled, Refunded, Chargeback).

Sample Data:

```
Transaction_ID, Date, Time, Customer_ID, Product_ID, Amount, Payment_Type,
Country, Merchant_ID, Status
```

```
TX10001, 2023-04-01, 12:34:56, C1001, P001, 200.00, Credit Card, USA, M001, Completed
TX10002, 2023-04-01, 13:21:09, C1002, P002, 300.00, PayPal, UK, M002, Refunded
TX10003, 2023-04-01, 14:15:23, C1003, P003, 450.00, Debit Card, Canada, M003, Chargeback
```

Data Generator:

Use the following python script to generate more data, the one provided is just a sample.

```
import pandas as pd
import numpy as np
from faker import Faker
import random

fake = Faker()

# Set random seed for reproducibility
np.random.seed(42)

# Generate data
num_records = 1000
dates = [fake.date_between(start_date='-1y', end_date='today') for _ in range(num_records)]
times = [fake.time() for _ in range(num_records)]
customer_ids = [f'C{1000 + i}' for i in range(num_records)]
product_ids = [f'P{random.randint(1, 100)}' for _ in range(num_records)]
amounts = np.random.uniform(50, 500, num_records).round(2)
payment_types = np.random.choice(['Credit Card', 'Debit Card', 'PayPal'], num_records)
countries = np.random.choice(['USA', 'UK', 'Canada', 'Australia'], num_records)
merchant_ids = [f'M{random.randint(1, 50)}' for _ in range(num_records)]
statuses = np.random.choice(['Completed', 'Pending', 'Cancelled', 'Refunded', 'Chargeback'], num_records)

# Create DataFrame
df = pd.DataFrame({
    'Transaction_ID': [f'TX{10000 + i}' for i in range(num_records)],
    'Date': dates,
    'Time': times,
    'Customer_ID': customer_ids,
    'Product_ID': product_ids,
    'Amount': amounts,
    'Payment_Type': payment_types,
    'Country': countries,
    'Merchant_ID': merchant_ids,
    'Status': statuses
})

# Save to CSV
df.to_csv('financial_transactions.csv', index=False)

print("Dataset generated and saved to 'financial_transactions.csv'")
```

Instructions:

1. **Install Necessary Libraries**: Ensure you have `pandas`, `numpy`, and `Faker` installed in your Python environment. You can install these using pip if you don't already have them:
   Bash
   ```
   pip install pandas numpy faker
   ```
2. **Run the Script**: Run the provided Python script. This will generate a file named `financial_transactions.csv` in your current directory containing 1,000 synthetic records of financial transactions.
3. **Customization**: You can modify the ranges, choices, and the number of records as necessary to better fit the specifics of the scenario you want to simulate.

Tasks for the Candidate:

1. **Data Cleaning and Preprocessing:**
   - Handle missing values, outliers, and duplicate entries.
   - Convert data types if necessary (e.g., converting string dates and times to datetime objects).
2. **Exploratory Data Analysis:**
   - Analyze the distribution of transactions over time.
   - Explore the relationship between transaction amount and other variables like Payment_Type and Status.
   - Identify any patterns or trends that could indicate fraudulent activity or other anomalies.

3. **Predictive Modeling:**
    - Build a model to predict the likelihood of a transaction being fraudulent.
    - Evaluate the model's performance using appropriate metrics (accuracy, precision, recall, F1-score, ROC-AUC).
4. **Insights and Recommendations:**
    - Provide insights based on the analysis.
    - Recommend strategies to mitigate risks associated with fraudulent transactions.

Data Privacy Note:

Please ensure that all data used in this assessment is fictional and does not contain any real personal or financial information.

Submission Guidelines:

- **Format:** Submit your proposal, case study analysis, and scenario responses in a markdown document in a github repository. Source code in python to be submitted in the same github repository as well.
- **Deadline:** Please submit your completed assessment within 5 days of receipt.

Evaluation Criteria:

- **Technical Proficiency:** Ability to apply data techniques effectively and efficiently.
- **Innovativeness:** Creativity in problem-solving and in proposing innovative solutions.