Chosen Topic: Credit Cards

Specific Problem:

As a data scientist, the problem I want to propose is:

"Predicting the likelihood of fraud in credit card transactions using historical data."

The goal is to help banks or payment processors identify and prevent fraudulent transactions more effectively, protecting both users and the organization.

Applying the Data Science Methodology:

1. Defining the Problem

- **Description**: A bank or payment processor wants to reduce the number of fraudulent transactions in their system. Fraudulent transactions harm the security of users and cause financial losses. The client's goal is to develop a model that can identify fraud with high accuracy in real time.
- **Objective**: Build a predictive model that, given a set of features related to credit card transactions, can tell if a transaction is fraudulent or not.

2. Collecting Data

- **Data to collect**:
    - History of credit card transactions, both legitimate and fraudulent.
    - Transaction attributes: transaction amount, geographic location, time of day, card type, merchant, etc.
    - Additional info on the user: card usage frequency, usage history, credit limits, and more.
- **Data sources**: Historical transaction databases from the bank or card processors.
- **Data size**: Fraudulent transactions are usually a small fraction of all transactions, so a large dataset is needed to get useful results (imbalanced class problem).

3. Data Preparation and Cleaning

- **Tasks to do**:
    - Remove duplicate or incorrect transactions.
    - Impute missing values (like when some transaction variables aren't recorded).
    - Normalize the data, for example, scale transaction amounts.
    - Handle outliers: transactions with unusually high amounts or from rare locations.
    - Transform categorical variables: convert categories like merchant or card type into numerical or dummy variables.

- **Objective**: Ensure the data is clean and ready for modeling, with a special focus on fraudulent transactions, which are usually a minority compared to legitimate ones.

4. Exploratory Data Analysis (EDA)

- **Actions**:
    - Create visualizations to see the distribution of fraudulent and legitimate transactions.
    - Identify patterns, such as high-risk transactions (like ones at unusual times, in remote locations, or with large amounts).
    - Look for correlations between features to find factors that could increase the likelihood of fraud.
- **Objective**: Gain a deeper understanding of key variables that might be related to fraudulent transactions.

5. Modeling

- **Algorithms to consider**:
    - **Decision Trees** and **Random Forests** to understand feature importance.
    - **Classification models** like **Logistic Regression** to predict the probability of fraud.
    - **Advanced algorithms** like **XGBoost** or **Neural Networks** to improve model accuracy.
- **Tasks**:
    - Split data into training and testing sets.
    - Train several models and fine-tune hyperparameters.
    - Deal with class imbalance using techniques like **oversampling** (SMOTE) or adjusting class weights.
- **Objective**: Find the model that best predicts fraud and generalizes well to new data.

6. Model Evaluation

- **Key metrics**:
    - **Accuracy**: the percentage of correctly classified transactions.
    - **Recall (sensitivity)**: the model's ability to detect all fraudulent transactions.
    - **F1-score**: combines precision and recall into a single metric, useful for imbalanced classes.
    - **ROC-AUC**: to evaluate the model's ability to distinguish between fraudulent and non-fraudulent classes.

- **Objective**: Ensure the model not only predicts well but minimizes false positives and false negatives, which is key to preventing losses without blocking too many legitimate transactions.

7. Deployment

- **Actions**:
    - Deploy the model in a real-time system that can monitor and flag suspicious transactions.
    - Develop an automatic alert system that notifies the fraud team or even the customer when a transaction is marked as suspicious.
    - Include an additional layer of manual review for cases where the probability of fraud is not clear.
- **Objective**: Prevent fraud in real time and improve security for both users and the bank.

8. Monitoring and Maintenance

- **Actions**:
    - Monitor the model's performance in production to ensure it continues to work properly with new data.
    - Periodically update the model with new transactions to improve accuracy.
    - Adjust parameters or retrain the model if fraud patterns change.
- **Objective**: Ensure the system remains effective over time, adapting to new fraud techniques.

9. Communicating Results

- **Actions**:
    - Present the model's results to the client, highlighting its ability to detect fraud and its accuracy.
    - Create visual reports showing how the number of detected fraudulent transactions evolves and the impact on financial losses.
    - Explain limitations and give recommendations on how to improve the system in the future.
- **Objective**: Make sure the results are easy to understand for non-technical stakeholders and give them a solid basis for decision-making.

10. Optimization and Continuous Improvement

- **Actions**:
    - Periodically review the results obtained and the fraud cases that were missed to improve the model.

    - Consider new data sources or additional features to enhance fraud detection.
    - Test new algorithms or combinations of models (ensemble methods) to optimize performance.
- **Objective**: Keep improving the model and make sure it continues to provide value over time.