CREDIT CARD FRAUD DETECTION

PHASE 1: PROBLEM DETECTION AND DESIGN THINKING

PROBLEM DEFINITION

- The problem of credit card fraud detection involves identifying and preventing unauthorized or fraudulent transactions made using credit cards. The goal is to develop algorithms and systems that can accurately distinguish between legitimate and fraudulent transactions in real-time to minimize financial losses for both cardholders and financial institutions.

DESIGN THINKING: DATA SOURCE

- In the context of design thinking, when tackling the problem of credit card fraud detection, it's essential to consider various data sources to gain a holistic understanding of the problem and to generate innovative solutions. Here are some potential data sources you can explore
- 1. TRANSACTION DATA: Collect historical credit card transaction data from financial institutions, including details like transaction amount, timestamp, location, merchant, and whether each transaction was classified as legitimate or fraudulent. Gathering information about cardholders, such as their demographics, spending habits, and transaction histories. This data can provide insights into typical behavior patterns.
- 2.FRAUD INVESTIGATION RECORD: Analyze historical fraud reports and complaints from customers to identify common fraud patterns and pain points experienced by cardholders. If available, examine records of past fraud investigations, including how they were resolved and what techniques were used to identify fraud transaction.

DESIGN THINKING: DATA SOURCE

- 3.Industry Benchmarks: Research industry benchmarks and best practices in fraud detection to gain insights into what works well in the field.
- 4.Regulatory Guidelines: Review regulatory guidelines and requirements related to fraud detection and prevention in the financial industry to ensure compliance with legal standards.
- 5.Expert Interviews: Conduct interviews with experts in the field of fraud detection, data science, and cybersecurity to gain insights and validate ideas.

DATA PREPROCESSING

- Data Cleaning:- Handle missing values: Identify and fill in missing values, possibly using techniques like mean imputation or more advanced methods.-Outlier detection: Detect and deal with outliers that could skew the model. Outliers in transaction amounts, for example, could indicate potential fraud.
- Data Sampling: .If the dataset is highly imbalanced employ techniques like oversampling or under sampling to balance the dataset.
- Data Splitting: Split the dataset into training, validation, and test sets to evaluate the model's performance effectively.

FEATURE ENGINEERING

- Transaction Aggregation: Aggregate transaction amounts and frequencies over various time windows to capture spending patterns.
- Cardholder Behavior: Average transaction amount , Frequency of transactions, The typical merchant category. Variability in transaction amounts.
- Time Series Features: Employ time series analysis techniques to extract features like moving averages, exponential smoothing, or autocorrelation.
- Fraud Indicator Features: Add features that may indicate potential fraud, like unusually large transactions or transactions from high-risk geographic areas.

MODEL SELECTION

- Logistic Regression: works well for linearly separable data.
- Gradient Boosting Algorithms: Effective for handling imbalanced datasets, capture complex patterns, and achieve high accuracy.
- Support Vector Machines (SVM):Effective for separating data in high-dimensional spaces, especially when classes are not linearly separable.
- K-Nearest Neighbors (KNN): Simple and effective for local pattern detection. Can work well for certain types of fraud detection.
- Naïve Bayes: Simple and efficient for text or categorical data. Can work well for certain types of fraud detection.

MODEL TRAINING

- Data Splitting: Split your preprocessed dataset into three subsets: training, validation, and test sets. A common split might be 70% for training, 15% for validation, and 15% for testing.
- Model Persistence: Once you are satisfied with the model's performance, save the trained model so that it can be deployed and used for real-time fraud detection.
- Validation: Continuously monitor the model's performance on the validation set during training. This allows you to detect issues like overfitting or under fitting and adjust hyper parameters accordingly.
- Feature Selection/Extraction: If necessary, further refine your feature set based on the model's performance during training. Feature selection methods can help identify the most relevant features for the model.

EVALUATION

- Feedback Loop: Establish a feedback loop to continuously monitor the model's performance in a production environment and make necessary adjustments as fraud patterns evolve.
- Documentation: Document the evaluation results, including performance metrics, visualizations, and any significant findings. This documentation is essential for regulatory compliance and model transparency.
- Communication: Communicate the evaluation results and model performance to relevant stakeholders, including business leaders, data scientists, and compliance teams.
- Compliance and Regulations: Ensure that your model's performance aligns with legal and regulatory requirements, especially if it involves handling sensitive customer data.