April 3, 2025

Data Methodology

Applying the Data Science Methodology: A Case Study on Credit Cards

By: Jeya Prakash I

Next, you will play the role of the client and the data scientist.

Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. (*3 points*)

You are required to:

1. Describe the problem, related to the topic you selected.
2. Phrase the problem as a question to be answered using data.
3. For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

ANSWER:

Client (Me):

"I'm managing a credit card company, and we're seeing an increase in fraudulent transactions. This is costing us money and damaging our reputation. We need to find a way to proactively identity potentially fraudulent transactions in real-time to minimize losses and protect our customers. We want to be able to flag transactions that deviate significantly from a user's normal spending habits."

Data Scientist (Also me):

"Understand. Fraudulent transactions are a serious concern. We can leverage machine learning and data analysis to build a robust fraud detection system. We'll analyze transaction patterns, identify anomalies, and develop a model that can flag suspicious activities in real-time. This will involve analyzing a number of feature related to each transaction."

Problem Description:

The problem is to develop a system that can accurately detect fraudulent credit card transactions in real-time. This involves analyzing transaction data to identify patterns and anomalies that indicate fraudulent activity. The goal is to minimize false positives (flagging legitimate transactions as fraudulent) and false negatives (failing to detect actual fraud).

Problem Question to be Answered using Data:

"Can we build a machine learning model that accurately identifies fraudulent credit card transactions in real-time by analyzing transaction features such as transaction amount, location, time, merchant category, and user spending history, with a high degree of precision and recall?"

Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. (*5 points*):

1. Analytic Approach
2. Data Requirements
3. Data Collection
4. Data Understanding and Preparation
5. Modeling and Evaluation
6. You can always refer to the labs as a reference with describing how you would complete each stage for your problem.

ANSWER:

Problem Question:

"Can we build a machine learning model that accurately identifies fraudulent credit card transactions in real-time by analyzing transaction features such as transaction amount, location, time, merchant category, and user spending history, with a high degree of precision and recall?"

1. Analytic Approach:

Classification Problem: This is fundamentally a binary classification problem. We need to classify each transaction as either "fraudulent" or "not fraudulent."

Supervised Learning: We'll use supervised learning algorithms, as we have labeled data (historical transactions marked as fraud or legitimate).

Real-time Prediction: The model needs to perform predictions in real-time or near real-time, so efficiency is crucial.

Anomaly Detection: We'll also explore anomaly detection techniques to identify transactions that deviate significantly from a user's typical behavior.

Algorithms: We'll consider algorithms like:

- Logistic Regression
- Random Forest
- Gradient Boosting Machines (e.g., XGBoost, LightGBM)

- Neural Networks (especially for real-time processing)
- Isolation Forest or One-Class SVM for anomaly detection.

2. Data Requirements:

Transaction Data:

- Transaction ID
- User ID
- Transaction Amount
- Transaction Date and Time
- Merchant Name/Category
- Transaction Location (IP address, GPS coordinates)
- Card Information (masked)

User Profile Data:

- User's typical spending patterns
- User's location history
- User's device information

Fraud Labels:

- A flag indicating whether a transaction was fraudulent.
- Time when the fraud was confirmed.

External Data (Optional):

- Geographic data (e.g., population density)
- Economic indicators (e.g., regional fraud rates)

3. Data Collection:

Database Extraction:

- Extract transaction data from the credit card company's database.

API Integration:

- If necessary, integrate with external APIs to gather additional data (e.g., geolocation data).

Data Warehousing:

- Store the collected data in a data warehouse for efficient access and processing.

Data Streaming:

- Set up data streams to capture real-time transaction data.

4. Data Understanding and Preparation:

Exploratory Data Analysis (EDA):

- Analyze the distribution of transaction amounts, dates, and locations.
- Identify patterns and anomalies in the data.
- Visualize the data to gain insights.

Data Cleaning:

- Handle missing values (e.g., imputation)
- Remove duplicate or inconsistent data.
- Encode categorical variables (e.g., one-hot encoding).

Feature Engineering:

1. ***Create features like:***
    - Time since last transaction
    - Frequency of transactions in a given location
    - Deviation from user's typical spending amount
    - Distance between transaction location and use's home location
2. ***Create rolling averages of spending.***

Data Splitting:

- Split the data into training, validation, and test sets.
- Because of the time based nature of the data, the split should be done in a way that the model is trained on older data, and tested on newer data.

Handling Imbalanced Data:

- Fraudulent transactions are typically rare, so we'll need to address the class imbalance (e.g., oversampling, under sampling SMOTE).

5. Modeling and Evaluation:

Model Selection:

Train and compare the performance of different classification algorithms.

Hyperparameter Turning:

Optimize the hyperparameters of the chosen model using techniques like cross-validation.

Model Evaluation:

- Use metrics like precision, recall, F1-score, and AUC-ROC to evaluate the model's performance.
- Pay close attention to recall, as it's crucial to minimize false negatives (missed fraud).
- Create a confusion matrix to understand the performance of the model.

Real-time Implementation:

- Deploy the model in a real-time environment.
- Monitor the model's performance and retrain it as needed.
- Create alerts for when the model flags a transaction as fraud.

Feedback Loop:

Implement a system for gathering feedback from fraud analysts to improve the model's accuracy.