Milestone 1: Identifying Datasets for Credit Card Fraud

Sreenivasulu Somu

Department of Data Science, Bellevue University

DSC540-T301 Data Preparation (2251-1)

Ms. Catherine Williams

September 28, 2024

Milestone 1: Identify Datasets for Credit Card Fraud

Project Subject Area

This project aims to analyze Credit Card Fraud patterns and their relationship to the consumer complaints and healthcare provider payments, providing insights into potential financial misconduct across different sectors.

Data Sources

- Flat File:
  - Description: Credit Card Fraud Detection dataset from Kaggle.
  - Link: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- API:
  - Description: Open Payments API providing data on payments to healthcare providers.
  - Link: https://openpaymentsdata.cms.gov/
- Website:
  - Description: Consumer Complaint Database from Data.Gov.
  - Link: https://catalog.data.gov/dataset/consumer-complaint-database

Relationships

To establish relationships among the datasets, we will try to link the Credit Card Fraud dataset with the consumer complaints database using common fields like transaction dates and amounts. We will connect to the Open Payments API data by correlating payment amounts and dates, allowing us to identify potential fraud patterns across financial transactions and healthcare payments. This integration will provide a comprehensive analysis of Credit Card Fraud and its implications.

Project approach/plan

The project approach is taken by preprocessing and cleaning the data from all the 3 sources, ensuring consistency in the date formats and transaction amounts. We shall then merge the datasets based on the identified relationships, creating a complete view of transactions, complaints, and healthcare payments.

The analysis shall focus on identifying patterns of fraudulent activity, particularly looking at correlations between credit card fraud, consumer complaints, and unusual healthcare provider payments. The visualization tools will be utilized to create interactive dashboards, allowing for an easy exploration of trends and patterns in the data.

Concerns/challenges

- **Data quality and consistency:** Merging of data from diverse sources may lead to inconsistencies and missing values.
- **Privacy concerns:** Handling some of the sensitive financial and healthcare data requires adherence to data protection regulations.
- **False positives:** Distinguishing between a genuine transactions and fraudulent ones could be challenging, potentially leading to misclassification.
- **Data volumes:** Processing and analyzing large datasets may require significant computational resources.

Ethical Implications

- **Privacy protection:** Handling some of the sensitive financial and healthcare data raises significant privacy concerns. Strict measures should be implemented to anonymize and protect individual identities.

- **Bias in fraud detection:** The ML models used for fraud detection might accidentally incorporate biases, potentially leading to an unfair treatment of certain groups.
- **Transparency:** There could be an ethical obligation to be transparent about the methods used in fraud detection, especially if they may impact individuals or businesses.
- **Ethical use of findings:** The insights derived from this analysis should be utilized with caution, ensuring that no actions are taken that might unjustly affect individuals or groups.