subramanian540_Project_Milestone1_Updated

April 11, 2023

1 Project: CreditCard Fraud Detection Data

2 DSC 540 - Data Preparation

3 Week3 and Week 4

4 Project Milestone 1

5 Data Sources Used

API : https://www.fraudlabspro.com/developer/api/screen-order?ref=apilist.fun

This REST API will detect all possibles fraud traits based on the input parameters supplied. The more input parameter supplied, the higher accuracy of fraud detection.

CSV File : https://www.kaggle.com/datasets/kartik2112/fraud-detection?resource=download

This is a simulated credit card transaction dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.

web: https://datahub.io/machine-learning/creditcard

The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset present transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

6 Accomplish the Milestone 1 and Interpretation of the data

The three sources of data that I plan to use for the Credit card fraud detection project are a REST API, a CSV file, and a web-based dataset. The REST API provided by FraudLabs Pro is a tool that can be used to detect possible fraud traits based on input parameters. The CSV file from Kaggle contains simulated credit card transactions with legitimate and fraud transactions between Jan 2019 and Dec 2020. The web-based dataset from Datahub.io contains credit card transactions made by European cardholders in Sep 2013, where we have 492 fraud cases out of 284,807 transactions.

The Fraud Detection dataset contains anonymized credit card transaction data, including the transaction amount, timestamp, and various features such as the merchant ID, cardholder information, and location. This information can be used to train a machine learning model to predict whether or not a transaction is fraudulent based on patterns and correlations in the data.All three datasets contain information that can be used to detect potential fraud, such as transaction details, customer information, merchant information, and indicators of fraudulent activity.

To connect the data from these sources, I plan to first analyze the CSV file and the web-based dataset separately to understand the features and patterns of fraudulent transactions. This will include analyzing transaction amounts, merchant types, time of day, and other relevant factors. Then, I will use the REST API to validate and verify the suspicious transactions found in the CSV file and the web-based dataset. This will help me to identify additional fraud traits and strengthen my understanding of what features contribute to fraudulent transactions.To tackle this project, I would first explore the FraudLabs Pro API documentation and understand its capabilities and limitations. I would then examine the CSV and web based dataset to ensure that it does not contain any sensitive information such as cardholder names or billing addresses. If it does, I would remove these fields to ensure the privacy of the individuals involved.

Next, I would import the Fraud Detection datase csv files into a programming language such as Python and write code to iterate through each row of the file and use the FraudLabs Pro API to screen each order for fraud. This would involve making API calls to the FraudLabs Pro API and analyzing the response to determine if an order is legitimate or not. I will also use visualization techniques to better understand the data and identify any trends or patterns that are not immediately obvious.

One ethical implication of this project is the potential for false positives or false negatives. If the screening process is too strict, legitimate orders may be flagged as fraudulent and rejected. Another challenge may be the quality of the data. Data cleaning and preprocessing may be necessary to remove outliers or errors that could negatively impact the accuracy of the analysis. In addition, the synthetic dataset may not accurately represent real-world transactions, which could limit the effectiveness of any fraud detection methods developed using this dataset.Overall, this project has the potential to improve the efficiency and accuracy of fraud screening processes, but it is important to approach it with caution and consideration for the ethical implications and potential challenges.

7 References

https://www.fraudlabspro.com/

https://app.datacamp.com/

https://kaggle.com/datasets/

[ ]: