Mishra540_Project_Milestone_01

June 18, 2023

[1]:
```
# DSC540,Summer 2023 - T302 Data Preparation(2237-1)
# Assignment: Project Milestone 01
# Author by:  Debabrata Mishra
# Date: 2023-06-18

# Topic - Credit Card Transactional & Demographic Data
```

1 Project Description

In this project, the data consists of credit card transactions along with demographic information details such as merchant information, timestamps, and contextual features. The fraud labels indicate whether each transaction is fraudulent or legitimate. The objective is to gather data from the three sources, perform data cleaning and formatting, and generate user-defined fields as needed. Next, the relationships between the data sources will be established, and the data will be stored for further analysis. Finally, Python will be used to create visualizations to gain insights from the data.

2 Data Sources

2.1 Flat File

Descrition:

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 across 693 unique merchant ids. It has 23 columns and transactional information along with demographic details of Merchants and Card Holders.

Link: https://www.kaggle.com/code/nathanxiang/credit-card-fraud-analysis-and-modeling/input?select=fraudTrain.csv

2.2 API

Description:

The Google Reverse Geocoding API is a service provided by Google Maps that allows you to convert geographic coordinates (latitude and longitude) into human-readable addresses. It enables you to retrieve detailed location information based on the provided coordinates. When you make a request to the Reverse Geocoding API, it sends back a response in JSON or XML format containing the

address components and other relevant details associated with the given coordinates. The API can provide a range of information, including the street address, city, state, postal code, country, and more

Link : https://maps.googleapis.com/maps/api/geocode/json?latlng={lat},{lng}&key={api_key}

2.3 Website

Descrition:

The dataset consists of credit card transactions made by European cardholders in September 2013. The dataset covers a two-day period and contains a total of 200K+ transactions and 31 columns. The dataset is highly imbalanced, with fraud transactions accounting for only 0.172% of the total transactions.

The dataset primarily includes numeric input variables resulting from a PCA (Principal Component Analysis) transformation. Unfortunately, due to confidentiality concerns, the original features and additional background information about the data are not provided. The dataset includes principal components labeled as V1, V2, … V28, which are the outcomes of the PCA transformation. The 'Time' and 'Amount' features are exceptions and have not undergone the PCA transformation. The 'Time' feature represents the number of seconds elapsed between each transaction and the first transaction recorded in the dataset. The 'Amount' feature represents the monetary value of each transaction. The 'Amount' feature can be useful, particularly for approaches involving example-dependent cost-sensitive learning. The response variable, labeled as 'Class,' indicates whether a transaction is fraudulent (1) or not (0).

Link: https://datahub.io/machine-learning/creditcard/r/0.html

3 Relationships

The data from each source is connected as follows:

The flat file contains the core transaction data, including merchant information and fraud labels. I will get the BIN numbers from the Card numbers , create an amount range for Fraud and Non Fraud transactions.

The API will be used to get the merchant address details by providing the Latitude Location of Merchant and Longitude Location of Merchant from the each transactional records of flat file. I will define a function takes the latitude, longitude coordinates along with the API key as input parameters. It constructs the API URL with the provided coordinates, makes a GET request to the Google Geocoding API, and parses the JSON response. The function then extracts the formatted address from the response and returns it.

The website data will provide insights of distribution of Amount for Fraud and Nit fraud transactions along with Time. Then build the relationship with flat file data.

4 Plan to tackle the project

The primary focus is on collecting the credit card transaction data, along with associated information, from the defined sources (CSV, Web and API). The data will be carefully cleaned and

formatted to ensure consistency and accuracy. During this process,I will create some additional fields based on specific requirements to build relationships or for desired analysis.

Once the data is prepared, the relationships between the different data sources will be established. This may involve linking transaction details, merchant information, and contextual features to gain a comprehensive view of each credit card transaction. Proper data storage techniques will be employed to efficiently manage and access the data for analysis.

Finally, using Python I will create visualizations that provide meaningful insights from the data. These visualizations may include charts, graphs, and other interactive representations to help identify patterns, trends, and anomalies related to fraudulent transactions. Python's data visualization libraries, such as Matplotlib and Seaborn, will be leveraged to generate informative and visually appealing outputs.

5 Ethical implications & Challenges

Throughout the project, it is important to adhere to ethical considerations, including data privacy and security. Protecting sensitive information and ensuring compliance with relevant regulations is essential. Additionally, challenges such as dealing with imbalanced datasets, maintaining data integrity, and effectively communicating the results through visualizations may need to be addressed.