Credit Card Data Fraud

Jean Pacheco

University of Central Florida

Machine Learning

Final Project Report

Credit Card Data Fraud

Abstract

Credit card fraud is one of the many common challenges the financial world faces. This project seeks to provide aid to financial institutions to prevent fraud from occurring. It utilizes a database containing credit card transactions that describe different aspects of each transaction. This ranges from the location of the transaction to personal information of the cardholder. Hence, the research makes initial predictions that certain aspects, such as age, may be a key factor as to why and how individuals fall victim to such incidents. The project preprocesses this information to train classification and regression models to identify key factors that have the most correlation to fraudulent transactions. Once found, it provides problem-solving steps to deter the frequency of fraudulent transactions. Ultimately, it finds that the hour of the day when a transaction occurs could indicate higher chances of fraudulent activities. Therefore, it suggests implementing stricter verification processes to validate transactions.

Credit Card Data Fraud

Introduction

Credit card fraud is a daily occurrence that targets all walks of life. According to Bankrate, just in 2021 there were 389,737 credit card frauds reports which amounted to $32.34 billion (E. Ortiz, 2021). It disregards which socioeconomic class the victim belongs to. Financial banks face this issue as a challenge to their security systems and customer retention potential, making it one of the pillar objectives for these institutions. To maximize their customer's trust, banks must ensure to secure each transaction that their cardholders make to prevent any variation of identity theft or unintended asset loss. Therefore, institutions must assess where the focus is in preventing credit card fraud from occurring.

Hence, this research project collects data containing a combination of authentic and fraudulent credit card transactions. The objective is to find out which aspect of these transactions can indicate a strong correlation between a credit card transaction being fraudulent or not. It overlooks variables, such as personal information of the cardholder, where the transaction has occurred, and so on. Additionally, other variables are created from the existing ones to create higher correlations with the target at hand. With wrangled, cleansed, and prepared data, the research will indicate a pattern of how these fraudulent transactions may occur.

Important Definitions and Problem Statement

This research collects a database of credit card transactions that provides numerous indicators of information. These indicators describe: 1) Credit card information, such as credit card numbers. 2) Personal information, such as the cardholder's name, date of birth, and gender. 3) Transaction timestamp when the transactions occurred. 4) Location of transactions; in which

state the transactions were issued. 5) Merchant information, such as business name. 6) Transaction category, such as shopping, gas station, grocery store purchases, and so on.

With the database provided, the problem for this research to evaluate is to find which indicator is a common target for credit card fraud criminals. However, the challenge the proposed study faces are the uncertainty of how secure the credit card processors are. Whether someone has tampered with the devices, or maliciously collected credit card numbers to make fraudulent transactions. Also, all credit card numbers in the database are valid, therefore it is firsthand not known whether the numbers are unofficially fabricated.

As a prediction target, the research predicts that the indicator for personal information is most likely to be the critical target for credit card theft. Security measures are based on this indicator, criminals can impersonate victims and steal their belongings if they obtain their personal information without the victims' awareness. The model proposed breaks down this indicator and converts date of birth to age, to have a different perspective in age groups and evaluate age groups' likeliness to be prone to credit card theft. Also, it investigates gender trends whether which gender group population indicates more frequent occurrences of fraudulent transactions.

Overview of proposed approaches/systems

The research suggests a model that targets classification and regression tasks for this problem. It needs to recollect categorical information to make binary decision for the outcome, whether the transaction is fraudulent or valid.

The model follows a Machine Learning System Pipeline as follows: 1) Identify the problem, separating the training data with the test data for machine learning purposes. 2)

Analyzing, identifying patterns, and exploring the data. 3) Wrangling, preparing, and cleansing the data. 4) Modeling, predicting, and solving the problem. 5) Visualizing, reporting, and presenting the problem-solving steps as a solution for the problem.

Technical details of proposed approaches/systems

In the first step of the Machine Learning System Pipeline, the model focuses on identifying the problem and separating the training data with the test data. In this instance, the model collects two datasets. One consists of the training data about credit card transactions while the other contains the test data of similar transactions. Therefore, the model does not need to separate the data for machine learning purposes. Instead, it identifies the target variable, being the transaction's fraudulence status 'is_fraud'. This indicator describes whether the transaction was valid or fraudulent.

It transforms the variable 'age' to determine the date of birth of the cardholder. It does so by scanning through the dates within the 'date of birth' variable and comparing it to current day. By predating this number, it generates the exact age of everyone, creating the variable 'age'. It also transforms the 'trans_date_trans_time' variable. The model separates this variable into two, named 'weekday' and 'hour. With two new variables, the model can create correlations with the target variable and indicate which day of the week and hour are fraudulent transactions most likely to occur.

Thereafter, the model seeks to find which indicators generate relevant correlations with the target variable 'is_fraud'. The variables found are as follows: 'gender', 'age', 'weekday', 'state', 'hour', and 'category'. The analysis of the dataset reveals several important insights related to the occurrence of credit card fraud. Specifically, the 'gender' variable indicates a higher

likelihood of fraudulent transactions from male cardholders, as evidenced by a correlation score of 0.006426, compared to female cardholders with a score of 0.005262 (Figure 1). Additionally, the 'age' variable highlights a strong correlation between fraudulence and cardholders above 65 years old, with a correlation score of 0.0113, while individuals at the age of 97 exhibit a perfect correlation score of 1.000 (Figure 2). Therefore, age may warrant stricter scrutiny when assessing the validity of transactions from senior cardholders. Moreover, the 'state' variable shows that fraudulent activities are highly concentrated in Delaware (DE), with a correlation score of 1.000, while other states exhibit correlation scores ranging from 0 to less than 1 (Figure 3). The 'weekday' variable reveals that fraudulent transactions are more likely to occur on Fridays, with a correlation score of 0.007086 (Figure 4), while the 'hour' variable indicates that fraudulent activities tend to occur during later hours of the day, with a peak at 10PM and a correlation score of 0.028829 (Figure 5). Lastly, the 'category' variable highlights that online shopping is the most vulnerable area for credit card fraud (Figure 6). Based on the strength of their correlation scores with the target variable, the model ranks the variables in a descending order: 'hour' (0.098484), 'weekday', 'age', 'state', 'category', and 'gender' (-0.223520) (Figure 7). Taken together, these findings underscore the need for increased surveillance and vigilance during select times of the day to effectively combat credit card fraud.

Hence, sending out two-factor authentication confirmation prompts to credit card holders at the time of transactions exhibits a commendable degree of merit and warrants serious consideration.

Experiments

Upon obtaining the relevant variables, the model employs a systematic approach whereby it eliminates all other insignificant variables from the database. Given that the model frames the problem at hand as a classification and regression task, it proposes various machine learning algorithms and applies them to the test dataset to evaluate the effectiveness of the preprocessing methods. Specifically, the Logistic Regression, Gaussian Naïve Bayes, Stochastic Gradient Descent, and Perceptron algorithms yield a score of 99.42, whereas the K-nearest neighbor algorithm achieves a slightly higher score of 99.63. Notably, both the Decision Tree and Random Forest algorithms exhibit the highest score of 99.79. However, based on the research's preference for a more robust approach that combines the outcome of multiple decision trees, the study relies on the Random Forest algorithm as the preferred choice (Figure 8).

Conclusion

Credit card fraud poses a significant threat to the financial world, with the potential to impact anyone, and therefore demands critical attention from financial institutions. To ensure customer retention and gain their trust, these institutions must prioritize securing their credit card transactions to the greatest extent possible. Considering this, this research offers useful insights into the time of day when transactions are processed and the steps that institutions can take to prevent suspicious activity. By implementing multi-factor verification prior to the completion of a transaction, institutions can effectively safeguard their clients' financial well-being and prevent any potential inconveniences from occurring.

References

E. Ortiz, "Credit Card Fraud Statistics," Bankrate, 06-Jul-2021. [Online]. Available: https://www.bankrate.com/finance/credit-cards/credit-card-fraud-statistics/#fraud

Data retrieval

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Code

https://github.com/jcp1109/Credit-Card-Fraud/blob/main/Credit_Card_Fraud.ipynb

Illustrations (Cont. on following page)

Figure 1

	gender	is_fraud
1	M	0.006426
0	F	0.005262

Figure 2

	age	is_fraud
78	97	1.000000
66	85	0.030047
71	90	0.014511
46	65	0.013756
72	91	0.013494
...	...	...
30	49	0.003338
26	45	0.003332
28	47	0.003246
20	39	0.003188
79	98	0.000000

80 rows × 2 columns

Credit Card Data Fraud

Figure 3

	state	is_fraud
8	DE	1.000000
39	RI	0.027273
0	AK	0.016981
33	NV	0.008382
5	CO	0.008141
37	OR	0.008012
42	TN	0.007975
29	NE	0.007448
21	ME	0.007210
30	NH	0.007127

Figure 4

	weekday	is_fraud
5	5	0.007086
4	4	0.006844
3	3	0.006554
6	6	0.006106
2	2	0.005835
0	0	0.004853
1	1	0.004648

Figure 5

	hour	is_fraud
22	22	0.028829
23	23	0.028374
1	01	0.015349
0	00	0.014940
2	02	0.014652
3	03	0.014239
5	05	0.001423
7	07	0.001327

Figure 6

	category	is_fraud
11	shopping_net	0.017561
8	misc_net	0.014458
4	grocery_pos	0.014098
12	shopping_pos	0.007225
2	gas_transport	0.004694
9	misc_pos	0.003139

Figure 7

	Feature	Correlation
4	hour	0.098500
5	weekday	0.058183
3	age	0.016813
2	state	0.002316
0	category	-0.219341
1	gender	-0.223129

Figure 8

	Model	Score
0	Random Forest	99.79
6	Decision Tree	99.79
2	KNN	99.63
1	Logistic Regression	99.42
3	Naive Bayes	99.42
4	Perceptron	99.42
5	Stochastic Gradient Decent	99.42