International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

CREDIT CARD FRAUD ANALYSIS USING PREDICTIVE MODELLING

Dr .S. MURALIKRISHNA[1]

[1]Associate professor Department Of ECE Bomma institute of technology and science Telangana state

ABSTRACT

Billions of dollars of loss are caused every year by fraudulent credit card transactions. The design of efficient fraud detection algorithms is key for reducing these losses, and more and more algorithms rely on advanced machine learning techniques to assist fraud investigators. The design of fraud detection algorithms is however particularly challenging due to the no stationary distribution of the data, the highly unbalanced classes distributions and the availability of few transactions labeled by fraud investigators. At the same time public data are scarcely available for confidentiality issues, leaving unanswered many questions about what is the best strategy. In this thesis we aim to provide some answers by focusing on crucial issues such as: i) why and how under sampling is useful in the presence of class imbalance (i.e. frauds are a small percentage of the transactions), ii) how to deal with unbalanced and evolving data streams (non-stationary due to fraud evolution and change of spending behavior), iii) how to assess performances in a way which is relevant for detection and iv) how to use feedbacks provided by investigators on the fraud alerts generated. Finally, we design and assess a prototype of a Fraud Detection System able to meet real-world working conditions and that is able to integrate investigators' feedback to generate accurate alerts.

Index Terms— credit card, fraud detection, online shopping, e-commerce , logistic regression

INTRODUCTION

The online shopping growing day to day. Credit cards are used for purchasing goods and services with the help of virtual card and physical card where as virtual card for online transaction and physical card for offline transaction. In a physical-card based purchase,the cardholder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of purchase, an attacker has to steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company.

In online payment mode, attackers need only little information for doing fraudulent transaction (secure code, card number, expiration date etc.). In this purchase method, mainly transactions will be done through Internet or telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details.

Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyses the spending patterns on every card and to figure out any inconsistency with respect to the "usual" spending patterns. Fraud detection based on the analysis of existing

purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds.

Since humans tend to exhibit specific behavioristic profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patternsis a potential threat to the system.

Existing System

This was on k-means Algorithm implementation, Only the two features with the most variance were used to train the model. The model was set to have 2 clusters, 0 being non- fraud and 1 being fraud. We also experimented with different values for the hyper parameters,but they all produced similar results. Changing the dimensionality of the data (reducing it to more dimensions than 2) also made little difference on the final values.

Disadvantages

The Clustering doesn't produce the less accuracy when compared to Regression methods in scenarios like credit card fraud detection. Comparatively with other algorithms k-means produce less accurate scores in prediction in this kind of scenarios

Proposed System

Our goal is to implement machine learning model in order to classify, to the highest possible degree of accuracy, credit card fraud from a dataset gathered from Kaggle. After initial data exploration, we knew we would implement a logistic regression model for best accuracy reports. Logistic regression, as it was a good candidate for binary classification. Python skarn library was used to implement the project, We used Kaggle datasets for Credit card fraud detection, using pandas to data frame for class ==0 for no fraud and class==1 for fraud, matplotlib for plotting the fraud and non fraud data, train_test_split for data extraction (Split arrays or matrices into random train and test subsets) and used Logistic Regression machine learning algorithm for fraud detection and print predicting score according to the algorithm. Finally Confusion matrix was plotted on true and predicted.

Advantages

- The results obtained by the Logistic Regression Algorithm is best compared to any otherAlgorithms.
- The Accuracy obtained was almost equal to cent percent which proves using of Logisticalgorithm gives best results.

Problem Statement

Credit card fraud stands as major problem for word wide financial institutions. Annual lost due to it scales to billions of dollars. We can observe this from many financial reports. Such as (Bhattacharyya et al., 2011) 10th annual online fraud report by Cyber Source shows that estimated loss due to online fraud is $4 billion for 2008 which is 11% increase than $3.6 billion loss in 2007and in 2006, fraud in United Kingdom alone was estimated to be £535 million in 2007 and now costing around 13.9 billion a year (Mahdi et al., 2010). From 2006 to 2008, UK alone has lost £427.0 million to £609.90 million due to credit and debit card fraud (Woolsey &Schulz, 2011). Although, there is some decrease in such losses after implementation of detection and prevention systems by government and

bank, card-not- present fraud losses are increasing at higher rate due to online transactions. Worst thing is it is still increasing un-protective and un-detective way.

Over the year, government and banks have implemented some steps to subdue these frauds butalong with the evolution of fraud detection and control methods, perpetrators are also evolving their methods and practices to avoid detection. Thus an effective and innovative methods need to be develop which will evolve accordingly to the need.

Scope

The online shopping growing day to day. Credit cards are used for purchasing goods and services with the help of virtual card and physical card where

as virtual card for online transaction and physical card for offline transaction. In a physical-card based purchase, the cardholder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of purchase, an attacker has to steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In online payment mode, attackers need only little information for doing fraudulent transaction (secure code, card number, expiration date etc.). In this purchase method, mainly transactions will be done through Internet or telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details. Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information.

The only way to detect this kind of fraud is to analyses the spending patterns on every card and to figure out any inconsistency with respect to the "usual" spending patterns. Fraud detection based on the analysis of existing purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behavioristic profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system.

Objective

Billions of dollars of loss are caused every year by fraudulent credit card transactions. The design of efficient fraud detection algorithms is key for reducing these losses, and more and more algorithms rely on advanced machine learning techniques to assist fraud investigators. The design of fraud detection algorithms is however particularly challenging due to the non-stationary distribution of the data, the highly unbalanced classes distributionsand the availability of few transactions labeled by fraud investigators. At the same time public data are scarcely available for confidentiality issues, leaving unanswered many questions about what is the best strategy.

In this thesis we aim to provide some answers by focusing on crucial issues such as:

why and how under sampling is useful in the presence of class imbalance (i.e. frauds are asmall percentage of the transactions)

how to deal with unbalanced and evolving data streams (non-stationarity due to fraudevolution and change of spending behavior)

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

how to assess performances in a way which is relevant for detection

how to use feedbacks provided by investigators on the fraud alerts generated

LITERATURE SURVEY

Literature survey

For maintaining associate info system, fraud detection is an important part of it. By ancient access management mechanisms, the management system will give intrusion hindrance to an extent, they are syntactically correct however transactions are semantically damaged.

[1]. Chung et al. say that in the info system the misuse detection is not self-addressed and proposed DEMIDS, based on the audit logs it derives user profiles. The sphere of the theoryof games has been explored for issues starting from auctions to chess and its application to the domain of knowledge warfare looks promising.

[2]. The theory of games in IW was brought by prophet et al. To predict future attacks.

[3] And also the challenges and variations during this domain, one will utilize a well-developed theory of games algorithms.

REQUIREMENTS ANALYSIS

Hardware Requirements

For developing the application following are the Hardware Requirements :

- RAM : 4GB and Higher
- Processor : Intel i3 and above
- Hard Disk : 500GB(Minimum)

Software Requirements

For developing the application following are the Software Requirements :

- OS : Windows or Linux
- Python IDE : python 2.7.x and above Pycharm IDE Required , jupyter notebook.
- Language : Python Scripting Setup tools and pip to be installed for 3.6 and above

Technologies and Languages used to Develop

Python

Functional Requirements

In this module, there are n numbers of users are present. User should register before doing some. After registration successful he can login by using valid user name and password. Login successful he will do some operations like view login user profile details, search city and view historical places in that city, and user can give tweet and ratings, view previous visit user history, and user add trips, view all previous users added trip details

Modules

Tensor flow

Tensor Flow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google.

Tensor Flow was developed by the Google Brain team for internal Google use. It wasreleased under the Apache2.0 open-source license on November 9, 2015.

Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object,and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

featuresincluding these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

IMPLEMENTATION

PYTHON

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++or Java. It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many operating systems. C Python, the reference implementation of Python, is open source software and has a community- based development model, as do nearly all of its variant implementations. C Python is managed by the non-profit Python Software Foundation. Python features a dynamic type system and automatic memory management. It supports multiple programming including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.

Interactive Mode Programming

Invoking the interpreter without passing a script file as a parameter brings up the following prompt

–

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document orit can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to befree from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that theuser will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

- Convey information about past activities, current status or projections of the
- Future.
-

SCREENS AND REPORTS

In the above Screen We can click the upload credit card dataset and then loaded thedataset

Upload Credit Card Data Set: The file can be upload here.

Conclusion

This machine learning fraud detection tutorial showed how to tackle the problem of credit card fraud detection using machine learning. It is fairly easy to come up with asimple model, implement it in Python and get great results for the Credit Card FraudDetection task on Kaggle.

Future work

Volume 3.

This process is used to detect the credit card transaction, which are fraudulent or genuine. Data mining techniques of Predictive modeling, Decision trees and LogisticRegression are used to predict the fraudulent or genuine credit card transaction. In predictive modeling to detect and check output class distribution. The prediction model predicts continuous valued functions. We have to detect 148 may be fraud and other are genuine. In decision tree generate a tree with root node, decision node and leaf nodes. The leaf node may be 1 becomes fraud and 0 otherwise. Logistic Regression is same as linear regression but interpret curve is different. To generalize the linear regression model, when dependent variable is categorical and analyzes relationship between multiple independent variables.

Bibliography

[1] Salazar, Addison , et al. "Automatic credit card fraud detection based on non-linearsignal processing

[2] Delamaire , Linda, H. A. H. Abdou, and John Pointon. "Credit card fraud and detectiontechniques: a review." *Banks and Bank systems* 4.2 (2009): 57-68.

[3] Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1.1 (1986): 81-106.

[4] *Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man-Machine Studies.* ***27*** *(3): 221. doi:10.1016/S0020-7373(87)80053-6.*

[5] K. Karimi and H.J. Hamilton (2011), "Generation and Interpretation of Temporal Decision Rules", International Journal of Computer Information Systems and IndustrialManagement Applications,