Abstract

Fraud is one of the major ethical issues in the credit card industry.The main aims are, firstly, to identify the different types of credit card fraud,and, secondly,to review alternative techniques that have been used in fraud detection. The sub-aim is to present,compare and analyze recently published findings in credit card fraud detection.This article defines common terms in credit card fraud and highlights key statistics and figures in this field. Depending on the type of fraud faced by banks or credit card companies,various measures can be adopted and implemented. The proposals made in this paper are likely to have beneficial attributes in terms of cost savings and time efficiency. The significance of the application of the techniques reviewed here is in the minimization of credit card fraud. Yet there are still ethical issues when genuine credit card customers are misclassified as fraudulent.

Contents

Chapter 1

Introduction

1.1 Background

Now a day the usage of credit cards has dramatically increased. As credit card becomes the most popular mode of payment for both online as well as regular purchase, cases of fraud associated with it are also rising. In this paper, we model the sequence of operations in credit card transaction processing using a Logistic regression and show how it can be used for the detection of frauds. An Logistic regression model is initially trained with the normal behavior of a cardholder. If an incoming credit card transaction is not accepted by the trained Logistic regression model with sufficiently high probability, it is considered to be fraudulent. At the same time, we try to ensure that genuine transactions are not rejected. We present detailed experimental results to show the effectiveness of our approach and compare it with other techniques available in the literature.

1.2 Motivation and Social Impact

While performing online transaction using a credit card issued by bank, the transaction may be either Online Purchase or transfer .The online purchase can be done using the credit or debit card issued by the bank or the card based purchase can be cat egorized into two types Physical Card and Virtual Card. In both the cases if the card or card details are stolen the fraudster can easily carry out fraud transactions which will result in substantial loss to card holder or bank. In the case of Online Fund Transfer a user makes use of details such as Login Id, Password and transaction password. Again here if the details of the account be miss used then, as a result, it which will give rise to fraud transaction. Credit card fraud is a wide - ranging term for th eft and fraud committed using a credit card or any similar payment mechanism as a fraudulent source of funds I a transaction. The purpose may be to obtain goods without paying, or to obtain unauthorized funds from an account. Credit card fraud is also an a djunct to identity theft. The fraud begins with either the theft of the physical card or the compromise of data associated with the account, including the card account number or other information that would routinely and necessarily be available to a merc hant during a legitimate transaction. The compromise can occur by many common routes and can usually be conducted without tipping off the card holder, the merchant or the issuer, at least until the account is ultimately used for fraud. A simple example is that of a store clerk copying sales receipts for later use. The rapid growth of credit card use on the Internet has made database security laps es particularly costly; in some cases, millions of accounts have been compromised.

1.3 Objectives and Outcomes

- the fraud transaction risks using credit card is a main problem which should be avoided.
- In proposed system, we present a Logistic regression a well-established statistical method for predicting binomial or multinomial outcomes.
- The detection of the fraud use of the card is found much faster than the existing system.
- In case of the existing system even the original card holder is also checked for fraud detection. But in this system no need to check the original user as we maintain a log.
- The log which is maintained will also be a proof for the bank for the transaction made.
- We can find the most accurate detection using this technique.
- This reduce the tedious work of an employee in the bank

Chapter 2

Literature Survey

2.1 Existing Techniques

2.1.1 Neural Network

A neural network based fraud detection system was trained on a large sample of labelled credit card account transactions and tested on a holdout data set that consisted of all account activity over a subsequent two-month period of time. The neural network was trained on examples of fraud due to lost cards, stolen cards, application fraud, counterfeit fraud, mail-order fraud and NRI (non-received issue) fraud. The network detected significantly more fraud accounts (an order of magnitude more) with significantly fewer false positives (reduced by a factor of 20) over rule-based fraud detection procedures.

2.1.2 Hidden Markov model

It does not require fraud signatures and yet is able to detect frauds by considering a cardholder's spending habit. Card transaction processing sequence by the stochastic process of an HMM. The details of items purchased in Individual transactions are usually not known to any Fraud Detection System(FDS) running at the bank that issues credit cards to the cardholders. Hence, we feel that HMM is an ideal choice for addressing this problem. Another important advantage of the HMM-based approach is a drastic reduction in the number of False Positives transactions identified as malicious by an FDS although they are actually genuine. An FDS runs at a credit card issuing bank. Each incoming transaction is submitted to the FDS for verification. FDS receives the card details and the value of purchase to verify, whether the transaction is genuine or not. The types of goods that are bought in that transaction are not known to the FDS. It tries to find any anomaly in the transaction based on the spending profile of the cardholder, shipping address, and billing address, etc. If the FDS confirms the transaction to be of fraud, it raises an alarm, and the issuing bank declines the transaction.

2.1.3 BLAST-SSAHA

BLAST-SSAHA Hybridization is one of the most effective and cheapest ways to detect a credit card fraud transaction. BLAST stands for Basic Local Alignment Search Tool whereas SSAHA stands for Sequence Search and Alignment by Hashing Algorithm. the working of BLAST - SHAHA Hybridization in Credit Card Fraud Detection System. Whenever a new transaction is initiated, BLAST - SHAHA Hybridization decides whether the transaction is authentic or fake.Thus; it avoids the loss of money and impedes the frauds ter from making the fraud transaction successful. A Time Amount (TA) sequence is been created by merging the incoming sequence with the sequence that are in the Customer

2.2 Comparison of Existing Techniques

Sr.no	Supervised detection technique	Unsupervised detection technique
1	In this paper, we discuss about the Supervised learning where you have input variables(x) and an output variable(Y) and you use an algorithm to learn the mapping function from the input to the output.Y = f(X)	Unsupervised learning is where you only have input data (X) and no corresponding output variables.
2	TThe goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.	The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.
3	It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.	These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.
4	Supervised learning problems can be further grouped into regression and classification problems. Classification: A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease". Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight".	Unsupervised learning problems can be further grouped into clustering and association problems. Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior. Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.
5	Examples of Supervised learning:neural networks, fuzzy neural nets, and combinations of neural nets,Bayesian learning neural network	Examples of Unsupervised learning: Peer Group Analysis and Break Point Analysis

Table 2.1: comparison

2.3 Logistic Regression

The two data mining approaches, are support vector machines and random forests, together with the well known logistic regression, as part of an attempt to detect the credit card fraud. It is well-understood, easy to use, and it is most commonly used for data-mining. Thus it provides a useful baseline for comparing performance of newer methods.Data mining tasks has more and more statistical model that involves discriminant analysis,regression analysis,multiple -logistic regression,etc. Logistic regression(LR) is useful for situations in which we want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a linear regression model but is suited to models where the dep endent variable is dichotomous. Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model and it is applicable to a broader range of research situations than feature analysis.

2.3.1 Steps

- Descriptive Analyses: to gain better insight into the data.
- Univariable Analyses: to test unconditional associations of the variables with the outcome.
- Testing of Collinearity: to test associations/correlations between explanatory variables.
- Multivariable Analyses: to test association of a variable after adjusting for other variables or confounders.
- Model diagnostics: to assess whether the final model fulfils the assumptions it was based on.

2.3.2 Algorithm

- Logistic Function: 1 / (1 + $e^{-}value$)
- Logistic regression equation: y = $e^{(}b0 + b1 * x)/(1 + e^{(}b0 + b1 * x))$

  Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.
- Predicts Probabilities: P(X) = P(Y=1|X)
- Learning the Logistic Regression Model
- Making Predictions with Logistic Regression:y = $e^{(}b0 + b1 * X)/(1 + e^{(}b0 + b1 * X))$

Chapter 3

Implementation

3.1 Flow of Work

- Descriptive Analysis

  During descriptive analysis we aim to gain an understanding of the distribution of data for each variable. Inspect distributions of all quantitative variables by creating histograms or box-and-whisker plots, and of all categorical variables by creating frequency tables or bar-charts.These analyses will give you a preliminary idea about the association of the explanatory variables with the outcome.

- Another step is Univariable Analyses

  During univariable analyses we test the association of one explanatory variable at a time with the outcome without worrying about other variables or confounders (unconditional association). This is essential in order to shortlist variables for multivariable analysis, especially if there are a large number of explanatory variables. It also excludes the variables from further analysis that do not show any significant association with the outcome on their own as they are not likely to be associated with the outcome after adjusting for other variables.

- Testing of collinearity

  If two explanatory variables are highly correlated with each other, they can cause problems during multivariable analysis because they are explaining almost the same variability in the outcome. Therefore, it is beneficial to examine associations/correlation between explanatory variables and exclude one of a pair of highly correlated variables before conducting multivariable analysis.

- Multivariable Analyses.

This is the real model building stage; however, you are likely to get erroneous results if you jump straight to multivariable analysis without carrying out the previous steps. In this step we test associations of variables with the outcome after accounting for other variables and confounders.

- Model Diagnostics

  After verifying goodness-of-fit tests, it is time to scrutinise whether your model meets the assumptions on which it was based. The most important of these (for logistic regression) is the assumption of linearity for any continuous variable in the model. Consult some resources to get information about how to evaluate these assumption and consult an experienced statistician, if this condition is not met.

3.2 Data collection and Data sets

This project uses python as a language with different header.In this model we first train the data with some data set and check the accuracy after each completion. I have taken data from online source "candle" site.

3.3 Results obtained

First we train the model by giving a data set and dividing given set into 5 K-folds. WE then test the four folds on the 5th one ,this is done 5 times.By doing so we get the Accuracy=1. Once the training is completed then the model is tested with different set of value and at the end accuracy is measured. So, in other step we test our model with different data set and get the accuracy of about 99 percent

Chapter 4

Results and Discussion

4.1 Discussion on Result Obtained

Fraud detection involves identifying Fraud as quickly as possible once it has been perpetrated. Fraud detection methods are continuously develope d to defend criminals in adapting to their strategies. The development of new fraud detection methods is made more difficult due to the severe limitation of the exchange of ideas in fraud detection. Data sets are not made available and results are often no t disclosed to the public. The fraud cases have to be detected from the available huge data sets such as the logged data and user behavior. At present, fraud detection has been implemented by a number of methods such as data mining, statistics, and artific ial intelligence. Fraud is discovered from anomalies in data and patterns. The detection of the fraud use of the card is found much faster that the existing system. In case of the existing system even the original card holder is also checked for fraud detection. But in this system no need to check the original user as we maintain a log. The log which is maintained will also be a proof for the bank for the transaction made. We can find the most accurate detection using this technique. This reduce the tedious work of an employee in the bank

4.2 Comparison of Results (with other researchers)

- **Fuzzy Darwinian**

  The Fuzzy Darwinian fraud detection systems improve the system accuracy.

- **Neural Network**

  The Neural Network based CARDWATCH shows good accuracy in fraud detection and processing Speed.

- **HMM**

  The fraud detection rate of Hidden Markov model is very low compare to other methods.

- **BLAST-SSAHA**

  The processing speed of BLAST-SSAHA is fast enough to enable on-line detection of credit card fraud.

- **BLAH-FDS**

  BLAH-FDS can be effectively used to counter frauds in other domains such as telecommunication and banking fraud detection.

Chapter 5

Conclusion and Future Work

- Currently, building a precise, accessible and simple handling credit card risk monitoring system is one of the key tasks for the merchant banks, organization to improve merch ants risk management level in an automatic, scientific and adequate wa y. In this paper, we demonstrate various techniques used in credit card fraud detection and their advantages with data mining techni ques including neural networks, and confidence value calculation. Further more studies are encouraged to improve the fraud detection basis to set more suitable weight and cost factor with both good tested accuracy and detection accuracy. More e fficient credit card fraud detection system / model an important r equirement for any card issuing b ank. Credit card fraud detection has drawn number of techniques, system, and models that have been proposed to counter credit fraud and lot of interest from the research community. The neural network based CARDWATCH shows m uch great accuracy in fraud detection and processing speed is also high but it is limited to one - network per customer. The Fuzzy Darwinian fraud detection systems (FDFDS) improve accuracy of the system . Since The Fraud detection rate (FDR) of Fuzzy Darwini an fraud detection systems in terms of true positive (TPR) is 100presents good results in detecting fraudulent transactions. The Fraud detection rate (FDR) of Hidden Markov model (HMM) is very low as compare d to other existing methods. P rocessing spe ed of BLAST - SSAHA is fast enough to enable on - line detection of credit card fraud. All the techniques of credit card fraud detection discussed in this survey paper have its own strengths as well as weaknesses and adva ntages along with disadvantages. S urve y of such kind will enable us to build a hybrid approach for fraudulent credit card transactions identification . In daily life, e very field of the daily life, credit card fraud has become much more important and popular. Building an accurate and efficient credit card fraud detection system t o improve security of the financial transaction is one of the key tasks for the financial institutions. In this paper , we determine 13 classification methods were used to build fraud detecting models / system . This work d emonstrates the advantages of applying the data mining techniques including ANN and LR , BN techniques to the credit card fraud detection problem for the purpose of reducing the banks or financial risk s . Yet, as the distribution of the training data sets b ecome more biased, then the performance of all models decrease in catching the fraudulent transactions.

- As a future work; instead of making performance , the cost based ones comparisons just over the prediction accuracy and TPR/FPR, these comparisons will b e extended to include the comparisons over other performance metrics.