PES University, Bengaluru

UE20CS312 -Data Analytics

CREDIT CARD FRAUD DETECTION SYSTEM

Literature Review

Submitted By: -

S J Tharun	**PES2UG20CS289**
Sanjay Sagarad	**PES2UG20CS312**
Sathyanarayana R K	**PES2UG20CS315**

Abstract

Credit cards, which enable cashless transactions, are a common form of offline and online payment. Making money and doing other activities is simple, practical, and fashionable. Along with technological advancement, credit card theft is also on the rise. Additionally, it may be claimed that as worldwide communication has improved, economic fraud has dramatically increased. Every year, billions of dollars in losses are attributed to these fraudulent activities. These operations are carried out so tastefully that they resemble real business transactions.

Simple pattern-related techniques and other less sophisticated approaches won't thus be effective. All banks now require an effective fraud detection tool in order to reduce confusion and restore order.

Keywords—Classifier; logistic regression; accuracy; smoothing; artificial intelligence; cross validation.

INTRODUCTION

The goal of fraud, according to its definition, is to obtain material or financial benefit by deceit. Based on this, the two key strategies for preventing fraud-related loss are fraud detection and prevention. Fraud detection is the process of identifying fraudulent transactions carried out by ***fraudsters***, whereas fraud prevention is the proactive strategy for preventing the occurrence of fraudulent activities. Several types of payment cards are presently readily accessible, including credit, charge, debit, and prepaid cards. In certain nations, they are the most widely used form of payment . Indeed, improvements in how we manage money have been made possible by developments in digital technologies.
Particularly with regard to payment techniques that have transitioned from being physical activities to digital activities utilising technological means . This has completely changed how monetary policy is practised, as well as how both big and small businesses conduct their commercial operations. The use of a credit card fraudulently to make a purchase of a good or service is known as credit card fraud. Physical or digital execution of these transactions is both possible . The credit card is actually present when a transaction is made in person. Conversely, digital transactions happen via the phone or internet.

Typically, a cardholder will use a website or phone call to submit their card number, card verification number, and expiration date. Credit card use has significantly increased as e-commerce has grown rapidly in recent years. Approximately 317 million credit card transactions were made in Malaysia in 2011; by 2018 , this figure had risen to 447 million. According to, global credit card theft in 2015 hit a record high of $21.84 billion. With more people using credit cards, there have been an increasing number of fraud occurrences. Despite the use of several verification techniques, the quantity of credit card fraud instances has not greatly decreased .

The banking industry has been severely impacted by the recent rise in credit card theft. Merchants are the ones that suffer losses as a result of credit card theft since they are responsible for paying all associated costs, such as card issuer fees, administrative costs, and other charges [9]. The merchants

are responsible for any losses, which results in higher product pricing and lower discounts. Therefore, minimising this loss is crucial.

A. Problem Statement

World Bank estimates that 10,000 transactions using credit cards happen every second worldwide. Credit cards are becoming the main targets of fraud due to the high transaction frequency.

Credit card businesses have been actively battling fraud since the Diners Club issued its first credit card in 1950. Credit card fraud causes billions of dollars in direct losses each year. Fraud cases can happen in a variety of situations, such as at ports of sale (POSs), online, or over the phone, often known as card-not-present (CNP) instances or transactions involving lost or stolen cards.

Thus, the total cost of credit card fraud in 2015 was $21.84 billion, of which $15.72 billion was paid for by issuers.

According to data from the European Central Bank, in 2012, CNP transactions accounted for the majority (60%) of fraud, with POS terminals accounting for the remaining 23%. Fraud has a huge economic cost both worldwide and locally in Malaysia. In 2016, there were 383.8 million credit cards, 107.6 million debit cards, and 4.1 million charge cards. By 2018, those numbers had risen to 447.1 million, 245.7 million, and 5.2 million.

When using credit, debit, and charge cards, the overall proportion of fraudulent payments was 0.0186% in 2016 and rose to 0.0256% in 2018. Fraudsters have possibilities because of the possibility for substantial financial advantages and the dynamic structure of financial services.

Numerous studies have demonstrated that as the use of credit cards for financial transactions increases, so does the prevalence of fraud. Because of the attackers' or hackers' growing ability, increased the issue since these individuals can abuse security flaws to access users' sensitive information using their credit details for nefarious purposes, such as deceit.Fig. 1 provides a correct definition of this issue. The typical situation while doing credit card fraud

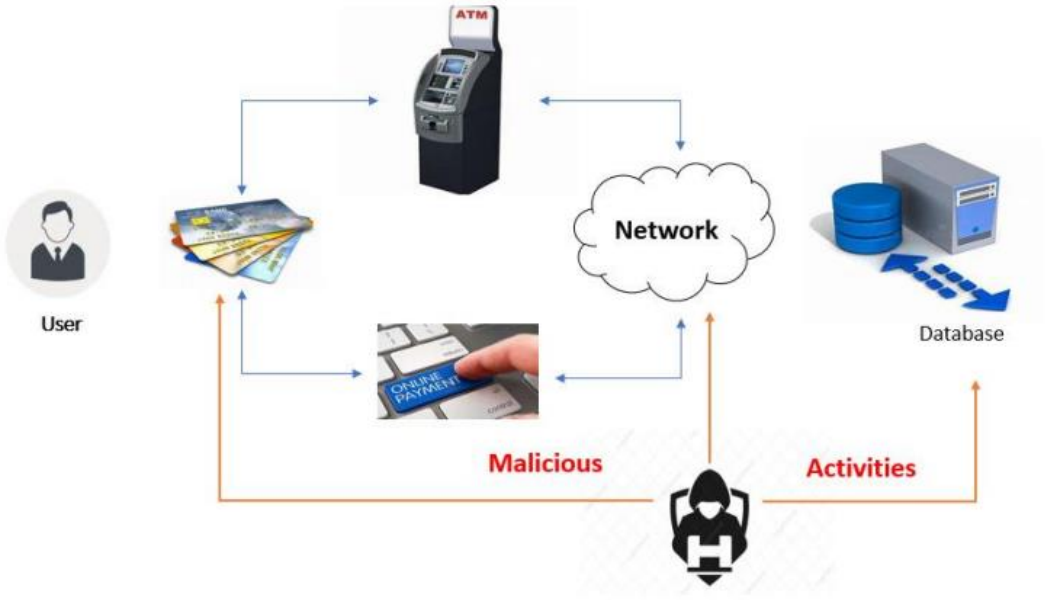

Fig. 1. General Scenario of Online Fraud.

Source:google

On the basis of the empirical evidence, the following research questions are developed to guide this study and meet its objectives.

• How can a fraud detection system be built using AI that can deal with imbalanced data effectively?

• How can we smooth (or clean) the data before using it for training the machine to ensure high detection accuracy?

- How can the bank or the service provider prevent these frauds using the available data?

B. Deliverables

The following is a summary of the work's deliverables:

- A data analytical system for fraud detection is proposed. The system uses logistic regression to build a classifier called the LogR classifier. The LogR classifier has the ability to deal with imbalanced data and adapt to the behaviour of the user by employing the cross-validation technique.
- Two major strategies are employed to clean the data in order to achieve high accuracy detection. Missing data are handled by the mean-based technique.

C. Background

A. Confusion matrix

The performance of the classification models for a certain set of test data is evaluated using a matrix called the confusion matrix. Only after the real values of the test data are known can it be determined. Although the matrix itself is simple to understand, some of the associated terms could be unclear.

- For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.
- The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with the total number of predictions.
- Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.
- It looks like the below table:

n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

Source: java point

B. Logistic regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

Based on the number of categories, Logistic regression can be classified as:

Binomial: Target variable can have only 2 possible types: “0” or “1” which may represent “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc.
Multinomial: Target variable can have 3 or more possible types which are not ordered (i.e. types have no quantitative significance) like “disease A” vs “disease B” vs “disease C”.
Ordinal: It deals with target variables with ordered categories. For example, a test score can be categorized as: “very poor”, “poor”, “good”, “very good”. Here, each category can be given a score like 0, 1, 2, 3.

The general form of logistic regression equation is defined as:

$$\log [y/1-y]=a_0+a_1\times x_1+a_2\times x_2+\cdots a_k\times x_k$$

Reference:

- Research paper from google scholar
- Geekesforgeeks.com
- Javapoint.com

Click Here for GitHub repository.

https://github.com/tharunj-droid/mini-project-Data-analysis