International Research Journal of Engineering and Management Studies (IRJEMS)
Volume: 03 Issue: 04 | April -2019 ISSN: 1847-9790 || p-ISSN: 2395-0126

Big Data For Fraud Detection on E-Commerce Application

Mayuri Gawatre[1], Madhuri Thakare[2], Maheshwari Yergude[3] ,Bhavana Kinhekar[4]

[1] *mayurigawatre@gmail.com*
[2]*tmadhuri555@gmail.com*
[3]*mahiyergude@gmail.com*
[4]*bhavanakinhekar@gmail.com*

Abstract: *The concept of exchanging goods and services over the Internet has seen an exponential growth in popularity over the years. The Internet has been a major breakthrough of online transactions, leaping over the hurdles of currencies and geographic locations. The increase in online transactions has been added with an equal increase in the number of attacks against security of online systems. Auction sites and e-commerce web applications have seen an increase in fraudulent transactions. Some of these fraudulent transactions that are executed in e-commerce applications happen due to successful computer intrusions on these web sites. Although a lot of awareness has been raised about these facts, there has not yet been an effective solution to adequately address the problem of application-based attacks in e-commerce.*

***Key Words*: FDS, HDFS**

1.Introduction

The volume of electronic transactions has risen significantly in last year's, mainly due to the popularization of electronic commerce (e-commerce), such as online retailers (e.g., Amazon.com, eBay, Ali Express.com). We also observe a significant increase in the number of fraud cases, resulting in billions of dollars losses each year worldwide. Therefore it is important and necessary to developed and apply techniques that can assist in fraud detection and prevention, which motivates our research. This work aims to apply and evaluate computational intelligence techniques (e.g., Data mining and machine learning) to identify fraud in electronic transactions, more specifically in credit card operations performed by Web payment gateways. In order to evaluate the techniques, we apply and evaluate them in an actual dataset of the most popular Brazilian electronic payment service.The Internet has been a major breakthrough of online transactions, leaping over the hurdles of currencies and geographic locations. However, the anonymous nature of the Internet does not promote an idealistic environment for transactions to occur. The increase in online transactions has been added with an equal increase in the number of attacks against security of online systems.The aim of this project is to develop a Fraud Detection System based on anomaly intrusion detection E-commerce application. The goal is to reduce the number of fraudulent transactions perpetrated through computer intrusion attacks in e-commerce sites. The objective is to use data mining models to detect anomalies as a second line of defense, when preventive methods fail.

1.1 Project Objective

To develop an online ecommerce system
To implement intrusion detection system using big data
To implement hybrid database scheme for intrusion detection

1.2 Project Perspective

In this, we have the online Ecommerce application and depend on that how we are detecting the fraud happen in Transaction on that Ecommerce application. It is a web-based system. It can access using Internet Explorer 8.0, MozillaFirefox 2.0 and Google Chrome browser.

Interfaces:

It contains two interfaces:

User interfaces: In it users are able to view the Ecommerce application. He is able to access shopping site, can browse the different categories, browse many products and add any number of products in his cart. Once he want to do shopping, then he will go for sign in or registration. As user log in to the application he will be able to go to the cart go for shopping select the payment option and complete his transaction.

Admin interface: Admin is able to view all the users ,as user placed the order, admin can be able to dispatch the order, confirm the order and he will be able to check for weather the transaction made by the user is correct or not . He will be able to access the data can able to view the number of fraudular transactions are made by the users, also the graph of fraudular transaction made in different payment categories.

2. Literature Survey

Due to the importance of the fraud detection problem, we may distinguish several works that discuss this subject. Thomas et al. (2004) [3] propose a very simple decision tree that is used to identify general fraud classes. They also propose a first step towards fraud taxonomy.Vasiu and Vasiu (2004) [4] propose a taxonomy for computer fraud and, to build it, employ a five-phase methodology.

According to the authors, the taxonomy presented was prepared from a fraud preventing perspective and may be used in various ways. For them, this methodology can be useful as a tool for awareness and education, and can also help those responsible for combating frauds associated with IT to design and implement policies to reduce risks. Chaue al (2006) [5] propose a methodology called 2-Level Fraud Spotting (2LFS) to model the techniques that fraudsters often use to carry out fraudulent activities and to detect offenders preventively. This methodology is used to characterize the auction users on-line as honest, dishonest, and accomplices. Methodologies that characterize fraud are essential for the first phase of the process, since they are the starting point to create a model of the problem and define the best technique for its solution.There are several researches that develop methods to detect fraud [6], [2], [7] and we can realize that these methodologies can differ significantly due to the peculiarities of each fraud type. However, what can be noticed is that the data mining techniques have been widely used in fraud detection regardless of the methodology adopted. This is because these techniques allow the useful information ex- traction in databases with large volumes of data. Phua et al. [8] conducted an exploratory study of numerous articles related to fraud detection using data mining and explained these methods and techniques. These algorithms are based on some approaches such as supervised strategy with labeled data, unsupervised strategy with unlabeled data and hybrid approachIn the hybrid approach (supervised and unsupervised) there are researches using data labeled with supervised and unsupervised algorithms to detect fraud in insurance and telecommunications. Unsupervised approaches have been used to segment data into groups to be used in supervised approaches. Williams and Huang [9] apply a three step process: k-means for detecting groups, C4.5 for decision making, and statistical summaries and visualization tools to evaluate the rule. It is important to note that the choice of which approach to be used depends on the methodology and the available database.These related works have helped us, indicating promising strategies for detecting and preventing fraud. As the datasets are different, mainly due to the very unbalanced data of our scenario, it is not possible to directly compare the results, but they provide an idea of the efficiency of these approaches.The 12th annual online fraud report byCyber Source [10] shows that, for most of the current decade,merchant online fraud losses continued to increase, reaching a peak of $4 billion in 2008.Bayesian Networks (BN) are directed acyclic graphs that represent dependencies between the variables of a probabilistic model, where each node in the graph represents a random variable and the arcs represents the relationships between these variables [11] Bhatla et al [12] said that the rate at which Internet credit card fraud occurs is 12 to 15 times higher than face-to-face transactions.According to Siddhartha Bhattacharyya et al. [13] with the growth in credit card transactions, as a share of the payment system, there has also been an increase in credit card fraud, and 70% of U.S.consumers are noted to be significantly concerned about identity fraud.Due to the importance of the fraud detection problem, we may distinguish several works that discuss this subject[13],[14], [15]. Netmap [17] describes how the clustering algorithm is used to form well-connected data groups and how it led to the capture of the real insurance fraudsters.

3. System Analysis

3.1 Market Analysis

Fraud Prevention and Fraud Detection are the two classes under which these processes are generally defined. Fraud Prevention is the process of implementing measures to stop fraud from occurring in the first place. Prevention is considered the first line of defense, where most fraud is halted at the very beginning. There are different types of Fraud Prevention techniques which can be associated with e-commerce applications, such as Internet security systems for credit card transactions, passwords and tokens to name but a few. However, in practice Fraud Prevention techniques are not perfect and sometimes a compromise must be reached between expense and inconvenience (e.g. to a customer) on one hand, and effectiveness on the other . Nonetheless, Fraud Prevention can sometimes fail due to vulnerabilities in the system and here is where Fraud Detection is needed. Fraud Detection is the process of identifying fraud as quickly as possible once it has been perpetrated, with minimal damage conceivable. The processes falling under this class are said to be the second line of defense. When preventive methods fail, Fraud Detection kicks-in. Fraud Detection is an evolving discipline because of the fact that once a detection method becomes known, criminal minds will adapt their strategies and try others methods to circumvent it. In addition, new criminals enter the field, with different capabilities and different mind-sets. If a detection method becomes known by attackers, it does not mean that it is no longer needed. On the contrary, some inexperienced attackers might not be aware of any detection methods that were successful.

3.2 Fraud Analysis

In the study of market analysis it is given that the current market fraud happen on the transaction are increasing day by day, so to reduce the fraud in the market this system is design. In this system we are using the tracker which tracks every action of the end user, is the tracker

found that suspicious or fraud in the transaction it will update our tracker file. As there are many users attempting transaction so the tracker will track every end user action so that system will be able to find fraud happens in a transaction.

Three types of Transactions:

Payment Gateway: On e-commerce application in payment gateway the user is supposed to write the correct details. He should enter proper bank and card details so that the transaction which he wants to make will be successfully done, otherwise the tracker will track his action and store it as a fraud action in our system.

Net Banking: On e-commerce application if user selects net banking, then he should know his proper transaction id of bank and password. He should enter a proper bank detail and his id and password if he fails then system will track his action and show it as a fraud.

Cash on Delivery: In ecommerce application if user selects cash on delivery option, then the system will generate an otp and that otp will be send to the registered mail id of the user. If the user enter the correct otp then his order will be placed successfully otherwise he will be out of the service and tracker will track the action as a fraud action.

4. System Design

Systems design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering.

4.1 Architectural Design

The architectural design of a system emphasizes the design of the system architecture that describes the structure, behavior and more views of that system and analysis.

- Logical design

The logical design of a system pertains to an abstract representation of the data flows, inputs and outputs of the system. This is often conducted via modeling, using an over-abstract (and sometimes graphical) model of the actual system. In the context of systems, designs are included. Logical design includes entity-relationship diagrams (ER diagrams).

- Physical design

The physical design relates to the actual input and output processes of the system. This is explained in terms of how data is input into a system, how it is verified, how it is processed, and how it is displayed. In physical design, the following requirements about the system are decided.

1. Input requirement,
2. Output requirements,
3. Storage requirements,
4. Processing requirements,
5. System control and backup or recovery.

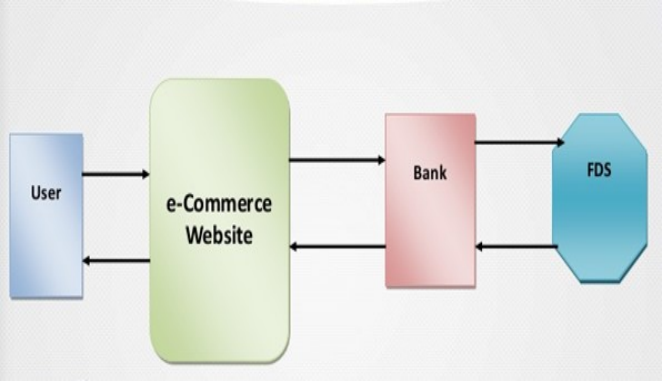

Figure4.1: Architecture of Fraud Detection

The architecture of fraud detection consists of four main blocks:

User: In this block the end user is going to visit the e-commerce application by logging in to system by entering correct login id and password. He can also register himself there.

E-commerce website: Here user can browse different category of products, add them to cart and continue shopping.

Bank: As user want to place the order then he will select the payment option in his respective bank.

FDS (Fraud Detection System): Here as user entered his bank details, if it is correct then he will be able to place the order successfully, otherwise the tracker will be activated in FDS. It will track his activity and give user his transaction failure report.

4.2 Input and Output Design

Input Design

In an information system, input is the raw data that is processed to produce output. During the input design, the developers must consider the input devices such as PC, MICR, OMR, etc.

Therefore, the quality of system input determines the quality of system output. Well-designed input forms and screens have following properties –

- It should serve specific purpose effectively such as storing, recording, and retrieving the information.
- It ensures proper completion with accuracy.
- It should be easy to fill and straightforward.

- It should focus on user's attention, consistency, and simplicity.
- All these objectives are obtained using the knowledge of basic design principles regarding –

Objectives for Input Design

- To design data entry and input procedures
- To reduce input volume
- To design source documents for data capture or devise other data capture methods
- To design input data records, data entry screens, user interface screens, etc.
- To use validation checks and develop effective input controls.

Output Design

The design of output is the most important task of any system. During output design, developers identify the type of outputs needed, and consider the necessary output controls and prototype report layouts.

Objectives of Output Design

- To develop output design that serves the intended purpose and eliminates the production of unwanted output.
- To develop the output design that meets the end users requirements.
- To deliver the appropriate quantity of output.

To form the output in appropriate format and direct it to the right person

4.3 Data Flow Analysis

Data-flow analysis is a technique for gathering information about the possible set of values calculated at various points in a computer program. A program's control flow graph (CFG) is used to determine those parts of a program to which a particular value assigned to a variable might propagate. The information gathered is often used by compilers when optimizing a program. A canonical example of a data-flow analysis is reaching definitions.

Application flowchart:

The application flow chart in figure 5.2 shows the program flow. The user have to register first. Then he can go for the brows products, add that search products to the cart and proceed towards place order by selecting payment option. As the option is correct and user entered correct details, then the transaction is successful and the order is placed successfully. If entered detail is wrong or incorrect the transaction is failed by the system and again backtracks to the place order page and the tracker will start tracking to the activity of the user. For all tracking purpose we use the Hive database and Mysql for storing the products and other details.

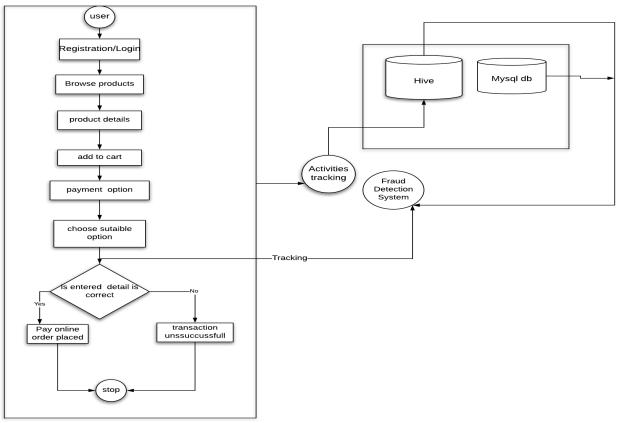

Figure4.2:FlowChart of Fraud Detection

- **External Entity**

An external entity sends or receives data from the system. It can represent a person, a machine, an organization etc, that is external to the system being modeled. Flows outgoing from external entities go to processes.External entities are based on organization units (Organization Units (BPM)) with an External Entity stereotype.

- **Process**

A process is an activity, which transforms and manipulates input data to produce output data. For example, in a model about the publication of books, selecting a manuscript is a process. Data is sent to the selection process in the form of a manuscript. During selection, the manuscript is transformed either into a manuscript that goes directly to the printer, or into a manuscript that must wait before it is printed. Processes are based on standard BPM processes. Flows from processes can go to external entities, data stores, split/merges, or other processes.

- **Data Flow**

Data flow shows the transfer of information from one part of the system to another. The symbol of the flow is the arrow. The flow should have a name that determines what information is being moved. Exceptions are flows where it is clear what information is transferred through the entities that are linked to these flows. Material shifts are modeled in systems that are not merely informative. Flow should only transmit one type of information (material).

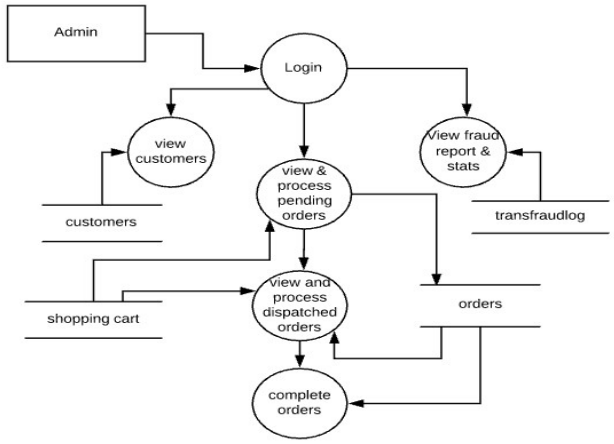

Figure4.3:DFD of User Admin

5. Advantages

- From end user side:
  - Easy to find product
  - Secure transaction is made
- The system stores previous transaction patterns for each user.
- Based upon the user spending ability and even country, it calculates user's characteristics.
- The system is more secured with OTP (One Time Password) implementation.
- IP address tracking at every transaction.
- Security questions for payment limit crossed.
- More than 20-30 % deviation of user's transaction (spending history and operating country) is considered as an invalid attempt and system takes action.

6. Conclusion

Big Data for Fraud Detection on Ecommerce Application focuses on the impact of Big Data Analysis on the current marketing operations and how BDA changes customer's online behavior. In this project using big data we build different fraud detection models to predict fraud in online transactions, more specifically credit card operations, net banking and also using cash on delivery. Credit card companies shall be able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. It includes-deeply understand the changes among customers' needs, respond to customers and supply chain quickly, get feedbacks of products from customers easily, develop the comprehensive understanding of the products and services, improve the strategy to fit the market speedily.

7. Future Scope

We design a system to detect fraud. This system is capable of providing the features required to detect fraudulent activity and transaction. As technology changes, it becomes difficult to track the behavior and pattern of fraudulent transactions. We have just detected the fraudulent activity but we have not prevented. Preventing known and unknown fraud in real time is not easy but it is feasible. The proposed architecture is basically designed to detect credit card fraud in online payments, and emphasis is made to provide a fraud prevention system to verify a transaction as fraudulent or legitimate. For implementation purposes it is assumed that issuer and acquirer bank is connected to each other. If this system is to be implemented in real time scenario then exchange of best practices and raising consumer awareness among people can be very helpful in reducing the losses caused by fraudulent transactions. Further enhancement can be done by making this system secure with the use of certificates for both merchant and customer and as technology changes new check scan be added to understand the pattern of fraudulent transactions and to alert the respective card holders and bankers when fraud activity is identified. The dataset are available on day to day processing may become outdated, it is necessary to have updated data for effective fraud behavior identification. To this extent, the incremental approach is necessary in making the system to learn from past as well as present data and capable of handling the both. Fraudster uses different new techniques that are instantaneously growing along with new technology makes it difficult for detection.

REFERENCES

1. Bolton, Richard J., and David J. Hand. "Statistical fraud detection: A review." Statistical science (2002): 235-249.
2. Maranzato, Rafael, Adriano Pereira, Alair Pereira do Lago, and Marden Neubert. "Fraud detection in reputation systems in e-markets using logistic regression." In Proceedings of the 2010 ACM symposium on applied computing, pp. 1454-1455. ACM, 2010.
3. Thomas, B., J. Clergue, A. Schaad, and M. Dacier. "A comparison of conventional and online fraud." In CRIS, vol. 4, pp. 25-27. 2004.
4. Vasiu, Lucian, and Ioana Vasiu. "Dissecting computer fraud: from definitional issues to taxonomy." In 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the, pp. 8-pp. IEEE, 2004.
5. Chau, Duen Horng, Shashank Pandit, and Christos Faloutsos. "Detecting fraudulent personalities in networks of online auctioneers." In European Conference on Principles of Data Mining and Knowledge

Discovery, pp. 103-114. Springer, Berlin, Heidelberg, 2006.
6. Fawcett, Tom, and Foster Provost. "Adaptive fraud detection." Data mining and knowledge discovery1, no. 3 (1997): 291-316.
7. Barse, Emilie Lundin, Hakan Kvarnstrom, and Erland Jonsson. "Synthesizing test data for fraud detection systems." In 19th Annual Computer Security Applications Conference, 2003. Proceedings. pp. 384-394. IEEE, 2003.
8. Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler. "A comprehensive survey of data mining-based fraud detection research." arXiv preprint arXiv:1009.6119 (2010).
9. Williams, Graham J., and Zhexue Huang. "Mining the knowledge mine." In Australian Joint Conference on Artificial Intelligence, pp. 340-348. Springer, Berlin, Heidelberg, 1997.
10. Caldeira, Evandro, Gabriel Brandao, and Adriano CM Pereira. "Fraud analysis and prevention in e-commerce transactions." In 2014 9th Latin American Web Congress, pp. 42-49. IEEE, 2014.
11. Caldeira, Evandro, Gabriel Brandao, Hudson Campos, and Adriano Pereira. "Characterizing and evaluating fraud in electronic transactions." In 2012 Eighth Latin American Web Congress, pp. 115-122. IEEE, 2012.
12. Maes, Sam, Karl Tuyls, Bram Vanschoenwinkel, and Bernard Manderick. "Credit card fraud detection using Bayesian and neural networks." In Proceedings of the 1st international naiso congress on neuro fuzzy technologies, pp. 261-270. 2002.
13. V. P. Tej Paul Bhatla and A. Dua, Understanding Credit Card Frauds, 2003.
14. S. Bhattacharyya, S. Jha, K. Tharakunnel, Westland, and J. Christopher, “Data mining for credit card fraud: A com-parative study,” Decis. Support Syst., vol. 50, pp. 602–613, February 2011.
15. P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, “Detection of financial statement fraud and feature selection using data mining techniques,” Decision Support Systems, vol. 50, no. 2, pp. 491–500, 2011.
16. E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” Decision Support Systems, vol. 50, no. 3, pp. 559–569, 2011.
17. Netmap, “Fraud and crime example brochure,” 2004.