ANOMALY DETECTION IN E-COMMERCE

JOSEPH, EKEMINI CHARLES (B.Sc.)

BAS/CSC/160156

A PROJECT WORK PRESENTED TO THE DEPARTMENT OF PHYSICAL SCIENCES, FACULTY OF SCIENCE, IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR DEGREE OF SCIENCE (B.Sc.) IN COMPUTER SCIENCE OF BENSON IDAHOSA UNIVERSITY, BENIN CITY, EDO STATE, NIGERIA

JULY, 2021

DECLARATION

I, JOSEPH, EKEMINI CHARLES, hereby declare that this project work is entirely my work and has not been submitted for any degree and it is concurrently being submitted. Works of other authors have been duly acknowledged.

_____ _____

JOSEPH, EKEMINI CHARLES DATE

CERTIFICATION

This is to certify that the work in this project work titled 'ANOMALY DETECTION IN E-COMMERCE' has been performed by me under supervision of Mr. Prince Collins Igbinigie.
No part of this project was presented for another degree elsewhere at any institution to the best of my knowledge. All information utilized and their sources have been acknowledged by means of reference.

_____ _____

JOSEPH, EKEMINI CHARLES **DATE**
(Project Student)

_____ _____

MR. PRINCE COLLINS IGBINIGIE **DATE**
(Project Supervisor)

_____ _____

DR. (MRS) OJUH DIVINE OSAMIROMWEN **DATE**
(Head of Department)

DEDICATION

I dedicate this project to God Almighty for the strength, knowledge and wisdom to put this project work together. I also dedicate this work to my family who have encouraged me all the way till this point to finish this project. Thank you and may God bless you all.

ACKNOWLEDGEMENTS

I want to express by gratitude to God and those who gave me possibility to complete my project work.

A special appreciation to my supervisor Mr Prince Collins Igbinigie who has made tremendous contributions and advice for my work

I am grateful to the Head of Department of Physical Sciences, Dr. (Mrs.) Divine Osamiromwen Ojuh, and my highly esteemed and distinguished lecturers Mrs. G.O. Iyawe, Mr. A.E. Odion, Dr. K.O. Obahiabgon, Mrs. A. Inyang, Mrs. J.U. Obadoni, Dr. M.S.U. Osagie, Mr. S. Obadan, Mr. I.B. Eraikhuemen, Engr. O.I. Akhidemo, Dr. O. Eguasa, Mr. B.E. Ibude, and Mrs. O. Faith.

And to my family and friends who have been my greatest support, Thank you.

ABSTRACT

E-commerce commonly described as an electronic or online method of carrying out transaction has been adopted by most organization, enterprises, businesses, and individuals. And this has greatly influenced economic activity positively, however using this medium for transactions comes with great risk, as lots of online fraud is being carried out daily which includes credit card fraud, phishing, and business compromise with huge financial losses running into billions of dollars. Efforts in curbing fraud related cases in e-commerce is still ongoing and very massive, however researches in detecting anomalous transactions in e-commerce has not received a great attention, as fraud detection research on e-commerce is only limited to the determination of features or attributes which will be used to determine the nature of fraud or non-fraud transactions in e-commerce. The need to stem the tide of the rising prevalent cases of ecommerce related frauds led to the conception of this project. This project therefore adopts a machine learning technique to detect abnormal transaction in an e-commerce platform. A logic regression model was developed using data's of customer spending at different times. The data's where obtained from the internet and was used to build the model. The Dataset was trained and tested for anomalous transactions. The project was developed using Python programming language, and was developed in a Jupyter environment.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background of study

Electronic Commerce (E-Commerce) is growing rapidly in worldwide marketplace nowadays and it has tremendous potential to drive the technology growth in the future. Along with the continuous development of the E-commerce, it can be connected, anywhere, anytime via e-mobile with wireless service. In spite of the challenging economy, the use of e-channel platforms –Internet banking, Mobile Banking, ATM, POS, Web, etc. has continued to experience significant growth. According to NIBSS 2015 annual fraud report, transaction volume and value grew by 43.36% and 11.57% respectively, compared to 2014. Although e-fraud rate in terms of value reduced by 63% in 2015, due, in part, to the introduction of BVN and improved collaboration among banks via the fraud desks; the total fraud volume increased significantly by 683% in 2015 compared to 2014. Similarly, data released recently by NITDA (Nigeria Information Technology Development Agency) indicated that Nigeria experienced a total of 3,500 cyber-attacks with 70% success rate, and a loss of $450 million within the last one year.

Also, the activities and transactions made from e-commerce platforms have generated huge amount of data. There exists a lot of anomalous transaction hidden in them. How to monitor and inspect these transactions become the key concerns and challenges. E-commerce applications are prime targets for criminal attacks. (Bolton, *et al*., 2002) explain that new types of fraud have emerged such as mobile telecommunications fraud and computer intrusion whilst traditional fraud, for instance money laundering, has become easier. Finding the best possible way against fraud is crucial. Different processes have to be implemented in order to protect clients from attackers perpetrating fraud. Fraud Prevention and Fraud Detection are the two (2) classes under which these processes are generally defined.

Anomaly Detection System (ADS) monitors the behavior of a system and flag significant deviations from the normal activity as an anomaly. Anomaly detection in e-commerce is used for identifying attacks on online businesses, financial details, business compromise, and other fraud related activities on e-commerce. Typically, ecommerce transaction anomaly action can be divided into two categories. The first category is that the system cannot be satisfied with the

process of normal trading, format and other rules, such as duplicate transactions and tampered transaction. The second category is that the system satisfies the requirements of normal trading, but the transactions also have a certain fraudulent characteristics. For example, attacker steals the user transaction information for a trading, or legitimate users make an act of malicious overdraft. In general, according to a specific implementation mechanism of payment system, which has the ability to detect and prevent the first category of transaction anomaly action. However, for the second category, the transactions which process and provide authentication information generally comply with the payment system. Thus, the payment systems are difficult to detect this type of transactions anomaly action. Obviously, the second type of execution of transaction will undoubtedly pose an enormous risk for the financial institutions and user. This thesis also focuses on this type of transaction.

This project is dedicated towards resolving these challenges. The proposed machine learning based transaction anomaly detection method can detect the anomaly action at the business and operation level. This thesis can be a roadmap for other data mining practitioners and web administrators on how to develop more effective of anomaly detection models to e-commerce transaction for protecting the online trading safety. Due to the fact that the existing algorithms are not capable of properly describing the behaviors profile of the user, causing higher false positive rate of data-base anomalous detecting. Even though part of the existing approaches are capable of achieving

1.2 Statement of problem

Anomaly detection refers to detecting patterns in a given data set that do not conform to an established normal behavior. The patterns thus detected are called anomalies and translate to critical and actionable information in several application domains. E-commerce applications are prime targets for criminal attacks. (Bolton, *et al*., 2002) explain that new types and forms of fraudulent activities are emerging at intervals. These types of attacks and fraud are increasing during the process of E-commerce transaction. Furthermore, due to the increase in credit card usage and convenient mode of online money transaction, fraudsters are also finding more opportunities to commit fraud, which affects banks and card holders to great financial losses (Sherly, k., *et al*., 2010). As the online money transaction is increasingly popular, more and more sensitive user information frequently transmit and store on the public Internet, the confidentiality

of user account information has to face greater threats. With the rapid development of the e-commerce and e-payment, the problem of online transaction fraud has become increasingly prominent, compared with traditional areas online transaction is larger volume of data's daily and fund transfer. This makes it rather impossible to detect any form of abnormalities in any transaction in ecommerce. These problems above however formed the problem statement of this project.

1.3 Aims and objectives

The aims and objectives of this project is geared towards detecting anomaly transactions in ecommerce using machine learning technique. Other objectives of this study is to;

i. Implement a machine learning algorithms that can be used to detect anomalies in an ecommerce transaction.
ii. Develop a model using logic regression algorithm that detects anomalous transactions on an e-commerce platform using data set of a user'sspending at different times.
iii. Clean, train and test dataset containing transactions of users from an e-commerce platform.
iv. To detect attacks on ecommerce in a fast and effective way by studying transaction anomaly with machine learning methods.
v. Review existing and related researches in the areas of detecting abnormalities in ecommerce transitions and other related frauds.

1.4 Significance of study

Anomaly Detection System (Lee, W. K., *et al*., 1999) is a proactive security protection technology. It is an important part of the information security architecture. The development of an Anomaly Detection model can be well used to resolve the prevalent problems of frauds and related attacks on ecommerce applications. This project is of great significance and importance to ecommerce platform owners, as it will eliminate or reduce cases of financial losses due to

attacks, credit card and other financial information theft, and business compromise perpetuated by fraudsters. Also this project will be of great benefit to ecommerce users or customers whose credit card details are most times stolen, hence losing their monies to online criminals, as it will be able to curb incessant cases of financial information theft during customer transactions. Lastly this project will be beneficial to researchers and stakeholders in the area of online fraud and attack detection as it will provide necessary tools needed in fighting the growing scourge of attacks on ecommerce platforms

1.5 Scope of study

The scope of this study is limited to developing a model using regression algorithm to detect anomalous transactions on ecommerce applications, using dataset of aperson's spending or transactions on ecommerce channels at various time.

1.6 Definition of terms

Anomaly Detection: Anomaly detection (also outlier detection) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.

E-Commerce: Popularlyreferred to as electronic commerce (sometimes written as eCommerce) is a business model that lets firms and individuals buy and sell things over the internet

Logic Regression: Logistic regression is a kind of statistical analysis that is used to predict the outcome of a dependent variable based on prior observations.

Customer behavior: is the study of how individual customers, groups or organizations select, buy, use, and dispose ideas, goods, and services to satisfy their needs and wants.

Credit card: A credit card is a thin rectangular slab of plastic or metal issued by a financial company, that lets cardholders borrow funds with which to pay for goods and services.

Machine learning: Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed

Fraud Detection: Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses.

CHAPTER TWO

LITERATURE REVIEW

2.1 Conceptualization of anomaly detection in ecommerce

As the e-commerce sales grow in global retail sector year by year, detecting anomalies that occur in the most important key performance indicators (KPI) in real-time has become a critical requirement for e-commerce companies. Such anomalies that may arise from software updates, server failures, or incorrect price entries cause substantial revenue loss in the meantime until they are detected with their root-causes. E-commerce applications are prime targets for criminal attacks. (Bolton, *et al*., 2002) explain that new types of fraud have emerged such as mobile telecommunications fraud and computer intrusion whilst traditional fraud, for instance money laundering, has become easier. Finding the best possible way against fraud is crucial.

Different processes have to be implemented in order to protect clients from attackers perpetrating fraud. Fraud Prevention and Fraud Detection are the two (2) 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 (Bolton, *et al*., 2002). 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 (Bolton, *et al*., 2002). Nonetheless, Fraud Prevention can sometimes fail due to vulnerabilities in the system and here is where Fraud Detection is needed.

Anomaly Detection is the process of identifying abnormalities from a given set of records of data's 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, thus, giving us the edge to detect them.

Although continuous security improvements are installed in auction sites or e-commerce web applications, in practice research has shown that almost every online system has some type of vulnerability which can be exploited to commit fraud (Jaquith, 2002). Even though online identity verification methods are nowadays more sophisticated, illegal activity is still not totally prevented and criminal activity is still successful. Fraud Detection systems are often derived from statistical tools which are based on the comparison of observed data with expected values. In contrast, an Intrusion Detection System (IDS) is often used to detect Computer Intrusion. The observed data in these statistical tools is generally based on behavior profiles, including past behavior. The problem, with statistical tools is the way expected values are derived, since it is determined by the environment context for which the statistical tools are built for. There cannot be a general, all-encompassing-statistical tool; a one-for-all tool.

Statistical Fraud Detection methods are categorized into Supervised and Unsupervised methods. Supervised methods involve techniques which observe training data and construct models based on what has been analyzed. In most cases, the observed data includes both fraudulent and legitimate cases. However, adaptability is of a concern with these statistical methods, since they can only be used to detect fraud which has previously occurred. Unsupervised methods, on the contrary, do not require any training data, but try to find dissimilarities in unlabeled data. Sometimes there are cases when training data is not available, or is very hard to get, thus giving rise to unsupervised methods. One of the difficulties with this type of statistical method is accuracy because they commonly create high volumes of false-positives and false-negatives. Any fraudulent case detected, involves a considerable amount of analysis and resources to identify the real cause and security implications. False-positives are of particular concern since a lot of time and resources are wasted to analyses cases which in reality were genuine.

Intrusion Detection Systems (IDS) are used to detect any kind of attack launched against entire computer networks. IDSs can also be used to detect web-based attacks by configuring them with a number of signatures that support the data of known attacks (Kruegel, *et al*., 2005). The problem with IDS is that it is very hard to keep signature sets updated with the latest known vulnerabilities. Furthermore, new vulnerabilities must first be discovered before signatures can

be applied to protect against them, at which stage it might be too late. In addition, when custom e-commerce applications are developed in-house, new vulnerabilities might be introduced especially when business updates are installed. In practice, it is a very time-intensive and error-prone activity to develop ad-hoc signatures to detect attacks against these applications, apart from the fact that substantial security expertise is also required.

2.2 Anomaly Detection

Anomaly detection has focused on the investigation of undesirable behavior changes. These negative changes are, at times, interchangeable with the term anomaly in machine learning research. Path-based Failure Detection is an example of a method previously used to identify negative changes (Chen, M. Y., *et al*., 2014) that was later embedded in other advanced approaches. The macro approach has focused on application component interactions rather than simple monitoring or code-level debugging. The interaction approach is significant because tracing interaction behavior could help identify specific application patterns. The macro approach has increasingly extensive applications and has been renamed Pinpoint, after its anomaly determinant method was changed from statistical analysis to data mining (Chen, M. Y., *et al*., 2002).

Though past studies have tried to detect anomalies like application component failures, they have not concentrated exclusively on anomaly data analysis. Instead, they have tended to focus on identifying failure symptoms themselves. Thus, these approaches have not been well suited to the fluid modern enterprise environment. A model that contains specific features such as decreased effort, increased quickness, and high intelligence would be more well-suited to contemporary circumstances. The anomaly state can be defined as the existence of an undesired object or an unexpected behavior. Anomalies are patterns occurring beyond ordinary expectations. In the e-Commerce context, anomalies are undesirable behaviors or threats associated with criminals (Das, K., 2009). In statistics, they are often referred to as outliers. Visualization helps to identify anomalies because they are positioned far from the means. Figure 1.1 presents typical e-Commerce services anomalies.

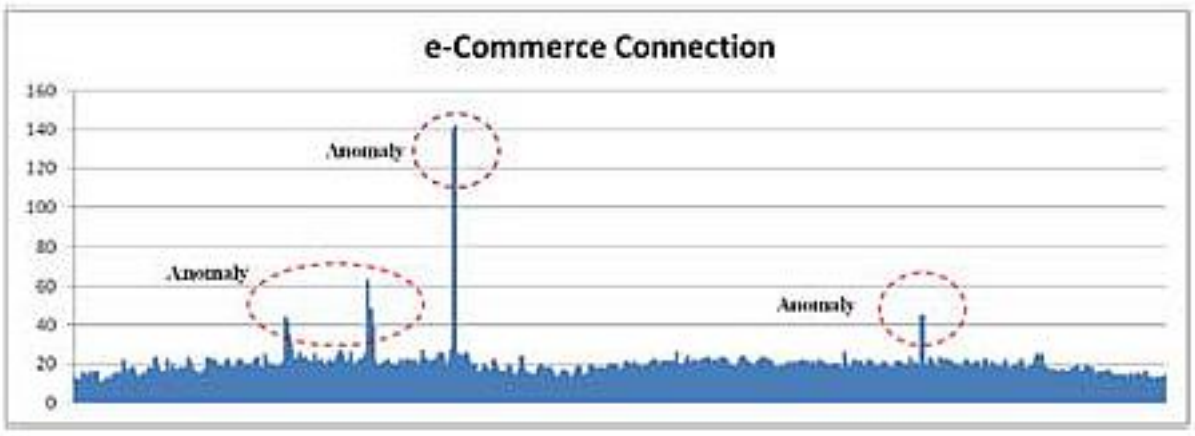

Fig 1.1 A Typical E-Commerce Anomaly Pattern (Ibidunmoye, O.,*et al*., 2015)

Anomalies can be divided into 4 types: Point, Collective, Contextual (Chandola, V.,*et al*.,2009) and Pattern Anomalies.

Point anomaly

Point anomalies are data points that deviate from mean groups or solid dots outside of normal groups (Ibidunmoye, O.,*et al*., 2015). Figure 1.2 shows a typical point anomaly. They are commonly used to provide insight regarding application latency or system resource utilization.

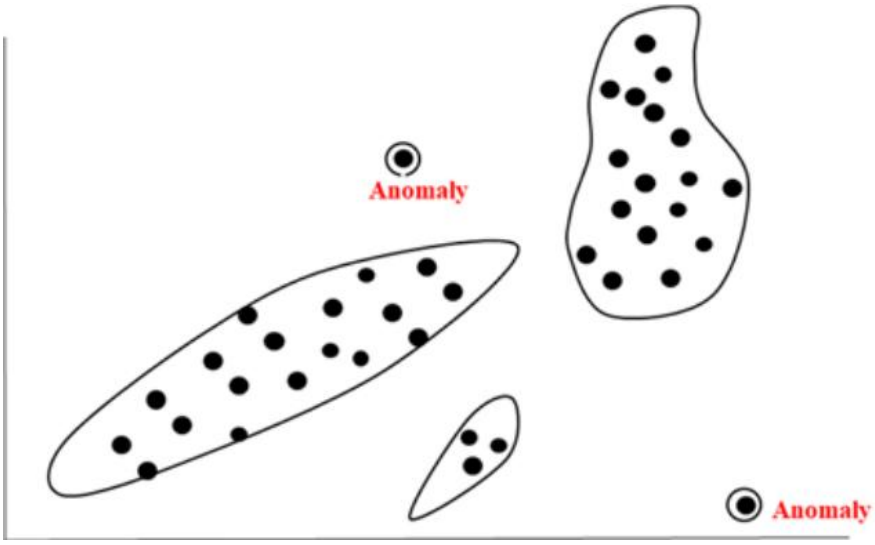

Fig 1.2 Showing Point Anomalies (Ibidunmoye, O.,*et al*., 2015)

Collective anomalies: Collective anomalies are homogeneous data points or groups of anomalies representing sudden changes of throughput, as shown in Figure 1.3

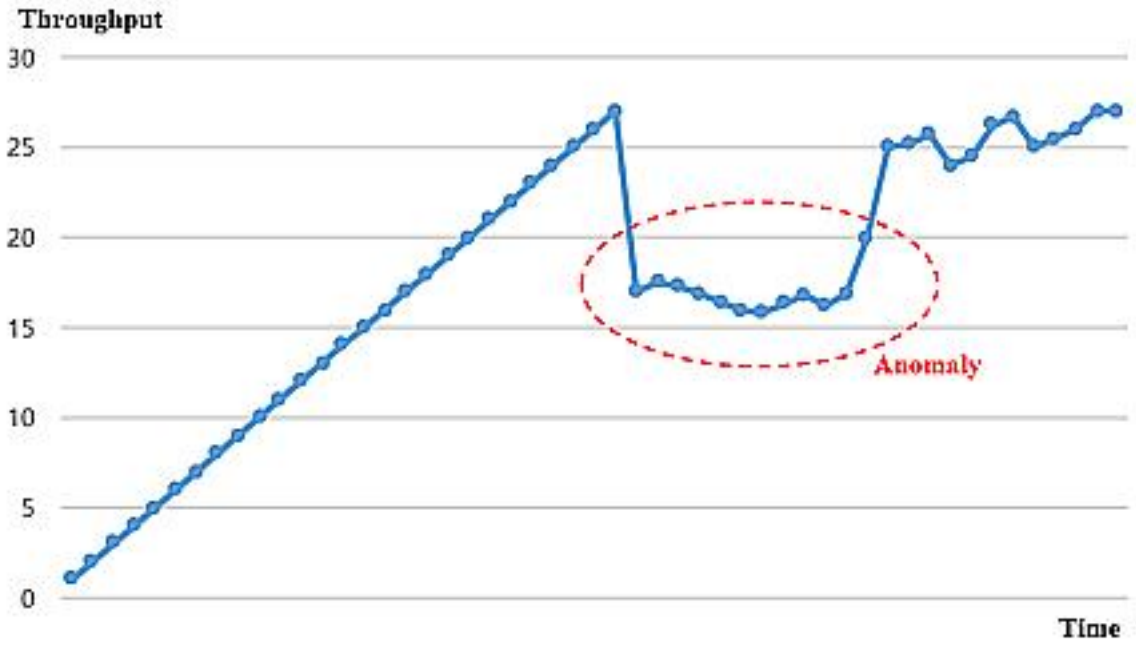

Fig 1.3 Collective anomalies (Kim, S., 2015).

Contextual anomalies

Contextual anomalies are anomalies detected under specific circumstances. Figure 1.3 displays the contextual collective anomaly of throughput when time is a contextual attribute. These sudden drops in the throughput within certain timeframes are identified as contextual collective anomalies Kim, S., (2015).

Pattern anomalies

A performance trend can identify anomalous behaviors, which recur in similar patterns. These days, the role of anomaly detection has expanded and it has acquired research applications beyond already known domains. Technically, it is used to detect network intrusions and machine defects or errors while, socially, it contributes to the investigation of criminal activity including fraud detection in Anti-Money Laundering, Insurance Claims, and so on. The healthcare and

medical domains are additional representative research areas where anomaly detection techniques play important roles (Patcha, A., *et al*., 2007)

2.3 Review of related works

The implementation of information technology in the field of AML began in the 1990s. Nowadays, transactions can be easily done online. Modern technology has allowed money laundering to become an online crime, so the money laundering detection technology also needs to be established. In April 1990 (FATF Report, 1990-1991), the Financial Action Task Force (FATF) issued a report which estimated the amount of money laundered globally from 1990 to 1991 was around $800 billion to $2 trillion. The proceeds in drug industry had reached more than $300 billion, and most of which was laundered in the US financial market. To solve the money laundering problem, an AML computer automatic monitor system is urgently needed to detect anomaly money transfers and money laundering. For example, the American Financial crime enforcement network (FINCEN) Artificial Intelligence System which was developed by Senator, T. E., et al., (Senator, T.E., *et al*., 1995). This system used Bayesian models to determine the level of suspicious transactions, and then further analyzed the high level suspicious transaction data based on the previous results. It commendably integrated a variety of artificial intelligence technologies and software agents to identify the potential money laundering problem on the transaction reports. The tested results clearly showed that Artificial intelligence computer analysis system can greatly enhance the work efficiency and is an essential method for AML.

Two years later, (Stofella, P., 1997) developed a DBInspector model. This model was employed in the anti-money laundering activities performed by the supervision department of the Italian central bank. It focused on the employment of high performance database and 3D data visualization technologies for the construction of a data mining environment. The DBInspector software environment was an integrated and open set of tools for the analysis and inspection of large databases. The users were allowed to interact with different data to process and visualize data flows in this environment. The DBInspector system was implemented in the high performance computing and networking environments, such as government, financial, and industrial organizations which maintained and managed large databases. Specifically, database servers based on parallel technology had allowed the possibility of real time inspection and

analysis of relational data stores in a large scale, which has not been possible in the past because of weak performance of available technologies.

In 2001, a decision support system based on data mining techniques in e-banking was proposed by (Ionita, I.,*et al*., 2001). They used data mining technologies in banking domain because it was suitable due to the nature and sensitivity of bank data and real time complex decision process. The main concern in this design was to make good decisions in order to minimize the risk level and anomaly transactions associated to the bank. Next year, (Syeda,*et al*., 2002) used parallel granular neural network (PGNNs) to improve the speed of data mining and knowledge discovery process for credit card fraud and anomaly detection. Based on the implemented system, PGNNs algorithm could reasonably improve the 10-fold processing speed. However, this method had a limitation in the system that it needed more processors to solve the load imbalance problem, otherwise a high anomaly detection error would occur.

The Wolfsberg Group was formed by a group of international banks to share ideas on how to fight global money laundering using artificial intelligence (AI) System. In 2002, this group pointed out that International money laundering had been about the Mafia, drug smuggling, and arms deals involving sums in excess of $500 billion per year. Unlike many types of financial fraud, money laundering could range from a single transaction to the culmination of months of complex transactional activities. Additionally, the bank would not share past cases because of the sensitivity of information. To eliminate these increasingly complex transactional activities, (Kingdon, J., *et al*., 2002) designed a bank transaction data monitoring and analysis system which could automatically detect payment fraud and money laundering in the financial sector through the system. Two years later, (Kingdon, J., 2004) designed another artificial intelligence system to automatically identify a set of customer behavior patterns. This system can efficiently identify customers' abnormal trading behaviors and make decisions dynamically, adaptively, and timely. People can use a machine with intelligence to judge context. This identification process could operate on the scale, and resolve problems with transparency and justification. The author believed this new generation of adaptive operational analytics could provide new possibilities from risk management to service provision.

In 2009,(Kundu, A.,*et al*., 2009) considered that E-commerce transaction anomalies are interspersed with genuine transactions. The simple pattern matching method which people

frequently use cannot guarantee the accuracy of the detection results. Therefore, a combined anomaly detection model should be developed as misuse detection techniques. According to this paper, they used a BLAST-SSAHA Hybridization model for credit card transaction anomaly detection which was anheuristics that improved the performance of the sequence alignment algorithm. It combined two sequence alignment algorithms: the Basic Local Alignment Search Tool (BLAST) which was the most popular heuristic approach, and the Sequence Search and Alignment by Hashing Algorithm (SSAHA) which was one of the fastest algorithms for sequence alignment. The BLAST-SSAHA Hybridization model included a two-stage sequence alignment. The first stage was profile analyzer (PA) which used past transaction record sequences to determine the similarity of an incoming sequence of transactions. The second stage was deviation analyzer (DA), during which stage the abnormal transactions were tracked on a deviation analyzer for possible alignment with past anomaly behaviors. The test result showed that the processing speed was fast enough to satisfy online detection of credit card transaction anomaly while maintaining high accuracy. The proposed BLAST-SSAHA hybridization approach can be effectively used to counter fraud in other domains such as telecommunication and banking fraud detection.

In the same year, (Wang, N., *et al*., 2009) built a model based on similar coefficient sum to predict whether a credit card transaction is anomaly or not. This method focused on the outlier detection based on the similar coefficient sum. By computing the similar coefficient sum of every two objects, the anomaly record would be found. During anomaly detection process, two types of mistakes occurred because the anomaly data was far less than the normal data in this data set. The first type of mistake is that the abnormal transactions were mistakenly considered as normal transactions, which called the first class error or False Negative error.

2.4 Data Analysis Techniques

E-commerce transaction anomaly detection can usually be based on the following two assumptions: The first assumption is that there is a clear distinction between existing normal transactions and anomaly transactions. The second assumption is that the abnormal transaction action only occupies a small proportion in all transactions. In this thesis, I will focus on dealing with the transaction records under the second assumption. This chapter will discuss the following three types of E-commerce transaction anomaly detection technology: statistical classification

methods, data mining classification methods, and clustering methods. Statistical classification methods are the collections of methods which take a statistical approach to classify data instances. The traditional statistical classifier is to construct an underlying probabilistic model which provides a probability measure of class membership instead of mere classification (Michie, D.,*et al*., 2004). The traditional statistical classifiers usually use distribution of the data points such as normal distribution and Poisson distribution to model E-commerce transaction anomaly detection, and then test the abnormal inconsistency. The statistical approaches frequently use two modern statistical techniques such as k-Nearest Neighbor (k-NN) and Naive Bayes (NB), and a classical statistical technique like Logistic Regression (LR). However, there are some limitations in these statistical approaches. For example, the data distribution in reality often does not match any of the known distribution. In addition, statistical approaches are not suitable for multidimensional data model and the results are usually unsatisfactory.

Data mining classification methods are gaining significant momentum nowadays in most industry sectors, and widely applied in anomaly detection field. Data mining is the process of finding knowledge from a large amount of data. It encompasses an interdisciplinary collection of methods and techniques from numerous scientific disciplines to achieve the purpose, such as machine learning, mathematics and artificial intelligence (Ngai, E.,*et al*., 2009). The term actually represents a collection of processes including data preparation, data mining and result evaluation, and in the most cases it is referred as the essential process during which the knowledge is extracted from the embedding data (Han, J.,*et al*., 2001). Consequently, the use of computer-based classification algorithms has the advantage on processing power over traditional statistical methods.

CHAPTER THREE

SYSTEM DESIGN AND METHODOLOGY

3.1 Introduction

This chapter introduces the concept of the design architecture of a model for detecting anomaly transactions in an ecommerce application. It presents the methodology adopted for the design and modeling of the system, and also analyze the existing methods of detecting fraudulent transactions in an ecommerce platform. The research methodology includes data collection tools used, machine learning algorithm and techniques adopted for developing the proposed model. This chapter will define approach in achieving the development of the system while emphasizing on the problem statement, and the system details such as systems architectural design, user interfaces, the database logical and physical designs, and data management.

3.2 Research methodology

Research methodology refers to the specific methods or techniques adopted to identify, select, process, and analyze information about a related issue. However, offers an overview about the research design, types of data to be collected, sampling design and relevant interpretation towards the conducting of respective research and necessary statistical tools selected for proposed hypotheses of the research. It allows direct evaluation of the study's overall validity and reliability, and also answers two main questions which are, how was the data collected or generated, and how was it analyzed? The method of data gathering for the proposed model for detecting anomalous transactions was obtained from transactions of users from an ecommerce application, hence formed the dataset.

3.3 Data collection method

This research employs the schedule method of data collection, as it is a common tool used for investigation that requires the researcher to fetch data's spanning for a period of time. is one of the very commonly used tools of data collection in scientific investigation. Young, P.V. says "The schedule has been used for collection of personal preferences, social attitudes, beliefs, opinions, behavior patterns, group practices and habits and much other data". For this project, data's where obtained from the transactions of a POS mobile terminal within different period of

time. 101 rows of data containing transactions description, time, amount, and the corresponding transaction type either it's a legitimation transition or not a legitimatetransactions.

3.4 System analysis

Systems analysis is a problem solving technique that involves the collection of factual data, understand the processes involved, identifying problems and recommending feasible suggestions for improving the system functioning. This involves studying the business processes, gathering operational data, understand the information flow, finding out bottlenecks and evolving solutions for overcoming the weaknesses of the system so as to achieve the organizational goals. System Analysis also includes sub-dividing of complex process involving the entire system, identification of data store and manual processes. Continuous monitoring and review of transactions is one method that is currently used in most business to detect anomalous transactions. Other method of detecting abnormal transactions is the use of third party applications and they are set to a default range of values. Any amount or time exceeding the value taken as a

3.5 Analysis of the existing system

Existing method of monitoring and detecting abnormal transactions is dependent on manual human efforts, and it's achieved by reviewing existing transactions made by a user and also comparing a function of the various fields with fixed criteria known as triggers (referred to as absolute analysis), or a set of transactions that might be tagged legitimate or regular, so as to detect any form of irregularities. Continuous monitoring and review of transactions is one method that is currently used in most businesses to detect anomalous transactions. Other method of detecting abnormal transactions is the use of third party applications and they are set to a default range of values. Any amount or time exceeding the default value are marked and tagged abnormal.

3.6 Analysis of the Proposed System

This proposed system adopts the supervisory medium and machine learning to detect and learn abnormal activity. The system utilizes a database of known fraudulent/legitimate cases from which to construct a model which yields a suspicion score for new cases. The proposed system is

an automated system that identify and detects anomalous transactions using machine learning techniques from dataset containing transactions of users on an Ecommerce application at various time of the day and also the corresponding remark on the transaction to either a genuine transaction or not genuine transaction. Machine learning is carried out in two half and it includes training and testing or execution. The database containing the transactions of users on an ecommerce platform is first trained and then used in execution of program. The system employs a supervised learning technique to train and test the model.

3.7 Input analysis

The inputs where designed so as to meet the system requirement. The inputs to the system are transactions of users containing the transaction date and time, transaction amount, and the remarks relating to those transactions. Each data is cleaned and validated to ensure they are not empty or NA data.

3.8 WEAKNESS OF THE PRESENT SYSTEM

The present system is currently used on portals and online platforms where illegal or fraudulent transactions are carried out daily, and the method has been faulted in many ways which includes.

i. It is usually very slow to act on reported abnormal activity thereby making people lose their money before necessary actions is carried out.
ii. The overall existing system has a very low efficiency and sometimes prone to error and mistakes, thereby exposing users or customers to great risks.
iii. Due to lack of profiling of illegitimate or related cases. Fraudster always comes out with new techniques and methods
iv. The system is not adaptable and does not profile cases of recent and related abnormal or fraudulent activity.
v. High cases of false positive due to the inability to distinct between legitimate and non-legitimate transactions, thereby increasing cases of false triggers

3.9 Justifications for the new system

The proposed system has a number of advantages as compared to the existing system in no small measure. Some of the advantages are enumerated below.

i. Real time detection of abnormal activities or transactions, thereby reducing the time lag in detecting abnormalities.
ii. Higher efficiency and precision in triggering illegal or abnormal transaction detection alarm thereby eliminating false positive alarms.
iii. It has the ability to continuously detect new form of illegal transactions based on previous data set, as profiles of previous normal and abnormal transactions is used to analyze all forms of entry
iv. The system is adaptive and evolves over time, with improved machine learning techniques to learn new form of fraud
v. Eliminates the need and high cost of purchasing third party application to monitor user's transaction entry.

CHAPTER FOUR

IMPLEMENTATION

4.1 Programming

Programming is the act of writing out a set of instructions which are written in a sequence so as to get a particular task done. Programming consists of different language and syntax, which brought together, can achieve a particular function. This chapter presents an overview and description on the choice of platform, tools and programming languages, software and hardware requirements, and the different modules and interfaces that were implemented for the proposed system. Implementation is the realization of an application by following through several execution phases of planning, modeling, design and analysis. The model design and implementation was achieved using Python programming language in a Jupyter Environment using logic regression technique.

4.2 The general steps for developing a Python application

These steps are often series of processes or major activities required to install both the hardware and software of the system. Below are the necessary steps required for developing a model that detects anomalous transactions in an ecommerce platform using python programming language.

i. Analyzing
ii. Coding
iii. Debugging
iv. Testing
v. Installation
vi. Documentation
vii. Training plan

4.2.1 Analyzing or Understanding the Problem

This is usually the first step in building any software, it enables the developer a thorough understanding and identification of the problem for which is the program or software is to be

developed. All the factors like Input/output, processing requirement, memory requirements, error handling, interfacing with other programs have to be taken into consideration in this stage.

4.2.2 Program Design

This is next stage done to achieve the design. Here the software developer makes use of tools like algorithms and flowcharts to develop the design of the program. All the instructions which are to be performing at different stages are listed in sequences. And are shown in simple English language

4.2.2.1 Data flow Diagram

It is a block tool that shows the flow or steps/stages which are to be executed in a program. All the steps which are written in the second stage are now presented in a diagrammatic manner so as to make it easily understandable. Making of DFD helps us in increasing our process of program development because it facilitates us to define the logic, detecting and removing errors in a program design.

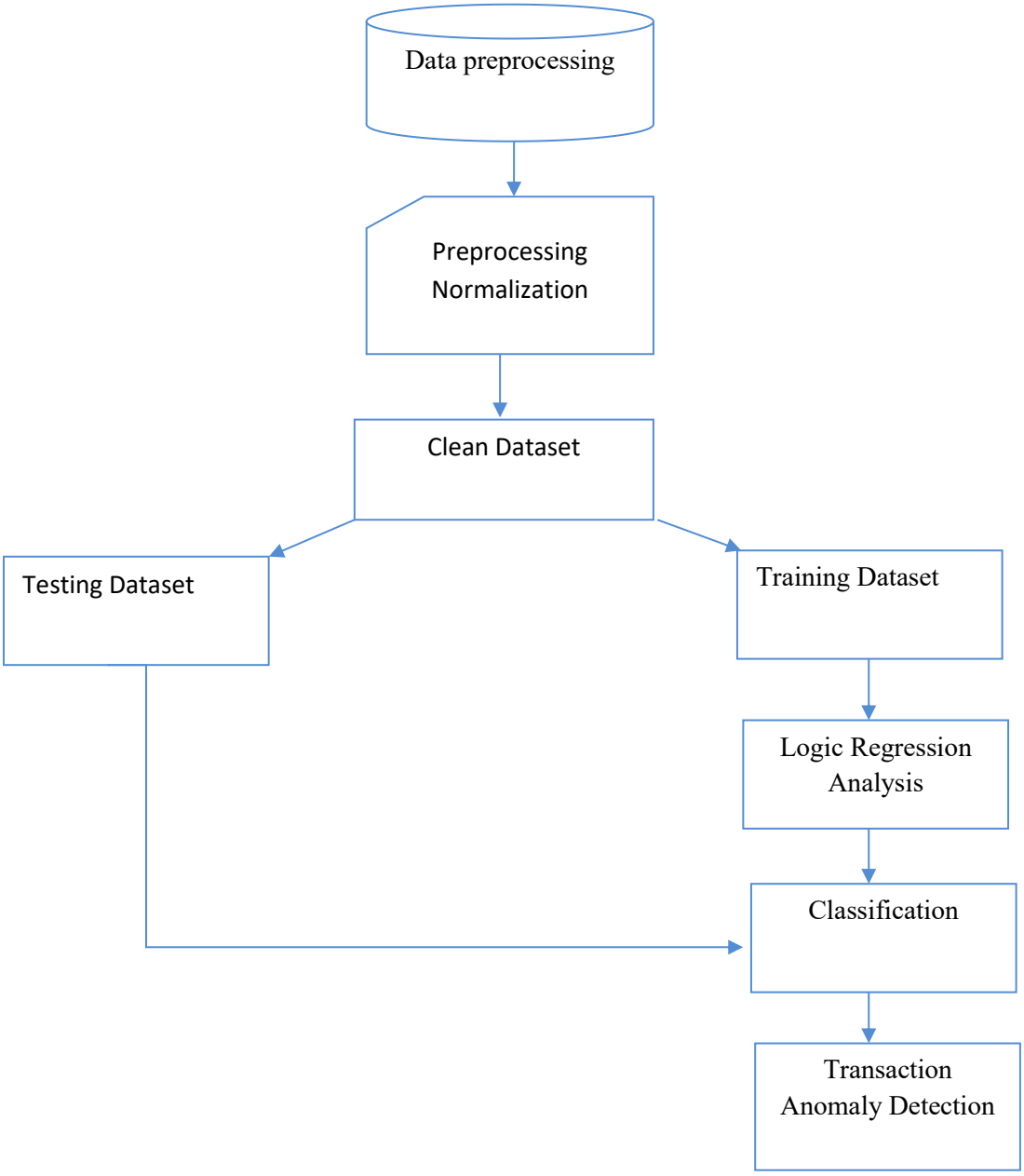

Fig 2.1 Data flow Diagram of the proposed Anomalous transactions detection system

4.2.3 Coding

This step involves writing the actual computer program which is usually the instructions written in a particular computer language.In this step programmer writes the instructions in a computer language to solve the problem. The coding process depends upon the information obtained

previous steps, and the choice of language depends upon the requirements and facilities available with a language.

4.2.4 Debugging

This stage of program development is an important process, wherein all the errors in the programs are detected and corrected. Debugging is also known as program validation. Some common errors which might occur in the programs include:

i. None initialization of variables.
ii. Reversing of order of operands.
iii. Confusion of numbers and characters.
iv. Inverting of conditions e.g. jumping on zero instead of on not zero.

4.2.5 Testing

In this stage the developed application is tested so to ensure that there is no bug. This is done by entering dummy data (includes usual, unusual and invalid data) to check the behavior and result of the program towards the given data.

4.2.6 Documentation

This usually the last stage in application development and it is an essential step in the program development, though most programmers neglect this stage by giving many reasons. It provides users and the programmer the necessary information to maintain the software in future, and also may help the programmer to correct the problems that may occur in the program.

4.3 Development Language used

The study used Python programming language to develop the model to detect and classify depression using dataset of tweets, posts, and comments of twitter users. Below are other programming language used in achieving the research objectives

PYTHON: Python is high-level programming, interpreted, object oriented language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together.

JUPYTER: Exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.It supports over 100 programming languages called ("kernels" in the Jupyter ecosystem) including Python, Java, R, Julia, Matlab, Octave, Scheme, Processing, Scala, and many more.

4.4 Program Design

The system is designed to detect anomalous transactions using data's obtained from transactions done in an ecommerce platform at intervals. A model was developed using logic regression algorithm. Data's of about 100 rows where obtained and loaded into the pandas dataframe in jupyter environment, this will enable the processing and transformation of the data from one data type to another. The data are cleaned using different libraries to remove characters, hyperlinks, 0 (zero) value, empty rows or column The model was trained and tested, hence was found to have an accuracy score of 96%. This shows the model as capable of detecting anomalous transactions in an Ecommerce platform at real time.

4.5 Input Design specification

The input to the system is a dataset obtained from a person transaction at intervalsusing different payment methodchannels such as transfer, cash or Bitcoin in an ecommerce platform. The dataset contains13000 rows of transactions made at various intervals.

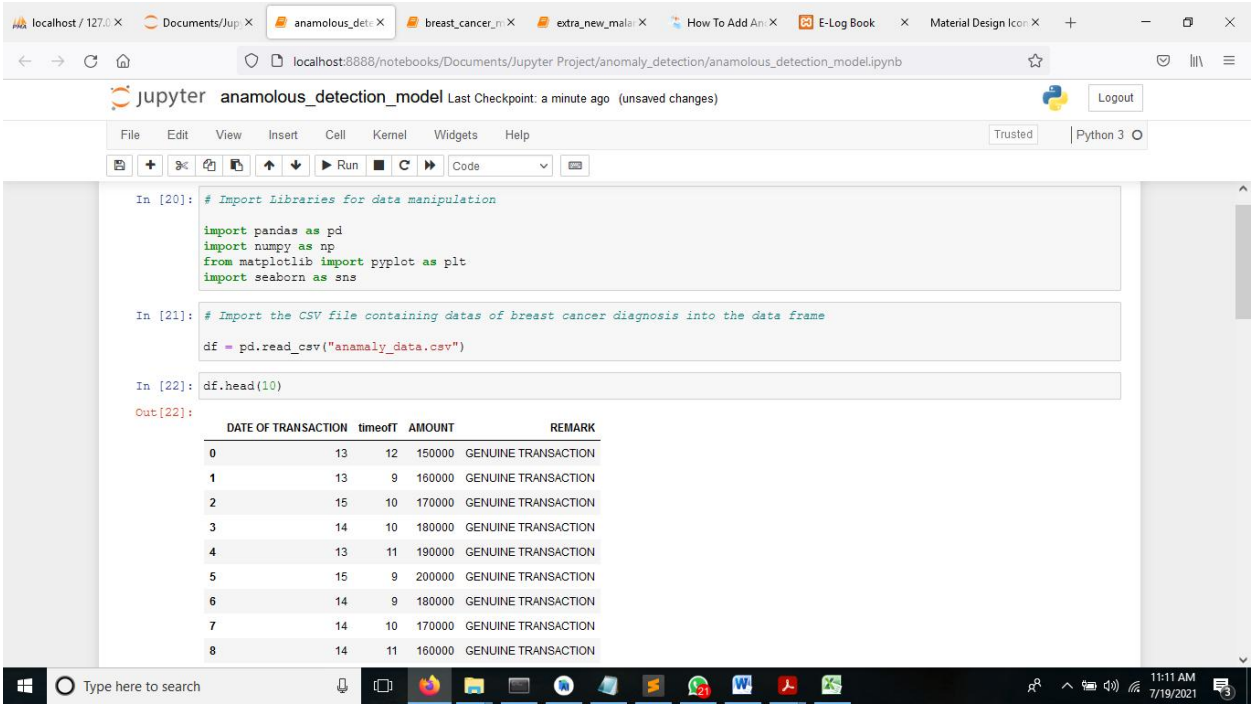

Fig 2.2 Dataset input of the model to detect anomalous transaction

4.6 Output Specification and Design

The output design of the developed system consists of charts, pictures, and accuracy score of model after model was trained and tested.

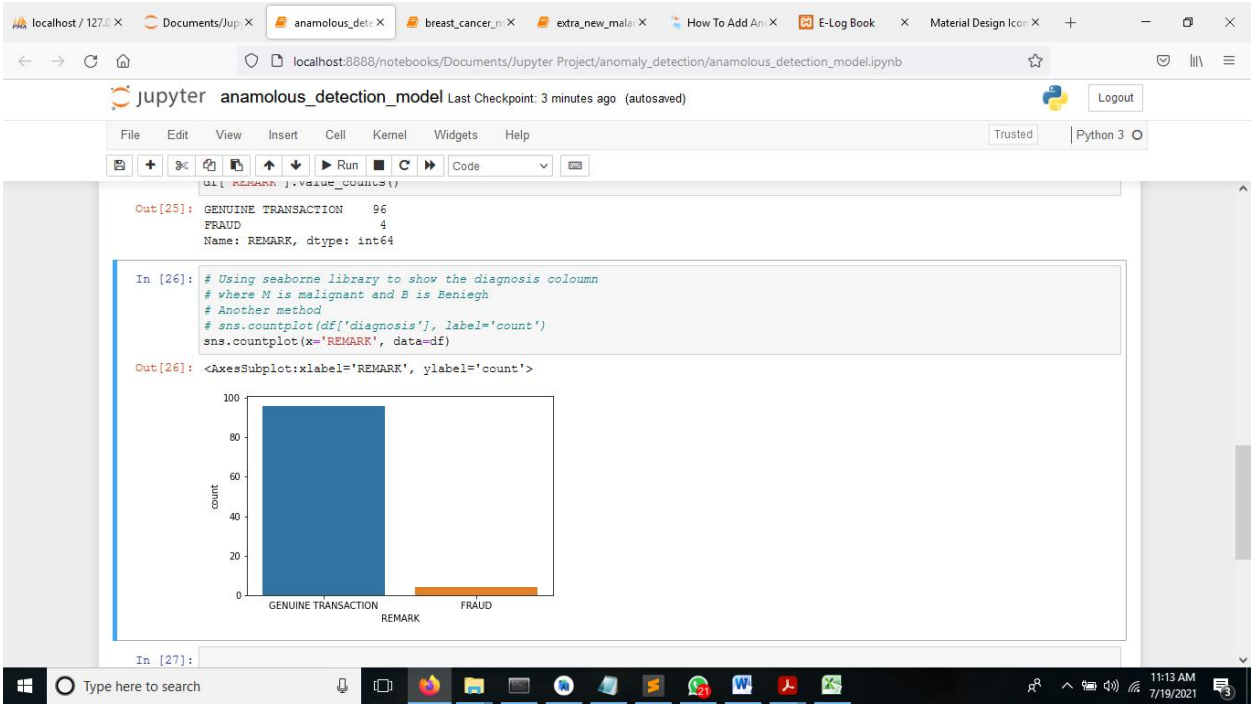

Fig 2.3 Showing graphical representation of categorical data of Genuine and fraudulent transactions

4.7 System Requirement

The system requirements provide an overview of the least criteria the hardware and software must obtain before its usage. It is to provide a detailed overview of the hardware and software product, its parameters and goals to the wide range of users. This describes the project's target audience and its user interface, hardware and software requirements.

4.8 Hardware Requirement

The hardware requirements are as follows:

I. A minimum of 150GB hard disk drive.

II. At least a Pentium III 800 MGHz MMX Intel Processor

III. Minimum of 4GB Random Access Memory.

IV. A CD-ROM Drive.

V. A super Video Graphic Adapter (SVGA) Monitor.

VI. A stabilizer and an uninterruptible Power Supply Unit (UPS).

VII. A keyboard and a mouse.

4.9 Software Requirement

I. The minimum software requirements for running this application are:

II. Microsoft Windows (XP, Vista, Windows 7, Windows 8).

III. Anaconda 2.1

IV. Jupyter 2.1 and above

V. Python 2.7 and above

VI. Browsers (Mozilla firefox 9.2.1/Google chrome 21.0.1180.81 or any other browser that supports HTML5)

VII. Anti-virus

4.10 SYSTEM TESTING

Testing is the last stage for every software design and implementation and it presents an interesting anomaly for the software engineer where he attempts to build software from an abstract concept to a tangible product. The model was tested repeatedly for errors that may arise during development. Data's are collected during system testing and this data's forms the trend for building a better software

4.10.1 Test Plan

A test plan documents the strategy that will be used to verify and ensure that a product or system meets its design specifications and other requirements. But in the course of this research work, the design verification and compliance test was used.

4.10.2 Unit Test

Each unit of the new system was tested (test run) individually alongside with the old system in other to identify areas of further enhancement and development.

4.10.3 System Test

The entire system was as well tested (test run) in general alongside with the old system in other to identify areas of further enhancement and development.

4.10.4 Packaging (Integration)

The software will be designed using visual basic. After which will be complied and packed for easy installation in any computer system and further use. The complied software will be transferred in to a CD.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Summary

Anomalous transaction detection is becoming important topic of research, as different types of attacks on bank accounts, ecommerce platform and other financial transactions are increasing at an alarming rate. Anomalous detection is a complex issue that requires a substantial amount of planning before throwing machine learning algorithms at it. Nonetheless, it is also an application of data science and machine learning for the good, which makes sure that the customer's money is safe and not easily tampered with.

This proposed a robust framework to process large volume of data containing financial transactions from an ecommerce platform, the functionality of the framework can be extended to extract real time data from different desperate sources. The extracted data is then used to build strong analytical model. To improve the analytical accuracy of fraud prediction, we have implemented three different analytical techniques. These analytical models are run on a financial transactions dataset and accuracy of the analytical model is evaluated with help of confusion matrix. Decision tree classifier algorithm was chosen as machine learning technique to build the model.

5.2 Conclusion

In this project, an anomaly detection framework was designed for an online store, to detect anomalous transactions from a customer transaction. Our models were able to detect the most important anomalies effectively. Besides detecting anomalies, we developed an approach that relies on the anomaly scores from a density model to explain the anomalies detected.A machine learning technique was employed in detecting anomalous transactions in an Ecommerce platform. The results obtained after successfully training and testing the model using dataset obtained from one of the new Generation Ecommerce platform was found to have an accuracy score of 96%. This clearly shows that the system is capable of detecting anomalous transactions at various hours of the day.

5.3 Recommendation

In this project, the dataset used in building the model had a limited number of features necessary to equally to identify a fraudulent transactions. It is recommendation for future works that focus should be on having a high number of features so as to effectively predict abnormal transactions and reduce cases of false positives.

REFERENCES

Baulier, G. D., Cahill, M. H., Ferrara, V. K., and Lambert, D. (2000) Automated fraud managementin transaction-based networks, Dec. 19 2000 US Patent 6,163,604.

Burge, P., and Shawe-Taylor, J. (2001). An unsupervised neural network approach to proling the behavior of mobile phone users for use in fraud detection. Journal of parallel and distributed computing, 61(7):915{925, 2001.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence and Research, 16, 2002, pp. 321-357.

Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley.

Dietterich, T. G. and Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Machine Learning, 255, 0-13.

Ingham, K. (2006) HTTP-delivered attacks against web servers. Retrieved December 14, 2011, from http://www.i-pi.com/HTTP-attacks-JoCN-2006/

Jaquith, A. (2002). The Security of Applications: Not All Are Created Equal, @Stake, Inc. Retrieved July 27, 2011, from http://www.securitymanagement.com/archive/library/atstake_tech0502.pdf

Katzgrau, K. (2008). KLogger. Retrieved September 15, 2011, from http://codefury.net/projects/klogger/

Kim, Jaekwon, Youngshin Han, and Jongsik Lee. (2016) "Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process." Advanced Science and Technology Letters 133 (2016): 79-84.

Kingdon, J., & Feldman, K.S., (2002), “Data monitoring and analysis system for bank transactions. constructs aggregate profiles of received data and investigates t0 identify its characteristic patterns of behavior[C]”, SEARCHSPACE LTD (SEAR-Non-standard).

Kruegel, C., Vigna, G., & Robertson, W. (2005). A multi-model approach to the detection of web-based attacks. Computer Networks, 48(5), 717-738. http://dx.doi.org/10.1016/j.comnet.2005.01.009

Larouche, F. (2007). SQL Power Injector Product Information. Retrieved December 17, 2011, from http://www.sqlpowerinjector.com/

Lee, S. Y., Low, W. L., & Wong, P. Y. (2002). Learning fingerprints for a database intrusion detection system. In Computer Security—ESORICS 2002 (pp. 264-279). Springer Berlin Heidelberg.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111{3119, 2013.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, and X. Sun (2011). The application of data mining techniques in nancial fraud detection: A classication framework and an academic review of literature. Decision support systems, 50(3):559{569, 2011.

Sharma, Shiven, *et al*. (2018) "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance." 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018.

Stofella, P., (1997), “The DBInspector Project” [J] // Proceedings of the IEEE international workshop on research issues in data engineering, AI Magazine, (5):73-75.