International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Anomaly Detection in Network Traffic Using Unsupervised Machine Learning Approach

Shraddha.D.Shegokar[1], Prof. P. P. Rane[2], Prof. S. A. Vyawhare[3]

[1]*Department of Computer Science and Engineering, Rajarshi Shahu College of Engineering, Buldhana 443001,Maharashtra India*

[2]*Department of Computer Science and Engineering, Rajarshi Shahu College of Engineering, Buldhana 443001,Maharashtra India*

[3]*Department of Computer Science and Engineering, Rajarshi Shahu College of Engineering, Buldhana 443001,Maharashtra India*

ABSTRACT -The advent of IoT technology and the increase in wireless networking devices has led to an enormous increase in network attacks from different sources. To maintain networks as safe and secure, the Intrusion Detection System (IDS) has become very critical. Intrusion Detection Systems (IDS) are designed to protect the network by identifying anomaly behaviors or improper uses. Intrusion Detection systems provide more meticulous security functionality than access control barriers by detecting attempted and successful attacks at the endpoint of within the network. Intrusion prevention systems are the next logical step to this approach as they can take real-time action against breaches. To have an accurate IDS, detailed visibility is required into the network traffic. The intrusion detection system should be able to detect inside the network threats as well as access control breaches. IDS has been around for a very long time now. These traditional IDS were rules and signature based. Though they were able to reduce false positives they were not able to detect new attacks. In today's world due to the growth of connectivity, attacks have increased at an exponential rate, and it has become essential to use a data-driven approach to tackle these issues. In this paper, the KDD dataset was used to train the unsupervised machine learning algorithm called Isolation Forest. The data set is highly imbalanced and contains various attacks such as DOS, Probe, U2R, R2L. Since this data set suffers from redundancy of values and class imbalance, the data preprocessing will be performed first and also used unsupervised learning. For this network traffic-based anomaly detection model isolation forest was used to detect outliers and probable attack the results were evaluated using the anomaly score

Keywords-anomaly detection, isolation forest, machine learning, intrusion detection system, KDD Cup, NSL-KDD.

1. INTRODUCTION

A sudden spike or dip in a metric is an anomalous behavior and both cases need attention. Detection of anomaly can be solved by supervised learning algorithms if we have information on anomalous behavior before modeling, but initially without feedback its difficult to identify that points.Anomaly detection is important and finds its application in various domains like detection of fraudulent bank transactions, network intrusion detection, sudden rise/drop in sales, change in customer behavior, etc. So we model this as an unsupervised problem using algorithms like Isolation Forest,One class SVM and LSTM. Here we are identifying anomalies using isolation forest.

PROBLEM STATEMENT: -

In recent years, the number of unknown attacks has increased rapidly both from inside and outside the organization. So, it has become imperative to provide customers and users secure access to the network and at the same time keeping the network attack free. That is why IDS (Intrusion Detection System) was introduced.

PROBLEM DEFINITION:-

The extensive increase of attacks has the potential for extremely negative im-pacts on individuals and society. Therefore, intrusion detection in netwrok traffic has recently become an emerging research that is attracting tremendous attention. Therefore there is a need to build a platform to detect anomalies.

PROPOSED SOLUTION: -

The idea is that we will apply isolation forest algorithm on a unlabelled dataset to see the accuracy .

- IF builds an ensemble of random trees for a given data-set and anomalies are points in the tree structure with the shortest average range
- As the data set is highly imablanced isolation forest can be used .
- While using isolation forest it is impoprtant to keep the n-estimators and the contamination parameters below the specific value.

2. LITERATURE REVIEW

Title:- Anomaly detection in Network Traffic Using Unsupervised Machine learning Approach

Author:- Aditya Vikram, Mohana

Description:- In this paper we get a brief description on what is IDS(intrusion detection system),what is isolation forest , what are the parameters that affect the network in day to day life and a short information on different types of attacks. Nowadays, the number of networking devices is increasing at an exponential rate, and the workplace has a lot of devices that handle sensitive data communication. In recent years, the number of unknown attacks has increased rapidly both from inside and outside the organization.So, it has become imperative to provide customers and users secure access to the network and at the same time keeping the network attack free. There are a few different ways to avert digital attacks, one of which is by utilizing Intrusion Detection Systems. IDS is one segment of system security that ensures information and data security, by checking the traffic on a bundle of information to identify an interruption or anomaly. The IDS used in this study is anomaly-based. The anomaly detection or outlier detection system assume that abnormal behavior is malicious. The idea is to train the machine learning model to learn normal behavior and then look for abnormal behavior or anomalies and raise alerts accordingly. Anomaly detection is perfect for such problems. IDS have been around for decades. The initial models relied on heuristics and thresholds, which helped to reduce the false positive and false negatives. These systems were notable to detect new attacks and due to the increasing number of wireless devices and growth of cloud computing the frequency of attacks has increased exponentially and it has become essential for companies to use a data-driven approach. However, there are some issues associated with the machine learning- based approach. These include detecting normal instances as false positives,which might seem benign but in reality, it leads to wastage of time and resources. Isolation Forest It is a tree-based anomaly or outlier detection algorithm. It is based on the logic that isolating an outlier in a tree structure is possible rather than use any density measure like one class SVM. In IF, it remains as not analyzed or do profiling of the normal points. IF builds an ensemble of random trees for a given data set and anomalies are points in the tree structures with shortest average path length. Useful when the large imbalance is there and scattered da

points are there. Besides, when different feature space is there the outlier data points have lesser depths on the tree structure as compared to normal data points this is because it is easier to isolate and differentiate. To build an outlier detection model using the isolation forest algorithm, it is essential to carefully select n estimators and contamination parameters

Title:- Anomaly Detection

Author:- AVI Networks

Description:- In this paper we get to know what are the different types of anomalies , the different techniques used for anomaly detection and its use cases. Anomaly detection is the identification of rare events, items, or observations which are suspicious because they differ significantly from standard behaviors or patterns. Anomalies in data are also called standard deviations, outliers, noise, novelties, and exceptions. In the network anomaly detection/network intrusion and abuse detection context, interesting events are often not rare—just unusual. For example, unexpected jumps in activity are typically notable, although such a spurt in activity may fall outside many traditional statistical anomaly detection techniques. Many outlier detection methods, especially unsupervised techniques, do not detect this kind of sudden jump in activity as an outlier or rare object. However, these types of micro clusters can often be identified more readily by a cluster analysis algorithm. There are three main classes of anomaly detection techniques: unsupervised, semi-supervised, and supervised. Essentially, the correct anomaly detection method depends on the available labels in the dataset. Supervised anomaly detection techniques demand a data set with a complete set of “normal” and “abnormal” labels for a classification algorithm to work with. This kind of technique also involves training the classifier. This is similar to traditional pattern recognition, except that with outlier detection there is a naturally strong imbalance between the classes. Not all statistical classification algorithms are well-suited for the inherently unbalanced nature of anomaly detection. Semi- supervised anomaly

detection techniques use a normal, labeled training data set to construct a model representing normal behavior. They then use that model to detect anomalies by testing how likely the model is to generate any one instance encountered. Anomaly Detection In Netwrok Traffic Using Unsupervised Machine Learning 6 Unsupervised methods of anomaly detection detect anomalies in an unlabeled test set of data based solely on the intrinsic properties of that data. The working assumption is that, as in most cases, the large majority of the instances in the data set will be normal. The anomaly detection algorithm will then detect instances that appear to fit with the rest of the data set least congruently

Title:- Anomaly Detection-A Survey

Author:-VARUN CHANDOLA, ARINDAM BANERJEE and VIPIN KUMAR

Description: - At an abstract level, an anomaly is defined as a pattern that does not conform to expected normal behavior. A straightforward anomaly detection approach, therefore, is to define a region representing normal behavior and declare any observation in the data which does not belong to this normal region as an anomaly. But several factors make this apparently simple approach very challenging: —Defining a normal region which encompasses every possible normal behavior is very difficult. In addition, the boundary between normal and anomalous behavior is often not precise. Thus an anomalous observation which lies close to the boundary can actually be normal, and vice-versa. —When anomalies are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear like normal, thereby making the task of defining normal behavior more difficult. —In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future. —The exact notion of an anomaly is different for different application domains. For example, in the medical domain a small deviation from normal (e.g., fluctuations in body temperature) might be an anomaly, while similar deviation in the stock market domain (e.g., fluctuations in the value of a stock) might be considered as normal. Thus applying a technique developed in one domain to another is not straightforward. —Availability of labeled data for training/validation of models used by anomaly detection techniques is usually a major issue. —Often the data contains noise which tends to be similar to the actual anomalies and hence is difficult to distinguish and remove. Due to the above challenges, the anomaly detection problem, in its most general form, is not easy to solve. In fact, most of the existing anomaly detection techniques solve a specific formulation of the problem. The formulation is induced by various factors such as nature of the data, availability of labeled data, type of anomalies to be detected, etc. Often, these factors are determined by the application domain in which the anomalies need to be detected. Researchers have adopted

1. 3. SYSTEM OVERVIEW

Project Scope

The purpose of Software Requirement Specifications (SRS) is to provide a detailed overview of the system. SRS provides a description of the system as well as lists any assumptions made while developing the system and all the constraints faced by the system. It also specifies the hardware and software requirements of the system.iple domains like money transaction ,Money investing application,Education sector**.**

System Requirement

The system requirement specification of our project will have the entire necessary requirement which will be a baseline of our project. The software requirement specification will incorporate functional and nonfunctional requirements, system architecture, data flow diagrams, UML diagrams, experimental setup requirements and performance metrics.

Hardware Requirements

A hardware interface is needed to run the software. Python IDLE and other necessary libraries is required which is minimal requirement.

- Processor: Pentium IV 2.6 Ghz
- 1 GB ram
- Monitor: 15 VGA color

Software Requirements

- Operating System : Windows 7/8.1/10 Linux
- Python

- Features Of Python
- Easy to Learn and Use

Python is easy to learn as compared to other programming languages. Its syntax is straightforward and much the same as the English language. There is no use of the semicolon or curly-bracket, the indentation defines the code block. It is the recommended programming language for beginners.

a) Expressive Language

Python can perform complex tasks using a few lines of

code. A simple example, the hello world program you simply type print("Hello World"). It will take only one line to execute, while Java or C takes multiple lines.

b) Cross-platform Language

Python can run equally on different platforms such as Windows, Linux, UNIX, and Macintosh, etc. So, we can say that Python is a portable language. It enables programmers to develop the software for several competing platforms by writing a program only once.

c) Object-Oriented Language

Python supports object-oriented language and concepts of classes and objects come into existence. It supports inheritance, polymorphism, and encapsulation, etc. The object-oriented procedure helps to programmer to write reusable code and develop applications in less code

4. PROJECT IMPLEMENTATION

4.1 Introduction

Anomaly detection deals with the identification of unusual patterns/behaviour that doesn't conform to the usual trend. It is applied in wide range of areas- Signal processing, Automation in manufacturing, Chemical reaction monitoring etc. Here we will narrow down to finding anomalous data points

4.2 Tools and Technologies Used

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

4.3 Verification and Validation for Acceptance

A. Verification

Software testing must follow approved methods and stan- dards; also, when tested, the models must meet these design specifica- tions. For this project, the Software Testing Plan describes the testing process for the software

B. Validation

- The process of determining the degree to which a model and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model.
- The process of determining the fitness of a model or simulation and its associated data for a specificpurpose.

4.4 Alorithm Details

4.4.1 Isolation Forest

Isolation forest detects anomalies by randomly partitioning the domain space. Yeah, you're heard me right- It works similar to Decision trees algorithm, where we start with a root node and keep on partitioning the space. In Isolation forest we partition randomly, unlike Decision trees where the partition is based on Information **gain.**

4.4.2 Steps To Build Isolation Forest

1. Select a feature at random from data. Let us call the random feature f.
2. Select a random value from the feature f. We will use this random value as a threshold. Let us call it
3. Data points where f ¡ t are stored in Node 1 and the data points where f t go in Node 2.
4. Repeat Steps 1–3 for Node 1 and Node 2.
5. Terminate either when the tree is fully grown or a termination criterion is met.

4.4.2.1

4.4.2.2 The following figure shows its mechanism for 1 Dimensional Data:

4.4.2.3 The feature to split on and threshold are

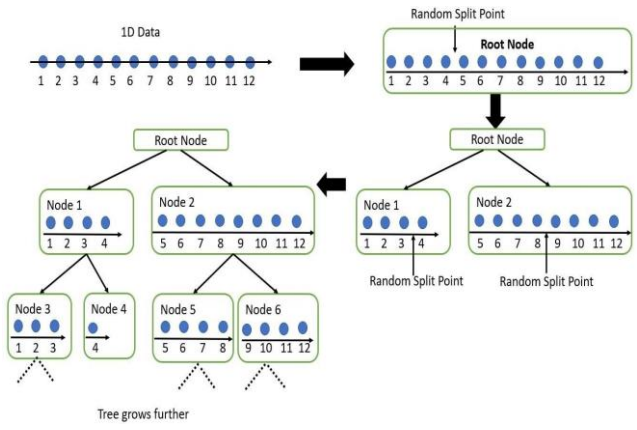

Figure 4.1: Splitting of Data

4.5 Working of Isolation Forest

Assume the data above has an anomaly. In that case, the anomalous point will be far away from the other data points. Isolation forests are able to isolate out anomalies very early on in the splitting process because the Random Threshold used for.

splitting has a large probability of lying in the empty space between the outlier and the data if the empty space is large enough. As a result, anomalies have shorter path lengths. After all, the split point(the threshold)is chosen at random. So, the larger the empty space, the more likely it is for a randomly chosen split point to lie in that empty would look in the presence of an Anomaly

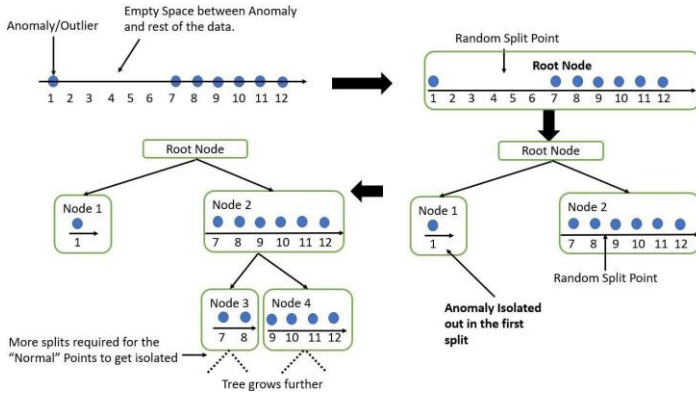

Figure 4.2: Splitting of Data with Anomaly

As we can see, due to the large space between the anomaly and the rest of the data, it is very likely that a random split will lie in this empty region.
Please note that the trees can grow either:

- Till there is exactly one data point in each leaf node. Or
- Till termination criterion regarding the minimum number of data points in a leaf node is reached.

Here is briefly how Isolation forests work:

- Construct an Isolation Tree either from the entire feature set or a randomly chosen subset of the feature set.
- Construct n such Isolation trees.
- Calculate an Anomaly score for each data point. The Anomaly score is a non-linear function of the Average path length over all Isolation trees. The path length is equivalent to the number of splits made by the Isolation tree to isolate a point. The shorter the Average path length, the larger are the chances of the point being an anomaly(as we saw earlier in the diagram

The Isolation forest in skearn has 4 important inputs:

- **n-estimators**: Number of Isolation trees trained.
- **max-samples**: Number of data points used to train each tree.
- **contamination**:Fraction of anomalous data points. For example, if we suspect 5percent of the data to be anomalous, we set contamination to 0.05
- **max-features**:Number of features to be used to train each tree(This is in contrast to Random Forests where we decide on a random subset of features for each split).

5. SYSTEM DESIGN

5.1 Introduction

Nowadays, the number of networking devices is increasing at an exponential rate, and the workplace has a lot of devices that handle sensitive data communication. In recent years, the number of unknown attacks has increased rapidly both from inside and outside the organization.So, it has become imperative to provide customers and users secure access to the network and at the same time keeping the network attack free. There are a few different ways to avert digital attacks, one of which is by utilizing Intrusion Detection Systems. IDS is one segment of system security that ensures information and data security, by checking the traffic on a bundle of information to identify an interruption or anomaly. The IDS used in this study is anomaly-based. The anomaly detection or outlier detection system assume that abnormal behavior is malicious. The idea is to train the machine learning model to learn normal behavior and then look for abnormal behavior or anomalies and raise alerts accordingly. Anomaly detection is perfect for such problems. IDS have been around for decades. The initial models relied on heuristics and thresholds, which helped to reduce the false positive and false negatives. These systems were notable to detect new attacks and due to the increasing number of wireless devices and growth of cloud computing the frequency of attacks has increased exponentially and it has become essential for companies to use a data-driven approach. However, there are some issues associated with the machine learning- based approach. These include detecting normal instances as false positives,which might seem benign but in reality, it leads to wastage of time and resources.

System Architecture

A description of the program architecture is presented

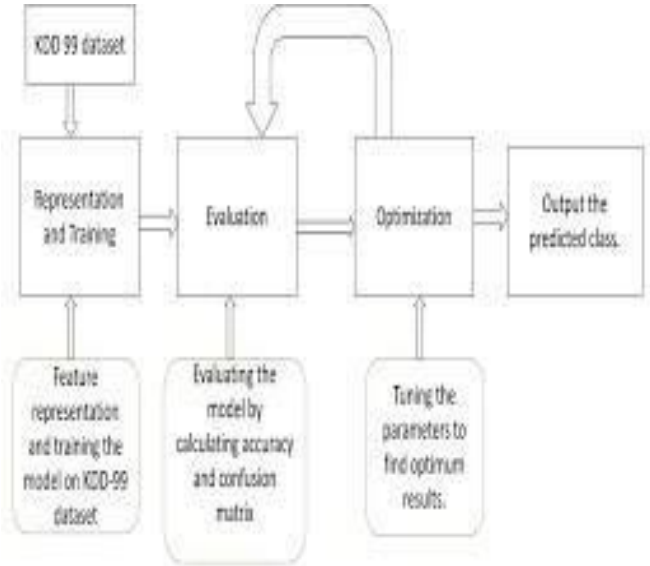

Figure 5.1: Architecture

6. MATHEMATICAL MODEL

6.1 Anomaly Score

Anomaly score is given by the following formula- where n- Number of data points c(n)- It is the average path length of unsuccessful search in a Binary search tree.We grow an isolation tree by randomly choosing a feature and randomly partitioning. This is very similar to the Binary Search tree. Thus we can approximate the

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

average path length of for a node termination with the unsuccessful search in a Binary Search tree. Thus we use c(n) as the reference.

Fig 6.1 formula for calculate anomaly score

- when $E(h(x)) \to c(n), s \to 0.5;$
- when $E(h(x)) \to 0, s \to 1;$
- and when $E(h(x)) \to n-1, s \to 0$

Fig 6.2 caption

It is always better to represent score between 0 to 1 because the score can now be interpreted as a probability. For example, say for a data point if we get the anomaly score as 0.8, then we can interpret such that the point has a probability of 80percent to be an anomalous point.

E(h(x))- Average of path lengths from the Isolation forest

1. As score is closer to 1, then it is an anomalous point
2. As the score is closer to 0, it a normal observation
3. A score near 0.5, indicates it doesn't have much distinction from normal observations

6.2 Flow-Chart Diagram

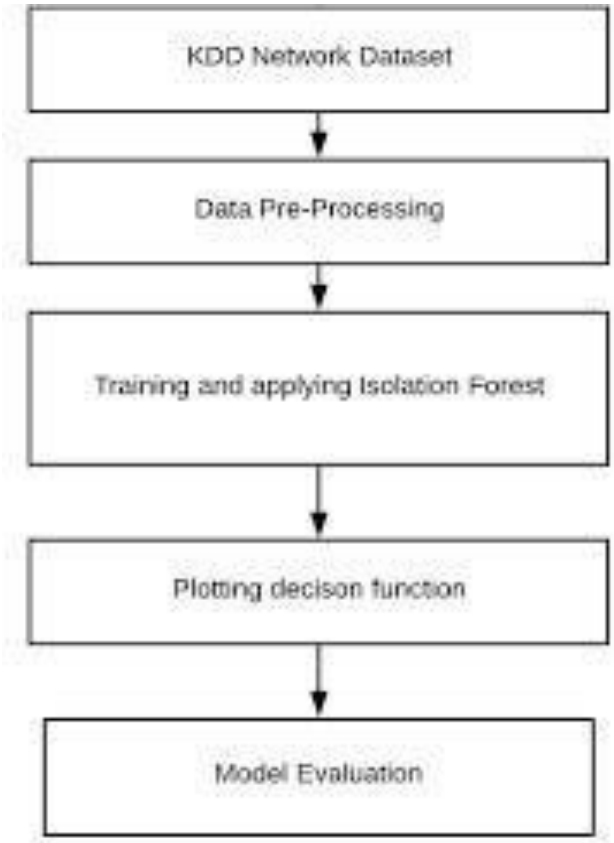

Figure 6.2: Flow Chart Diagram

6.3 Plotting Datase

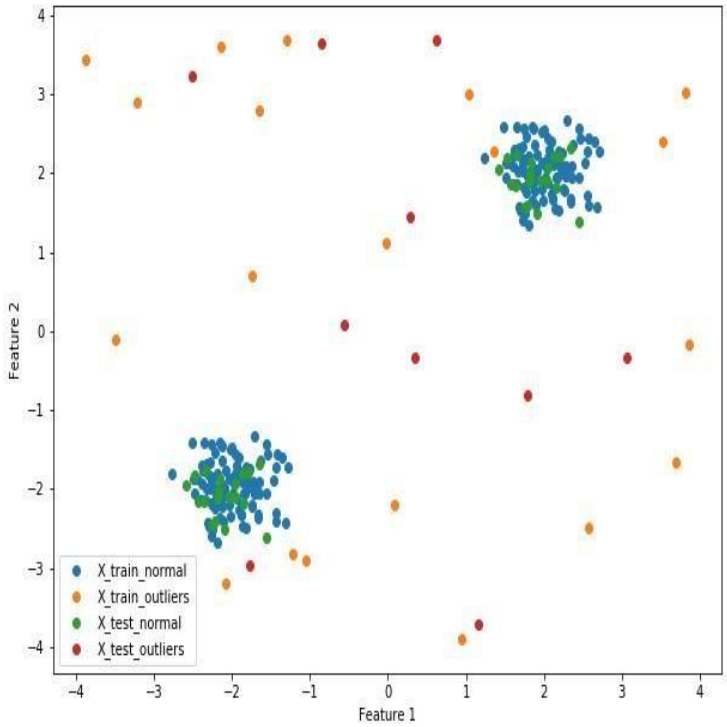

Figure 6.3: Caption

6.4 Visualising training predictions

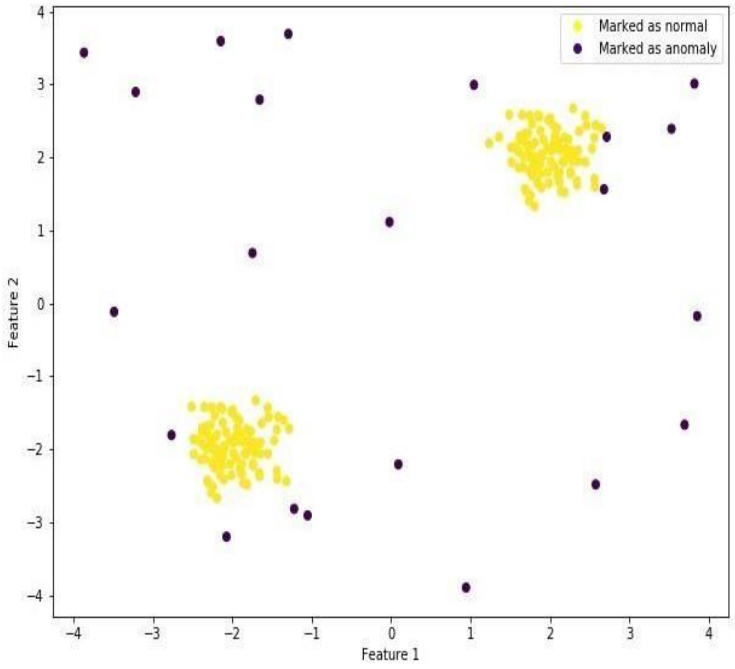

Figure 6.4: Caption

6.5 Visualising test prediction

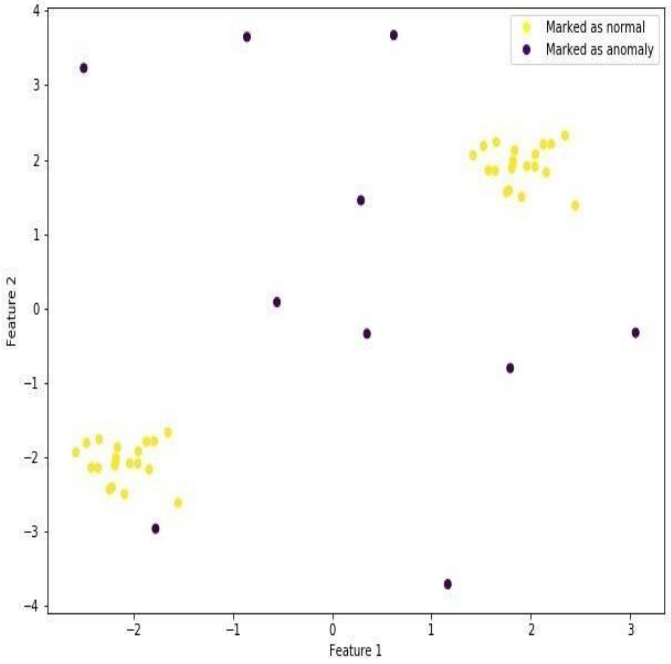

Figure 7.1: Caption

7. Advantages

7.1 Automated KPI analysis

- For most businesses, KPI analysis is still a manual
- task of sorting through all of their digital channel's data across different dashboards. Depending on how much data the company collects, this can be an
- incredibly time-consuming task.
- But, when using an anomaly detection system, AI algorithms are constantly scanning through all your data across all your dashboards and analysing metrics 24/7. This means you'll no longer have to check BI tools like Google Analytics constantly to know what's going on in your KPIs.
- Instead, the anomaly detection will alert users straightaway when it finds anomalies and unusual behaviour – good or bad – in their metrics, so you can use these insights in your strategy without delay.

7.2 Prevention of security breaches and threats:

- With hacker attacks now taking place every 39 seconds, online security has never been more critical. And anomaly detection is one of the best ways to prevent

security breaches and threats to your business and website.

- According to IBM, the average time to detect a breach was 206 days in 2019-which is a lot of damage.But, with anomaly detection, security breaches can be detected as soon as they happen, because the AI is constantly scanning your data and will pick up on anything unusual immediately

7.3 Discovery of hidden performance opportunities:

- Currently, digital teams can spend spend hours and hours each week searching through data for ways to improve digital performance. If anomaly detection is applied, this kind of repetitive work can be eliminated, freeing up time to plan and execute more performance-driving strategies=
- Not to mention the fact that, with AI conducting thorough analysis, many things that were hidden in your data will be uncovered.

8. Disadvantages

- So by using ML approach anomaly detection we can detect anomalies faster than any normal traditional method and provide a better security to our data.
- It is not always certain that the obtained results will be useful since there is no label or output measure to confirm its usefulness.
- One cannot accurately define the sorting and output of an unsupervised task. It is heavily dependent on the model and in-turn on the machine.
- The results often have lesser accuracy.

9. Conclusion

An unsupervised machine-learning model was built due to highly imbalanced data. The AUC score was computed is 98.3 percent. The “n estimators” parameter was kept at 100. The “contamination” parameter value was 4percent of the total number of samples or 0.04. There is tremendous growth in the different types of network attacks and thus organizations are developing Intrusion Detection System (IDS) that are not only highly efficient but also capable of detecting threats in realtime. Anomaly detection has great promise in this area, as it is efficient to train and detects anomalies with low false positive and false negative rates. In the implementation, it has been found that the anomaly detection process can be improved using various values of the available parameters for these algorithms. Also, it could be concluded that a more complete and clean data set leads to better results. The contamination parameter is very important in deciding the proportion of anomalies that could be detected. It is important to realize that machine learning, deep learning application is fairly new in the network security domain, and therefore there are still challenges related to scalability and efficiency.

10.0.1 10. Future Scope

The accuracy of the model can be further improved if the machine learning algorithms were combined and a hybrid model was prepared. Feature normalization can also be used to increase accuracy. Different feature selection algorithms can also be used to select certain features that can influence results better. Deep learning techniques have been proven to showhigher accuracy and are robust . Due to the growing number of attacks from within an organization, it has become highly important to analyze the behavior and detect anomaly in real-time with high efficiency. This can be achieved using user and entity behavior analytics along with machine learning methods. Unsupervised machine learning can also be used along with supervised to build a hybrid system that can give better results. Parallelization is the classic computer science answer to performance problems. In the future, the model could be improved to intake real-time data and recommend attacks due to variation in network traffic.

11. REFERENCES

[1] G. Karatas et al., "Deep Learning in Intrusion Detection Systems" 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), Turkey,2018.

[2] H. Azwar et al., "Intrusion Detection in secure network for Cybersecurity systems using Machine Learning" 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences ,Bangkok, Thailand, 2018.

[3] Y. Chang et al., "Network Intrusion Detection Based on Random Forest and Support Vector Machine," IEEE International Conference on Computational Science and Engineering (CSE), Guangzhou, 2017\

[4] Brao, Bobba et al., "Fast kNN Classifiers for Network Intrusion Detection System", Indian Journal of Science and Technology. 2017.

[5] 5M. Z. Alom et all., "Network intrusion detection for cyber security using unsupervised deep learning approaches", 2017 IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, 2017.

[6] Mukkamala et al., "Intrusion detection using neural networks and support vector machines", International Joint Conference 2012.

[7] Azwar, Hassan et all.,"Intrusion Detection in secure network for Cybersecurity systems using Machine Learning and Data Mining", 2018.

[8] Jeya, P et al., "Efficient Classifier for R2L and U2R Attacks", International Journal Comput. Appl. (2012)

[9] Mohana, NK Srinath "Trust Based Routing Algorithms for Mobile Adhoc Network", International Journal of Emerging Technologies and Advanced Engineering (IJETAE), volume 2, issue 8, pp. 218-224, IJETAE.