IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 7.047 ⁜ Vol. 10, Issue 4, April 2022

DOI: 10.17148/IJIREEICE.2022.10479

Anomaly detection in network traffic using unsupervised machine learning approach

Prathamesh Kulkarni [1], Himanshu Samariya [2], Akash Sitoke[3], Aman Chandre [4], Prof. Sagar Dhanake[5]

Student, Computer Engineering, DYPIET, Pune, India[1, 2, 3, 4]

Assistant Professor, Computer Engineering, DYPIET, Pune, India[5]

Abstract: A sudden spike or dip in a metric is an anomalous behaviour and both the cases needs attention. Detection of anomaly can be solved by supervised learning algorithms if we have information on anomalous behaviour before modelling, but initially without feedback it's 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 behaviour, 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.

I. INTRODUCTION

One of the most prominent ways for making money for middle-class investors is an investment in Stock. After that, it is the actual trading business of high-class investors and traders. A company's share price is the most important point for investors which always fluctuates up and downwards.

Eyes always need on live price of share market and instant decision making is necessary to prevent loss of money and eventually to gain money. For this, you have to make a study of the company's financial history and future agenda.

Dependent on the overall study related to the market and company you can decide to invest. But you have limits to study because one cannot be sure that study and analysis are correct. Company's market history, the tendency of maintaining business in any period or slack, policies, and announcements are the key points of Stock Rate. It is a difficult field of work and needs a lot of experience to be a successful investor.

Prediction of stock has a predominant application which is pulchritudinous for the entire stock market investing circumference. The proposed theoretical predicament lays a groundbreaking establishment when it comes to efficiently predict the stock market peripheral.

II. 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.

- Title:- Anomaly Detection Author:- AVI Networks Description:- In this paper we get to know what are the different type of anomalies , the different techniques used for anomaly detection and its use cases.

- 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.

III. 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

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 7.047 ⁜ Vol. 10, Issue 4, April 2022

DOI: 10.17148/IJIREEICE.2022.10479

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 Requirements:

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 non-functional requirements, system architecture, data flow diagrams, UML diagrams, experimental setup requirements and performance metrics.

IV. PROJECT IMPLEMENTATION

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.

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.

Verification and Validation for Acceptance:

- **Verification:**

Software testing must follow approved methods and standards; also, when tested, the models must meet these design specifications. For this project, the Software Testing Plan describes the testing process for the software.

- **Validation:**

1. 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.
2. The process of determining the fitness of a model or simulation and its associated data for a specific purpose.

Algorithm Details:

- **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.

Steps To Build Isolation Forest:

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

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 7.047 ⁜ Vol. 10, Issue 4, April 2022
DOI: 10.17148/IJIREEICE.2022.10479

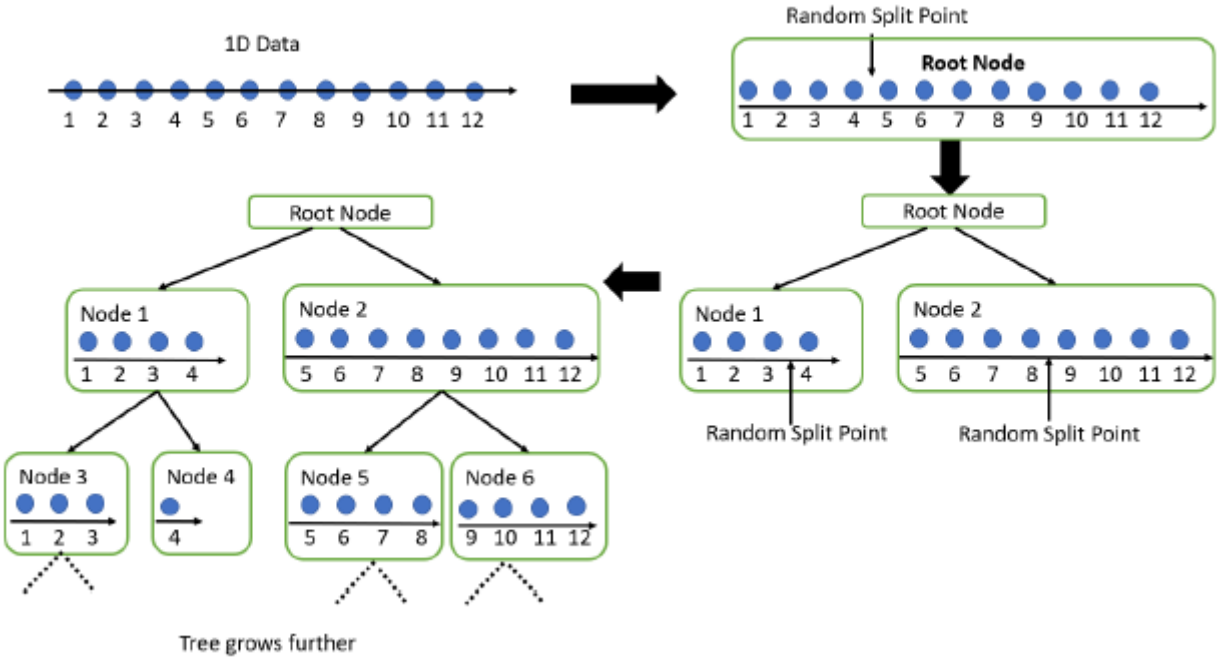

Fig.1:- Spitting of Data

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 region. Let us take a look at how an Isolation tree would look in the presence of an Anomaly.

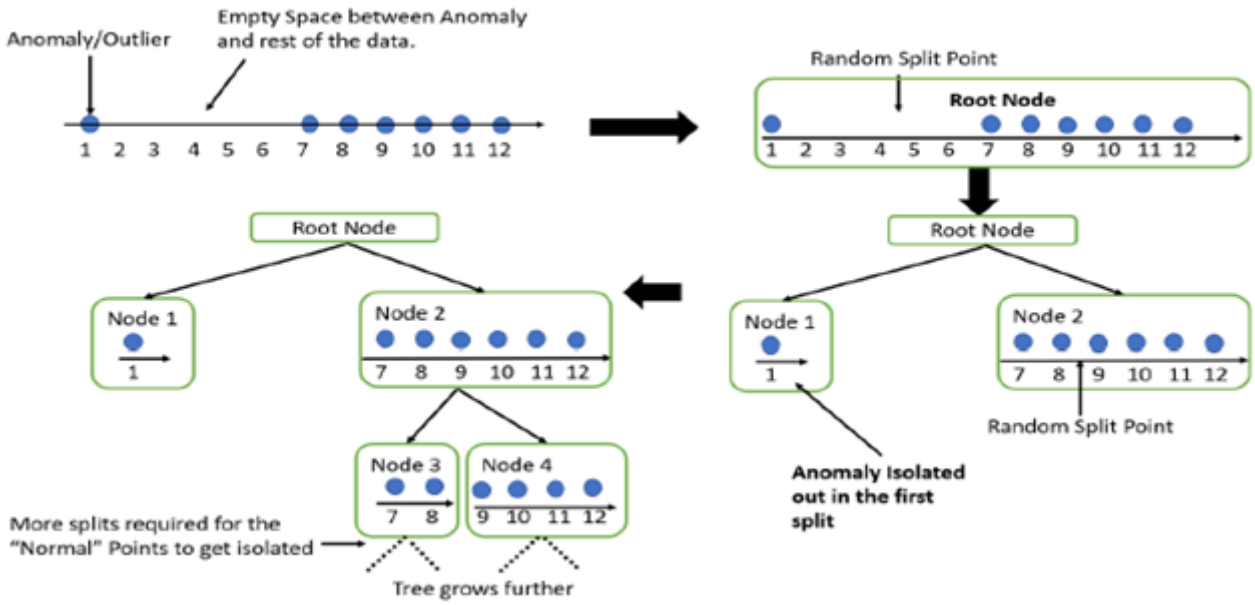

Fig.2:- Spitting of Data with Anomaly

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 7.047 ⁜ Vol. 10, Issue 4, April 2022

DOI: 10.17148/IJIREEICE.2022.10479

V. PROJECT IMPLEMENTATION PLAN

Project is carried out in different phases. There are number of activity/tasks implemented in step by step. Planning is essential for initiating our proposed work. It helps to get the details to plan the future estimates and implementation process. Data collection is one of the important process through which we get the relevant information for our project work. It also helps to get the process of analysis and to make comparative analysis. Initial settings like setting the protocols and creating the variables generates result and perform all operations like real system do Post. The data collected from the Result generation is in the raw format, that data will be processed and results are generated from it. We need to test whether the simulation result we got is matching to the real world or not and validate it. It takes 20 percent of the time.

VI. SYSTEM DESIGN

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. 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 anomaly detection or outlier detection system assume that abnormal behavior is malicious. 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.

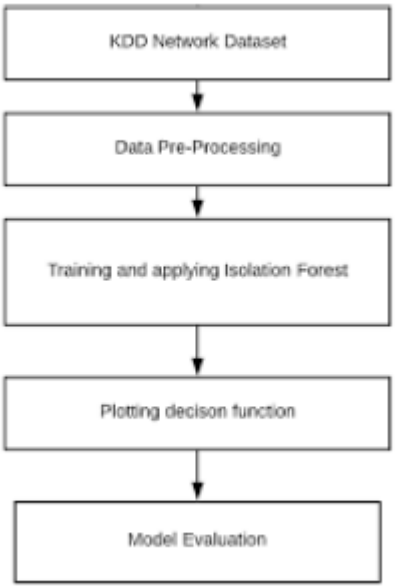

Fig.3:- Flow Diagram

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 7.047 ⁜ Vol. 10, Issue 4, April 2022
DOI: 10.17148/IJIREEICE.2022.10479

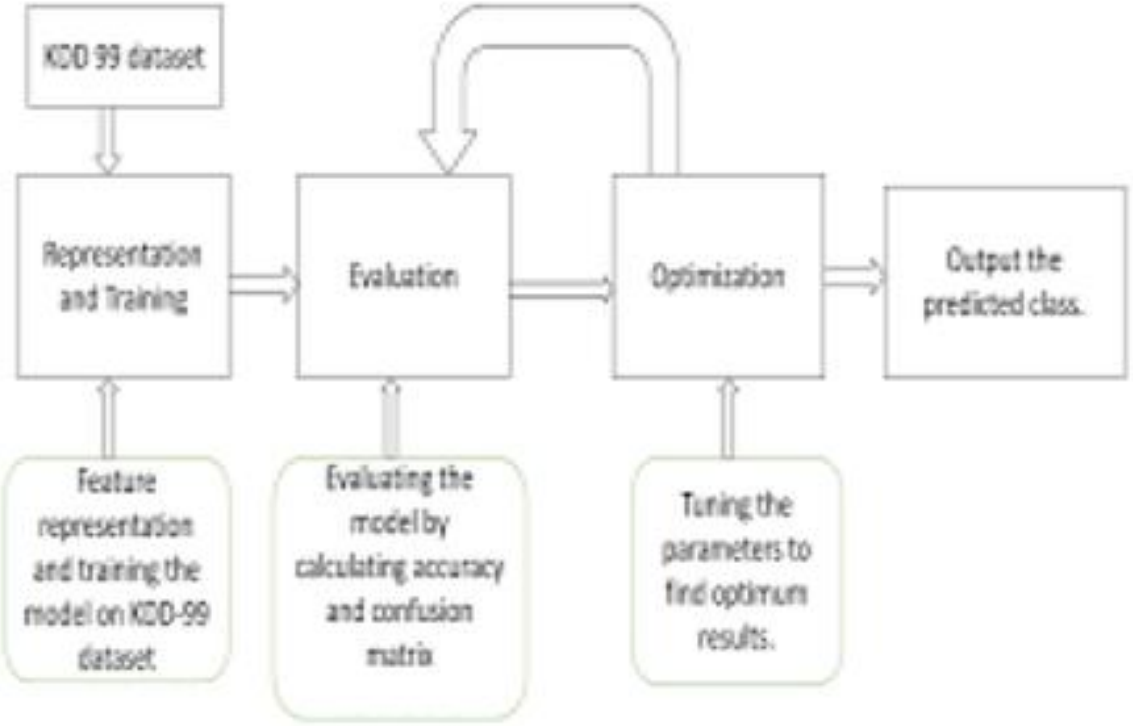

Fig.4:- System Architecture

VII. MATHEMATICAL MODEL

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.
It is always better to represent score between 0 to 1 because the score can now be interpreted as a probability.
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.

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

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

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 7.047 ⁜ Vol. 10, Issue 4, April 2022
DOI: 10.17148/IJIREEICE.2022.10479

VIII. EXPERIMENTAL RESULTS

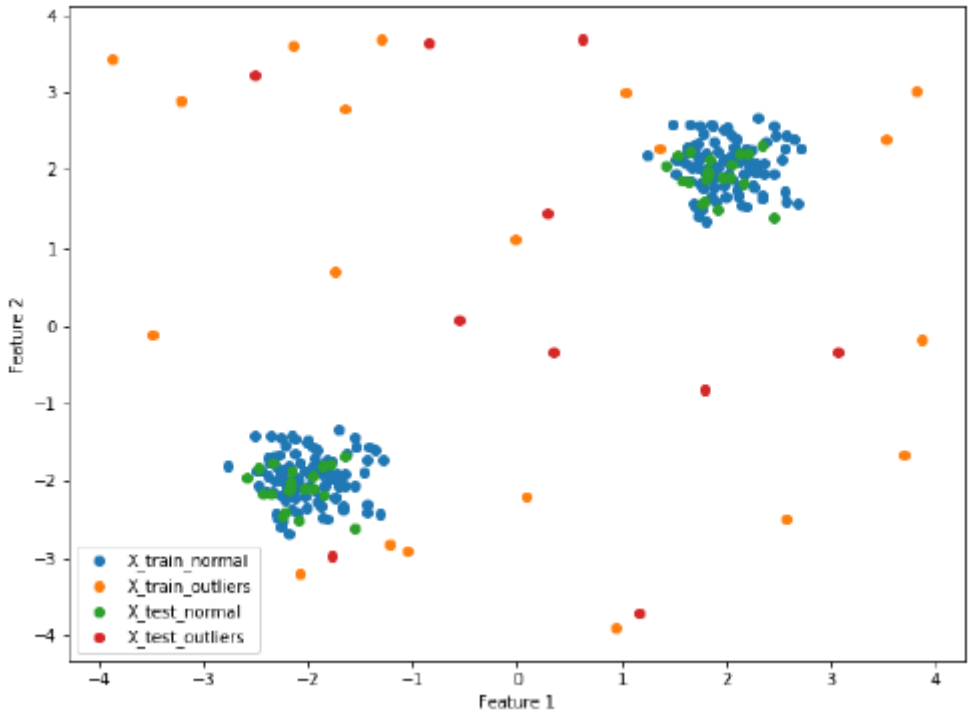

Fig.5:- Plotting Dataset

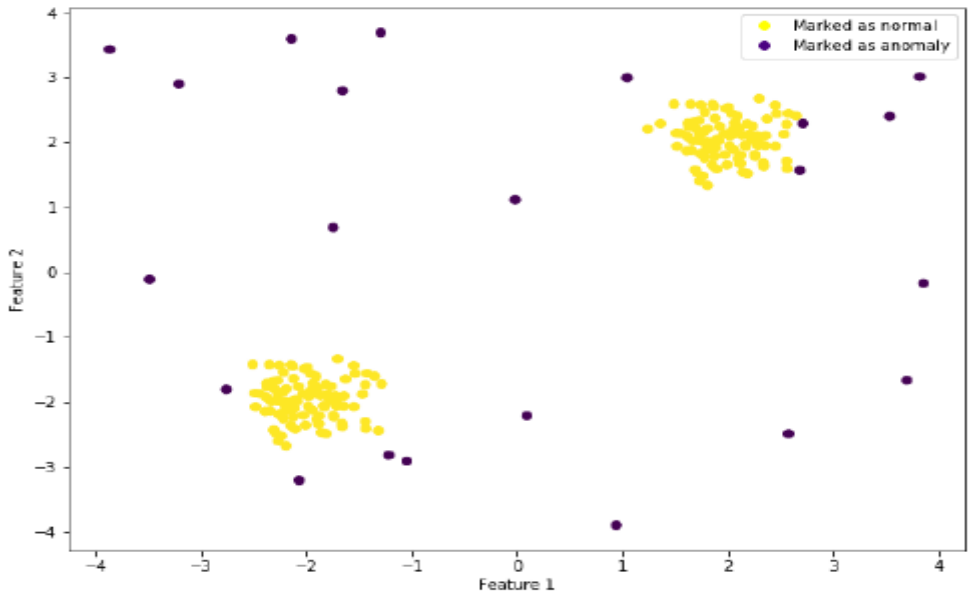

Fig.6:- Visualizing training predictions

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 7.047 ⁜ Vol. 10, Issue 4, April 2022
DOI: 10.17148/IJIREEICE.2022.10479

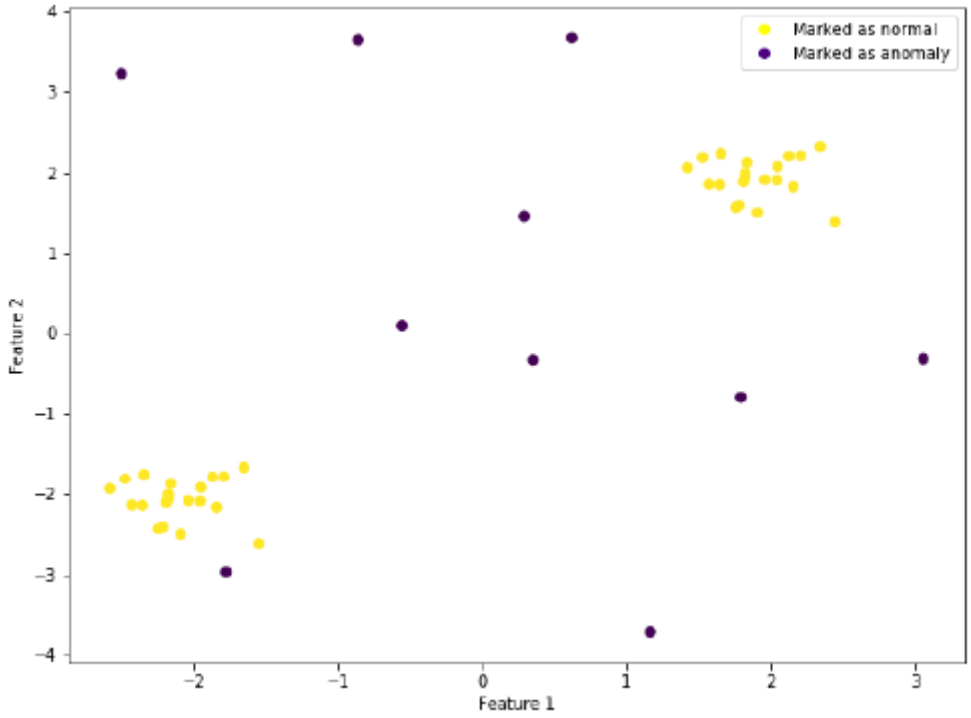

Fig.7:- Visualizing test predictions

IX. ADVANTAGES AND DISADVANTAGES

Advantages:

- 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.

- 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.

- 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.

- Faster Results:

Finding anomalies in data manually can be extremely time-consuming. But, not only do anomalies take a long time to find, anomalies can also take a long time to actually surface using traditional reporting techniques.

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.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 7.047 ⁜ Vol. 10, Issue 4, April 2022

DOI: 10.17148/IJIREEICE.2022.10479

X. CONCLUSION AND FUTURE SCOPE

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 real-time. 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.

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. Due to the growing number of attacks from within an organization, it has become highly important to analyse the behaviour and detect anomaly in real-time with high efficiency. This can be achieved using user and entity behaviour 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.
.

ACKNOWLEDGMENT

The completion of our project brings with it a sense of satisfaction, but it is never complete without them those people who made it possible and whose constant support has crowned our efforts with success. One cannot even imagine our completion of the project without guidance and neither can we succeed without acknowledging it. It is the great pleasure that we acknowledge the enormous assistance and excellent co-operation to us by the respected personalities.

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] M. 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.