Internship Report
on

ML/DL IN INTRUSION DETECTION SYSTEM

Submitted in partial fulfilment of the requirement for the internship of
BTech IV Semester
in
Computer Science and Engineering
By

Aryan Tuteja

(Uni Roll no: 2014373)

&

Harshit Sati

(Uni Roll no: 2015204)

Under the supervision of
Dr Priya Matta
Associate Professor
Department of Computer Science & Engineering,
Graphic Era Deemed to be University

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GRAPHIC ERA DEEMED TO BE UNIVERSITY-248002
Jun-Jul 2021

ACKNOWLEDGEMENT

We are extremely thankful to our honourable President Sir, **Prof. (Dr.) Kamal Ghanshala** of **GRAPHIC ERA DEEMED TO BE UNIVERSITY** for providing all kinds of educational and infrastructural support to work in this project, without which this project would not have been possible.

We hereby like to express our sincere gratitude and respect to **Dr. Priya Matta, Associate Professor, Department of Computer Science & Engineering, Graphic Era Deemed to be University** for her stimulating guidance and continuous supervision, monitoring and constant encouragement throughout the project completion. The blessing, help and guidance given by her from time to time shall go a long way in the journey of life we are about to embark on.

We are obliged to our project team members for the valuable information provided by them in their respective fields. We are grateful for everyone's cooperation during the period of our internship.

Aryan Tuteja
&
Harshit Sati

CERTIFICATE

GraphicEra
Deemed to be University
Accredited by NAAC with Grade A
NBA Accredited Program in CSE, ECE & ME
Approved by AICTE, Ministry of HRD, Govt. of India

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the internship work entitled **ML/DL IN INTRUSION DETECTION SYSTEM** is a bonafide work carried out by **Aryan Tuteja (2014373) & Harshit Sati (2015204)** in partial fulfilment of the requirement for the award of summer internship of BTech IV semester, Computer Science and Engineering, **Graphic Era Deemed to be University**, Dehradun. This work is original work accomplished by the abovementioned interns, and has been approved as it satisfies the academic requirements with respect to the work prescribed for the BTech IV Internship. The time duration for this summer internship was 15 Jun -31 July.

Signature of the Supervisor

Table of Figures

Table of Contents

1. Birth of Artificial Intelligence

Mentions of AI can be found in the old story books filled with myths and legends , but it was only until John McCarthy coined the term "Artificial Intelligence" in the mid-1950s did the research under this field actually boomed and AI became a world wide phenomena due to his astounding contributions in the field and development of the Lisp language.

1.1 Artificial Intelligence

Artificial Intelligence is a computer system able to perform tasks that ordinarily require human intelligence, many of these artificial intelligence systems are powered by a set of rule-based instances or machine learning while some of them are powered by deep learning all while trying to mimic human behavior.

Figure 1. Al-Jazari's programmable automata (1206 CE)

1.1.1 Types of AI

- Reactive Machines
- Limited Memory
- Theory of Mind
- Self-awareness

1.1.2 Applications of AI

- Robotics
- Gaming Industry
- Searching for Exoplanets.

1.2 Machine Learning

It is a branch of Artificial intelligence, that is a Broader notion of building computational artifacts such as applied computational statistics that learn over time based on experience through data and algorithms.

1.2.1 Types of Machine Learning

- Supervised
- Unsupervised
- Reinforcement learning

1.2.2 Applications of Machine Learning

- Predicting House prices in an area
- Recommendation systems
- Predicting Survivors of Titanic

1.3 Deep Learning

Deep learning is an improved machine learning technique for feature extraction, perception and learning of machines. Deep learning algorithms perform their operations using multiple consecutive layers. The layers are interlinked and each layer receives the output of the previous layer as input.

1.3.1 Applications of Deep Learning

- Self Driving Cars
- Object Recognition
- Covid-19 Detection through lung X-Ray images of Patients

1.4 Deep Learning vs Machine Learning

Deep learning requires more data to tune the parameters to reach ideal conclusions whereas Machine learning can work on less amount of data as features are decided by the expert.

Deep learning requires more resources than Machine learning models which can run on ordinary CPU's.

Training of Deep learning algorithms takes more time than training over data through machine learning algorithms.(1)

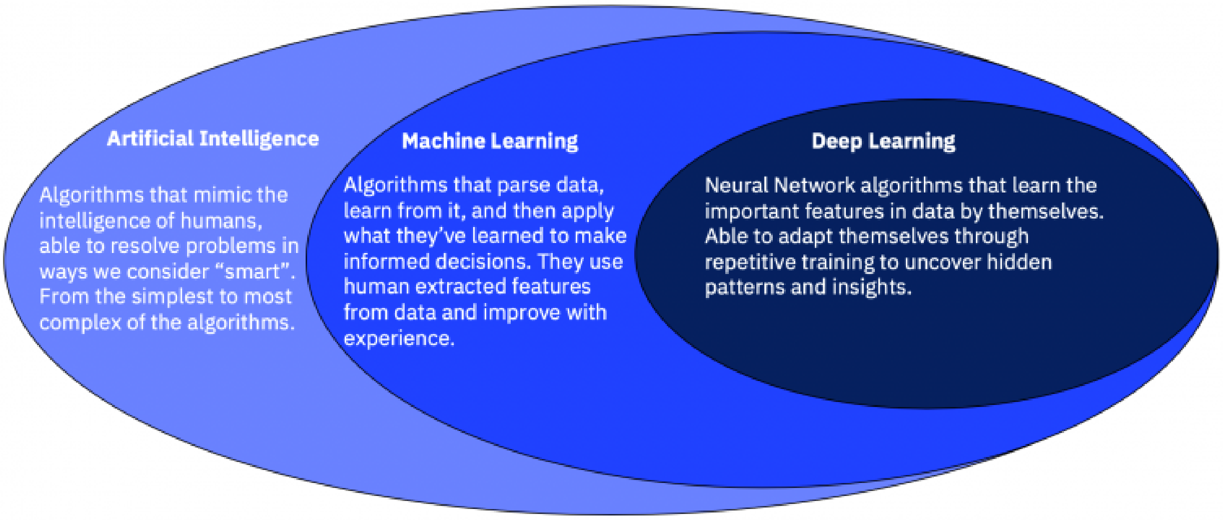

Artificial Intelligence
Algorithms that mimic the intelligence of humans, able to resolve problems in ways we consider "smart". From the simplest to most complex of the algorithms.
Machine Learning
Algorithms that parse data, learn from it, and then apply what they've learned to make informed decisions. They use human extracted features from data and improve with experience.
Deep Learning
Neural Network algorithms that learn the important features in data by themselves. Able to adapt themselves through repetitive training to uncover hidden patterns and insights.

Figure 1.2 Property of IBM

2 Birth of Cyber Security

Cybersecurity's history began with a research project during the 1970s, which was then known as the ARPANET (The Advanced Research Projects Agency Network). A researcher named Bob Thomas created a computer program which was able to move inside ARPANET's network, leaving a small trail wherever it went. He named the program 'CREEPER', because of the printed message that was left when travelling across the network: 'I'M THE CREEPER: CATCH ME IF YOU CAN'.

Ray Tomlinson – the man who invented the world's first email – later designed a program which took CREEPER to the next level, making it self-replicating and the first ever computer worm. Fortunately, he then wrote another program classified as the world's first antivirus called Reaper which chased CREEPER, deleting it from every system.

2.1 Intrusion

A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources and almost always jeopardize the security of networks and/or their data. In order to proactively detect and respond to network intrusions, organizations and their cybersecurity teams need to have a thorough understanding of how network intrusions work and implement network intrusion, detection, and response systems that are designed with attack techniques and cover-up methods in mind.

2.1.1 Types of Intruders

- Outside Intruder: (Masquerader)
- Inside Intruder:(Misfeasor)

2.2 Intrusion Detection System

An Intrusion Detection System (IDS) dynamically monitors the actions of a specific environment, for example, the network traffic, syslog records or system calls of a given operating system, in order to determine if those actions are a legitimate use or a symptom related to a given attack.

2.2.1 Types of Intrusion Detection System

These systems are usually classified into:

Network-based Intrusion Detection Systems (NIDS) : An NIDS works on feature vectors that comprise summarized information related to network traffic within a specified time interval.

Host-based Intrusion Detection Systems (HIDS) : A HIDS is located on a specific host and monitors information related to the system.

2.3 Machine Learning in IDS

It is a branch of Artificial intelligence, that is a Broader notion of building computational artifacts such as applied computational statistics that learn over time based on experience through data and algorithms and is used for intrusion detection.

2.3.1 Types of ML algorithms used in IDS

- Decision Tree
- Random Forest Classifier
- SVM
- k-NN Classifier
- k-Means Clustering

2.4 Deep Learning

Deep learning is an improved machine learning technique for feature extraction, perception and learning of machines. Deep learning algorithms perform their operations using multiple consecutive layers. The layers are interlinked and each layer receives the output of the previous layer as input and is also used to detect intrusion in the system.

2.4.1 Some DL algorithms used in IDS

- Multi-Layer Perceptron
- Recurrent Neural Network (RNN) : LSTM
- Deep belief network (DNB)
- Convolutional Neural Networks (CNN)

2.4.2 Brief explanation about some methods used for IDS

Random Forest Classifier : A Random Forest (RF) consists of a large number of Decision Trees that operate together as an ensemble. Each Decision Tree is a decision tool that works in a tree-like model of decisions and outcomes. For a given dataset entry, each tree of a Random Forest model predicts a given class and the most voted one is elected as the model’s output. The underlying theory behind Random Forests is the wisdom of crowds.

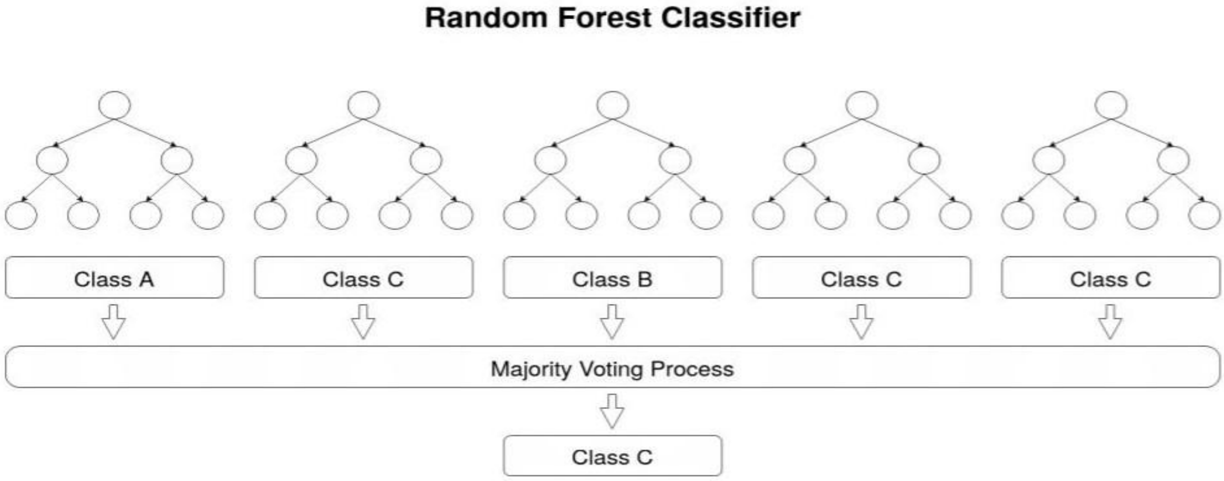


Fig 1.3 A diagram of the working of Random Forest Classifier

Multi-Layer Perceptron : A Multi-Layer Perceptron (MLP) is a type of FeedForward Network which can be represented as an acyclic graph with no feedback connections (the outputs of the model are not fed back into itself). An MLP comprises three or more layers, having one input layer, one or more hidden layers and an output layer, in which each layer has multiple neurons that can be represented in mathematical notation.

Hidden and Output layers activation for MLP Rectified Linear Unit (ReLU) was selected as the activation function of the hidden layers. This activation function has proven to be very computationally efficient and it was one of the main breakthroughs in neural network history for reducing the vanishing and exploding gradient phenomenon. Softmax was used for the output layer since it assigns decimal probabilities to the prediction of each class in a multi-classification problem.

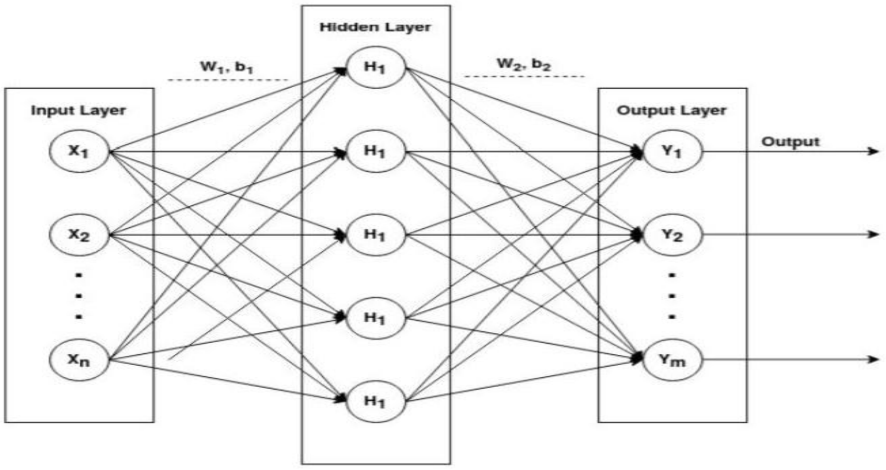


Figure 2. Architecture of a Multi-Layer Perceptron.

Fig 1.4 Diagramatic representation of MLP

Long-Short Term Memory : A Long-Short Term Memory (LSTM) is a type of Recurrent Neural Network (RNN). An RNN contains feedback connections that allow information to travel in a loop from layer to layer. These networks store information about past computations through a hidden state that represents the network memory. Therefore, the output, ot, for a given input, xt, at a given timestep, t, is influenced by the inputs of its previous timesteps, xt−1, xt−2, ..., xt−n, where n defines the total number of prior timesteps. This characteristic allows RNNs to be very suited to working with sequential data.

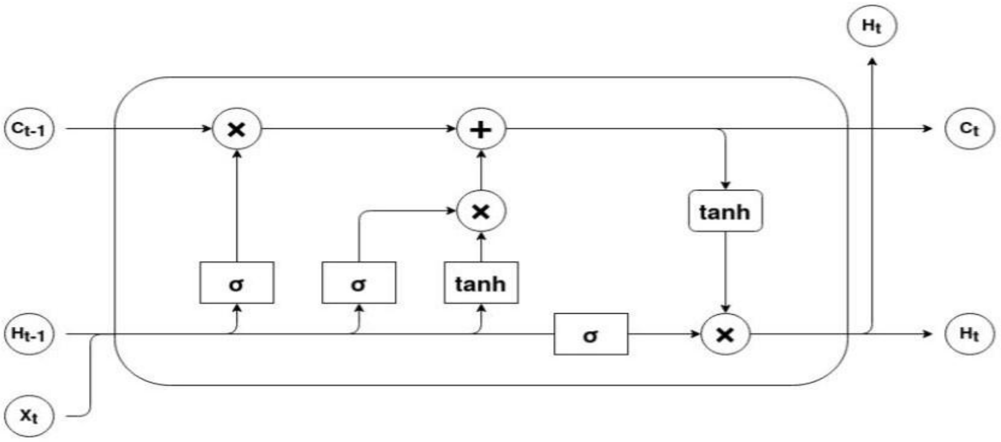


Figure 4. Basic LSTM cell architecture (adapted from [34]).

Fig 1.5 Diagramatic representation of LSTM

3 Analysis of CSE-CIC-IDS2018

With the rise in cyber attacks all over the world it is getting important to find out how and what type of attacks are being launched against an organization.
In this chapter we'll discover the latter i.e. type of attack.

3.1 Dataset

In the CSE-CIC-IDS2018 dataset, we use the notion of profiles to generate datasets in a systematic manner, which will contain detailed descriptions of intrusions and abstract distribution models for applications, protocols, or lower level network entities. These profiles can be used by agents or human operators to generate events on the network. Due to the abstract nature of the generated profiles, we can apply them to a diverse range of network protocols with different topologies. Profiles can be used together to generate a dataset for specific needs. We will build two distinct classes of profiles:

B-profiles: Encapsulate the entity behaviours of users using various machine learning and statistical analysis techniques (such as K-Means, Random Forest, SVM, and J48). The encapsulated features are distributions of packet sizes of a protocol, number of packets per flow, certain patterns in the payload, size of payload, and request time distribution of a protocol. The following protocols will be simulated in our testbed environment: HTTPS, HTTP, SMTP, POP3, IMAP, SSH, and FTP. Based on our initial observations, the majority of traffic is HTTP and HTTPS.

M-Profiles: Attempt to describe an attack scenario in an unambiguous manner. In the simplest case, humans can interpret these profiles and subsequently carry them out. Idealistically, autonomous agents along with compilers would be employed to interpret and execute these scenarios. For attacks we considered six different scenarios [3]

The dataset was from 02-14-2018, it contains 79 columns and 1048575 rows, the target class consisted of three types of attacks namely: Benign, FTP-BruteForce, SSH-Bruteforce.

3.2 Model Building

We started with data exploration where we found out that the dataset contained no null values, it had both categorical and numerical features.

We further observed that the class distribution of targets was imbalanced (as shown in the figure) Benign attacks were the highest reported.

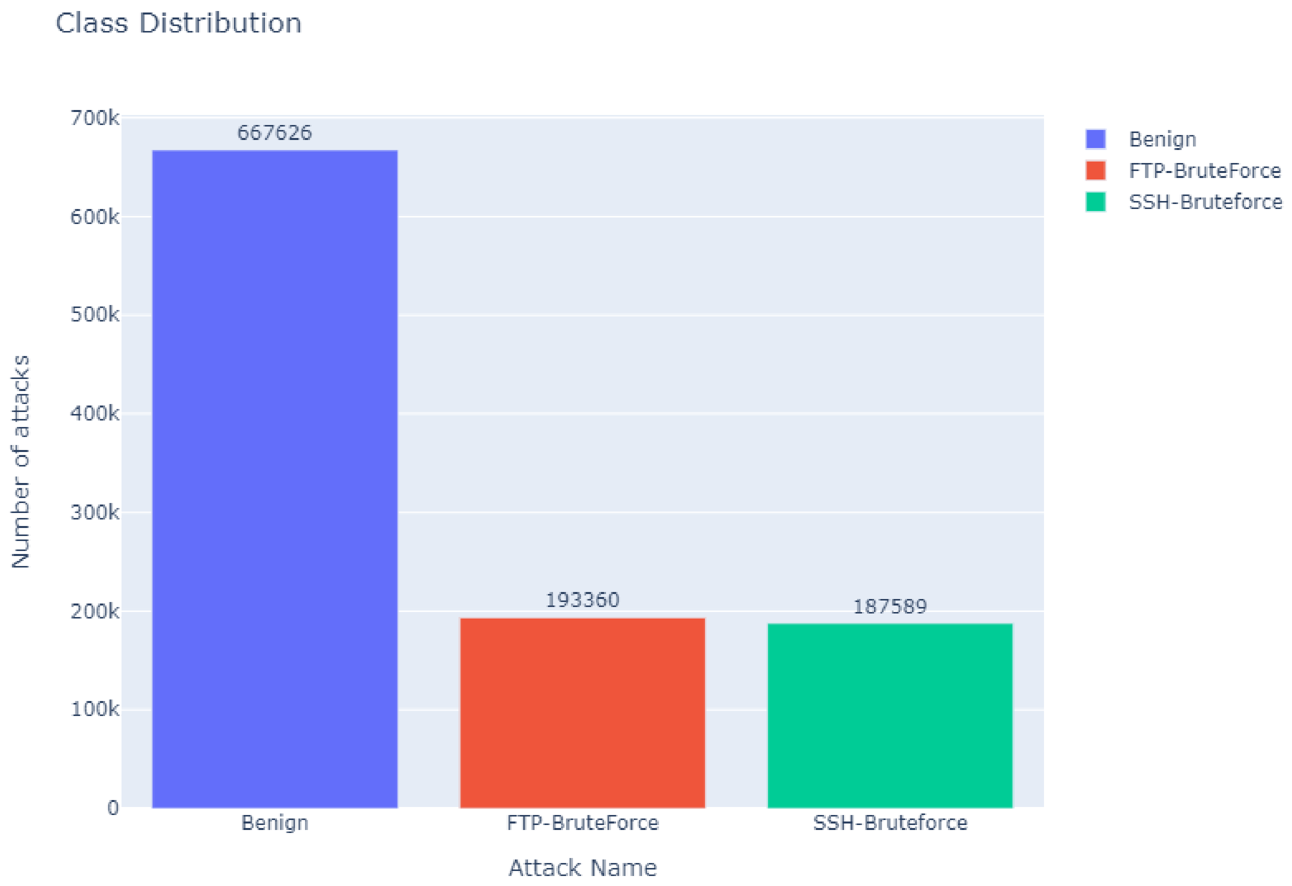


Fig 1.6 Barchart of an imbalanced class

We further noticed that FTP-Bruteforce and SSH-Bruteforce had very high correlation hence we discarded SSH-Bruteforce cases and kept benign and FTP-Bruteforce

We further balanced the class to only have 5000 cases per target type.

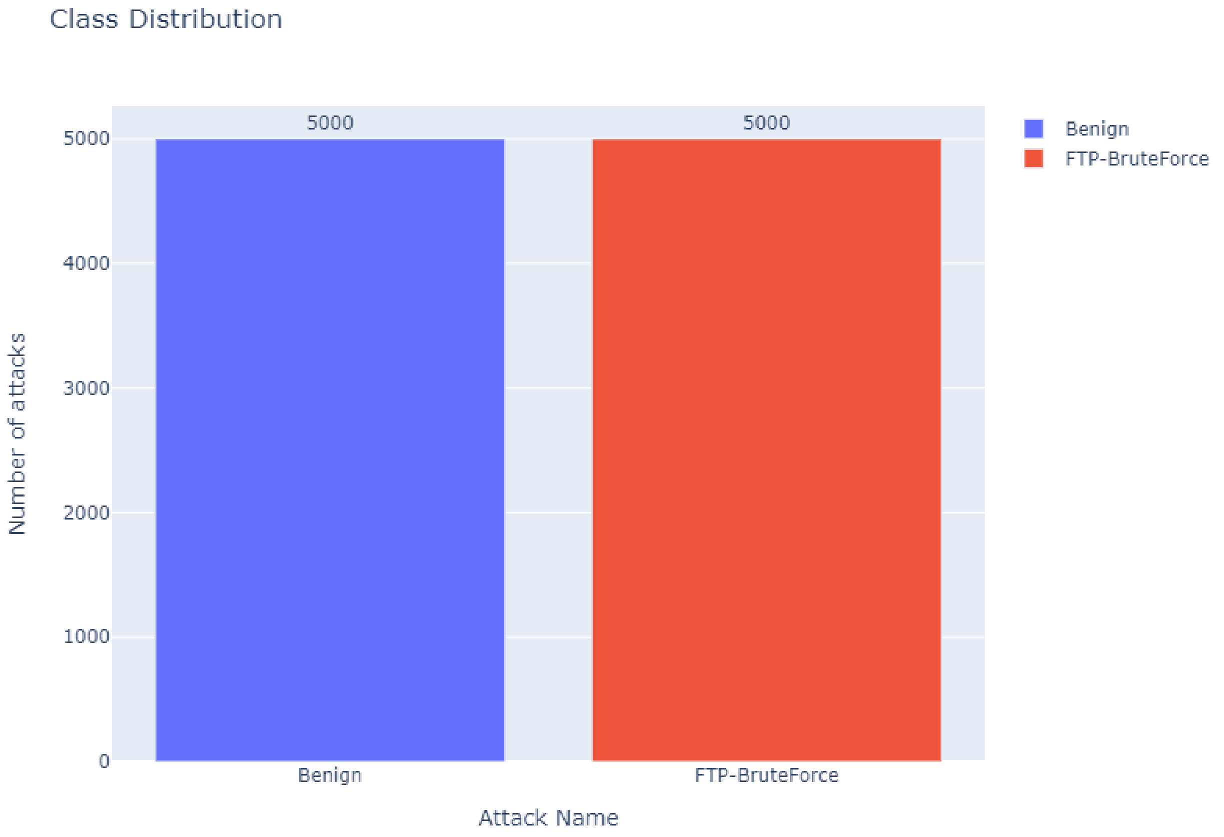


Fig 1.7 Barchart of a balanced class

We then converted the categorical attributes to numerical.

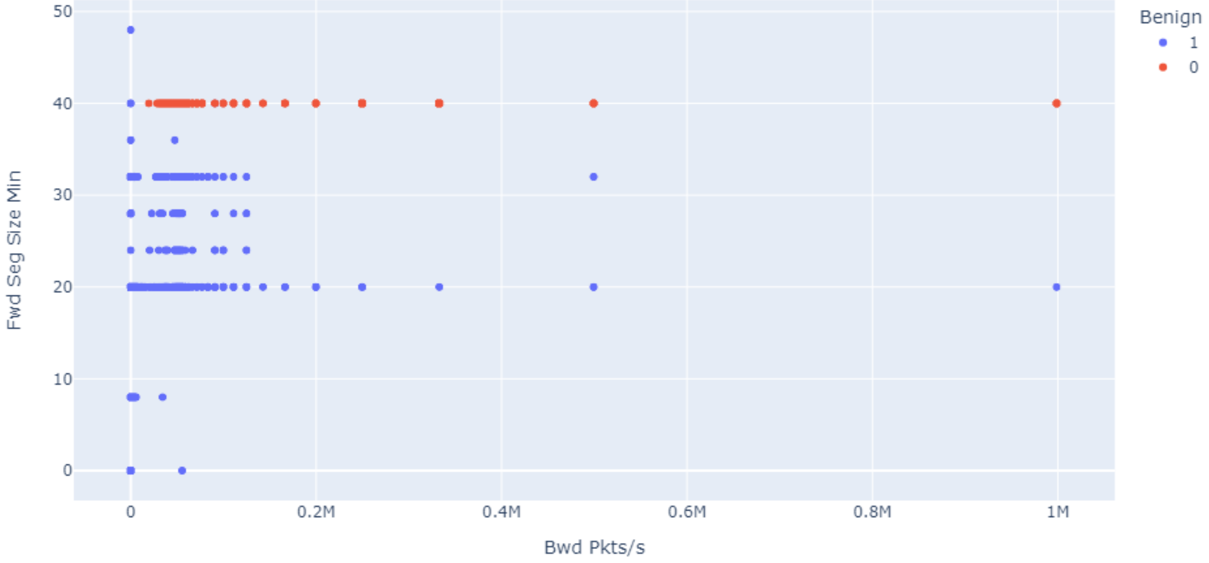


Fig 1.8 Scatterplot of two random features

We decided on using Logistic regression in our model, with standard scaler and onehotencoder as preprocessors.

3.3 Model Training

We dropped one of the target columns (FTP-BruteForce) so as to avoid dummy trap variables and make it easier for the model to fit to the datapoints.

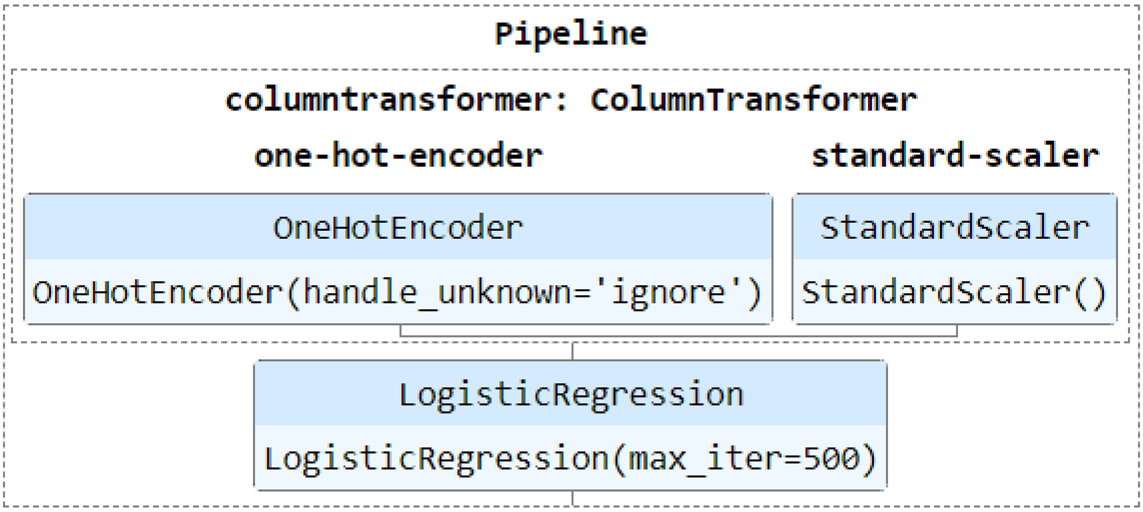

Pipeline
columntransformer: ColumnTransformer
one-hot-encoder
standard-scaler
OneHotEncoder
OneHotEncoder(handle_unknown='ignore')
StandardScaler
StandardScaler()
LogisticRegression
LogisticRegression(max_iter=500)

Fig 1.9 representation of the pipeline used

We found out that some of the columns had no contribution to the target according to pearson's coefficient, hence we dropped those features.

3.4 Model Evaluation

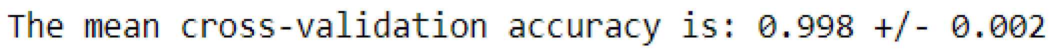
Cross validation is a technique where different folds of a datasets are trained and tested against one another. We used cross validate to find out the best model with the best accuracy.

Conclusion

There’s a thin line between Machine learning and Deep learning algorithms, where deep learning algorithms decide their own features/set of parameters much like unsupervised learning, Machine learning algorithms are fed predefined features.

The deep learning algorithm like the LSTM model proved to be most promising for detecting such patterns in data as compared with other algorithms of deep learning as well as machine learning.

With the model achieving 0.99% accuracy we conclude that we can further move onto Multi class classification and that Logistic regression is a wise choice for the Binary classification in this case.



```
The mean cross-validation accuracy is: 0.998 +/- 0.002
```

References

1) *University of Brunswick, Canada, CSE-CIC-IDS2018 https://www.unb.ca/cic/datasets/ids-2018.html*
2) G. Karatas, O. Demir and O. Koray Sahingoz, "Deep Learning in Intrusion Detection Systems," *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, 2018, pp. 113-116, doi: 10.1109/IBIGDELFT.2018.8625278
3) *Oliveira, N.; Praça, I.; Maia, E.; Sousa, O. Intelligent Cyber Attack Detection and Classification for Network-Based Intrusion Detection Systems. Appl. Sci. 2021, 11, 1674. https://doi.org/10.3390/app11041674*