International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Data Science Based Cyber Security System

Mohammed Naif

Student, Ramaiah University of Applied Sciences, Bangalore, India

ABSTRACT

Cyber-attacks are unauthorised attempts to access computer systems and take, reveal, modify, impair, or destroy data. The state of cyberspace predicts vulnerability for the coming Internet and its growing user base. The vast amount of data collected by device sensors that might be used for targeted attacks raises further concerns with new ideal models. There is a need to consider new models and calculations that rely on information descriptions other than task-explicit procedures because many surviving approaches, models, and algorithms have provided the foundation for cyber-attack prediction. However, its non-direct data handling design can be modified to understand the many network traffic information depictions to characterise the organisation assault. In this study, we describe cyber-attack forecast as a grouping problem, in which networking areas must use machine learning methods to predict the type of network attack from a given dataset. The analysis of a dataset using supervised machine learning (SMLT) to catch some data, such as variable identification, univariate examination, bivariate examination, multivariate examination, missing value, and so on. A near report on machine learning had been finished to determine which calculation is most accurate at foreseeing the types of digital assaults. We divide attacks into four categories: DOS attacks, R2L attacks, U2R attacks, and Probe attacks. The results demonstrate that the suggested machine learning method is viable and can achieve the highest levels of precision in terms of entropy estimates, accuracy, recall, F1 Score, sensitivity, specificity, and entropy.

Keywords: Data Science, Cyber Security, Machine Learning, Attack, Prediction, Dos attack, R2L attack, U2R attack, Probe attack

1. Introduction

This project aims to determine which characteristics are most helpful in predicting business attacks such as DOS, R2L, U2R, Probe, and combinations of attacks as well as to identify general trends that may prove helpful to us in selecting hyperparameters and modelling attacks. Utilise machine learning approaches in order to develop a capacity that can predict the discrete class of fresh information. The archive is a learning exercise designed to: Apply the fundamental principles of machine learning to a dataset that is easily accessible; assess and interpret my results; and validate my interpretation in light of noticed dataset. Create a scratch pad that serves as a computational record, archives my viewpoint, and studies the network connection regardless of whether it is under attack or not in order to look over the data. Identify and investigate quantifiable and speculative outcomes that follow the common pattern for all regiments.

2. Proposed System

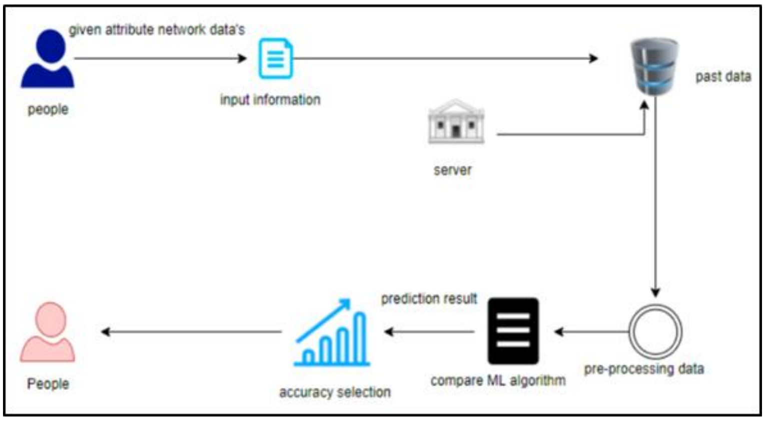
To create a machine learning model for anomaly detection, the proposed model is being put forth. Network interruptions, shady activities, misrepresentation exercises, and other unexpected occurrences that can be of exceptional relevance but are hard to spot can all be detected with anomaly detection. Applying reputable information science techniques, such as variable recognisable proof, which is the dependant and free factors, is how the machine learning model is operated. The information is then seen according to experiences with it at that point. The model is created using historical data from which computations have learned information and developed a variety of calculations that may be used for more accurate tests. Calculations and comparisons are done for the performance metrics.

Benefits:

- Machine learning can automate the process of anomaly detection.
- To improve the model, performance metrics are compared.

3. Architecture

Fig. 1 - Architecture of the System



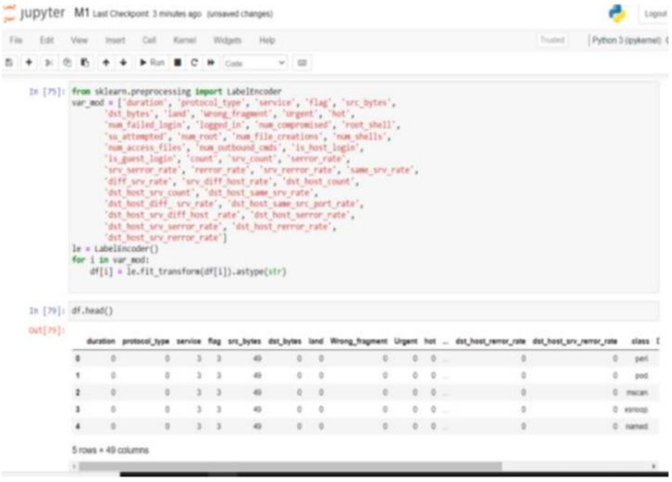
4. Modules

- Data validation process by each attack [Module-01]
- Performance Measurements of DoS attacks [Module-02]
- Performance Measurements of R2L attacks [Module- 03]
- Performance Measurements of U2R attacks [Module- 04]
- Performance Measurements of Probe attacks [Module- 05]
- Performance Measurements of Overall Network attacks [Module-06]
- GUI based prediction results of Network attacks [Module-07]

4.1. Data Validation Process by each attack

Importing the library bundles and stacking the provided dataset. By examining the information's shape, type, and copy values, we may analyse the variable's recognisably proof. An example of data withheld from model preparation is an approval dataset, which is used to measure model competence when adjusting models and systems that you may use to employ approval and test datasets while evaluating your models. Cleaning up and organising information by renaming the provided dataset, deleting sections, and so on to analyse the uni-variate, bi-variate, and multi-variate processes. Depending on the dataset, different ways and techniques will be used to clean the information. The primary goal of information cleaning is to identify and get rid of errors and anomalies in order to increase the value of information for research and independent direction.

Fig. 2 – Data Analysis



```
In [75]: from sklearn.preprocessing import LabelEncoder
         var_mod = ['duration', 'protocol_type', 'service', 'flag', 'src_bytes',
                'dst_bytes', 'land', 'wrong_fragment', 'urgent', 'hot',
                'num_failed_login', 'logged_in', 'num_compromised', 'root_shell',
                'su_attempted', 'num_root', 'num_file_creations', 'num_shells',
                'num_access_files', 'num_outbound_cmds', 'is_host_login',
                'is_guest_login', 'count', 'srv_count', 'serror_rate',
                'srv_serror_rate', 'rerror_rate', 'srv_rerror_rate', 'same_srv_rate',
                'diff_srv_rate', 'srv_diff_host_rate', 'dst_host_count',
                'dst_host_srv_count', 'dst_host_same_srv_rate',
                'dst_host_diff_ srv_rate', 'dst_host_same_src_port_rate',
                'dst_host_srv_diff_host _rate', 'dst_host_serror_rate',
                'dst_host_srv_serror_rate', 'dst_host_rerror_rate',
                'dst_host_srv_rerror_rate']
         le = LabelEncoder()
         for i in var_mod:
             df[i] = le.fit_transform(df[i]).astype(str)

In [79]: df.head()
Out[79]:
   duration protocol_type service flag src_bytes dst_bytes land Wrong_fragment Urgent hot ... dst_host_rerror_rate dst_host_srv_rerror_rate   class
0         0             0       3    3        49         0    0              0      0   0 ...                    0                        0    perl
1         0             0       3    3        49         0    0              0      0   0 ...                    0                        0     pod
2         0             0       3    3        49         0    0              0      0   0 ...                    0                        0   mscan
3         0             0       3    3        49         0    0              0      0   0 ...                    0                        0  xsnoop
4         0             0       3    3        49         0    0              0      0   0 ...                    0                        0   named

5 rows × 49 columns
```

4.2 DOS Attack

A denial-of-service attack (DoS attack) is a type of cyberattack in which the perpetrator tries to render a machine or organisational asset inaccessible to its intended users by momentarily or continuously disrupting management of a host connected to the Internet. Foregoing management is frequently accomplished by overloading the targeted machine or asset with useless requests in an effort to overwhelm frameworks and prevent some or all real requests from being satisfied. A distributed denial-of-service assault (DDoS attack) involves a large number of sources of traffic flooding the victim. This really makes it impossible to halt the attack by obstructing a single source. A DoS or DdoS attack is similar to a crowd barricading a store's section entrance, preventing actual customers from entering and disrupting business.

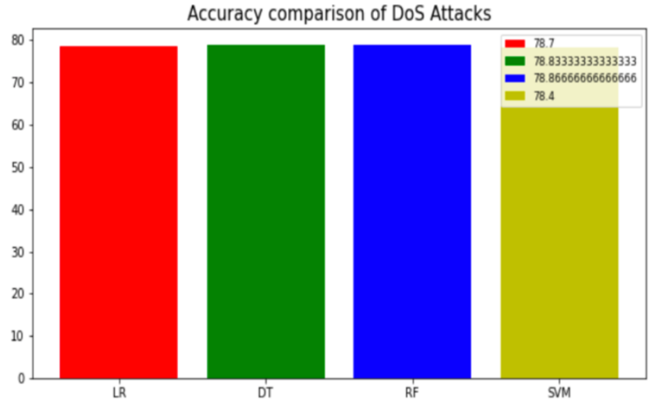
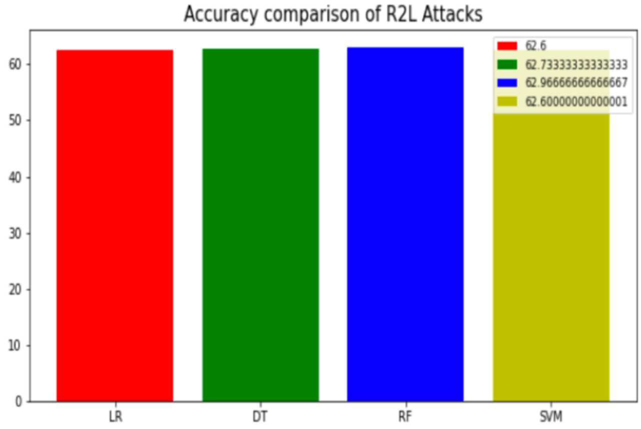


Fig. 3 - Accuracy Comparison of DoS Attacks

4.3 R2L Attack

Today, it is crucial to maintain general security to provide protected and trusted data communication between various associations. But there is always a risk of interruptions and exploitation when receiving information via the internet or from another organisation. Recognition of assaults is a fundamental aspect of managing these risks. Examining, Denial of Service (DoS), and Remote To User (R2U) attacks are a few of the assaults that daily affect a huge number of PCs worldwide. Finding these attacks and protecting PCs from them is a key research area for experts all around the world.

Fig. 4 - Accuracy Comparison of R2L Attacks



4.4 U2R Attack:

An attacker is known to launch a remote to local attack (r2l) in order to gain unauthorised access to a victim machine throughout the entire organisation. User to root attacks (u2r) are similarly often reported when they successfully access a nearby machine while illegally obtaining the root privileges. Of all U2R attacks, buffer overflow is the most well-known. To gain access to a PC asset as a root client, this class begins by sniffing around for passwords on a normal client. Recognising these attacks and protecting PCs against them is an important research topic for analysts worldwide.

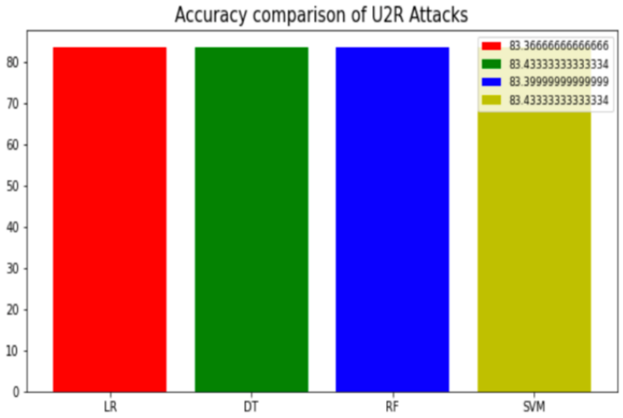


Fig. 5 - Accuracy Comparison of U2R Attacks

4.5 Probe Attack

By observing the real silicon execution of a device, probing attacks are an intrusive method for getting around security measures. In an invasive attack, sensitive data is concentrated by directly accessing a designated device's internal wiring and connections. A probe is an attack that is intentionally designed to be distinguished and reported with a glaring "unique finger impression" in. the report by its aim. The attacker then makes use of the cooperative framework to acquire access to the indicator's protected capabilities and area from this report. Here, the assault aims to compile information on the target machine or organisation in order to describe it. Information about the target could reveal useful information, such as open ports, its IP address, hostname, and working framework. Network Probe is a powerful organisation screen and convention analyzer that can monitor network traffic continuously and help you quickly identify the causes of any organisational bottlenecks.

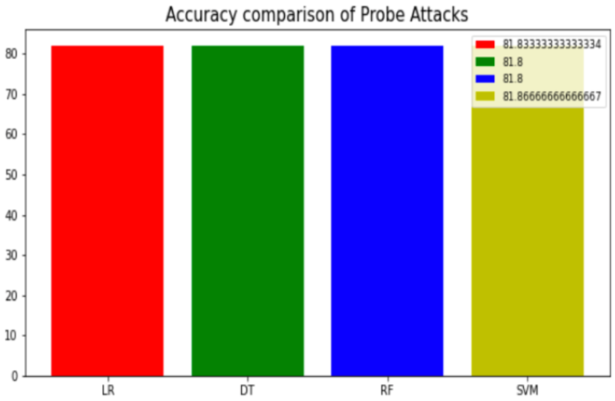


Fig. 6 - Accuracy Comparison of Probe Attacks

4.6 Overall Network attacks

Attacks are carried out in a variety of ways throughout time, making them tougher to recognise. Protectors are expected to associate the various phases of an assault, which may have taken place over a longer period of time, with other assaults that are comparable to it. There are two steps to complex attacks: inquiry and double-dealing. Finding flaws and filtering and testing a framework are all part of the investigation process. A series of phishing attacks followed by exfiltration attacks is an example of a puzzling attack that combines research with double-dealing. In the beginning, attackers would try to obtain information about the organisation they intend to attack, such as the names of important representatives. A phishing attack often involves sending an email that purports to be from a reliable source and tricking the recipient into clicking on a URL that installs malware on the recipient's system. The software then gains a backdoor into the client's system to plan a more complex attack. Both the types of catchphrases used in the email (much like with spam emails) and the characteristics of URLs recalled for the message can be used to identify phishing scams. A crisscross pattern between an anchor and the content of a link, URLs that include IP addresses, and the age of a connected region are features that have been successfully used to identify phishing attacks.

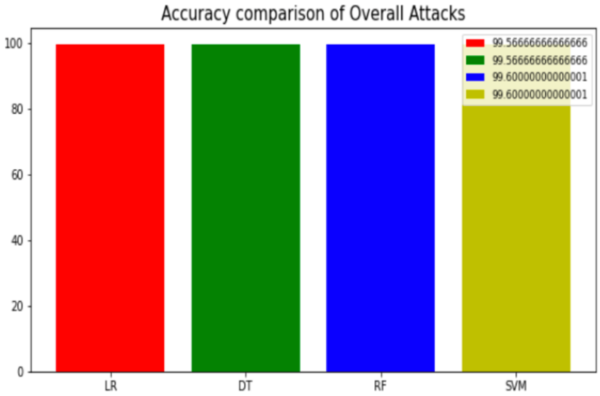


Fig. 6 - Accuracy Comparison of Overall Network Attacks

4.7 GUI based prediction results of Network attacks:

Instead of message-based UIs, composed order names, or message routes, the graphical user interface (GUI) enables users to interact with electronic devices through graphical symbols and aural markers that serve as important documentation. In reaction to the perceived high learning and adaptation requirements of command-line interfaces (CLIs), which demand that orders be composed on a PC console, GUIs were made familiar.

5. Future Enhancements

- To automate the identification of assaults on packet transfers that are time-based in the network sector.
- Automate the procedure by displaying the prediction result in a desktop or web application.
- To streamline the artificial intelligence (AI) implementation process.

6. Conclusion

The scientific exchange started with the management and cleaning of data, followed by missing value and exploratory research, and finally model structure and evaluation. The best precision on the open test set will be determined by comparing each calculation and type of organisational attack for future forecast outcomes by locating the best associations. This adds some of the supporting information for how to assess the organisational assault of each new association. To introduce a forecasting model that uses computerised reasoning as its guide in order to improve on human accuracy and supply with a degree of early detection. This model tends to suggest that area analysis and the use of AI techniques are beneficial in developing prediction models that can help with systems administration areas shorten the time it takes to detect and eliminate any human error.

References

Wentao Zhao, Jianping Yin,"A Prediction Model of DoS Attack's Distribution Discrete Probability,2008

Wenying Xu , Guoqiang Hu ," Distributed Secure Cooperative Control Under Denial-of-Service Attacks From Multiple Adversaries,2019

Seraj Fayyad, Cristoph Meinel, "New Attack Scenario Prediction Methodology",2013

Wentao Zhao, Jianping Yin and Jun Long,"A Prediction Model of DoS Attack's Distribution Discrete Probability",2008.

Jinyu W1, Lihua Yin and Yunchuan Guo," Cyber Attacks Prediction Model Based on Bayesian Network,2012

Xiaoyong Yuan , Pan He, Qile Zhu, and Xiaolin Li," Adversarial Examples: Attacks and Defenses for Deep Learning,2019

Preetish Ranjan, Abhishek Vaish,"Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network",2014