Business Context:

We are in a time where businesses are more digitally advanced than ever, and as technology improves, organizations' security postures must be enhanced as well. Failure to do so could result in a costly data breach, as we've seen happen with many businesses. The cybercrime landscape has evolved, and threat actors are going after any type of organization, so in order to protect your business's data, money and reputation, it is critical that you invest in an advanced security system.

Cyber security can be described as the collective methods, technologies, and processes to help protect the confidentiality, integrity, and availability of computer systems, networks and data, against cyber-attacks or unauthorized access.

a. Information Security vs. Cyber Security vs. Network Security:

Information security (also known as InfoSec) ensures that both physical and digital data is protected from unauthorized access, use, disclosure, disruption, modification, inspection, recording or destruction. Information security differs from cyber security in that InfoSec aims to keep data in any form secure, whereas cyber security protects only digital data.

Cyber security, a subset of information security, is the practice of defending your organization's networks, computers and data from unauthorized digital access, attack or damage by implementing various processes, technologies and practices. With the countless sophisticated threat actors targeting all types of organizations, it is critical that your IT infrastructure is secured at all times to prevent a full-scale attack on your network and risk exposing your company' data and reputation.

Network security, a subset of cyber security, aims to protect any data that is being sent through devices in your network to ensure that the information is not changed or intercepted. The role of network security is to protect the organization's IT infrastructure from all types of cyber threats including:

Viruses, worms and Trojan horses

a. Zero-day attacks
b. Hacker attacks
c. Denial of service attacks
d. Spyware and adware

Your network security team implements the hardware and software necessary to guard your security architecture. With the proper network security in place, your system can detect emerging threats before they infiltrate your network and compromise your data.

There are many components to a network security system that work together to improve your security posture. The most common network security components include:

a. Firewalls
b. Anti-virus software
c. Intrusion detection and prevention systems (IDS/IPS)
d. Virtual private networks (VPN)

Network Intrusions vs. Computer intrusions vs. Cyber Attacks

1. Computer Intrusions:

Computer intrusions occur when someone tries to gain access to any part of your computer system. Computer intruders or hackers typically use automated computer programs when they try to compromise a computer's security. There are several ways an intruder can try to gain access to your computer.

They can Access your

a. Computer to view, change, or delete information on your computer,
b. Crash or slow down your computer
c. Access your private data by examining the files on your system
d. Use your computer to access other computers on the Internet.

2. Network Intrusions:

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 cyber security 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.

Network Intrusion Attack Techniques:

Given the amount of normal activity constantly taking place on digital networks, it can be very difficult to pinpoint anomalies that could indicate a network intrusion has occurred. Below are some of the most common network intrusion attack techniques that organizations should continually look for:

Living Off the Land: Attackers increasingly use existing tools and processes and stolen credentials when compromising networks. These tools like operating system utilities, business productivity software and scripting languages are clearly not malware and have very legitimate usage as well. In fact, in most cases, the vast majority of the usage is business justified, allowing an attacker to blend in.

Multi-Routing: If a network allows for asymmetric routing, attackers will often leverage multiple routes to access the targeted device or network. This allows them to avoid being detected by having a large portion of suspicious packets bypass certain network segments and any relevant network intrusion systems.

Buffer Overwriting: By overwriting certain sections of computer memory on a network device, attackers can replace normal data in those memory locations with a slew of commands that can later be used as part of a network intrusion. This attack technique is a lot harder to accomplish if boundary-checking logic is installed and executable code or malicious strings are identified before they can be written to the buffer.

Covert CGI Scripts: Unfortunately, the Common Gateway Interface (CGI), which allows servers to pass user requests to relevant applications and receive data back to then forward to users, serves as an easy opening for attackers to access network system files. For instance, if networks don't require input verification or scan for backtracking, attackers can use a covert CGI script to add the directory label ".." or the pipe "|" character to any file path name, allowing them to access files that shouldn't be accessible via the Web. Fortunately, CGI is much less popular today and there are far fewer devices that provide this interface.

Protocol-Specific Attacks: Protocols such as ARP, IP, TCP, UDP, ICMP, and various application protocols can inadvertently leave openings for network intrusions. Case in point: Attackers will often impersonate protocols or spoof protocol messages to perform man-in-the-middle attacks and thus access data they wouldn't have access to otherwise, or to crash targeted devices on a network.

Traffic Flooding: By creating traffic loads that are too large for systems to adequately screen, attackers can induce chaos and congestion in network environments, which allows them to execute attacks without ever being detected.

Trojan Horse Malware: As the name suggests, Trojan Horse viruses create network backdoors that give attackers easy access to systems and any available data. Unlike other viruses and worms, Trojans don't reproduce by infecting other files, and they don't self-replicate. Trojans can be introduced from online archives and file repositories, and often originate from peer-to-peer file exchanges.

Worms: One of the easiest and most damaging network intrusion techniques is the common, standalone computer virus, or worm. Often spread through email attachments or instant messaging, worms take up large amounts of network resources, preventing the authorized activity from occurring. Some worms are designed to steal specific kinds of confidential information, such as financial information or any personal data relating to social security numbers, and they then relay that data to attackers waiting outside an organization's network.

Network Intrusion Cover-Up Methods

Once attackers have employed common network intrusion attack techniques, they'll often incorporate additional measures to cover their tracks and avoid detection. As mentioned above, using non-malware and living off the land tools have the dual advantage of being powerful while blending into business justified usage, thus making them hard to detect. In addition, below are three practices that are frequently used to circumvent cyber security teams and network intrusion detection systems:

Deleting logs: By deleting access logs, attackers can make it nearly impossible to determine where and what they've accessed (that is, without enlisting the help of an extensive cyber forensics team). Regularly scheduled log reviews and centralized logging can help combat this problem by preventing attackers from tampering with any type and/or location of logs.

Using encryption on departing data: Encrypting the data that's being stolen from an organization's network environment (or simply cloaking any outbound traffic so it looks normal) is one of the most straightforward tactics attackers can leverage to hide their movements from network-based detections.

Installing rootkits: Rootkits, or software that enables unauthorized users to gain control of a network without ever being detected, are particularly effective in covering attackers' tracks, as they allow attackers to leisurely inspect systems and exploit them over long periods of time.

3. Cyber Attack:

A cyber-attack is any type of offensive action that targets computer information systems, infrastructures, computer networks or personal computer devices, using various methods to steal, alter or destroy data or information systems.

Common cyber-attack types:

a. Denial-of-service (DoS) and distributed denial-of-service (DDoS) attacks
b. Man-in-the-middle (MitM) attack
c. Phishing and spear phishing attacks
d. Drive-by attack
e. Password attack
f. SQL injection attack
g. Cross-site scripting (XSS) attack
h. Eavesdropping attack
i. Birthday attack
j. Malware attack

Business Objective:

With the enormous growth of computer networks usage and the huge increase in the number of applications running on top of it, network security is becoming increasingly more important. All the computer systems suffer from security vulnerabilities which are both technically difficult and economically costly to be solved by the manufacturers. Therefore, the role of Intrusion Detection Systems (IDSs), as special-purpose devices to detect anomalies and attacks in the network, is becoming more important.

The research in the intrusion detection field has been mostly focused on anomaly-based and misuse-based detection techniques for a long time. While misuse-based detection is generally favoured in commercial products due to its predictability and high accuracy, in academic research anomaly detection is typically conceived as a more powerful method due to its theoretical potential for addressing novel attacks.

As part of this project, your task is to build network intrusion detection system to detect anomalies and attacks in the network.

There are two problems:

Binomial classification: Detect anomalies by predicting Activity is normal or attack

Multinomial Classification: Detecting type of activity by predicting Activity is Normal or Back or Buffer Over flow or FTP Write or Guess Password or Neptune or N-Map or Port Sweep or Root Kit or Satan or Smurf

Available Data:

Organization captured the data over the period of time for different types of attacks and provided the data in different files for different type of activities along with normal.

Tables: There are 10 tables for different type of attacks with same columns

a. Data_of_Attack_Back_Normal
b. Data_of_Attack_Back
c. Data_of_Attack_Back_BufferOverflow
d. Data_of_Attack_Back_FTPWrite
e. Data_of_Attack_Back_GuessPassword
f. Data_of_Attack_Back_Neptune
g. Data_of_Attack_Back_NMap
h. Data_of_Attack_Back_PortSweep
i. Data_of_Attack_Back_RootKit
j. Data_of_Attack_Back_Satan
k. Data_of_Attack_Back_Smurf

BASIC FEATURES OF EACH NETWORK CONNECTION VECTOR

1 Duration: Length of time duration of the connection
2 Protocol_type: Protocol used in the connection
3 Service: Destination network service used
4 Flag: Status of the connection – Normal or Error
5 Src_bytes: Number of data bytes transferred from source to destination in single connection
6 Dst_bytes: Number of data bytes transferred from destination to source in single connection
7 Land: if source and destination IP addresses and port numbers are equal then, this variable takes value 1 else 0
8 Wrong_fragment: Total number of wrong fragments in this connection
9 Urgent: Number of urgent packets in this connection. Urgent packets are packets with the urgent bit activated

CONTENT RELATED FEATURES OF EACH NETWORK CONNECTION VECTOR

10 Hot: Number of „hot" indicators in the content such as: entering a system directory, creating programs and executing programs
11 Num_failed _logins: Count of failed login attempts
12 Logged_in Login Status: 1 if successfully logged in; 0 otherwise
13 Num_compromised: Number of ``compromised'' conditions
14 Root_shell: 1 if root shell is obtained; 0 otherwise
15 Su_attempted: 1 if ``su root'' command attempted orused; 0 otherwise
16 Num_root: Number of ``root'' accesses or numberof operations performed as a root in the connection
17 Num_file_creations: Number of file creation operations in the connection
18 Num_shells: Number of shell prompts
19 Num_access_files: Number of operations on access control files
20 Num_outbound_cmds: Number of outbound commands in an ftp session
21 Is_hot_login: 1 if the login belongs to the ``hot'' list i.e., root or admin; else 0
22 Is_guest_login: 1 if the login is a ``guest'' login; 0 otherwise

TIME RELATED TRAFFIC FEATURES OF EACH NETWORKCONNECTION VECTOR

23 Count: Number of connections to the same destination host as the current connection in the

past two seconds

24 Srv_count: Number of connections to the same service (port number) as the current connection in the past two seconds

25 Serror_rate: The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in count (23)

26 Srv_serror_rate: The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in srv_count (24)

27 Rerror_rate: The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in count (23)

28 Srv_rerror_rate: The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in srv_count (24)

29 Same_srv_rate: The percentage of connections that were to the same service, among the connections aggregated in count (23)

30 Diff_srv_rate: The percentage of connections that were to different services, among the connections aggregated in count (23)

31 Srv_diff_host_ rate: The percentage of connections that were to different destination machines among the connections aggregated in srv_count (24)

HOST BASED TRAFFIC FEATURES IN A NETWORK CONNECTION VECTOR

32 Dst_host_count: Number of connections having the same destination host IP address

33 Dst_host_srv_ count: Number of connections having the same port number

34 Dst_host_same _srv_rate: The percentage of connections that were to the same service, among the connections aggregated in dst_host_count (32)

35 Dst_host_diff_ srv_rate: The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)

36 Dst_host_same _src_port_rate: The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count (33)

37 Dst_host_srv_ diff_host_rate: The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33)

38 Dst_host_serro r_rate: The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_count (32)

39 Dst_host_srv_s error_rate: The percent of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_srv_count (33)

40 Dst_host_rerror_rate: The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_count (32)

41 Dst_host_srv_rerror_rate: The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_srv_count (33)

Type Features:

Nominal: Protocol_type(2), Service(3), Flag(4)

Binary: Land(7), logged_in(12), root_shell(14), su_attempted(15), is_host_login(21), is_guest_login(22)

Numeric: Duration(1), src_bytes(5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23), srv_count(24), error_rate(25), srv_serror_rate(26), rerror_rate(27),srv_rerror_rate(28), same_srv_rate(29), diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

Hints about Data: Different attack data set have different number of observations. This data is an example of imbalance data.

Data Preparation:

You are required to append all the files and create new column called attack based on the name of attack. While you are appending the files, you can take resampling of data based on the number of attacks.

For Binomial classification, you can create attack variable with attack vs. normal

For Multinomial classification, you can create attack variable with normal vs**.** Back **vs.** Buffer Over flow **vs.** FTP Write **vs.** Guess Password **vs.** Neptune **vs.** N-Map **vs.** Port Sweep **vs.** Root Kit **vs.** Satan **vs.** Smurf