Cyber Threats Prediction Using Machine Learning

Chetan Pathade[1], Tanvi Bhosale[2]

[1,2]*Research Scholar, Dept. of Computer Engineering, Savitribai Phule Pune University, Maharashtra, India*

Abstract - *Cyber Threats damage computer systems and the network with or without user consent; hence prediction of the cyber threats is very crucial in these scenarios. We know that all computers are connected through various networks, so predicting cyber threats will be very helpful to prevent future data loss or disaster. Prediction is one of the approaches from which we can know the output based on the input which is provided. There is an existing approach in which the model is built on certain algorithms and that model is trained with a certain dataset. Based on the model training the model should predict the outcome of the given input. These predictions are done using Machine Learning algorithms; which will help to predict better results from the perspective of cyber threats. We have explored the work done by various researchers on cyber threat predictions and in addition to which we will be presenting our work. For this, we will use different methodologies which will help get better results for the prediction of cyber threats. As a result, it will be very helpful to get prior information about the cyber threats from the past learning experience of the model. And thus, easily prevent data loss from these cyber threats.*

Key Words: **Cyber Threats; Machine Learning; Cyber Security; Predictions; Algorithms; Python; Anaconda.**

1.INTRODUCTION

1.1 Overview

We all know that in today's world, the internet and computers are a very necessary and important part of life. Due to the wide range of internet availability, cyber threats are increasing day by day. So we are finding the solution for it by predicting the cyber threats before they take place. If we can predict the threats of the system or of the applications which are developed, then the prediction model will give us a proper prediction of whether the existing systems have threats against any cyber-attacks or not.

So, in this paper, we will be covering:

- What models can we use to predict cyber threats?
- How should the model be trained with the help of the clean dataset?
- What are the existing approaches for the prediction of threats?
- What dataset should be ideal for the prediction?
- What tools/languages can we use for prediction?
- Which methodology will be relevant among the existing ones? Why?
- How will it improve the prediction results compared to older ones? (Outcomes)

Primarily, let's see the basic concept of Cyber Threats and their types.

1.2 Cyber Threats

In the world of computer security, the threat is the potential vulnerability that results in a harmful impact on the computer systems or applications. It can occur due to either intentional or accidental events. When we consider the intentional events they can be referred to as individual attackers or criminal organizations.

On the other hand, accidental events come under the possibility of the computer malfunctioning or a natural disaster such as fire, earthquake, tornado etc. According to National Information Assurance Glossary (NIAG), the threat is defined as any event or circumstance with the potential to have a huge impact on the system or the infrastructure through sensitive data disclosure, unauthorised access, modification of information and the denial of service (DoS).

The main pillar of security is the CIA triad i.e Confidentiality, Integrity and Availability. On these three pillars, security is defined. When any one of these pillars gets affected in terms of impact then there is a high threat possibility in that particular system or application.

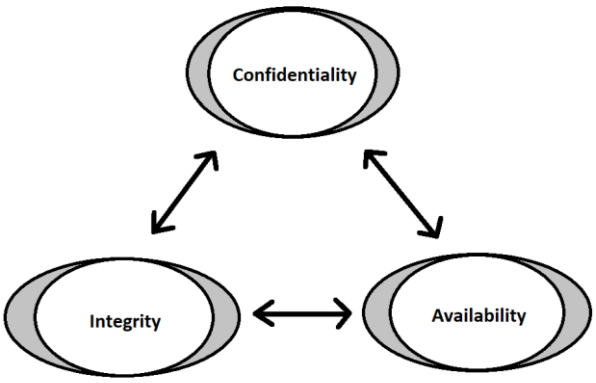

Fig.1.1 CIA Triad

1.3 Types of cyber threats

There are mainly 9 types of cyber threats as follows:

A) Malware:

Malware refers to malicious software which is generally a file or code which is spread over the network and infects, steals any sensitive information or conducts any behaviour that the attacker wants virtually or physically. There are various types of malware shown in the figure.

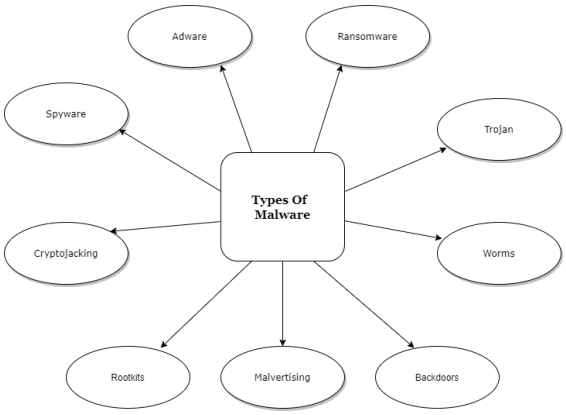

Fig.1.2 Types of Malware

B) Phishing:

It is a category of social engineering attack and it's frequently used to steal credit card information, login credentials or user data. Example: Email Phishing, Spear Phishing, Clone Phishing, Voice Phishing, SMS Phishing etc.

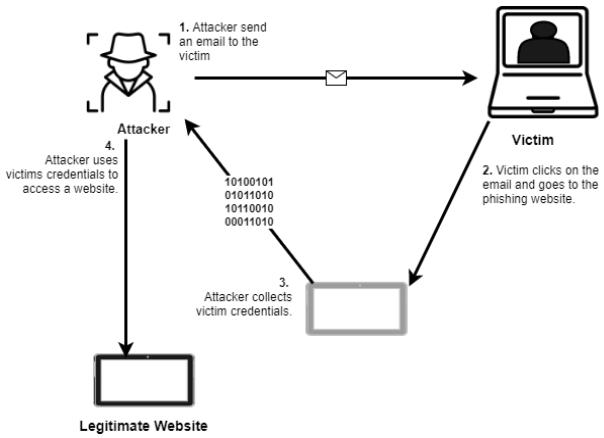

Fig.1.3 Phishing

C) Man in the Middle Attack:

It is also referred to as monster-in-the-middle, person-in-the-middle, machine-in-the-middle, monkey-in-the-middle. In this, the attacker can alter the communication between two parties who believe that they are directly communicating with each other. If the message is not encrypted then the attacker can modify the message or get the credentials that are sent over HTTP instead of HTTPS.

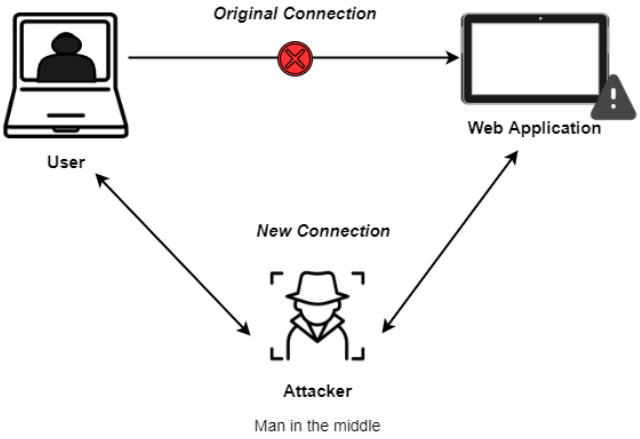

Fig.1.4 MITM (Man-in-the-middle)

D) Denial of Service Attack:

It makes the network or machine unavailable to its genuine users. In this, the attacker sends a large number of requests to the server which leads to the crashing of the server so that the genuine user can not access it.

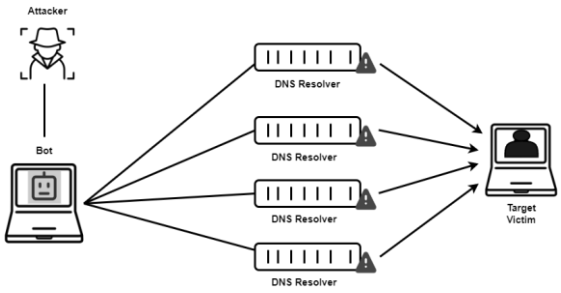

Fig.1.5 DOS (Denial of Service)

E) SQL injection:

It is a web security threat in which it permits the attacker to get involved with SQL queries that the numerous applications make to the database. From this threat, the attacker is able to view or retrieve the data from the database.

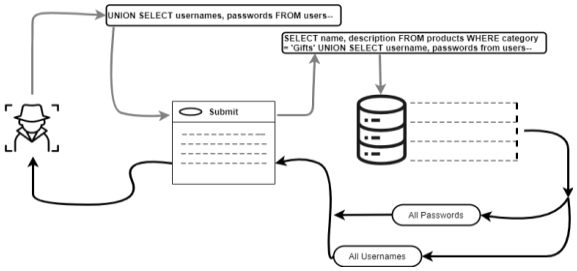

Fig.1.6 SQL Injection

F) Zero-day Exploit:

This is the threat that is unknown and the patch against it is not developed yet. Until the mitigation is developed hackers can exploit it and can take the advantage of this threat.

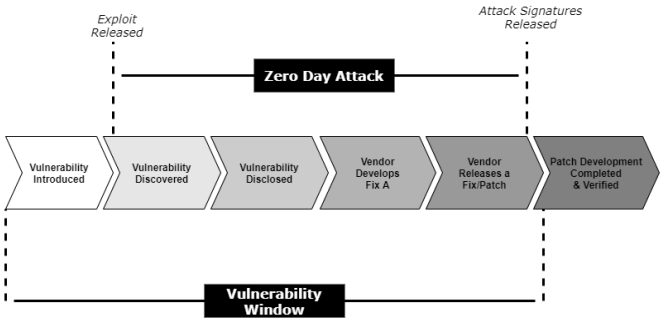

Fig.1.7 Zero-day Exploit Cycle

G) Advanced Persistent Threat:

This is done by groups such as nation-state or state-sponsored for gaining unauthorised access to networks or systems. Also, they remain undetected for an extended period.

H) Ransomware:

Ransomware is a type of malware that encrypts a user's personal data. It blocks the access until the ransom is not paid. The attacker asks for the money for the decryption of the data.

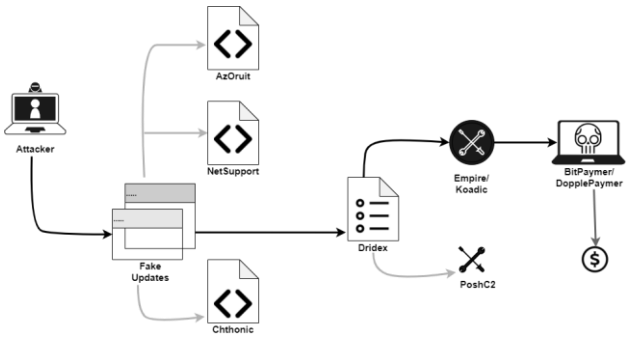

Fig.1.8 Ransomware

I) DNS Attack:

DNS means Domain Name System which resolves domain names into the specific IP address which is associated with it. In this threat, the attacker takes advantage of a vulnerable domain name system (DNS). There are various types of DNS threats such as domain hijacking, DNS flood attack, DNS tunnelling etc.

2. Related Work

Dong-Jie Wu, Ching-Hao Mao, Te-En Wei, Hahn-Ming Lee, Kuo-Ping Wu, Presents the understanding of the detection of android malware through API call tracing and manifest. In the outcome of their approach, the recall rate turned out to be better than tools like Androguard that was published in Blackhat 2011. This tool (DroidMat) focused on the analysis of android malware. They use the K-mean clustering algorithm in this and the number of clusters is decided by SVD (Singular Value Decomposition) method. According to their research, DroidMat was 2x time-efficient than Androguard in terms of time. [1]

Jong-Hyun Kim, Sun-Hee Lim, Seunghwan Yun, Underlines the model which predicts botnet-based cyber threats. In their prediction model, primarily they used Botsniffer and BotMiner for the detection of botnets. Furthermore, they defined the prediction model for the threats estimation. In conclusion, they monitor communication with the C&C server, botnets and zombies and measure the possible domain's threats. [2]

A.M.S.N Amarasinghe, W.A.C.H Wijesinghe, D.L.A Nirmana, Anuradha Jayakody, A.M.S Priyankara, Clarifies the work on cyber threats and vulnerability detection, prevention and prediction systems based on artificial intelligence (AI). They have divided their work into three phases such as

detection, prevention and prediction. On the basis of a rich database, they evaluate the results and the final prediction is done by the logistic regression. [3]

Tahia Infantes Morris, Liam M. Mayron, Wayne B. Smith, Margaret M. Knepper, Reg Ita and Kevin L. Fox, Present the ontology-driven framework that consists of a dynamic knowledge base. The results were a dynamic capability for the management of cyber missions that provides on-demand cyber information to professionals, policymakers and analysts. [4]

Hafiz M.Farooq, Naif M. Otaibi, Takes a shot at cyber threat prediction using the optimal machine learning algorithms. While using various prediction, classification and forecasting algorithms they proposed machine learning algorithms that are optimal based on analytical and empirical evaluations. They used Decision Trees, Ensemble Learning, Deep Learning, Classification and Regression etc. [5]

Adam Dalton, Bonnie Dorr, Leon Liang, Kristy Hollingshead, In this work demonstrate how to improve the Cyber-Attack prediction through information foraging. The accuracy of cyber-focused forecasting systems can be improved by information foraging. In this research, they described a framework for Information Foraging for Algorithm Discovery (IFAD). The results describe that cognitive augmentation and information foraging, which is useful in the development of tools to anticipate cyber threats. [6]

Vishal Mehta, Pushpendra Bahadur, Manik Kapoor, Dr. Preeti Singh, Dr. Subhadra Rajpoot, Proposed the research about using the honeypot and Machine Learning how they did the threat prediction. Using Honeypot as the source of the data and different machine learning algorithms the architecture of the frameworks predicts the threats. This prediction is very precise using OSSEC as Host Intrusion Detection System (HIDS) and SNORT. [7]

Kunal Rashmikant Dalal, Mayur Rele, Clarifies the work on how the threat detection model is based on the Machine Learning Algorithm. In this research, the prediction of malware must be executed in the sandbox environment based on machine learning algorithms. Machine learning methods along with the sandbox technique are dominant tools that prevent cyber attacks. This method is proven to be efficient and autonomous in advanced threat detection systems. [8]

3. Objective of Present Work

The main objective of this work is to determine how the present work is going on and how we can improve that, using the latest technologies. So that, it will help to get accurate and precise results to predict the correct cyber threat using advanced machine learning techniques. Our work will help to get better results efficiently and accurately to predict various cyber threats.

4. Methodology

When we consider the prediction techniques it really depends on the dataset and its size. We can say that the bigger your data the better your model gets trained and it will give more accurate results. For prediction there are various methods as follows:

A) Multiple Linear Regression:

To predict the outcome of a responsive variable Multiple Linear Regression (MLR) uses several explanatory variables. It is an extension of Linear Regression that uses only one explanatory variable. The formula for Multiple Linear Regression:

$$yi = \beta 0 + \beta 1xi1 + \beta 2xi2 + ... + \beta pxip + \epsilon$$

Where,

i) yi=dependent variable

ii) xi=explanatory variables

iii) β0=y-intercept (constant term)

iv) βp=slope coefficients for each explanatory variable

v) ϵ=the model's error term (also known as the residuals)

(for i = n observations)

B) k-Nearest Neighbors (kNN):

kNN is also known as the non-parametric method, the association between an independent variable and the continuous outcome by averaging the observation in the same neighbourhood. The size of the neighbourhood can be set by two methods:

i. Set by analyst

ii. Can be chosen using cross-validation

C) Regression Tree:

Binary recursive partitioning is the base on which a regression tree is built. This is an iterative process and splits data into branches. This process goes on until it does not resolve the parts into further branches.

D) CART Algorithm:

The Classification and Regression Tree algorithm works on the Gini impurity index by splitting the training set into

two subsets using a threshold value(tk) and a single feature(k). The algorithm searches for a pair (k, tk) that produces the purest subset and continuous splitting until it reaches maximum depth. Hence, explaining the working of how target variables are predicted on the basis of other variables. CART algorithm tries to reduce the cost function to minimal which is formulated as,

J (k, tk) = (mleft /m) Gleft + (mright /m) Gright

Where,

i) Gleft =measure the impurity of the left subset

ii) Gright =measure the impurity of the right subset

iii) mleft =number of instances in left subset

iv) mright =number of instances in right subset

E) K-means Clustering:

K-means algorithm iteratively keeps forming the clusters internally from the defined data, in such a way that the sum of the euclidean distance between the data points and the centroid is the minimum. To get the optimal output it is necessary to choose the number of clusters for the algorithm. One of the most efficient ways of doing this is using the Elbow method, in which a curve is plot WCSS (Within Cluster Sum of Squares) vs the number of clusters. And the value of cluster number at the sharpest bend is considered to give the best result.

F) Logistic Regression:

Logistic regression algorithm considers the dependent variable which can be binary or multinomial to determine the probability of the likelihood of the event happening. Logistic regression algorithm uses Sigmoid as the activation function to convert the output into the categorical format.

This algorithm can be formulated as,

y = b0 + b1x , threshold (t)

and,

y = 1 when y>=t

y = 0 when y<t

log(p / 1-p) = y, where p gives us the probability of success.

5. Tools/Libraries

A) Anaconda -

Anaconda is an open-source platform available for professionals and beginners to code and implement programs in languages like Python, R, etc. and simplify package management and deployment. Provides various tools for domains like Data Science, Machine learning and Deep learning.

B) Jupyter Notebook -

Jupyter Notebook is an open-source web application that allows creating and sharing documents integrated with live coding, computational output, visualizations, and explanatory text document.

C) Google Colab Notebook -

Google Colab notebooks is a free cloud notebook platform integrated with google drive, providing writing and executing codes on browsers. Also, giving free access to computing resources like GPUs and TPUs.

D) Google Colab Notebook -

Google Colab notebooks is a free cloud notebook platform integrated with google drive, providing writing and executing codes on browsers. Also, giving free access to computing resources like GPUs and TPUs.

E) Pandas -

Pandas is a Python library that enables data manipulation by converting any extension of the database into a tabular format.

F) Scikit-Learn -

It is the most important library provided by the Python programming language, which helps implement various machine learning algorithms, various classification and regression algorithms.

G) Matplotlib -

It is an interactive visualization library of the Python programming language.

6. Noteworthy Improvements

For prediction, we use various algorithms but when it comes to accuracy then we have to choose the correct and precise algorithm. For this, either we have to check one by one implementation of all algorithms and check which one is giving better results or we have to use some advanced methods which will help to get desired results. This

process is called Model selection and can be done by various evaluating factors like:

1) Resampling Method - In this method, the rearranging of data samples is carried out to ensure the model is generalized well. A few ways of doing it are:

a) Random split

b) Time-based split

c) K fold cross-validation

d) Stratified K fold

e) Bootstrap

2) Probabilistic measure - It measures the model performance and model complexity.

The above option is very time consuming and it will require various factors for comparison and then we have to choose one according to our requirements. But there are several advanced methods that we can use in our existing algorithm and we can get our desired outcomes, such as :

A) Bagging -

Bagging is known as the Bootstrap Aggregation, an ensemble learning method used to reduce the variance in the noise data. It helps to improve the performance and accuracy of the algorithms. Basically, works towards avoiding overfitting of data and bootstrapping are the method of randomly creating data samples out of the dataset by replacement. Bagging deals very efficiently with high dimensional data.

B) Boosting -

Boosting is an ensemble learning method that follows iteratively running the model to identify the weak predictions and combine them eventually into a single strong prediction rule. Famous techniques of Boosting are:

i) AdaBoost (Adaptive Boosting)

ii) Gradient Boosting

iii) XGBoost

7. CONCLUSIONS

In this overview, we surveyed that if we use advanced algorithms or methods we can get more accurate results. If we include the resampling method or probabilistic measure method in our existing algorithm then model performance increases with high prediction results. Keep note that the dataset which you are using for the prediction should be organised, sanitized and it should be minimal bias, it will help the model to perform well and predict accurate results so that we can mitigate the cyber threats. It will definitely help to reduce cyber threats well in advance prior to the actual attack so that we can mitigate the risk associated with it. To keep the organization and other infrastructure safe, this method and improvements will definitely help to reduce and mitigate the risk.

REFERENCES

[1] Dong-Jie Wu, Ching-Hao Mao, Te-En Wei, Hahn-Ming Lee and Kuo-Ping Wu, “DroidMat: Android Malware Detection through Manifest and API Calls Tracing”, 2012 Seventh Asia Joint Conference on Information Security.

[2] Sun-Hee Lim, Seunghwan Yun, Jong-Hyun Kim and Byung-gil Lee, “Prediction Model for Botnet-based Cyber Threats”.

[3] A.M.S.N. Amarasinghe, W.A.C.H. Wijesinghe, D.L.A. Nirmana, Anuradha Jayakody and A.M.S. Priyankara, “AI Based Cyber Threats and Vulnerability Detection, Prevention and Prediction System”, 2019 International Conference on Advancements in Computing (ICAC) December 5-6, 2019. Malabe, Sri Lanka.

[4] Tahia Infantes Morris, Liam M. Mayron, Wayne B. Smith, Margaret M. Knepper, Reg Ita, and Kevin L. Fox, “A perceptually-relevant model-based cyber threat prediction method for enterprise mission assurance”, 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), Miami Beach, FL.

[5] Hafiz M. Farooq and Naif M. Otaibi, “Optimal Machine Learning Algorithms for Cyber Threat Detection”, 2018 UKSim-AMSS 20th International Conference on Modelling & Simulation.

[6] Adam Dalton, Bonnie Dorr, Leon Liang and Kristy Hollingshead, “Improving Cyber-Attack Predictions Through Information Foraging”, 2017 IEEE International Conference on Big Data (BIGDATA).

[7] Vishal Mehta, Pushpendra Bahadur, Manik Kapoor, Dr. Preeti Singh and Dr. Subhadra Rajpoot, “Threat Prediction Using Honeypot and Machine Learning”, 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015).

[8] Kushal Rashmikant Dalal and Mayur Rele, “Cyber Security: Threat Detection Model based on Machine learning Algorithm”.