

Cyberattack Analysis and Prevention Methods

Rajya Chivaluri
Wentworth Institute of Technology
School of Computing and Data Science

Abstract—This report is designed to outline the process of a project focused on cybersecurity and the protection of data from a cyberattack. The project uses data analysis and modeling concepts to analyze various vulnerabilities and recommend tools. The project then uses machine learning to analyze possible attack opportunities that present themselves to a piece of technology, and then configures the type of the attack and what defense mechanism can be used to handle said attack.

Keywords—Cybersecurity, cyberattack, phishing, ransomware, DDoS, malware, intrusion, logical regression

I. INTRODUCTION

Over the years, technology has become more accessible, but along with that, so has our data and privacy. There is a strong positive relationship between the increase in technology use and cyberattacks, including phishing attacks, ransomware attacks, DDoS attacks and unfortunately more. It impacts not only the individual users but also larger companies that use technology/the internet to complete their work. Thankfully, there are ways to solve these issues and experts work around the clock to strengthen all technology to fight off these attacks. This project will analyze the frequency of various types of cyberattacks, vulnerabilities to these attacks, assess the success of current tools being used against them, and lastly, investigate how treacherous an attack can be, based on its outcome.

II. DATASETS

A. Source of dataset

This dataset is derived from Kaggle, a platform full of datasets to be used for various analytical projects. Although the dataset, titled “Cyber Security Attacks”, was not uploaded by a trusted individual, it seems to be reliable enough for the sake of this project. It is maintained well and even lists the source at the end of each cell to show where the information about each attack came from.

B. Character of the dataset

The dataset contains 25 categorical variables, including malware indicators and (cybersecurity) attack types. With 40,000 incidents, the dataset provides a wide range of values for informative uses on the severity of cyberattacks in general, whilst showing the more fine-tuned details of each attack, for analytical purposes. Minimal data munging was used as this dataset was already clean.

C. Source of dataset

This dataset is derived from Kaggle, a platform full of datasets to be used for various analytical projects. The dataset, titled “Cyber Crimes Dataset”, contains a vast amount of information about various instances of cyberattacks. Though the dataset may not be completely reliable, this data is reliable enough to use as data to train the machine learning model.

D. Character of the dataset

This dataset contains 15 different columns, each as a different categorical variable, including target system of the cybercrime and the outcome of the attack. There are a total of 10,000 values in the dataset, all quite varied in their respected variables. Minimal data munging was used as this dataset was already clean, but some feature engineering was completed to take out some columns before completing the ML model.

E. Source of dataset

This dataset is derived from Kaggle, a platform full of datasets to be used for various analytical projects. The dataset, titled “Cyber incidents 2005 to 2020”, contains information about all cyber incidents multiple types of organizations have been affected by. As continuously mentioned, although the reference source the dataset is derived from may not be the most reliable, the sources of each attack data point is listed in its corresponding row.

F. Character of the dataset

The dataset contains 12 total columns with a total of 479 values, each as a different cyber incident. The columns contain categorical variables of each incident, including its description, its victims/target, and the type of incident.

III. METHODOLOGY

A. Exploratory Data Analysis

Exploratory data analysis (or EDA) will be used to analyze the frequency of various cyberattacks. The hope is that after completing this analysis, it will be known which attacks are the most frequent, making them a higher priority to handle. EDA

will also be used to look at the success rates of previous tools and measures that have been taken to either a) prevent cyberattacks or b) handle them once they have already occurred. This will help figure out where the gaps in safety are.

B. Correlation Analysis

The statistical method of correlation analysis will be used to understand the most frequent causes of cyberattack causes and technological vulnerabilities that make one prone to a cyberattack. Once this analysis is completed, it should be seen which vulnerabilities need to be handled first and foremost.

C. Logical Regression

The machine learning method of logical regression, when implemented, will use algorithms designed to take a potential cyberattack and predict its outcome. To do so, SciKit will be used to develop and carry out said algorithms. This would be useful in understanding what tools and resources should be allocated towards an attack, if its outcome will be a success. If its predicted outcome is a failure, less or no resources should be wasted on the potential attack.

IV. RESULTS

A. Cyberattack Frequency and Success Rates of Tools

After using exploratory data analysis to assess the frequency of various cyberattacks, it was found that out of the three types, the most frequent attack type/category was a DDoS (Distributed Denial of Service) attack. This type of attack, when used, floods any sort of online resource/tool with tons of unnecessary or even malicious traffic, causing for innocent users to lose access to that source. The relative frequency of this attack, assessed from the dataset used, was roughly 33.57%. The other two attack categories, malware attacks and intrusion attacks, had relative frequencies of approximately 33.26% and 33.16%, respectively. This goes to show that more tools should be implemented to specifically handle DDoS attacks, as they are the most frequent out of the popular cyberattacks. With that being said, exploratory data analysis was also used to dive into which tools are the strongest towards various attacks. The table below outlines various attack types (left) and the tool that worked best against it (right).

Brute Force	Intrusion Detection System (IDS)
Cross-Site Scripting	Antivirus
DDoS	Multi-Factor Authentication (MFA)
Malware	Security Information and Event Management (SIEM)
Phishing	Endpoint Detection
Ransomware	Endpoint Detection
SQL Injection	Endpoint Detection
Zero-Day Exploit	Security Information and Event Management (SIEM)

B. Cyberattack Vulnerability

After using correlation analysis, chi-square testing to be more exact, to assess what qualities in a system made it the most vulnerable to an attack, it was found that it was not actually the system at all and the sponsor of the attack, instead. Although weak systems are more susceptible to attacks, even strong systems with adamant sponsors to these attacks can fall victim to cyberattacks. In order to handle this and strengthen companies to prevent them from dealing with these, it would be imperative for the management to work with cybersecurity professionals to assess who may be out to attack a specific company and allocate more funding towards protecting them.

Vulnerability	(Approximate) P-Value
Affiliations	0.5031
Response	0.2294
Victims	0.2337
Sponsor	0.00000006997
Attack Type/Category	0.9748

C. Predicting Outcome of Cyberattacks

The machine learning method of logistic regression was used to predict the result of one categorical variable, which is the outcome of the cyberattack. If the prediction comes out to say that the attack will be a failure, then less resources should be used to handle said attack. But this tool is especially useful when the predicted outcome is a successful attack, as it will allow potential victims handle the attack by securing their system before it is too late.

V. DISCUSSION


Although the logistic regression model is useful, since it is trained off of one dataset thus far, it is just guessing 50/50 outcomes from the current attacks it knows of. In order to actually be useful, it would need to be trained on more cyberattack data in general.

Aside from the faults of the machine learning model, there is a quite bit of future work that would be done on this project. Currently, the model is trained to predict the *outcome* of an incoming cyberattack. By using different methods to create ML models, such as classification, the model can be worked on to predict the type of a cyberattack as it is incoming, so that it is known what tools should be used for the best outcome. This will further improve anybody's protection against cyberattacks, as they already know what attacks are most common (DDoS), what tools are the most useful against various attack types, including DDoS attacks (MFA), and what outcome could come out of an incoming attack.

VI. CONCLUSION

In conclusion, this project has looked at cyberattacks, as important as they are, very generally. To reiterate the purpose of this project, it was completed to educate those reading about which cyberattack is generally the most frequent, which vulnerability of a system is most prominent, and which current tools are the most useful against various types of attacks. It was found that DDoS attacks were the most frequent, meaning that more resources towards multi-factor authentication should be pushed to those who have either experienced DDoS attacks before or are prone to it. In terms of who/what is the most vulnerable to any attack, it was found that although weak systems are usually the most vulnerable, even a strong system with a determined sponsor makes for an even worse combination. Lastly, a heatmap/table was created outlining a few types of cyberattacks and their respective tools that proved to be the most useful in handling said attack. Lastly, using machine learning (and logistic regression more specifically), a model was created to predict the outcome of an incoming attack, to help assess how much effort and how many resources should be allocated towards the system being attacked. As mentioned prior, this project only scrapes the surface of cyberattack analysis and prevention, and there is a lot more work to be done.

REFERENCES

- [1] "Cyber Security Attacks," *www.kaggle.com*. <https://www.kaggle.com/datasets/teaminciribo/cyber-security-attacks/data>
- [2] Shakirul_09, "Cyber Crimes Dataset," *Kaggle.com*, 2024. <https://www.kaggle.com/datasets/shakirul09/cyber-crimes-dataset>
- [3] PrivacyMatters, "Cyber incidents 2005 to 2020," *Kaggle.com*, 2020. <https://www.kaggle.com/datasets/fireballbyedimyrnmom/cyber-incidents-up-to-2020/data> (accessed Apr. 14, 2025).
- [4] "Chi-Square Test in Python: A Technical Guide," *www.stratascratch.com*. <https://www.stratascratch.com/blog/chi-square-test-in-python-a-technical-guide/>
- [5]  A. C.J., "Logistic Regression in Python - Theory and Code Example with Explanation | ASPER BROTHERS," Aug. 25, 2021. <https://asperbrothers.com/blog/logistic-regression-in-python/>