Malicious Web Page Prediction

Nadine Safwat - 900212508

INTRODUCTION

Web security has become a huge concern in the digital age, with malicious web pages posing significant threats to data privacy and system integrity. This project aims to develop a machine learning model to detect malicious web pages effectively. Multiple models will be trained on a dataset comprising numerous webpage features, such as URL length, the IP address, domain registration details and more. Testing will then be done to select the most competent and accurate model available so that it can be further developed into a full system.

LITERATURE REVIEW

As the internet became a primary source for data and communication a whole world of security issues were created alongside, thus creating the field of cybersecurity. Malicious web pages are some of the most common examples on how the internet can be harmful and so many companies have already explored the task of classifying web pages as dangerous so that their users can be protected.

The most common approach, the blacklist approach [1], is one that does not use machine learning. Quite simply this method utilizes a database that is consistently updated with blocked URLs that are usually collected when a large group of people on the internet block the same URLs. It was then enhanced by using key features like filenames or pathnames of request URLs or attribute values of tags from the source code as a way to detect malicious web pages. This method was able to detect 84.36% of malicious web pages and had a zero false positive rate [2].

As one of the biggest search engines available today, Google has some extremely strong detection systems to protect their users against malicious webpages, one of them being HTML file analysis [3]. It is Google's patented method, by analyzing elements of HTML files and uses deep learning to compare its features against a threshold file. If the score of the webpage is low then it is considered malicious and Safe Browsing [4] will warn the user that they may be visiting an unsafe web page. Google claims that their systems protect over five billion devices every day however they do not guarantee 100% protection [4].

Other popular approaches heavily rely on machine learning. These methods use lexical properties like word length, word frequency, high frequency words etc. Then machine learning models are applied so that these features can be reviewed and categorized and generalized and the system can detect whether the website is malicious or not [1]. All of these methods have bragging detection rates from 95% to 99% [5] and so this is the method that will be explored for the project.

However we can see that no method has a 100% detection rate as there are many limitations in this field. Mainly the fact that the internet is constantly evolving and new malicious pages are made every day. Even the highly accurate machine learning models struggle to adapt to the ever changing tactics of cybercriminals and new models and training systems have to be created [6].

PROJECT OVERVIEW

The main aim of this project is to go through the process of training and testing machine learning models and exploring how one model can be better than another based on the context of the dataset. By the end of this project the application should be able to take in the URL of any website, extract the required fields needed for the model and accurately predict whether this web page is harmful or safe. The work will be divided into 4 phases explained below.

Phase 1 will begin by cleaning and engineering the data from the selected dataset so that it is sufficient for all the models to be best used on. This is a necessary stage so that the chances of accuracy and success are boosted. This data will then be split 80:20 to a training and testing set respectively. We allocated 80% to the training set so that the model can be given a chance to sufficiently understand the trends in the data and create appropriate predictors. The training set is split again to a smaller training set and a

validation set so that we can avoid over or underfitting. This can be done using the k-fold cross validation method which will partition the data into k-subsets so that one subset can be used for validation while the rest are used for training. Then we choose another subset and repeat these steps k-times so that we can then average all the partitions.

The next phase is the pilot study where any preprocessing or cleaning will be completed so that a full analysis can be done on all the features. The required models will be applied to the readied dataset and a report on the distributions over the major features and their correlation with the label is written up. To do this the true positive rate (TPR) and the true negative rate (TNR) must be calculated for each model. seeing as this is a binary classification problem, it is likely that logistic regression will be the most suitable model.

Based on the results from the pilot study, the most suitable model is chosen and the design for the final model will begin with an emphasis on making the system user friendly while still ensuring the purpose of the project is satisfied. The final phase will bring the full implementation of the system, which will be done using python and the model can be deployed.

DATASET OVERVIEW

DATASET OF MALICIOUS AND BENIGN WEBPAGES

This is the chosen dataset for this project as it has the required features as well as an appropriate size. On top of the statistics there are also references as to where the data has been gathered from which shows that the data available is reliable. It has also conveniently been split into respective testing and training sets. The dataset contains extracted attributes from websites that can be used for Classification of webpages as malicious or benign. There are 11 features, some of which are the URL, the IP address, the domain, if the website uses HTTPS or not, and most importantly whether the webpage is malicious or not, which is the label. There are 1.2 Million records in the training set and just over 350K records in the testing set. Kaggle has given the dataset a usability rating of 8.82

This dataset is relevant as it has almost all the features that can accurately predict if the web page is malicious or not and it has more than enough points for the models to train on. The only limitation I found in the set is that only 2% of the data points are classified as malicious which will cause changes in the next two phases. However, this dataset was the most suitable for the goal of this project.

TWITTER-BOT DETECTION DATASET

This dataset is what initially motivated the topic for this project. It is designed so that a machine learning model can look at tweets and recognize if it comes from a bot. It has features that can be strong indicators of bots or malicious activities, including hashtags, retweet count, follower count, verification, and the bot label, however there are only 50K data points which isn't enough to train and test a machine learning model suitably. Kaggle has given the dataset a usability rating of 10

This dataset is relevant to the topic as twitter bots are simply another form of cyberattacks on the internet so the approach between detecting bots and detecting malicious URLs is extremely similar. If there had been more tweets gathered this dataset would have been an extremely good candidate to use.

BOT_DETECTION

This dataset is also created to be able to detect bots on a website which can slow down a site or even steal sensitive data which is relevant to the topic as bots can be some of the worst forms of malware on the internet. The dataset has 19 features some of which are the IP address, domain name and device type. There are also 1.05M data points which makes it an extremely available dataset to use. Kaggle has also given it a usability rating of 8.24. However many of the available features are not extremely usable for Malware detection, and there are too many NULL values in the dataset, which is the main reason this dataset was not chosen for the project.

MALICIOUS URLS DATASET

This dataset has a collection of 651,191 URLs out of which 428103 are benign or safe, 96457 are defacement URLs, 94111 phishing URLs, and 32520 malware URLs. This dataset has a usability rating of 10 but only two columns: the URL and the category of the webpage. This means that there aren't enough features to be able to

classify any new URLs into any category so the dataset could not be used for this specific project.

REFERENCES

[1] Swarnkar, M., Sharma, N., Kumar Thakkar, H. (2023). Malicious URL Detection Using Machine Learning. In: Thakkar, H.K., Swarnkar, M., Bhadoria, R.S. (eds) Predictive Data Security using AI. Studies in Computational Intelligence, vol 1065. Springer, Singapore. https://doi.org/10.1007/978-981-19-6290-5_11

[2] Rao, R.S., Pais, A.R. (2017). An Enhanced Blacklist Method to Detect Phishing Websites. In: Shyamasundar, R., Singh, V., Vaidya, J. (eds) Information Systems Security. ICISS 2017. Lecture Notes in Computer Science(), vol 10717. Springer, Cham. https://doi.org/10.1007/978-3-319-72598-7_20

[3] US20210092130A1 - detecting malicious web pages by analyzing elements of Hypertext Markup Language (HTML) files (no date) Google Patents. Available at: https://patents.google.com/patent/US20210092130A1/en.

[4] Google. Available at: https://safebrowsing.google.com/.

[5] Utku, A. and Can, U., 2021. Machine Learning-Based Effective Malicious Web Page Detection. International Journal of Information Security Science, 11(4), pp.28-391 https://dergipark.org.tr/en/download/article-file/2556308

[6] Ha, M., Shichkina, Y., Nguyen, N., Phan, TS. (2023). Classification of Malicious Websites Using Machine Learning Based on URL Characteristics. In: Gervasi, O., et al. Computational Science and Its Applications – ICCSA 2023 Workshops. ICCSA 2023. Lecture Notes in Computer Science, vol 14112. Springer, Cham. https://doi.org/10.1007/978-3-031-37129-5_26