Determining Website Legitimacy Based on Phishing Website Data Using Decision Tree Method

Jaken Whipp
Hood College, Frederick, MD, USA
jaw32@hood.edu

ABSTRACT

Decision trees are a type of data mining model which predicts the classification of an item based on a given set of parameters. A key feature of decision trees is their ability to handle the classification of multiple classes. Phishing attacks are cyber-attacks in which somebody mimics a legitimate organization to obtain personal information. The use of fraudulent links is one of the most common sources of phishing attacks. In this study, I develop a decision tree model based on dataset of phishing website data from the UCI machine learning repository. When passed website data as parameters, the decision tree model had an accuracy of 86.94%. Towards the end of the paper, I discuss possible approaches to improving the accuracy of the model.

1. INTRODUCTION

Phishing attacks are one of the most common, and most efficient, cyber-attacks producing a loss of $44,213,707 in the year of 2021 alone [1]. In recent years, the amount of phishing scams has increased more than 34 percent from 2020 to 2021 [1]. Two factors which contribute to the increase in phishing attacks is their success and their simplicity. Phishing attacks involve the attacker who mimics legitimate organizations and sends fraudulent links, which obtain sensitive information, and the unsuspecting user who has their information taken. However, it can sometimes be difficult to determine whether a link is fraudulent or legitimate as sometimes the fake messages can look, almost, the same as messages from the company themselves. Some phishing attacks are so intricate that the only difference between a real and fake message is that in the fake one the button redirects and takes your information instead of taking you to the companies' website. As scammers become more proficient with the creation and use of phishing attacks, something must be done to prevent these attacks from taking place. If websites can be correctly detected as phishing attacks, preventative measures can be taken to keep them from being sent out to unsuspecting people. The purpose of this study is to create a decision tree model which will be able to predict whether a website is legitimate or fraudulent based on a given set of parameters about the site. The remainder of the paper is organized as follows. In Section II, an overview of existing research related to decision trees is provided. Section III highlights details pertaining to the dataset used within the study as well as the proposed methodology for the creation of the decision tree model. The experimental results are discussed in Section IV with Section V containing the paper's conclusion.

2. LITERATURE REVIEW

In the past, several studies have utilized decision tree models with promising results. One study [2] aimed to classify sleep stages based on using multi-class Support Vector Machines (SVMs) alongside a decision tree approach. The results of this study showed that a dendrogram-based SVM achieved an overall accuracy of 0.88 which was comparable to the scoring of professional who had an overall accuracy of 0.92.

Another study [3] which utilized a decision tree approach was one aimed at classifying ground water quality in the Ardebil Region in Iran. The classification performance was evaluated using Correctly Classified Instances (CCI) and a kappa statistic. The overall average performance for CCI was 0.88 with the kappa statistic metric being 0.83. Additionally, the results demonstrated that the decision tree approach was more precise and efficient when compared to Principal Component Analysis (PCA).

In [4], four different decision tree algorithms were evaluated to test their effectiveness at classifying breast cancer. The study found that a Priority based decision tree classifier had the best performance with an accuracy of 93.63% and took less time to build than other models.

Previous studies utilizing decision tree models demonstrate their effectiveness for classification and provide hope that a decision tree approach aimed at determining website legitimacy may prove beneficial. However, there are also existing studies aimed at classifying phishing websites. A similar study [5] investigated the same dataset by using an Associative Classification (AC) method known as Multi-label Classifier based Associative Classification (MCAC) to determine if it could be applied to the issue of phishing problems. They found that MCAC can detect phishing websites with great accuracy.

A different study [6], also utilizing the same dataset as this study, investigated phishing websites by using multilayer perceptron (MLP), a type of neural network, trained with Hybrid Salp Swarm Jaya (HSSJAYA). Results were compares with several other MLPs trained by either the Cuckoo Algorithm (CA), Genetic Algorithm (GA), or Firefly Algorithm (FFA). They found that the MLP trained with HSSJAYA was able to determine phishing websites.

3. METHODOLOGY

Data was obtained from the UCI Machine Learning Repository [7], which contained detailed information regarding legitimate, suspicious, and illegitimate websites. In total, the dataset consisted of information about 1353 websites, which included 702 phishing websites, 548 legitimate websites, and 103 suspicious websites. Each website within the dataset had 10 attributes associated with it: Slow Frequency Hopping (SFH), Pop-Up Window, Secure Socket Layer (SSL) Final State, Request URL, URL of Anchor, Web Traffic, URL Length, Age of Domain, whether it had an IP Address, and finally whether the website was phishing, legitimate, or suspicious. Values for most attributes were either a -1, 0, or 1, with -1 representing phishing, 0 representing suspicious, and 1 representing legitimate. The only attributes which were exceptions to this rule were Age of Domain and whether it had an IP Address. The Age of Domain attribute consisted of values of -1 or 1 and the IP Address attribute consisted of values of 0 and 1. No information is provided explaining why there is a missing value for these attributes nor is there information provided which further explains their meaning. Additionally, the dataset did not provide any information regarding how attribute values were decided.

Records within the dataset looked like: 1,-1,1,-1,1,-1,-1,1,0,-1

The dataset was contained within an Attribute-Relation File Format (arff) which made the data require some pre-processing for it to properly be utilized. Once read from the file, the numeric values were all strings and were different than anticipated. Where values of -1, 0, or 1 were expected, values of b'-1', b'0', and b'1' were found instead. As such, pre-processing was necessary to remove the unnecessary b's and quotation marks. Additionally, the strings containing just the numbers were then converted to integers so that the library which creates the model could properly utilize the data. Once the necessary modifications had been made to the data so that it could be used, the original dataset was split into a training and testing set with a ratio of 7:3.

The creation of the decision tree model was facilitated by the pandas, sklearn, and scipy python libraries. By using pandas and scipy, the dataset was able to be properly read and loaded into a data frame. The library which did the bulk of the work creating the decision tree model was the sklearn library. The decision tree model was trained using the training set and the Gini index classification criteria.

4. RESULTS

The model's performance was evaluated by generating predictions for the testing set by providing it the parameters, generating a classification based on the parameters, and comparing the classification to the actual legitimacy of the website. A depiction of the decision tree model is shown in Fig. 1.

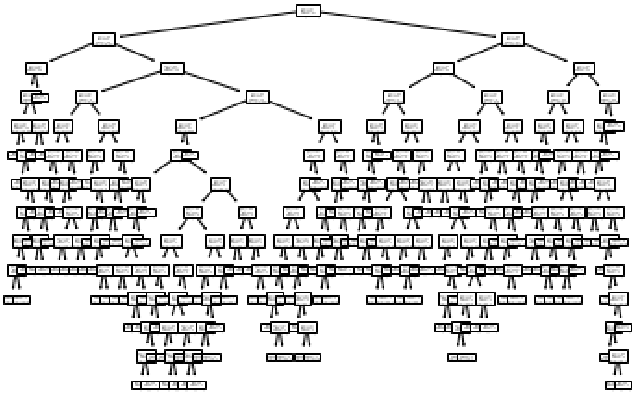

Fig 1. An overview of the trained decision tree model

The model correctly predicted the legitimacy of websites within the testing set 86.94% of the time. An accuracy within the 86th percentile is comparable to previous studies which investigated decision tree models. However, this would mean that roughly 14% phishing links were incorrectly classified by the model, which has several potential meanings. Of the 14% of links which were incorrectly classified, it would not include just phishing sites being marked as either safe or suspicious but would also include legitimate sites being incorrectly classified. Both scenarios hinder people as it either puts them at risk of being on the receiving end of a phishing attack or missing out on a link which may have been important or beneficial to them. Ideally, should a system be implemented to take preventative measures against phishing links, the accuracy should be as close to perfect as possible.

There may be several reasons the accuracy of the decision tree model was not higher. Firstly, the model's performance could have been hindered by the "suspicious" classification. Due to the size of the dataset, and the lack of information regarding how attributes received their values, removing suspicious websites and attribute values would be difficult. By choosing to maintain the suspicious values, it turns the classification problem from choosing between two options to choosing between three options. Having an additional option would make classification more error prone. As such, if the dataset had focused solely on whether a website was a phishing or legitimate site and not considered suspicious sites, the model may have performed better. A larger dataset may also improve the accuracy of the model. The dataset used in the study contained only 1353 cases with those cases being split by a 7:3 ratio for training and testing. As such, the data used to train the model was even smaller than it could have been. If the study were to be re-examined using a larger dataset, the accuracy of a decision tree model trained with the Gini index may improve.

5. CONCLUSION

This study aimed to develop and evaluate a decision tree model for predicting the legitimacy of websites. It incorporated website data found at the UCI Machine Learning Repository. The experimental results showed that a decision tree model trained with the Gini index produced an accuracy of 86.94%. While the results were comparable to other studies utilizing decision trees, the accuracy of the model could be better. Further investigation of the decision tree model approach explored in this study could focus solely on phishing and legitimate, ignoring suspicious, and use a larger dataset.

6. REFERENCES

[1] P. Bischoff, "The state of phishing in the US: Report and statistics 2021," Comparitech, 28-Sep-2022. [Online]. Available: https://www.comparitech.com/blog/information-security/state-of-phishing/#:~:text=5%2C831%20over%2060s%20were%20victims,lost%20to%20%249.2%20million%20lost. [Accessed: 4-Oct-2022]. Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

[2] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, and K. Jerbi, "Learning Machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," Journal of Neuroscience Methods, vol. 250, pp. 94–105, 2015.Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.

[3] S. M. Saghebian, M. T. Sattari, R. Mirabbasi, and M. Pal, "Ground water quality classification by decision tree method in Ardebil Region, Iran," Arabian Journal of Geosciences, vol. 7, no. 11, pp. 4767–4777, 2013. Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.

[4] P. Hamsagayathri and P. Sampath, "Performance Analysis of Breast Cancer Classification Using Decision Tree Classifiers," International Journal of Current Pharmaceutical Research, vol. 9, no. 2, p. 19, 2017.

[5] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining,"

Expert Systems with Applications, vol. 41, no. 13, pp. 5948–5959, 2014.

[6] E. Erdemir and A. A. Altun, "Website phishing technique classification detection with HSSJAYA based MLP training," Tehnicki vjesnik - Technical Gazette, vol. 29, no. 5, 2022. Series. ACM, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738.

[7] [7] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php, last accessed on November 29, 2022