Optimizing Phishing Attack Identification with Supervised Machine Learning Techniques

Olaniyi A. Ayeni, Hope O. Akinyemi
Department of Cybersecurity
Federal University of Technology
Nigeria

Abstract

Phishing is a technique under Social Engineering attacks which is most widely used to get sensitive information from users, such as login credentials i.e. usernames, passwords or other security tokens, credit and debit card information, etc.[4] In most cases, Phishing does not require high technicality making it the third most performed attack according to multiple sources. Despite many solutions being proposed to completely eradicate or at least mitigate about 80% of Phishing attacks, Phishing continues to be prevalent. Vishakha and Sahil [24] were unable to automatically detect web pages and the systems were limited to system applications alone. This work therefore aims to design a system for the detection of phishing attacks, Implement the design and evaluate the performance of the algorithm. Jupyter Notebook was the medium used for analysis. The two algorithms used were two-class logistic detection and Naïve Bayes. The dataset (phishing.csv) obtained from Kaggle was split into training and test datasets in the ratio 80:20. The result of two class logistic regression was 94.2% compared to 85% test accuracy by the Naïve Bayes algorithm [2]. The combination of multiple machine learning algorithms to create more sophisticated security could provide a more accurate detection but also a fairly negative effect in complexity and technicality in handling. With more training experts could be groomed to align with its sophistication.

1. Introduction

Phishing is an internet attack mostly propagated by email. It has grown rampant in recent years with the goal being to trick the recipient into believing that the message is legit and urgently requires attention like a request from their bank, from their place of work, An urgent question from their boss or/and to click a link or download an attachment. Phishing can be very difficult to detect especially if the targeted victims are the non-technical staff of the organization. In phishing, the attackers masquerade as a trusted entity, often a real or close person, or the organization where the victim works or does business with. It is one of the oldest types of cyberattacks, dating back to the 1990s, and it is still one of the most prevalent, with phishing messages taking dynamic turns to evade being detected. Phishing has thrived due to the urgency of the message carries, for example, a newly employed staff receives a mail from his direct boss asking him to re-upload a document he has sent earlier. The new staff being eager to impress his new boss might not necessarily think of confirming over a voice call before getting himself phished. In other cases, the big figures in an organization such as the Executive officers, and Directors can be the target since that would be of more advantage financially, this is a type of phishing called Whaling. A phishing mail could contain a malicious link that could corrupt the system, leak sensitive data, allow backdoor and many more.

2. Objectives

Many methods have been introduced in recent years to curb phishing attacks. However, they have failed to completely remediate the prevalent situation, several machine learning and deep learning algorithms have been developed to detect phishing sites but can only be used by a technically oriented user. The elements derived from Technology Threat Avoidance Theory have also been incorporated with the game theory to develop a phishing detection game which may not work in complex cases, thereby reducing its efficiency. The objectives of this paper are to:

a) Design a system for detection of phishing attack.

b) Implement the design in (a).

c) Evaluate the developed system order [3].

3. Comparing related works

Arachchilage and Love [18] developed a framework to thwart phishing attacks. The given game design framework using a theoretical model derived from Technology Thread Avoidance Theory (TTAT) enhanced users' avoidance behaviour to

 | DOI: 10.20533/jitst.2046.3723.2024.0104

thwart phishing threats. Worked well with its human friendliness and attractive approach to make sure everyone is kept busy through the addition of learning but may not be reliable enough to combat the ever-dynamic phishing attack

Mohith Gowda HR et al. [19] improvised and introduced detection methods into the browser architecture named 'Embedded Phishing Detection Browser' (EPDB) due to the failure of other methods aiming to preserve the existing user experience while improving the security of an individual system or network as the case may be. This developed a technique that can be easily used by everyone to detect non-legitimate websites accurately in real-time using EPDB. This method may not be proactive enough to be trusted.

Whittaker, C., et al **[16]** aimed to develop a phishing detection medium by describing the design and performance characteristics of a scalable machine learning classifier. The writer aimed at reducing phishing through the blacklisting of malicious mail researched a scalable machine learning classifier that can be used to automatically maintain a blacklist of phishing pages. The method employed was building a phishing classifier. Despite this research's precision in reducing the number of false positives, it was incapable of detecting pages disguised as a non-phishing page with non-phishing links.

Garera et al. [7] studied the structure of URLs used in phishing attacks. This paper focused on improving URL-based detection of phishing sites having noticed that other detection techniques were not robust enough. This paper identified several fine-grained heuristics that can be used to distinguish between a phishing URL and a benign URL. These heuristics are used to model a logistic regression classifier. It then discovered several new and generic features for recognizing phishing URLs. The major contribution of this work is a large-scale measurement study conducted on Google Toolbar URLs but the new generic features identified may not be enough to guarantee 100 percent safety from Phishing due to the ever-evolving cyberworld.

Chen et al. [6] researched to show that 60 % of internet users do not use the web browsers' built-in phishing-website detection tools. Therefore, it addresses the issue of inappropriate trust by arguing that trust in detection tools should be calibrated by closely aligning trust to the capability of the tool. This paper suggests that calibrating users' trust in detection tools should play an important role in promoting protection against phishing websites, It used a theoretical approach to identify the salient factors (or trust calibrators) in calibrating trust in phishing-website detection tools and the outcomes of calibrated trust in such tools for individual users.

As shown in Table 1, Goud and Mathur [8] suggested that phishing attacks can be handled based on source code URL or image. This research designed the model based on URL features, This research tried to find the most important attributes that can determine whether it is a phishing website or not. Some of the websites' access is marked as unauthorized. This research tried to find the most important attributes that can determine whether it is a phishing website or not. The major goal of feature selection is to reduce the computation time of the complex models. In this research, the feature selection to detect the phishing website is solved using the ensemble algorithms. It proposed a model of feature engineering.

Vishakha and Sahil [20] built up a framework to identify the phishing site and this framework utilizes effective methods to recognize the phishing sites. The phishing sites can be distinguished depending on some significant attributes like URL and Domain Identity. RNN (Recurrent Neural Network) was primarily used in NLP. This is because the RNNs cannot take care of the long-term dependency of a sentence. To detect and predict phishing websites, an intelligent, flexible and effective system that is based on using Machine Learning algorithm was used but findings cannot automatically detect the web page and the compatibility of the Application with the web browser, Furthermore, it is limited to being a system application. Table 1 compares the various methods used by different authors in implementing phishing detection.

Table 1. Comparison of related works and the method used

Author	Method used
Vishakha P.R. and Sahil S.J. (2020)	URL and doman identirty
Hossain, S., et al (2020)	Random forest
Salahdine, F., et al (2021	Rule-based, white and blacklist, heuristic, and hybrid
Abedin, N. F., et al (2020)	Random Forest
N.Swapna Goud et al (2021)	Feature Engineering
Mohith Gowda HR, et al (2020)	Embedded Phishing Detector

4. Methodology

To execute this project, Machine Learning would be used specifically, Two class logistic regression. Two-class Logistic regressions is an outstanding technique insight that is utilized to foresee the likelihood of a result and is particularly prominent for arrangement undertakings. The calculation predicts the likelihood of an event or an occasion by fitting information to a strategic capacity. This technique is used to create a phishing detection model which predicts only two outcomes that are spam or ham, n model which predicts only two outcomes that is spam or ham.

4.1. Machine Learning

Machine learning is an evolving branch of computational algorithms that are designed to emulate

 | DOI: 10.20533/jitst.2046.3723.2024.0104

human intelligence by learning from the surrounding environment. They are considered the working horse in the new era of the so-called big data. Machine learning offers useful ways to approach problems that otherwise defy manual solutions [11]. It addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical fields, Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications. The effect of machine learning has been felt broadly across computer science and a range of industries concerned with data-intensive issues, such as consumer services, the diagnosis of faults in complex systems, and the control of logistics chains. Machine-learning methods have been developed to analyze high-throughput experimental data in novel ways.

Machine learning has been applied successfully in diverse fields ranging from pattern recognition, computer vision, spacecraft engineering, finance, entertainment, and computational biology to biomedical and medical applications [12]. A diverse array of machine-learning algorithms has been developed to cover the wide variety of data and problem types exhibited across different machine-learning problems. Machine-learning algorithms can be viewed as searching through a large space of candidate programs, guided by training experience, to find a program that optimizes the performance metric [14]. Its algorithms vary, in part by how they represent candidate programs such as decision trees, mathematical functions, etc and in part by how they search through this space of programs.

4.2. Logistic Regression

Logistic regression is a supervised machine learning algorithm developed for learning classification problems. A classification learning problem is when the target variable is categorical [5]. The goal of logistic regression is to map a function from the features of the dataset to the targets to predict the probability that a new example belongs to one of the target classes.

Logistic regression makes it possible to test whether two variables are linearly related and to calculate the strength of the linear relationship if the relationship between the variables can be described by an equation of the form $Y = a + bx$, where Y is the variable being predicted and X is a variable whose values are being used to predict Y. Logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables [15].

The datasets used for this project were acquired from the Mendeley Dataset which is a collection of publicly available lists of phishing and legitimate websites from which the features presented in the dataset were extracted. It is compiled in a .csv file format. The datasets comprise the features extracted from the collection of website addresses and consist of about 1000 rows and 50 columns. In preprocessing, data is converted or consolidated so that the mining process result can be applied or may be more efficient.

5. Implementation and Results

To implement, we found out the percentage distribution of the legitimate and the phishing websites from our data sets, with the result well represented through a bar graph, it is safe to say both the phishing websites and the legitimate sites are in almost the same proportion visually.

Data Preprocessing

To appropriately set up the data as input for a particular Data Mining algorithm, a series of techniques known as data pretreatment is applied. Preprocessing data is typically a required step. It creates fresh data that matches a Data Mining method from previously meaningless data. First of all, poorly prepared data may not be received for operation by the Data Mining method, or it will undoubtedly disclose mistakes while it is running. In the best of circumstances, the algorithm will provide findings, but they won't make sense or be regarded as reliable information [9]. The data is transformed or combined in this preparation stage so that the mining process outcome can be used or could be more effective. A few of the subtasks that fall under the category of data transformation are feature building, normalization, discretization, generalization, aggregation, and smoothing. Most of the procedures in data transformation, including data cleaning, will be broken out into separate tasks since they are referred to as a broad data preparation family.

Training and Result

Here, we used a data-oriented means to train and learn, which essentially takes several N features for model training the model and the metrics are re-evaluated. This uses a logistic regression model to perform a repetitive training process to determine the perfect number of features that can be utilized to find the model that best fits without many adjustments.

As shown in the legend below, 0.5 is the average correlation in the positive direction; for the first 10 columns, the highest value is NumDots at 0.29, which is far from an average correlation with the label (see Figure 1).

 | DOI: 10.20533/jitst.2046.3723.2024.0104

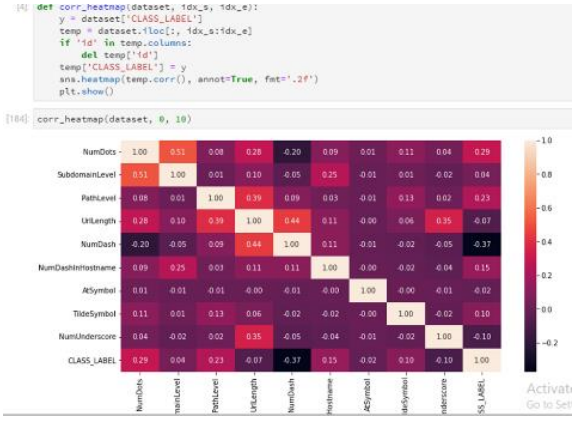


Figure 1. Correlation for the first 10 columns

As shown in the legend above, 0.5 is the average correlation in the positive direction; for the first 10 columns, the highest value is NumDots at 0.29, which is far from an average correlation with the label (see Figure 1).

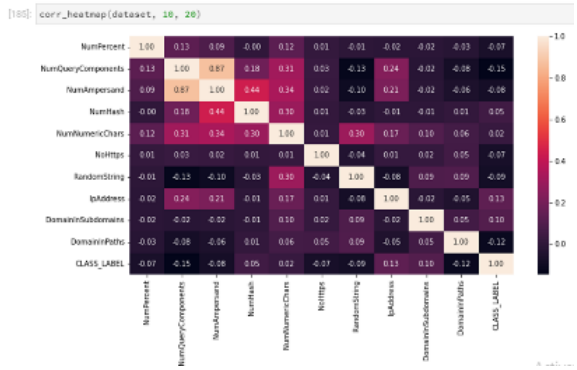


Figure 2. Correlation between columns 10 and 20

There are still no strong, or even medium-level correlation features, both positive and negative, in the next 10 rows with CLASS_LABEL. The highest being 0.13 from ipAddress (see Figure 2).

The legend shows results for rows 40 – 50, At -0.54, PctExtNullSelfRedirectHyperlinksRT is the sole column within this group exhibits some association with labels, albeit negatively this time. This could indicate that as the percentage of null self-redirect hyperlinks increases, so does the likelihood of phishing attacks (see Figure 3).

Training the logistic model would help easily predict if an email is phishing or spam by rightly considering historical data and, in this case, our dataset. We started the loop from 10 as we will train from the 10th feature to the 50th feature to find the optimal number of features needed for this problem (see Figure 4).

The training produced a considerably high level of accuracy and precision, showing that it is phishing detection ability would almost accurately detect when being phished (see Figure 5).

Figure 3. Correlation between columns 40 and 50

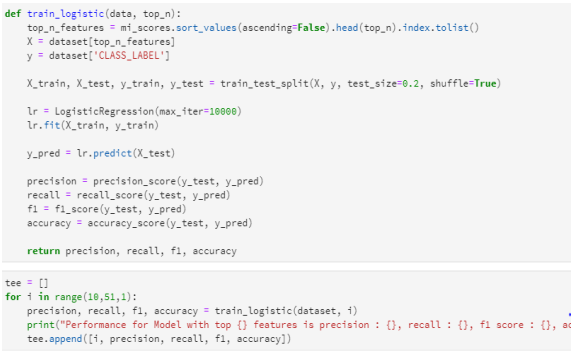


Figure 4. Training of the dataset

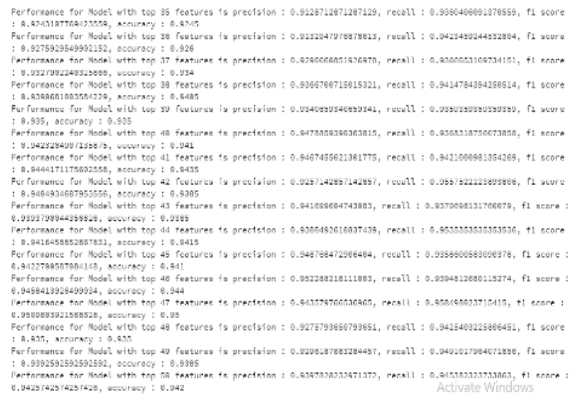


Figure 5. Training Result

The model was unstable as more features were added during training; so, we found the 39th feature to be the one with the best metrics and all-around performance; it came closest with the lowest score of 0.947162. The result is represented graphically in Figure 6.

Additionally, recall is also constantly performing well, but our model tends to have problems with precision scores. Compared with the Naïve Bayes model (see Figure 7).

 | DOI: 10.20533/jitst.2046.3723.2024.0104

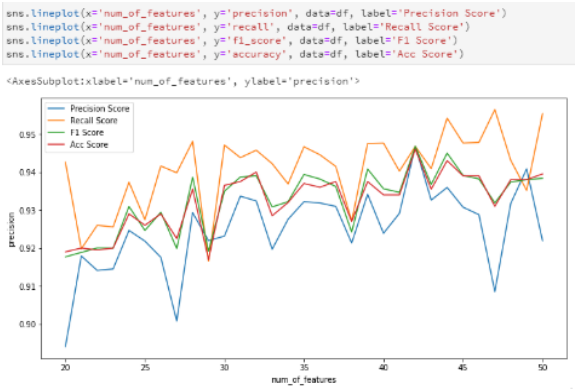


Figure 6. Graphical representation of data result

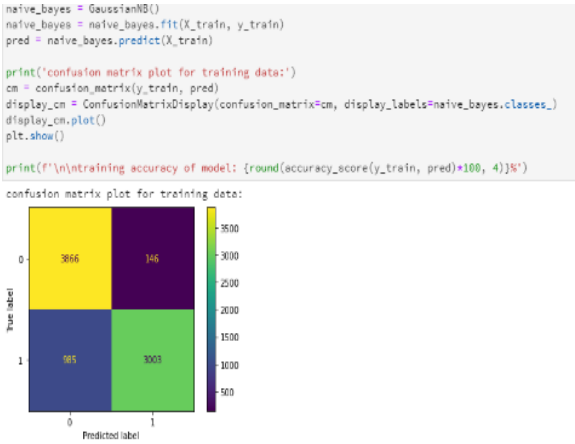
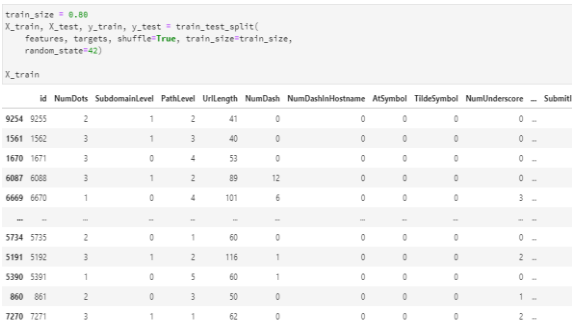


Figure 7. Comparative modelling



	id	NumDots	SubdomainLevel	PathLevel	UrlLength	NumDash	NumDashInHostname	AtSymbol	TildeSymbol	NumUnderscore	...	SubmitI
9254	9255	2	1	2	41	0	0	0	0	0	...	
1561	1562	3	1	3	40	0	0	0	0	0	...	
1670	1671	3	0	4	53	0	0	0	0	0	...	
6087	6088	3	1	2	89	12	0	0	0	0	...	
6669	6670	1	0	4	101	6	0	0	0	3	...	
...	...	...	...	...	...	...	...	...	...	...	...	...
5734	5735	2	0	1	60	0	0	0	0	0	...	
5191	5192	3	1	2	116	1	0	0	0	2	...	
5390	5391	1	0	5	60	1	0	0	0	0	...	
860	861	2	0	3	50	0	0	0	0	1	...	
7270	7271	3	1	1	62	0	0	0	0	2	...	

Figure 8. Bayes Training

Since Machine Learning models work on past observations, the accuracy of the model could be subject to its exposure to training data. Meanwhile, the testing dataset tests the performance of the trained data to determine if it needs more training or has been trained effectively enough, so using the 20:80 proportion is not new (see Figure 8).

6. Conclusion

In this paper, We suggested utilizing machine learning to detect phishing attacks. Using the dataset, three classifiers are trained and evaluated. A parametric research is carried out for each classifier, and the best findings are reported for assessment. Among the evaluated algorithms, Naïve Bayes demonstrated the best accuracy at 94.2%. Comparatively, N. F. Abedin, et al., [21] in "Phishing Attack Detection using Machine Learning Classification Techniques", the performance of the three widely used machine learning classifiers is compared. Among these three classifiers, random forest performance is the highest with a precision of 97% (see Table 2).

Table 2. Comparing Machine Learning Performance with Similar Authors

Author	Recall	Precision	F1-score	Accuracy
This work, 2023	94.5%	93.9%	94.2%	94.2%
Basnet, R., Mukkamala, S., and Sung, A. H. [10]	97%	96.6%	96.8%	96.8%
Niels P., Dean M., Panayiotis M., et al. [11]	94%	96%	95%	95%

In the future, features will be changed to increase accuracy. Systems based on deep learning models and neural networks that can handle vast amounts of data can be built to identify phishing attacks from logged datasets.

Future research to increase the system's accuracy will also take feature reduction approaches into consideration.

7. References

[1] Afroz, S., and Greenstadt, R. (2011, September). Phishzoo: Detecting phishing websites by looking at them. In 2011 IEEE Fifth International Conference on Semantic Computing (pp. 368-375). IEEE.

[2] Ayeni O.A and Akinyemi H.O.D. (2023). Phishing Detection Using Machine Learning

[3] Ayeni O.A (2022). A Supervised Machine Learning Algorithm for Detecting Malware.

[4] Basnet, R., Mukkamala, S., and Sung, A. H. (2014). Detection of phishing attacks: A machine learning approach. Soft computing applications in industry, 373-383.

[5] Bisong, E., and Bisong, E. (2019). Logistic regression. Building machine learning and deep learning models on Google Cloud Platform: A comprehensive guide for beginners, 243-250.

[6] Chen, Y., Zahedi, F. M., Abbasi, A., and Dobolyi, D. (2021). Trust calibration of automated security IT artefacts: A multi-domain study of phishing-website detection tools. Information and Management, 58(1), 103394.

[7] Garera, S., Provos, N., Chew, M., and Rubin, A. D. (2007, November). A framework for detection and measurement of phishing attacks. In Proceedings of the 2007 ACM workshop on Recurring malcode (pp. 1-8).

 | DOI: 10.20533/jitst.2046.3723.2024.0104

[8] Goud, N. S., and Mathur, A. (2021). Feature Engineering Framework to detect Phishing Websites using URL Analysis. International Journal of Advanced Computer Science and Applications, 12(7).

[9] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: Springer.

[10] HR, M. G., MV, A., S, G. P., and S, V. (2020). Development of an anti-phishing browser based on random forest and rule of extraction framework. Cybersecurity, 3 (1), 20.

[11] Jain, A. K., and Gupta, B. B. (2018). Towards detection of phishing websites on client-side using a machine learning-based approach. Telecommunication Systems, 68, 687-700.

[12] Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[13] Kiruthiga, R., and Akila, D. (2019). Phishing website detection using machine learning. International Journal of Recent Technology and Engineering, 8(2), 111-114.

[14] Lavesson, N., and Davidsson, P. (2007). Evaluating learning algorithms and classifiers. International Journal of Intelligent Information and Database Systems, 1(1), 37-52.

[15] Menard, S. (2002). Applied logistic regression analysis (No. 106). Sage.

[16] Whittaker, C., Ryner, B., and Nazif, M. (2010, February). Large-Scale Automatic Classification of Phishing Pages. In Ndss (Vol. 10, p. 2010).

[17] Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., and Cheung, D. (2009, December). Naive Bayes classification of uncertain data. In 2009 Ninth IEEE International Conference on data mining (pp. 944-949). IEEE.

[18] Arachchilage, N.A.G., Love, S. (2013). A game design framework for avoiding phishing attacks. Computers in Human Behavior, Volume 29, Issue 3, ISSN 0747-5632. DOI: 10.1016/j.chb.2012.12.018.

[19] Mohita Gowda, H.R., Adithya, M.V., Prasad, G., Vinay, S. (2020). Development of an anti-phishing browser based on random forest and rule of extraction framework. Cybersecur 3, 20 (2020).DOI: 10.1186/s42400-020-00059-1.

[20] Ratnaparkhi, V.P., and Jambhulkar, S.S. (2020). Framework for detection and prevention of phishing websites using machine learning approach.

[21] Abedin, N. F., Bawm, R., Sarwar, T., Saifuddin, M., Rahman, M. A., and Hossain, S. (2020, December). Phishing attack detection using machine learning classification techniques. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 1125-1130). IEEE.

[22] Salahdine, F., El Mrabet, Z., and Kaabouch, N. (2021, December). Phishing attack detection is a machine learning-based approach. In 2021 IEEE 12th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON) (pp. 0250-0255). IEEE.

[23] Hossain, M. L., Shams, S. N., and Ullah, S. M. (2024). Comparative study of machine learning algorithms for wind speed prediction in Dhaka, Bangladesh. Sustainable Energy Research, 11(1), 23.

 | DOI: 10.20533/jitst.2046.3723.2024.0104