SecureSense: A Data-Driven Framework for Phishing Attack Prevention

Abstract:

Since the outbreak of the COVID pandemic, our patterns of living, learning, and working have fundamentally shifted online, resulting in several new cyber security issues. As a result, concerns over phishing attempts have increased significantly. Phishing is a sort of fraud in which the offender sends emails or utilizes other communication channels while acting as a reputable company or person in order to acquire sensitive information such as login passwords or account information. It is an attempt to steal money or an individual's identity by deceiving individuals into providing personal information on websites that seem legitimate but are in fact fraudulent, such as credit card numbers, bank account information, and passwords. The cybercriminal contacts them, and they often send fraudulent messages with links to phishing websites that look to originate from reputable firms, friends, or acquaintances. Due to its efficacy, phishing is a prevalent kind of cybercrime. Cybercriminals have found success using emails, SMS, direct messages on social media, and video games to lure customers into divulging their personal information.

I. Introduction:

According to CBS news, the Illinois Department of Employment Security (IDES), international criminals continue to attempt to exploit unemployment systems across the country by contacting individuals via phishing emails, SMS messages, and social media posts, resulting in identity theft ("Phishing Scams Linked to IDES on the Rise., 2021). In another instance, a phishing effort impersonates the U.S. Department of Labor in an attempt to get login credentials. Initially, the attackers developed a false version of the Department of Labor website by copying and pasting authentic HTML and CSS code from the actual website, then utilized a legitimate email server to transmit phishing emails to avoid being discovered by security defenses (Whitney, 2022). In addition, they created new domains that might circumvent security checks and hide from threat intelligence. Lastly, the attackers offered consumers what looked to be a real government website but redirected them to a phishing form from which they could collect their login credentials (Whitney, 2022).

Email is the most common delivery route for phishing attempts. The fraudster will create a fictitious domain that impersonates a legitimate business and send tens of thousands of generic requests. These con artists build a domain using the actual company's name in the URL. In order to determine whether or not an email is phishing, we will use machine learning algorithms to evaluate websites and emails, including domains, subdomains, and any unique UI elements and UX interactions. Combining three supervised learning models, namely the Logistic Regression model, the Binary Decision Tree model, and the Random Forest, we will apply them to two datasets: the Webpage's dataset Phishing Legitimate Full.csv from Mendeley Data.

II. Methodology

1. Dataset Description:

The dataset includes 48 features which were collected from 5000 legitimate websites and 5000 phishing legitimate websites which were published on Mendeley Data. Those websites were downloaded from January to May 2015, and from May to June 2017. The author of this dataset is Choon Lin Tan from Universiti Malaysia Sarawak Faculty of Computer Science and Information Technology. Phishing webpages were taken from PhishTank and OpenPhish. Legitimate web pages were taken from Alexa and Common Crawl.

The first column of the dataset contains the ID numbers of each website. From the second column to the 48th column are the 48 features which provide a comprehensive overview of the legitimacy of the website. Some of the key features are URL Length, Presence of the "@" symbol, IP Address in the URL, Use of redirection, Use of HTTPS, and Presence of suspicious words. In the Decision Tree algorithm model, 20% of the dataset was used for testing and 80% for training. Other models used 30% of the dataset for testing and 70% for training purposes.

2. Spearman Coefficient

We use Spearman Coefficient to measure the strength and direction of association between a pair of columns in the dataset. Additionally, it is critical to comprehend monotonic function in order to comprehend Spearman's rank correlation. A monotonic function is one that, when the independent variable changes, never either rises or decreases (Gupta, 2022).

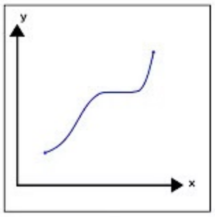


Figure 1

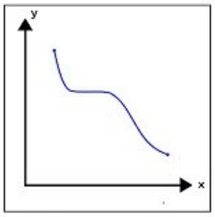


Figure 2

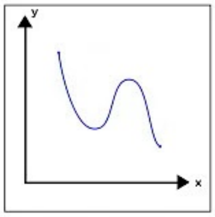


Figure 3

Figure 1: A monotonically increasing function: the variable X increases, and the variable Y never decreases

Figure 2: A monotonically decreasing function: the variable X increases, but the variable Y never increases.

Figure 3: A function that is not Monotonic: the X variable increases, the Y variable sometimes decreases, and sometimes increases.

The formula for Spearman Coefficient:

$$p = 1 - \frac{6\,\Sigma d_i^{\,2}}{n(n^2 - 1)}$$

Where p is the Spearman correlation coefficient, *di* is the difference between the paired ranks of each observation and n is the number of observations.

In fact, *P* will always be a value between -1 and 1. Hence, the link between the two variables is stronger the further p is from zero. If it is positive, the likelihood is for the other variable to increase as the first one does. If it's negative, the likelihood is for the other variable to decrease as the first one rises.

3. Decision Tree

Due to its simplicity and ease of visualization, Decision Tree is one of the most resilient and often used classification methods for classification tasks (Agarwal, S). A DT consists of terminal nodes that provide a benchmark for a single attribute and interior nodes that hold the goal value for an attribute. A DT is comparable to a flowchart in that each internal node applies a certain criterion to test an attribute and each branch represents the result of the test. For example, given a set of inputs, each input would recursively traverse the DT, and at each level, it would either be further categorized via further traversal or labeled.

The Decision Tree classifier for phishing detection underwent several important steps to achieve optimal performance. Firstly, the dataset was split into 20% for testing and 80% for training, ensuring that the model was trained on a sufficient amount of data. Secondly, we conducted data preprocessing to eliminate noise and extract relevant characteristics from the URLs, improving the model's accuracy. The decision tree was then simplified by selecting unique values for the features, resulting in a more interpretable model with fewer nodes, which reduced computational complexity during training. To assess the model's efficacy, we used accuracy scores and confusion matrices, which allowed us to identify potential issues and areas for improvement in the model's predictions.

Moreover, we applied pruning techniques to the decision tree to further optimize its performance. By selecting the alpha value that provides the best balance between classification accuracy and tree complexity, we were able to cut off branches that didn't improve the model's ability to classify new data accurately, resulting in a simplified and easier-to-understand tree that performs better when generalizing to new data. This pruning process also lowered the computational cost of analyzing the decision tree, as a smaller tree with key features required fewer computations to make predictions.

Overall, these methods helped develop a reliable and precise phishing detection model that can successfully spot potential phishing URLs.

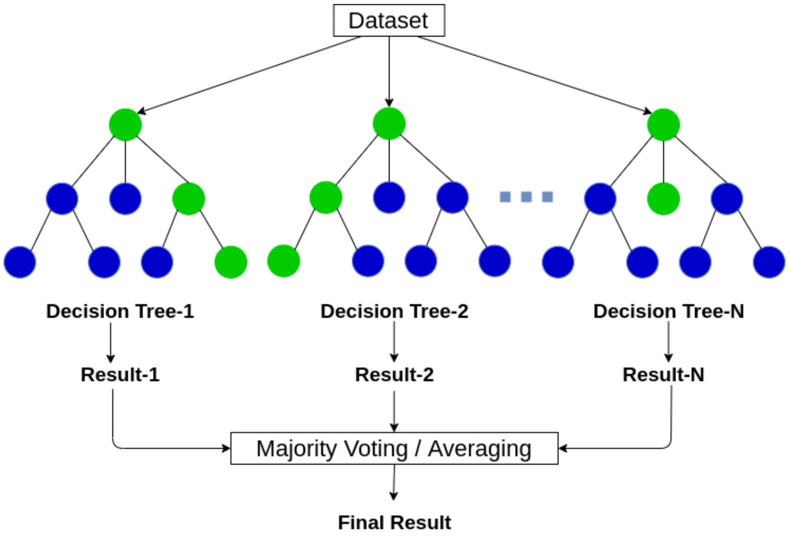
The formula for the Gini Index

$$Gini = 1 - \sum_{i=1}^{n} \quad (p_i)^2$$

Where p_i is the probability of an object being classified to a particular class

4. Random Forest Classification

Random forest is a useful classification and regression approach for problems requiring the categorization of data into classes. Despite the fact that random forest is a collection of decision trees, the manner in which they function varies considerably: Random forests, which are constructed from smaller data sets and deliver conclusions based on average or majority rating, avoid overfitting. Random forest constructs a decision tree from randomly selected observations, then calculates the average outcome. It utilizes no formulae and is slower than decision trees (R, Sruthi E, 2022).

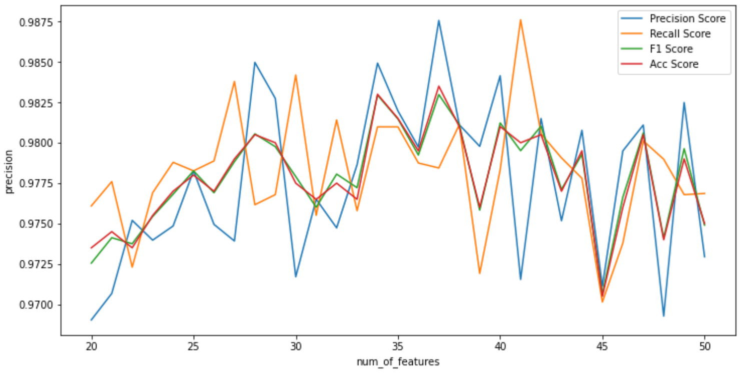


Applying Decision Trees in Random Forest:

To produce the necessary forecast, the random forest uses the bagging method (Mbaabu, 2020). Bagging requires using multiple samples of data (training data) as opposed to a single sample. Predictions are made using features and observations from a training dataset. Depending on the training data that the random forest algorithm receives, the decision trees generate a variety of results. The highest ranking of these outputs will be chosen as the final output (Mbaabu, 2020).

Proficient in Random Forest:

A random forest generates accurate predictions that are simple to comprehend. It is capable of both classification and regression tasks, and it can handle large datasets effectively. In comparison to the decision tree method, the random forest algorithm offers a higher level of accuracy in outcome prediction (Mbaabu, 2020).



5. Logistic Regression

Regression analysis is a component of logistic regression. Finding the relationship between a dependent variable (often referred to as the "Y" variable) and either one independent variable (the "X" variable) or a collection of independent variables is done using regression analysis, a sort of predictive modeling approach. Multiple regression is the process of predicting or explaining the result of the dependent variable using two or more independent variables (Thanda, Anamika, 2022). Linear regression and logistic regression are the two main categories of regression analysis.

On the basis of a number of independent factors, it is used to forecast a binary outcome. There are just two conceivable results in a binary outcome: either the event occurs (1) or it doesn't (0). Independent variables are those variables or elements that could have an impact on the result or dependent variable (Thanda, Anamika, 2022).

The formula for Logistics:

$$P = \frac{1}{1 + e^{-(\alpha + bX)}}$$

where e is the natural log’s base, a and b are the model's parameters

6. Web scraping

Web scraping is a technique used to extract data from websites. In the context of our phishing detection model, web scraping is used to gather URLs from both legal and phishing websites. The URLs are preprocessed to obtain the necessary features after being gathered through web scraping. This step is important as it gives meaningful information about the URLs that can be used to classify them as phishing or not phishing. The URLs are converted into a set of useful features using feature engineering and feature extraction methods.

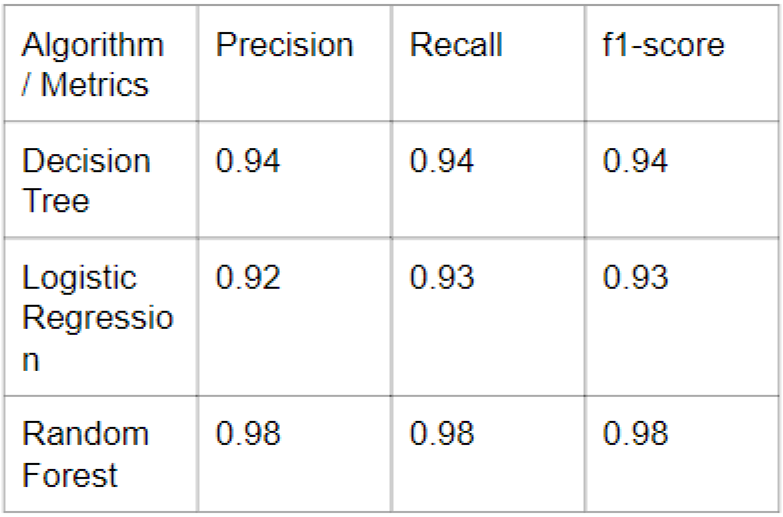
For instance, since many phishing URLs are housed on phony or dubious domains, the domain name of a URL is an important piece of information. In addition, since phishing URLs frequently have longer lengths than genuine ones, the length of the URL can also play a role. Additionally, the appearance of specific keywords like "login," "password," or "bank" can be a sign that a phishing effort has been made.

We used a total of 48 features, including domain-based features, URL-based features, and lexical features, in our phishing detection model. The URL's domain name, including the number of dots, the length, and the use of top-level domains, was one of the domain-based characteristics (TLDs). The length of the URL, the inclusion of symbols and digits, and the existence of subdomains or redirection were all considered URL-based characteristics. The inclusion of keywords or particular patterns in the URL, such as the use of "https" or "www," were considered lexically-based characteristics.

In the phishing detection model, feature extraction and modeling are critical phases because they assist in converting raw data into meaningful features that can be used to create a reliable classifier.

III. Result

Compare and contrast the three different models, we have learnt that Random Forest produces the most promising result. Something we should note about Random Forest is that the model produces very little to no overfitting unlike Decision Tree where overfitting is an inherited problem. Logistic regression has the lowest prediction accuracy out of the three models despite its inclination for overfitting. Decision Tree and Logistic Regression are both very simple and easy to understand which make it quick to implement. However both models have inherited problems. On the bright side, Logistic Regression does have a high performing recall. Decision Tree in our case is the middle child. It produces acceptable results with certain limitations. Decision Tree is extremely inclined against overfitting as noted earlier in the report, therefore, it's crucial that we prune the tree before producing the final model. Given its simplicity and computational advantages, we consider Decision Tree to be reliable in predicting phishing attempts given its limitations are dealt with.



Algorithm / Metrics	Precision	Recall	f1-score
Decision Tree	0.94	0.94	0.94
Logistic Regression	0.92	0.93	0.93
Random Forest	0.98	0.98	0.98

IV. Discussion

The lessons learnt are being documented and put to use in order to help business organizations strengthen their resilience and confidence by using phishing simulations and training. In this project, we are able to obtain an accuracy of more than 92%, recall greater than 93% with an f1 score less than 98%, in which our machine learning technique has proved its high-expected effectiveness.

Two out of three of our models relied on simplicity and quick computational runtime to predict the result. Random Forest does take us a step further in terms of computational runtime because it is somewhat more complicated as compared to the two other models. However, given that phishing attempts are getting incredibly advanced and much more difficult to detect, we would propose deep learning models such as Convolution Neural Networks, Long Short Term Memory Networks or the rising star, Transformer model. All of the mentioned have demonstrated exemplary performance in phishing detection.

The dataset played an important role in shaping our models in the early stages. For instance, with the Decision Tree model, since our features have little correlation with each other, the result was an incredibly complex and deep tree. Similarly, in the Logistic Regression model, the Spearman Correlation was unable to detect strong correlations between the features up until the last few features. With such in mind, we propose giving more attention to our dataset by preprocessing and only use the necessary features to reduce complexity during post-training. Furthermore, to have a more well-rounded detection model, it’s important that we do not limit the language to only English as phishing could come in any language.

V. Reference

Ampadu, Hyacinth. “Random Forests Understanding.” AI Pool, 1 May 2021, ai-pool.com/a/s/random-forests-understanding.

Agarwal, S. (n.d.). A Report on Decision Tree, Random Forest and Deep Forest. edu.authorcafe.com. Retrieved November 24, 2022, from https://edu.authorcafe.com/academies/7920/a-report-on-decision-tree-random-forest-and-deep-forest

IBM Cloud Education. “What Are Convolutional Neural Networks?” IBM, 20 Oct. 2020, www.ibm.com/cloud/learn/convolutional-neural-networks.

Ginni. “What Are the Approaches to Tree Pruning?” Tutorials Point, 22 Nov. 2021, www.tutorialspoint.com/what-are-the-approaches-to-tree-pruning.

Gupta, P. (2017, May 17). Decision Trees in Machine Learning - Towards Data Science. Medium. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Gupta, Aryan. “Spearman's Rank Correlation: The Definitive Guide to Understand: Simplilearn.” Simplilearn.com, Simplilearn, 15 Nov. 2022, www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation.

Lawton, George, et al. “What Is Logistic Regression? - Definition from Searchbusinessanalytics.” SearchBusinessAnalytics, TechTarget, 20 Jan. 2022, www.techtarget.com/searchbusinessanalytics/definition/logistic-regression#:~:text=The%20main%20advantage%20of%20logistic,the%20data%20are%20linearly%20separable.

Mbaabu, Onesmus. “Introduction to Random Forest in Machine Learning.” Section, 11 Dec. 2020, www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

Ronaghan, Stacey. “The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-Learn and Spark.” Medium, Towards Data Science, 1 Nov. 2019, towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3.

R, Sruthi E. “Random Forest: Introduction to Random Forest Algorithm.” Analytics Vidhya, 21 June 2022, www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.

Thanda, Anamika, et al. “What Is Logistic Regression? A Beginner's Guide [2022].” CareerFoundry, 4 Oct. 2022, careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/.

Whitney, Lance, et al. “Phishing Attack Spoofs US Department of Labor to Steal Account Credentials.” TechRepublic, 19 Jan. 2022, www.techrepublic.com/article/phishing-attack-spoofs-us-department-of-labor-to-steal-account-credentials/.

“Phishing Scams Linked to IDES on the Rise.” CBS News, CBS Interactive, 16 June 2021, www.cbsnews.com/chicago/news/phishing-scams-ides-unemployment-insurance/.