GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -103-111

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

MACHINE LEARNING FOR IMPROVED IDENTIFICATION OF FRAUDULENT INTERNET DOMAIN NAMES

Sai Dutta Abhishek Dash, Department of CSE, GIET University,
21cse144.saiduttaabhishekdash@giet.edu
Jyotirmayi Sadangi, Department of CSE, GIET University,
21cse686.jyotirmayisadangi@giet.edu
Jatin Kumar Pradhan , Department of CSE, GIET University,
21cse152.jatinkumarpradhan@giet.edu
Bidush Kumar Sahoo , Department of CSE, GIET University, bidushsahoo@giet.edu
Raghvendra Kumar, Department of CSE, GIET University, raghvendra@giet.edu

Abstract: This research introduces a new machine learning-based system for detecting phishing attempts. It aims to overcome common cybersecurity challenges, including the lack of real-world phishing examples, the limitations of traditional methods, and the constantly changing nature of new attack types like zero-day attacks. Our approach combines carefully designed features that allow us to effectively detect techniques used to hide malicious intent and prevent common evasion tactics such as URL shortening. Through extensive testing, we have shown significant improvements in our system's ability to identify phishing attempts, highlighting the importance of educating the public about these threats. We also believe in the importance of making resources readily available for further research in cybersecurity. By advocating for this and providing our solution, we hope to contribute to the development of stronger phishing detection methods and emphasize the need for education and accessible tools to protect against online dangers.

1. Introduction

In the digital age, phishing attempts provide a persistent and increasing menace. A staggering 220% surge in these attacks has occurred during the COVID-19 pandemic 1. Phishing attacks deceive people or organizations into disclosing sensitive information by using deceit and social engineering. Phishing attacks can have disastrous effects, including compromised personal data, financial loss, and reputational harm.

Despite efforts to stop phishing, it can be challenging to identify these assaults efficiently due to several issues:

- Sophisticated Techniques: Attackers conceal the genuine purpose of their phishing websites using sophisticated techniques, making them difficult to detect.
- User Vulnerabilities: Inadvertent activities by users who are ignorant of possible hazards frequently result in security breaches.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -103-111

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

- Limited Research Resources: It is difficult to investigate and create more effective detection techniques because there aren't many readily available instances of phishing websites.
- Reliance on Data Quality: Machine learning techniques exhibit potential in identifying phishing attempts; nevertheless, their efficacy is mostly contingent upon the Caliber of training data and the identification of pertinent attributes.

Our goal in this study is to tackle these issues by concentrating on two main areas:

- Feature engineering: To detect phishing attempts, we will meticulously design and choose a combination of lexical, domain-specific, and URL-based features.
- Model Selection: We will investigate how well a Random Forest classification model works to detect phishing websites with accuracy and minimal false positives.

By constructing an extensive dataset of phishing and authentic URLs for further studies, we want to further the field of phishing detection through our efforts.

- Determining the essential characteristics that are extremely predictive in differentiating between reputable and phishing websites.
- Offering countermeasures to attackers' cloaking and evasion tactics.
- Stressing the value of international cooperation and user education in the fight against phishing scams.

2. Related Work

The field of phishing detection using machine learning is an active area of research with a growing body of published studies. Here, we explore three key areas within this field:

1.**Feature Engineering:** The process of selecting and constructing relevant features to improve model performance.

2. **Algorithms and Performance**: The various machine learning algorithms used for phishing detection and their effectiveness.

3.**Ensemble Methods:** Techniques that combine predictions from multiple models to enhance overall system performance.

Feature Engineering

Effective feature engineering is crucial for accurate phishing detection. Research has shown the value of lexical features in classification tasks:

Machine Learning-Based Phishing URLs Detection with Lexical Features emphasizes the importance of lexical features in classification. To detect phishing URLs, the study highlights

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

characteristics including URL length, hyphenation, digit counts, subdomain analysis, and the existence of trusted top-level domains (TLDs).

We extend this methodology in our study by including these lexical features and investigating others, such as the existence of an explicit protocol (e.g., "http" or "https"). The objective is to improve our model's detecting skills even more.

2.1 Performance and Algorithms

Phishing detection relies heavily on machine learning methods. Two noteworthy studies have investigated various algorithmic techniques:

1. Shanigoz et al. [4] used Random Forest algorithm for detecting phishing websites, achieving a detection accuracy of 98.90%. This high accuracy rate indicates that the algorithm is effective for classification problems, specifically in detecting phishing URLs.
2. Nguyet Quang Do et al. [3] focused on deep learning algorithms for phishing detection in their study. While the specifics of their results are not provided in the context, it is inferred that they explored the possibilities of neural networks, like deep learning models, for detecting phishing attacks.

Our project's main goal is to assess Random Forest classifier performance utilizing an engineered feature set that was motivated by Sahingoz et al. Although we focus primarily on a single robust classifier, we also recognize that ensemble approaches may have the following advantages:

The paper "Phishing URLs Detection Using Machine Learning-Based Lexical Features" emphasizes the advantages of ensemble approaches, which integrate predictions from many models. Further research in this area is possible as it is a promising approach for improving the effectiveness of phishing detection systems. We hope to contribute to the ongoing work in creating efficient phishing detection technologies by comprehending these important areas and making use of findings from previous investigations.

3. Methodology

List(s) of Phishing Websites: taken from https://github.com/mitchellkrogza/Phishing.Database & https://openphish.org. To facilitate progressive benchmarking & iterative training, the dataset was split up into 2000 smaller datasets.

Valid URLs: Obtained from https://moz.com and many SEO ranking sites that provide a list of the top websites based on traffic. There were about 5000 valid domains in all.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -103-111

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

Test URLs: To ensure that the model's generalization performance is evaluated on unseen data, a separate test set of phishing and valid URLs was employed for evaluation.

3.1 Pre-processing

Iterative Setting Up:
The subsequent actions were performed for every training iteration:

1. To enable performance analysis as training data volume rose, a subset of the phishing dataset was loaded and mixed with the entire legitimate dataset.

2. The quantity of phishing samples was matched by oversampling the valid URL examples in each iteration to mitigate any potential class imbalance

3. To ensure compliance with the machine learning model, the categorical "trusted_tld" feature was encoded using a Label Encoder.

3.2 Feature Engineering

The subsequent characteristics were meticulously chosen based on their established ability to discriminate in phishing detection:

- Length of URL: Authentic URLs typically have more regular length patterns than phishing URLs, which may obfuscate their URLs with excessive length.
- Hyphen Count: To imitate real domain names or generate aesthetically bewildering subdomains, phishing URLs typically utilize hyphens.
- Digit Count: Abnormally high or low numbers in a URL may indicate phishing efforts.
- Subdomain Count: To conceal their actual destination, phishing URLs may use several subdomains.
- Trusted TLD: One way to identify reputable websites is to look for common trusted top-level domains (like.com and.org).
- Protocol existence: Since phishing URLs may omit the protocol, it is helpful to explicitly check for the existence of "http" or "https" to distinguish properly constructed URLs.

3.3 Model Architecture

Model Selection
Because of its computational efficiency, robustness against overfitting, and good handling of both numerical and categorical information, a Random Forest classifier was selected.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -103-111

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

Principles of Random Forests

The way Random Forests work is by building a group of decision trees, each of which is trained using a different subset of characteristics and data. These trees are used to aggregate predictions, which lowers variance and increases generalization.

Hyper-parameters

The hyperparameters listed below were applied:

* There are 100 estimators (trees) in total.
* Random state: 42 (to ensure repeatability).

The other hyperparameters were maintained at their initial settings.

3.4 Benchmarking Procedure

Iterative Training and Evaluation

Iterative training and evaluation of the model was done. Every iteration included the following:

1. Loading the complete legitimate dataset together with a portion of the phishing dataset.
2. Pre-processing the information (features encoding, oversampling).
3. Using this dataset to train the Random Forest model.

Evaluation Metrics

The following evaluation metrics being used to evaluate the model.

1. **Accuracy:** The percentage of correctly categorized instances out of all evaluated examples is measured by accuracy. It offers a general evaluation of how well the model is in distinguishing between phishing and authentic cases.

   Importance: Since accuracy provides a broad picture of the model's performance, it is essential. On the other hand, it might not be adequate on its own, particularly in unbalanced datasets where the proportion of phishing cases varies greatly from those of legal ones.
2. **Precision:** It refers to the proportion of accurately classified positive instances (phishing) in the model. It focuses on how relevant the phishing cases that have been found are.

   Importance: since it gauges how well the model prevents false positives. To prevent misclassifying valid occurrences as phishing efforts, high precision implies a low false positive rate.
3. **Recall:** The percentage of accurately categorized positive instances out of all real positive instances is measured by recall, which is sometimes referred to as sensitivity or true positive rate. It evaluates how well the model captures real-world cases of phishing.

   Importance: Recall is crucial since it assesses how well the model detects most phishing attacks while reducing false negatives. A high recall score means that a significant percentage of real phishing attempts are successfully captured by the model.
4. **Prediction Time:** The amount of time the model takes to predict new data instances is referred to as prediction time. In practical applications, it plays a crucial role, particularly in web-based settings where prompt identification of phishing efforts is necessary.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -103-111

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

Importance: Prediction time matters since it establishes how quickly the model can identify phishing attempts. A reduced prediction time allows for the quick detection and prevention of phishing attacks, improving user security and experience.

5. **F1-Score:** A balanced indicator of a model's performance, the F1-score is the harmonic mean of precision and recall. It provides a thorough assessment of the model's efficacy by integrating recall and precision into a single statistic.
   Importance: The F1-score offers a fair evaluation of the model's performance by considering both false positives and false negatives. When there is an imbalance between the positive and negative classes, it is especially helpful.

4. Results and Discussion

4.1. Performance Presentation

Accuracy vs. Data

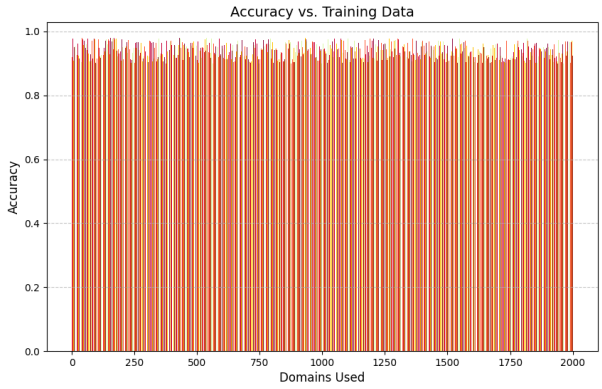

Figure 1: Accuracy vs Training Data

The graph in Figure-1 clearly shows that accuracy and the amount of training data have a strong positive relationship. Here are some key observations from the graph:

1. Accuracy increases rapidly until around 1000 phishing data files, reaching a level of 95%.
2. After this point, the rate of improvement in accuracy slows down, indicating a possible trend towards reaching a plateau.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -103-111

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

F1-Score vs. Data

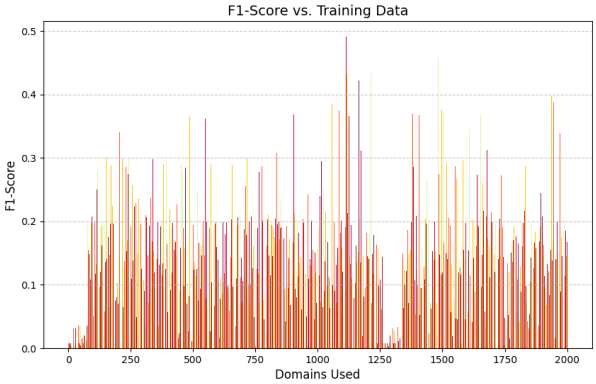

Figure 2: F1 score vs Training data

The Figure-2 graph shows that as the amount of training data increases, the F1-score also increases. This is like what we observed with accuracy. However, in addition to this overall trend, the graph also has some ups and downs, suggesting that the F1-score may be sensitive to specific variations within the phishing datasets.

Precision vs. Data

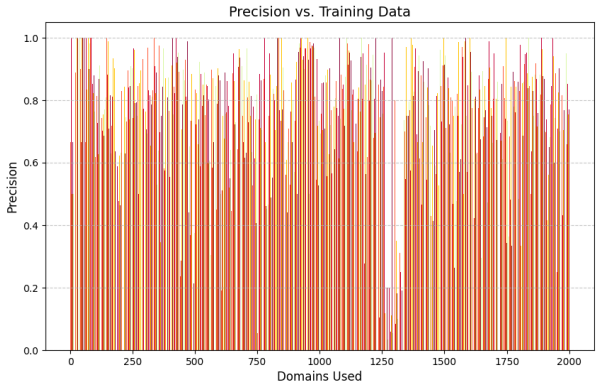

Figure 3: Precision vs Training Data

The Figure-3 graph exhibits a positive correlation with the accuracy and F1-score graphs, demonstrating a general upward trajectory as training data is incorporated. Nonetheless, sporadic fluctuations suggest that the integration of certain data points may have inadvertently introduced noise or posed challenges for the model, rendering it more vulnerable to false alarms.

Prediction Time vs. Data

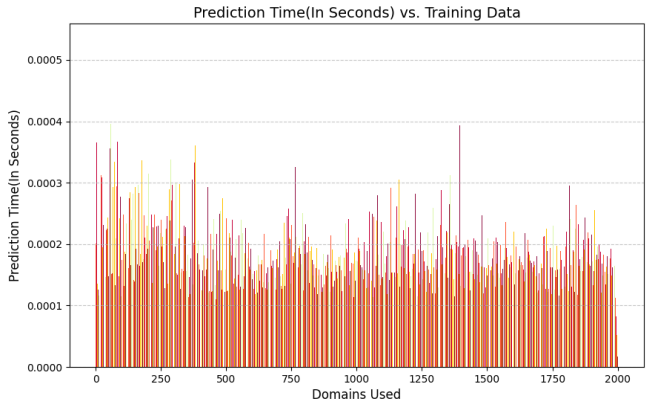

Figure 4: Prediction Time vs Training data

Interestingly, in Figure-4 the prediction time shows a generally decreasing trend with more training data, indicating improved prediction efficiency. However, there are fluctuations, and a potential plateau is observed with increasing dataset sizes.

GIET UNIVERSITY JOURNAL, 2024, VOL.3, NO.3, PP. -103-111

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

Recall vs. Data

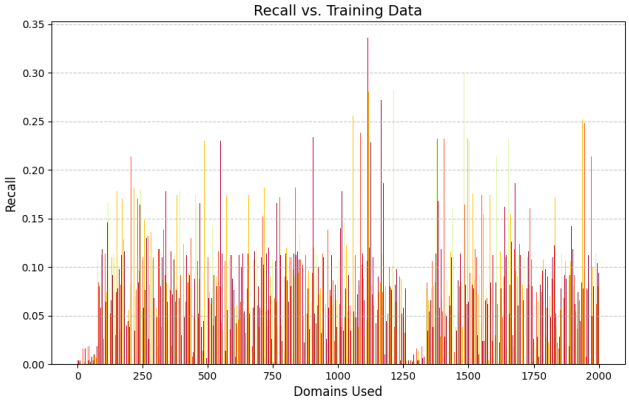

Figure 5: Recall vs training data

The recall graph in Figure-5 shows a clear upward trend as the amount of training data increases, indicating that the model is getting better at identifying phishing URLs with more exposure to different examples. This trend highlights the important idea that having a larger and more diverse dataset can greatly improve the model's ability to recall accurately. However, even though there is an overall trend, there are also noticeable changes in the graph, suggesting that the model may be affected by variations in the phishing datasets. These changes could be caused by different factors such as changes in the complexity or types of phishing attacks in the dataset.

5 Analysis

Data is Key

The positive trends in accuracy, F1-score, precision, and recall underscore the significance of large, balanced datasets for effective phishing detection. Additionally, the improvement in prediction time with more data signals that the model learns to make inferences more efficiently.

Diminishing Returns

Beyond around 1000 data files, performance gains slowdown in accuracy, F1-score, precision, and recall. Similarly, the improvement in prediction time plateaus, suggesting a potential saturation point with the current model and feature set.

Precision and Recall Fluctuations

Fluctuations in precision and recall graphs might indicate a need for addressing potential noise or inconsistencies in the data, highlighting the importance of data quality and pre-processing techniques.

Prediction Time vs. Performance

The observed trends suggest a possible trade-off between achieving the highest accuracy, precision, recall and the fastest prediction time, as gains in performance metrics seem to come with some fluctuations in prediction efficiency, especially with larger datasets.

7 Conclusion

Manuscript received 28th April, 2024; revised 3rd May, 2024; Accepted 6th May, 2024

The usefulness of machine learning—more especially, Random Forest classifiers—when combined with extensive feature engineering for phishing detection was examined in this study. The experimental results illustrated several important conclusions:

Data Correlation and Performance

The quantity of phishing data utilized for training the model positively correlated with its accuracy, precision, recall, and F1-score. Accuracy increased to 95% when 1500 phishing data files were included.

Reduction in Returns

Fascinatingly, after training on roughly 2500 data files, performance increased, but these metrics' pace of progress slowed down. This suggests that there might come a time when the present model and feature set are fully optimized possible, necessitating future modifications to achieve even greater benefits.

Prediction Time Considerations

The time it takes for the model to make predictions also needs to be considered. Fortunately, the prediction time only increased slightly as we used larger datasets, stabilizing after around 1000 domains. This shows that the model can handle larger amounts of data without sacrificing too much efficiency in making predictions.

8. References

1. R. Zieni, L. Massari, M. C. Calzarossa. "Phishing or not phishing? A survey on the detection of phishing websites." *IEEE Access* 11 (2023): pp.18499-18519.

2. I. Kara, M. Ok, A. Ozaday, “Characteristics of understanding URLs and domain names features: The detection of phishing websites with machine learning methods” ,*IEEE Access*, 10, 2022, pp.124420-124428.

3. N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, H. Fujita, “Deep learning for phishing detection: Taxonomy, current challenges and future directions”, *IEEE Access*, 10, 2022. 36429-36463.

4. O.K. Sahingoz, E. BubEr, E. Kugu, “DEPHIDES: Deep learning based phishing detection system”. *IEEE Access*, 12, 2024.8052-8070

5. M. Sánchez-Paniagua, E.F. Fernández, E. Alegre, W. Al-Nabki, V. González-Castro, “Phishing URL detection: A real-case scenario through login URLs”, *IEEE Access*, 10, 2022.42949-42960.