**Malicious Websites: A threat to modern network**

Hrishikesh Anil Patil

Department of OBIAS, University of Cincinnati

IS 8044: Projects in Information Systems Security Analytics

Truong (Jack) Luu, PhD

26 April 2024

**Malicious Websites, A threat to modern network**

In the era of ubiquitous internet connectivity, the world wide web has become an indispensable part of our daily lives, serving as a platform for communication, information exchange, and various online activities. However, alongside its numerous benefits, the internet has also given rise to malicious activities, posing significant threats to individuals, businesses, and organizations alike. Among these threats, malicious websites have emerged as a major concern, compromising user privacy, facilitating cybercrimes, and enabling the spread of malware, phishing scams, and other nefarious activities.

The ability to accurately identify and differentiate between benign and malicious websites has become crucial for maintaining a secure online environment. Malicious websites can be used as entry points for launching various types of cyber-attacks, such as distributing malware, stealing sensitive information, or carrying out financial fraud. These attacks can have devastating consequences, including data breaches, financial losses, and damage to reputation and credibility. Furthermore, the proliferation of malicious websites poses a significant threat to individuals, as unsuspecting users may inadvertently fall victim to phishing scams, drive-by downloads, or other malicious activities. Such incidents can lead to identity theft, financial fraud, and the compromise of personal data, causing immense distress and potential legal implications. In the corporate and organizational realm, the presence of malicious websites can result in data breaches, intellectual property theft, and disruptions to business operations, ultimately leading to substantial financial losses and reputational damage. Safeguarding against these threats is crucial for maintaining trust and ensuring the confidentiality, integrity, and availability of sensitive information.

Given the severe consequences of malicious website activities, developing effective methods for identifying and classifying these websites has become a critical endeavour for cybersecurity professionals, researchers, and organizations alike. By accurately detecting and mitigating the risks posed by malicious websites, we can enhance online safety, protect sensitive information, and foster a more secure and trustworthy digital environment.

The primary objective of this study is to develop an effective and reliable method for distinguishing between malicious and benign websites. By leveraging machine learning techniques, specifically exploratory data analysis (EDA) and regression analysis, we aim to uncover patterns and characteristics that can accurately classify websites into their respective categories.

Through an in-depth EDA, we seek to gain valuable insights into the dataset, exploring the distributions, correlations, and relationships among various features that may contribute to the identification of malicious websites. This analysis will provide a comprehensive understanding of the data, facilitating informed feature selection and preprocessing steps. Post that, we will employ regression analysis techniques to build predictive models capable of classifying websites as either malicious or benign based on their inherent characteristics. The study will evaluate and compare the performance of different regression algorithms, such as logistic regression, decision trees, or ensemble methods, to determine the most effective approach for this classification task.

The models developed through this study, leveraging exploratory data analysis and regression techniques, will play a pivotal role in enhancing signature detection capabilities for malicious websites. By accurately identifying the distinctive patterns and characteristics associated with malicious activities, these models will contribute to the creation of robust signatures or fingerprints. These signatures can then be integrated into security systems, enabling proactive monitoring, early detection, and timely mitigation of potential threats posed by malicious websites. Ultimately, the findings of this research will fortify the cybersecurity landscape, empowering organizations, and individuals to stay ahead of emerging threats and maintain a secure online presence.

The study utilizes a comprehensive dataset containing diverse features related to websites, such as URL characteristics, server information, Length, network traffic patterns, and DNS query times, aiming to capture a holistic representation for distinguishing malicious from benign websites. The approach involves an in-depth exploratory data analysis (EDA) to gain insights through statistical summaries, visualizations, and correlation analyses, followed by the application of regression analysis techniques like logistic regression, classifiers, or ensemble methods. Feature selection, model training, cross-validation, and performance evaluation using metrics like accuracy and F1-score will be employed to develop robust predictive models capable of accurately classifying websites based on their inherent characteristics.

**Literature Review**

Detecting malicious websites has been an active area of research in the cybersecurity domain, with numerous techniques and approaches proposed over the years. Traditional methods often relied on blacklists or whitelists, which maintain a database of known malicious or benign URLs, respectively. However, these methods are reactive and struggle to keep up with the ever-evolving landscape of cyber threats. More recent research has focused on leveraging machine learning techniques to proactively identify malicious websites. Approaches such as support vector machines (SVMs), decision trees, and neural networks have been employed to analyse various features extracted from websites, including content-based features (e.g., HTML structure, JavaScript code), network-based features (e.g., DNS queries, network traffic patterns), and URL-based features (e.g., length, character distribution).

Researchers have explored the use of static and dynamic analysis techniques to extract relevant features from websites. Static analysis involves examining the website's code and structure without executing it, while dynamic analysis involves running the website in a controlled environment and monitoring its behaviour. Both approaches have their advantages and limitations, and hybrid techniques combining static and dynamic analysis have also been proposed. Despite the advancements in the field, several gaps and limitations persist. Many existing studies focus on specific subsets of features or rely on labelled datasets that may not accurately represent the diversity and complexity of real-world malicious websites. Additionally, the constantly evolving nature of cyber threats demands continuous adaptation and refinement of detection techniques.

This study aims to address some of these limitations by leveraging a comprehensive dataset that encompasses a diverse range of features, including content-based, network-based, and URL-based characteristics. By employing exploratory data analysis and advanced regression techniques, the study seeks to uncover previously undiscovered patterns and relationships that can improve the accuracy and robustness of malicious website detection. Furthermore, the study will investigate the effectiveness of different regression algorithms and feature selection strategies, aiming to develop a generalized approach that can adapt to the ever-changing landscape of malicious websites. By addressing these gaps, the study contributes to the ongoing efforts in enhancing cybersecurity measures and fostering a safer online environment.

The regression models trained in this study will subsequently be employed for signature detection of malicious websites. By accurately identifying the distinctive patterns and characteristics associated with malicious activities, these models will contribute to the creation of robust signatures or fingerprints. These signatures can then be integrated into security systems, enabling proactive monitoring, early detection, and timely mitigation of potential threats posed by malicious websites. (Check **References** for List of studies took into consideration)

**Description of Dataset:**

*Source of the data:* The dataset was obtained from Kaggle, a platform that hosts data science and machine learning datasets contributed by various individuals and organizations. The dataset was originally curated and uploaded to Kaggle by Christian Urcuqui. It was collected and compiled 6 years ago. Even though our dataset is 6 years old, it still provides very good reference point of how malicious websites behave today.

*Number of instances:* 1781 rows, 20 columns

*Features or attributes present in the data:*

|  |  |
| --- | --- |
| *Column* | *Description* |
| URL\_LENGTH | The length of the URL |
| NUMBER\_SPECIAL\_CHARACTERS | The number of special characters present in the URL |
| CHARSET | Categorical value and its meaning is the character encoding standard (also called character set). |
| SERVER | Categorical value and its meaning is the operative system of the server got from the packet response |
| CONTENT\_LENGTH | it represents the content size of the HTTP header |
| APP\_BYTES | Number of bytes transferred. |
| DNS\_QUERY\_TIMES | The number of DNS packets generated during the communication between the honeypot and the server |
| TYPE | Categorical variable, its values represent the type of web page analysed, specifically, 1 is for malicious websites and 0 is for benign websites |

*Class distribution (malicious vs. benign):*

Malicious websites: 216 (12%)

Benign websites: 1565 (88%)

**Exploratory Data Analysis (EDA)**

*Data Preprocessing - Handling missing values*: The dataset contained missing values in several features, including 'CONTENT\_LENGTH', 'SERVER', 'CHARSET', and 'DNS\_QUERY\_TIMES'. Given the relatively small size of the dataset (1800 rows), instead of removing instances with missing values, several imputation techniques were employed to retain the maximum number of rows. For the 'CONTENT\_LENGTH' feature, missing values were imputed using the K-Nearest Neighbors (KNN) imputation technique. For the 'SERVER' feature, missing values were imputed based on the most frequent 'SERVER' value for each unique 'CHARSET' value, leveraging the relationship between these two features. Missing values in the 'CHARSET' feature were imputed with the mode value of the 'CHARSET' column. Lastly, for the 'DNS\_QUERY\_TIMES' feature, missing values were imputed with the mean value of the column.

Before building predictive models, we thoroughly explored the dataset to uncover patterns and relationships that could aid in distinguishing malicious from benign websites. This exploratory phase provided valuable insights guiding our subsequent approaches.

We calculated summary statistics and created feature distributions, identify outliers, and explore potential correlations. Correlation coefficients quantified relationships between features. To investigate specific questions, like the link between server types and website maliciousness, we used bar charts, and statistical tests like chi-square. We also employed outlier detection methods to ensure anomalous values didn't adversely affect modelling. Insights from this thorough EDA informed our data preprocessing, including feature selection, handling missing values, and transformations. These steps prepared the dataset for effective regression modelling to distinguish malicious and benign websites.

*How does URL Length Affect Malicious Weather?* The box plot of URL length by type reveals that malicious websites (Type 1) tend to have longer URLs on average compared to benign websites (Type 0). While there is considerable overlap between the two categories, the presence of outliers, particularly in the benign category, indicates that some benign websites also feature longer URL lengths. However, the median URL length for malicious websites is higher, which suggests that URL length could be a distinguishing feature when analysing website security. This finding aligns with the hypothesis that malicious URLs may use longer lengths to obfuscate dubious content or mimic legitimate websites. (Refer **Exhibit 2** for Box Plot of URL Length of Malicious vs Non-Malicious Websites.)

*How does the type of server correlate with the likelihood of a website being malicious?* In our statistical analysis to determine the relationship between server types and the likelihood of a website being malicious, we conducted a chi-square test. The test yielded a chi-square statistic of 408.08 and a p-value of 3.24e-11, indicating a significant association. This suggests that server type could be a key indicator of malicious websites.



*Is there a significant difference in DNS query times between malicious and benign websites?* Our investigation into DNS query time differences between malicious and benign websites involved normality and variance checks, followed by a t-test. Despite both categories significantly deviating from normality (malicious p ≈ 3.12e-15, benign p ≈ 2.03e-43), the Levene's test suggested equal variances (p ≈ 0.40), validating our use of a standard t-test. The t-test resulted in a statistic of 2.91 and a p-value of approximately 0.0037, indicating a statistically significant difference in DNS query times between the two groups. This points to DNS query times as a potential feature of interest for distinguishing between malicious and benign websites.

A black text on a white background

Description automatically generated

*Do malicious websites typically have longer URL lengths compared to benign websites?* In our analysis of URL lengths, a t-test was conducted to compare the lengths between malicious and benign websites. The results yielded a t-test statistic of 6.93, with a highly significant p-value of approximately 5.91e-12. These findings robustly suggest that malicious websites tend to have longer URLs than benign ones, supporting the notion that URL length could be a predictive characteristic in cybersecurity analysis.

A black text on a white background

Description automatically generated

*Do malicious websites typically have longer URL lengths compared to benign websites?* In our examination of application byte size, we observed that benign websites had a higher mean of application bytes (approximately 3216.8 bytes) compared to malicious websites (around 1283.6 bytes). To further analyse this difference, we applied a log transformation to the application byte feature to normalize the data and reduce the impact of outliers. The resulting boxplot of the log-transformed application bytes by website type visually underscores this disparity, with benign websites generally showing higher values. This suggests that benign websites tend to engage in more data exchange, whereas malicious websites may be more conservative in the amount of data they transfer, possibly to remain undetected or due to simpler malicious payloads.



**Regression Analysis**

*Can a predictive model accurately classify websites as malicious or benign based on available features?* A core question driving our investigation was the model's ability to accurately classify websites as malicious or benign based on available features. The Random Forest algorithm, known for its robustness and high accuracy in classification tasks, was chosen to construct the predictive model.

*Model Development and Validation:* Our dataset, after preprocessing, consisted of various numeric and categorical featuresranging from URL length to server type, which were encoded appropriately for model input. We divided our dataset into a training set, which constituted 70% of the data, and a testing set, which made up the remaining 30%. This split was crucial in validating the model's predictions against previously unseen data, ensuring the robustness of our findings.

We implemented standardization on the dataset to scale the features, addressing potential bias due to differing ranges. This process is particularly important for algorithms that operate based on distances or assume normality, ensuring that each feature contributes equally to the outcome.

*Model Evaluation*

Upon training, the model was subjected to performance evaluation using the test set. The classification report unveiled an overall accuracy of 96%, with precision scores for the benign and malicious classes standing at 96%. This high precision indicates that the model's predictions were reliable and had few false positives. However, the recall for the malicious class was 71%, suggesting that while the model was precise, it missed some malicious websites, marking them as benign. This is reflected in the F1-score for the malicious class, which, at 0.82, shows room for improvement in balancing precision and recall.

The confusion matrix provided further insights, revealing that out of 535 websites, the model correctly identified 460 as benign and 52 as malicious. Notably, 21 malicious websites were misclassified as benign, highlighting a potential area for model improvement as these false negatives are critical in the context of website security.

A screenshot of a computer screen

Description automatically generated

*Feature Importance Analysis*

Understanding the features that influenced the model's decisions can provide valuable insights into the nature of malicious websites. Our analysis revealed that the amount of data sent by the website (SOURCE\_APP\_BYTES) and the diversity of remote TCP ports used (DIST\_REMOTE\_TCP\_PORT) were the most predictive features. The presence of specific server types and the complexity of the URL, indicated by the number of special characters, also played significant roles. Interestingly, the CONTENT\_LENGTH and the size of the application data (APP\_BYTES) were among the top features, suggesting that the content and data exchange patterns are markers of website intent.

A white screen shot of a computer

Description automatically generated

The results affirm the potential of machine learning model to distinguish between malicious and benign websites based on key characteristics. The high precision of our model signifies its utility in scenarios where the cost of false positives is high. However, the model's tendency to miss certain malicious sites as evidenced by the lower recall necessitates further refinement to improve its sensitivity to malicious activities.

This study underscores the complexity of malicious website detection and the necessity for balanced datasets that encompass the nuanced behaviours of web threats. The insights drawn from feature importance highlight opportunities for enhanced feature engineering and the need for domain-specific knowledge to identify subtle indicators of malicious intent.

**Conclusions**

*Predictive Modelling*

In conclusion, while our predictive model demonstrates a high degree of accuracy and precision, it also highlights the intrinsic challenges in cyber threat detection — particularly the need for balanced sensitivity to all classes. Future work may explore hybrid models or advanced ensemble techniques to elevate the recall for malicious websites without compromising on precision. Additionally, investigating underrepresented patterns in the data and incorporating more sophisticated features, such as those derived from website content analysis or behaviour-based indicators, could lead to further advancements in the field.

*Comparative Analysis of Random Forest and Ensemble Models in Website Classification*

In our study, we also evaluated the effectiveness of a standalone Random Forest model and compared it with an ensemble model that integrates Logistic Regression, Random Forest, and SVC. The objective was to assess whether combining different algorithms could enhance the predictive accuracy and reliability over using a single model approach.

*Random Forest Model Performance*

The Random Forest model demonstrated robust overall accuracy, achieving a score of 96%. This model was particularly effective in identifying benign websites with a precision of 96% and almost perfect recall, indicating a strong capability in correctly classifying non-malicious websites. However, while the model also showed high precision for malicious websites, its recall of 71% suggested room for improvement, as approximately 29% of malicious websites were not correctly identified.

*Ensemble Model Performance*

The ensemble model, employing soft voting to integrate the predictions of Logistic Regression, Random Forest, and SVC, achieved an accuracy of approximately 93.6%. This model excelled in precision, particularly for malicious websites, achieving a perfect score of 100%, thus eliminating false positives entirely. However, its recall for malicious sites was lower than that of the Random Forest model, at 53%. This indicates that while the ensemble model is extremely reliable when it flags a site as malicious, it fails to detect a significant portion of actual malicious websites.

*Analysis and Implications*

The comparison highlights the trade-offs between different modelling approaches. The Random Forest model, while slightly less accurate overall compared to the ensemble, provided a better balance between precision, and recall for malicious websites. In contrast, the ensemble model, despite its perfect precision, sacrificed recall, suggesting that it could be overly conservative in labelling websites as malicious.

This distinction is crucial in practical applications. For environments where the cost of missing a malicious website is high, the Random Forest's higher recall might be preferable. Conversely, in scenarios where false positives — incorrectly labelling benign sites as malicious — are more disruptive, the ensemble's perfect precision would be advantageous.

*Comparative Analysis*

Our findings illustrate the complexity of model selection in predictive analytics for cybersecurity. While ensemble methods can enhance decision accuracy through diverse approaches, they may also introduce new challenges, such as decreased sensitivity (recall). Future work might explore adaptive techniques that dynamically adjust the decision thresholds based on the operational context or develop hybrid models that further refine the balance between precision and recall.

**References**

Eshete, B., Villafiorita, A. and Komminist Weldemariam (2013). BINSPECT: Holistic Analysis and Detection of Malicious Web Pages. Springer eBooks, [online] pp.149–166. doi: <https://doi.org/10.1007/978-3-642-36883-7_10.>

Ma, J., Saul, L., Savage, S. and Voelker, G. (n.d.). Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. [online] Available at: <https://cseweb.ucsd.edu/~jtma/papers/beyondbl-kdd2009.pdf.>

Canali, D., Cova, M., Vigna, G. and Kruegel, C. (n.d.). Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages. [online] Available at: <https://sites.cs.ucsb.edu/~chris/research/doc/www11_prophiler.pdf.>

Cova, M., Kruegel, C. and Vigna, G. (n.d.). Detection and Analysis of Drive-by-Download Attacks and Malicious JavaScript Code. [online] Available at: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=eb5a092458d50ad1119ca0ef10e416db24235e9b>

Rieck, K., Holz, T., Willems, C. and Pavel Laskov (2008). Learning and Classification of Malware Behavior. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/28358465_Learning_and_Classification_of_Malware_Behavior>

Hason, N., Amit Dvir and Chen Hajaj (2020). Robust Malicious Domain Detection. [online] ResearchGate. Available at: <https://www.researchgate.net/publication/342431428_Robust_Malicious_Domain_Detection>

Gopinath P, Sangeetha S, Rajendran, B. and Bapu Bindhumadhava (2020). Malicious Domain Detection Using Machine Learning On Domain Name Features, Host-Based Features and... [online] ResearchGate. Available at: <https://www.researchgate.net/publication/341907043_Malicious_Domain_Detection_Using_Machine_Learning_On_Domain_Name_Features_Host-Based_Features_and_Web-Based_Features>

Abubakr Sirageldin, Baharudin, B.B. and Low Tang Jung (2014). Malicious Web Page Detection: A Machine Learning Approach. Lecture notes in electrical engineering, [online] pp.217–224. doi:<https://doi.org/10.1007/978-3-642-41674-3_32.>

Mindrila, D. and Balentyne, P. (2013). *The Chi Square Test*. [online] Available at: <https://www.westga.edu/academics/research/vrc/assets/docs/ChiSquareTest_LectureNotes.pdf>

Investopedia. (2024). *P-Value: What It Is, How to Calculate It, and Why It Matters*. [online] Available at: <https://www.investopedia.com/terms/p/p-value.asp>

Kirasich, K., Smith, T. and Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*, [online] 1(3), p.9. Available at: <https://scholar.smu.edu/cgi/viewcontent.cgi?params=/context/datasciencereview/article/1041/&path_info=Report.pdf>

Dietterich, T. (n.d.). *Ensemble Methods in Machine Learning*. [online] Available at: <https://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf>

West, R.M. (2021). Best practice in statistics: The use of log transformation. *Annals of clinical biochemistry*, [online] 59(3), pp.162–165. doi: <https://doi.org/10.1177/00045632211050531>

Nist.gov. (2024). *1.3.5.10. Levene Test for Equality of Variances*. [online] Available at: <https://itl.nist.gov/div898/handbook/eda/section3/eda35a.htm#:~:text=Levene's%20test%20(%20Levene%201960)%20is,used%20to%20verify%20that%20assumption>

‌

‌

‌

‌

‌

‌

**Exhibit 1A**: Post Logarithmic Transformation of Application Bytes for Malicious Websites

A graph of different colored squares

Description automatically generated with medium confidence

**Exhibit 2**: Box Plot of URL Length for Malicious Websites (Type = 1)

A graph of a box plot

Description automatically generated

**Exhibit 3**: Some visualizations with Dataset.

A graph of blue rectangular bars

Description automatically generated with medium confidence

A graph of a bar graph

Description automatically generated with medium confidence