Developing a Machine Learning Model for Malicious URL Detection

S Sri[illegible]ri (ENG21CS344) , Priyanka Datta(ENG21CS0309),
Department of Computer Science and Engineering,
Dayananda Sagar University, Bengaluru

Abstract: The proliferation of malicious URLs poses a significant threat to cybersecurity, leading to financial losses and compromised data security. This research endeavors to develop robust mechanisms for early detection and prevention of such threats through the utilization of machine learning techniques. A meticulously curated dataset comprising a diverse range of malicious URLs is employed to train and evaluate various machine learning models. The dataset encompasses instances of spam, phishing, drive-by downloads, and other forms of cyber threats, providing a comprehensive representation of the challenges faced in cybersecurity. Methodologies for dataset construction, feature selection, and labeling criteria are rigorously discussed, ensuring the quality and relevance of the data utilized for model development. Furthermore, a comparative analysis of multiple machine learning algorithms, including RandomForest, DecisionTree, XGBoost, and ExtraTrees classifiers, is conducted to evaluate their efficacy in detecting malicious URLs. Evaluation metrics such as accuracy, precision, recall, and F1 score are employed to assess the performance of each model. The study highlights the importance of feature engineering, including character occurrences, URL structure anomalies, HTTPS usage, presence of IP addresses, and URL shortening, in improving the predictive capabilities of the models. Additionally, the research presents a method for model serialization using pickle, enabling the deployment of trained models for real-time URL classification. Ultimately, this research contributes to the advancement of cybersecurity defenses by providing practitioners with reliable tools for identifying and mitigating malicious URLs proactively. By leveraging machine learning algorithms and a comprehensive dataset, the proposed approach aims to bolster the resilience of digital ecosystems against evolving cyber threats, thereby safeguarding users' online experiences and organizational assets.

Keywords: Malicious URLs, Cybersecurity, Machine Learning, Feature Engineering, Classification Algorithms, Evaluation Metrics, URL Detection, Phishing, Spam.

1. Introduction

In today's digitally interconnected world, the proliferation of malicious URLs poses a pervasive and ever-evolving threat to cybersecurity, challenging individuals, organizations, and society at large. From phishing scams to malware dissemination and financial fraud, malicious URLs serve as conduits for a myriad of cyberattacks, resulting in substantial financial losses, compromised privacy, and reputational damage. As cybercriminals continuously devise new techniques and exploit vulnerabilities in online systems, the need for proactive detection and mitigation of these threats has become paramount. In response to this imperative, the integration of machine learning techniques offers promising avenues for enhancing cybersecurity defenses by enabling automated analysis, detection, and response to malicious URLs in real-time.

At the heart of this endeavor lies the development of a machine learning-based model specifically tailored for the detection and classification of malicious URLs. This ambitious project seeks to harness the power of advanced algorithms and data-driven methodologies to accurately identify and categorize malicious URLs, thereby mitigating cybersecurity risks and safeguarding digital ecosystems. The project's scope encompasses the creation of a comprehensive dataset comprising diverse samples of malicious URLs, meticulously curated to represent various threat categories and attack vectors prevalent in the cybersecurity landscape.

A key aspect of the project involves feature engineering, wherein relevant information is extracted from URLs to enable effective discrimination between benign and malicious entities. Features such as character occurrences, URL structure anomalies, presence of IP addresses, HTTPS usage, and URL shortening are among the critical indicators leveraged by the machine learning models to discern malicious intent and identify potential security threats. Moreover, the integration of domain-specific knowledge and expertise into the feature selection process enhances the models' ability to adapt to evolving cyber threats and minimize false positives.

To achieve its objectives, the project entails the exploration and evaluation of multiple machine learning algorithms, including but not limited to RandomForest, DecisionTree, XGBoost, and ExtraTrees classifiers. Through rigorous experimentation and performance evaluation, the project aims to identify the most effective model for malicious URL detection, leveraging metrics such as accuracy, precision, recall, and F1 score to assess

model efficacy and generalizability. Additionally, the project involves the serialization of trained machine learning models using the pickle library, enabling their deployment for real-time URL classification and threat detection in operational cybersecurity environments.

The interdisciplinary nature of the project underscores its significance in advancing the state-of-the-art in cybersecurity by integrating machine learning techniques, domain expertise, and high-quality datasets. By developing robust and scalable models capable of accurately identifying and mitigating cyber threats, the project contributes to the enhancement of cybersecurity defenses and the protection of critical infrastructure against evolving malicious activities. Furthermore, the project's outcomes have implications beyond academia, as they hold the potential to inform and guide the development of practical cybersecurity solutions and frameworks deployed by organizations and cybersecurity practitioners worldwide.

In summary, the project represents a holistic and multifaceted approach to addressing the pervasive threat of malicious URLs in the digital age. By leveraging machine learning algorithms, advanced feature engineering techniques, and domain expertise, the project aims to empower cybersecurity professionals with effective tools and methodologies for detecting, analyzing, and responding to malicious URLs, thereby fostering a safer and more secure online environment for users and organizations alike.

2. Related Works

Related works in malicious URL detection encompass a diverse array of approaches and methodologies, reflecting the multifaceted nature of cybersecurity challenges. Research papers, academic studies, and industry reports offer valuable insights and contributions to advancing detection techniques. Chen et al. explore machine learning for detecting malicious URLs in social media, emphasizing contextual information and user behavior. Jagatic et al.'s survey provides an overview of phishing detection techniques, discussing both rule-based and machine learning methods. Wang et al. present a machine learning-based approach for phishing URL classification, leveraging lexical, structural, and behavioral features. Liu et al. compare feature engineering techniques' effectiveness in malicious URL detection, focusing on selection, reduction, and transformation methods. Rajasegarar et al. delve into deep learning, specifically recurrent neural networks, achieving state-of-the-art performance in malicious URL detection. Zhang et al. investigate ensemble learning for detection, combining multiple classifiers to improve accuracy and robustness. Li et al. explore anomaly-based detection using unsupervised learning, employing clustering and outlier detection to identify suspicious URL activity. These works collectively contribute to the field, offering insights and methodologies for researchers, practitioners, and policymakers to develop effective solutions for detecting and mitigating malicious URLs

3. Proposed Work

The intended project seeks to advance malicious URL detection through the development of a machine learning-based model tailored for accurately identifying and categorizing malicious URLs. Leveraging a meticulously curated dataset encompassing diverse samples of malicious URLs, the project will extract relevant information from URLs and enhance the models' discriminative power. By evaluating multiple machine learning algorithms, including RandomForest, DecisionTree, XGBoost, and ExtraTrees classifiers, the project seeks to identify the most effective model for malicious URL detection. Furthermore, the project entails the serialization of trained machine learning models using the pickle library, enabling their deployment for real-time URL classification and threat detection in operational cybersecurity environments.

3.1 First Phase: Data Collection and Preprocessing

The initial phase of the project will focus on acquiring a comprehensive dataset of URLs encompassing both benign and malicious examples. This dataset will serve as the foundation for training and evaluating the machine learning models. Various sources, including cybersecurity repositories, threat intelligence feeds, and public datasets, will be explored to gather a diverse range of URL samples spanning different threat categories such as phishing, malware, and drive-by downloads. Following data acquisition, extensive preprocessing steps will be undertaken to ensure the quality and integrity of the dataset. This includes data cleaning to remove duplicates, irrelevant entries, and

inconsistencies, as well as normalization techniques to standardize the format and structure of URLs. Additionally, each URL will be labeled according to its threat category, facilitating supervised learning in subsequent phases.

3.2 Second Phase: Feature Engineering and Model Development

In the second phase, the focus will shift to feature engineering and model development. Feature extraction methods will be devised to transform raw URL data into informative feature vectors that capture relevant characteristics indicative of malicious intent. This involves analysing the structure, syntax, and content of URLs to identify discriminative features such as character occurrences, presence of IP addresses, HTTPS usage, URL shortening, and lexical patterns. Leveraging domain-specific knowledge and expertise, a comprehensive set of features will be curated to provide rich contextual information for training the machine learning models. Subsequently, multiple machine learning algorithms, including RandomForest, DecisionTree, XGBoost, and ExtraTrees classifiers, will be implemented and trained using the curated dataset and extracted features. Hyperparameters will be optimized, and model performance will be evaluated using appropriate metrics such as accuracy, precision, recall, and F1 score to determine the most effective model for malicious URL detection

- **RandomForest Classifier:** RandomForest is a method in machine learning that uses multiple decision trees to create a strong model. Each tree is trained separately on a part of the dataset, and the final prediction is based on the combined predictions of all trees. It's great for handling data with many dimensions and complex relationships, making it ideal for tasks like classifying URLs. RandomForest is especially good at finding intricate patterns and anomalies in data, which helps in accurately spotting malicious URLs. It also comes with tools to analyze which features are most important for classification, giving researchers valuable insights.

- **DecisionTree Classifier:** The DecisionTree classifier is a machine learning algorithm that builds a tree-like structure to represent the decision-making process based on input features. At each node of the tree, the classifier selects the feature that best splits the data into homogeneous subsets, thereby maximizing information gain or minimizing impurity. DecisionTree classifiers are highly interpretable and intuitive, making them suitable for understanding the underlying logic behind classification decisions. However, they are prone to overfitting, especially when dealing with high-dimensional or noisy data. Nonetheless, DecisionTree classifiers offer a fast and efficient solution for URL classification tasks, particularly in scenarios where interpretability is paramount.

- **XGBoost Classifier:** The XGBoost classifier is an advanced implementation of gradient boosting, a machine learning technique that builds a sequence of decision trees iteratively to minimize a predefined loss function. XGBoost is renowned for its scalability, efficiency, and superior performance in a wide range of classification tasks, including URL classification. By incorporating regularization techniques and optimized tree construction algorithms, XGBoost effectively mitigates overfitting and improves generalization accuracy. It also provides native support for handling missing values and categorical features, enhancing its versatility and applicability to real-world datasets. XGBoost's ensemble learning approach enables it to capture intricate relationships and nonlinear dependencies in the data, making it a popular choice for achieving state-of-the-art results in malicious URL detection.

- **ExtraTrees Classifier:** The ExtraTrees classifier, short for Extremely Randomized Trees, is a variant of the RandomForest algorithm that introduces additional randomness during the tree-building process. Unlike

traditional RandomForest, which selects the best split among a subset of features at each node, ExtraTrees randomly selects splits without considering feature importance. This randomization strategy helps prevent overfitting and reduces variance, making ExtraTrees robust to noisy or high-dimensional data. Additionally, ExtraTrees can be trained efficiently in parallel, making it suitable for large-scale URL classification tasks. Despite its simplicity, ExtraTrees often achieves competitive performance compared to more complex models, making it a valuable asset in the ensemble of classifiers used for detecting malicious URLs.

4. Working of the Project

Upon launching the application, users are greeted with a straightforward user interface consisting of a text input field where they can enter the URL they wish to check. After inputting the URL, users simply click the "Check" button.

Behind the scenes, the application utilizes a pre-trained Random Forest model, which has been previously trained on a dataset containing features extracted from URLs. These features include counts of specific characters, the presence of certain patterns (such as HTTP/HTTPS, IP addresses), and other indicative factors of malicious intent.

Once the "Check" button is clicked, the application retrieves the entered URL and extracts relevant features using a feature extraction function. This function processes the URL, extracting key characteristics that can be indicative of malicious behavior.

Subsequently, the application employs the loaded Random Forest model to make predictions based on the extracted features. The model predicts whether the input URL is likely to be benign or malicious.

Finally, the application updates the user interface to display the prediction result. If the model predicts that the URL is benign, the application informs the user that the URL is "Not malicious." Conversely, if the model predicts that the URL is malicious, the application notifies the user that the URL is indeed "Malicious."

Figure 1: Flowchart

5. Results and Discussion

5.1 Dataset

This dataset compiles 651,191 website links. Of these, 428,103 links are considered safe, while 96,457 are labeled as defacement, 94,111 as phishing, and 32,520 as malware. Link to dataset https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset.

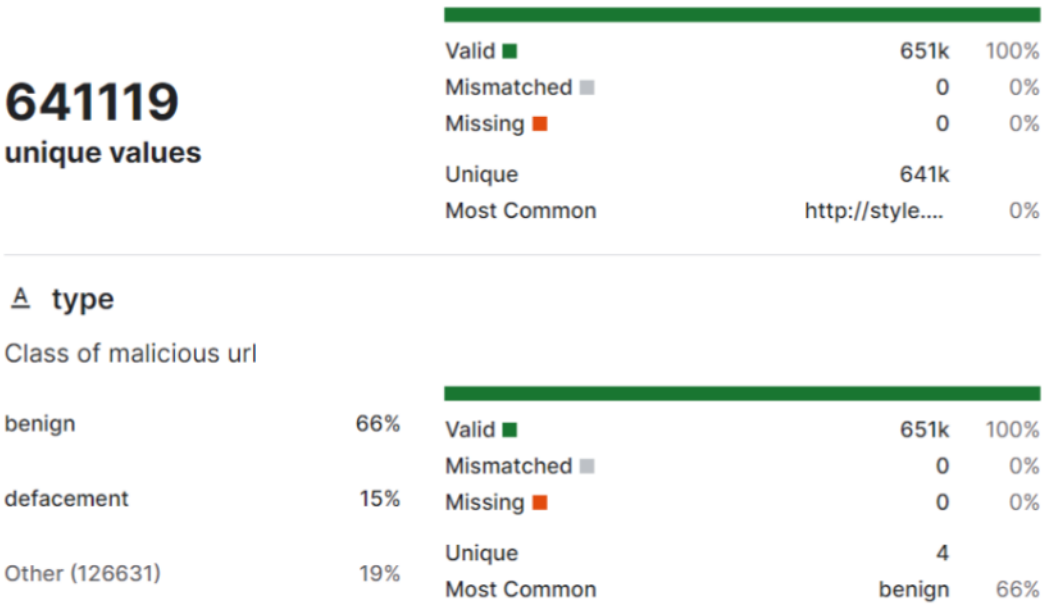

Figure 2: Dataset Description

5.2 Results:

The RandomForestClassifier exhibited commendable performance in the evaluation phase, achieving an impressive accuracy score of 93.2%. This metric indicates the proportion of correctly classified instances out of the total instances in the dataset, showcasing the model's ability to accurately discern between benign and malicious URLs. A high accuracy score signifies the effectiveness of the RandomForestClassifier in accurately identifying and categorizing URLs, thereby bolstering its suitability for real-world application in cybersecurity contexts. This robust performance underscores the model's reliability and efficacy in detecting malicious URLs and mitigating potential cyber threats. The Malicious URL Checker application offers a streamlined solution for users to swiftly ascertain the safety of URLs by leveraging machine learning techniques. Upon inputting a URL, the application employs a trained RandomForestClassifier model to analyze key features extracted from the URL, including character patterns, presence of IP addresses, and URL shortening indicators. This process enables the application to make an informed prediction regarding the URL's maliciousness, promptly informing users whether the URL is deemed safe or potentially harmful. By providing real-time threat assessment capabilities, the Malicious URL Checker enhances user awareness and aids in safeguarding against cyber threats, thereby contributing to a safer online browsing experience.

GitHub repository:
https://github.com/Sri-Hari2003/URL-classification

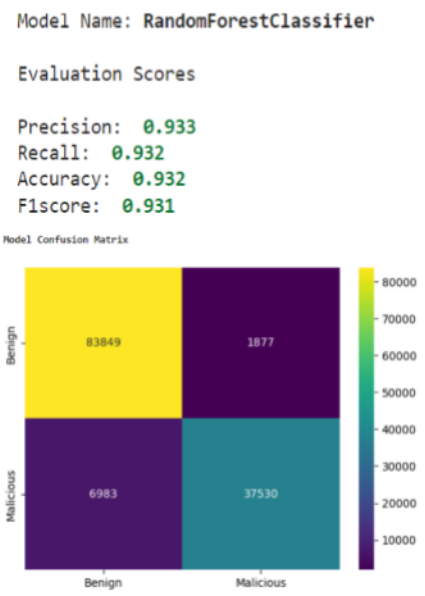

Figure 3: Metrics & Confusion matrix

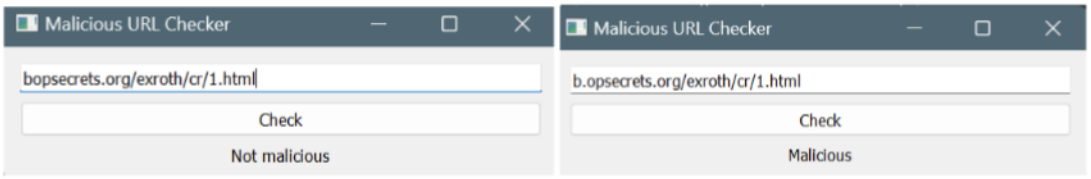

Figure 4: Output Prediction

6. Conclusions and Future Scope

The development of the Malicious URL Checker application represents a significant advancement in bolstering cybersecurity measures for users navigating the digital landscape. By harnessing machine learning algorithms, particularly the RandomForestClassifier model, the application effectively identifies potential threats posed by URLs in real-time. Through a streamlined user interface, individuals can swiftly assess the safety of URLs and take appropriate precautions to mitigate risks associated with malicious web content. The robust performance of the application underscores its utility in enhancing user awareness and fortifying defenses against cyber threats, thereby fostering a safer online environment. While the Malicious URL Checker application demonstrates promising capabilities in its current iteration, there exist several avenues for future enhancements and expansion. Firstly, integrating additional machine learning models and refining feature extraction techniques could enhance the application's accuracy and reliability in detecting a broader spectrum of malicious URLs. Moreover, incorporating dynamic analysis capabilities to assess URL behavior in real-time could provide deeper insights into evolving cyber threats. Furthermore, enhancing the application's user interface and compatibility across different platforms would ensure broader accessibility and usability for diverse user demographics. Additionally, continuous monitoring and updates to the underlying threat intelligence database would enable the application to adapt to emerging threats and maintain efficacy in safeguarding users' online experiences. Overall, through ongoing innovation and development, the Malicious URL Checker application holds immense potential to evolve into a pivotal tool in the ongoing fight against cybercrime and malicious online activities.

7. References

1) Alsaedi, Mohammed, et al. "Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning." *Sensors*, vol. 22, no. 9, 2022, p. 3373[2].

2) Cavnar, W.B., and J.M. Trenkle. "N-Gram-Based Text Categorization." *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161–175[2].

3) Chiramdasu, R., et al. "Malicious URL Detection using Logistic Regression." *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*, 2021[2].

4) Ding, C. "Automatic Detection of Malicious URLs using Fine-Tuned Classification Model." *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, 2020[2].

5) He, S., et al. "An Effective Cost-Sensitive XGBoost Method for Malicious URLs Detection in Imbalanced Dataset." *IEEE Access*, vol. 9, 2021, pp. 93089–93096[2].

6) Kuyama, M., et al. "Method for Detecting a Malicious Domain by Using Whois and DNS Features." *Proceedings of the Third International Conference on Digital Security and Forensics (DigitalSec2016)*, 2016[2].

7) Li, T., et al. "Improving Malicious URLs Detection via Feature Engineering: Linear and Nonlinear Space Transformation Methods." *Information Systems*, vol. 91, 2020, p. 101494[2].

8) Patil, D.R., and J.B. Patil. "Malicious URLs Detection Using Decision Tree Classifiers and Majority Voting Technique." *Cybernetics and Information Technologies*, vol. 18, 2018, pp. 11–29[2].

9) Phung, N.M., and M. Mimura. "Detection of Malicious JavaScript on an Imbalanced Dataset." *Internet of Things*, vol. 11, 2020, p. 100185[2].

10) Vinayakumar, R., et al. "Evaluating Deep Learning Approaches to Characterize and Classify Malicious URL's." *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 2, 2018, pp. 1333–1343[2].

RE-2022-272540-plag-report

ORIGINALITY REPORT

9%	6%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	researchbank.swinburne.edu.au Internet Source	1%
2	Submitted to National Economics University Student Paper	1%
3	Submitted to Georgia State University Student Paper	1%
4	www.qodenext.com Internet Source	1%
5	www.springerprofessional.de Internet Source	1%
6	ijrpr.com Internet Source	1%
7	Submitted to University of Huddersfield Student Paper	1%
8	Submitted to University of Lancaster Student Paper	<1%
9	Submitted to Taylor's Education Group Student Paper	<1%

10	www.ijraset.com Internet Source	<1%
11	Abdijalil Abdullahi, Mohamed Ali Barre, Abdikadir Hussein Elmi. "A machine learning approach to cardiovascular disease prediction with advanced feature selection", Indonesian Journal of Electrical Engineering and Computer Science, 2024 Publication	<1%
12	Shahram Esteki, Ahmad R. Naghsh-Nilchi. "SW/SE-CNN: semi-wavelet and specific image edge extractor CNN for Gaussian image denoising", Neural Computing and Applications, 2024 Publication	<1%
13	rocys.ici.ro Internet Source	<1%
14	S Geetha, Yusuf Mohammed Khan, Rohan Sujay, Sai Pavan Yoganand, Rohan B. "Fraudulent URL and Credit Card Transaction Detection System Using Machine Learning", 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS), 2023 Publication	<1%
15	mdpi-res.com Internet Source	

		<1%
16	web.cs.dal.ca Internet Source	<1%
17	Juwairiyyah, Anila Macharla, G. Kiran Kumar, D. Malathi Rani. "Detection and classification of malicious URLs based on machine learning models", 7th IET Smart Cities Symposium (SCS 2023), 2023 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

RE-2022-272540-plag-report

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7