Phishing URL Detection (COMP3125 Individual Project)

Raymond Eichner
School of Computing & Data Science

This project develops a machine learning system that determines whether URLs are phishing or legitimate. This was accomplished with two different models which were Logistic Regression and Random Forest. Both models achieved great performance, but Random Forest showed slower results. This model can be help in an organization to use as a defensive tool to detect malicious links automatically.

Keywords URL Classification, Cybersecurity, Machine Learning, Detection

I. Introduction

Phishing is one of the largest and most common cybersecurity attacks that organizations must defend against. Attackers attempt to make links look legitimate to try and trick users to click on them. A machine learning based detection system provides an automated protection by analyzing URLs and identifying malicious patterns without the user needing to try and determine that on their own. This project focuses on determining if URLs are either phishing or legitimate using supervised learning. I started by converting the URLs into numerical features, which I then used to train and test Logistic Regression and Random Forest models which was implemented using the scikit-learn library [1]. The goal of this project is to evaluate if a model can accurately detect phishing links.

II. Datasets

A. Source of dataset

The dataset used in this project was from Mendeley Data, titled *Phishing URL dataset* [2]. It is a large dataset with labeled URLs as either phishing or legitimate. This dataset was picked because of its size, and credibility.

B. Character of the datasets

The dataset was in CSV format and contains a little more than 450,000 URL samples. Each row has a URL which is a string and its corresponding label. The first column name is URL which is the full website URL as text. The second column name is type which is a string that tells us either phishing or legitimate.

III. Methodology

This project utilizes two different supervised machine learning models which are Logistic Regression and Random Forest to classify URLs as phishing or legitimate. Each model was trained using TF-IDF vectorized URL text, allowing the models to learn patterns.

A. TF-IDF Vectorization

URLs are unstructured text, so they had to be converted into numeric form before model training can begin. This was done using TF-IDF.

B. Logistic Regression Model

Logistic Regression is a classification algorithm that estimates the probability of a sample belonging to a certain class. It is lightweight, quick to train, and performs well when the dataset that is used is large. In this project the model was trained on the TF-IDF output to determine whether a URL belong to the phishing or legitimate class.

C. Random Forest Model

Random Forest is a learning models that uses multiple decision trees. Each tree makes its own prediction, and the result is determined by a majority vote. For this project a forest of 300 trees was trained. Accuracy was high, but Random Forest required more computing time, which makes it less efficient, but still valuable for its high accuracy.

IV. Results

Both the Logistic Regression and Random Forest were trained on the TF-IDF vectorized URL dataset and they were both evaluated using a test split of 20%. Accuracy was used as a performance metric, also four test URLs were feed to the model to demonstrate its effectiveness.

A. Model Accuracy

Random Forest performed slightly better having a 99.7% accuracy, but it was much slower to train and had much slower prediction time due to the numerous large tree architecture. Logistic Regression also had strong performance with an accuracy of 99.6% and was much faster to train.

B. Classification Examples

Several URLs were passed in to evaluate the model's performance with real examples. Both models were able to successfully identify the URLs and classify them appropriately.

C. Results

Logistic Regression produced a very fast model and was very accurate. Random Forest was slightly more accurate than Logistic Regression but was much slower to train and get results.

V. Discussion

Future improvements could include adding more legitimate login URLs to reduce false positives, testing on examples found in the real world, and testing additional algorithms to look for improvements in speed and accuracy.

VI. Conclusion

This project successfully demonstrated that machine learning can classify URLs as phishing or legitimate with high accuracy. Both logistic Regression and Random Forest can produce very accurate models, but Randon Forest can require

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

more computing time. These results show that URL based phishing detection can be automated making it a valuable tool in cybersecurity. This system could be used in email filters, or browsers to prevent users from clicking on malicious links.

Acknowledgment

I would like to thank my professor and the School of Computing and Data Science for providing me with the tools and support needed throughout this project. Their guidance and resources made completion of this work possible.

References

[1] "TfidfVectorizer – scikit-learn documentation," scikit-learn.org, Avaliable: https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[2] J. K. S. Kaitholikkal and A.B., "Phishing URL Dataset," Mendeley Data, Version 1, Apr. 2024. Available: https://data.mendeley.com/datasets/vfszbj9b36/1

[3] "Understanding Logistic Regression," GeeksforGeeks. Avaiable: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/

[4] "Random Forest Algorithm in Machine Learning," GeeksforGeeks. Avaiable: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/