PHISHING URL DETECTION USING MACHINE LEARNING ALGORITHMS

PRESENTED BY:

CYNTHIA NGETHE

PAMELA CHEBII

AGGREY TIMBWA

OMARA WALDEA

RICHARD MACHARIA

CAPSTONE PROJECT DOCUMENTATION SUBMITTED IN PARTIAL FULFILLMENT FOR THE AWARD OF DATA SCIENCE CERTIFICATE.

TABLE OF CONTENTS

SECTION ONE:
BUSINESS UNDERSTANDING

1.1 Introduction

Imagine you're a small business owner who relies on a website to manage customer relationships, handle transactions, and communicate with clients. Like many users, you're constantly online—navigating links from emails, social media, and suppliers. But as your business grows, so does the threat landscape around you. Cybercriminals are getting smarter, and phishing attacks are more sophisticated than ever, often hiding behind seemingly legitimate URLs. One wrong click, and you could be handing over sensitive data, from financial details to client information, putting your business and customers at serious risk.

Phishing attacks are one of the most prevalent and damaging forms of cybercrime today (Jain & Gupta, 2022). These attacks deceive individuals into divulging sensitive information such as usernames, passwords, and credit card numbers by masquerading as legitimate websites(Aljofey et al., 2020). Despite growing awareness, phishing remains a major cybersecurity threat due to the increasing sophistication of attackers who craft Universal Resource Locators (URLs) and websites that are difficult to distinguish from legitimate ones (Banu & Banu, 2013). The rapid growth of the internet and the ease of setting up fraudulent websites have made this problem even more critical, especially for organizations and individuals relying heavily on online platforms. These attacks often lead to substantial monetary loss(Adebowale et al., 2023).

Cybercriminals often use subdomain spoofing techniques by creating a subdomain resembling a legitimate website to convince users they are accessing a legitimate website (Vijayalakshmi et al., 2020). Malicious actors can use transposition to swap adjacent characters in a domain name to create a visually similar but phishing website. Top-level domains (TLDs), the last segment of a domain name, are extensively used by malicious actors to create a domain using a TLD that resembles well-known and trusted extensions to trick users into visiting fake websites (Adebowale et al., 2023).

The global impact of phishing attacks will continue to intensify, and thus, a more efficient phishing detection method is required to protect online user activities. To address this need, this study

focused on the design and implementation of a machine-learning phishing detection solution that leverages the universal resource locator.

1.2 Problem Statement

The challenge of detecting phishing websites lies in the ability to differentiate between legitimate and malicious URLs. Machine learning offers a promising solution by identifying subtle patterns in URLs and website content that distinguish phishing sites from legitimate ones. However, creating an effective detection model requires a comprehensive analysis of multiple URL features, such as structure, domain information, and HTML behaviour, which are often overlooked in traditional detection systems.

This project seeks to address the problem of phishing detection by developing a machine learning model that leverages 44 distinct features from URLs, including structural, domain-based, and behavioural characteristics, to distinguish phishing sites from legitimate ones. The goal is to achieve high accuracy and improve the generalizability of phishing detection models.

1.3 Objectives

1.3.1 Major Objective

This project aimed to develop a machine learning model capable of accurately detecting phishing URLs based on a wide range of features, including URL structure, domain attributes, content behaviour, and external reputation factors. It leveraged a diverse set of 46 features extracted from different URLs.

1.3.2 Specific Objectives

i. To extract features from labeled urls

ii. To perform feature engineering on the phishing detection dataset to identify the most relevant and discriminative features for classifying URLs as phishing or legitimate.

iii. To iteratively build, train, and optimize a series of increasingly sophisticated models, exploring various machine learning algorithms to achieve the highest predictive accuracy and robustness.

iv. To design and develop a web interface that allows users to input URLs and receive real-time predictions on phishing risks.

1.4 Stakeholders

The primary stakeholders involved in and impacted by this project are:

i. Business Users (Small to medium enterprises): owners, employees, and administrators who frequently navigate emails and online transactions and need a way to verify URLs for security

ii. Individual users (General Public): Everyday internet users who want a quick, reliable way to avoid phishing attacks and protect their data

iii. Cybersecurity Teams: Professionals in organizations who may incorporate this tool as part of a broader security protocol to preempt phishing threats

1.5 Success Criteria & Key Performance Indicators

To gauge the success of this project, we will track the following metrics:

i. **Model Accuracy**: Achieve a minimum accuracy of 90% on test data, aiming for a high true-positive rate (correctly identifying phishing URLs).

ii. **False Positive Rate**: Maintain a low false-positive rate to avoid incorrectly labelling legitimate URLs as phishing, which could decrease user trust.

1.6 Expected Business Impact

i. **Financial Protection**: Minimizes the risk of financial loss from phishing scams, especially for small businesses and individuals.

ii. **Improved User Trust and Confidence:** Builds user trust by offering a reliable tool for verifying URL safety.

iii. **Enhanced Cybersecurity Strategy:** Adds an additional layer of security to prevent phishing attacks, which is beneficial for organizations incorporating this tool into their cybersecurity protocols.

SECTION TWO:
DATA UNDERSTANDING

2.1 Data Source and Form of Data

The Mendeley phishing detection URLs dataset was used in this project. The original dataset was stored in text format. This dataset contains only URLs and their corresponding target variable. The data was initially divided into four subsets: a training set (further categorized into benign, meaning

legitimate, and malignant, representing phishing URLs) and a test set, which also had separate subsets for benign and malignant URLs. These subsets were later combined into a single dataframe, with "URL" as the sole feature and the labels for benign and malignant URLs as the target variable. The complete dataset comprised 6,568,184 observations. From this, a random sample of 10,000 observations was selected for processing. This smaller subset was chosen because extracting features from URLs is a memory-intensive and time-consuming process.

To the best of our knowledge, no known models have been developed using this specific dataset. While previous models for phishing detection exist, they were developed using other datasets, such as the Kaggle and UCI Machine Learning datasets. A key distinction between our dataset and these datasets is that we had to extract features from the raw URLs for modeling, whereas the Kaggle and UCI datasets already provided pre-extracted and well-labelled features.

SECTION THREE:
DATA PREPARATION

3.1 Feature Extraction

A total of 44 features were extracted from the URLs. These features capture various characteristics of the URLs and their associated metadata, such as:

- **General URL Properties**: url_length, num_subdirectories, num_query_params, path_length, num_slashes, shortened_url, url_is_random, and digit_ratio_in_url.
- **Domain Characteristics**: domain_length, tld, domain_entropy, has_hyphen, domain_age, days_to_expiry, is_expired, registration_duration, registration_type, and expiration_risk.
- **Security Indicators**: is_https, has_ip_address, has_www, has_redirect, and contains_homograph_chars.
- **Content Similarity and Relevance**: title, description, title_similarity_bin, description_similarity_bin, url_similarity_score, url_title_match_score, and domain_title_match_score.

- **Phishing Indicators**: common_phishing_words, typosquatting_distance, path_suspicious_keywords, and query_suspicious_keywords.
- **Social Engineering Clues**: has_brand_name_in_domain, has_social_net, char_repetition, and title_is_random.

Based on domain knowledge, most of these features are well-documented as potential indicators or avenues for phishing attacks.

3.2 Data Cleaning

The dataset underwent cleaning before modelling. This involved either dropping columns with a high number of missing values or removing rows with features that had fewer missing values. The creation_date and expiry_date columns were dropped because both had 9,552 missing values, rendering them irrelevant for analysis.

3.3 Encoding the Data

The dataset included three types of columns: Boolean, categorical, and numeric. We encoded all columns into a binary format to suit the classification problem. A column transformer pipeline was used to preprocess the data, incorporating techniques such as target encoding, one-hot encoding, and standard scaling.

3.4 Analysis Visualization

Various visualizations were created to examine the relationship between different features and the target column. These included a correlation heatmap, pair plots for selected features, a dual line plot of URL similarity scores for benign and phishing URLs, and a bar graph illustrating the distribution of URLs based on whether they are live or not, among others.

SECTION FOUR:
MODELLING

4.1 Techniques and Models Developed

The supervised learning technique was used in this project because we already had a target variable that classifies URLs as either phishing or legitimate. Since this is a classification problem, several classification algorithms were developed. The DummyClassifier was used as the baseline model.

Other models developed include Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), XGBoost, and the Perceptron model.

SECTION FIVE:
EVALUATION

5.1 Evaluation Metrics

The metrics used to evaluate the performance of the models include Precision, Recall, F1-Score, Accuracy, and AUC curves. The baseline model, DummyClassifier, had the lowest performance at 54%, while the SVM model, which had been processed using Principal Component Analysis (PCA), achieved the best performance at 94%. Iterative development and testing led to improvements in the models' performance.

SECTION SIX:
DEPLOYMENT

6.1 Deployment

The final report will be presented in the form of a PowerPoint presentation. Additionally, there is a web application platform that captures URLs and classifies them as either phishing or legitimate. The application extracts features from the URL, compares them with the model, and then provides the classification result. The platform consists of two pages: one serving as the homepage and the other for capturing the URL input and displaying the classification result.

SECTION SEVEN:
TOOLS/METHODOLOGIES

7.1 Methodologies

Google Colaboratory was used for feature extraction, preprocessing, and model development, as the machine's memory could not support the intensive tasks required. A GitHub repository was utilized for version control and collaboration among team members. The GitHub Projects Kanban Board was used for project management, as it allowed for tracking tasks assigned to different team members.

7.2 Tools

Several Python libraries were used to gather, clean, explore, and model the data. Some of the libraries include pandas, NumPy, scikit-learn (sklearn), BeautifulSoup (bs4), asyncio, urllib.parse, RapidFuzz, Seaborn, and Matplotlib, among others.

References

Adebowale, M. A., Lwin, K. T., & Hossain, M. A. (2023). Intelligent phishing detection scheme using deep learning algorithms. *Journal of Enterprise Information Management*, *36*(3), 747–766. https://doi.org/10.1108/JEIM-01-2020-0036

Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics (Switzerland)*, *9*(9), 1–24. https://doi.org/10.3390/electronics9091514

Banu, M. N., & Banu, S. M. (2013). A Comprehensive Study of Phishing Attacks. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *4*(6), 783–786.

Jain, A. K., & Gupta, B. B. (2022). A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterprise Information Systems*, *16*(4), 527–565. https://doi.org/10.1080/17517575.2021.1896786

Vijayalakshmi, M., Mercy Shalinie, S., Yang, M. H., & Raja Meenakshi, U. (2020). Web phishing detection techniques: A survey on the state-of-the-art, taxonomy and future directions. *IET Networks*, *9*(5), 235–246. https://doi.org/10.1049/iet-net.2020.0078