From Bytes to Bait: Data Science in Phishing Detection

Emily Gelchie and Colin Fitzgibbons

Providence College

CSC 485-001: Data Science Capstone

Dr. Michael Chou

May 8, 2024

Abstract

Phishing is a prevalent cybercrime whereby attackers impersonate legitimate entities through various communication channels, such as emails, phone calls, or text messages. Such schemes aim to manipulate victims into divulging confidential personal or corporate information, leading to significant security breaches. To combat this growing threat, an advanced predictive model has been developed to differentiate between phishing and legitimate website links accurately. Utilizing two extensive datasets, the model applies a comprehensive analytical framework to evaluate multiple characteristics of these websites. The resulting classifications, phishing or legitimate, are determined through sophisticated data science techniques, enhancing both the theoretical understanding and practical detection capabilities in cybersecurity. This research not only pushes the boundaries of phishing detection but also bolsters defenses against cyber threats.

Keywords: Phishing, Cybersecurity, Internet Safety, Data Science, Data Analysis, Machine Learning, Predictive Modeling.

From Bytes to Bait: Data Science in Phishing Detection

As the integration of the Internet into daily life has become omnipresent, cybercrimes have increased, inflicting significant damage on individuals, businesses, and entire communities. Among these, phishing is a prevalent cybercrime wherein attackers deceive victims into relinquishing personal or secure information through convincingly counterfeit websites, emails, text messages, and phone calls. The consequences of falling prey to phishing attacks are severe and varied, ranging from the theft of sensitive information such as passwords and banking details to financial losses and the potential installation of malicious software. Such breaches can compromise entire computer networks, disable security protocols, and open the floodgates to further unauthorized access and data theft. The repercussions of phishing extend beyond immediate financial damage, historically tarnishing the reputations of affected organizations and eroding trust among clients and customers. In 2023 alone, approximately 298,878 individuals in the United States reported incidents of phishing attacks, underscoring the widespread nature of this threat (statista.com). Despite the sophistication of these schemes, many phishing attempts can be thwarted through straightforward online safety measures. These include verifying the authenticity of the sender of communications, steering clear of untrustoworthy links, and maintaining up-to-date cybersecurity software. By implementing such preventive strategies, individuals and organizations can significantly mitigate the risk of phishing attacks, safeguarding their informational assets and reinforcing their operational integrity.

In this project, we employ feature importance techniques to analyze two datasets comprising website links, their associated characteristics, and their classifications as either phishing or legitimate. By utilizing various feature importance models, we aim to identify the most statistically significant attributes that will inform the development of a predictive model.

This model will incorporate a user-interactive component, enabling it to evaluate a website link and determine its legitimacy based on its distinct characteristics. The model's design is iterative, intending to enhance its performance continuously. The model is expected to achieve progressively higher accuracy, precision, and recall levels through ongoing learning from the characteristics extracted from each analyzed link. This adaptive approach ensures that the model remains effective in the dynamic landscape of cybersecurity threats, providing a robust tool for identifying phishing attacks.

Dataset Information, Preprocessing, and Feature Importance

Initial Datasets

For this research, two distinct datasets were sourced from Kaggle, a prominent data science community platform. The first dataset, "Dataset for Link Phishing Detection," amalgamates two separate datasets to enhance phishing detection capabilities. Authored by Winson, this dataset comprises 19,432 data entries and is accessible at (https://www.kaggle.com/datasets/winson13/dataset-for-link-phishing-detection?rvi=1). The second dataset, titled "Web Page Phishing Detection," was originally compiled and uploaded to Mendeley Data before being shared on Kaggle by the author, Manish KC. This dataset includes 11,482 entries related to website links and can be found at (https://www.kaggle.com/datasets/manishkc06/web-page-phishing-detection). Both datasets are structured similarly, comprising various website characteristics across multiple columns. These characteristics include but are not limited to, URL length, hostname details, and the presence of various characters within the links, such as dots, dashes, slashes, hyphens, colons, semicolons, and spaces. Further, the datasets enumerate attributes such as word length variables, the Google

index status of the page, the presence of popup windows, the count of hyperlinks, subdomain details, web traffic volume, the age and registration length of the domain, DNS records, and the page rank. The culmination of each dataset is a decisive 'status' column that classifies each website link as either 'phishing' or 'legitimate.' This structured and detailed dataset assembly is crucial for developing a robust predictive model to discern safe and malicious website links accurately.

Data Cleaning and Preprocessing

In the preprocessing phase of this project, the datasets underwent thorough cleaning to ensure data quality and reliability. Initially, both datasets were scrutinized by our team for duplicate records, which were removed to prevent bias in the model's learning phase. Our team handled missing data points by employing appropriate imputation methods to maintain the integrity of the datasets. Key preprocessing steps included converting categorical variables into numerical format through encoding techniques, allowing for better integration with machine learning algorithms. Specifically, variables such as 'status' were mapped from categorical to binary numerical values, where 'legitimate' was encoded as 0 and 'phishing' as 1. Furthermore, we determined selected features based on their importance from preliminary analysis; these included 'google_index,' 'web_traffic,' 'nb_hyperlinks,' 'domain_age,' 'url_length,' 'popup_window,' and 'page_rank.' These features were extracted by the team and formed the subset of data used for training the predictive models. To normalize the feature scales and improve the algorithm's performance, feature scaling was applied, ensuring that no single attribute would dominate the model due to its range or variance. This rigorous preprocessing regimen was pivotal in preparing the data for the subsequent modeling phase, aimed at accurately predicting phishing attempts.

While preparing the datasets for predictive modeling, it was crucial to standardize the column names across the 'Web Page Phishing Detection' and 'Dataset for Link Phishing Detection' to facilitate seamless integration and analysis. This process was executed using Tableau Prep, a robust data preparation tool that enhances data readiness for analytics. The initial step involved importing the datasets into Tableau Prep and implementing a 'clean' step, pivotal for editing and renaming the columns. During this phase, column names in the 'Web Page Phishing Detection' dataset were meticulously edited to mirror those in the 'Dataset for Link Phishing Detection.' Ensuring uniformity in variable names across datasets is essential as it simplifies the data handling processes in subsequent stages of machine learning modeling. This uniformity allows for the application of the same operations and algorithms to both datasets without the need for repetitive, individual adjustments, thereby streamlining the predictive modeling process and reducing potential errors associated with data management.

Exploratory Data Analysis

With initial exploratory data analysis in the project, we began looking for key columns and features of the datasets. Two variables we were able to focus on initially were the variables of web traffic and URL length, both of which were able to be visualized with distribution plots. For web traffic, we were able to produce a bar chart of phishing and legitimate traffic, as well as a distribution bar and area chart, both of which confirm a result of phishing websites containing higher amounts of web traffic than that of legitimate sites, with an average of 0.091054 for phishing sites and 0.068589 for legitimate sites. For the URL length, the distribution area chart validated that phishing sites were known to have longer URLs compared to legitimate ones.

In the preliminary phase of our exploratory data analysis, we focused on identifying key features within the datasets that could indicate phishing activities. Initial investigations centered

around web traffic and URL length variables, visualized using distribution plots. Our analysis revealed that phishing websites exhibit higher web traffic than legitimate sites, with phishing sites averaging 0.091054 and legitimate sites averaging 0.068589. This pattern was distinctly visualized through bar charts and area distribution plots, highlighting a notable disparity in web traffic volumes between the two categories. Furthermore, an analysis of URL length demonstrated that phishing sites typically feature longer URLs than legitimate sites, as evidenced by the distribution area chart, supporting the hypothesis that malicious entities employ more complex URLs to disguise nefarious links. Building upon these findings, we extended our analysis to a more detailed exploration of the second dataset, dataset_link_phishing.csv. A significant discovery from this dataset was the prevalent use of special characters in phishing URLs. This trend was quantitatively supported by histograms displaying a pronounced skew toward phishing links. Moreover, the analysis of SSL certificate implementation revealed that a substantial proportion of phishing sites either lacked SSL certificates or had improperly configured them, as depicted in pie charts comparing the SSL states of phishing versus legitimate sites. These findings were instrumental in enhancing our understanding of the technical tactics used by phishing operations.

In this research, we encountered challenges inherent to analyzing datasets predominantly structured for categorization, particularly in the context of phishing detection. While rich in categorical variables, such datasets often lack continuous metrics that facilitate the extraction of straightforward trends and insights. This characteristic significantly constrained the depth of traditional statistical analysis that could be applied, as many standard data exploration techniques are better suited to datasets with a broader spectrum of variable types. Recognizing these limitations, our project pivoted towards enhancing the machine learning methodologies

employed. We concentrated on refining our machine learning models to handle the categorical nature of the data better, aiming to improve classification accuracy and robustness. Advanced predictive analytics techniques were implemented to optimize model performance, including feature engineering and hyperparameter tuning. This strategic shift was essential to accommodate the data's categorical focus and achieve a more precise and reliable prediction of phishing activities. Incorporating ensemble methods and cross-validation strategies further enabled us to mitigate overfitting and improve the generalizability of our predictive models across unseen data, thus elevating the efficacy of our phishing detection system.

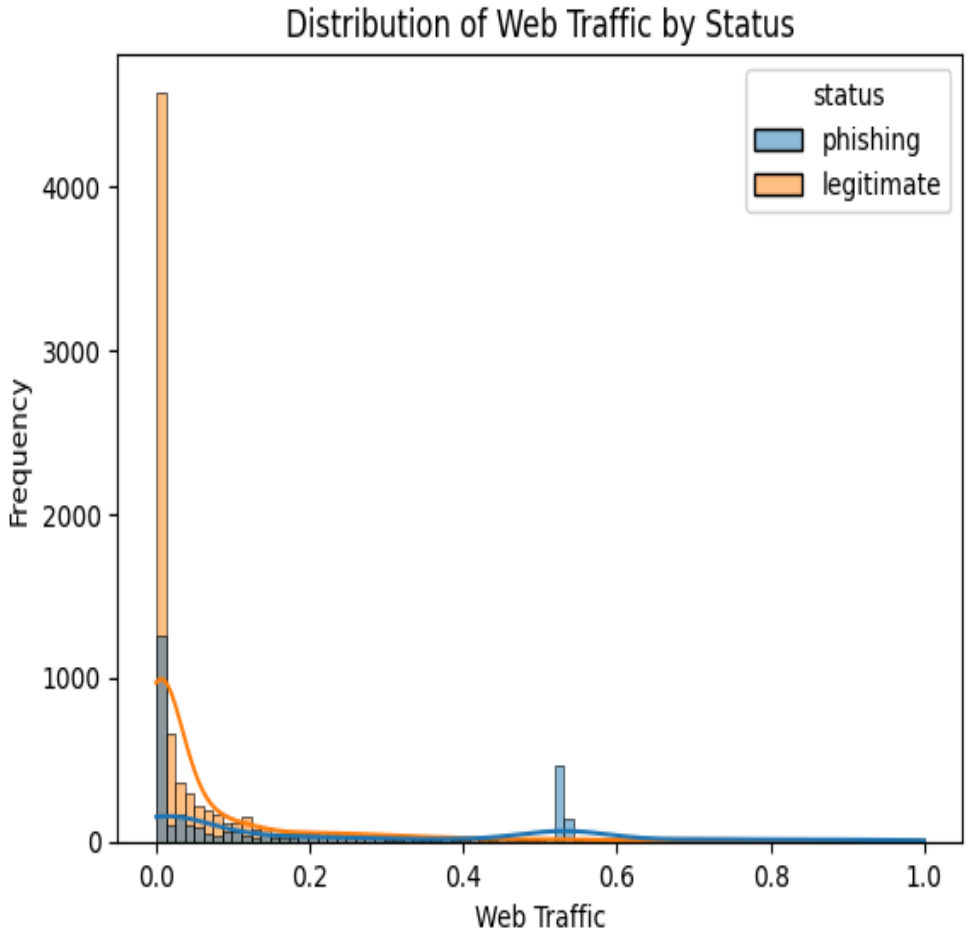

Figure 1. Distribution Chart of the amount of web traffic compared to the frequency of phishing and legitimate sites having lower or higher amounts.

Figure 2. Bar chart comparison of the amount of web traffic on phishing sites vs. legitimate sites.

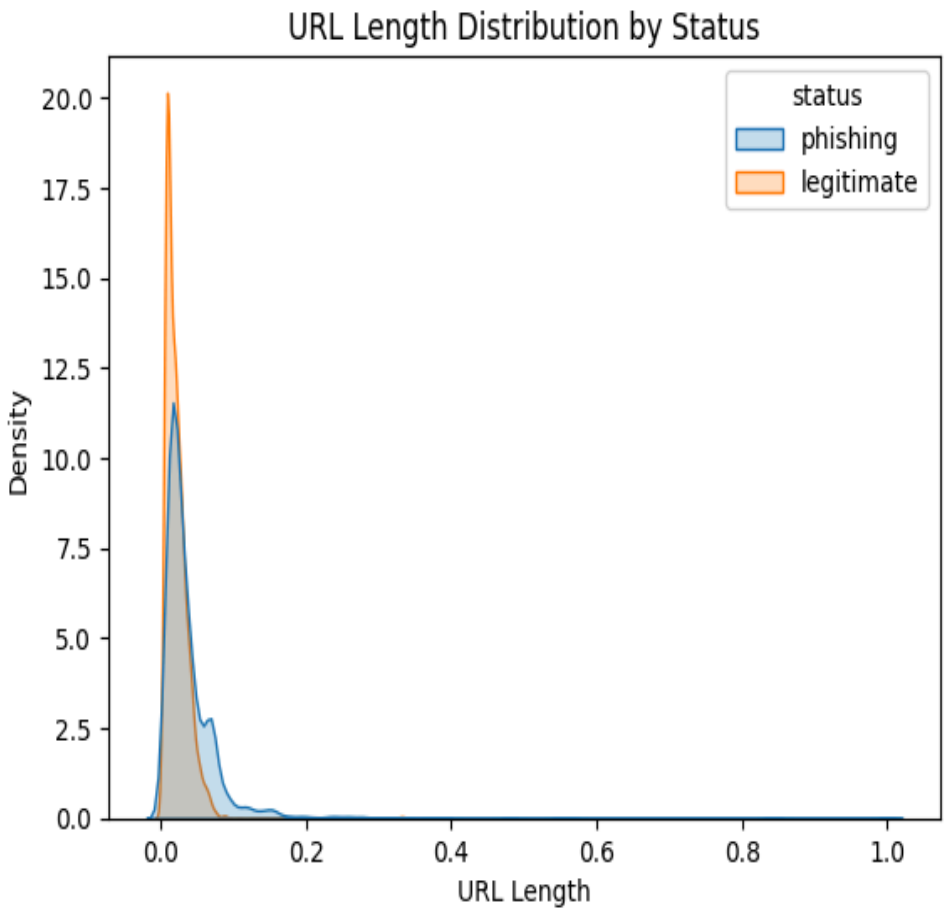

Figure 3. Area distribution chart of the density of the length of a URL between legitimate and phishing websites.

Data trends and correlations were also visualized with heat maps. Our analysis utilized heat mapping to investigate correlations between web features and their association with website status. Initially, we generated a comprehensive heat map encompassing all dataset columns. Due to the extensive number of variables, this initial visualization lacked clarity and readability. We conducted a feature importance analysis to identify the most relevant variables to address this, creating a more focused and interpretable heat map. This refined heat map included key variables such as 'google_index,' 'web_traffic,' 'nb_hyperlinks,' 'domain_age,' 'url_length,' 'popup_window,' 'page_rank,' and 'status_numeric.'

Our analysis revealed significant findings from this streamlined visualization. Notably, the 'google_index' demonstrated a strong positive correlation with 'status_numeric' (0.73), indicating that websites indexed by Google are generally less susceptible to being classified as phishing sites. In contrast, 'page_rank' was moderately negatively correlated with 'google_index' (-0.39) and 'status_numeric' (-0.51), suggesting that websites with lower page ranks are more likely to be phishing threats. Additionally, the correlation between 'domain_age' and 'page_rank' (0.59) suggests that older domains typically enjoy higher page rankings, which could imply a reduced likelihood of being involved in phishing activities.

These insights underscore the importance of selected features in predicting phishing, highlighting how specific web characteristics interrelate with the likelihood of a website being deemed safe or malicious. The focused heat map not only facilitated a clearer understanding of these dynamics but also reinforced the value of feature selection in enhancing the efficacy of the predictive model we developed.

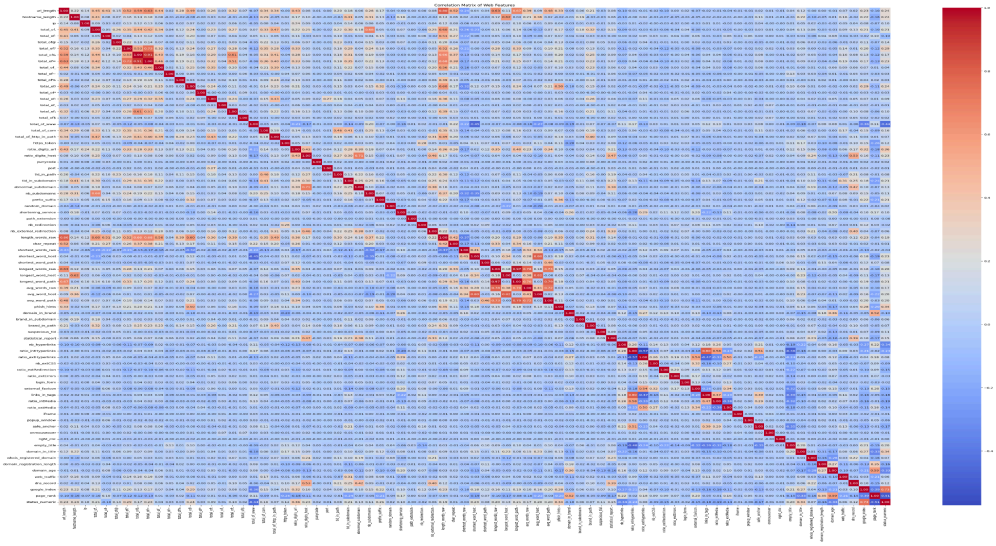

Figure 4: Initial heat map for dataset containing all columns as variables.

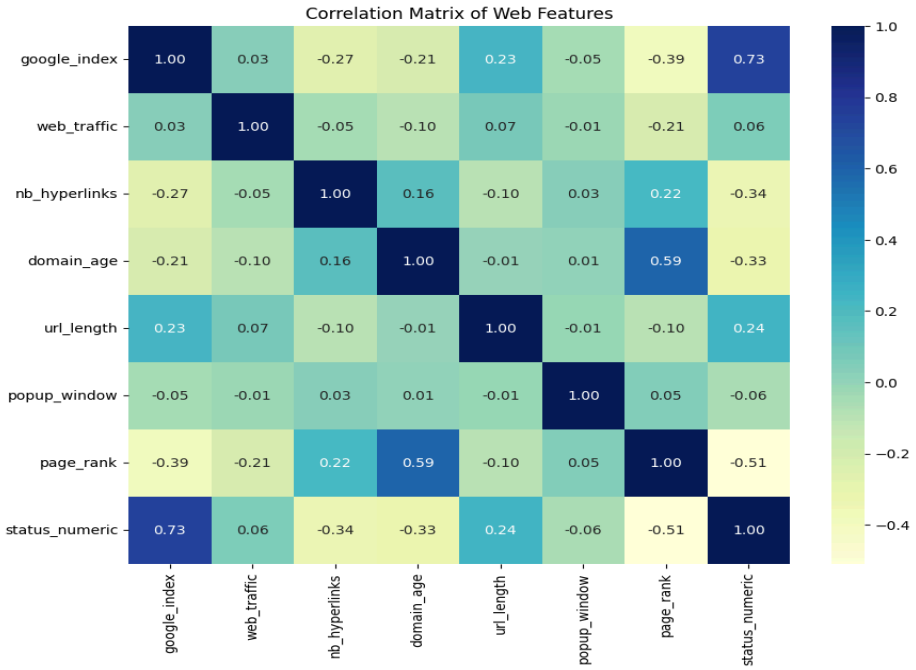

Figure 5: Revised heat map with columns selected after feature importance model evaluations.

These results were also seen in a dendrogram from performing hierarchical clustering. The hierarchical clustering dendrogram depicted provides a sophisticated analysis of the relationships among various web features pertinent to phishing detection. This visualization is instrumental in deciphering the associations and similarities between the features, which, in turn, elucidates their collective impact on the predictive accuracy of phishing detection algorithms.

Employing a color-coded schema to demarcate distinct clusters, the dendrogram suggests that features within the same color grouping exhibit higher similarity and interconnectedness than those in separate clusters. Notably, the dendrogram reveals a significant linkage between 'google_index' and 'status_numeric,' positioned in close proximity at a minimal hierarchical distance. This proximity underscores a robust positive correlation, affirming the hypothesis that Google's indexing status strongly indicates a website's legitimacy. The dendrogram also identifies a critical cluster comprising 'url_length,' 'nb_hyperlinks,' and 'popup_window,'

delineated in green. This cluster suggests a functional convergence among these features, typically manipulated in phishing schemes to craft deceptive web interfaces. Specifically, 'url_length' and 'nb_hyperlinks' may reflect the complexity of URLs that confuse or mislead users, while 'popup_window' may indicate aggressive tactics commonly employed by malicious sites. Additionally, the clustering of 'page_rank' and 'domain_age,' highlighted in blue, suggests that older domains often possess higher page ranks. This association implies that these features, when combined, are likely indicative of less susceptibility to phishing, valuable for models that prioritize domain longevity and reputation as markers of legitimacy. Moreover, 'web_traffic' and 'domain_age' form another notable cluster, shown in purple, indicating a correlation where older domains typically accrue more traffic, likely due to sustained trust and recognition over time. This refined clustering analysis is paramount for optimizing feature selection and understanding feature interdependencies. By concentrating on tightly clustered features, predictive models can achieve enhanced precision and recall, focusing on the most informative predictors of phishing activities. The dendrogram's hierarchical structure also assists in pinpointing potentially redundant features, thereby streamlining the predictive model to improve computational efficiency and model interpretability.

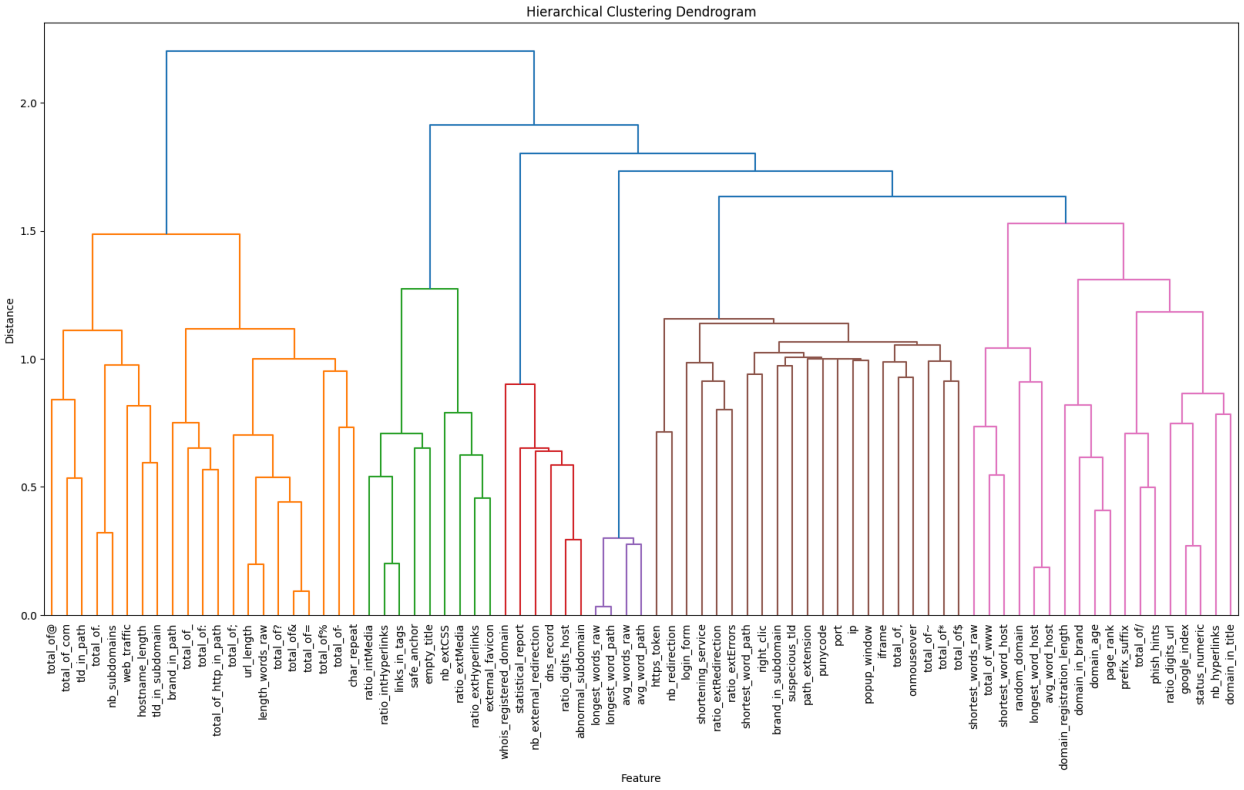

Figure 6: Hierarchical Clustering dendrogram representing the dissimilarity between features within clusters.

Feature Importance Modeling

To ascertain the relative importance of variables in determining the status of a website link as either phishing or legitimate, we employed multiple predictive models designed to assess feature importance across various dataset columns. Although the outcomes of these feature importance models varied due to the intrinsic differences in their algorithms, there was a notable consensus among many models regarding ranking key features. This convergence underscores the critical nature of certain variables that significantly enhance the performance of our predictive models. The identified features, consistently ranked near the top, are instrumental in refining the predictive accuracy of our models. By prioritizing these essential variables over the more indiscriminate use of all available data, our approach not only streamlines the modeling

process but also significantly improves our ability to distinguish between phishing and legitimate links.

Gradient Boosting with XGBoost Model

We chose XGBoost (Extreme Gradient Boosting) for its efficacy in iteratively refining its predictive capability by learning from previous iterations. This methodological choice is grounded in XGBoost's sophisticated decision-making framework, which incrementally optimizes feature selection based on reducing prediction error. Each feature is assigned a score based on the improvement 'gain' that its inclusion brings to the model's accuracy. This gain is computed across all trees in the ensemble, allowing for a nuanced assessment of each feature's impact. Among the top 20 features identified by XGBoost, the 'Google index' variable emerged as particularly significant, demonstrating a much higher importance score than other features. This feature's prominence can be attributed to several factors: its relevance and trustworthiness, as sites indexed by Google, are often vetted and deemed more trustworthy; its discriminatory power, effectively differentiating between phishing and legitimate sites; and the data-driven insights provided by XGBoost, which prioritizes features that split the data most effectively, reducing entropy or impurity in the resulting subsets. The iterative boosting process enhances the 'Google index' feature's influence on the model's decisions, as each tree in the ensemble evaluates and re-evaluates its contribution towards minimizing prediction errors. This dynamic refinement process explains why the 'Google index' stood out in XGBoost and across various other feature importance assessments and corroborative analyses, such as the heat map visualizations, highlighting its strong positive correlation with legitimate website classifications.

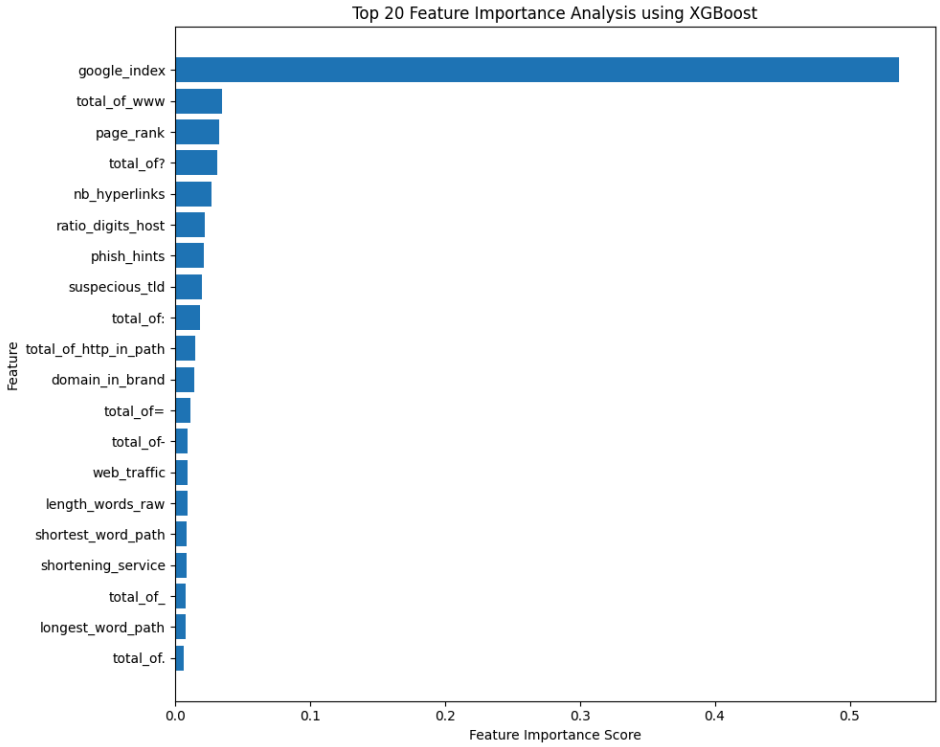

Figure 7: Google index featured as the most important variable with XGBoost feature importance.

RandomForestClassifier Feature Importance

The Random Forest Classifier was deployed to ascertain the significance of various features in predicting the legitimacy of website links. Unlike gradient-boosting models, which may exhibit a pronounced preference for a single feature, the Random Forest approach distributes importance more evenly among a broader array of features. This method involves constructing numerous decision trees, each trained on a randomly selected subset of the data, ensuring that each tree develops distinct insights into the feature set. The decision-making process within the Random Forest Classifier is both robust and comprehensive. Each tree in the forest makes a prediction, and the final output is determined by the majority vote among all trees, thereby enhancing the model's generalizability and reducing the risk of overfitting. The significance of each feature within this ensemble is evaluated based on its ability to reduce impurity across the trees. Specifically, a feature's importance is measured by observing how

much the accuracy of splits it provides improves the predictive power of the model—essentially, how effectively it helps in "cleaning" the data to reach a more accurate classification.

In the analysis of our dataset, the 'Google index' emerged as the most significant feature, consistent with findings from the gradient-boosting model, underscoring its critical role in distinguishing between phishing and legitimate sites. Other highly important features were 'page rank,' 'number of hyperlinks, and 'web traffic.' These features collectively contribute to a nuanced understanding of a website's profile ' page rank' may reflect a site's established credibility, 'number of hyperlinks' could indicate the interconnectedness or complexity of the site, and 'web traffic' serves as a proxy for the site's popularity and user engagement. This broad spectrum of features highlighted by the Random Forest Classifier underscores its utility in providing a holistic view of the factors contributing to website legitimacy. By leveraging the collective decision-making power of multiple decision trees, the Random Forest model affirms the pivotal importance of the 'Google index' and illuminates other significant variables that enhance the predictive model's accuracy and reliability.

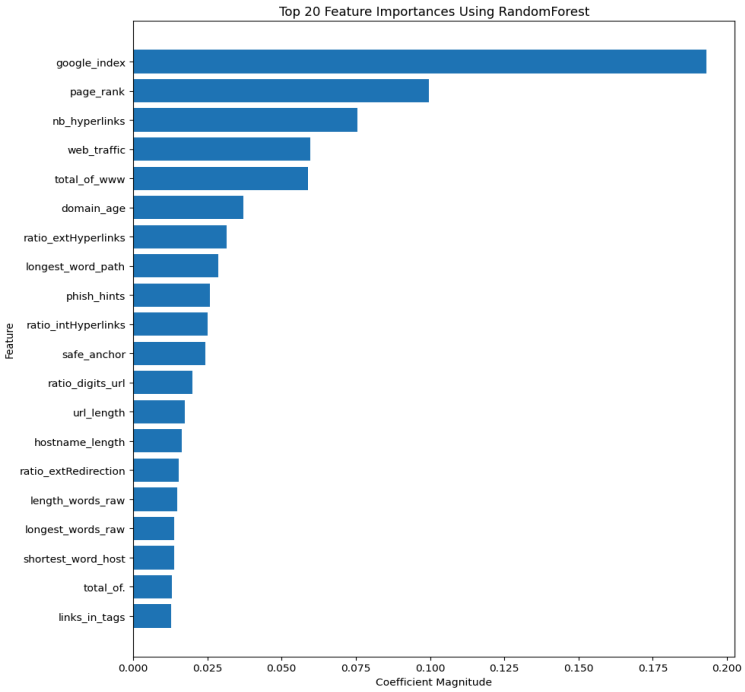

Figure 8: Google index holding a significantly higher coefficient magnitude, along with page rank being an important feature.

Absolute Correlation with Target Feature Importance

The absolute correlation with the target model was employed to systematically quantify the relationships between each feature in the dataset and the target variable—the status of the website link. This methodological approach utilized the Pearson Correlation Coefficient to measure the linear relationship between each feature and the website's status, emphasizing those features with higher absolute correlation coefficients. Such features exhibit a stronger and more significant linear relationship with the site’s legitimacy status, indicating their pivotal role in predictive modeling. Distinct from the previous models, this absolute correlation model highlighted several unique features as particularly influential. 'Popup_window' emerged as the most significant feature, suggesting that the presence or absence of popup windows in a website strongly indicates its legitimacy or lack thereof. Following closely were 'total_of_underscores', which relates to the URL structure and may indicate suspicious patterns often associated with

phishing sites, and 'shortest_words_raw,' which measures the length of the shortest word in the raw text data—possibly reflecting on the sophistication or crudeness of the site's content. Additionally, 'ratio_extErrors' was identified as critical, indicating that external errors linked to the website could be a strong predictor of a phishing attempt.

Such errors involve links to external sites that fail or other problematic elements that are less common in legitimate domains. The absolute correlation model offers a different perspective than gradient boosting and Random Forest models by focusing on these features with higher correlation coefficients. This approach underscores the importance of direct, linear relationships between features and a website's phishing status, providing a clear and quantifiable metric for feature importance.

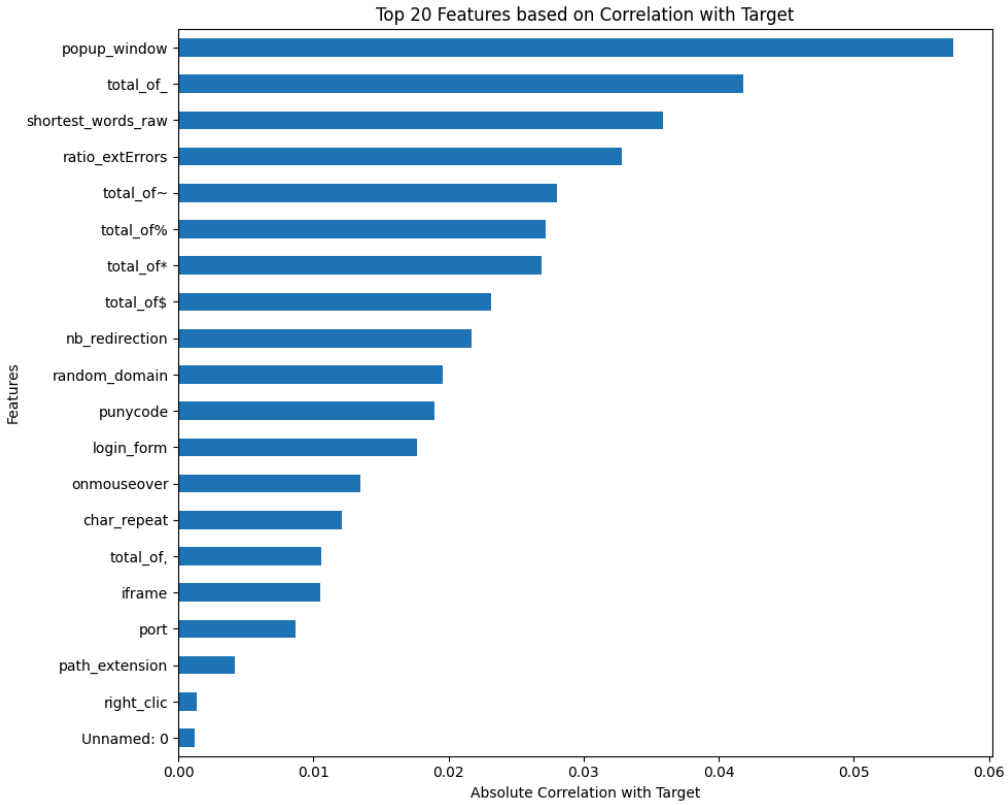

Figure 9: Popup window showcased as having the strongest absolute correlation with the status of a link.

Logistic Regression Feature Importance

Logistic regression was employed to ascertain the influence of various features on the likelihood of a website being classified as phishing. The model calculated coefficients for each feature, which quantify their impact on the predicted probability of the outcome variable—namely, whether a website is phishing. These coefficients, representing the relationship between the independent variables and the log odds of the site being phishing, provide a clear measure of each feature's influence on the model's decisions. Notably, the logistic regression model assigned significant importance to 'URL length' and 'number of hyperlinks', as indicated by their larger coefficient magnitudes. This suggests that these features strongly influence the model's ability to predict phishing websites. A longer URL might indicate a deceptive attempt to mimic a legitimate site or hide malicious parameters within the link itself. Similarly, a higher number of hyperlinks could suggest either an attempt to manipulate search engine algorithms or to engage users with malicious links. The logistic regression model stands out for its ability to balance interpretability, efficiency, and effectiveness, making it particularly valuable for practical cybersecurity applications where understanding the weight and impact of each feature is crucial. Unlike other models that may prioritize a broader array of features, logistic regression highlights specific features with the most substantial direct influence on the outcome, as evidenced by the top features identified. This focus allows for a more targeted approach in phishing detection, where key indicators such as URL characteristics play a pivotal role in distinguishing between legitimate and malicious sites.

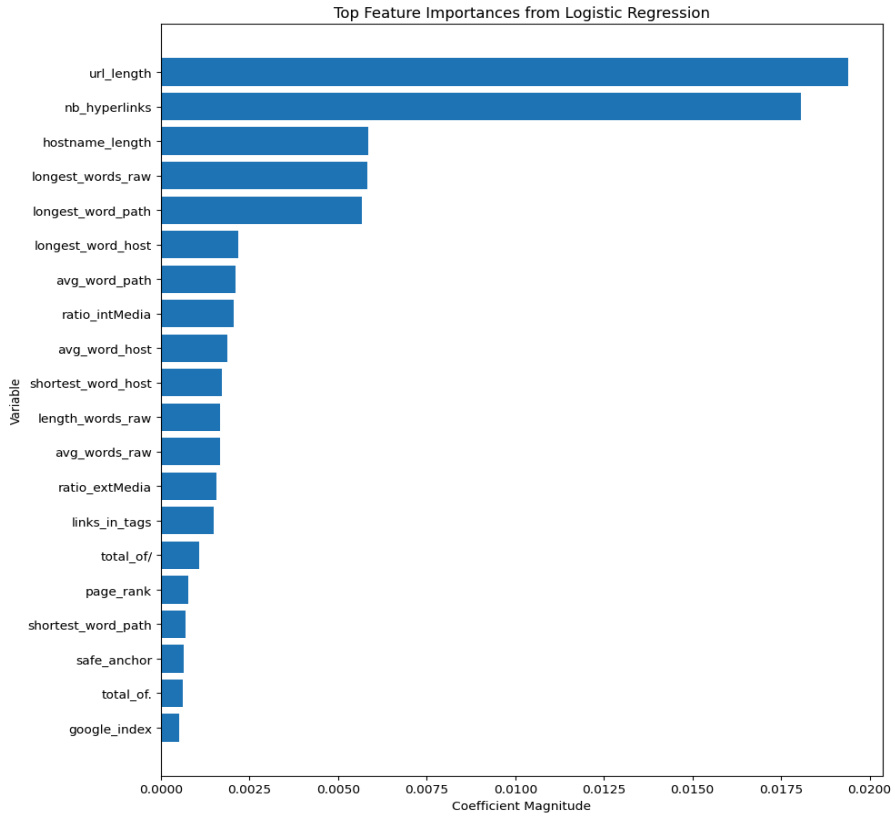

Figure 10: URL Length and Number of Hyperlinks holding significantly stronger coefficient magnitudes in logistic regression.

Permutation Importance Feature Importance

The permutation importance model, utilized as this study's final feature importance prediction method, is based on a unique evaluation mechanism. This model assesses the importance of each feature by measuring the effect on the model's accuracy when the values of that feature are randomly shuffled. This shuffling process disrupts the relationship between the feature and the outcome, allowing for an assessment of how model performance deteriorates without the structure provided by the feature. In this evaluation of permutation importance, the model identifies the 'number of hyperlinks' as the most significant feature, highlighting its pivotal role in predicting website legitimacy. This feature's preeminence is underscored by its strong influence on model accuracy; when hyperlinks are permuted, the model's ability to

correctly classify websites as phishing or legitimate declines sharply. Similarly, 'domain age' and 'URL length' emerged as critical features, affirming their substantial weights observed in logistic regression and Random Forest models. These findings are consistent with the understanding that older domains typically carry a history that either corroborates their legitimacy or exposes their longstanding malicious use, while URL length can indicate a phishing site attempting to mimic complex legitimate URLs to deceive users. The permutation importance model, therefore, not only corroborates the findings from other predictive models but also enhances the robustness of the feature evaluation by directly measuring the impact of each feature's removal on the accuracy of the predictive outcome.

Weight	Feature
0.1760 ± 0.0091	nb_hyperlinks
0.0427 ± 0.0050	domain_age
0.0371 ± 0.0047	url_length
0.0035 ± 0.0012	hostname_length
0.0025 ± 0.0009	longest_word_path
0.0025 ± 0.0015	longest_words_raw
0.0014 ± 0.0022	domain_registration_length
0.0013 ± 0.0006	web_traffic
0.0010 ± 0.0012	ratio_extMedia
0.0005 ± 0.0005	shortest_word_host
0.0004 ± 0.0020	Unnamed: 0
0.0003 ± 0.0005	avg_word_path
0.0003 ± 0.0003	avg_words_raw
0.0001 ± 0.0002	total_of_www
0.0001 ± 0.0002	page_rank
0.0001 ± 0.0002	shortest_words_raw
0.0001 ± 0.0001	google_index
0.0001 ± 0.0001	ratio_extHyperlinks
0.0001 ± 0.0001	total_of-
0.0001 ± 0.0001	phish_hints
… 65 more …	

Figure 11: The list of permutation importance features, suggesting number of hyperlinks, domain age, and URL length significantly impact model performance.

Machine Learning for Link Detection

In this project, two variations of the Random Forest Classifier were employed alongside a Stochastic Gradient Descent Classifier to optimize phishing detection. The initial Random Forest model served as the foundational predictive model, utilizing key features identified as significant through regression analyses. This model was crucial for establishing baseline predictive capabilities. A subsequent version of the Random Forest was developed as a user-interactive model designed to incorporate user feedback. However, this iteration did not adapt based on user interactions as anticipated. We integrated an SGD Classifier to refine the model through user feedback to address this limitation. Despite its potential for adaptability, the SGD model initially underperformed in accuracy compared to the Random Forest. Given these challenges, a hybrid approach was adopted, combining the robust prediction capabilities of the Random Forest with the adaptive potential of the SGD Classifier. This strategy aimed to harness the strengths of both models to maximize accuracy and responsiveness in the phishing detection system.

Random Forest Model 1

For our first machine learning model, we chose to compile a Random Forest Classifier model for its ability to work with the large datasets we used, higher accuracy rating, low overfitting rate, and less proneness to bias within. Based on all the feature importance models, we selected the top features as the features to integrate into the model, being the Google index, the amount of web traffic, the number of hyperlinks, the age of a website's domain, the length of the URL, the popup window, and the page ranking, all compared against the status of the website. From training and testing with the first dataset, the Random Forest model produced an accuracy score of 98.6%, with a root mean squared error of 0.117 and a mean absolute error of

0.013. The first predicting round also generated a precision score of 0.9829, a recall score of 0.9899, and an F1 score of 0.9864. Because this model was so accurate, we wanted to test the prediction capabilities using a second dataset to ensure that our model was not overfitting. Retesting with the secondary dataset, structured to be nearly identical to the first, the model was 99.7% accurate, with a root mean squared error of 0.044, a mean absolute error of 0.0020, a precision score of 0.997, a recall score of 0.998, and an F1 score of 0.997.

Random Forest Model 2

The second machine learning model deployed in this project is a Random Forest Classifier, specifically designed to enhance the phishing detection system by interacting directly with users. The model's primary function is to evaluate URLs provided by users and predict whether they are phishing or legitimate based on a set of extracted features. This approach leverages the RandomForestClassifier from the `sklearn.ensemble` library, renowned for its robustness and accuracy in handling complex classification problems.

The model begins by extracting key characteristics from the user-provided URLs. This involves normalizing the URL to include an HTTP/HTTPS prefix, parsing the URL to extract components such as the domain and path, and calculating a set of features including the URL length, number of subdomains, presence of HTTPS, and various counts of special characters that might indicate suspicious patterns (e.g., '-', '@', '?'). The model also checks for the presence of top-level domains within the path or subdomains and identifies common phishing indicators like URL shortening services and numerical sequences in domains. An innovative aspect of this feature extraction is the attempt to detect popup scripts within the page's HTML content, which can be indicative of phishing attempts.

The extracted features are then utilized to train the Random Forest model. Training involves dividing the dataset into training and testing subsets, with the model learning to classify URLs based on the patterns observed in the training data. After training, the model's performance is evaluated using standard metrics such as accuracy, precision, recall, and F1 score, alongside a confusion matrix to visually represent the classification outcomes. Once deployed, the model allows for real-time interaction with users. When a URL is entered, the model predicts its status based on the learned patterns and immediately seeks user feedback on the accuracy of its prediction. This interaction is crucial as it was intended to help the model learn from its mistakes by incorporating user feedback into future model adjustments.

Despite achieving a high accuracy rate of 94%, the model exhibited limitations in its learning capability. Specifically, it struggled to adapt based on user feedback when its predictions were incorrect. This limitation stemmed from the initial design, which did not incorporate mechanisms to dynamically update the model's learning based on new data provided through user interactions. As a result, while the model excelled in static prediction tasks, its inability to evolve with new data and correct its biases or errors led to dissatisfaction with its performance in dynamic, real-world scenarios. To address this, further enhancements are needed to integrate continuous learning capabilities, possibly through online learning algorithms or reinforcement learning techniques, which could allow the model to adjust its parameters in response to new information and improve over time effectively.

Stochastic Gradient Descent Classifier Model

The next approach in enhancing our phishing detection system involved implementing the Stochastic Gradient Descent Classifier, particularly favored for its ability to handle online learning—updating its model incrementally as new data arrives. This feature is critical in

dynamic environments like cybersecurity, where threats constantly evolve, and models must adapt swiftly.

The Stochastic Gradient Descent Classifier in this project is configured to use a logistic regression loss function (`log_loss`), which enables probability estimation for the binary classification of URLs as either phishing or legitimate. This setup allows the model to learn incrementally from each piece of user feedback, adjusting its parameters in response to new information. This dynamic learning process is facilitated by the `partial_fit` method of the classifier, which is designed to update the model with a single iteration over a provided subset of data, allowing for adjustments without retraining from scratch.

In practical application, the system prompts the user to classify a URL, then uses the extracted features to predict the category of the URL. Following the prediction, user feedback is solicited to confirm or refute the model's decision. This interaction is crucial as it informs the model about its performance in real-time scenarios. If the feedback indicates an incorrect prediction, the model immediately adjusts using the corrected label, enhancing its accuracy progressively with each interaction. This feedback loop is intended to refine the model's predictive capabilities continually.

Despite these advanced features, initial comparisons showed that the SGD Classifier's accuracy was lower than that of the more robust Random Forest model. While the SGD's adaptability to new data and errors presents a significant advantage in maintaining model relevance over time, it initially lacked the precision and reliability provided by the Random Forest. This discrepancy highlighted a trade-off between the adaptability of SGD and the static accuracy of traditional ensemble methods.

Thus, while the SGD Classifier brought valuable learning dynamics to the phishing detection system, enabling it to evolve with user input and real-time data, it required further tuning to reach the accuracy levels of established models like the Random Forest. This exploration into different modeling approaches underscores the complexity of designing effective cybersecurity measures, where both accuracy and adaptability must be balanced to address continuously changing threats effectively.

Final Hybrid Model

Our final product is a sophisticated hybrid machine learning model that synergistically combines the strengths of the Random Forest Classifier and the Stochastic Gradient Descent Classifier to enhance phishing detection under varied levels of predictive certainty. The primary component of this model, the Random Forest Classifier, serves as the foundational predictive mechanism, employing a diverse array of extracted URL features—such as URL length, the presence of HTTPS, and the number of subdomains. These features are integral to identifying potential phishing attempts based on established patterns derived from historical data. To augment the primary model, we integrated the SGD Classifier, configured with a logistic regression loss function to facilitate incremental learning. This secondary model is activated particularly in scenarios where the Random Forest's confidence in its predictions does not surpass a predefined uncertainty threshold, set at 0.7 for this study. Under these conditions, the SGD Classifier's role is to provide a secondary assessment, which is continually refined through real-time user feedback. This feedback mechanism is critical as it allows for the dynamic adjustment of the model based on direct user interactions, thereby enhancing the model's adaptability and accuracy in live environments. Operationalizing this model involves initial training on an 80/20 train/test split, followed by continuous re-calibration of the SGD component

using the `partial_fit` method, which updates the classifier with new data as user feedback is received. This approach ensures that the model remains responsive to new and evolving phishing strategies that may not have been present in the initial training set.

The efficacy of this hybrid model is demonstrated by its base accuracy rate of 94%, attributed to the Random Forest Classifier. However, the unique advantage of this system lies in its session-based adaptability, where accuracy can improve incrementally as the SGD model adjusts based on user-validated inputs. This feature is particularly valuable in a cybersecurity context, where the threat landscape is continually evolving, requiring detection systems that are not only accurate but also adaptable. Overall, the integration of robust classification with adaptive learning represents a significant advancement in the application of machine learning techniques to cybersecurity, offering improved responsiveness to emerging threats through an innovative feedback-oriented approach. This model exemplifies the potential of hybrid systems to leverage the strengths of multiple machine learning methodologies to achieve superior performance in complex, real-world applications.

Conclusion

Challenges Faced

Our project, characterized by its collaborative nature, involved multiple contributors, leading us to adopt Google Colab as our primary platform to facilitate accessible and simultaneous work. However, this choice brought about its own set of challenges. One significant issue was the limitations inherent in Google Colab concerning simultaneous code editing. Contributors often found themselves unable to edit code concurrently in real-time, which hindered the fluidity of collaborative development. Additionally, we encountered recurring issues

with document saving, which posed a risk of losing code or failing to capture the latest modifications during collaborative sessions.

Another substantial challenge was the compromised accuracy of our user-interactive model, primarily due to our restricted access to all the data points used in training the machine learning model. This restriction stemmed from the lack of necessary API permissions, which prevented us from fully utilizing the data, thereby limiting the model's learning potential and overall effectiveness. Furthermore, the task of identifying crucial features across multiple regression models proved to be particularly daunting. Each regression model employs a distinct algorithmic approach to assess and weigh feature importance, leading to variability in the features deemed significant by each model. This variability introduced complexities in deciding which features to include in the final machine learning model, as it was challenging to reconcile differing evaluations of feature importance across models.

Lastly, the binary nature of the features within our categorical dataset posed significant hurdles for conducting exploratory data analysis and creating effective visualizations. Traditional visualization techniques often fell short in providing clear insights due to the limited variability in binary data. This limitation necessitated the adoption of alternative analytical approaches, such as specialized data transformation and visualization techniques, to extract and illustrate meaningful insights from the dataset. These challenges, while substantial, provided valuable learning opportunities and highlighted the importance of adaptability and innovative problem-solving in the realm of collaborative machine learning projects.

Next Steps

Building on the successes and challenges of our current project, several strategic steps are outlined to further enhance the effectiveness and scope of our phishing detection system. Firstly,

integrating an Application Programming Interface will be critical. This integration aims to augment the user-centric machine learning model by providing comprehensive access to all relevant data variables from user-inputted sites, thus aligning more closely with the training dataset and enhancing accuracy. Furthermore, to address the evolving nature of phishing tactics, our project will expand to include Large Language Models to detect phishing attempts across various communication mediums such as text messages, phone calls, and emails. This expansion will allow the model to adapt to the multifaceted forms of phishing attacks encountered by users in different contexts. In addition to broadening the scope of detection, there is a clear need to expand our dataset. By incorporating more data points into our model, we can significantly improve the predictive accuracy and robustness of our system. More comprehensive data will provide a richer foundation for training and refining the machine learning algorithms. Lastly, we plan to develop a user-friendly application that consolidates our findings and tools into a single, accessible platform. This application will serve as a vital resource for users to verify the legitimacy of websites easily, significantly enhancing the practical utility of our research.

Key Takeaways

Overall, this project was able to highlight key aspects of website links in order to determine if a site is a phishing website, or if it is a legitimate website. Through feature importance as well as predictive modeling, we were able to filter our dataset of link features to eight specific features that can be seen as significant when determining if a website is malicious or not. Through building multiple regression models, we were able to identify key characteristics of phishing sites, especially with a user interactive feature to allow for skeptical users to input a link and have the model classify it as phishing or legitimate, supplemented by reinforcement learning. Overall, this project significantly increases phishing knowledge, awareness, and

protection though highlighting key features of websites, as well as an ever-improving model for detection.