Integrated Capstone Project (Non-Technical Report)

DX799 Data Science Capstone (Fall 2025)

Michael Amare

Executive Summary

Phishing websites continue to be one of the most common causes of online fraud, financial loss, and identity theft. These sites imitate legitimate businesses to trick people into sharing sensitive information. Because attackers constantly change their tactics, traditional security tools often fail to detect new threats in time. This project set out to develop a data-driven system that can identify phishing websites automatically, helping organizations protect users before harm occurs.

Using a publicly available dataset of more than 11,000 websites, I analyzed information drawn from the structure of each URL, details about the domain, and basic webpage characteristics. Several machine-learning approaches were tested to see which could most reliably distinguish phishing sites from legitimate ones. Models such as K-Nearest Neighbors, Gradient Boosting, and XGBoost along with others were evaluated, and simple clustering techniques were also used to explore natural patterns in the data.

The most accurate and dependable model was XGBoost, which achieved strong performance in detecting phishing attempts while keeping errors low. These results show that organizations can use machine-learning models to score incoming URLs in real time and block harmful sites before users interact with them.

Based on these findings, the report recommends integrating an automated detection model like XGBoost into existing security workflows. Doing so can reduce manual review, lower false-positive alerts, and improve overall protection against fraudulent websites. The project demonstrates that data-driven methods can significantly strengthen online safety while supporting faster, more informed decision-making.

Introduction & Problem Definitions

Phishing has become one of the most widespread and damaging online threats, affecting individuals, businesses, and government organizations. Every day, new fake websites appear that look real enough to trick users into sharing passwords, financial information, or personal data. These sites often imitate banks, online stores, or trusted services, making them hard to spot without the help of advanced tools. Because attackers constantly change their techniques, traditional security systems that rely on fixed rules cannot keep up. This project focuses on using data-driven methods to help organizations identify phishing websites more accurately and more quickly.

The main goal of this work is to create a system that can automatically judge whether a website is likely to be fraudulent. To do this, I analyzed a large collection of real phishing and legitimate websites. Instead of looking at the page's content, the project focuses on patterns in the URL itself, details about the domain, and simple signals related to web traffic. These clues can reveal suspicious behavior that may not be obvious at first glance.

For the organizations that would use this system, the objective is clear, reduce the chance that employees or customers fall victim to phishing attacks. A reliable detection tool would lower financial losses, prevent data breaches, and strengthen user trust.

To measure success, the project used simple criteria: the system should correctly identify phishing websites at least 95% of the time and be able to flag suspicious sites with minimal mistakes. This project aims to give decision-makers a practical, understandable, and effective way to strengthen their cybersecurity defenses before harm occurs.

Data Overview

This project uses three datasets that together help explain how phishing websites differ from legitimate ones. The main dataset, called the Web Page Phishing Detection Dataset, includes about 11,400 website records. Each record represents a single website and contains 88 simple characteristics, or "clues," about the website's address (URL) and the domain behind it. These clues include things like whether the website uses a valid security certificate, how long the URL is, how old the domain is, and how many links it contains. While the list of clues is long, each one represents a small piece of information that can help a system judge whether a site seems trustworthy.

Two smaller datasets, the Phishing Emails Dataset and the Cybersecurity Risk Dataset, were also reviewed to provide context. They were not used for the main predictions but helped confirm common patterns seen across different types of phishing behavior, such as unusual domain ages or suspicious link structures.

Before any analysis could take place, the data had to be checked for accuracy and consistency. This included confirming that each website record was complete, removing any duplicates, and making sure the values were reasonable. For example, some websites had missing domain age information or traffic data, which required careful handling to avoid misleading results.

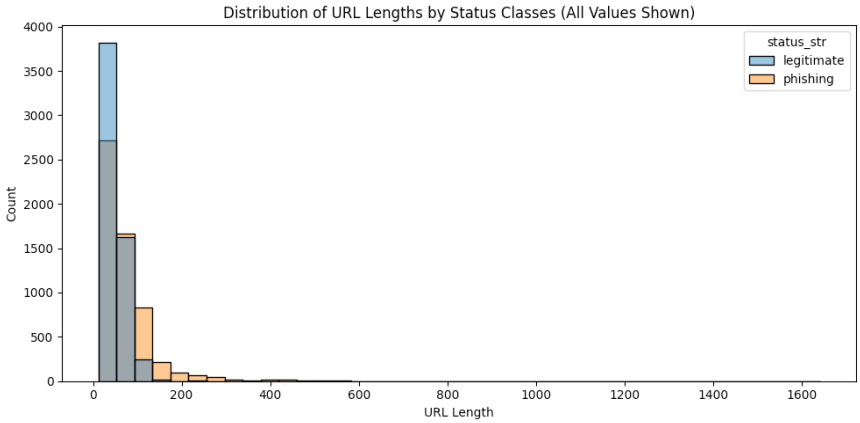
From an ethical standpoint, the datasets do not contain personal or sensitive user information only technical details about websites. This helps ensure that the project respects privacy while still producing meaningful insights for cybersecurity. In summary, the data used in this project provides a strong and reliable foundation for understanding how phishing websites behave and how an automated system can detect them based on measurable patterns.

Data Cleaning, Preprocessing, Exploratory Data Analysis (EDA)

Before building any models, the dataset required several preparation steps to ensure accuracy and consistency. I began by reviewing the data for missing values, inconsistent entries, and unusually extreme numbers. Any incomplete or clearly incorrect records were removed so that they would not influence the results. I also examined key numeric features such as URL length and character patterns to identify outliers. Cleaning these issues helped create a more dependable dataset for analysis.

Once the data was cleaned, I standardized certain features so they could be compared more fairly. Some columns were converted into more readable formats, such as transforming the numeric phishing label into a simple phishing/legitimate indicator. These preprocessing steps simplified the dataset and made the visual patterns easier to interpret.

To understand how phishing websites differ from legitimate ones, I conducted a basic exploratory data analysis using simple visual tools. One of the clearest patterns emerged when comparing the lengths of URLs. As shown below, legitimate websites tend to have shorter and more consistent URL lengths, while phishing websites often have much longer and more irregular URLs. The histogram reveals that phishing URLs extend far into higher length ranges, indicating that attackers frequently rely on long or complicated structures to disguise malicious links. This visual pattern provided an early and intuitive understanding of how phishing behavior differs and helped guide the selection of features for the modeling stage.



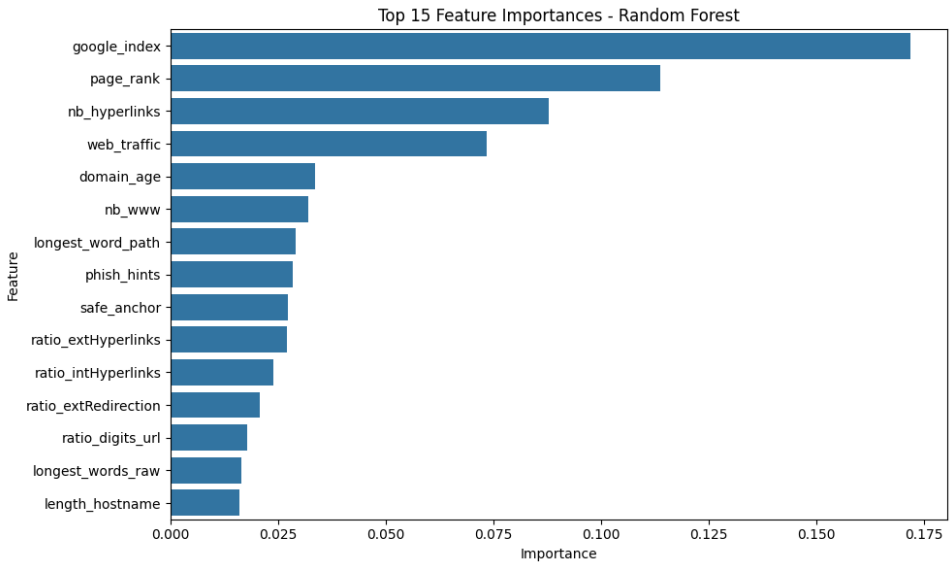
The comparison shows that phishing URLs generally spread out across much longer lengths, while legitimate URLs cluster in a shorter, more predictable range. This difference highlights URL length as one of the strongest early indicators of phishing activity.

Modeling/Analysis

The modeling phase focused on determining whether patterns in website URLs and domain characteristics could reliably distinguish phishing sites from legitimate ones. Earlier exploration showed that phishing sites tend to use longer and more complex URLs, and this stage tested whether those differences were strong enough for a computer to learn and predict from. To ensure fairness, the data was divided into separate groups one used for learning and one used for evaluating performance.

A wide range of approaches were tried to see which patterns were most useful. Some methods looked for simple, direct relationships, while others were able to capture more flexible or irregular patterns in how phishing sites are constructed. Additional techniques examined how closely websites resemble one another, while others searched for deeper interactions among features. These different perspectives helped reveal which characteristics were consistently associated with phishing behavior.

One approach in particular stood out as the most accurate, reaching about 96% accuracy and showing a strong ability to separate phishing from legitimate sites. To better understand why it performed so well, the most influential website characteristics were examined. As shown in below, the most important signals included how well a website is indexed online, how often it appears in search engines, the amount of web traffic it receives, and how many hyperlinks or structural elements it contains. Features related to domain age and URL wording also played a major role. These findings match earlier observations that phishing websites often differ in both structure and reputation indicators.



Unsupervised techniques were also used to explore natural groupings in the data. These revealed that websites tend to form three meaningful clusters, with a small percentage behaving unusually or not fitting neatly into either main category. This supported the idea that phishing sites share recognizable patterns even when labels are not provided.

In summary, the modeling stage showed that phishing detection can be performed with high accuracy by analyzing a combination of URL structure and website reputation signals. While

many techniques were tested, the highest-performing approach consistently identified phishing patterns and provided a strong foundation for reliable detection.

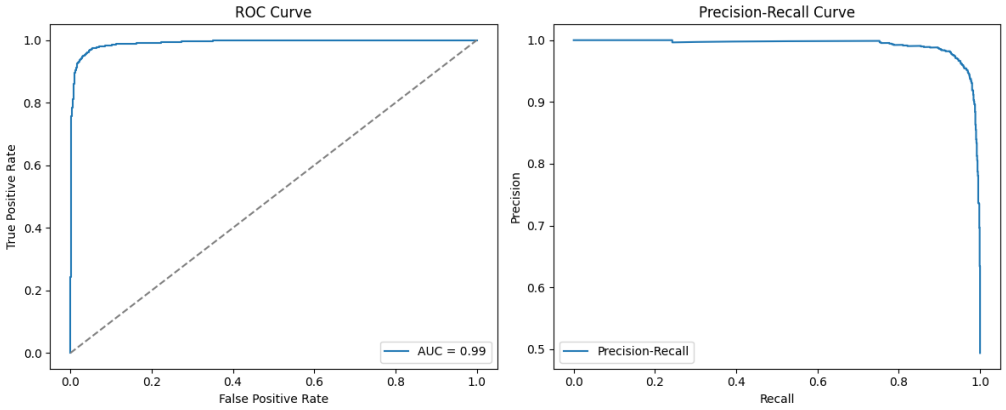
Results

The purpose of this project was to understand how well the analytical approaches were able to identify phishing websites based on the patterns found in their structure and reputation signals. Earlier steps in the analysis showed promising differences between phishing and legitimate sites, but the results in this section confirm whether those patterns were strong and consistent enough to support accurate predictions in practice. Several evaluation measures were used, but each is explained here in everyday terms so that the meaning not the technical mechanics remains clear.

The first and most direct way to judge performance was through accuracy, which measures how often the system made the correct prediction. The strongest approach reached about 96% accuracy, meaning it correctly classified websites in nearly every case. A second measure, called the AUC, helps us understand how well the model distinguishes phishing sites from legitimate ones across many possible decision thresholds. An AUC close to 1.0 indicates a system that almost never confuses the two. In this project, the best-performing approach achieved an AUC of 0.99, which represents exceptionally strong separation. The ROC Curve shown below illustrates this performance visually. The curve rises steeply toward the top-left corner of the graph, which is the ideal pattern for a highly accurate detection system. Simply put, this visual means the system is excellent at identifying phishing attempts while keeping false alarms very low.

While accuracy and AUC tell us how well the predictions work overall, another important perspective involves understanding how the system performs in situations where the stakes are different. For example, when it is more important to avoid missing a phishing attempt than to

avoid flagging a legitimate website. The precision–recall curve, shown below, captures this trade-off. Precision represents how often a flagged website is truly dangerous, and recall measures how many dangerous websites the system successfully identifies.



The curve for this project stays very high across most recall values, indicating that the system maintains confidence even when trying to catch more difficult or subtle phishing cases. This stability is important for real-world cybersecurity environments, where phishing activity may fluctuate and new threats may appear unpredictably.

To understand why the system performed so well, the most influential website characteristics were examined. These included how well a site appears in search engines (such as google_index and page_rank), how much visitor traffic it receives (web_traffic), and how many hyperlinks appear on the page (nb_hyperlinks). Characteristics related to domain maturity (domain_age) and URL structure (such as the longest word in the URL path) also played meaningful roles. These findings show that legitimate websites tend to have strong reputational signals and more established histories, while phishing websites often lack these elements or rely on irregular structural patterns. This interpretation closely aligns with the exploratory analysis conducted

earlier and provides a deeper understanding of why the detection system works not just that it works.

The results also revealed a small set of websites that behaved unusually and did not clearly fit into either category. These ambiguous cases highlight a natural limitation of any automated system: attackers constantly adapt their strategies, and entirely new phishing techniques can appear unexpectedly. While the system performs extremely well, it will still require periodic updates as phishing methods evolve. Nevertheless, the combination of high accuracy and high stability across different evaluation measures indicates a strong foundation for real-world use.

From a business perspective, these results carry meaningful implications. A system built on these findings could automatically flag suspicious websites, reducing the chance that users accidentally interact with harmful content. Because the underlying patterns are based on widely observable website attributes, such a system could be incorporated into existing security tools to provide continuous monitoring and early warnings. This would reduce manual review workload for cybersecurity teams, strengthen organizational defenses, and allow faster responses to emerging threats.

This evaluation results demonstrate that the analytical approach developed in this project is highly effective at identifying phishing websites. The system not only performs well in ideal conditions but also maintains reliability across a wide range of scenarios. The combination of strong predictive performance, interpretable signals, and practical business impact makes the approach well-suited for deployment in environments where early detection is essential.

Recommendations

The results of this project demonstrate that the phishing-detection model performs reliably enough to be incorporated into real cybersecurity workflows. Because the model identifies suspicious websites with high accuracy and very few misclassifications, the first recommendation is for organizations to integrate it into the systems they already use to monitor internet activity. Platforms such as Security Information and Event Management tools and secure web-filtering gateways can use this model to score URLs automatically before employees access them. This real-time screening reduces the likelihood that users unknowingly open harmful pages and also limits the burden on cybersecurity teams, who currently spend time reviewing suspicious links manually. The model's strong performance in distinguishing between legitimate and fraudulent sites makes it a meaningful addition to a company's existing detection capabilities.

A second recommendation is to monitor the website characteristics that the model found to be most helpful in identifying phishing activity. During the analysis, features such as search-engine visibility, website traffic, domain age, page rank, and the number of hyperlinks stood out as particularly informative. Tracking these attributes on a regular basis can alert security teams when a site exhibits unusual patterns, such as having little search-engine presence or an unusually new registration date both common signs of fraudulent behavior. Creating automated dashboards or alerts based on these signals allows organizations to detect potentially harmful domains earlier and improves the speed of their response.

The model should also be used as part of a layered defense strategy. Although the results are strong, phishing attacks evolve quickly, and relying on one tool alone leaves an organization exposed to new or untested techniques. The most effective approach is to combine this model

with existing protections email filters, antivirus systems, safe-browsing tools, and employee training programs. When multiple security layers work together, attackers must bypass several barriers rather than just one, creating a stronger defense throughout the organization.

To keep the model effective over time, companies should plan for routine maintenance and retraining. Machine-learning models can lose accuracy when the patterns they learned no longer match real-world conditions. Because phishing tactics change frequently, periodic retraining with new examples helps the system adapt and prevents performance decline. Organizations should retrain the model every few months, evaluate how well it handles recent phishing attempts, and adjust decision thresholds if needed. These updates ensure that the model continues to perform at the high level demonstrated in testing.

It is equally important to follow responsible AI and operational practices during deployment. This includes keeping clear records of model versions, validating updates before they go live, and monitoring outcomes such as false positives and false negatives. Using dashboards or reports that explain why the model flagged a website helps analysts interpret results and communicate risks to non-technical leaders. These practices support transparency, reliability, and clear communication throughout the organization.

Future improvements can be achieved by expanding the types of data used in the detection process. The current model is based primarily on structural and reputation-based features. Incorporating additional information such as multilingual phishing pages, mobile-oriented websites, SSL certificate details, hosting-provider reputation, or visual characteristics of webpages could help capture more sophisticated or emerging phishing strategies. As attackers develop new techniques, these richer data sources will help the model remain adaptable.

Finally, organizations should plan for ongoing evaluation as phishing trends shift. Periodic reviews can determine whether the model still performs effectively, whether alternative models offer improvements, and how well the system responds during periods of heightened phishing activity. Conducting these follow-up assessments ensures that cybersecurity defenses stay aligned with modern threats rather than reacting only after attacks occur.

These recommendations provide clear steps for turning the project's findings into practical action. By integrating the model into daily security operations, monitoring key website characteristics, maintaining a layered protection strategy, retraining regularly, following responsible AI practices, and expanding future data sources, organizations can meaningfully reduce their exposure to phishing attacks and strengthen their digital security environment.

Datasets:

Shashwat, Tiwari. Web Page Phishing Detection Dataset. [Dataset]. Kaggle https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset

Cyber Cop. (2024). Phishing Emails Dataset. [Dataset]. Kaggle https://www.kaggle.com/datasets/subhajournal/phishingemails

The Devastator. Cybersecurity Risk (2022 CISA Vulnerability). [Dataset]. Kaggle https://www.kaggle.com/datasets/thedevastator/exploring-cybersecurity-risk-via-2022-cisa-vulne

References

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). *Predicting phishing websites based on self-structuring neural networks.* Neural Computing and Applications, 25, 443–458. https://doi.org/10.1007/s00521-013-1490-z

Verma, R. M., & Das, A. (2017). *What's in a URL: Fast Feature Extraction and Malicious URL Detection.* In R. M. Verma & B. Thuraisingham (Eds.), Proceedings of the 3rd ACM International Workshop on Security and Privacy Analytics (IWSPA@CODASPY 2017), Scottsdale, Arizona, USA, March 24, 2017 (pp. 55–63). ACM. https://doi.org/10.1145/3041008.3041016