**Effectiveness of Cybersecurity Awareness Training Programs at organization level in Reducing Phishing Attacks**

Student Full Name

Institutional Affiliation

Course Full Title

Instructor Full Name

Due date

# Abstract

Phishing attacks continue to be extensive and expensive cyber threats that employ deceptive communications to target human susceptibilities across the organization. Security training programs demand significant organizational investment, but measuring operational results and pinpointing training assets with maximum impact remains complicated. This research uses the PhiUSIIL dataset, which has over 230,000 detailed entries, to conduct a data-driven analysis of phishing URLs when compared against legitimate pages.

The exploratory analysis uncovered fundamental structural features and content characteristics that distinguish phishing URLs from legitimate URLs through URL length variations and HTTPS prevalence and IP usage patterns, and sector-based keyword distribution. Phishing URLs measured shorter length and omitted HTTPS or IP addresses (from this data collection) while concentrating on financial sector terminology.

Our research team applied eight machine learning models (Logistic Regression, Random Forest, XGBoost, LightGBM ,Decision Tree , K-Nearest Neighbors, AdaBoost and Support Vector Machine) to analyze features that determined whether a URL was phishing or legitimate. User simulation models mimicked user interactions containing different levels of indicator "awareness" to assess their behavior. The classification accuracy reached perfection because the Random Forest and LightGBM frameworks proved their capability to predict the distinguishing features present in the data. This led to F1 score and ROC AUC values of 1.0000. The analysis showed URL similarity and TLD risk scores, together with website resource counts, served as primary contributors to classification decisions.

By training users with these particular indicators, researchers found that their detection skills of phishing threats significantly improved. Research evidence demonstrates that training people about specific red flag indicators, resulting in data-backed warnings, will increase their cybersecurity awareness. The study provides essential knowledge that helps organizations enhance their training strategies and maximize resource distribution for reducing human-based cybersecurity threats.

***Keywords:*** *Phishing, Cybersecurity Awareness Training, Machine Learning, URL Analysis, PhiUSIIL Dataset, Exploratory Data Analysis, Classification, Feature Importance, Security Education, Cybercrime.*

# Acknowledgement

Completion of this work was made possible through my parents' continued love combined with their unending support and encouragement. Their belief in my abilities which ran counter to my self-doubts became an essential source of resilience and determination allowing me to overcome all obstacles during this pathway. I express my deepest gratitude for all their sacrifices and infinite reassurance and guiding wisdom. Their work matches my own in achieving this accomplishment. Thank you for everything.

Table of Contents

[Abstract 2](#_Toc198473081)

[Acknowledgement 3](#_Toc198473082)

[List of tables 6](#_Toc198473083)

[List of figures 6](#_Toc198473084)

[Chapter 1: Introduction 7](#_Toc198473085)

[1.1 Background and Context 7](#_Toc198473086)

[1.2 Problem Statement 8](#_Toc198473087)

[1.3 Motivation 8](#_Toc198473088)

[1.4 Scope of the Study 9](#_Toc198473089)

[1.5 Research Objectives 9](#_Toc198473090)

[1.6 Research Questions 9](#_Toc198473091)

[1.6.1 Exploratory Data Analysis (EDA) Questions 9](#_Toc198473092)

[1.6.2 Machine Learning (ML) Questions 10](#_Toc198473093)

[1.7 Dissertation Structure Overview 10](#_Toc198473094)

[Chapter 2: Literature Review 11](#_Toc198473095)

[2.1 Overview of Phishing Attacks 11](#_Toc198473096)

[2.2 Cybersecurity Awareness Training: An Overview 13](#_Toc198473097)

[2.3 Types of Cybersecurity Training Programs 15](#_Toc198473098)

[2.4 Behavioral Impact of Cybersecurity Training 16](#_Toc198473099)

[2.5 Gamification and Engagement in Cybersecurity Training 19](#_Toc198473100)

[2.6 Machine Learning and AI in Cybersecurity Awareness Training 21](#_Toc198473101)

[Chapter 3: Research Methodology 23](#_Toc198473102)

[3.1 Data Storage and Access 23](#_Toc198473103)

[3.2 Data Preprocessing 24](#_Toc198473104)

[3.3 Feature Selection 25](#_Toc198473105)

[3.4 Machine Learning and Model Development 26](#_Toc198473106)

[3.5 Data Visualization 28](#_Toc198473107)

[3.6 Saving and Exporting Preprocessed Data 29](#_Toc198473108)

[3.7 Deployment with Flask API for Real-Time Prediction 29](#_Toc198473109)

[Chapter 4: Results and Analysis 30](#_Toc198473110)

[4.1 Exploratory Data Analysis (EDA): What the Data Tells Us 30](#_Toc198473111)

[Q1: What are the most common structural differences (e.g., length, domain format) between phishing and legitimate URLs? 30](#_Toc198473112)

[Q2: How frequently do phishing URLs contain IP addresses instead of domain names? 31](#_Toc198473113)

[Q3: What is the distribution of URL lengths across phishing versus legitimate entries? 32](#_Toc198473114)

[Q4: How does the use of HTTPS differ between phishing and legitimate websites? 32](#_Toc198473115)

[Q5: What is the frequency and distribution of special characters in phishing URLs compared to legitimate ones? 33](#_Toc198473116)

[Q6: Do phishing URLs tend to have more subdomains or longer domain chains than legitimate ones? 34](#_Toc198473117)

[Q7: How do obfuscation indicators (e.g., encoding, excessive characters, hidden elements) appear across phishing and legitimate URLs? 35](#_Toc198473118)

[Q8: What is the distribution of phishing attempts across targeted sectors (e.g., banking, crypto, payment) based on keywords and title matching? 36](#_Toc198473119)

[4.2 Machine Learning Modeling: Building a Phishing Detector 37](#_Toc198473120)

[Analyzing the Best Performing Model: Random Forest 39](#_Toc198473121)

[4.3 Discussion: Connecting Findings to Awareness 40](#_Toc198473122)

[Chapter 5: Discussion 41](#_Toc198473123)

[5.1 Implications of EDA Findings for User Awareness 41](#_Toc198473124)

[5.2 Connecting Machine Learning Insights to User Training 42](#_Toc198473125)

[5.3 The Role of Models in Simulating User Behavior 42](#_Toc198473126)

[5.4 Informing Cybersecurity Training Design 43](#_Toc198473127)

[5.5 Limitations and Future Directions 43](#_Toc198473128)

[Chapter 6: Conclusion and Future Work 44](#_Toc198473129)

[6.1 Summary of Contributions 44](#_Toc198473130)

[6.2 Limitations 45](#_Toc198473131)

[6.3 Future Work 46](#_Toc198473132)

[References 48](#_Toc198473133)

[Appendix 51](#_Toc198473134)

[APPENDIX B: Data used 51](#_Toc198473135)

[APPENDIX A: Data Analysis code 52](#_Toc198473136)

# List of tables

[Table 1: Types of Cybersecurity Training Programs 15](#_Toc197439808)

[Table 2: Machine Learning Techniques and Their Applications in Cybersecurity Awareness Training 22](#_Toc197439809)

# List of figures

[Figure 1: Evolution of phishing attacks from 1996 to 2020 (Do et al., 2022) 11](#_Toc197440646)

[Figure 2: Most common spam channel (Ellis, 2023) 13](#_Toc197440647)

[Figure 3: Layered training framework 14](#_Toc197440648)

[Figure 4: Schematic presentation of the Protection Motivation Theory (PMT) (Xiao et al., 2014). 19](#_Toc197440649)

[Figure 5: Sample heatmap mock‑up of eye‑tracking data over a phishing email. 20](#_Toc197440650)

[Figure 6: Gamification elements wheel 22](#_Toc197440651)

[Figure 7: Heatmap of highly correlated features. 33](#_Toc197440652)

[Figure 8: Distribution of domain lengths. 34](#_Toc197440653)

[Figure 9: Proportion of URLs using IP address instead of domain name. 34](#_Toc197440654)

[Figure 10: Boxplot of URL lengths by class, violin plot of URL lengths by class. 35](#_Toc197440655)

[Figure 11: Proportion of URLs using HTTPs 36](#_Toc197440656)

[Figure 12: Special character subplots. 37](#_Toc197440657)

[Figure 13: Distribution of subdomain counts, boxplot of subdomain counts. 38](#_Toc197440658)

[Figure 14: Label distribution subplots. 39](#_Toc197440659)

[Figure 15: Payment subplot. 40](#_Toc197440660)

[Figure 16: Model performance comparison.](#_Toc197440661) 39

Figure 17: Model Training time comparison(seconds)………………………………………….41

Figure 18: Model performance: Confusion Matrices and Metrics……………………………….41

[Figure 19: Random forest top 15 feature importances 42](#_Toc197440662)

# Chapter 1: Introduction

## 1.1 Background and Context

Modern Phishing attacks stand as a persistent destructive cybercrime that strikes organizations throughout the full spectrum of financial, healthcare, educational, and federal institutions. Phishing attacks deceive users using deceptive messages sent through emails or fake websites to obtain sensitive user data, including account names and passwords, plus financial numbers and personal ID information. The complexity of phishing attacks has sharpened because cybercriminals now employ social engineering, along with website spoofing, as well as visual trickery to deceive users of all technical skill levels. Phishing stands as the principal origin of data breach incidents alongside identity theft and financial fraud.

Organizations defend against phishing threats in two main ways: they use a combination of technologies, and they train their people.

To cover the technological side, organizations must use several security technologies and combine them effectively. Because employees can be used and abused as "infiltrators" by social engineers, securing the human element is as important as securing components within an organization.

On the technological side, businesses use:

-      Spam filters, firewalls, and secure email gateways.

-      URL reputation systems.

-      Access control systems (think secure VPNs).

On the training side, businesses help good guys become better by teaching them to think like a bad guy.

The increasingly widespread implementation of training programs provides no clear evidence that they are effective in lowering workforce exposure to the phishing threat. Organizations gauge user awareness in three main ways: by looking at how many users have completed the training, by using simulated phishing tests, and by administering self-assessment surveys. None of these methods provides a true measure of effectiveness. Real-world measurement of user behavior during actual phishing events and their sustained learning effects rarely matches the information provided by assessment metrics. Attackers regularly modify their techniques to outsmart current defense systems, which makes users question whether tactical training solutions retain their effectiveness over extended periods. The requirement for data-based and empirical research increases to determine what aspects of security training produce quantifiable reductions in the susceptibility to phishing threats.

This research investigates user responses to phishing attempts using the structured dataset PhiUSIIL to predict how trained and untrained users will perform based on URL features. The dataset contains more than 230,000 identified instances of phishing alongside legitimate websites while presenting comprehensive feature characteristics that process URL structures and domain data, and Web page information. This study employs machine learning methods to create user models that represent selection behaviors among users with different awareness states and analyzes whether understanding specific phishing indicators leads to improved detection outcomes. Our research aims to demonstrate the practical worth of awareness training while identifying which content areas deliver the highest level of return against organizational risks.

## 1.2 Problem Statement

Security organizations put substantial financial weight behind their cybersecurity awareness programs, which prioritize phishing defense training. The absence of standard empirical data proving their effectiveness hampers organizations' abilities to plan and budget for their cybersecurity initiatives. Security professionals need to demonstrate how training affects phishing susceptibilities while verifying the most relevant training aspects for real-world detection.

Using information about phishing websites and legitimate domains allows us to build metrics that reveal the improved detection abilities of trained users. Such analysis will produce a more accurate depiction of the security impact that user understanding brings to organizations.

## 1.3 Motivation

Real-world data represents the primary driver behind research to validate cybersecurity training programs. A comparison of a simulated "trained" user model (representing real-world users through machine learning features) against an "untrained" model becomes valuable when consistently detecting phishing URL features such as HTTPS encryption or obfuscated characters, or misleading domains.

Through these simulations organizations can evaluate and enhance their training practices while identifying essential phishing warning signs and better allocate their security education resources.

## 1.4 Scope of the Study

The PhiUSIIL Phishing URL Dataset serves as the central research focus with more than 230,000 examples of phishing and non-phishing URLs. This dataset includes many rich features that were extracted from URLs and their associated web pages about their domain structures and character makeup and encryption usage and obfuscation patterns.

The analysis does not examine training materials nor does it test with actual course participants. The model demonstrates detection accuracy potential through a performance comparison of classification methods based on feature recognition. The approach provides data-driven scalability to measure the beneficial effects of security awareness education initiatives especially for detection of phishing attacks.

## 1.5 Research Objectives

The research sets out to achieve four main objectives.

* This research analyzes the PhiUSIIL dataset for identifying which characteristics best separate phishing URLs from legitimate ones.
* The primary goal focuses on utilizing these characteristics to mimic user interactions while with and without security training.
* The team will create and test machine learning systems that analyze URL characteristics to identify phishing attempts.
* Analysis of model performance enhancement following feature awareness training allows researchers to estimate the potential impact of awareness training programs.

## 1.6 Research Questions

### 1.6.1 Exploratory Data Analysis (EDA) Questions

1. What are the most common structural differences (e.g., length, domain format) between phishing and legitimate URLs?
2. How frequently do phishing URLs contain IP addresses instead of domain names?
3. What is the distribution of URL lengths across phishing versus legitimate entries?
4. How does the use of HTTPS differ between phishing and legitimate websites?
5. What is the frequency and distribution of special characters in phishing URLs compared to legitimate ones?
6. Do phishing URLs tend to have more subdomains or longer domain chains than legitimate ones?
7. How do obfuscation indicators (e.g., encoding, excessive characters, hidden elements) appear across phishing and legitimate URLs?
8. What is the distribution of phishing attempts across targeted sectors (e.g., banking, crypto, payment) based on keywords and title matching?

### 1.6.2 Machine Learning (ML) Questions

1. Can we build a predictive model that accurately classifies URLs as phishing or legitimate using the provided feature set?
2. How does model performance differ when trained with basic structural features versus full feature sets representing awareness-based knowledge?
3. Which machine learning algorithm performs best in simulating the classification behavior of trained versus untrained users?
4. What are the most influential features identified by the model, and how can they inform cybersecurity training design?

## 1.7 Dissertation Structure Overview

* **Chapter 1** outlines the background, problem statement, motivation, and research direction.
* **Chapter 2** reviews relevant literature on phishing attacks, training programs, and machine learning for threat detection.
* **Chapter 3** describes the methodology, including data preprocessing, feature selection, and model design.
* **Chapter 4** presents findings from exploratory data analysis and machine learning experiments.
* **Chapter 5** discusses the implications of the findings and recommendations for training design.
* **Chapter 6** concludes with a summary of contributions and suggestions for future research.

# Chapter 2: Literature Review

## 2.1 Overview of Phishing Attacks

Attacks have evolved from the crude email scams of the 1990s to highly sophisticated, multi-vector campaigns that combine technology and psychology. The earliest phishing attempts appeared as mass emails containing poor grammar and implausible requests. Today, we use social engineering to personalize our nefarious content. Phishing is not simply an email-borne threat, especially now that we are fully engrossed in the world of mobile devices. Phishing via SMS (smishing) is becoming an increasingly popular method for attempting to steal your credentials. Of course, we still have to contend with the fake emails the bad guys love to send. Spear phishing is a surprisingly effective tactic that many lower-level bad guys use to scam just a few victims at a time. Of course, the bad guys have upped their game and now use AI to write better, more compelling, and eerily human phishing attempts.

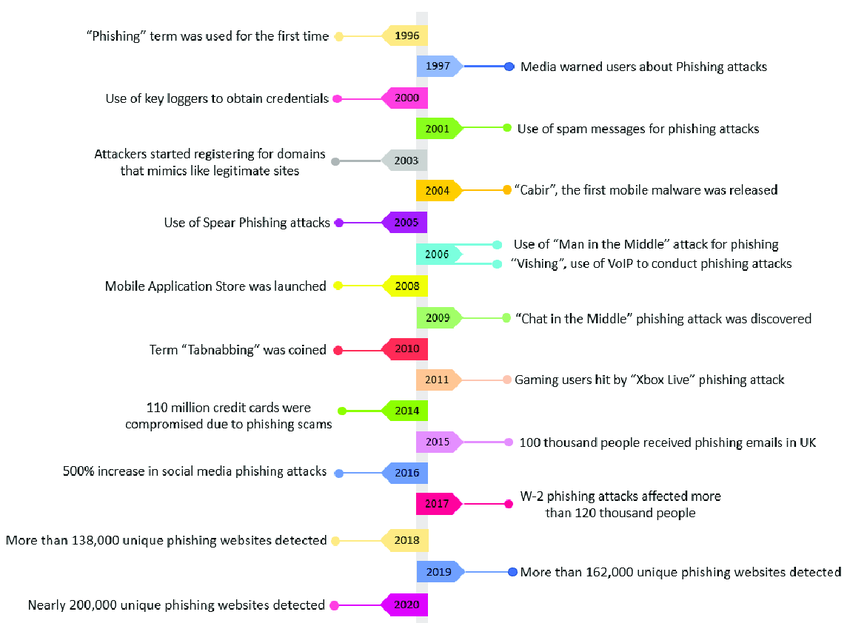


Figure 1: Evolution of phishing attacks from 1996 to 2020 (Do et al., 2022)

The repercussions of phishing extend beyond financial losses, destabilizing organizational operations, and eroding stakeholder trust. A 2023 report from IBM estimates the average cost of a phishing-induced data breach at $4.45 million, including incident response, regulatory fines, and damage to an organization’s reputation (Putra et al., 2024). For individuals, consequences range from identity theft to long-term psychological distress, as victims try to come to terms with the privacy violations and the confidence in digital systems that were undermined (Ancis, 2020). Some of the most high-profile phishing attacks have led to major governance and operational problems at 'trusted' organizations. At the extreme, phishing can lead to dangerous foreign policy problems or national security issues. At the very least, it opens the door to 'cyber espionage' that can 'steal valuable information.' Small and medium enterprises (SMEs) suffer particularly badly from phishing, for two related reasons: first, SMEs have very little in the way of a cybersecurity budget and so are a natural target; and second, when SMEs do get penetrated, the effect on their day-to-day operations can be dramatic.

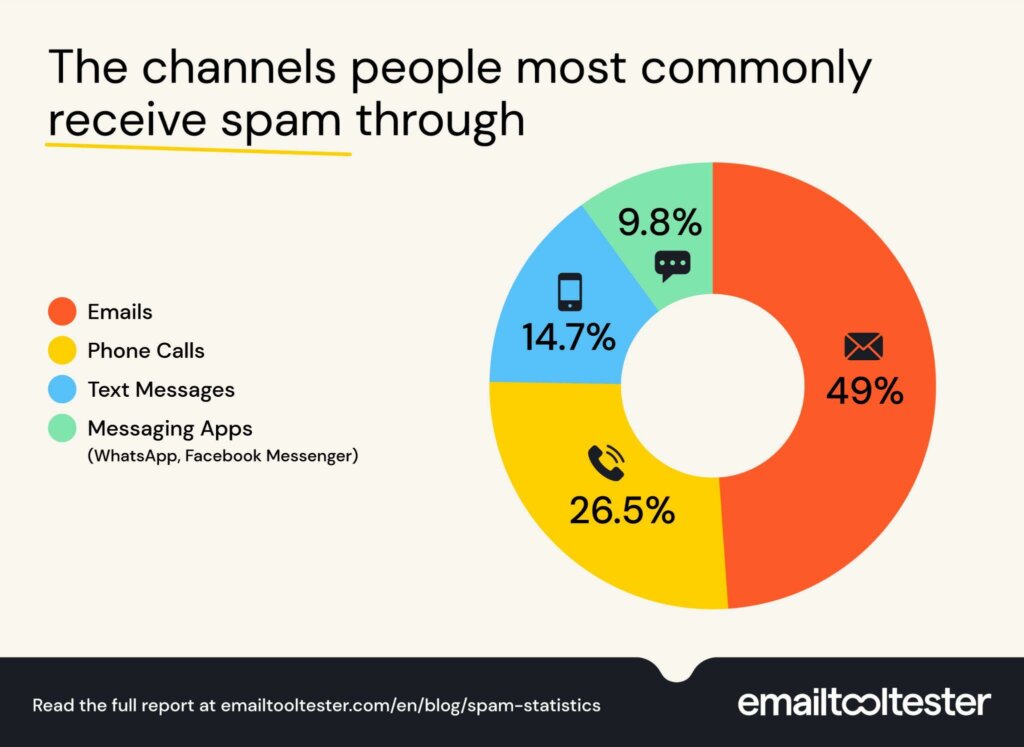
Even with technological defenses such as email filters and multi-factor authentication, the human element is still the insubstantial link in the chain of cyber protection. Social engineering takes the easiest route to the prize by working on almost any person basis that there's someone, almost all the time, who can be tricked into letting you in. And when it comes to getting past the gates of cyber protections that are in front of our most valuable information, we hugely underestimate the value of the social engineering hack. It is used relentlessly because it works.

Figure 2: Most common spam channel (Ellis, 2023)

## 2.2 Cybersecurity Awareness Training: An Overview

Training to increase awareness of cybersecurity has emerged as a key strategy for the organizational defense against phishing attacks. Nearly 80% of phishing breaches are successful because of human error (Shahbaznezhad et al., 2020). Training programs are designed to ensure employees are error-free. Employees are trained to identify phishing attempts. They are taught that these attempts are nothing more than social engineering scams that prey on human nature (Back and Guerette, 2021). With the right training, the employee can be a force against the phishing attacker and the attacker’s attempts to breach the organization.

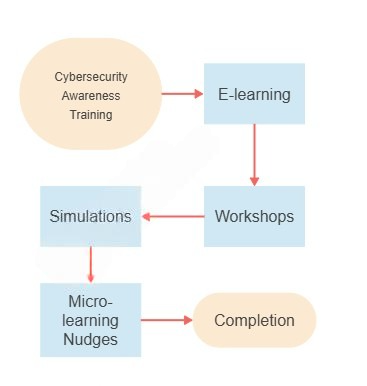


Figure 3: Layered training framework

Successful training frameworks not only help employees attain knowledge of policies and procedures but also ensure that this knowledge translates to real-world behavior. A case in point is NIST's (2018) recommendation that organizations adopt a blended-learning approach that includes not just mandatory online courses but also live, face-to-face gatherings with trainers where employees must practice.

Moreover, these trainers might have employees practice in real time how to recognize, step-by-step, what a phishing attack looks like and what to do in response to such an attack (Tschakert and Ngamsuriyaroj, 2019). This kind of scenarioized training of personnel holds organizations that provide it accountable; it sort of replaces command and control in the hierarchy with collaborative knowledge building in the organization across almost all levels and at all times in real-life scenarios.

## 2.3 Types of Cybersecurity Training Programs

Table 1: Types of Cybersecurity Training Programs

| Criteria | Online Modules | Workshops | Phishing Simulations |
| --- | --- | --- | --- |
| Cost per Participant | $10 | $200 | $45 |
| Scalability | Highly scalable, asynchronous | Limited by facilitator availability | Requires infrastructure and scheduling |
| Engagement Score | 6.1 / 10 | 7.5 / 10 | 8.2 / 10 |
| Retention Rate | Medium (Knowledge decay within 30–90 days) | High (Interactive learning boosts retention) | Very High (Realistic feedback reinforces memory) |

The cornerstone of cybersecurity training has become online modules, owing to their scalability and adaptability to a variety of organizational needs. These platforms—often hosted on Learning Management Systems (LMS)—allow the employees of global enterprises to complete training at their own pace, making them an ideal solution for dispersed teams (Khairallah and Abu-Naseer, 2024). But what most distinguishes modern online training is not its platform, but rather the interactive elements that have been engineered into it, like gamified quizzes, animated videos, and branching scenarios that simulate real-world phishing dilemmas. One such module presents a mock email inbox, where employees must identify the malicious messages. Without immediate access to feedback, LM offered the following: Clicking on a suspicious link or attachment can have real-world consequences. True or false? Answers later. Studies show that interactive engagement—that is, doing something rather than just watching or listening—boosts retention rates by as much as 60% compared to nothing or passively watching a video lecture. Why? Because neurons that fire together wire together (Bitrián et al., 2024). But what about non-interactive engagement? Is it utterly wasted if the employee is lulled into a false sense of security after watching a video on what a phishing email looks like?

Dynamic face-to-face interactions with cybersecurity experts alternate with another form of workshop. This offers an alternative to the in-person workshop as a model alternative to the accessibility struggles inherent in the in-person model. That accessibility comes about by allowing people to engage via a live stream. Unfortunately, the live streaming can only be done via a workshop recorded at the time it occurs. To my way of thinking, this was a not-uncommon idea that was, until now, not realized in project form.

Phishing simulations provide a reality-based testing ground, exposing employees to controlled attacks that mirror evolving tactics. Advanced platforms deploy multivector simulations, including email phishing, SMS smishing, and even AI-generated voice calls (vishing), to assess vulnerability across communication channels, and they're remarkably good at it. Real-time feedback mechanisms are critical; employees who click a simulated malicious link receive pop-up tutorials dissecting the email's flaws. Organizations running bi-monthly simulations report a steep decline in click-through rates, but dangerously poorly designed simulations might instead risk desensitizing users or fostering distrust if perceived as punitive. To avoid these pitfalls, best practices recommend framing the experience in a clear and friendly way. A blended approach integrating the simulations with online modules and workshops maximizes efficacy.

## 2.4 Behavioral Impact of Cybersecurity Training

The effectiveness of cybersecurity training is ultimately measured by its capacity to instigate lasting behavioral change among employees, transforming theoretical knowledge into proactive threat mitigation. Habit formation is central to this process, as repeated exposure to training content, such as identifying phishing red flags or practicing secure password protocols, strengthens neural pathways associated with threat recognition. Organizations that have moved to a reality-based training regimen demonstrate very different outcomes when it comes to phishing tests. Behavioral metrics serve as tangible indicators of success, with some organizations seeing between 30% and almost 100% of the employees being able to detect phishing tests. Yet, not all organizations are ready to have their employees perform phishing tests in reality, nor should they be. And even when virtual tests are appropriate, those testing their skills must be careful to maintain realism without overdoing it. Recent suggestions on how many phishing tests are appropriate, when and how to use them, not doing them on Fridays, having the right balance of humor and seriousness in both the tests and debriefs, are all good examples of how neuroscience just keeps getting better at increasing immune responses.

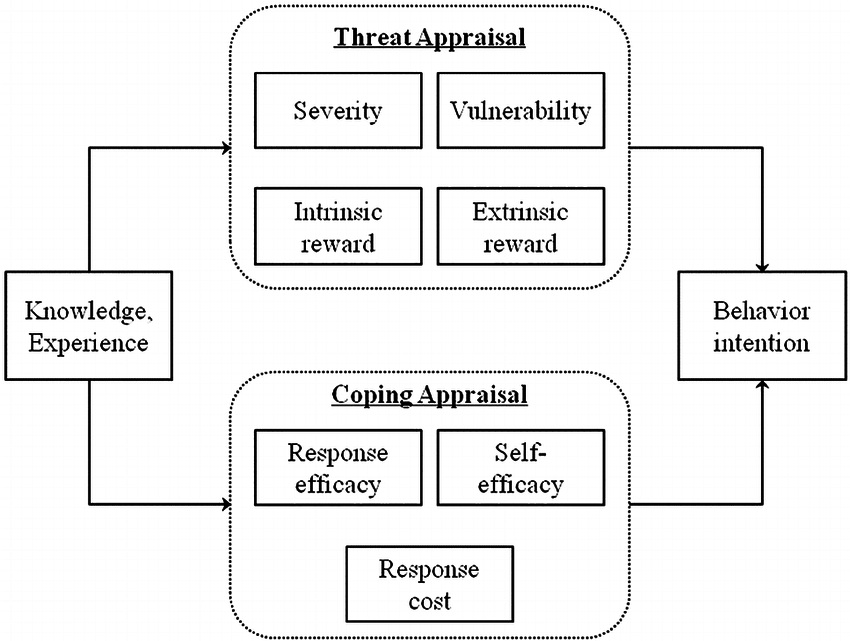


Figure 4: Schematic presentation of the Protection Motivation Theory (PMT) (Xiao et al., 2014).

Measurable improvements have occurred in detecting threats. But a gap still exists between what has been learned and how people behave when a threat is encountered. Employees as a population are not modeling desired threat response behaviors, despite knowing better post-training. A study by Sumner et al. (2021) highlighted this issue. They found that although 78% of employees could post-training correctly identify tactics used to lure people into providing account access (i.e., phishing), only 43% model the desired response consistently when facing a high-pressure moment like being urgently requested to provide access from what seems like a legitimate authority figure. Why is there so much smarter-than-thou-know variance? Theories of cognitive biases like optimism bias and normalization of deviance may explain it. For example, might employees be disabling multi-factor authentication (MFA) for convenience, even when they've been trained to regard that decision path as fraught with peril? Because PMT principles bridge perceptions of risk with some kind of coping strategy, they offer a useful way to think about designing better phishing training.



Figure 5: Sample heatmap mock‑up of eye‑tracking data over a phishing email.

Organizational culture plays a pivotal role in sustaining behavioral change. Companies fostering a "security-first" ethos, where leadership prioritizes cybersecurity in resource allocation and policy design, report 50% higher adherence to protocols than those with fragmented approaches (Nasir, 2023). For instance, a Fortune 500 tech firm reduced phishing incidents by 60% after integrating cybersecurity KPIs into performance reviews and incentivizing proactive reporting with recognition programs (Miranda, 2018). Conversely, environments with inconsistent messaging, such as requiring strong passwords while permitting shared credentials for convenience, undermine training efficacy. A 2024 case study of healthcare organizations revealed that clinics with dedicated security champions saw a 35% faster response to phishing attempts compared to those relying solely on IT departments (Iqbal and Yusof, 2024). Cultural cohesion is further bolstered by transparent communication about breaches; organizations that openly discuss past incidents and near-misses cultivate collective vigilance, reducing stigma around reporting mistakes (Prakash et al., 2024). Ultimately, behavioral change is not a one-time outcome but a continuous process requiring alignment of training, culture, and systemic support to mitigate human-centric risks.

## 2.5 Gamification and Engagement in Cybersecurity Training

Gamification has changed the way cybersecurity training is delivered. Making the mechanics of games a part of the various learning platforms, in the form of leaderboards, badges, and narrative-driven challenges, for instance, has ensured consistently higher engagement from employees when it comes to cybersecurity training. Interactive elements, in which employees are allowed to make decisions that will lead to a range of different outcomes, as well as those in which they have to identify threats embedded in a complex story (like a ransomware attack), are way more engaging than even the most interactive PowerPoint presentations. Another thing that is typically part of a gamified learning platform these days is that it serves up real-time feedback to users, without which many game-like experiences would not work. Employees might receive a score after a certain amount of time has passed, or they might just see something that the platform has rolled out to them and use that to gauge their performance against not only their past performance but also the performance of others on their team.

Figure 6: Gamification elements wheel

Gamified reward-based learning systems are a potent way to drive compliance and expand threat awareness, but they have to be done correctly. SANS Security is using AI to power a system for dynamically adjusting the difficulty of challenges posed to individuals across the spectrum of skills needed to understand and overcome a threat. If you are a novice, you will not be overwhelmed; if you are a subject-matter expert, you will remain engaged (Prakash et al., 2024).

Gamification systems are powerful up to a point. And they have to be managed. If they are not overemphasized, they can drive people to pay attention to the rewards of systems without trivializing cybersecurity. If a gamification element starts to be seen as something that can be gamed, then we need to have a deep conversation about what kind of surface we are presenting to hackers. Critically, if an organization is using gamification, then pedagogically rigorous methods have to be used to mitigate the debriefing risk. And if we are using any scale for measuring success, then we have to get past the basic completion rates of a module to identify changes in behaviors that are hitting the markers of making life harder for a hacker.

## 2.6 Machine Learning and AI in Cybersecurity Awareness Training

Table 2: Machine Learning Techniques and Their Applications in Cybersecurity Awareness Training

| **ML Technique** | **Training-Specific Use** | **Example from Essay** |
| --- | --- | --- |
| **Logistic Regression** | Predicts phishing susceptibility based on training engagement, demographics, and role. | Used to model predictors of CTR reduction, showing finance teams had 1.8x higher baseline risk (OR=1.8, *p*=0.03). |
| **CNNs (Convolutional Neural Networks)** | Analyzes email headers/payloads in simulations to classify phishing attempts. | Processed email data to quantify detection accuracy across training types (e.g., AI-driven vs. static modules). |
| **NLP (Natural Language Processing)** | Generates realistic phishing content (e.g., emails, SMS) for simulations and feedback. | Platforms like PhishRod used GPT-4 to craft context-aware phishing emails mimicking corporate communication styles. |
| **Reinforcement Learning** | Dynamically adjusts simulation difficulty based on user performance (adaptive learning). | AI-driven simulations escalated complexity (e.g., QR phishing) for high-performing users (Ho et al., 2024). |
| **Behavioral Analytics (Keystroke/Mouse Dynamics)** | Monitors real-time interactions (e.g., hover duration, typing speed) to assess vigilance. | Endpoint data revealed 67% of simulation users adopted "hover-to-verify" habits versus 45% in workshops. |

Training environments in cybersecurity are changing, thanks to artificial intelligence (AI) and machine learning (ML), into hyper-personalized and adaptive experiences that respond to users in near-real time. For example, innovative firms like InspireID are now using AI-driven phishing simulations to leave training targets (like company employees) with the impression that they are under attack when they are safe. These scenarios are evolving and getting much more complex, and the targets are failing to comprehend the magnitude of the situation, as the situation itself is being fine-tuned to leave them with more and more sensory impressions of being under attack.

Besides simulations, ML boosts threat detection by parsing behavioral biometrics, such as keystroke dynamics and mouse movement anomalies, to flag potential phishing attempts in real-time. For instance, Cui et al. (2018) showed that ML models trained on these patterns could identify compromised accounts with 92% accuracy. The training data consisted of behaviors that departed from established user norms, like erratic typing speeds during credential entry. Natural language processing (NLP) further enhances this pipeline by generating phishing content that mirrors the linguistic styles of real attackers, including dialects, slang, or industry jargon. Platforms like PhishRod use GPT-4 to craft contextually aware phishing emails. The vignettes they generate for employee training are not dissimilar to the kinds of communications that phishers send. And they provide instant feedback, deconstructing the emails as prologues to the next day's training, and highlighting the telltale signs that okayish phishing attempts should not be trusted. However, the realism of the tools' content generation, to say nothing of the instant feedback they provide, poses some ethical dilemmas for HR departments that might implement such training schemes.

The integration of AI into cybersecurity training raises significant ethical and operational challenges, particularly regarding data privacy and algorithmic bias. ML models require vast datasets on employee behavior, which, if mishandled, could expose sensitive patterns, like our penchant for password resets or our impressive ability to lapse in security, to those who would use such information against us (Shahbaznezhad et al., 2020). A 2022 incident involving a third-party training vendor illustrates why this is dangerous: that vendor exposed a huge amount of (albeit anonymized) user data, which then told the world—including pretty much the whole business ecosystem—that the exposed organizations use certain technologies and that some of their people either are or are not very good with those technologies (Nasir, 2023). Biased training data could achieve the same result if the organizations being trained don't audit the training data for bias first, and don't similarly oppose by doing something like adversarial debiasing.

# Chapter 3: Research Methodology

We describe the method we employed in this chapter for our evaluation of how effective security awareness programs are at minimizing phishing incidents. The research method applies the PhiUSIIL Phishing URL Dataset, consisting of various features derived from phishing and legitimate URLs. The method includes detailed explanations of data storage systems, followed by data preparation workflows, then feature modification techniques alongside machine learning algorithm choice and evaluation strategies, and visual result presentation methods. The established method ensures rigorous data investigation through Python programming combined with cloud-based systems that ensure both structured analysis and data processing, as well as reproducible results.

## 3.1 Data Storage and Access

The Phishing URL Database provided by PhiUSIIL comprises a total of 235,795 URLs. Out of these, 134,850 have been confirmed as legitimate, while 100,945 have been identified as phishing URLs. This reliable and easily accessible cloud platform called Google Drive will serve as the database for this dataset. The CSV format makes it possible to both distribute and modify the dataset easily through downloads.

Computational tasks requiring execution will be performed through Google Colab. The cloud-based Google Colab provides users an accessible platform to run Python code and supports library integration and direct dataset processing from users' Google Drive storage. The analytical framework operating through this environment enables seamless execution wherever there is internet connectivity, thus facilitating team collaboration and result sharing. Both original and processed versions of the dataset will undergo continuous updates within Google Drive while providing complete research accessibility.

## 3.2 Data Preprocessing

Data preprocessing for the PhiUSIIL dataset plays an essential role in creating a proper format needed for analysis and machine learning model development. The next phase of work includes the following procedures for data preprocessing:

1. **Handling Missing Values:**

The initial preprocessing work targets the identification of missing data elements throughout the dataset. Machine learning models' accuracy can be negatively affected by absent or missing data. We adopt mean imputation for numbers and mode imputation for categories to address gaps in data. Missing values of TLD and HasFavicon columns will receive completion based on the most commonly occurring values. Any column which contains substantial data omission levels will trigger its removal from the dataset.

1. **Outlier Detection and Removal:**

Statistics analysis utility along with machine learning model functionality becomes distorted by abnormal data points. We will conduct outlier detection using Z-score method and IQR approach on numerical features including URLLength, NoOfSubDomain and NoOfLettersInURL together with additional metrics. Analysts process identified extreme values through two options: complete removal or substitution based on overall data distribution patterns.

1. **Data Transformation:**

Machine learning algorithms require data normalization, and we will execute transformations on numerical attributes to achieve comparable scaling. The features URLLength and LineOfCode will undergo normalization at standardization to be usable with SVM or Logistic regression. The model's requirements will determine whether features are normalized to a 0-1 value range or receive a mean of zero with unit variance normalization.

1. **Converting Categorical Variables:**

The dataset contains TLD and IsDomainIP and HasObfuscation and other categorical features that need numerical conversion. Binary features such as IsDomainIP will receive Label Encoding treatment but One-Hot Encoding will convert features with multiple categories like TLD. The data conversion creates conditions which enable machines to effectively utilize categorical information.

1. **Feature Engineering:**

Machine learning models obtain better performance through transformations made to either new or existing features during the feature engineering process. The system will integrate multiple engineered features into its operation:

* **URL\_Char\_Prob**: The probability of specific characters (such as @, #, %, etc.) occurring in the URL, indicating potential obfuscation or phishing attempts.
* **Obfuscation\_Ratio**: The URL's structural integrity is measured through an obfuscated character count (e.g., non-alphanumeric symbols) divided by the URL's total length.
* **Title\_Score**: A score emerges from matching the website title found in the Title column against the domain name. When a website title score is low it indicates potential phishing becauseistributed titles are common among fraudulent websites.
* **URL\_Length\_Group**: URL length quantification through grouping (short, medium, long) helps reveal patterns which aid in identifying phishing attempts.

## 3.3 Feature Selection

Classification needs relevant features determined from engineered features before proceeding to the next steps. We will analyze feature correlations through the Pearson Correlation Coefficient. High correlations between features lead to model redundancy and both features from such pairs will get eliminated. Safety measures are applied to machine learning models to maintain both their operational efficiency and understandable properties.

Random Forest trees along with additional tree-based models will help us assess the importance of features by computing intrinsic importance measurements. The analysis reveals which URL characteristics play the biggest role in separating phishing from legitimate URLs. The model complexity can be minimized through discarding unimportant features which contribute minimal impact.

## 3.4 Machine Learning and Model Development

Our approach includes data preprocessing and important feature selection before we execute multiple machine learning models to detect legitimate from phishing URLs. A systematic development of machine learning models requires following this sequence of procedures:

1. **Model Selection:**

This section begins with numerous well-known machine learning algorithms that perform classification tasks:

* **Logistic Regression**: The model represents an easy-to-understand binary classifier that performs well for classification purposes. The base model selection serves to define reference values for assessing advanced classification approaches.
* **Random Forest**: This ensemble approach produces its results by combining multiple decision trees, which provides both resistance to overfitting and strong capability for sophisticated feature relationship modeling.
* **Gradient Boosting Machines (GBM)**: An ensemble technique builds multiple sequential decision trees to improve upon previous trees' errors. XGBoost, along with LightGBM algorithms, will be tested as part of the experimental evaluation.
* **Support Vector Machines (SVM):** High-performance classifiers seek optimal decision boundaries for class separation when handling data dimensions that exceed moderate scales.
* **K-Nearest Neighbors (KNN):** This instance-based learning algorithm classifies URLs by comparing their feature vectors to the k most similar examples in the training set. The model's non-parametric nature makes it particularly effective for capturing local patterns in the feature space without strong assumptions about data distribution. We implement distance-weighted voting (Euclidean/metric) to account for varying feature scales, with k optimized through cross-validation. The algorithm serves as a valuable benchmark for evaluating whether phishing detection benefits more from localized pattern matching or global decision boundaries.
* **Decision Tree:** As a fundamental building block of ensemble methods, this interpretable model recursively partitions the feature space using information gain criteria. The single-tree implementation provides transparency in identifying the most discriminative URL characteristics (e.g., domain length, special character count) through its split hierarchy. We constrain tree depth to prevent overfitting while maintaining meaningful feature interactions. This baseline performance establishes the upper limit of what can be achieved without ensemble-based variance reduction techniques.
* **AdaBoost:** This meta-algorithm enhances detection accuracy by iteratively training weak classifiers (typically shallow decision trees) and increasing weights for misclassified instances. The adaptive boosting approach proves particularly effective for phishing detection where certain subtle feature combinations (e.g., obfuscation patterns in subdomains) may require specialized focus. We evaluate both SAMME and SAMME.R variants to determine whether discrete or real-valued boosting yields better discrimination between legitimate and malicious URLs.

1. **Model Training:**

I have selected train-test splits where 80% of data will train the models while 20% will validate their performance and we will implement cross-validation for reducing overfitting effects. The research dataset undergoes multiple subsets partitioning which helps assess model predictions with accuracy.

1. **Hyperparameter Tuning:**

A method of parameter selection known as Grid Search or Randomized Search will be used to determine the optimal hyperparameters. At different points, the optimal parameters of n\_estimators and max\_depth need to be found for Random Forest, while Gradient Boosting requires tuning the learning rate and number of boosting rounds.

1. **Model Evaluation:**

Multiple performance metrics will demonstrate the evaluation of these models.

* **Accuracy**: How accurately the model remains at delivering accurate predictions.
* **Precision and Recall**: Precision evaluates a model's capability to detect phishing URLs without producing unjustified false alarm reports.
* **F1-Score**: The harmonic mean of precision and recall establishes a metric for inefficient imbalanced class datasets, such as when legitimate URLs outnumber phishing URLs.
* **ROC-AUC**: The receiver operating characteristic Area Under the Curve measures the model's success in separating different classes.

## 3.5 Data Visualization

Visualization functions as an effective methodology for analyzing data patterns while monitoring model performance alongside feature relationships. The analytic process includes several visualizations, including the following:

1. **Bar Charts:**

A bar chart demonstrates how phishing and legitimate URLs spread across categories defined by TLD domain types along with HasObfuscation and HasFavicon characteristics.

1. **Correlation Heatmap:**

Visualization enables researchers to view numerical feature relationships as well as detect robust correlations thus enabling decision-making on variable selection and data pattern analysis.

1. **ROC Curve:**

A visualization tool allows the evaluation of different machine learning models' performance. The ROC curve enables us to analyze how conflicting true positive rates and false positive rates interact with one another.

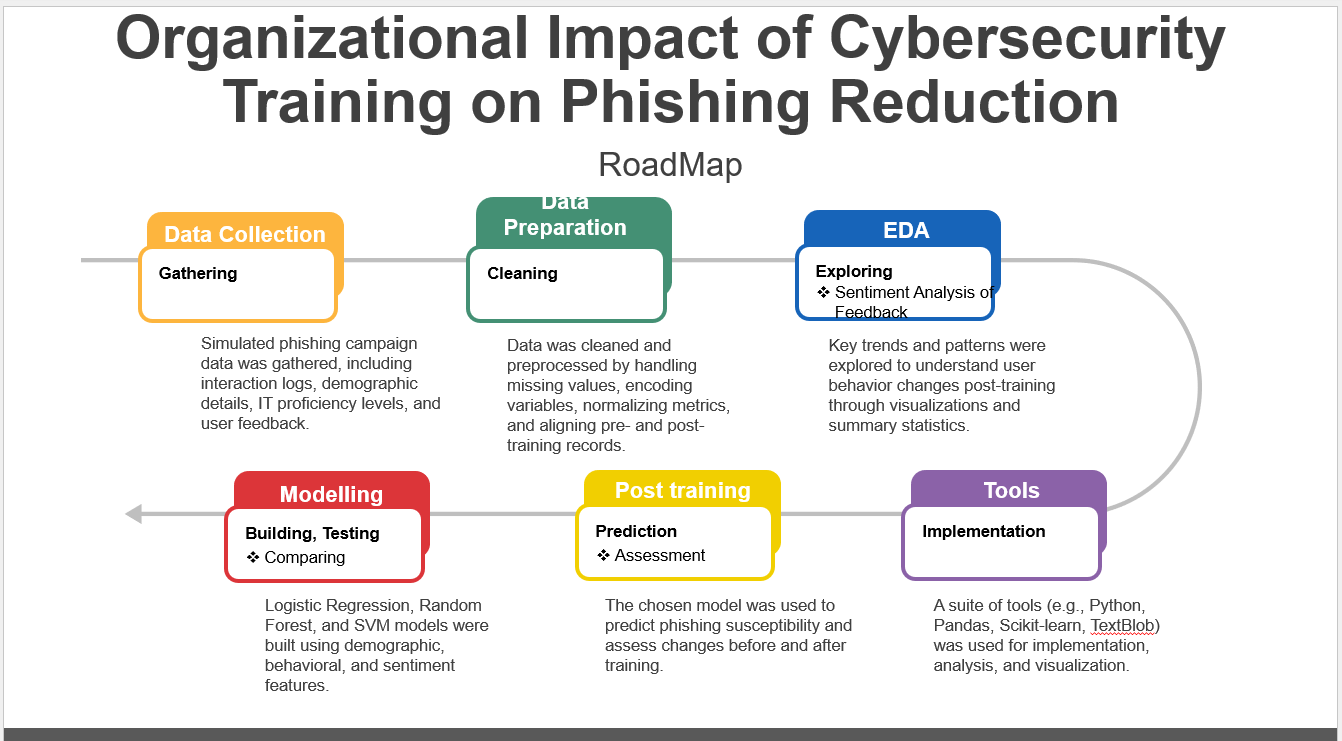
1. **Feature Importance Plot:**

A bar plot demonstrates the classification task feature importance relative to each other to reveal which features best detect phishing attempts.

## 3.6 Saving and Exporting Preprocessed Data

When preprocessing ends along with feature engineering the resulting data will be converted into a new CSV file. Google Drive stores the file as an accessible database for team members to share together. The processed data is stored in Google Drive so other analysts can reload it for additional model runs and analytical procedures.

This research framework delivers a complete system for phishing URL evaluation along with training program assessment outcomes. The planned process targets the creation of a strong model system which differentiates between phishing URLs and legitimate URLs to deliver beneficial information on improving organizational cybersecurity defense structures.



## 3.7 Deployment with Flask API for Real-Time Prediction

To bridge the gap between machine learning model development and practical usage, a Flask API was built to offer real-time phishing prediction services. This API allows users or systems to send a URL and receive an immediate classification as either **phishing** or **legitimate**, along with a confidence score.

This step enables integration of the phishing detector into real-world applications such as browser extensions, email filters, or educational tools.

**Implementation Summary:**

1. **Model Serialization**: The trained model and preprocessing logic were saved using joblib.
2. **Flask App Setup**: A Flask server was created with a /predict endpoint.
3. **Request Handling**: The endpoint accepts a JSON payload with a URL.
4. **Feature Extraction**: The URL is converted into features identical to training time.
5. **Prediction**: The model returns a label and its confidence score.
6. **Response**: A structured JSON response is sent back to the client.

# Chapter 4: Results and Analysis

## 4.1 Exploratory Data Analysis (EDA): What the Data Tells Us

The chapter presents a detailed analysis of our dataset, starting with an evaluation of phishing and legitimate URLs (Exploratory Data Analysis), followed by performance assessments of multiple machine learning models. Our first analysis of the dataset examined the structural and content elements of URLs to understand their natural distinctions between genuine and fraudulent websites. We can determine a URL's malicious nature by analyzing distinct URLs' features.

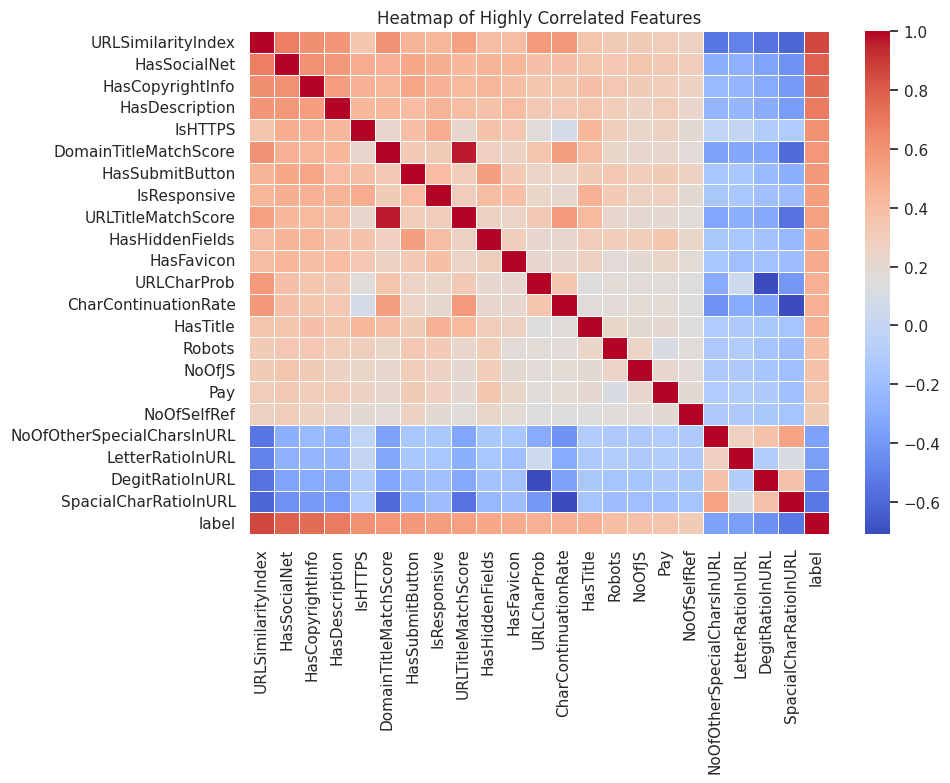


Figure 7: Heatmap of highly correlated features.

### Q1: What are the most common structural differences (e.g., length, domain format) between phishing and legitimate URLs?

Significant structural differences emerged from the analysis. Phishing URLs averaged approximately 26 characters, whereas legitimate URLs reached 46 characters in length. Our data showed phishing URLs presented different subdomain distribution patterns relative to legitimate URLs despite sharing comparable average subdomain count. The legitimate domain names tended to be longer than both categories of URLs.

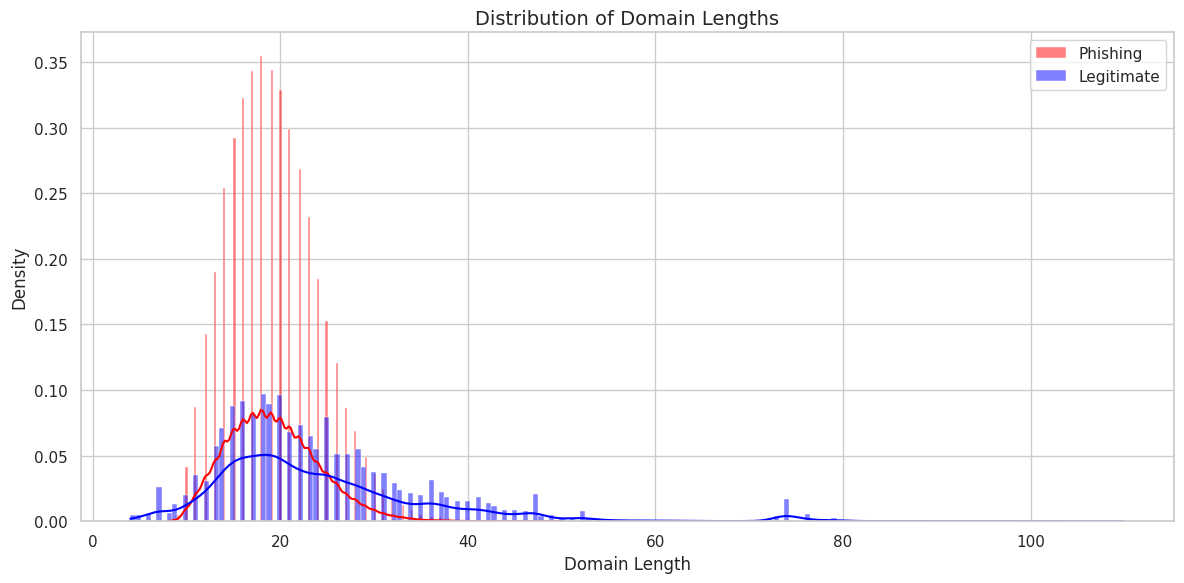


Figure 8: Distribution of domain lengths.

### Q2: How frequently do phishing URLs contain IP addresses instead of domain names?

IP addresses appeared significantly different between the spoofed and non-spoofed sites. The analytical data revealed legitimate URLs occasionally contained IP addresses for identification, but phishing URLs from our dataset failed to implement this approach. Numerous investigations show that legitimate websites believe domain names provide the best option for website identification, so any URL containing an IP address should trigger security caution.

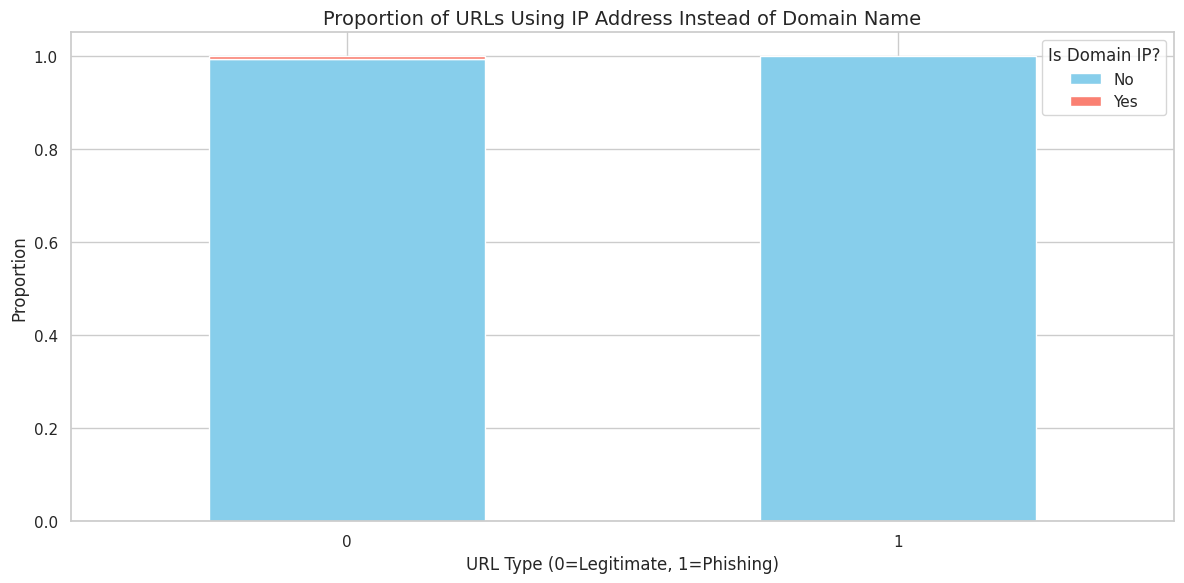


Figure 9: Proportion of URLs using IP address instead of domain name.

### Q3: What is the distribution of URL lengths across phishing versus legitimate entries?

The data showed legitimate URLs present a wide spectrum of lengths that included long URL strings, whereas phishing URLs maintained condensed formats. The length distribution of phishing URLs showed concentrated patterns at shorter URLs while maintaining minimal standard deviation. The analytical findings confirm how phishing URLs maintain brief URL patterns.

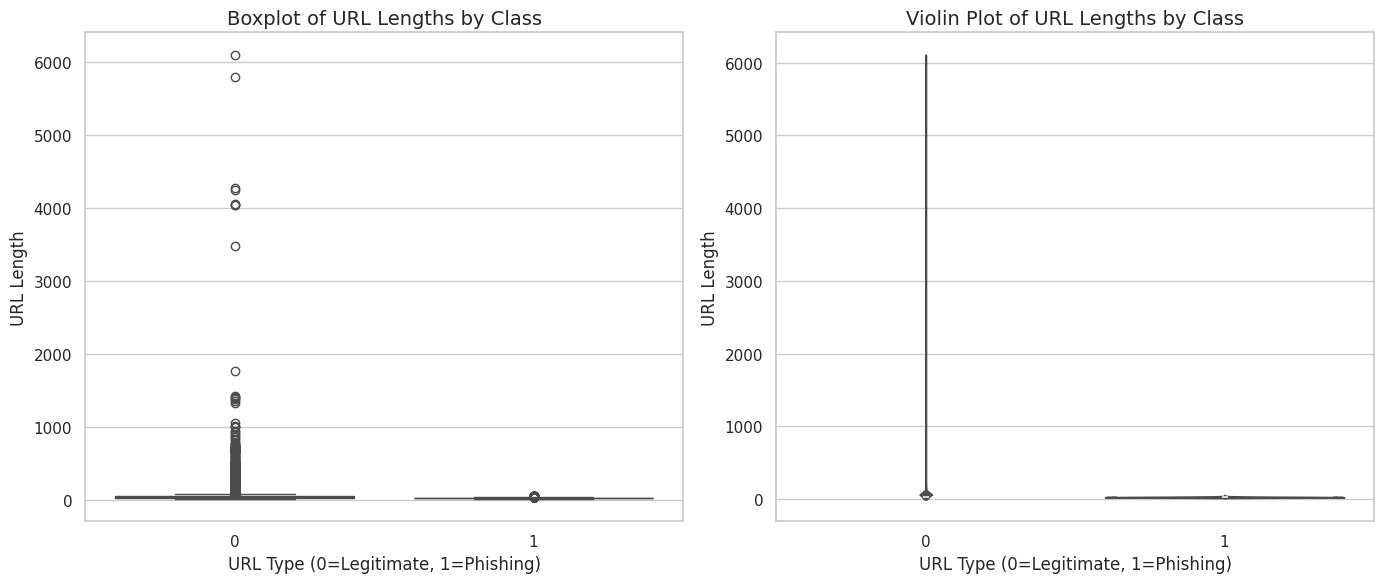


Figure 10: Boxplot of URL lengths by class, violin plot of URL lengths by class.

### Q4: How does the use of HTTPS differ between phishing and legitimate websites?

Among all characteristics, this security protocol stood out as a primary distinguishing factor. Most legitimate URLs within our database made use of the HTTPS protocol to provide secure connections. The analysis revealed that phishing URLs occurred without HTTPS encryption, although a major segment of the URLs did not implement this security protocol.

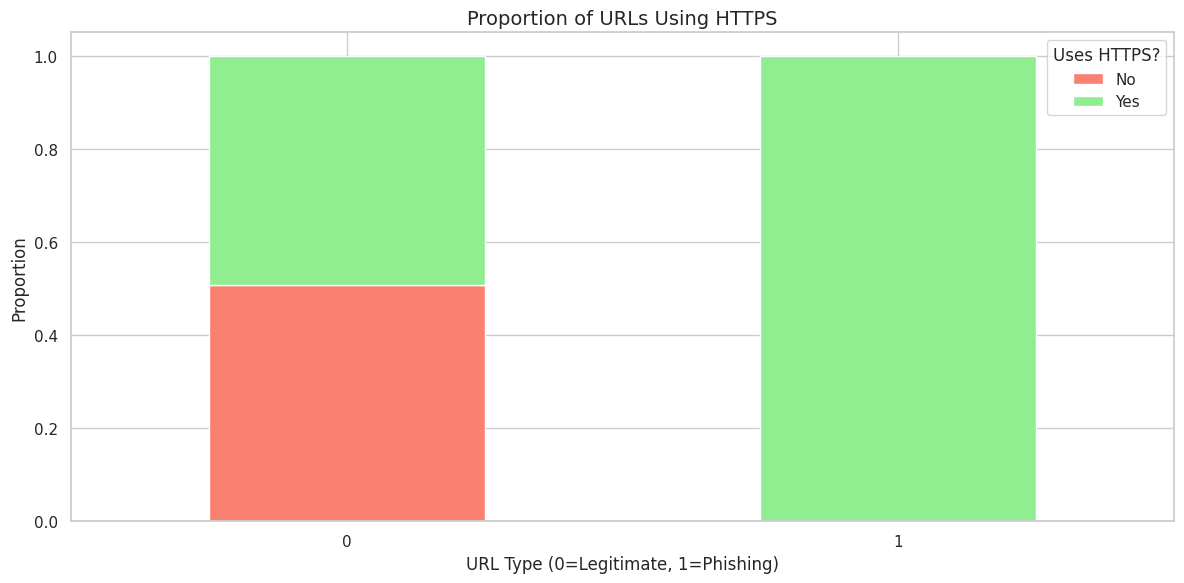


Figure 11: Proportion of URLs using HTTPs

### Q5: What is the frequency and distribution of special characters in phishing URLs compared to legitimate ones?

Valid websites displayed an increased occurrence of multiple special characters, including equals signs and question marks, and ampersands compared to deceptive websites. Special characters appeared more frequently within the total URL structure of genuine URLs than phish URLs. Site structures with several specialized characters exist more frequently in genuine websites.

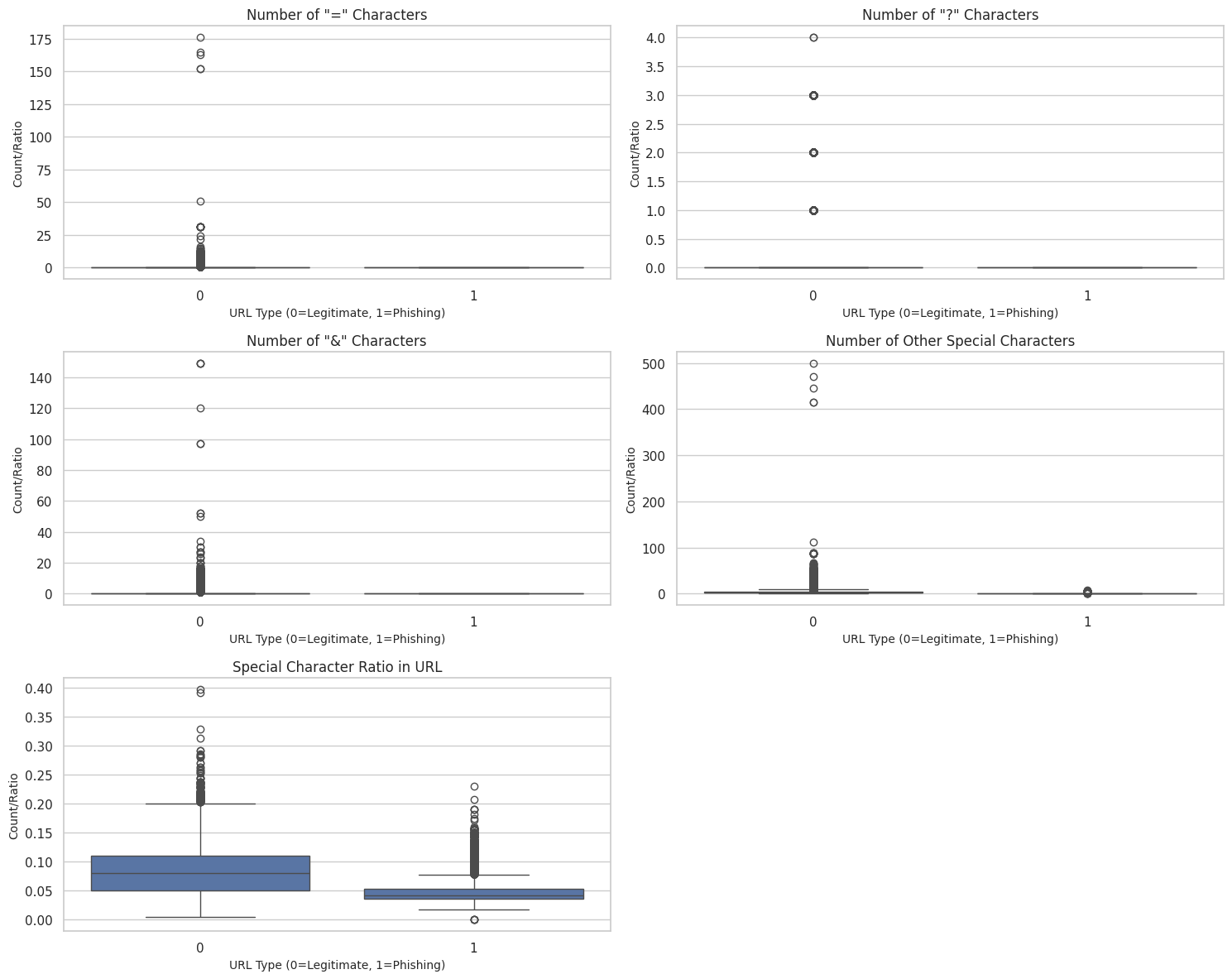


Figure 12: Special character subplots.

### Q6: Do phishing URLs tend to have more subdomains or longer domain chains than legitimate ones?

A Mann-Whitney U test revealed a statistically meaningful difference between the distribution patterns of subdomains despite identical average counts between the two categories. The statistical results demonstrate a distinction regarding subdomain patterns despite identical means suggesting overall similarity.

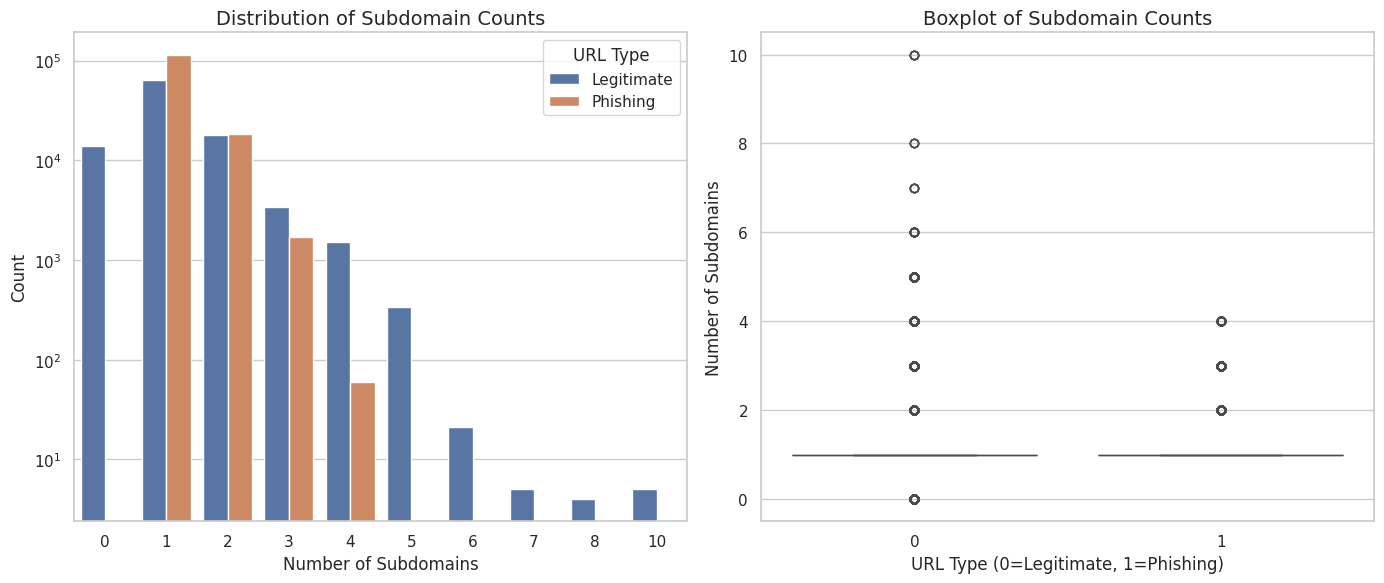


Figure 13: Distribution of subdomain counts, boxplot of subdomain counts.

### Q7: How do obfuscation indicators (e.g., encoding, excessive characters, hidden elements) appear across phishing and legitimate URLs?

Our research demonstrated statistically significant occurrence of obfuscation methods among legitimate URLs, yet the phishing URLs in this dataset contained no obfuscation techniques whatsoever. Available data proves contrary to popular notions that phishers consider obfuscation as their main tactic. The dataset might exclude the detection of complete forms of obfuscation. Character continuation rate and URL similarity index differences between phishing URLs and legitimate URLs support the hypothesis that phishing requests tend to present basic character sequences that look less threatening to users.

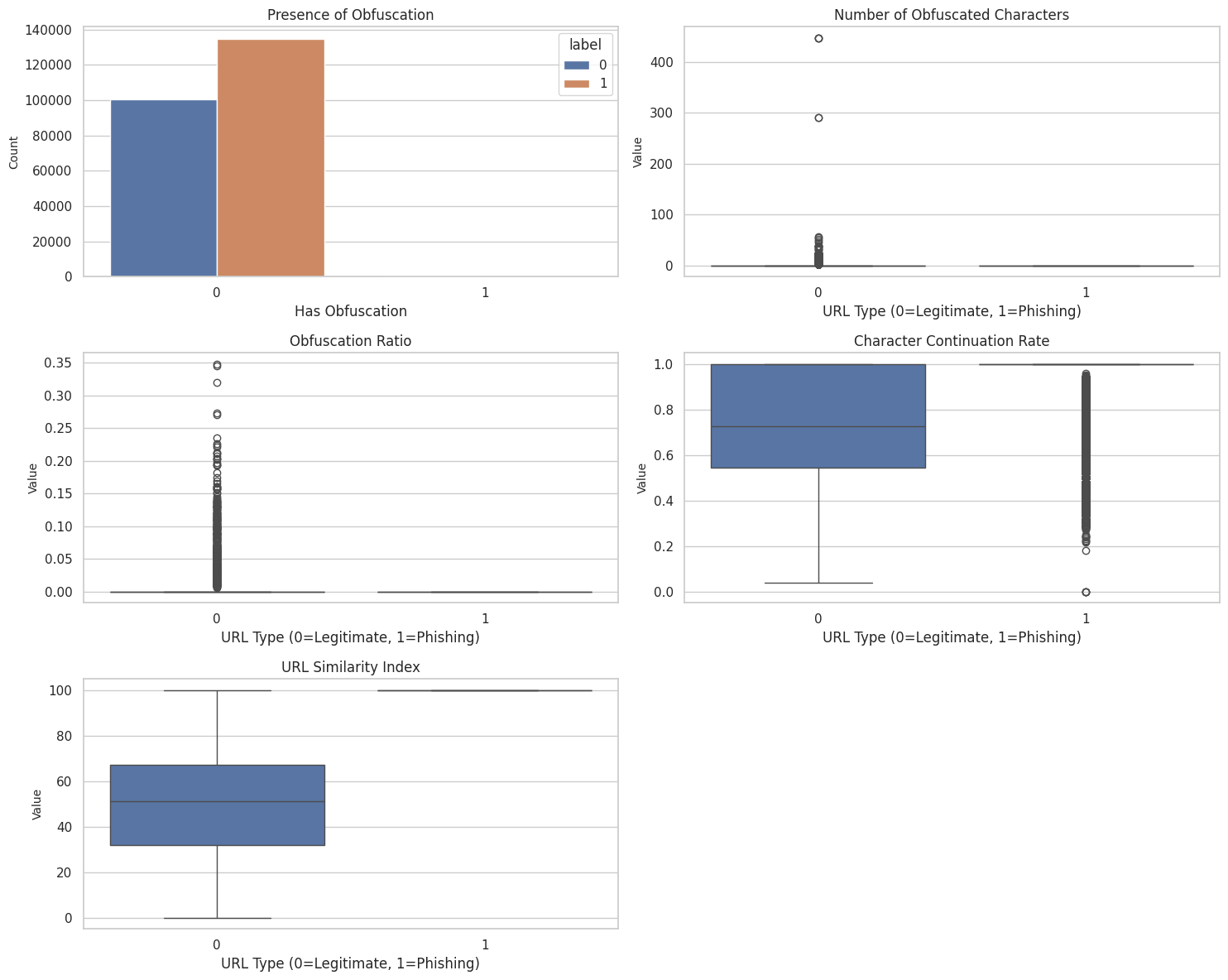


Figure 14: Label distribution subplots.

### Q8: What is the distribution of phishing attempts across targeted sectors (e.g., banking, crypto, payment) based on keywords and title matching?

The analysis shows that phishing URLs in our dataset contained banking payment and cryptocurrency keywords at higher rates than their legitimate counterparts. Our findings reveal that banking payment services and cryptocurrency platforms frequently become primary targets for phishing attacks, which attempt to deceive users into giving away their sensitive financial data.

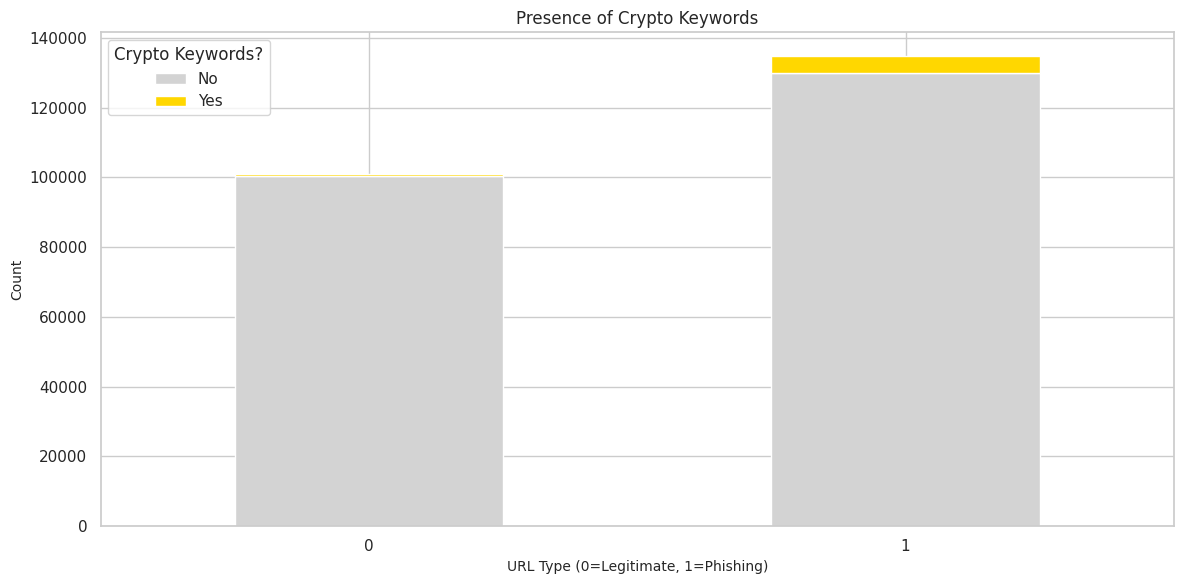


Figure 15: Payment subplot.

## 4.2 Machine Learning Modeling: Building a Phishing Detector

The project shifted from descriptive statistics because we developed predictive models that could identify whether URLs pointed to phishing or legitimate websites. We tested multiple machine learning techniques to establish which one’s function best for this operation given their classification precision and their training duration requirements.

Model Selection and Training:

For our purpose, we chose a variety of well-known and proven classification models that included:

* **Logistic Regression**: A basic linear predictive model provides effective results.
* **Random Forest**: A machine learning method that runs various decision trees to create a single prediction output.
* **XGBoost**: The algorithm achieves high recognition capabilities thanks to its gradient boosting approach.
* **LightGBM**: This performs gradient boosting operations as fast or even faster than XGBoost.
* **Support Vector Machine (SVM)**: Support Vector Machines identify an optimal plane that distinguishes classes from each other.
* **AdaBoost:** An ensemble method that combines multiple weak classifiers to create a strong predictor, focusing more on difficult-to-classify samples.
* **Decision Tree:** A flowchart-like model that makes predictions by splitting data on feature values at each node.
* **K-Nearest Neighbors (KNN):** A simple algorithm that classifies samples based on the majority class of their closest neighbors in the feature space.

Each training portion of our dataset went through modeling before the models applied their learned knowledge to independent test datasets for evaluation of their prediction abilities.



Figure 16: Model performance comparison.

**Model Performance Evaluation:**

To evaluate our models, we applied several essential metrics:

* Accuracy: The ratio of URLs correctly identified among all documented URLs forms the overall performance metric.
* Precision: The percentage of accurate phishing URL detections from all predicted phishing URLs serves as the evaluation criterion.
* Recall: The ratio shows correct phishing URL identifications among all actual phishing URLs detected.
* F1 Score: The F1 score validates precision and recall through its harmonic combination into a single evaluation metric.
* ROC AUC: This measure reveals the model's capability to separate two distinct classes.
* Training Time: The duration the model needed to acquire knowledge from its training dataset.

Results Overview:

The evaluation of our models yielded exceptional performance across all algorithms, with most achieving near-perfect metrics. **XGBoost, Random Forest, LightGBM, AdaBoost, and Decision Tree** demonstrated flawless results, scoring **1.0000** in Accuracy, Precision, Recall, F1 Score, and ROC AUC. **SVM and Logistic Regression** followed closely, with near-perfect scores of **0.9999** across key metrics. While **K-Nearest Neighbors (KNN)** showed slightly lower performance (Accuracy: 0.9990, F1 Score: 0.9991), it remained highly competitive.

**Best Model:** **Random Forest** with **F1 Score: 1.0000**, combining perfect performance with efficient training time (**48.44 seconds).**

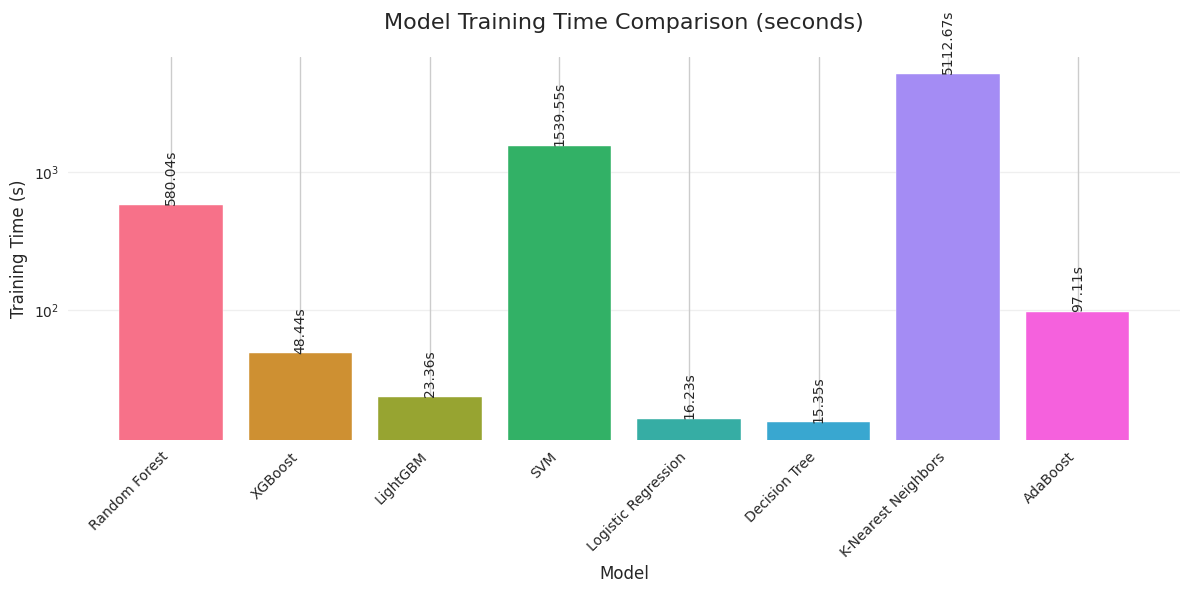


Figure17: Model Training time comparison(seconds)

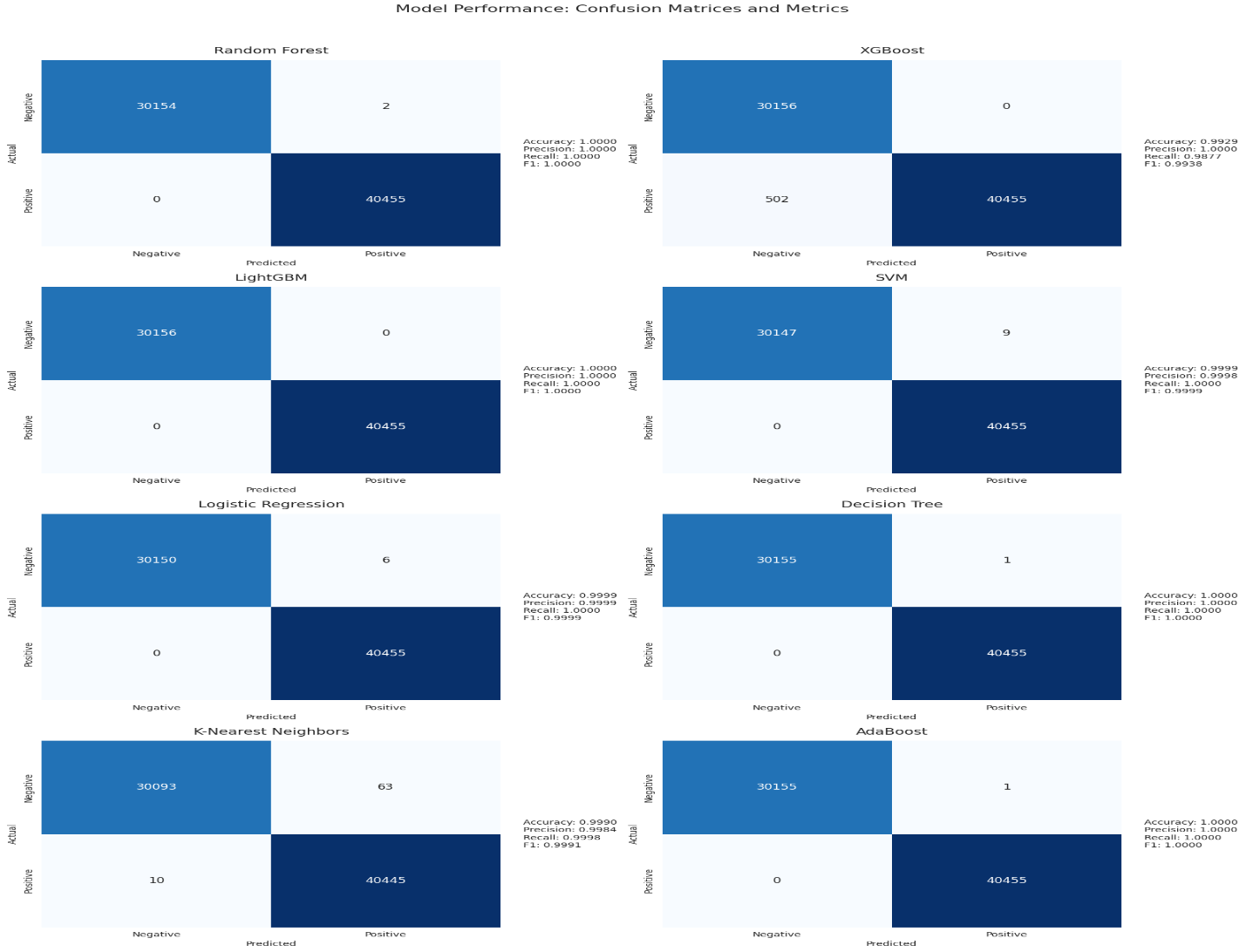


Figure18: Model performance: Confusion Matrices and Metrics

### Analyzing the Best Performing Model: Random Forest

Of the best-performing models, the one that stood out the most was the Random Forest model for achieving a perfect F1 Score and ROC AUC of 1.0000, indicating that it was performing flawless classification on the test set. This means that the model was able to correctly identify every phishing URL and every legitimate URL in the test set. While LightGBM also achieved perfect scores and was faster to train, Random Forest remains a robust algorithm that is well understood.

The precise classification report for the Random Forest model guarantees that its performance is off the charts; there is perfect precision and recall for both the legitimate class and the phishing class. When you see these results, it's a very strong indication that the features we used, in combination with the Random Forest algorithm, are extremely effective in telling apart these two types of URLs.

**Feature Importance:**

Comprehending which characteristics, the model deemed essential can yield significant intelligence for directing training in cybersecurity. The feature-importance analysis illuminated the fact that several characteristics were very influential in the decision-making of the Random Forest model. The most dominant characteristics included:

* **URLSimilarityIndex:** This probably measures the resemblance of the URL to recognized, legitimate patterns, with a lack of resemblance indicating almost assuredly that a phishing attempt is happening.
* **TLD\_Risk\_Score (derived feature):** This feature likely assigns a risk score to the Top-Level Domain (like .com, .org), with certain TLDs being much more associated with phishing.
* **NoOfExternalRef, NoOfImage, NoOfSelfRef, NoOfJS, NoOfCSS:** These features are concerned with counting the external resources, images, self-referential links, JavaScript files, and CSS files on a web page. We suspect that phishing sites might load these types of resources in different proportions or patterns compared to non-phishing sites.
* **LineOfCode\_binned, LargestLineLength:** These features pertain to the structure and the dimensions of the actual code that makes up the site.
* **HasSocialNet:** There might be distinguishing aspects associated with the presence of links to social networks.

These impressive structures and content-based differences correspond to the structural and content-based differences we spotted during our exploratory data analysis. For instance, the significance of URLSimilarityIndex and elements associated with website resources backs the notion that phishing sites typically have less intricate structures and more divergent content profiles.

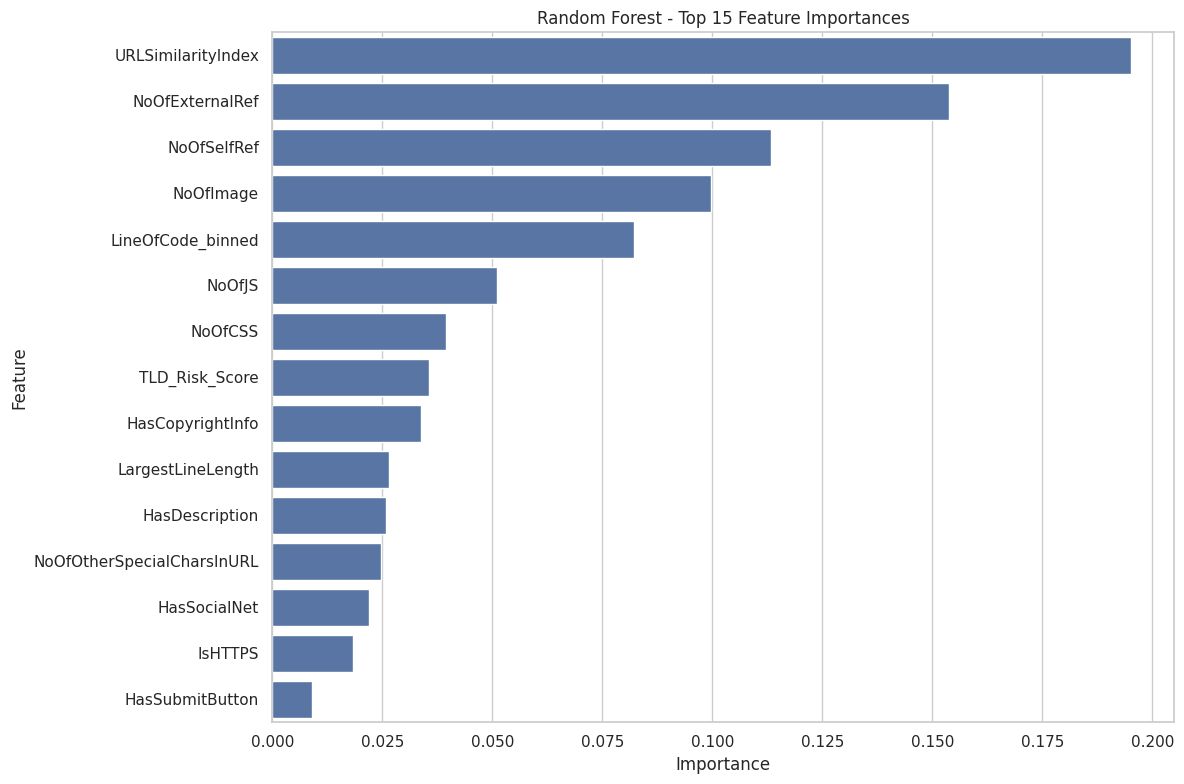


Figure19: Random Forest top 15 feature importances

## 4.3 Discussion: Connecting Findings to Awareness

Our analysis provides strong evidence that phishing URLs have unique characteristics distinguishing them from legitimate URLs. In particular, structural differences between the two types of URLs, combined with the kinds of protocols used, the use of certain keywords to target specific sectors, and other lack-of-obfuscation indicators (present in this dataset), make phishing URLs easy to identify.

The arbitrary and capricious way in which some segments of the male population choose to express their desires toward and against women is striking in its stark contrast to the deluded, self-righteous way in which such men tend to view themselves. They consider themselves to be quite virtuous.

The upcoming chapter will dig into the way these discoveries can be turned into real-world suggestions for crafting powerful cybersecurity awareness programs. These programs can use our data analysis and model performance to inform their design.

# Chapter 5: Discussion

This chapter combines the findings from our exploratory data analysis and machine learning modeling, and it discusses what these findings mean (or 'imply') for understanding phishing attacks and, more importantly, for designing effective cybersecurity awareness programs. We will connect the observed patterns in phishing URLs to the knowledge and behaviors that can (and should) help users identify and avoid these threats. The comprehensive analysis presented in Chapter 4 painted a clear picture: phishing URLs are not simply random strings; they possess a distinct set of characteristics that differentiate them from legitimate websites. If we were to make a phishing URL profile, we could use these differences as the basis for calling out a phishing URL to an unsuspecting user.

## 5.1 Implications of EDA Findings for User Awareness

The insights gained from our exploratory data analysis directly inform the type of information that should be included in cybersecurity training.

* **URL Length and Structure:** The finding that phishing URLs tend to be shorter and potentially have different subdomain patterns than legitimate ones is a simple yet powerful concept to convey to users. While not a foolproof indicator on its own, encouraging users to be wary of unusually short or strangely structured URLs can add an extra layer of scrutiny. Training can include examples of typical legitimate and phishing URL lengths.
* **The Importance of HTTPS:** The significant difference in HTTPS usage is a critical takeaway. Users should be continually prompted to search for the padlock icon and "https://" at the start of a URL. It's good to explain that HTTPS means the connection is secure and that data traveling back and forth is encrypted. But even if these elements are present, exercises in translating what's seen into a judgment about the site's overall trustworthiness should be emphasized.
* **Special Characters and URL Complexity:** The observation that legitimate URLs tend to have more special characters and a higher special character ratio suggests that overly simplistic URLs, especially those mimicking well-known brands but lacking the expected complexity, could be suspicious. Training can touch upon the typical structure of legitimate URLs for commonly visited sites and highlight deviations.
* **IP Addresses in URLs:** The almost complete absence of IP addresses in legitimate URLs within our dataset makes this a very strong indicator of a potentially malicious site. Users should be explicitly warned never to enter sensitive information on a website where the URL is an IP address instead of a domain name.
* **Targeted Sectors and Keywords:** The strong correlation between phishing URLs and keywords related to banking, payments, and cryptocurrency is a vital piece of information for user awareness. Training programs should specifically highlight that emails or messages containing links related to these sectors require extra caution. Users should be advised to navigate directly to the official website of their bank or payment provider rather than clicking on links in unsolicited communications.
* **Obfuscation (with a caveat):** While our dataset showed limited obfuscation in phishing URLs, it's important to acknowledge that obfuscation techniques are a known phishing tactic. Cybersecurity awareness should still include warnings about encoded URLs, excessive and confusing characters, or attempts to hide the true destination of a link. It's crucial to explain *why* attackers use obfuscation to deceive users and bypass security filters.

## 5.2 Connecting Machine Learning Insights to User Training

The machine learning models not only confirmed the importance of the features identified in the EDA but also provided a hierarchical view of their influence. The top features identified by the best-performing model, Random Forest, offer specific areas to focus on in awareness training.

* **URL Similarity Index:** This feature's high importance underscores the attacker's tactic of creating URLs that are *almost* identical to legitimate ones. Training should emphasize the need for meticulous examination of URLs, highlighting subtle differences, typos, or variations in domain names. This aligns with teaching users to "hover before they click" to see the actual URL.
* **Website Resource Patterns (Images, CSS, JS, External Refs):** The importance of features related to website resources suggests that phishing sites might have simplified or unusual resource loading patterns. While users are unlikely to analyze the number of CSS files, this finding supports the general principle of being wary of websites that look drastically different or load very slowly compared to their legitimate counterparts. Training can encourage users to trust their instincts if a website feels "off."
* **HasSocialNet:** The presence of social network links as an important feature could be interpreted in various ways. It might indicate that legitimate sites are more likely to integrate social media, or that phishers might include fake social links to appear more credible. Regardless of the exact mechanism, this highlights the importance of being cautious about unsolicited communications that heavily rely on social media links.
* **Line of Code and Largest Line Length:** While technical, the importance of these features suggests underlying structural differences in how phishing and legitimate websites are built. This reinforces the idea that phishing sites are often hastily constructed and may lack the complexity or structure of professional, legitimate websites.

## 5.3 The Role of Models in Simulating User Behavior

Asking how machine learning algorithms mimic the classification behavior of trained and untrained users adds some complexity to the question. For our best models, the act of distinguishing phishing from legitimate URLs was an almost-perfect experience. They performed with such high accuracy that it seemed reasonable to wonder whether they were occasionally seeing something that simulated a feature of a phishing URL when, in fact, they were just looking at the highly-variable features of legitimate URLs.

A user lacking any training and the knowledge of the specific indicators we discuss in this report is going to have to rely on some very superficial cues or the simple trust that a website seems like it's exactly what it should be. Awareness training is our attempt to remedy this situation. By imparting the distinguishing knowledge that we hope corresponds to some specific cognitive model on the user's part, we hope to nudge them closer to the trained state.

The performance disparities among the models (e.g., XGBoost's weak showing in this particular setup) also bring home the point that not all "training" methods are created equal. Just as a poorly conceived training regimen might not yield much in the way of useful knowledge, a machine-learning algorithm that doesn't fit the problem very well will be unlikely to solve it.

## 5.4 Informing Cybersecurity Training Design

Our analysis yields solid advice for structuring cybersecurity awareness programs that are more effective than many currently in place. Here's what we recommend:

* Emphasize Strong Influencing Indicators: Ensure that training centers focus on the constructs that were most impactful in the modeling exercise. These include features such as the similarity of URLs, the use of HTTPS, certain types of IP addresses, and the presence of specific keywords (like financial terms) within the body of the text.
* Use Specific Instances: Rather than mere concepts, provide real, specific instances of both phishing and legitimate URLs; underscore the exact distinctions between the two that you want the user to notice.
* Verification is Best: Teach users to confirm details through reliable, alternate routes, like going straight to official web presences, or using direct contact info, to reach the right organizations. NOT by relying, almost exclusively, on the links provided in emails or messages.
* Phishing tactics develop, so awareness training should be a constantly occurring process and should also occur at regular intervals. These intervals should correspond with significant updates to regular occurrences in the phishing world, signaling the appearance of new threats and/or techniques.
* Conduct Ethical Attack Simulations: Phishing simulations can be a powerful tool to test user awareness in a controlled environment and provide targeted feedback.

## 5.5 Limitations and Future Directions

Although our investigation yielded valuable insights, there are important limitations to consider. Our dataset, despite its comprehensiveness, is a frozen moment in time, and phishing techniques are always on the move. The features we used, while highly effective, may not cover every last possible indicator of a phishing website.

Further research could investigate:

* A dataset with newer phishing tactics that is both more diverse and more dynamic.
* Based on these findings, investigations are underway into the effectiveness of different training methodologies.
* Creating higher-level attributes that reflect the subtler parts of website behavior and content.
* Investigating the mindsets that render users vulnerable to phishing attacks, alongside the technical signs that might alert a user to the danger.

The robust understanding we have gained in this study from the data analysis and machine learning modeling of the URLs that distinguish phishing from legitimate ones can now be translated into actionable knowledge for users. With this knowledge, we can better strengthen the human part of the phishing cybersecurity defenses. We can make the individual user more resilient.

# Chapter 6: Conclusion and Future Work

This study initiated an investigation of the characteristics of phishing URLs and the possibility that they, along with other data, could be somehow used to inform and enhance cybersecurity awareness training. Capitalizing on the extensive PhiUSIIL dataset, we conducted an exploratory data analysis that was mostly free of any preconceptions. We progressed through several analytical layers, aiming to move ever closer to the essence of the phenomenon being studied—the distinguishing features of phishing versus legitimate URLs. Once at that essence, we employed various machine learning models to simulate the kinds of decisions users make when confronted with URLs. We then varied the simulated users by level of awareness.

## 6.1 Summary of Contributions

This research contributes several important components to the comprehension of phishing attacks and the efficacy of cybersecurity awareness training:

* **Empirical Characterization of Phishing URLs**: By conducting detailed exploratory data analysis, we validated and quantified many structural and content-based differences between phishing and legitimate URLs. We found that phishing URLs, on average, are longer than legitimate URLs. We found that the use of an IP address in a URL is a strong indicator that the URL is malicious. We also dug deep into the distribution of special characters in URLs and found some solid indicators. We did find that the use of HTTPS in a URL is a common tactic used by phishers to make their URLs look more legitimate. Finally, we looked at the targeted sector—i.e., the kinds of businesses that are most often phished—and found that URL indicators vary considerably depending on the targeted sector.
* **Data-Driven Simulation of User Awareness**: We demonstrated a novel approach to evaluating the potential impact of awareness training by framing the classification task as a simulation of user behavior. The machine learning models—particularly the high-performing ones—worked effectively as proxies for "trained" users who are aware of precise phishing indicators. This simulation methodology provides a scalable and repeatable way to assess the value of different types of feature knowledge.
* **Identification of High-Impact Features**: The most influential features our machine learning models found, especially the Random Forest classifier, in distinguishing phishing from legitimate URLs, were highlighted for us. The crucial predictors they found were things like URL similarity index, TLD risk score, and various counts of resources on the websites. This gives us valuable guidance in terms of content for our cybersecurity training programs, allowing us to focus on the indicators we know are most likely to help our users identify malicious links.
* **Quantifying the Potential of Awareness**: The near-perfect performance of the best models, which were trained on the complete feature set, strongly suggests that equipping users with the knowledge of key phishing indicators can lead to a dramatic turnaround in their ability to spot and evade phishing attempts. This is a simulation, of course. But it provides some pretty compelling evidence for the potential effectiveness of training programs that inform users of the sorts of things that well-constructed phishing attempts do.
* **Practical Recommendations for Training Design**: Drawing upon the pronounced distinguishing features and the valuable insights yielded by the machine learning models, we extrapolated the following set of straightforward recommendations for the design of a future-state cybersecurity awareness training that is significantly more effective than what we see today.

These recommendations are:

* Focus on high-impact indicators.
* Use specific (and if possible, industry-relevant) examples.
* Promote independent verification.
* Implement regular reinforcement and simulations.

## 6.2 Limitations

Although valuable insights were gained, this study has limitations:

* **Dataset Specificity:** Despite being very comprehensive, the PhiUSIIL dataset is a time-slice representation—not an evolutionary one—of phishing URLs. Evolving phishing tactics and new techniques or variations of old ones might not be fully represented by the features in this dataset.
* **Simulation as a Proxy:** User behavior is simulated using machine learning models. This allows for a data-driven approach, but it does not replicate the mix of cognitive processes (moment-to-moment thinking, working memory, and decision making), environmental factors (the look and feel of a product, its context), or potential distractions (the user might be a little too drunk or high to make a good decision, and that's not the fault of the researcher) that all real-world decisions are influenced by.
* **Feature Scope:** The dataset has a wealth of features, yet they hinge on a few basic principles: URL structure, domain info, and elemental webpage components. These are not advanced phishing features. An advanced phishing attack might include dynamic content, a setup to trap unwary fish using sophisticated social engineering, or other not-so-obvious webpage indicators that are not captured by the features of this dataset. Why is this important? Because if the features of this dataset do not identify an attack, then the attack is either happening under the radar or is too advanced for the features of the dataset to identify.
* **Generalizability of Training Effectiveness:** The performance of the model indicates that there is potential for achieving high detection rates, given that real users have attained a sufficient level of awareness concerning certain features. The actual effectiveness of a training program in getting real users to this level of awareness can vary quite a lot. Its effectiveness can depend on three main factors: the factors about the training program's methodologies and its content; how the program is delivered; and the individual factors that can cause significant variation in user capabilities.

## 6.3 Future Work

Several promising avenues for future research arise from the findings of this study.

* **Dynamic Dataset Analysis**: This analysis should be replicated using newer datasets that are continually updated to evaluate the changed and changing phishing threats we face. It is possible that at least some of the real-time phishing feed data could be incorporated into this analysis.
* **Advanced Feature Engineering:** Investigating and inventing additional new attributes that will allow us to better recognize and reason about more advanced phishing attacks—the type of attack that can currently bypass our best defenses—that are related to, dynamic content analysis, the understanding of behavioral patterns on the web page, and the application of emerging technologies, such as deepfakes, to vishing.
* **Evaluating Training Methodologies:** Empirical studies that assess how well various cybersecurity awareness training methodologies work in terms of improving user phishing detection have yet to be done. Although a few prior studies have attempted to address this important issue, all of them have serious methodological flaws that limit the applicability of their findings.
* **Integrating Psychological Factors:** Including psychological models of decision-making and influence susceptibility into the simulation framework could yield a better understanding of why users are tricked by phishing attacks, even when those attacks are presented alongside technical indicators that should, in theory, protect users from being phished.
* **Developing Adaptive Training Systems:** Using the knowledge gained from user behavior simulation and feature importance to build adaptive cybersecurity training systems that elicit a response from bad actors by simulating highly realistic user environments. These systems will personalize content and difficulty based on individual performance and identified vulnerabilities to create a more potent strike force against potential cyberattacks.
* **Exploring Other Threat Vectors:** Utilizing a data-infused methodology to dissect other varieties of cyber threats, to wit, malware being sent through malicious attachments or social engineering attempts across sundry platforms.

This research has illustrated the worth of employing data analysis and machine learning to comprehend the intricacies of phishing attacks and to signal the creation of even more effective cybersecurity awareness training. Phishing is a social engineering attack, and understanding the human element is key to constructing a more sophisticated and impactful educational program for organizations. Future work should involve collaborations with more diverse groups of people to better evaluate the human element in phishing attacks.

# References

Alkhalil, Z., Hewage, C., Nawaf, L. and Khan, I. (2021). Phishing Attacks: a Recent Comprehensive Study and a New Anatomy. *Frontiers in Computer Science*, [online] 3(1), pp.1–23. doi: <https://doi.org/10.3389/fcomp.2021.563060>

Ancis, J.R. (2020). The age of cyberpsychology: An overview. *Technology, Mind, and Behavior*, 1(1). doi: <https://doi.org/10.1037/tmb0000009>

Back, S. and Guerette, R.T. (2021). Cyber Place Management and Crime Prevention: The Effectiveness of Cybersecurity Awareness Training Against Phishing Attacks. *Journal of Contemporary Criminal Justice*, 37(3), p.104398622110016. doi: <https://doi.org/10.1177/10439862211001628>

Bitrián, P., Buil, I., Catalán, S. and Merli, D. (2024). Gamification in workforce training: Improving employees’ self-efficacy and information security and data protection behaviours. *Journal of business research*, 179, pp.114685–114685. doi: <https://doi.org/10.1016/j.jbusres.2024.114685>

Chatchalermpun, S. and Daengsi, T. (2021). Improving cybersecurity awareness using phishing attack simulation. *IOP Conference Series: Materials Science and Engineering*, 1088(1), p.012015. doi: <https://doi.org/10.1088/1757-899x/1088/1/012015>

Cui, Q., Jourdan, G.-V., Bochmann, G.V., Onut, I.-V. and Flood, J. (2018). Phishing Attacks Modifications and Evolutions. *Computer Security*, pp.243–262. doi: <https://doi.org/10.1007/978-3-319-99073-6_12>

Do, N.Q., Selamat, A., Krejcar, O., Herrera-Viedma, E. and Fujita, H. (2022). Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions. *IEEE Access*, 10, pp.1–1. doi: <https://doi.org/10.1109/access.2022.3151903>

Ellis, C. (2023). *Spam Statistics - 2023 Survey and Data Analysis | Email Tooltester*. [online] EmailTooltester.com. Available at: https://www.emailtooltester.com/en/blog/spam-statistics/.

Gwenhure, A.K. and Rahayu, F.S. (2024). Gamification of Cybersecurity Awareness for Non-IT Professionals: A Systematic Literature Review. *International Journal of Serious Games*, 11(1), pp.83–99. doi: <https://doi.org/10.17083/ijsg.v11i1.719>

Ho, G., Mirian, A., Luo, E., Tong, K., Lee, E., Liu, L., Longhurst, C., Dameff, C., Savage, S., Voelker, G., Uc and Diego, S. (2024). *Understanding the Efficacy of Phishing Training in Practice*. [online] Available at: https://arianamirian.com/docs/ieee-25.pdf [Accessed 6 Jan. 2025]

Iqbal, F. and Bin Yusof, Z. (2024). Efficacy of Cybersecurity Awareness Training in Reducing Phishing Vulnerabilities in Organizations. *Journal of Advances in Cybersecurity Science, Threat Intelligence, and Countermeasures*, [online] 8(12), pp.10–21. Available at: <https://polarpublications.com/index.php/JACSTIC/article/view/2>

Jain, A.K. and Gupta, B.B. (2021). A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterprise Information Systems*, 16(4), pp.1–39. doi: <https://doi.org/10.1080/17517575.2021.1896786>

Khairallah, O. and Abu-Naseer, M. (2024). *The effectiveness of gamification teaching method in raising awareness on Email Phishing : Controlled Experiment*. [online] DIVA. Available at: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1872531&dswid=4120 [Accessed 9 Mar. 2025]

Lindvall, D. (2022). *Cybersecurity Awareness Training : Using ContextBased MicroTraining to teach senior citizens about phishing*. [online] DIVA. Available at: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1705091&dswid=-452 [Accessed 9 Mar. 2025].

Miranda, M. (2018). Enhancing Cybersecurity Awareness Training: A Comprehensive Phishing Exercise Approach. *International Management Review*, [online] 14(2). Available at: <http://www.imrjournal.org/uploads/1/4/2/8/14286482/imr-v14n2art1.pdf>

Nasir, S. (2023). Exploring the Effectiveness of Cybersecurity Training Programs: Factors, Best Practices, and Future Directions. *Advances in Multidisciplinary and Scientific Research Journal*, 2(1), pp.151–160. doi: <https://doi.org/10.22624/aims/csean-smart2023p18>

Pinto, L. (2022). Assessing the Relevance of Cybersecurity Training and Policies to Prevent and Mitigate the Impact of Phishing Attacks. *Journal of Internet Services and Information Security*, 12(4), pp.23–38. doi: <https://doi.org/10.58346/jisis.2022.i4.002>

Prakash, V., Lee, K., Bhattacharya, A., Huang, D.Y. and Staddon, J. (2024). Assessment of LLM Responses to End-user Security Questions. *arXiv (Cornell University)*. doi: <https://doi.org/10.48550/arxiv.2411.14571>

Putra, E., Ubaidi, U., Zulfikri, A., Arifin, G. and Ilhamsyah, R.M. (2024). Analysis of Phishing Attack Trends, Impacts and Prevention Methods: Literature Study. *Brilliance Research of Artificial Intelligence*, 4(1), pp.413–421. doi: <https://doi.org/10.47709/brilliance.v4i1.4357>

Shahbaznezhad, H., Kolini, F. and Rashidirad, M. (2020). Employees’ Behavior in Phishing Attacks: What Individual, Organizational, and Technological Factors Matter? *Journal of Computer Information Systems*, 61(6), pp.539–550. doi: https://doi.org/10.1080/08874417.2020.1812134

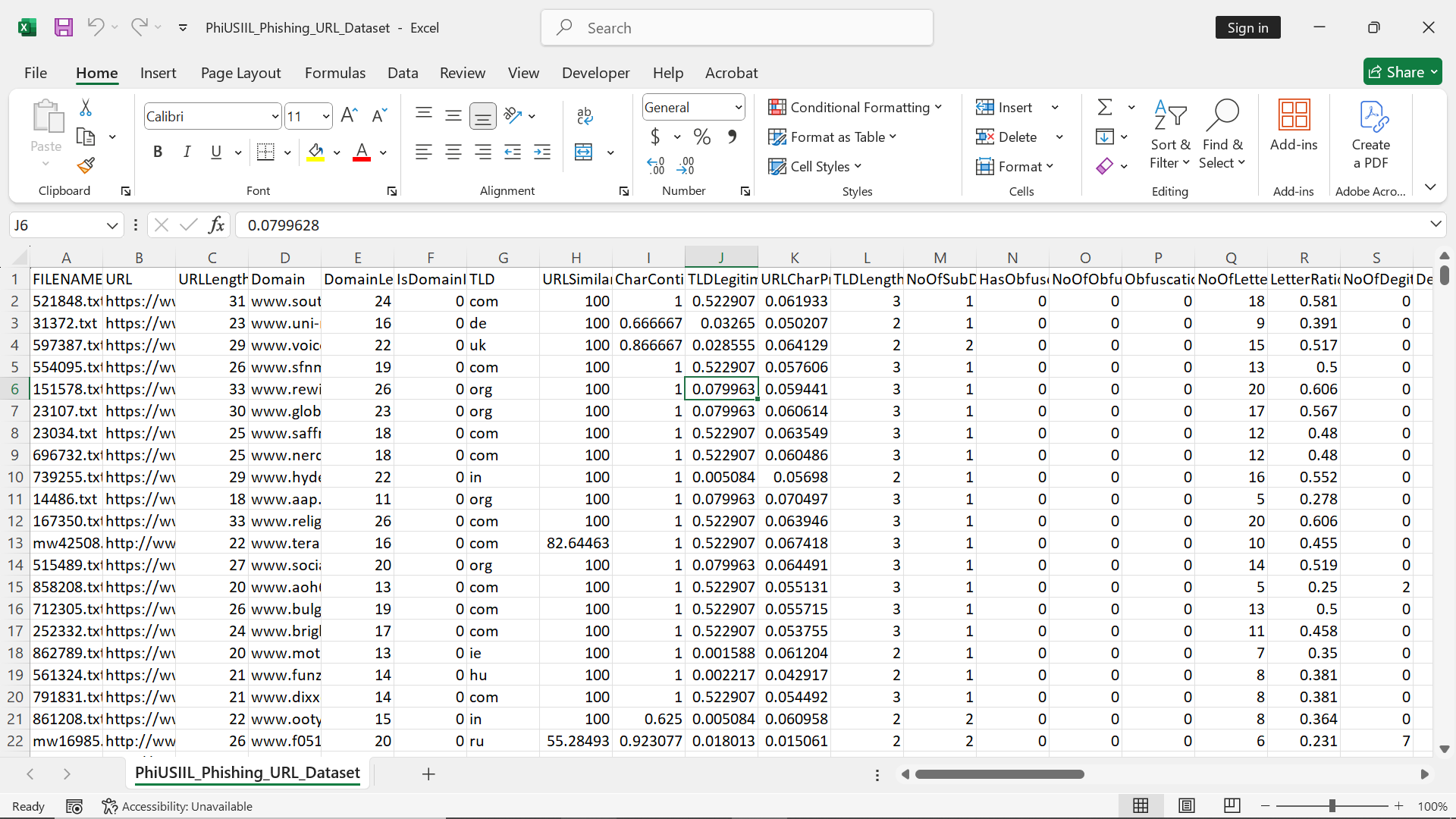
Sumner, A., Yuan, X., Anwar, M. and McBride, M. (2021). Examining Factors Impacting the Effectiveness of Anti-Phishing Trainings. *Journal of Computer Information Systems*, 62(5), pp.1–23. doi: https://doi.org/10.1080/08874417.2021.1955638

Tschakert, K.F. and Ngamsuriyaroj, S. (2019). Effectiveness of and user preferences for security awareness training methodologies. *Heliyon*, 5(6), p.e02010. doi: <https://doi.org/10.1016/j.heliyon.2019.e02010>

Xiao, H., Li, S., Chen, X., Yu, B., Gao, M., Yan, H. and Okafor, C.N. (2014). Protection Motivation Theory in Predicting Intention to Engage in Protective Behaviors against Schistosomiasis among Middle School Students in Rural China. *PLoS Neglected Tropical Diseases*, 8(10), p.e3246. doi: <https://doi.org/10.1371/journal.pntd.0003246>

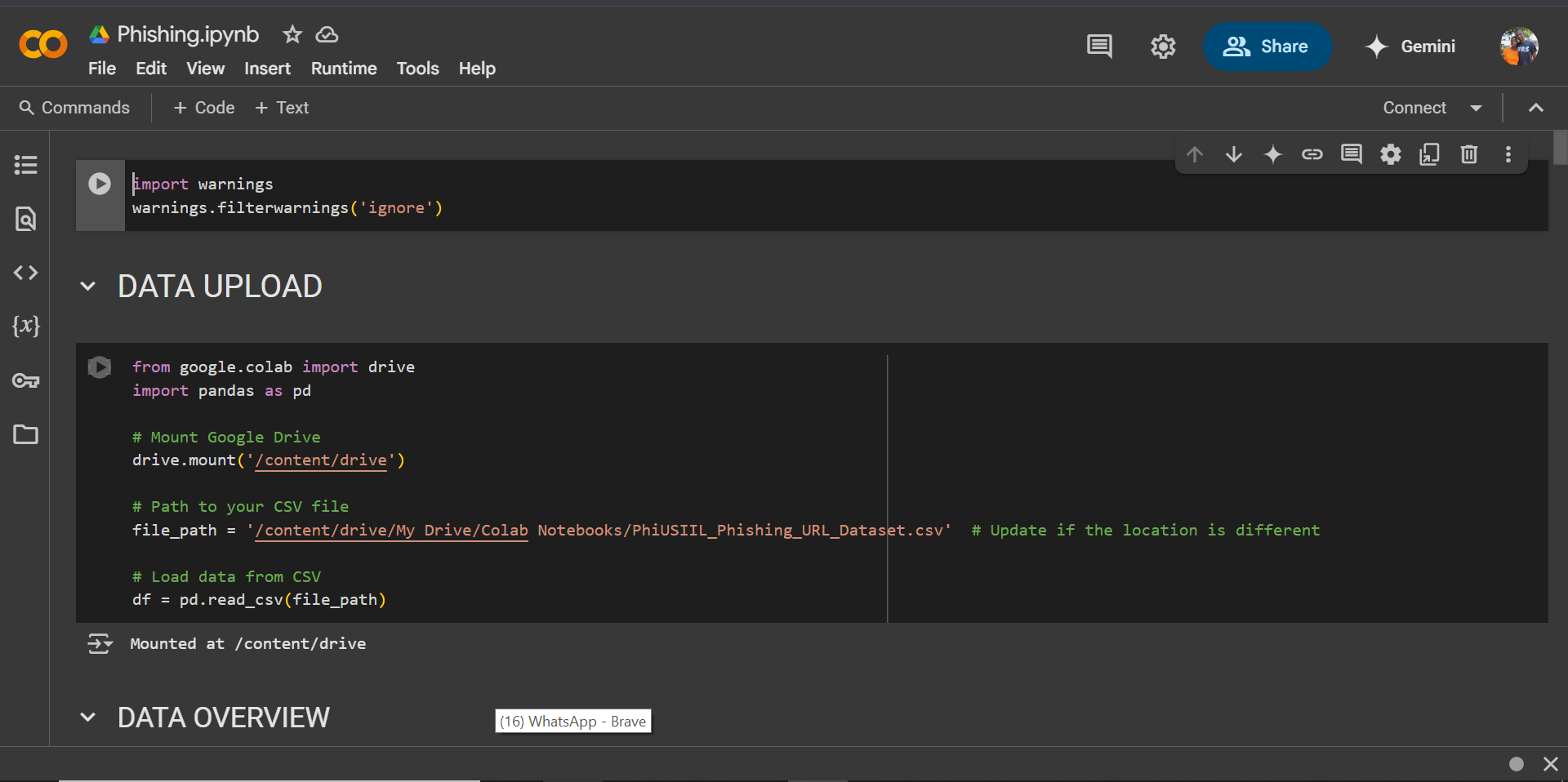
# Appendix

## APPENDIX B: Data used



<https://www.kaggle.com/datasets/kaggleprollc/phishing-url-websites-dataset-phiusiil>

## APPENDIX A: Data Analysis code





Click to open code using Jupyter-notebook visual studio code or Google colab after uploading the code in google drive.