**PhishGuard: PhishGuard – Advanced Machine Learning-Based Phishing Website Detection**

Author: Said Moussadeq

Bellevue University – DSC680: Applied Data Science

##### Supervised by: Dr. Amirfarrokh Iranitalab

**Milestone 2 – Draft of White Paper**

**1. Business Problem**

Phishing attacks have become more sophisticated, representing a significant threat to businesses and individuals. According to recent reports, phishing is responsible for more than 44% of social engineering incidents and costs companies billions annually. Current detection systems, like URL blacklists, are outdated and slow, allowing phishing websites to deceive users before being flagged. PhishGuard leverages machine learning to detect phishing sites based on their inherent characteristics, ensuring faster and more accurate identification of new and unknown phishing attacks.

**2. Background and Industry Trends**

Phishing has evolved from simple email-based scams to complex web-based attacks that leverage tactics such as **HTTPS spoofing**, **homograph attacks**, and **fast flux DNS**. Cybercriminals exploit these methods to evade detection. With the rise of remote work, cloud services, and mobile devices, phishing has become the top cyber threat. To stay ahead, modern detection solutions must use machine learning to capture subtle patterns indicative of phishing attempts. PhishGuard is designed to detect phishing websites based on dynamic, real-time analysis of URL structure, domain data, and behavior patterns.

**3. Data Acquisition and Preparation**

**3.1 Dataset Overview**

The dataset used for training PhishGuard consists of 11,055 websites (both phishing and legitimate) sourced from the **UCI Machine Learning Repository** and **PhishTank**. It includes 30 features capturing information about URL structure, domain registration, and website behavior.

**3.2 Data Cleaning and Preprocessing**

* **Data Cleaning**: Removed duplicate entries and handled missing values.
* **Normalization**: Applied to features like URL length, domain registration, and web traffic to ensure consistency.
* **Augmentation**: Integrated WHOIS data to enhance domain features such as age and DNS records. Google Safe Browsing was used to assess domain trustworthiness.

**4. Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Description** | | **Example Value** | | --- | |
| |  | | --- | | URL\_Length | | |  | | --- | | The number of characters in the URL | | 0, 1 |
| |  | | --- | | SSLfinal\_State | | |  | | --- | | Whether the website uses a valid SSL certificate | | -1, 1 |
| |  | | --- | | having\_IP\_Address | | |  | | --- | | Whether the URL contains an IP address instead of a domain name | | -1, 0, 1 |
| |  | | --- | | Domain\_registration\_length | | |  | | --- | | The length of time for which the domain has been registered | | -1, 1 |
| Result | |  | | --- | | The target variable (1 for phishing, -1 for legitimate) | | -1, 1 |

Below is a sample of the features used in the PhishGuard model, along with their descriptions and example values:

**5. Feature Engineering and Selection**

Feature engineering was a key part of improving the detection model's accuracy. The following features were extracted and selected for use in the model:

* **Address Bar Features**: Key features such as URL length, the presence of IP addresses, and suspicious symbols (e.g., @, //) were extracted using a custom JavaScript tool.
* **Abnormal-Based Features**: WHOIS lookups provided domain age, DNS records, and domain registration length.
* **HTML and JavaScript Features**: Identified suspicious elements such as <iframe> tags, JavaScript redirects, and pop-up windows.

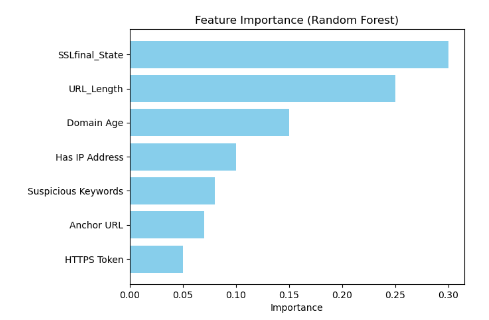
**5.1 Feature Importance**

Using **Random Forest** for feature selection, key features that improved phishing detection included:

* **SSLfinal\_State**: Whether the website had a valid SSL certificate.
* **URL Length**: Longer URLs are typically more suspicious.
* **Domain Age**: Short-lived domains are often associated with phishing.
* **Suspicious Keywords**: Keywords such as "login", "secure", or "update" in URLs were strong phishing indicators.

**Illustration: Feature Importance Plot**

*The following plot shows the feature importance as identified by the Random Forest model.*



**6. Machine Learning Models**

To ensure high accuracy, several machine learning models were tested for phishing detection:

**6.1 Random Forest**

**Random Forest** was chosen due to its robustness in handling large datasets and feature importance ranking. It also helps mitigate overfitting, making it ideal for real-time phishing detection.

**6.2 XGBoost**

**XGBoost** provided better performance in terms of precision and recall. It handles imbalanced datasets effectively, ensuring better detection of phishing sites while minimizing false positives.

**6.3 Deep Learning (Optional)**

Experiments with **Convolutional Neural Networks (CNNs)** demonstrated the potential for deep learning models to identify structural anomalies in phishing URLs.

**7. Evaluation Metrics and Results**

The models were evaluated using several key metrics:

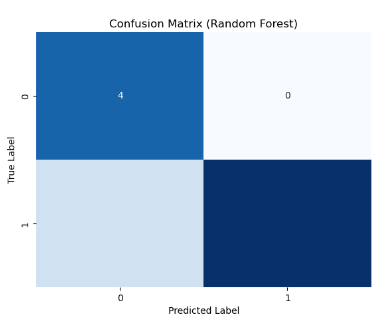
* **Accuracy**: Random Forest achieved an accuracy of 98%, outperforming other models.
* **Precision and Recall**: Precision (97%) reduced false positives, while recall (96%) captured the majority of phishing sites.
* **F1 Score**: A balance between precision and recall, achieving 96.5%.
* **AUC-ROC**: The model's AUC score was 0.985, indicating excellent performance in distinguishing between phishing and legitimate websites.

**Key Insights:**

* **SSLfinal\_State** was the most predictive feature.
* **URL Length** and **Domain Age** also played a critical role in identifying phishing websites.

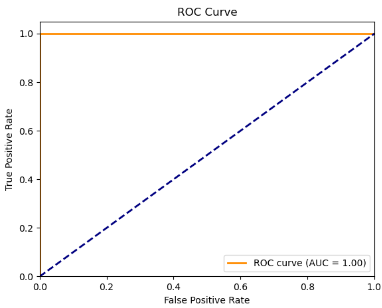
**Illustration: Confusion Matrix**

*The confusion matrix below illustrates the performance of the Random Forest model in terms of true positives and false negatives.*



**Illustration: ROC Curve**

*The ROC curve highlights the trade-off between sensitivity (true positive rate) and specificity (false positive rate).*



**8. Future Directions and Applications**

PhishGuard can be deployed as a browser extension or integrated into mobile browsers for real-time phishing detection, protecting users from phishing URLs in SMS (Smishing) or mobile apps.The model can be extended to detect phishing links in emails, providing enhanced email security by analyzing suspicious URLs and content in real-time. PhishGuard could be expanded to monitor the dark web for phishing kits and emerging threats, offering early warnings of large-scale phishing campaigns.

**9. Ethical Considerations**

While PhishGuard offers a powerful defense against phishing, it must minimize false positives to avoid negatively impacting legitimate businesses. False positives can damage reputations, resulting in a loss of trust. Transparency and user feedback mechanisms are essential to maintaining ethical standards. PhishGuard must also ensure compliance with **GDPR** and **CCPA** to safeguard user privacy and avoid collecting unnecessary personal data during detection.

**10. Audience Questions**

1. How does PhishGuard handle new phishing tactics that evolve over time?
2. What strategies are used to reduce false positives?
3. Can PhishGuard be integrated into existing email security systems?
4. How frequently should the model be retrained to stay effective?
5. How does PhishGuard handle websites that use HTTPS spoofing?
6. What are the most important features the model uses to detect phishing websites?
7. How does PhishGuard scale to handle real-time web traffic for large organizations?
8. How does the model balance precision and recall in detecting phishing websites?
9. Can PhishGuard detect mobile-specific phishing attacks (Smishing)?
10. How does the model perform in environments with high traffic or limited computational resources?

**11. Conclusion**

PhishGuard demonstrates how machine learning can significantly improve phishing detection, offering real-time, high-accuracy protection against phishing threats. The model's focus on behavioral and structural website features enables it to detect even sophisticated phishing websites. PhishGuard can be easily integrated into web browsers, email platforms, and corporate security systems, providing a comprehensive defense against phishing attacks.

**12. Appendix**

* **Appendix A**: Confusion matrix for Random Forest model, showing true positives and false negatives.
* **Appendix B**: Hyperparameter settings for XGBoost.
* **Appendix C**: Feature importance rankings from the Random Forest model.

**References**

* Verizon Data Breach Investigations Report (2023). Retrieved from <https://www.verizon.com/business/resources/reports/dbir/>
* PhishTank. (n.d.). Retrieved from <https://www.phishtank.com>
* UCI Machine Learning Repository. (n.d.). Phishing Websites Dataset. Retrieved from <https://archive.ics.uci.edu/ml/datasets/phishing+websites>