Synopsis of

PHISHING WEBSITE DETECTION

A Summer Project to be submitted by

ISHITA DIXIT (02504092022)
MANSI JOSHI (03804092022)
DEVIKA (07604092022)

for

Summer internship Machine learning and its Application in Cyber security

under the supervision of

Mr Santanoo Pattnaik

Indira Gandhi Delhi Technical University for Women
(Established by Govt. of Delhi vide Act 09 of 2012)
Kashmere Gate, Delhi - 110006

DEPARTMENT OF INFORMATION TECHNOLOGY

Title

Machine Learning-based Phishing Website Detection: Building a Robust Model to Differentiate Legitimate Websites from Phishing Websites

Introduction

In today's digital age, where online transactions and interactions have become an integral part of our lives, the threat of phishing websites has increased significantly. Phishing websites aim to deceive unsuspecting users by mimicking legitimate websites, often with the intention of stealing sensitive information such as usernames, passwords, and financial details. As a result, there is a growing need for robust models that can accurately differentiate between legitimate websites and their malicious counterparts.

The development of an effective model to identify phishing websites requires a comprehensive understanding of the characteristics and techniques employed by these malicious entities. Phishing websites often employ various tactics such as domain spoofing, deceptive content, and social engineering techniques to trick users into revealing their confidential information. These websites may appear identical to legitimate ones, making it difficult for users to discern the difference without additional assistance.

This project aims to explore the key factors involved in building a robust model that can accurately differentiate between legitimate and phishing websites. We will delve into the various approaches and methodologies employed in phishing detection, ranging from rule-based heuristics to machine learning techniques. Additionally, we will examine the importance of feature engineering, data collection, and model evaluation to enhance the performance and accuracy of the phishing detection model.

One of the primary challenges in building a robust model is the ever-evolving nature of phishing techniques. Malicious actors constantly adapt their methods to bypass existing security measures. Therefore, the model must be designed to handle new and previously unseen phishing attacks effectively. This necessitates the integration of real-time data feeds, continuous model updates, and active monitoring to stay one step ahead of the attackers.

Furthermore, the development of a robust model involves ethical considerations. It is essential to balance accuracy with minimizing false positives and false negatives to avoid inconveniencing legitimate website owners and users while maintaining a high level of security. The model should be designed with transparency and fairness in mind, ensuring it does not discriminate against any specific group or inadvertently label legitimate websites as phishing entities.

Motivation

There are several motivations for building a robust model to differentiate legitimate websites from phishing websites:

1. Protection against cyber threats: Phishing websites are a common method used by cybercriminals to deceive users and steal sensitive information such as login credentials, financial details, or personal data. By building a reliable model to identify phishing websites, we can help protect individuals and organizations from falling victim to these fraudulent activities.

2. Mitigating financial losses: Phishing attacks can lead to significant financial losses for individuals and businesses. By accurately detecting phishing websites, we can reduce the risk of users unknowingly providing their financial information to malicious actors, thus minimizing potential monetary damages.

3. Safeguarding personal information: Phishing attacks often target personal information, including social security numbers, addresses, and credit card details. By developing an effective model, we can contribute to the protection of individuals' privacy and prevent the misuse of their sensitive data.

4. Enhancing cybersecurity awareness: Building a robust model to differentiate legitimate websites from phishing websites can also raise awareness among users about the existence and techniques employed by cybercriminals. By educating individuals about the risks associated with phishing attacks, we can empower them to make informed decisions while browsing the internet and reduce the overall susceptibility to such attacks.

5. Strengthening cybersecurity measures: The development of a reliable model requires analyzing various features and indicators associated with phishing websites. This process can help identify common patterns, tactics, and techniques employed by cybercriminals, ultimately aiding in the improvement of cybersecurity measures. The insights gained from this project can be used to enhance existing security protocols and develop more effective defense mechanisms against phishing attacks.

Overall, building a robust model to differentiate legitimate websites from phishing websites is crucial for protecting individuals and organizations from cyber threats, minimizing financial losses, safeguarding personal information, promoting cybersecurity awareness, and strengthening overall security measures in the digital landscape.

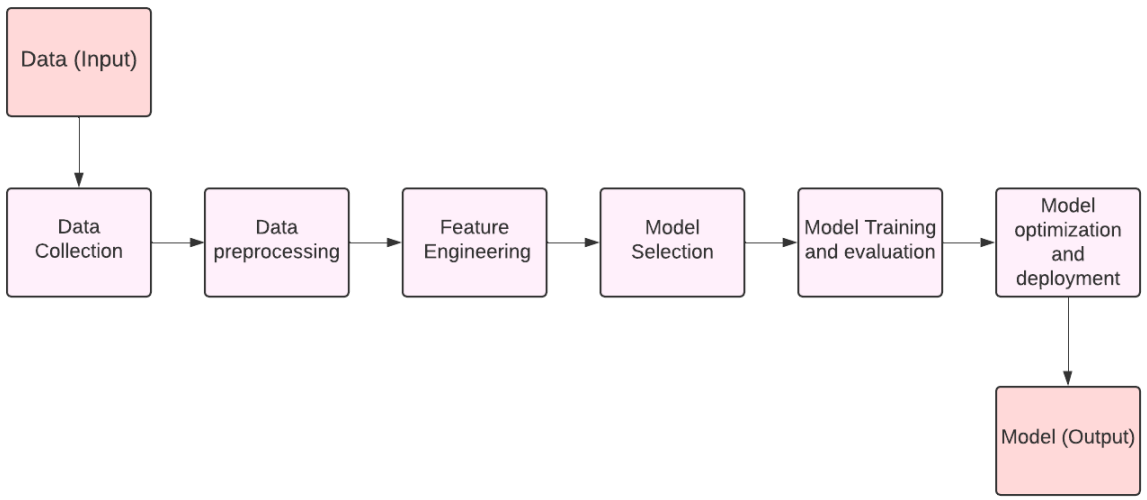
Related papers

s. no	Title of paper	Models used	Datasets	Accuracy
1	Phishing Website Detection using Machine Learning Algorithms	Decision Tree, Random Forest, Support Vector Machine	URLs of benign and phishing websites were collected from www.alexa.com and www.phishtank.com, respectively. The dataset comprises a total of 36,711 URLs, with 17,058 labeled as benign (0) and 19,653 labeled as phishing (1).	Decision Tree: 96.8 Random Forest: 96.84 SVM: 96.4
2	Detection of Phishing Websites using Machine Learning	Random Forest, Decision Tree	collected unstructured data of URLs from Phishtank website, Kaggle website and Alexa website, etc.	Random Forest: 97.29 Decision Tree: 95.9

Work problem statement

Build a precise and reliable machine learning-based model to distinguish between trustworthy and phishing websites. A reliable classification should be provided by the model after it has thoroughly examined website properties such as URL structure, domain reputation, content, and graphic components. To ensure excellent detection performance and generalizability, address issues including imbalanced datasets, feature selection, and optimisation. The goal is to strengthen cybersecurity defenses and shield people and businesses from falling for scams.

Methodology



Data Collection:

- Obtain the Phishing Dataset for Machine Learning from Kaggle https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning .
- Familiarize with the dataset description, attributes, and any available documentation to understand the data's structure and context.

Data Preprocessing:

- Analyze the distribution of classes (legitimate vs. phishing websites) to understand the class imbalance, if any.
- Handle missing values by imputation or removing instances with missing information, ensuring the integrity of the dataset.
- Normalize or scale relevant features to ensure consistency and prevent bias towards certain attributes.

Feature Engineering

- Conduct feature selection techniques, such as correlation analysis, information gain, or chi-square tests, to identify the most informative attributes.
- Choose a subset of features that provide significant discriminatory power between legitimate and phishing websites.

Model Selection

- Select appropriate machine learning algorithms suitable for the phishing website detection task.
- Consider algorithms like decision trees, random forests, logistic regression, support vector machines (SVM), or gradient boosting methods.

Model Training and Evaluation

- Split the preprocessed dataset into training and testing sets, maintaining the class distribution.
- Train the selected models using the training set while fine-tuning hyperparameters through techniques like cross-validation or grid search.
- Evaluate the models' performance using evaluation metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).
- Conduct additional analyses like confusion matrices to understand the models' behavior regarding false positives and false negatives.

Model Optimization & Deployment

- Perform model optimization techniques to improve the model's performance and generalization capabilities.
- Experiment with hyperparameter tuning, regularization methods, ensemble techniques (e.g., bagging or boosting), or sampling strategies (e.g., oversampling or undersampling) to enhance the model's accuracy and robustness.
- Implement the trained model into a functional application or system capable of detecting phishing websites in real-time.

Conclusion

The threat posed by phishing websites necessitates the development of robust models that can accurately differentiate between legitimate and malicious entities. By employing advanced techniques, staying up-to-date with emerging phishing strategies, and considering ethical implications, we can enhance the security of online interactions and protect users from falling victim to phishing attacks.

References

- https://www.sciencedirect.com/book/9780128029275/a-machine-learning-approach-to-phishing-detection-and-defense#book-description
- https://www.phishtank.com/
- https://www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms