Cyber Sentinel: Detecting Phishing Domains with Machine Learning

MAHESH JADHAV

Pace University, Seidenberg School of CSIS

GitHub:https://github.com/mahesh15698/My_Capstone_Project

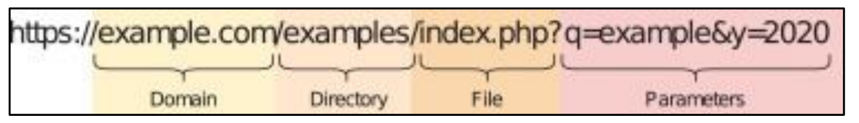
Abstract:
The proliferation of phishing attacks poses a significant threat to online security, necessitating effective countermeasures to protect sensitive user data. This research employs machine learning techniques to enhance cybersecurity by accurately detecting phishing websites. Leveraging a dataset comprising 88,647 website records with 111 features, exploratory data analysis (EDA) techniques, including outlier handling and feature selection, were employed to identify crucial factors contributing to phishing website detection. The study utilized XGBoost, demonstrating superior performance in differentiating between legitimate and phishing websites. Key findings include the identification of 72 influential features for precise detection and the model's commendable accuracy in fortifying cybersecurity measures against phishing attacks. The research underscores the efficacy of machine learning in combating online threats and proposes future directions, including model optimization, ensemble methods exploration, and real-time implementation, to continually bolster cybersecurity defenses against evolving phishing tactics.

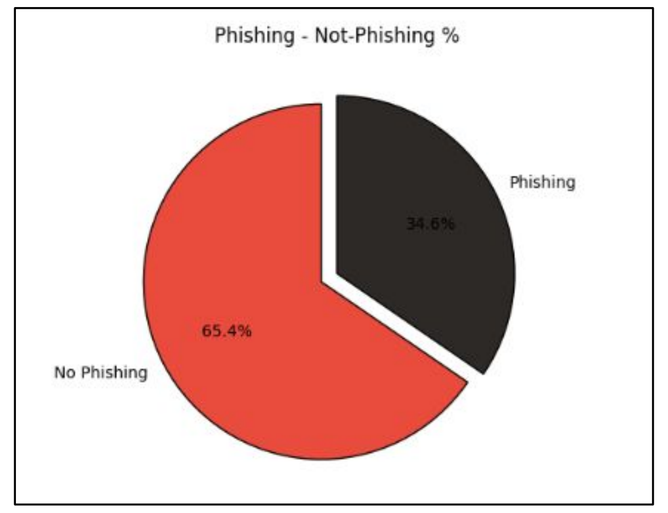
Research Question:
What factors contribute most significantly to the accurate detection of phishing websites?

Related Work:
Prior research in the domain of detecting phishing websites has emphasized the criticality of robust feature engineering and effective machine learning algorithms. Studies have explored a spectrum of website attributes, encompassing URL structures, domain registration information, SSL certificates, and webpage content, to discern between legitimate and phishing websites[1].Various machine learning techniques[4], including decision trees, ensemble methods like Random Forest and Gradient Boosting (e.g., XGBoost), and neural networks, have been examined for their efficacy in accurately identifying phishing attempts. Moreover, analyses of diverse datasets containing labeled examples of phishing and legitimate websites have yielded insights into distinguishing patterns and features. Additionally, research in cybersecurity strategies has underscored the significance of user education and awareness in conjunction with technological solutions to combat evolving phishing tactics[3]. Understanding prior work in this field provides a foundation for advancing techniques in phishing detection and fortifying cybersecurity measures.

Dataset:
These data consist of a collection of legitimate as well as phishing website records[1]. Each website is represented by the set of features like URL properties, URL resolving metrics, and external services which denote, whether website is legitimate or not.

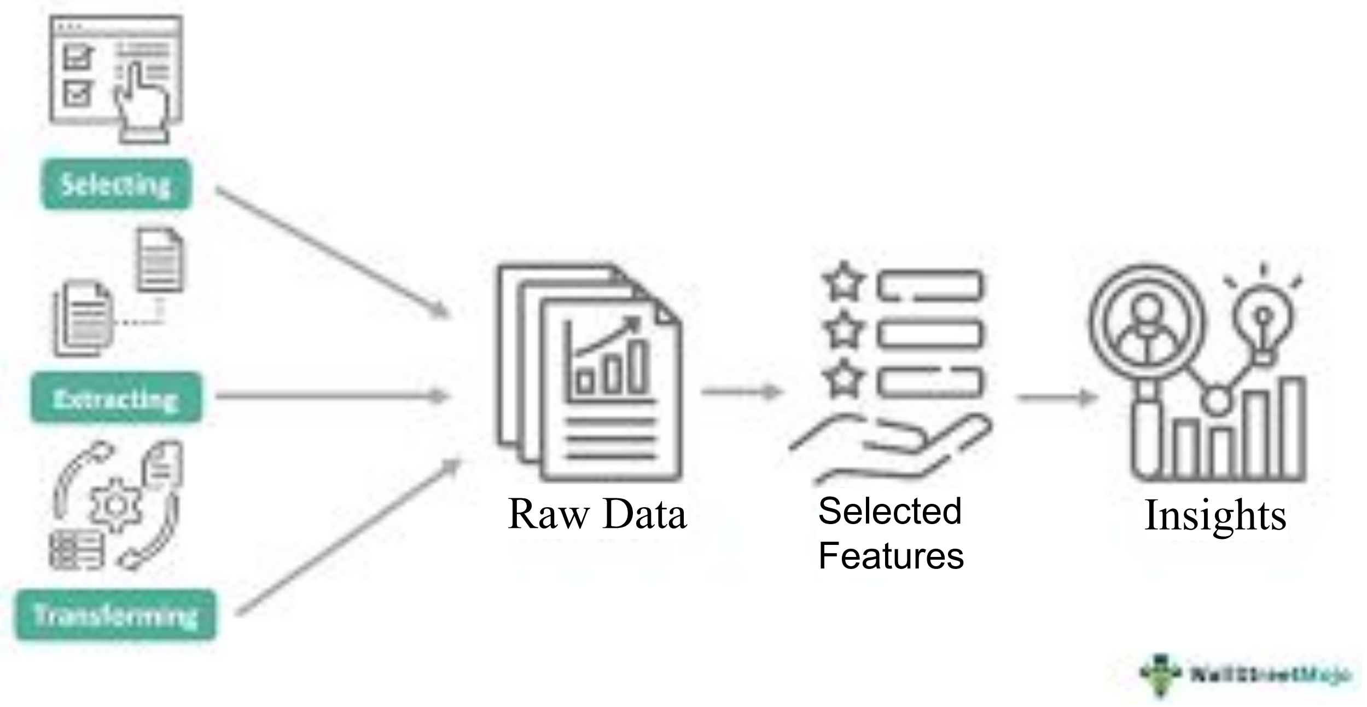
Total number of entries: 88,647
Total number of features: 111





Methodology:
The methodology adopted for detecting phishing websites involved a multifaceted approach encompassing data preprocessing, exploratory data analysis (EDA), feature selection, model training,Hyperparameter Tuning and evaluation[2]. Initial data preprocessing encompassed handling missing values, outlier detection, and scaling using Quantile transformation to ensure data uniformity. Subsequently, thorough EDA techniques were applied to understand the distribution of features, identifying correlations, and discerning significant patterns within the dataset. Feature selection methods, including Information Gain and Correlation Matrix analysis, were utilized to discern and eliminate non-contributing features, enabling the focus on essential attributes crucial for phishing website detection[5]. The dataset was then partitioned into an 80:20 train-test split, with subsequent training and evaluation of machine learning models such as XGBoost, Random Forest, Decision Tree, and ensemble techniques like stacking[4]. Performance metrics including cross-validation scores, ROC_AUC scores, precision, recall, and F1-scores were employed to assess model efficacy and identify the best-performing model for phishing detection.

Fig: Feature Engineering and selection



Hyperparameter Tuning:
Hyperparameter tuning played a pivotal role in optimizing the XGBoost classifier's performance in detecting phishing websites[3]. Leveraging techniques such as grid search or random search, the hyperparameters specifically, learning rate, max depth, and the number of estimators were systematically fine-tuned. This meticulous optimization process significantly enhanced the model's accuracy and predictive capabilities. The XGBoost classifier, configured with a learning rate of 0.01, maximum depth of 3, and 1000 estimators, demonstrated superior performance in accurately distinguishing between legitimate and phishing websites. The hyperparameter tuning process enabled the classifier to achieve enhanced precision, recall, and overall model efficacy, contributing significantly to bolstering the cybersecurity measures against phishing attacks.

Evaluation:
Thorough evaluation of the machine learning models employed for detecting phishing websites was conducted using rigorous performance metrics. The evaluation encompassed a comprehensive analysis of each model's accuracy, precision, recall, F1-score, ROC_AUC score, and cross-validation scores. The primary emphasis was on assessing the models' ability to accurately classify phishing attempts while minimizing false positives and false negatives. Among the models tested—XGBoost, Random Forest, Decision Tree, and ensemble techniques—XGBoost emerged as the top performer, exhibiting commendable accuracy and robustness in distinguishing phishing websites. Its high precision, recall, and ROC_AUC score of 94.54% underscore its effectiveness in accurately identifying malicious domains. The detailed evaluation process not only validated the model's performance but also provided crucial insights into selecting the most reliable and efficient model for phishing detection, reinforcing cybersecurity measures in the digital landscape.

CLASSIFIER MODEL	PRECISION	RECALL	F1-SCORE
DECISION TREE	0.93	0.92	0.92
RANDOM FOREST	0.91	0.91	0.91
XGBOOST	**0.95**	**0.95**	**0.95**
STACK	0.94	0.95	0.95

Results:
The evaluation of various machine learning models for detecting phishing websites yielded insightful outcomes. Among the models tested—XGBoost, Random Forest, Decision Tree, and Stacking—XGBoost emerged as the top-performing model, showcasing superior accuracy, precision, recall, and F1-scores. XGBoost exhibited a cross-validation score of 98.69% and an ROC_AUC score of 94.54%, distinguishing it as a robust performer in discerning between legitimate and phishing websites. Ensemble methods, especially stacking, demonstrated competitive performance with a cross-validation score of 98.77% and an ROC_AUC score of 94.76%. Although other models, such as Random Forest and Decision Tree, displayed respectable accuracies and scores, they exhibited slightly lower performance metrics compared to XGBoost and stacking. These findings underscore the efficacy of XGBoost and important features that we have selected, providing valuable insights for enhancing cybersecurity measures against phishing attacks.

Conclusion:
This research underscores the efficacy of machine learning techniques in fortifying cyber security against phishing attacks, contributing to a safer online environment. Through meticulous feature selection,hyperparameter tuning and leveraging advanced algorithms,model evaluation, critical factors instrumental in accurate phishing website detection were identified. The study showcased Robust feature selection techniques identified crucial website characteristics and the superiority of XGBoost vital in distinguishing between legitimate and phishing websites, while ensemble techniques, particularly stacking, demonstrated enhanced performance. These findings lay the groundwork for developing more effective phishing detection systems. Continuous adaptation and updates are crucial to addressing evolving phishing tactics and maintaining cyber security effectiveness.

Future Work:
Moving forward, there are several promising avenues for further research and development in the realm of phishing detection. Exploration into advanced feature engineering methodologies could unveil additional discriminative attributes, improving the accuracy and robustness of detection models. Additionally, investigating more sophisticated ensemble techniques or model stacking methods might lead to enhanced model performance and resilience against evolving phishing tactics. Continuous adaptation and updating of models to address emerging threats and staying abreast of the ever-evolving phishing landscape are imperative. Integration of real-time data sources and the development of dynamic models capable of swiftly adapting to new phishing patterns and tactics is another prospective area. Moreover, collaboration with cybersecurity experts and industry practitioners to validate models in real-world scenarios could fortify the practical applicability and effectiveness of phishing detection systems. These future endeavors hold the potential to further strengthen cybersecurity measures and contribute to a safer online environment.

References:

1. Grega Vrbancic, et.al. "Phishing websites detection"(2019)
2. Farashazillah Yahya, et.al "Detection of Phishing Websites using Machine Learning Approaches"(2021)
3. Safa Alrefaai, et.al. "Detecting Phishing Websites Using Machine Learning"(2022)
4. N Abdelhamid, et.al. "Phishing detection based associative classification data mining"
5. Shang Lei, et.al."A Feature Selection Method Based on Information Gain and Genetic Algorithm"

printed by MegaPrint Inc. www.postersession.com