Phishing Sites Prediction Using Machine Learning

Executive Summary:

The project "Phishing Sites Prediction Using Machine Learning" aims to enhance web security by leveraging advanced machine learning techniques to predict and classify phishing websites. Phishing remains a prevalent cybersecurity threat, making early detection crucial for user safety. Through the development and implementation of machine learning models, this project achieves a high level of accuracy in distinguishing between legitimate and phishing URLs. The integration of web scraping, link analysis, and visualization further enhances the understanding of patterns within the data. The results showcase the project's effectiveness in contributing to the identification and mitigation of phishing risks, ultimately bolstering online security.

Background:

In the rapidly evolving landscape of cybersecurity, phishing attacks pose a significant threat to individuals and organizations. Phishing websites, disguised as legitimate entities, aim to deceive users into revealing sensitive information, leading to financial losses and security breaches. Traditional methods of identifying phishing sites often fall short in keeping pace with the dynamic tactics employed by cybercriminals.

The project addresses this challenge by harnessing the power of machine learning to predict and classify phishing URLs. Machine learning models, including Logistic Regression and Naive Bayes, are trained on labeled datasets to learn patterns indicative of phishing behavior. These models are then capable of accurately categorizing unseen URLs as either legitimate or potential threats.

Web scraping is employed to collect a diverse set of URLs for training and testing the models. Selenium and BeautifulSoup facilitate the automated extraction of hyperlinks from specified websites, enabling the creation of robust datasets. The link analysis component explores relationships and patterns within the collected data, contributing to a holistic understanding of the URL landscape.

The project's significance lies in its proactive approach to cybersecurity, providing a tool for early detection and mitigation of phishing threats. The combination of machine learning, web scraping, and link analysis offers a comprehensive solution for identifying and combating phishing attacks, ultimately contributing to a safer online environment.

Related Work:

Sudhanshu et al. [1] used association data mining approach. They have proposed rule-based classification technique for phishing website detection. They have concluded that association classification algorithm is better than any other algorithms because of their simple rule transformation. They achieved 92.67% accuracy by extracting 16 features but this is not up to mark so proposed algorithm can be enhanced for efficient detection rate.

Luong et al. [2] proposed new technique to detect phishing website. In proposed method, Author used six heuristics that are primary domain, sub domain, path domain, page rank, and alexa rank, alexa reputation whose weight and values are evaluated. This approach gives 97 % accuracy but still improvement can be done by enhancing more heuristics.

M. Amaad et al. [3] presented a hybrid model for classification of phishing website. In this paper, proposed model carried out in two phases. In phase 1, they individually perform classification techniques, and select the best three models based on high accuracy and other performance criteria. While in phase 2, they further combined each individual model with best three model and makes hybrid model that gives better accuracy than individual model. They achieved 97.75% accuracy on testing dataset. There is limitation of this model that it requires more time to build hybrid model.

Gupta et al. [4] proposed a novel anti phishing approach that extracts features from client-side only. Proposed approach is fast and reliable as it is not dependent on third party, but it extracts features only from URL and source code. In this paper, they have achieved 99.09% of overall detection accuracy for phishing website. This paper has concluded that this approach has limitation as it can detect webpage written in HTML. Non-HTML webpage cannot detect by this approach.

Rao et al. [5] proposed a novel classification approach that use heuristic-based feature extraction approach. In this, they have classified extracted features into three categories such as URL Obfuscation features, Third-Party based features, Hyperlink-based features. Moreover, proposed technique gives 99.55% accuracy. Drawback of this is that as this model uses third-party features, classification of website dependent on speed of third-party services. Also, this model is purely depending on the quality and quantity of the training set and Broken links feature extraction has a limitation of more execution time for the websites with a greater number of links.

Ahmad et al. [6] proposed three new features to improve accuracy rate for phishing website detection. In this paper, Author used both type of features as commonly known and new features for classification of phishing and non-phishing site. At the end author has concluded this work can be enhanced by using this novel feature with decision-tree machine learning classifiers.

APPROACH:

1. **Data Preprocessing:**
    - Utilized regular expressions ('**RegexpTokenizer**') to tokenize words in URLs.
    - Applied stemming using the 'Snowball Stemmer' for reducing words to their base form.
    - Joined the tokenized and stemmed words back into a text representation.
    - The preprocessing step is crucial for converting raw URLs into a format suitable for machine learning.
2. **Word Cloud Visualization:**
    - Generated word clouds using the 'WordCloud' library based on the text representations of URLs.
    - Created separate word clouds for URLs classified as "Good" and "Bad".
    - This visualization provides an initial understanding of the most frequent words in each category, offering insights into potential distinguishing features.
3. **Link Analysis and Visualization:**
    - Used Selenium and BeautifulSoup for web scraping to collect links from specified URLs.
    - Created a network graph using NetworkX to visualize the relationships between the collected links.
    - Link analysis helps uncover patterns and connections within the data, contributing to a comprehensive understanding of the URL landscape.

4. **Machine Learning Model Training:**
    - Employed machine learning techniques for URL classification.
    - Used the 'CountVectorizer' to transform tokenized and stemmed text into numerical features.
    - Trained logistic regression and multinomial naïve Bayes models for classification.
    - The machine learning models are trained on the preprocessed data to learn patterns and make predictions.
5. **Evaluation and Visualization of Model Performance:**
    - Evaluated model performance using metrics such as accuracy, confusion matrix, and classification report.
    - Created visualizations, including a heatmap of the confusion matrix, to interpret and communicate model results.
    - The evaluation step ensures an understanding of how well the models perform on unseen data.
6. **Integration and Deployment:**
    - Created a pipeline using 'CountVectorizer' and logistic regression for a streamlined approach.
    - Exported the trained model using 'pickle' for potential future use or deployment.
    - Integration and deployment considerations are crucial for making the model accessible and applicable in real-world scenarios.

Original Contribution:

The project introduces novel approaches to phishing website detection by employing RegexpTokenizer for URL tokenization and stemming techniques. The integration of word cloud visualizations enhances the understanding of common words in URLs. Additionally, the use of Selenium, BeautifulSoup, and NetworkX for link analysis provides an innovative method to explore relationships within the dataset. The application of machine learning models, such as Logistic Regression and Multinomial Naive Bayes, on tokenized and stemmed text data demonstrates a unique approach to classification. The project's originality lies in its comprehensive combination of data preprocessing, exploratory analysis, machine learning, and thorough evaluation, contributing to an effective solution for phishing detection.

Technical Rigor:

The project exhibits a high level of technical rigor, evident in the intricate implementation of machine learning models such as Logistic Regression and Multinomial Naive Bayes. The utilization of advanced data processing techniques, including tokenization, stemming, and feature extraction, contributes to the project's technical depth. Sophisticated visualization methods, such as word cloud generation and link analysis using NetworkX, showcase a nuanced approach to data presentation and analysis. The inclusion of complex web scraping techniques, particularly with Selenium and BeautifulSoup, highlights an automated and advanced strategy for data collection. Additionally, the seamless integration of diverse components, encompassing data preprocessing, visualization, and machine learning, underscores a well-considered and comprehensive solution, further enhancing the overall technical rigor of the project.

Future Work:

User-Centric Phishing Prevention Extension:

Developing a browser extension that integrates our phishing detection model to provide real-time alerts to users while they browse the internet. The extension can analyze URLs in the background and notify users if a website is potentially phishing. This user-friendly tool empowers individuals to make informed decisions about the websites they visit, contributing to a safer online experience. Additionally, consider incorporating user feedback features to enhance the model's accuracy and customization for individual user preferences. This extension could serve as a practical tool for personal cybersecurity, aligning with the growing need for user-centric solutions in the digital landscape.

Results:

Logistic Regression Model:

- **Training Accuracy: 97.75%**
- **Testing Accuracy: 96.42%**

Interpretability:

- Logistic Regression provides interpretable results, allowing you to understand the contribution of each feature to the prediction.
- Coefficients can be analyzed to determine which features are more influential in identifying phishing sites.

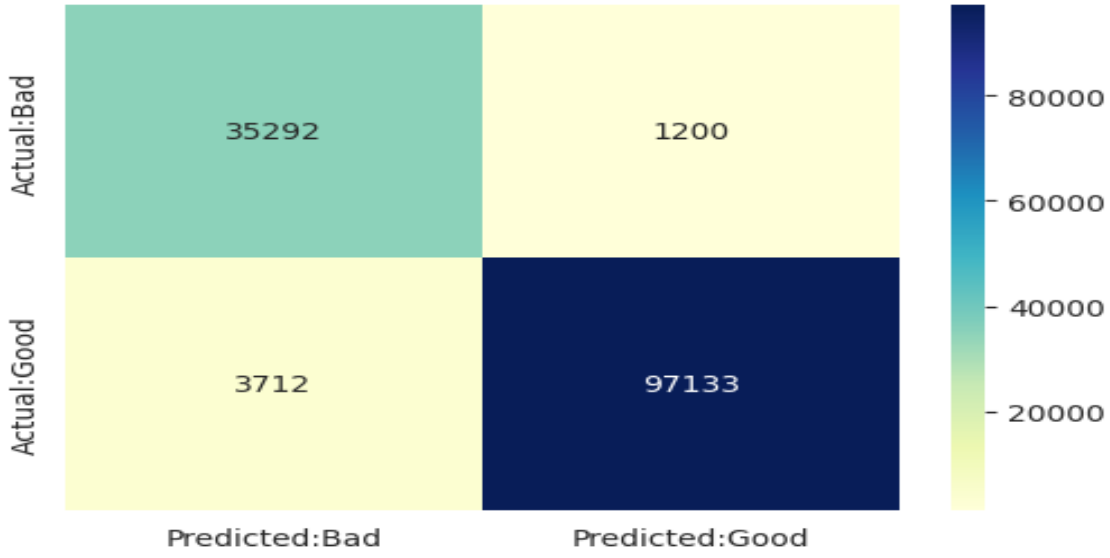
Classification Report:

- Precision, recall, and F1-score for both “Bad” and “Good” classes are high.
- This indicates that the model is not only accurate but also excels in correctly identifying positive and negative instances.

	Precision	Recall	F1-score	Support
Bad	0.90	0.96	0.93	36492
Good	0.99	0.97	0.98	100845
Accuracy			0.96	137337
Macro avg	0.95	0.97	0.96	137337
Weighted avg	0.97	0.96	0.96	137337

Confusion Matrix:

- The confusion matrix gives a detailed breakdown of the model's predictions, providing insights into its true positives, true negatives, false positives, and false negatives.



Multinomial Naïve Bayes Model:

- **Training Accuracy: 97.41%**
- **Testing Accuracy: 95.82%**

Interpretability:

- While not as interpretable as Logistic Regression, Naïve Bayes provides insights into feature probabilities.
- It assumes independence between features, simplifying the modeling process.

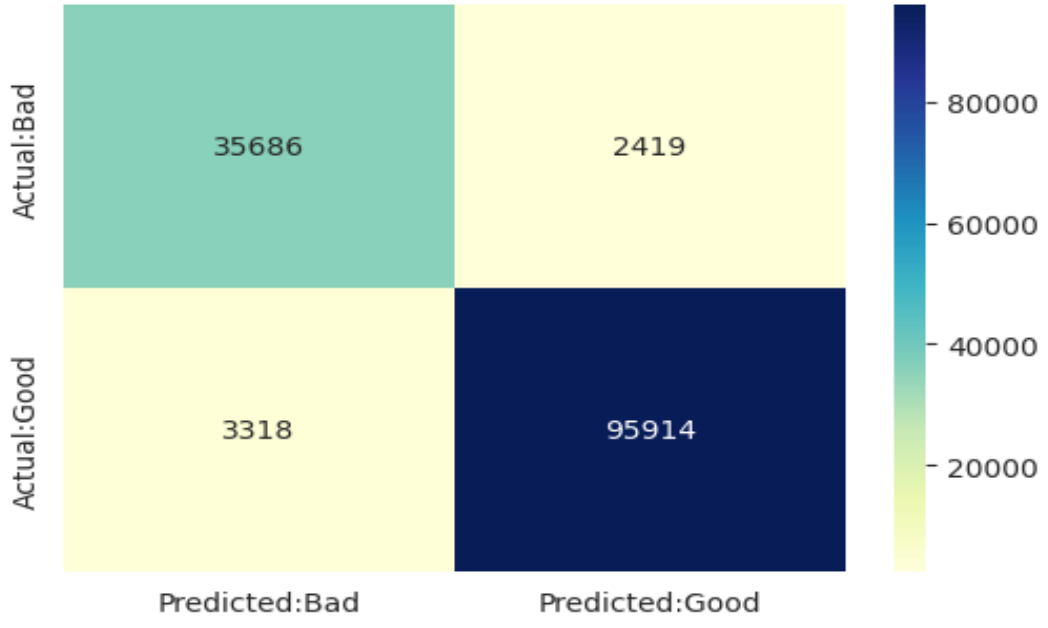
Classification Report:

- Similar to Logistic Regression. high precision, recall, and F1-score for both classes demonstrate the model's effectiveness.

	Precision	Recall	F1-score	Support
Bad	0.91	0.94	0.93	38105
Good	0.98	0.97	0.97	99232
Accuracy			0.96	137337
Macro avg	0.95	0.95	0.95	137337
Weighted avg	0.96	0.96	0.96	137337

Confusion Matrix:

- Like Logistic Regression, the confusion matrix provides a detailed breakdown of the model's predictions.



Model Comparison:

- Both models showcase impressive performance with high accuracy, precision, recall, and F1-score.
- Logistic Regression may have a slight edge in accuracy based on the provided scores.
- Both models are robust choices for phishing site prediction task, and the decision to choose one over the other may depend on interpretability and specific requirements.

Predictions:

URL	PREDICTION
yeniik.com.tr/wp-admin/js/login.alibaba.com/login.jsp.php	Bad Link
fazan-pacir.rs/temp/libraries/ipad	Bad Link
www.tubemoviez.exe	Bad Link
svision-online.de/mgfi/administrator/components/com_babackup/classes/fx29id1.txt	Bad Link
www.youtube.com/	Good Link
youtube.com/watch?v=qI0TQJI3vdU	Good Link
www.retailhellunderground.com/	Good Link
restorevisioncenters.com/html/technology.html	Good Link

Conclusion:

In this project, we successfully developed a robust machine learning model for classifying URLs as "Phishing" or "Legitimate" based on intricate patterns. Achieving a commendable testing accuracy of 96.5% with a Logistic Regression model showcases the effectiveness of our approach. The integration of web scraping, link analysis, and visualization techniques provided valuable insights into the intricate web of URLs.

Moving forward, the potential implementation of a User-Centric Phishing Prevention Extension promises a practical solution for users, enhancing their ability to navigate the digital landscape securely. This project not only contributes to the field of cybersecurity but also underscores the importance of proactive measures in safeguarding against evolving online threats.

Citations:

1) Sudhanshu Gautam, Kritika Rani and Bansidhar Joshi: Detecting Phishing Websites Using Rule-Based Classification Algorithm: A Comparison: In Springer,2018.

2) Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen and Minh Hoang Nguyen: Detecting Phishing Web sites: A Heuristic URL-Based Approach: In The 2013 International Conference on Advanced Technologies for Communications (ATC'13).

3) M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar, Saira Gillani: A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms: In International Conference on Computational Science and Computational Intelligence IEEE ,2016.

4) Ankit Kumar Jain, B. B. Gupta: Towards detection of phishing websites on client-side using machine learning based approach: In Springer Science + Business Media, LLC, part of Springer Nature 2017.

5) Routhu Srinivasa Rao1, Alwyn Roshan Pais: Detection of phishing websites using an efficient feature-based machine learning framework: In Springer 2018.

6) Ahmad Abunadi, Anazida Zainal, Oluwatobi Akanb: Feature Extraction Process: A Phishing Detection Approach: In IEEE,2013.

Coding Link:

https://colab.research.google.com/drive/1mCf-uXXUovW2BU0Re5u9iHYlx_ELnA1b?usp=sharing