Phishing website detection using machine learning(classification)

Eng/Mohamed Moheeb , DR/Mohamed Abo Rizka

Department of computer science ,college of computing and information technology, Arab Academy for Science and ,Technology &maritime transport

Introduction

- *Phishing attacks pose a significant threat to online security, targeting individuals and organizations worldwide. These malicious attempts aim to deceive users into revealing sensitive information such as login credentials, financial data, or personal details. As phishing techniques become increasingly sophisticated, there is a growing need for effective solutions to detect and prevent these attacks.*
- *The prediction of phishing websites plays a crucial role in safeguarding users against such threats. By utilizing machine learning algorithms and data analysis techniques, predictive models can analyze the characteristics and features of websites to determine their likelihood of being phishing sites. This proactive approach enables users and security systems to identify and avoid potential risks before falling victim to phishing attacks.*
- *Predicting phishing websites offers several benefits. First and foremost, it enhances online security by empowering individuals, organizations, and internet platforms to take preventive measures against phishing attacks. It helps reduce the risks associated with financial losses, data breaches, and reputational damage caused by falling victim to phishing scams.*

Objective

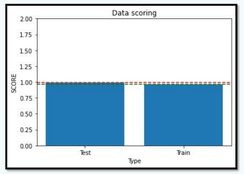
- *The main goal of this project is to develop a predictive model that can accurately identify and classify phishing websites. The project aims to address the problem of online security threats posed by phishing attacks and provide a solution to help individuals, organizations, and internet users in general*
- *Building Trust: Predicting phishing websites contributes to building trust between internet users and digital platforms or services. By implementing robust security measures and providing reliable protection against phishing attacks, the objective helps maintain user confidence and trust in online interactions.*
- *Enhancing Online Security: By accurately predicting phishing websites, the objective aims to enhance online security for individuals, organizations, and internet users in general. It seeks to prevent users from falling victim to phishing attacks and mitigate the risks associated with fraudulent activities.*
- *the project is helpful for all people who surfing the internet and all social media users because phishing URL helps hackers to access social media application like Facebook, Instagram, twitter, and others through tool called ZPhisher and this tool used on kali Linux. (Classification problem)*

Methods

Plotting for train and test score of all features

```
In [15]: clf.score(X_train, y_train)
Out[15]: 0.9910663801877191

In [16]: clf.score(X_test, y_test)
Out[16]: 0.9615558570782451
```

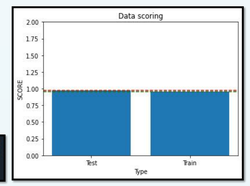


Choose important features from our dataset

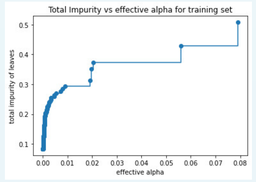
Features	class
HTTPS	0.714704
Anchor URL	0.692895
Prefix Suffix	0.348588
Website Traffic	0.346003
Sub Domains	0.298231
Request URL	0.253478
Links in Script Tags	0.248415
Server Form Handler	0.221380
Google Index	0.129000
Age of Domain	0.121402
PageRank	0.104593
Using IP	0.094033

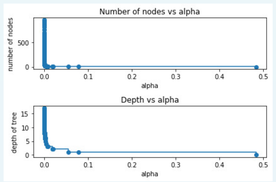
observed the overfitting in chosen important features data

```
TRAIN Accurecy : 0.9699197105054845
TEST Accurecy : 0.9525101763907734
```



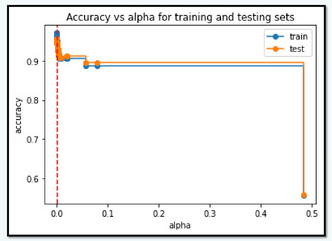
repair overfitting by helping of cost reduction and complexity pruning algorithm also reduces the depth of the tree.

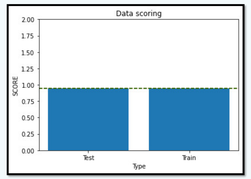




Methods Cont.

Get all possible alpha data based on train and test data and then create every possible tree and then get minimum distance between train and test data for each tree to get the best alpha






```
TRAIN Accurecy : 0.9460590297410381
TEST Accurecy : 0.9439167797376753
```

results

Based on the best alpha shows the best tree at which train, and test are the closest

URL input → our model will parse it → result

```
[1, 0, -1, 1, -1, -1, -1, 0, -1, 1, -1, -1]
# Decission Tree Classification :
RESULT : PHISHING
# Logistic Regression :
RESULT : NOT PHISHING
```

Decision tree VS logistic regression

Conclusion

In conclusion, predicting phishing websites is a challenging task that requires a combination of technological advancements, user awareness, and proactive measures. While no prediction method can guarantee 100% accuracy, several approaches can significantly reduce the risk of falling victim to phishing attacks.

User Education and Awareness: Educating users about the common signs of phishing, such as suspicious URLs, grammatical errors, and requests for personal information, is crucial. By increasing user awareness, individuals can become more cautious and better equipped to recognize and avoid phishing attempts.

While these measures can enhance the detection and prevention of phishing attacks, it is important for individuals and organizations to remain vigilant and exercise caution while browsing the internet. Combining technological defenses with user awareness and responsible online behavior is key to minimizing the risk posed by phishing websites.

References

https://scikit-learn.org/

https://www.kaggle.com/

Acknowledgement

all praise is due to Allah alone, the degree would not have been possible without the interminable support and guidance of my adorable supervisors, prof. Dr. Mohamed Abo rizka for his encouragement, continuous support, patience, motivation and invaluable guidance and Eng. Mohamed Moheeb for his continuous support, encouragement, caring kind words, listening and understanding us every time