

Web Phishing Detection using IBM Watson



Team Members -

Anamika Lochab - 18BCE10035

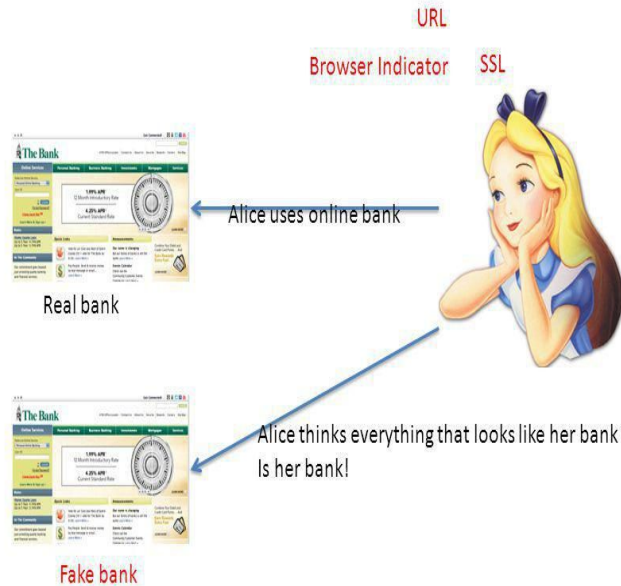
Arpita Pandey - 18BCE10060

Vitti Gupta - 18BCE10299

Vishwas - 18BCG10104

Sneha Rani - 18BOE10060

Introduction



- One of the biggest threats to web security is Phishing. Phishing is the technique of extracting user credentials by masquerading as a genuine website or service over the web.
- Phishing is popular among attackers since it is easier to trick someone into clicking a malicious link that seems legitimate than trying to break through a computer's defence systems.
- Typically a victim receives a message that appears to have been sent by a known contact or organization. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate content.
- The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.
- Phishing attacks costs internet users billions of dollars per year.

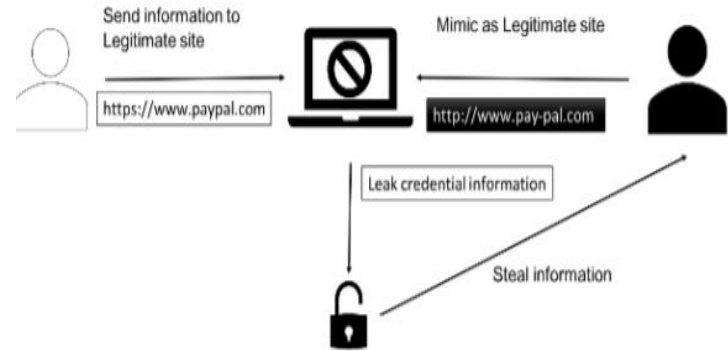
Non-Content based approaches:

- **URL based phishing detection** - Overall, URL-based methods perform faster than any other, including content and visual-similarity based approaches. More importantly, they work well on zero-hour phishing attacks, which are becoming a major concern in modern anti-phishing society.
- **Blacklisting** -Web browsers—such as Google Safe Browsing – that defend against phishing attacks by updating a list of black-listed sites. However, since their proposed system relies on third-party services (like Google) for searching domain name to compare top results, it results in poor performance. Furthermore, blacklist approaches encounter the major issue of zero-hour phishing attacks because newly created phishing sites are not in the list
- **Whitelisting** - Although whitelist-based methods seem effective for phishing detection, there is a limitation on getting legitimate sites all on the web. An abundant list of reliable websites is necessary for a robust system with high accuracy; otherwise, false positive rates increase due to a lack of white-listed websites information, which is practically impossible to collect all legitimate sites in the world.

Architecture of URL-Based Phishing

URL-based phishing attacks are mainly performed by embedding sensitive words or characters in a link that:

1. Mimic similar but misspelling words.
2. Contain special characters for redirecting.
3. Use shortened URLs.
4. Use sensitive keywords which seem reliable.
5. Add a malicious file in the link and so on

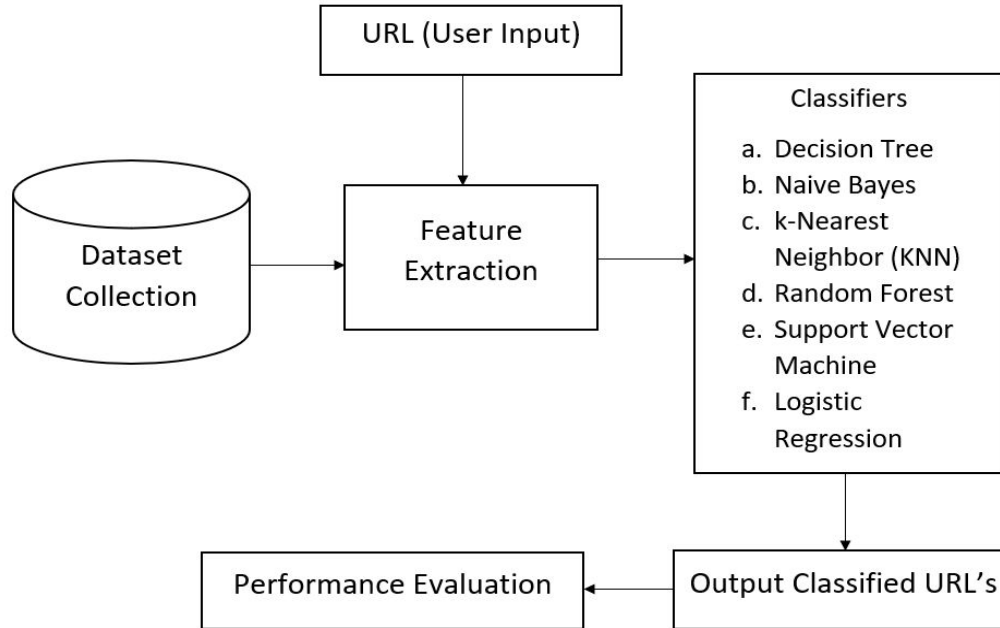


Machine learning classifiers and methods to detect a phishing website

Detecting and identifying Phishing Website is a really complex and dynamic problem. Machine learning has been widely used in many areas to create automated solutions. Depending on the application and nature of the dataset used we can use any classification algorithms mentioned below. As there are different applications, we can not differentiate which of the algorithms are superior or not. Each of classifiers have its own way of working and classification.

- Naive Bayes
- Logistic regression
- K-nearest neighbors
- SVM
- Decision tree
- Random Forest

Approach



The dataset of phishing and legitimate URL's is given to the system which is then pre-processed so that the data is in the useable format for analysis. The features have around 30 characteristics of phishing websites which is used to differentiate it from legitimate ones. Each category has its own characteristics of phishing attributes and values are defined. The specified characteristics are extracted for each URL and valid ranges of inputs are identified.

Sr. No	Feature name	Description
1	IP address	Whether domain is in the form of an IP address
2	Length of URL	Length of URL
3	Suspicious character	Whether URL has @, //
4	Prefix and suffix	Whether URL has –
5	Length of subdomain	Length of subdomain
6	Number of /	Number of / in URL
7	HTTPS protocol	Whether URL use https.
8	Phishing words in URL	Whether url has phishing terms
9	Number of .	Number of dots . in url

Algorithms

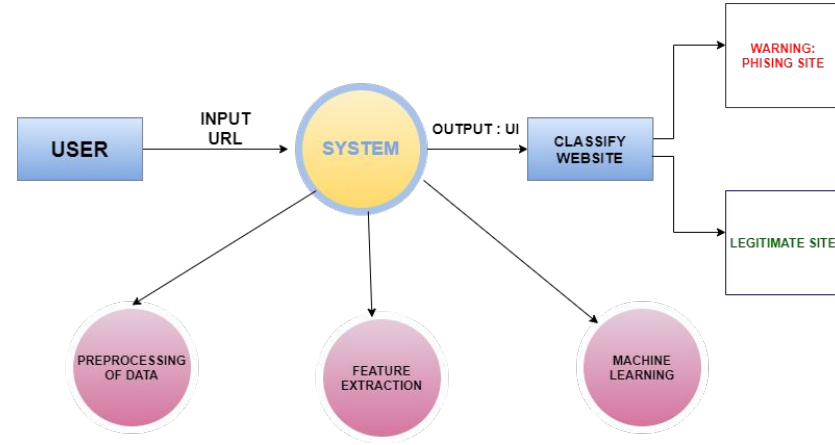
After this the data is trained we shall apply a relevant machine learning algorithm to the dataset. The machine learning algorithms.

Finally, we go with logistic regression with 91.67% accuracy to avoid overfitting.

	Model	Test Score
2	KNN	0.963365
5	Decision Tree	0.962913
3	SVM-rbf	0.940751
6	Random Forest	0.938942
0	Logistic Regression	0.916780
4	SVM-sigmoid	0.832655
1	Naive Bayes	0.615106

ADVANTAGES


- Use of new classification features and algorithms with improved accuracy
- More adaptable
- Increased accuracy and decreased false positive rate
- Provide secure and healthy online shopping and e-banking environment to the users
- Do not require changes in authentication platforms
- Do not rely on the user's ability to detect phishing
- Easy deployment of our phishing detection model to end users




Prediction


127.0.0.1:5000/predict


Guest



NILA


HOME


ABOUT



CONTACT

Phishing Website Detection using Machine Learning

Enter the URL to be verified

Predict

Windows search bar



System tray: Help, Cloud, 32°C, Network, Volume, ENG, 12:27, 03-08-2021

NILA



HOME



ABOUT



CONTACT

Phishing Website Detection using Machine Learning

Predict

You are on the wrong site. Be cautious!

[http://activate.facebook.fblogins.net/88adbao798283o8298398?
login.asp](http://activate.facebook.fblogins.net/88adbao798283o8298398?login.asp)



32°C Light rain

12:29
03-08-2021

Conclusion

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any of the machine learning algorithm with the classifier.

In this project, we built a mechanism to detect phishing websites. Our methodology uses not just traditional URL based or content based rules but rather employs the machine learning technique to identify not so obvious patterns and relations in the data. We have used features from various domains spanning from URL to HTML tags of the webpage, from embedded URLs to favicon. To check the traffic and status of the website. We were able to obtain an accuracy of more than 91% thus classifying most websites correctly and proving the effectiveness of the machine learning we are using Logistic Regression technique to attack the problem of phishing websites. We provided the output as a user-friendly web platform which can further be extended to a browser extension to provide safe and healthy online space to the users.



THANK
YOU