Phishing Website Detection Web

Souhardya Gayen

26/06/2023

Abstract

The predefined anti-phishing approaches based on machine learning techniques extract few features from other sources which detects the fraudulent websites comparatively slower and unfit for real-time execution. This paper presents a solution to this problem by deeply examining various characteristics of phishing as well as legitimate websites and analysing thirty outstanding features to distinguish phishing websites from legitimate URLs. These thirty features scan the URLs thoroughly and extract their values from the URLs of the websites and are thus independent of any third-party presence which gives an accuracy of 93%.
This results in a more realistic, reliable, resourceful, well-informed computational approach.

1.0 Introduction

Phishing is a social engineering attack that aims at exploiting the weakness found in the system at the user's end. For example, a system may be technically secure enough for password theft but the unaware user may leak his/her password when the attacker sends a false update password request through forged (phished) website. For addressing this issue, a layer of protection must be added on the user side to address this problem. A phishing attack is when a criminal sends an email or the url pretending to be someone or something he's not, in order to get sensitive information out of the victim. The victim in regard to his/her curiosity or a sense of urgency, they enter the details, like a username, password, or credit card number, they are likely to acquiesce. The recent example of a Gmail phishing scam that targeted around1 billion Gmail users worldwide.

Phishing attacks have become anxiety for the cyber world. It causes enormous problems for privacy and financial issues of internet users. Scammers, namely fishers, create false websites to feel and look like a genuine to deceive the people. They spoof emails to steal the identity of legitimate users. They gather personal covert information, password, account information, and credit card details for the transaction. Fishers always change their strategy to attack the system. Social engineering is one of the essential techniques the fishers use. Using this technique, they gather personal credentials from a trustworthy person. Phishers create false websites and spoof email in such a way that they are very similar and sometimes look like a real company website that comes from a source. Sometimes the attackers act like a real source and force the users to update the system. Moreover, they threaten the customer to suspend the account and demand ransom. Email spoofing is another technique used for phishing fraud. Customers are usually misled to disclose private information like passwords and credit card number. Thus phishing is mainly used to steal valuable information such as bank account, password, and credit card details. This type of scam is increasing rapidly, and individuals, business-people are losing their trust in online business. Thus, a negative impression of clients on online business was swarmed as they lost faith in online transactions. Even though encryption software is used to protect the information in the computers' storage, they are also vulnerable to attacks. In this paper, the detection of fishing was performed through ML.

As phishing attack allows attackers a foothold in corporate networks causing access to vital information, it is important to safeguard users from becoming victims of fraud. Thus phishing detection tools play a vital role in ensuring security. So we planned to work on this topic.

2.0 Problem Statement

Malicious links will lead to a website that often steals login credentials or financial information like credit card numbers. Attachments from phishing emails can contain malware that once opened can leave the door open to the attacker to perform malicious behaviour from the user's computer. Phishing causes social engineering attack. It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email, instant message, or text message.

The Harmful Effects of Phishing on Businesses:

- Loss of Data
- Damaged Reputation
- Direct Monetary Loss
- Loss of Productivity
- Loss of Customers
- Financial Penalties
- Intellectual Property Theft
- Loss of Company Value

3.0 Assessment

There are several methods used to assess the detection of phishing websites, including vulnerability assessments, penetration testing, and user awareness training.

Vulnerability assessments are used to identify weaknesses in an organization's security systems that could be exploited by attackers. This assessment typically involves scanning network and system infrastructure to detect vulnerabilities that could be used to launch a phishing attack.

Penetration testing, on the other hand, is a more targeted approach that involves simulating an attack on an organization's systems to identify weaknesses that could be exploited. This testing can help identify specific weaknesses in an organization's security system that could be used to launch a phishing attack.

Finally, user awareness training is an essential aspect of assessing the detection of phishing websites. By training users to recognize the signs of a phishing website, such as suspicious URLs or requests for sensitive information, organizations can help prevent attacks from succeeding.

Overall, assessing the effectiveness of measures used to detect phishing websites is critical for online security. By using vulnerability assessments, penetration testing, and user awareness training, organizations can identify weaknesses in their security systems and take proactive steps to prevent attacks. By continuously assessing and improving their detection capabilities, organizations can help protect themselves and their users from the harm caused by phishing attacks.

4.0 Target Specification

1. **Objective:** The main objective of the project is to develop a web-based application that can detect and prevent phishing attacks.
2. **User Interface:** The application should have a user-friendly interface that allows users to input a URL or a webpage, and the system should be able to determine whether it is a phishing website or not.
3. **Detection Techniques:** The application should use various detection techniques such as machine learning algorithms, heuristic analysis, blacklists, and whitelists to identify phishing websites accurately.
4. **Database:** The application should have a database that stores information about known phishing websites, including URLs, IP addresses, and other relevant information.
5. **Notification System:** The application should have a notification system that alerts users when they attempt to visit a known phishing website. The notification should include information about the potential risk of visiting the website and instructions on how to avoid it.
6. **Reporting System:** The application should have a reporting system that allows users to report suspected phishing websites. The system should also have a feedback mechanism that enables users to provide feedback on the accuracy of the detection system.

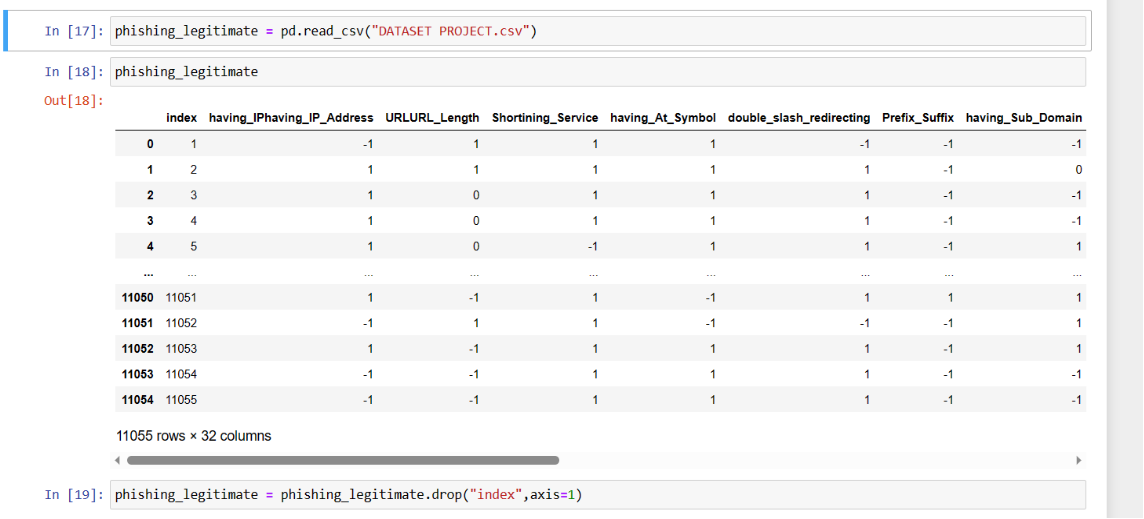
7. **Security:** The application should be designed with security in mind, including encryption of user data, secure login credentials, and protection against hacking attempts.
8. **Compatibility:** The application should be compatible with various web browsers and operating systems to ensure that it can be used by a wide range of users.
9. **Scalability:** The application should be scalable to accommodate future updates, increased traffic, and additional features.
10. **Maintenance:** The application should be easy to maintain and update to ensure that it remains effective and up-to-date with the latest phishing attack trends.

5.0 External Search

I use the Online shopper dataset for this project Dataset can be found here: https://www.kaggle.com/datasets/akashkr/phishing-website-dataset

Source:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0258361
https://www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8504731/
https://www.sciencedirect.com/topics/computer-science/website-phishing-detection
https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5#:~:text=Phishing%20is%20a%20form%20of%20fraud%20in%20which%20the%20attacker,email%20or%20other%20communication%20channels.

Let’s view our dataset:



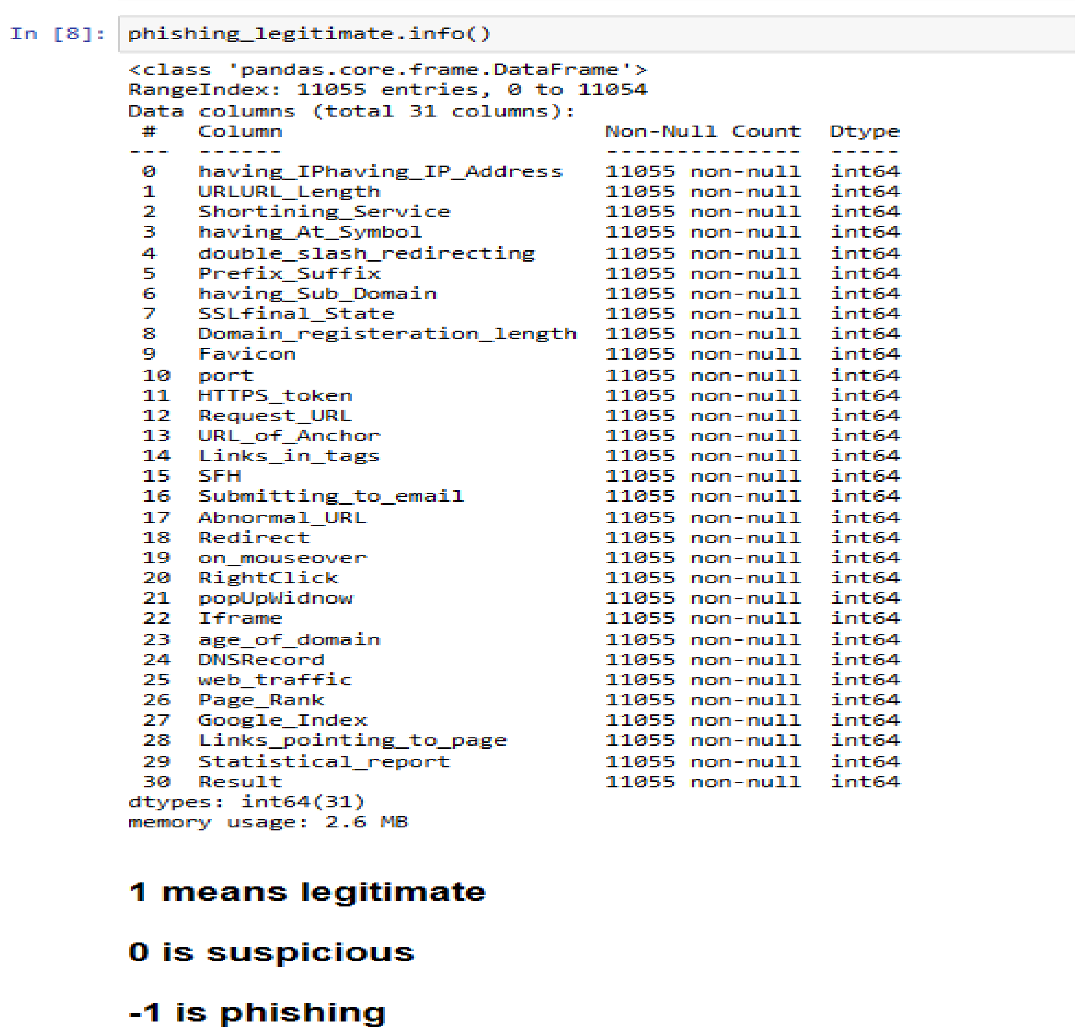
```
In [17]: phishing_legitimate = pd.read_csv("DATASET PROJECT.csv")

In [18]: phishing_legitimate
Out[18]:
```

	index	having_IPhaving_IP_Address	URLURL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain
0	1	-1	1	1	1	-1	-1	-1
1	2	1	1	1	1	1	-1	0
2	3	1	0	1	1	1	-1	-1
3	4	1	0	1	1	1	-1	-1
4	5	1	0	-1	1	1	-1	1
...	...	...	...	...	...	...	...	...
11050	11051	1	-1	1	-1	1	1	1
11051	11052	-1	1	1	-1	-1	-1	1
11052	11053	1	-1	1	1	1	-1	1
11053	11054	-1	-1	1	1	1	-1	-1
11054	11055	-1	-1	1	1	1	-1	-1

11055 rows × 32 columns

```
In [19]: phishing_legitimate = phishing_legitimate.drop("index",axis=1)
```

See some information about our dataset



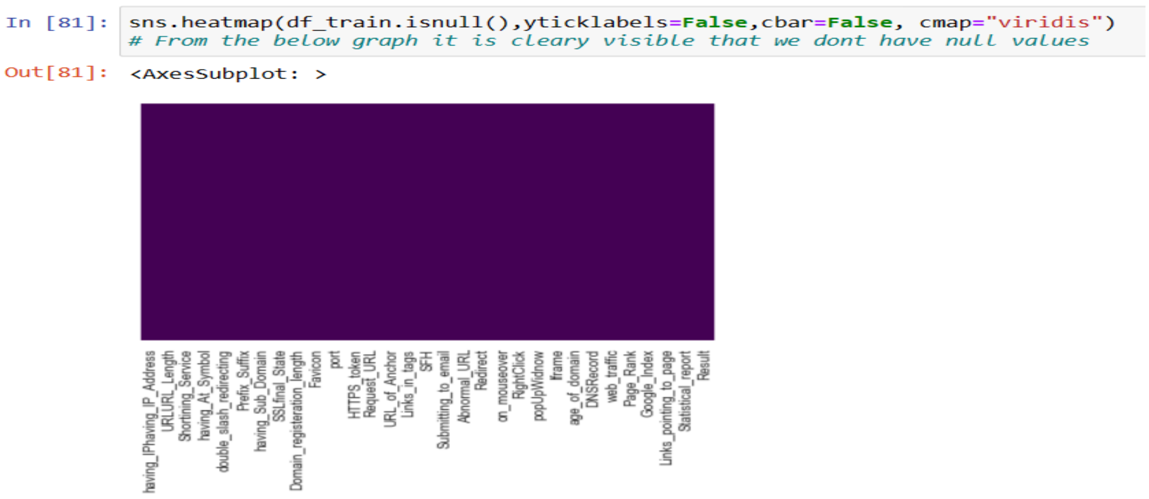
```
In [8]: phishing_legitimate.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11055 entries, 0 to 11054
Data columns (total 31 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   having_IPhaving_IP_Address         11055 non-null  int64
 1   URLURL_Length                      11055 non-null  int64
 2   Shortining_Service                 11055 non-null  int64
 3   having_At_Symbol                   11055 non-null  int64
 4   double_slash_redirecting           11055 non-null  int64
 5   Prefix_Suffix                      11055 non-null  int64
 6   having_Sub_Domain                  11055 non-null  int64
 7   SSLfinal_State                     11055 non-null  int64
 8   Domain_registeration_length        11055 non-null  int64
 9   Favicon                            11055 non-null  int64
 10  port                               11055 non-null  int64
 11  HTTPS_token                        11055 non-null  int64
 12  Request_URL                        11055 non-null  int64
 13  URL_of_Anchor                      11055 non-null  int64
 14  Links_in_tags                      11055 non-null  int64
 15  SFH                                11055 non-null  int64
 16  Submitting_to_email                11055 non-null  int64
 17  Abnormal_URL                       11055 non-null  int64
 18  Redirect                           11055 non-null  int64
 19  on_mouseover                       11055 non-null  int64
 20  RightClick                         11055 non-null  int64
 21  popUpWidnow                        11055 non-null  int64
 22  Iframe                             11055 non-null  int64
 23  age_of_domain                      11055 non-null  int64
 24  DNSRecord                          11055 non-null  int64
 25  web_traffic                        11055 non-null  int64
 26  Page_Rank                          11055 non-null  int64
 27  Google_Index                       11055 non-null  int64
 28  Links_pointing_to_page             11055 non-null  int64
 29  Statistical_report                 11055 non-null  int64
 30  Result                             11055 non-null  int64
dtypes: int64(31)
memory usage: 2.6 MB
```

1 means legitimate

0 is suspicious

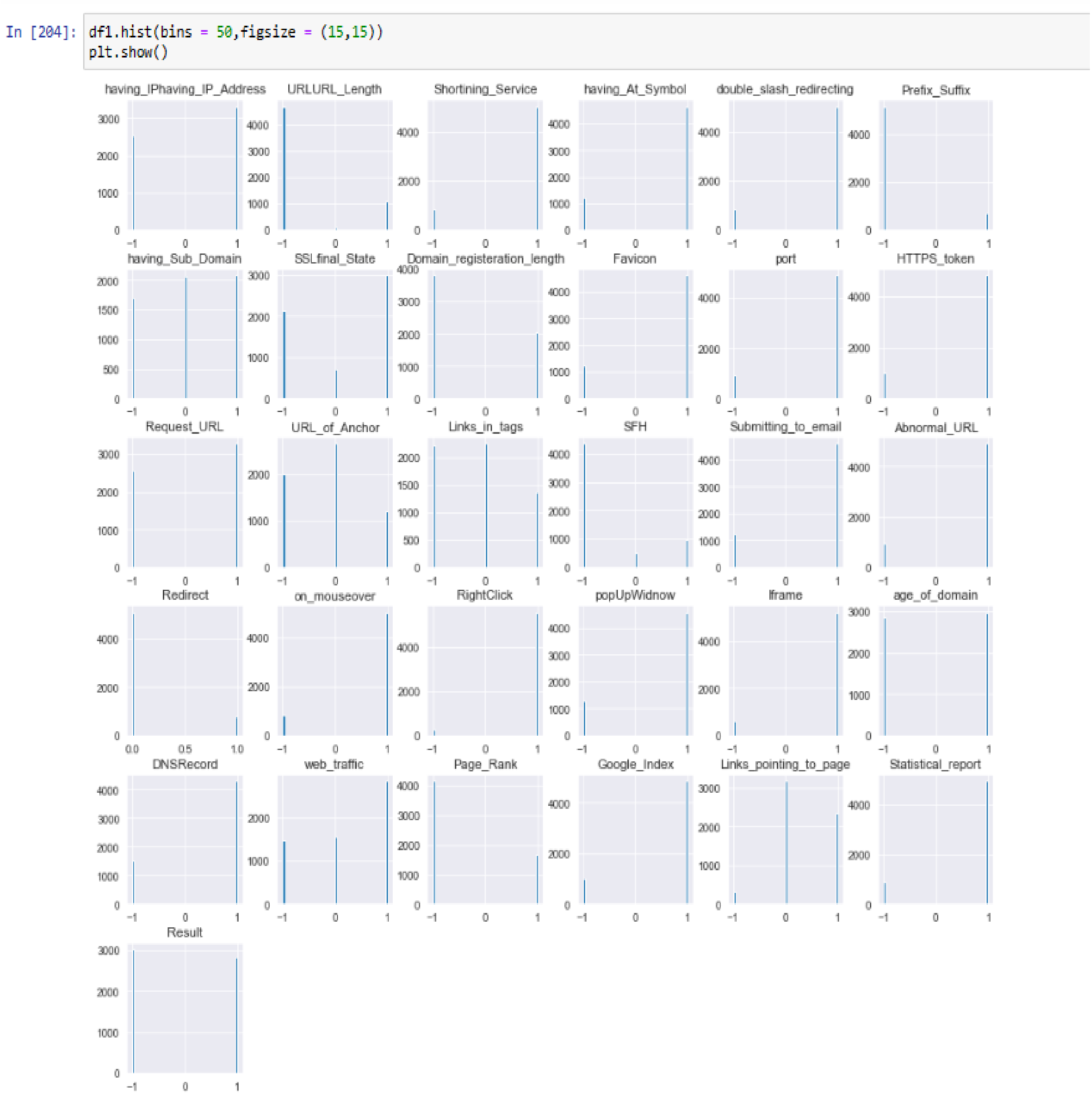
-1 is phishing

6.0 Benchmarking



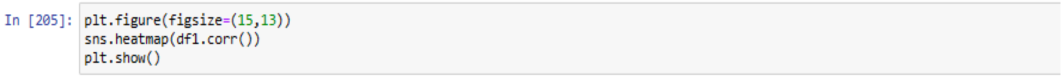
```
In [81]: sns.heatmap(df_train.isnull(),yticklabels=False,cbar=False, cmap="viridis")
         # From the below graph it is cleary visible that we dont have null values

Out[81]: <AxesSubplot: >
```

Histogram Plot For Phishing Data

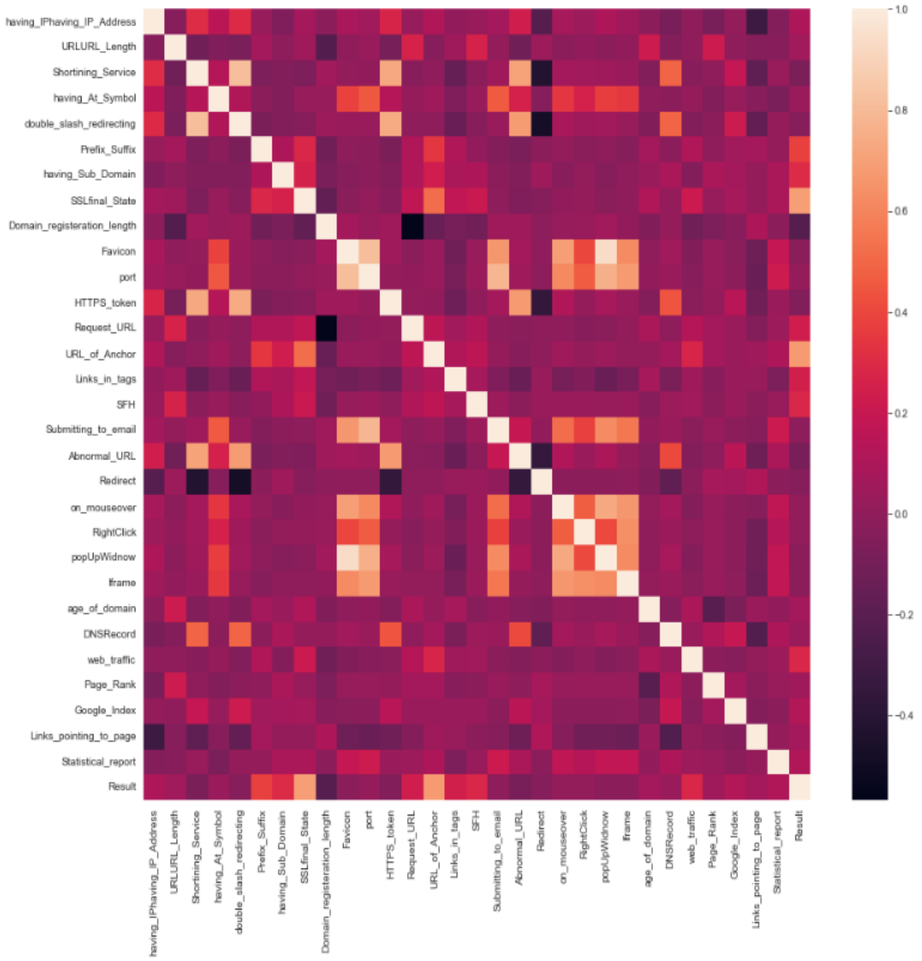


```
In [204]: df1.hist(bins = 50,figsize = (15,15))
          plt.show()
```

Correlation Matrix For Phishing Data



```
In [205]: plt.figure(figsize=(15,13))
          sns.heatmap(df1.corr())
          plt.show()
```



7.0 Applicable Regulations

Phishing is a type of online fraud that targets individuals and organizations by tricking them into giving away sensitive information, such as passwords or credit card numbers. To help prevent phishing attacks, there are several regulations in place that focus on the detection of phishing websites.

One of the most notable regulations is the Anti-Phishing Working Group (APWG), which is an international coalition dedicated to eliminating phishing and other online fraud. The APWG works to bring together industry, government, and law enforcement organizations to share information and best practices for detecting and preventing phishing attacks.

In addition to the APWG, another important regulation is the use of URL blacklists. URL blacklists are lists of known phishing websites that are automatically blocked by web browsers and other software. These lists are updated regularly to ensure that users are protected from the latest phishing threats.

Apart from these regulations, there are also technical measures that can be used to detect phishing websites. Machine learning algorithms can be used to analyze website content and identify suspicious patterns. Certificate validation can also be used to verify the authenticity of a website's SSL certificate.

Finally, user education is an important part of detecting phishing websites. Individuals should be aware of the signs of a phishing website, such as misspelled words, suspicious URLs, and requests for sensitive information. They should also be trained to recognize phishing emails and other forms of social engineering attacks.

In summary, the detection of phishing websites is critical for online security. Regulations such as the APWG and URL blacklists, along with technical measures and user education, are all essential tools for detecting and preventing phishing attacks. By working together, we can create a safer and more secure online environment for everyone.

8.0 Applicable Constraints

Detecting phishing websites can be constrained by technical limitations such as encryption and other security measures used by attackers, resource constraints for small and medium-sized businesses, legal restrictions on the use of certain detection technologies, and user behavior. Addressing these constraints is critical to improving anti-phishing capabilities and protecting organizations and individuals from the harm caused by phishing attacks.

9.0 Business Opportunity

The detection of phishing websites presents a significant business opportunity for companies offering anti-phishing solutions. As the threat of phishing attacks continues to grow, organizations are increasingly looking for effective ways to protect themselves and their users. This has created a growing market for anti-phishing solutions, including software, training, and consulting services.

10.0 Concept Generation

This section reflects on how I framed my concept with EDA technology, data cleaning, feature selection algorithms, Machine Learning algorithms and evaluated the performance. The following figure (Fig 1) gives a brief of the steps I have followed in our proposed methodology. It aims at building a website to detect whether the websites are fraud or trustworthy.

STEP 1: DATA COLLECTION

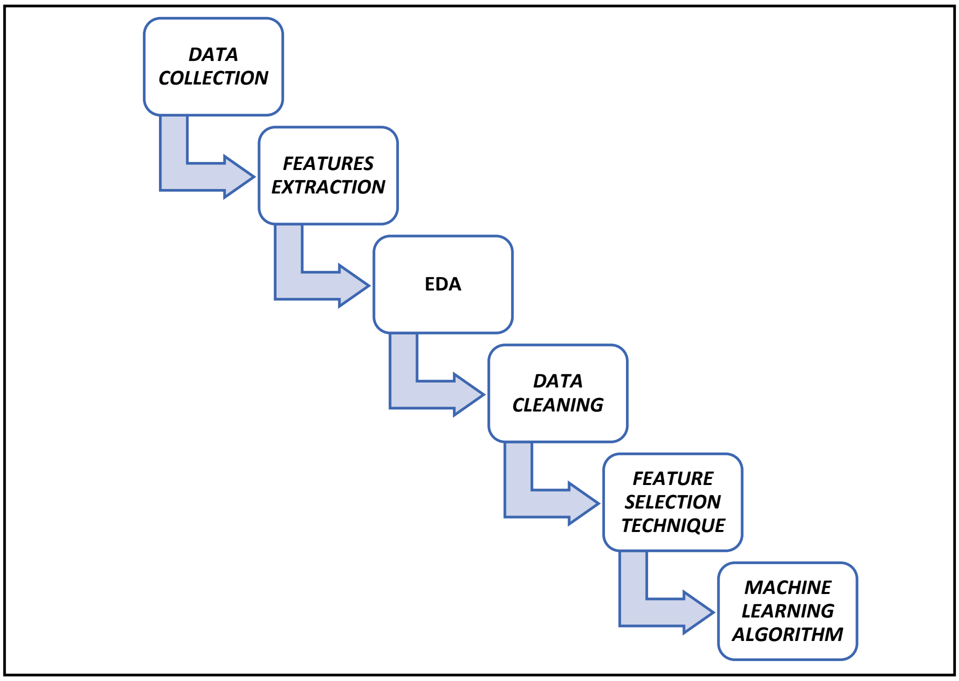
STEP 2: FEATURES EXTRACTION

STEP 3: EXPLORATORY DATA ANALYSIS

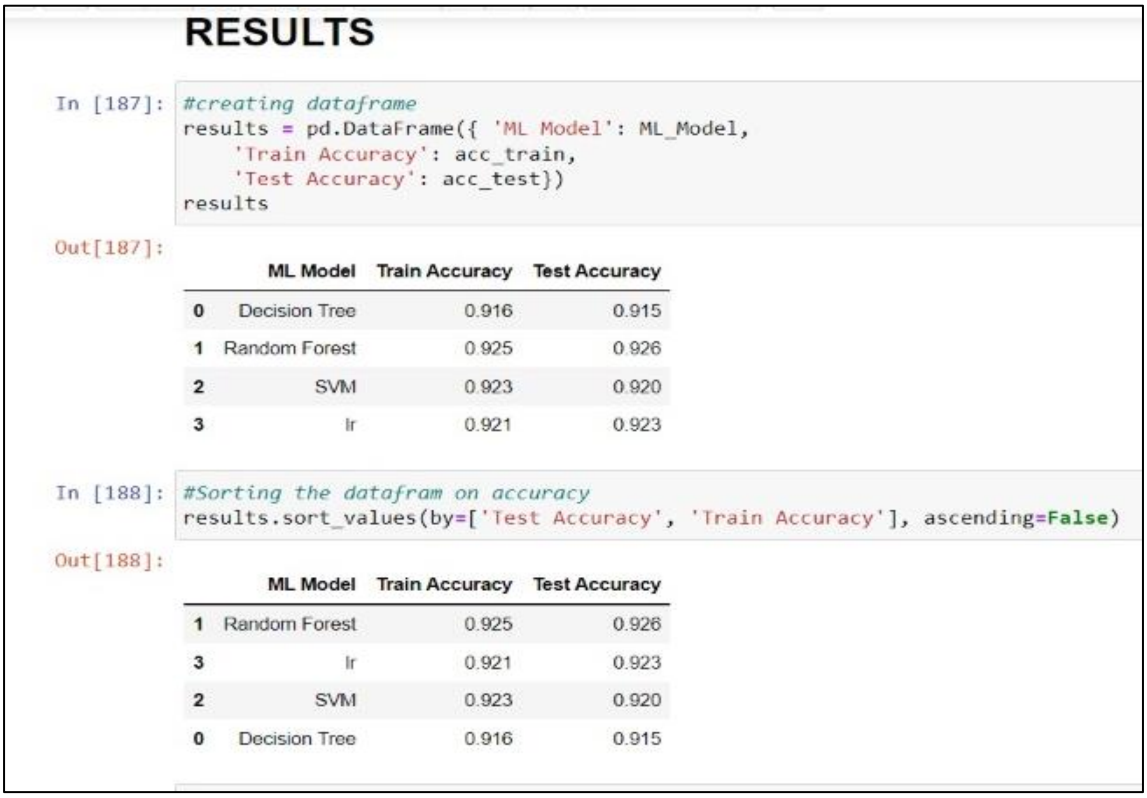
STEP 4: DATA CLEANING

STEP 5: FEATURE SELECTION TECHNIQUE

STEP 6: MACHINE LEARNING ALGORITHM



After applying the Machine Learning Algorithms, I have trained and tested the model. I have performed model evaluation to determine which model gives the better performance. My source code has produced the following result.



RESULTS

In [187]:
```
#creating dataframe
results = pd.DataFrame({ 'ML Model': ML_Model,
    'Train Accuracy': acc_train,
    'Test Accuracy': acc_test})
results
```

Out[187]:

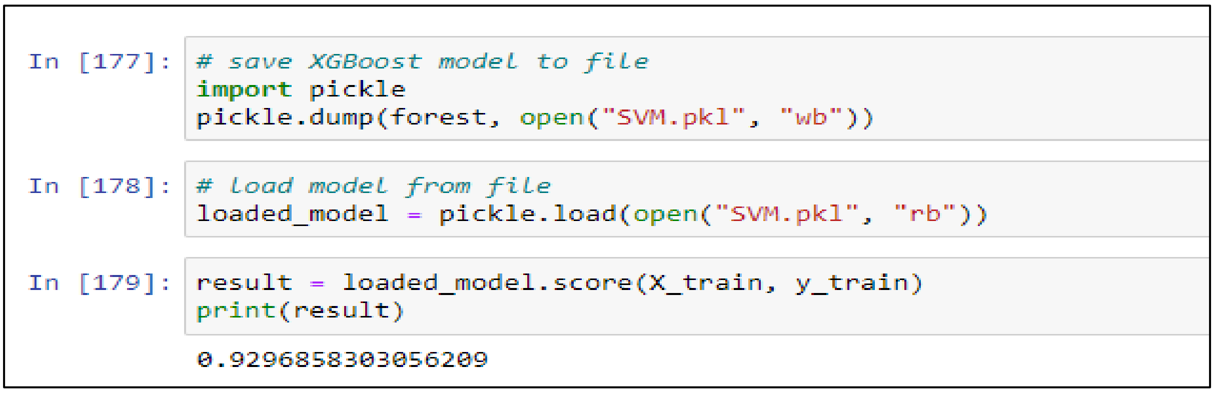
	ML Model	Train Accuracy	Test Accuracy
0	Decision Tree	0.916	0.915
1	Random Forest	0.925	0.926
2	SVM	0.923	0.920
3	lr	0.921	0.923

In [188]:
```
#Sorting the datafram on accuracy
results.sort_values(by=['Test Accuracy', 'Train Accuracy'], ascending=False)
```

Out[188]:

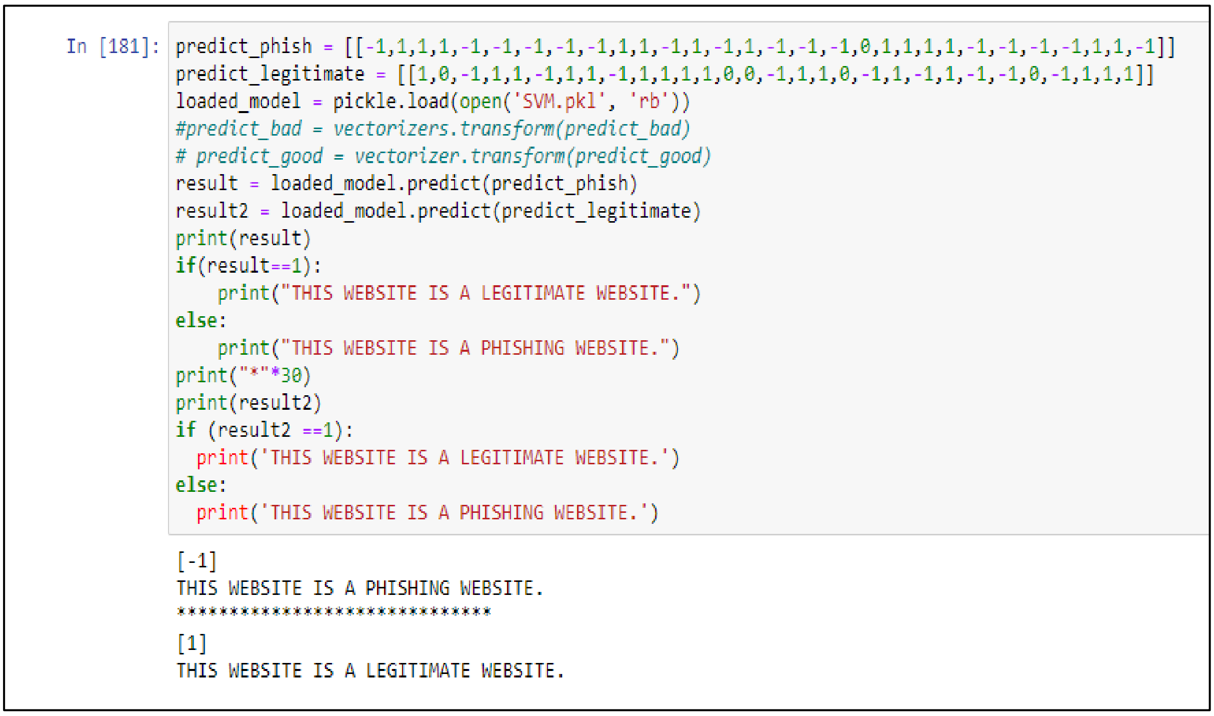
	ML Model	Train Accuracy	Test Accuracy
1	Random Forest	0.925	0.926
3	lr	0.921	0.923
2	SVM	0.923	0.920
0	Decision Tree	0.916	0.915

The Accuracy of the initial model is given below:



```
In [177]: # save XGBoost model to file
          import pickle
          pickle.dump(forest, open("SVM.pkl", "wb"))

In [178]: # load model from file
          loaded_model = pickle.load(open("SVM.pkl", "rb"))

In [179]: result = loaded_model.score(X_train, y_train)
          print(result)

          0.9296858303056209
```

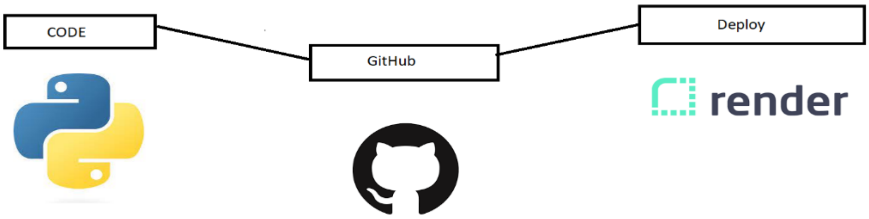
I have taken data from the data set and checked manually whether it is giving correct and appropriate result.



```
In [181]: predict_phish = [[-1,1,1,1,-1,-1,-1,-1,-1,1,1,-1,1,-1,1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,1,-1]]
          predict_legitimate = [[1,0,-1,1,1,-1,1,1,-1,1,1,1,1,0,0,-1,1,1,0,-1,1,-1,1,-1,-1,0,-1,1,1,1]]
          loaded_model = pickle.load(open('SVM.pkl', 'rb'))
          #predict_bad = vectorizers.transform(predict_bad)
          # predict_good = vectorizer.transform(predict_good)
          result = loaded_model.predict(predict_phish)
          result2 = loaded_model.predict(predict_legitimate)
          print(result)
          if(result==1):
              print("THIS WEBSITE IS A LEGITIMATE WEBSITE.")
          else:
              print("THIS WEBSITE IS A PHISHING WEBSITE.")
          print("*"*30)
          print(result2)
          if (result2 ==1):
            print('THIS WEBSITE IS A LEGITIMATE WEBSITE.')
          else:
            print('THIS WEBSITE IS A PHISHING WEBSITE.')

          [-1]
          THIS WEBSITE IS A PHISHING WEBSITE.
          ******************************
          [1]
          THIS WEBSITE IS A LEGITIMATE WEBSITE.
```

11.0 Concept Development

The concept can be developed by using The appropriate API (flask in this case) and upload it on Github and for its deployment. The cloud services has to be choosen accordingly to the need.



12.0 Final Report Prototype

The product takes the following functions to perfect and provide a good result.

Back-end

Model Development: This must be done before releasing the service. A lot of manual supervised machine learning must be performed to optimize the automated tasks.

1. Performing EDA to realize the dependent and independent features.
2. Algorithm training and optimization must be done to minimize overfitting of the model and hyperparameter tuning.

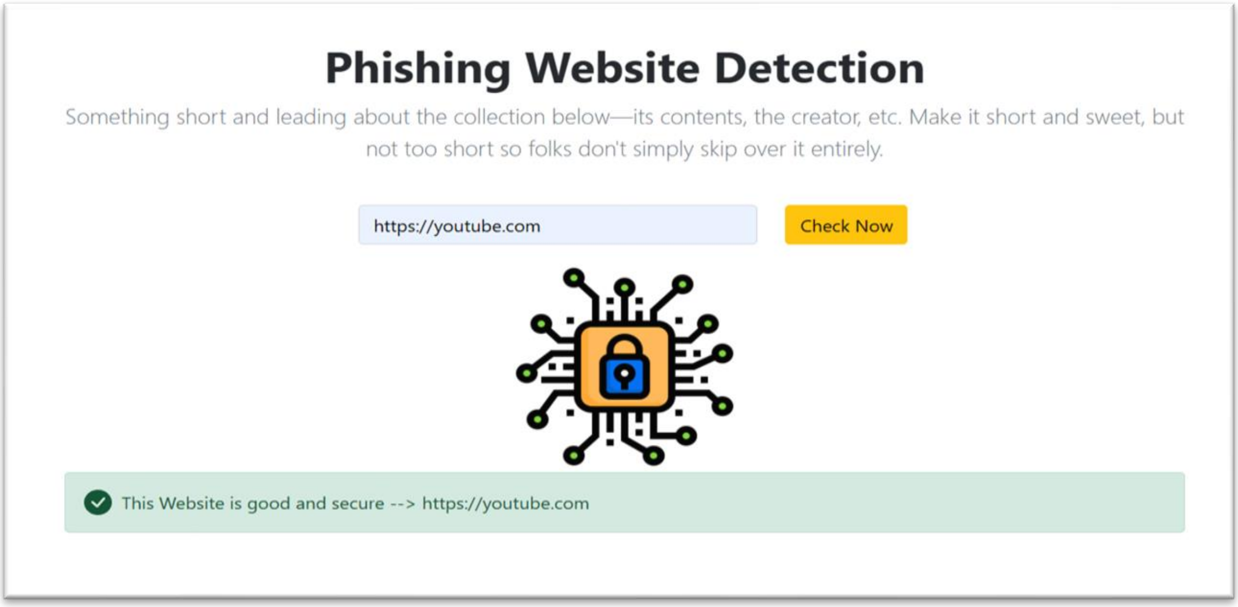
Front End

1. Different user interface: The user must be given many options to choose form in terms of parameters. This can only be optimized after a lot of testing and analysis all the edge cases.
2. Interactive visualization the data extracted from the trained models will return raw and inscrutable data. This must be present in an aesthetic and an "easy to read" style.
3. Feedback system: A valuable feedback system must be developed to understand the customer's needs that have not been met. This will help us train the models constantly.

13.0 Product Details

I created a **web app** using the **python, html** and use SVM model for training data and deployed this web app on the **render cloud platform.**

First, I have checked for legitimate website.



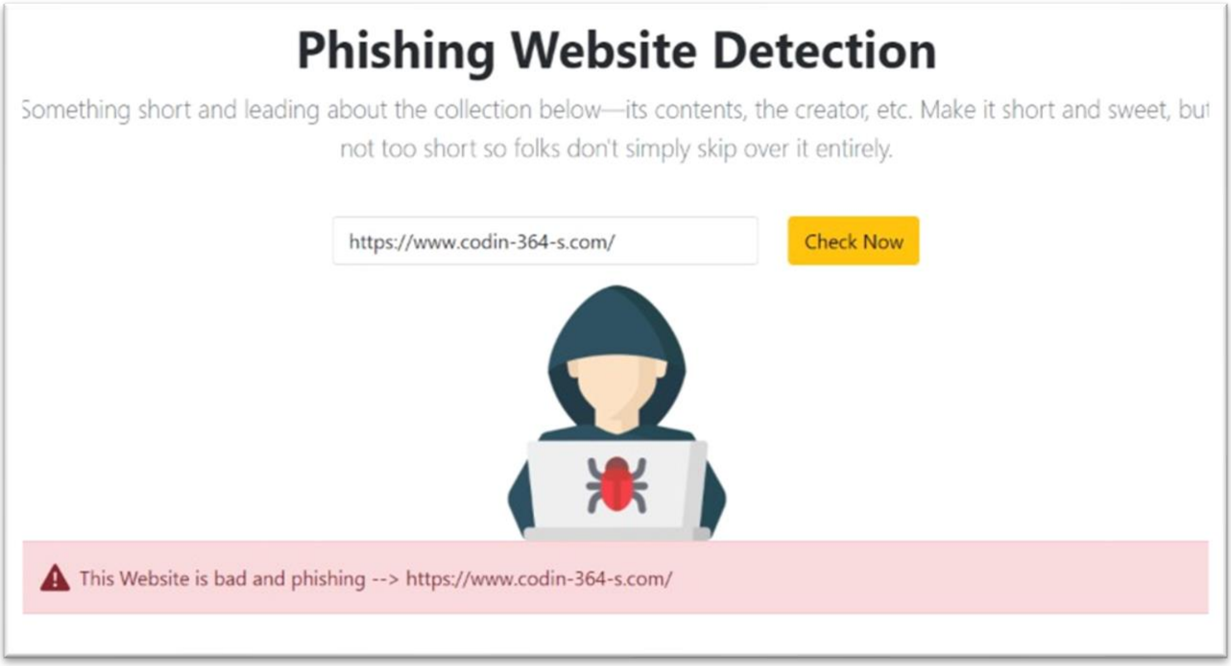
Phishing Website Detection

Something short and leading about the collection below—its contents, the creator, etc. Make it short and sweet, but not too short so folks don't simply skip over it entirely.

https://youtube.com

✓ This Website is good and secure --> https://youtube.com

Then I have checked for phishing websites.



Phishing Website Detection

Something short and leading about the collection below—its contents, the creator, etc. Make it short and sweet, but not too short so folks don't simply skip over it entirely.

https://www.codin-364-s.com/

⚠ This Website is bad and phishing --> https://www.codin-364-s.com/

14.0 Code Implementation

This is a github link :-
https://github.com/Souhardyagayen99/Detection-of-Phishing-Websites-using-Machine-Learning-Approach---ML-Project---Python

15.0 Conclusion

In conclusion, detecting and preventing phishing attacks is crucial for maintaining online security. There are various strategies and tools available to detect phishing websites, including technical solutions, user education, and targeted detection measures. While there are challenges and constraints to effective detection, investing in anti-phishing solutions is a wise decision for organizations and individuals looking to protect their sensitive information and prevent financial loss. By staying vigilant and taking proactive measures, we can work to stay safe and secure in an increasingly digital world.

16.0 References

- Published by Vaibhav Patil, Pritesh Thakkar Prof. S. P. Godse, Tushar Bhat and Chirag Shah of “Detection and Prevention of Phishing Websites using Machine Learning Approach”, Dept. of Computer Engineering in Sinhgad Academy of Engineering Pune, India 2018.
- Ankit Kumar Jain and B. B. Gupta, “Phishing Detection Analysis of Visual Similarity Based Approaches”, Hindawi 2017.

- Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang, “Bait Alarm Detecting Phishing Sites Using Similarity in Fundamental Visual Features”, INCS 2013.
- Y. Zhang, J. I. Hong, and L. F. Cranor. Cantina: a content-based approach to detecting phishing web sites. In WWW ’07: Proceedings of the 16th international conference on World Wide Web, pages 639–648, New York, NY, USA, 2007. ACM.
- Haijun Zhang, Gang Liu, Tommy W. S. Chow, and Wenyin Liu, “Textual and Visual Content-Based Anti-Phishing A Bayesian Approach”, IEEE 2011.