Phishing Website Detection

Sabikun Nahar	**170204093**
Md. Latifur Rahman	**170204102**
Md. Abir Hossain	**170204106**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Spring 2021

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

Phishing Website Detection

Submitted by

Sabikun Nahar	**170204093**
Md. Latifur Rahman	**170204102**
Md. Abir Hossain	**170204106**

Submitted To

Faisal Muhammad Shah, Associate Professor

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

ABSTRACT

This paper explores machine learning techniques and evaluates their performances when trained to perform against datasets consisting of features that can differentiate between a Phishing Website and a safe one. This capability of telling these sites apart from one another is vital in the modern-day internet surfing. As more and more of our resources shift online, one vulnerability and a leak of sensitive information by someone could bring everything down in a connected network.This paper's objective through this research is to highlight the best technique for identifying one of the most commonly occurring cyberattacks and thus allow faster identification and blacklisting of such sites, therefore leading to a safer and more secure web surfing experience for everyone. To achieve this, we describe each of the techniques we look into in great detail and use different evaluation techniques to portray their performance visually. After pitting all of these techniques against each other, we have concluded with an explanation in this paper that Random Forest Classifier does indeed work best for Phishing Website Detection.

Contents

ABSTRACT i

List of Figures iv

List of Tables v

1 Introduction 1

2 Literature Reviews 2

2.1 Overview . . . 2

2.2 Phishing Attacks . . . 2

2.3 Machine learning . . . 3

3 Data Collection & Processing 4

3.1 Structure and Component of a URL . . . 4

3.2 Feature Selection . . . 4

3.3 Ensemble Learning . . . 5

3.4 Model . . . 5

3.4.1 Logistic Regression . . . 5

3.4.2 K-Neighbors Classifier . . . 5

3.4.3 Bernoulli Naïve Bayes . . . 6

3.4.4 Support Vector Machine . . . 6

3.4.5 Random Forest Classifier . . . 6

4 Methodology 7

4.1 Overview . . . 7

4.2 Framework . . . 7

4.3 Implementation Procedure . . . 7

4.3.1 Dataset Preprocessing and Attributes . . . 7

4.3.2 Model Implementation . . . 11

5 Experiments and Results 12

5.1 Model Analysis . . . 12

5.1.1 K-Nearest Neighbor . . . 12

5.1.2 Bernoulli Naive Bayes 12
5.1.3 Random Forest Classifier 13
5.1.4 Random Forest Classifier (Tuned) 13
5.1.5 Support Vector Machine (Tuned) 13
5.2 Accuracy Comparison 14

6 Future Work and Conclusion **15**
6.1 Conclusion: 15
6.2 Future work 15

References **16**

List of Figures

4.1 Correlation with Result 10

List of Tables

4.1 Features and Rules to create the Dataset 9

5.1 Confusion Matrix for KNN 12
5.2 Confusion Matrix for Bernoulli Naive Bayes 12
5.3 Confusion Matrix for Random Forest Classifier 13
5.4 Confusion Matrix for Random Forest Classifier 13
5.5 Confusion Matrix for Support Vector Machine 13
5.6 Accuracy Comparison table 14

Chapter 1

Introduction

In this technological era, the Internet has made its way to become an inevitable part of our lives. It leads to many convenient experiences in our lives regarding communication, entertainment, education, shopping and so on. As we progress into online life, criminals view the Internet as an opportunity to transfer their physical crimes into a virtual environment. The Internet not only provides convenience in various aspects but also has its downsides, for example, the anonymity that the Internet provides to its users. Presently, many types of crimes have been conducted online. Hence, the main focus of our research is phishing. Phishing is a type of cybercrime where the targets are lured or tricked into giving up sensitive information, such as Social Security Number personal identifiable information and passwords. This obtainment of such information is done fraudulently. Given that phishing is a very broad topic, we have decided that this research should specifically focus on phishing websites.

Chapter 2

Literature Reviews

2.1 Overview

In this part cover a similar and related work, detailed description of the scientific articles is presented as a literature review. They used the data mining algorithm for classification.

2.2 Phishing Attacks

Claudio Marforio, Ramya Jayaram Masti, Claudio Soriente, Kari Kostiainen, Srdjan Capkun, in their article [1] they did a comprehensive study on the security vulnerabilities caused by mobile phishing attacks, including the web page phishing attacks, the application phishing attacks, and the account registry phishing attacks. They also discussed that existing schemes designed for web phishing attacks on PCs cannot effectively address the various phishing attacks on mobile platforms. And they proposed a lightweight anti phishing scheme for mobile device.

Neelam Choudhary, Ankit Kumar Jain, in their research [2] they presented a comprehensive analysis of mobile phishing attacks, their exploitation, some of the recent solutions for phishing detection. They classified the mobile phishing attacks in four groups, (i) Smishing - Phishing done by sending SMS messages is known as Smishing. (ii) ncomplete display of URLs - Mobile devices are smaller in size and have a smaller screen due to which it is not possible to see the complete URL when the user visits some website, and it becomes difficult to identify whether he is on the official website or some fake website. (iii) Vishing and Wi-Fi - Vishing is a voicemail phishing to get user's sensitive information. (iv) Availability to the app store - Phishing through installing malicious apps on mobile devices is known as

application phishing.

Diksha Goel, Ankit Kumar Jain in their research paper [3], they discussed and analyzed about mobile phishing attack, then they provided a taxonomy of various recently proposed solutions that can detect and defend mobile phishing attacks.

2.3 Machine learning

Supervised machine learning techniques are applicable in numerous domains. **S. B. Kotsiantis**, in his paper [4] he described various supervised machine learning classification techniques. He also discussed all the pros and cons of each individual algorithms and empirical comparisons of various bias options. He mentioned that SVMs and neural networks tend to perform much better when dealing with multi-dimensions and continuous feature. The paper provides a Table **??** compairing the learning algorithms.

Chapter 3

Data Collection & Processing

3.1 Structure and Component of a URL

A URL [5] is commonly known as the website address. It is composed of many different parts, as illustrated in Fig. 1. Fig. 1. Structure and Components of a URL. In the figure, the area labelled with „1 is the Hypertext Transfer Protocol (HTTP). The HTTP represents the protocol used to fetch resources and contents that are requested. The area labelled with „2 is the hostname. The hostname can be further divided into three parts, namely, subdomain (labelled with „3), domain (labelled with „4) and top-level domain (labelled with „5) which is also known as the web address suffix. The area labelled with „6 shows the path that can be typically referred to as a directory on the webserver. Finally, the area labelled with „7 holds the parameter (v) and value (AbcdEffGhlJ). The symbol „? before the parameter initialises the parameters inside the URL.

3.2 Feature Selection

Feature selection [6] plays a significant role during data analysis. The feature selection method aids in improving the accuracy of the prediction model in such that it reduces the number of features to only those that are critical in influencing the prediction. Specifically, this method helps in cleaning the initial dataset features by retaining only relevant and useful features. Thus, the feature selection [7] algorithm will disregard the features that do not have a high rank in feature importance. However, information loss has no critical effect if the data underwent the feature selection.

3.3 Ensemble Learning

The concept of ensemble learning is an ensemble of algorithms that use more than one learning models. The models [8] used to create an ensemble has its predictions combined to obtain the final prediction. Ensemble methods are useful and have three primary advantages. The application of this method can be used for a statistical reason, which is relevant to the lack of sufficient data used to represent the data distribution. Owing to the lack of such data, the hypotheses that provide a similar training accuracy can be used as one of the learning algorithms for the ensemble. Thus, these methods can help in risk reduction when a wrong model is selected by aggregating the available candidate models. In addition, the ensemble method can be used for computational purposes. Moreover, many learning algorithms, such as decision tree or neural network (NN) that work by executing a local search, are available. These methods will provide optimal solutions from a local perspective. The ensemble method can showcase its advantage in such scenarios because it can run multiple local searches in a parallel manner at different starting points. Finally, it can be used in representation purposes. Although the representation of the actual function cannot be implemented by a single hypothesis, it can be approximated by the combined hypotheses. This concept is similar to signal processing.

3.4 Model

3.4.1 Logistic Regression

the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc.

3.4.2 K-Neighbors Classifier

The K in the name of this classifier represents the k nearest neighbors, where k is an integer value specified by the user. Hence as the name suggests, this classifier implements learning based on the k nearest neighbors. The choice of the value of k is dependent on data.

3.4.3 Bernoulli Naïve Bayes

Bernoulli Naïve Bayes is another useful naïve Bayes model. The assumption in this model is that the features binary (0s and 1s) in nature. An application of Bernoulli Naïve Bayes classification is Text classification with 'bag of words' model.

3.4.4 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

3.4.5 Random Forest Classifier

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Chapter 4

Methodology

4.1 Overview

In this paper we proposed a model to detect phishing websites with Machine Learning algorithms and compared their results.

4.2 Framework

The theoretical framework represents our study about machine learning model and it's implementation on phishing dataset. The framework includes five phases. They are:

1.Find a solution to reduce the problem, like machine learning.
2.Gather information about Machine Learning algorithms in detection phishing.
3.Implement our proposed model.
4.Analyze our proposed system.

4.3 Implementation Procedure

4.3.1 Dataset Preprocessing and Attributes

The dataset is collected form UCI Machine Learning Repository and the name of the dataset is Phishing Websites Data Set [9]. Based on the features described in UCI Machine Learning Repository [9] we can see that the data is clustered into five main groups. They are,

(i) HTML and JavaScript based Features (ii) Domain based Features (iii) Abnormal based Features and (iv) Address Bar based Features. Some rules were applied on the features to determine if a website is phishing or not. For example, if a domains expiration time is equal or less then one year. This site is more likely to be a a phishing site or else it is a legitimate site. Some rules were applied to determine if the website neither legitimate nor phishing, these were categorized as Suspicious. We summarized the features into a table [4.1].
1, 0 and -1 Numerical notations were used to represent Legitimate, Suspicious and Phishing respectively.

1. Domain Based Features	
Age of Domain	6 months → Legitimate Otherwise → Phishing
DNS Record	No DNS record → Phishing Otherwise → Legitimate
Website Traffic	<100,000 → Legitimate >100,000 → Suspicious Otherwise → Phishing
Page Rank	<0.2 → Phishing Otherwise → Legitimate
Google Index	Indexed → Legitimate Otherwise → Phishing
Number of Links Pointing to Page	= 0 → Phishing [1 to 2] → Suspicious Otherwise → Legitimate
Statistical reports	identified as top phishing domain → Phishing Otherwise → Legitimate
2. HTML and JavaScript Based Features	
Web Forwarding	1 → Legitimate [2 to 4]→ Suspicious Otherwise → Phishing
Status bar customization	Mouse Movements affecting status → Phishing Otherwise → Legitimate
Right click disabling	Disabled → Phishing Otherwise → Legitimate
Pop-up windows	Pop windows with text fields → Phishing Otherwise → Legitimate
Iframe redirection	Active IFrame → Phishing Otherwise → Legitimate
3. Abnormal behavior-based features	
Abnormal behavior-based features	<22 percent → Legitimate [22 to 61] percent → Suspicious Otherwise → Phishing
URL anchor percentage	<31 percent → Legitimate Anchor [31 to 67] percent → Suspicious Otherwise → Phishing
Meta, script, and link (%) tags	<17 percent → Legitimate [17 to 81] percent → Suspicious Otherwise → Phishing
Server form handler (SFH)	Empty or Blank → Phishing a different domain → Suspicious Otherwise → Legitimate
Method of mail submission	Use mail() or mailto() → Phishing Otherwise → Legitimate
Abnormal URL	URL minus hostname → Phishing Otherwise → Legitimate
4. Address Bar Based Features	
Use of IP addresses	Domain + IP → Phishing Otherwise → Legitimate
Long URL	Length <54 → Legitimate Length 54 & 75 → Suspicious Otherwise → Phishing
URL shortening	TinyURL → Phishing Otherwise → Legitimate
URL having @ symbol	With @ → Phishing Otherwise → Legitimate
Redirect using // symbol	With // → Phishing Otherwise → Legitimate
- symbol in Domain name	Includes – symbol → Phishing Otherwise → Legitimate
Sub Domain and Multi Sub Domain	Domain part = 1 → Legitimate Domain part = 2 → Suspicious Otherwise → Phishing
HTTPS	Trusted Issuer & Age (1 year) → Legitimate HTTPS but Untrusted Issuer → Suspicious Otherwise → Phishing
Domain registration length	1 year → Phishing Otherwise → Legitimate
Favicon	Load from External Sources → Phishing Otherwise → Legitimate
Non-standard port	Preferred Port Number → Legitimate Otherwise → Phishing
HTTPS Token in Domain	Token present → Phishing Otherwise → Legitimate

Table 4.1: Features and Rules to create the Dataset

We used **Feature Selection** method as Data Prepossessing. First we observed the correlations between the feature columns, then we observed the correlation between feature columns and the Result column [4.1].

From that we learned that three feature columns named 'Popupwidnow', 'Favicon','Iframe'

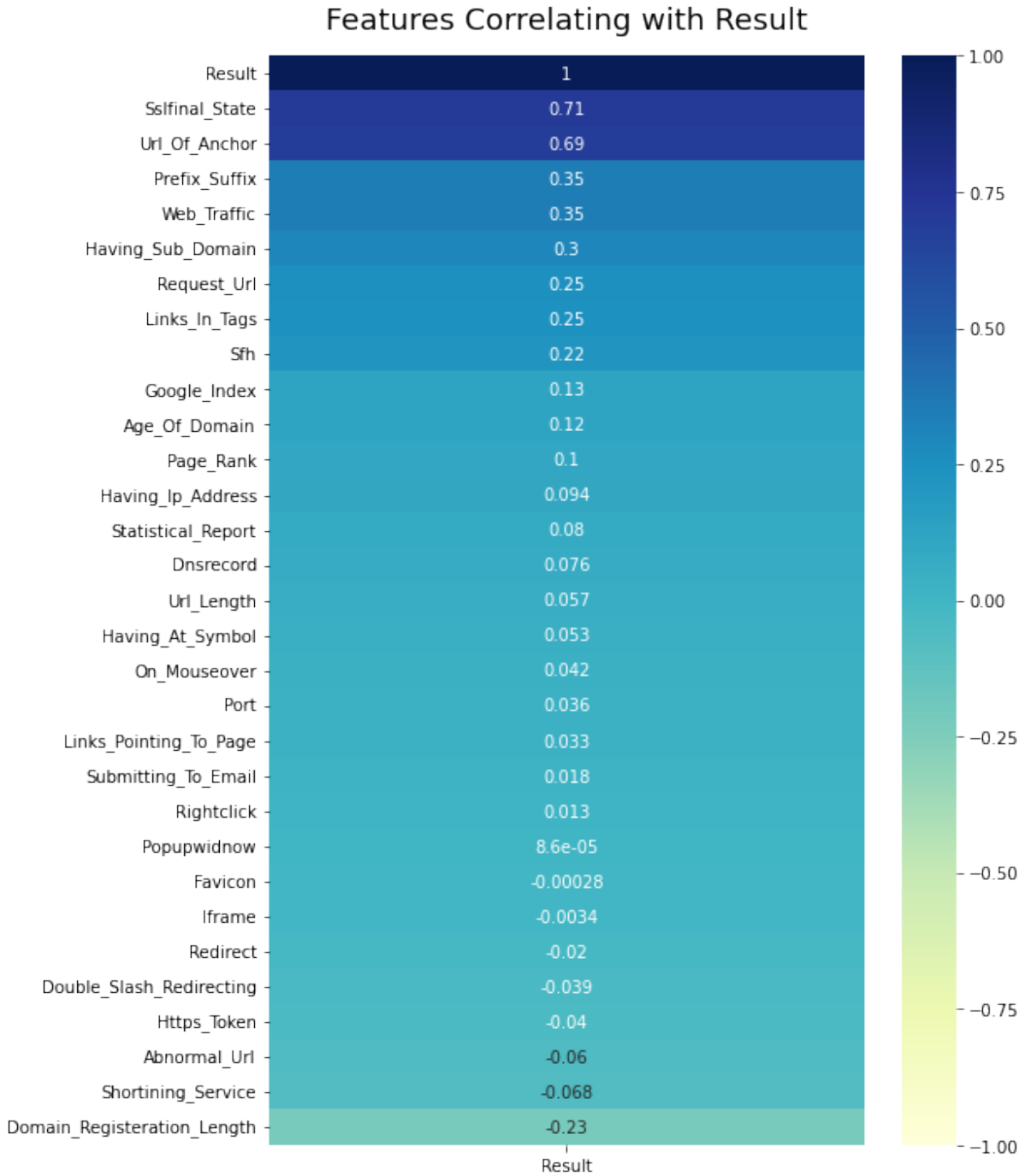

Figure 4.1: Correlation with Result

has very low correlation with the Result column. So we removed these columns from our features.

4.3.2 Model Implementation

We used four machine learning models:

- Logistic Regression
- K - Nearest Neighbors
- Bernoulli Naive Bayes
- Support Vector Machine

And one ensemble model:

- Random Forest

Then we tuned the Hyper-parameter for Random Forest with Random Search CV and Support Vector Machine with Grid Search CV.

Chapter 5

Experiments and Results

5.1 Model Analysis

5.1.1 K-Nearest Neighbor

We implemented K-Nearest Neighbor Algorithm using python's Scikit Learn library. We used neighbor value, k = 3 for the test.

		Predicted	
		Phishing (0)	Legitimate (1)
Actual	Phishing (0)	1370	101
	Legitimate (1)	86	1760

Table 5.1: Confusion Matrix for KNN

5.1.2 Bernoulli Naive Bayes

For Naive Bayes, we implemented Bernoulli Naive Bayes Classifier using python's Scikit Learn library.

		Predicted	
		Phishing (0)	Legitimate (1)
Actual	Phishing (0)	1308	163
	Legitimate (1)	145	1701

Table 5.2: Confusion Matrix for Bernoulli Naive Bayes

5.1.3 Random Forest Classifier

We implemented Random Forest Classifier using python's Scikit Learn library. We used max depth as 10 for the test.

		Predicted	
		Phishing (0)	Legitimate (1)
Actual	Phishing (0)	1357	114
	Legitimate (1)	47	1799

Table 5.3: Confusion Matrix for Random Forest Classifier

5.1.4 Random Forest Classifier (Tuned)

After tuning the Hyper-parameter for Random Forest Classifier we get the following result.

		Predicted	
		Phishing (0)	Legitimate (1)
Actual	Phishing (0)	1401	70
	Legitimate (1)	37	1809

Table 5.4: Confusion Matrix for Random Forest Classifier

5.1.5 Support Vector Machine (Tuned)

After tuning the Hyper-parameter for Support Vector Machine we get the following result.

		Predicted	
		Phishing (0)	Legitimate (1)
Actual	Phishing (0)	1376	95
	Legitimate (1)	42	1804

Table 5.5: Confusion Matrix for Support Vector Machine

5.2 Accuracy Comparison

From the Accuracy Table [5.6] we can see that the tuned Random Forest ensemble model gives the best performance in terms of accuracy. Therefore this model is the best model in detecting phishing website for the given dataset.

Model	Accuracy
Logistic Regression	92.7%
K Nearest Neighbor	94.3%
Bernoulli Naive Bayes	90.7%
Random Forest	95.1%
Random Forest(tuned)	96.7%
Support Vector Machine(tuned)	95.8%

Table 5.6: Accuracy Comparison table

Chapter 6

Future Work and Conclusion

6.1 Conclusion:

Certain classifiers that are more prone to overfitting than others are present, thus yielding higher accuracy rate if they are based on the dataset that they have been trained upon. To address the overfitting problem while focusing on increasing the prediction accuracy, the proposed solution model uses feature selection and ensemble learning where multiple learning models are combined to produce a prediction. By using multiple models, the prediction is not bias towards one model and is instead based on majority of predictions such that all predictions from each model influences the final ensemble prediction.

6.2 Future work

The phishing attacks are increasing day by day based on the literature review, though ample solutions are available. However, it is a bit challenge to educatethe users besides of detecting phishing attacks.

References

[1] L. Wu, X. Du, and J. Wu, "Effective defense schemes for phishing attacks on mobile computing platforms," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 6678–6691, 2016.

[2] N. Choudhary and A. K. Jain, "Comparative analysis of mobile phishing detection and prevention approaches," in *International Conference on Information and Communication Technology for Intelligent Systems*, pp. 349–356, Springer, 2017.

[3] D. Goel and A. K. Jain, "Mobile phishing attacks and defence mechanisms: State of art and open research challenges," *Computers & Security*, vol. 73, pp. 519–544, 2018.

[4] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," (NLD), p. 3–24, IOS Press, 2007.

[5] A. A. Ubing, S. K. B. Jasmi, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Phishing website detection: an improved accuracy through feature selection and ensemble learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, pp. 252–257, 2019.

[6] P. Sharma, "The ultimate guide to 12 dimensionality reduction techniques (with python codes)," *Analytics Vidhya*, p. 27, 2018.

[7] L. Rokach, "Ensemble-based classifiers," *Artificial intelligence review*, vol. 33, no. 1, pp. 1–39, 2010.

[8] R. Polikar, "Ensemble learning in ensemble machine learning: Methods and applications; zhang, c., ma, y., eds," 2012.

[9] D. Dua and C. Graff, "UCI machine learning repository," 2017.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Monday 9th May, 2022 at 11:01am.