Information Security

Final Project Report

PHISHING DETECTION USING MACHINE LEARNING

Submitted by: Syed Ahsan Akhtar, Mustafa Bawany, Sara Sameer

19K-1475 , 19K-1273 , 19K-1255

BS-CS

G

Submitted to: Sir Nadeem Kafi

Department of Computer Sciences

FAST-NUCES

Karachi Campus

Submitted on: 2th December 2022

FAST School of Management Sciences

NATIONAL UNIVERSITY OF COMPUTERS AND EMERGING SCIENCES

KARACHI CAMPUS

Introduction / What is the problem we are going to address?

Phishing is a form of social engineering attack that is frequently employed to steal user information, such as login credentials and credit card numbers. It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email, instant message, or text message. The recipient is then tricked into clicking a malicious link, which can lead to the installation of malware, the freezing of the system as part of a ransomware attack, or the revealing of sensitive information.

An attack can have devastating results. For individuals, this includes unauthorized purchases, the stealing of funds, or identity theft.

Considering the devastating effects of phishing attacks, we have designed an algorithm and worked on the Phishing factors dataset to determine the top features or factors that play an important role in phishing attacks.

Our solution uses two machine learning algorithms, namely linear logical regression, and the Random Forest Classifier.

Why is this Problem Important?

Today, everyone is highly dependent on the internet to perform routine activities like shopping, banking, booking, etc. These luxuries came with the cost of security threats. Across the web, unauthorized users have baited unsuspecting victims into handing over bank info, social security numbers, and more. Plus, cybercriminals have become even savvier with their disguises. These all attacks come under the umbrella of 'Phishing Attacks'. Thus, it is extremely important to address this problem, find the underlying factors that contribute to these attacks and equip the users with the knowledge to protect themselves from such attacks.

Some common phishing attacks across the web are as follows:

1. Sending spam emails to high-privilege account holders and tricking them into revealing their personal /organizational information.
2. Users tricked into clicking a link or opening an attachment might download malware onto their devices.
3. an attacker who can inject malicious content into an official site will trick users into accessing the site to show them a malicious popup or redirect them to a phishing website.

How is this problem currently addressed by others?

Before starting our project, we evaluated a few research papers and found out their way of working. According to **Detecting phishing websites using a machine learning technique**, the paper was published on Oct 11, 2021. The author **Ashit Kumar Dutta** employed frameworks of RNN—LSTM to identify the properties Pm and Pl in an order to declare an URL as malicious or legitimate.

The modified version of RNN is LSTM. It is a deep learning method, which prevents the gradient problem of RNN. The LSTM model is an effective predictive model. It generates an output based on an arbitrary number of steps. There are five essential components that enable the model to produce long—term and short—term data.

The commonality between our approach and this approach was the usage of similar evaluation metrics which were Accuracy score, F1 score, recall, and precision.

Another research article that we explored was **Phishing attack detection using Machine Learning** by **S.Pandiyan, S.P.Selvaraj, V . K.Burugari, J.B.P. Kanmani**, published on 11 July 2022 employed frameworks of Decision Tree Classifier, Random Forest Classifier, Multi-Layer Perceptron, XG Boost Classifier, SVM, and Cat Boost Classifier.

What is the way we are proposing to address the problem?

We are working on the dataset provided by Kaggle which has a total of 50 features, contributing to Phishing in general. We have utilized two models, a Random forest classifier and logistic regression to find the optimal number of features that contribute to phishing, hence we will be getting the top listed factors related to phishing.

We have the following evaluation metric:

- Accuracy Score
- F1 Score
- Support
- Precision
- Recall

What is your plan for completing the project and submitting it on time?

Our project was partitioned into two parts: Research and Implementation. For the research phase, we explored the vulnerabilities of organizations in the domain of 'Information Security ' and found out that Phishing Attack is the most alarming cybercrime and can devastate the organization if overlooked.

The next step was to find an appropriate dataset for this problem. For this project, we have used the 'Phishing Dataset for Machine Learning' available on Kaggle. This dataset contains 48 features extracted from 5000 phishing webpages and 5000 legitimate webpages, which were downloaded from January to May 2015 and from May to June 2017.

After performing the exploratory data analysis on the dataset, we started to explore the models that would yield best results.

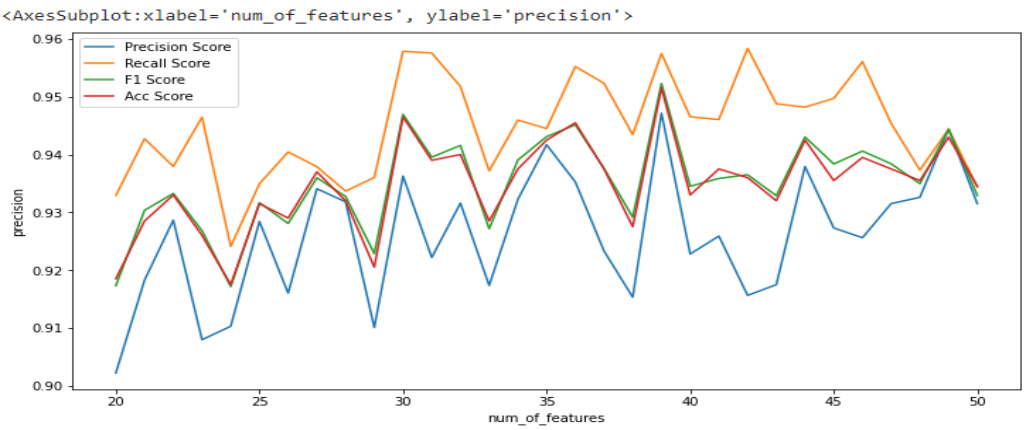
We set a 2 week deadline for this project, and distributed the tasks based on our expertises.

Methodology and Results:

We employed frameworks of Logistic Regression and Random Forest Classifier on the dataset provided by Kaggle of 50 features contributing to Phishing.

Our goal was to select an optimal number of features on which maximum accuracy,f1 score, recall, and precision would be produced.

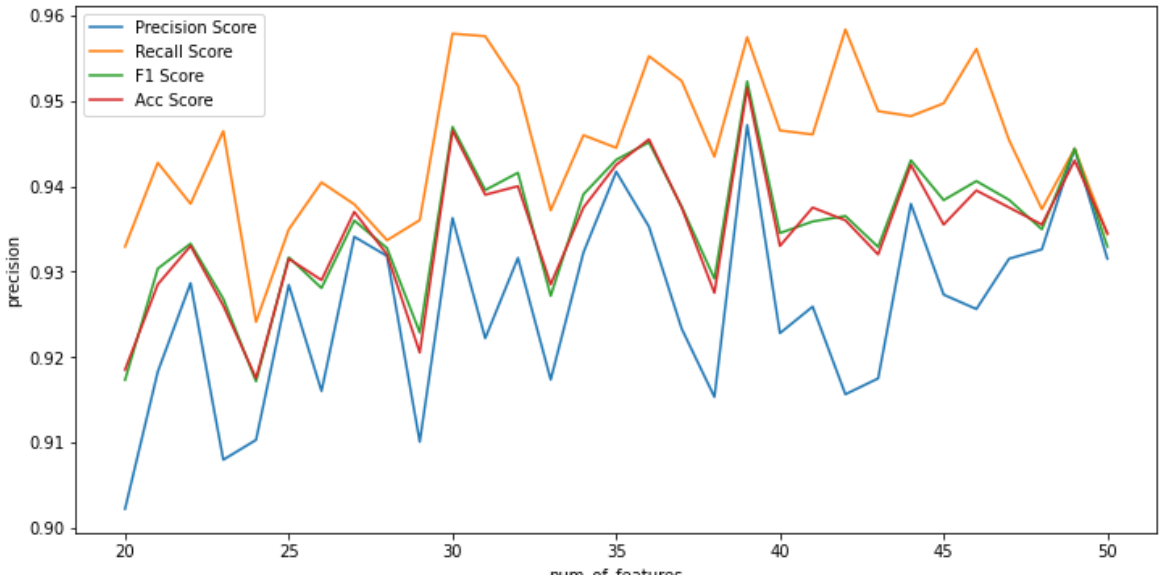
We considered the results of Logistic Regression as our baseline model results, and then we employed the power the of Random Forest Classifier to beat the results of Logistic regression.



We got the following evaluation metrics on a respective number of features. It can be seen that the optimal number of features in logistic regression is 39, and the Precision, Recall, F1_Score, and Accuracy of 39 features is 0.947162,0.957468, 0.952287,0.9515 respectively.

Now we trained the Random Forest Classifier and aimed to beat the results of our baseline model (Logistic Regression).

The Random Forest Classifier produced the following results:



So by visualizing the figure below, we can conclude that the best number of features for this model would be 32, one less than logistic regression, the reason why I chose 32 is because the number of features allowed the model to perform the best across all the evaluation metrics.

The model is now capable of predicting at up to 98% accuracy and also precision and recall, this shows the model has high confidence in predicting phishing or non-phishing site

The final results that we achieved from our model were:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	999
1	0.98	0.98	0.98	1001
accuracy			0.98	2000
macro avg	0.98	0.98	0.98	2000
weighted avg	0.98	0.98	0.98	2000