Concordia University

Engineering and Computer Science

INSE-6180 Security and Privacy Implications of Data Mining

PROJECT REPORT: PREDICTING PHISHING WEBSITES

SUBMITTED BY:
Ayush Arya (40084127)
Gurvinder Singh Bhai Ka(40070767)
Munish Sehdev (40083946)
Pulkit Wadhwa (40082832)

SUBMITTED TO:
Nizar Bouguila

ABSTRACT

Phishing as the word suggests is the way to catch a fish by providing a bait. It is a crime in the world of Internet. It is a way to capture private information of user by tricking someone and posing as a legitimate entity. It is very difficult to capture a complete phishing attack by a particular solution due to the involvement of many factors and criteria. Different criteria for Identifying Phishing attack are URL and Domain Identity, Security & Encryption and Web Address bar. Data Mining techniques are very helpful methods in discovering and spotting phishing websites. In this project, we will be applying different Machine Learning algorithm to Phishing problem. These algorithms will help us to identify all the facts and rule in order to classify whether a website is phishing website or not based on the criteria given above like for example like phishing websites use long URLs to hide suspicious part and use IP address, tiny URLs, http instead of https and various other methods so we can use them as features in our data set to classify the websites.

Table of Contents

1. INTRODUCTION

1.1 INTRODUCTION:

Phishing websites aims at targeting end users. It is a cyber-attack which tries to get the credentials of users such as user name, password and bank details etc. in this type of attack attackers try to create fake web pages by copying the original webpages. The thrown bait in most instances is either an email or messaging site and these redirect users to phishing websites. The attackers use n number of ways to make the web link lookalike the real one which tricks people and prompts them to open that link and the personal identities of people is being sacrificed. The attackers also steal sensitive information such as Social Insurance Number and another information like credit cards, etc. Main source of phishing is email spooking or instant messaging in which those fake weblinks are included which make it look like a legitimate link. The attackers make use of the weak current security protocols to do phishing attacks. Thus, it becomes very important to protect users from these phishing websites and attacks.

1.2 MOTIVATION:

The main motivation behind this is to apply data mining algorithms to detect and classify various phishing websites such as e-banking phishing websites, great benefits can be provided to people who are working in the field of information security, these benefits include classification, prediction and decision support system against these websites. Classification is a process of classifying data into one of the given classes, it is being used in various fields these days and play an imperative role in information security.

Classification is the process of knowledge discovery in database which can be defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.

DM is a step in the ***Knowledge Discovery*** Database process that consists of the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. DM Techniques can be categorized into two groups: descriptive and predictive. These techniques can be achieved by using various methods based on DM functions, used to specify the types of patterns to be mined. These functions include characterization, clustering, association, classification.

Data mining classification technology contains two parts: construction of classification model, and evaluation of model classification efficiency.

In the first part, the adopted classification algorithm is trained by a classified training data set in order to build classification predictive model. In the second part, testing data set is used to test classification efficiency of this model. We used the following algorithms: Decision Tree (DT), Support Vector Machine (SVM) and Naïve Bayesian Classifier (NBC).

1.3 OBJECTIVES

Our objective is to implement the algorithms such as Naive bayes,Decision tree, Random forest and logistic regression to predict whether a website is phishing website or not. We first implement these algorithms by using the sklearn libraries and we also implement on our own to calculate the accuracy of these algorithms.

1.4 BASIC ASSUMPTIONS/LIMITATIONS

- Unavailability of reliable training set that has been published publicly.
- There are no definitive features that Characterize phishing in web pages.
- Features that proved to be effective in prediction have been included

2. BACKGROUND AND LITERATURE REVIEW

Widespread use of internet and development of information technology has led to various kinds of cyber threats so it becomes very important to protect computers and people from these threats. There have been various studies on phishing detection and it is observed that various data mining techniques can be a handful in phishing detection.

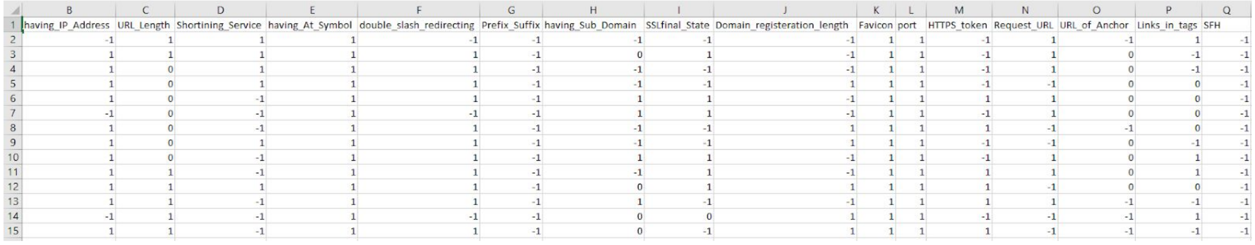
According to Wombat Security State of the Phish, 76% of businesses reported being a victim of a phishing attack in the last year. And "nearly 1.5 million new phishing sites are created each month" so it becomes very important to detect and prevent phishing attacks so we use data mining techniques for efficient and accurate prediction of these websites.

So various studies were conducted so that people could be protected from these attacks which required changing of web infrastructure and also other methods which included certain rules for web page creators to follow but it was hard to enforce these rules so it lead to other studies which were conducted with the help of various classification algorithms such as decision trees, neural networks, naive bayes classifiers, support vector machines and many others for phishing detection ,these are also used in various other areas such as for healthcare, market basket analysis, education, manufacturing engineering, fraud detection etc.

There were also various experiments conducted like phone and website phishing experiments by a group of people they conducted various studies and also used various data mining algorithms Kaytan and Hanbay[5]proposed determining phishing websites based on neural network their accuracy was measured up to 92.45% and also Li et al.[6]proposed a novel approach based support vector machine (SVM) to detect phishing websites

Dataset

Dataset is available on UCI machine learning Repository. A brief description of features is given below features included binary values and categorical values

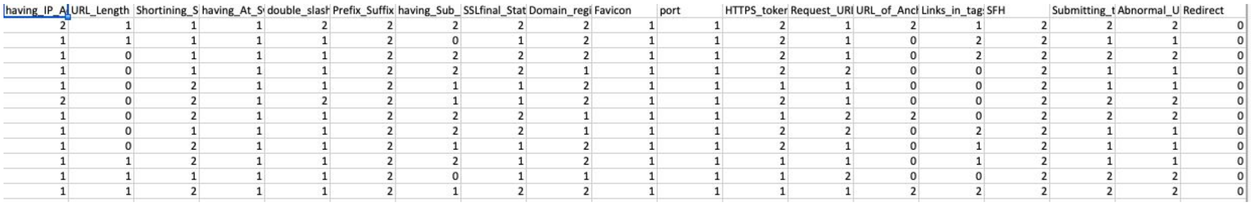


	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	having_IP_Address	URL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain	SSLfinal_State	Domain_registeration_length	Favicon	port	HTTPS_token	Request_URL	URL_of_Anchor	Links_in_tags	SFH
2	-1	1	1	1	-1	-1	-1	-1	-1	1	1	-1	1	-1	1	-1
3	1	1	1	1	1	-1	0	1	-1	1	1	-1	1	0	-1	-1
4	1	0	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	-1	-1
5	1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	0	-1
6	1	0	-1	1	1	-1	1	1	-1	1	1	1	1	0	0	-1
7	-1	0	-1	1	-1	-1	1	1	-1	1	1	-1	1	0	0	-1
8	1	0	-1	1	1	-1	-1	-1	1	1	1	1	-1	-1	0	-1
9	1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	-1	-1
10	1	0	-1	1	1	-1	1	1	-1	1	1	-1	1	0	1	-1
11	1	1	-1	1	1	-1	-1	1	-1	1	1	1	1	0	1	-1
12	1	1	1	1	1	-1	0	1	1	1	1	1	-1	0	0	-1
13	1	1	-1	1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1
14	-1	1	-1	1	-1	-1	0	0	1	1	1	-1	-1	-1	1	-1
15	1	1	-1	1	1	-1	0	-1	1	1	1	1	-1	-1	-1	-1

(screenshot of the original DataSet)

- **Address Bar Based Features**:Using IP address,Long URL,Using URL shortening services,subdomain and multi domains,HTTPS(12 features)
- **Abnormal Based Features**:Abnormal URL,submitting information to Email,Server form handler(6 features)
- **HTML and JavaScript based features**:Using pop up window,website forwarding,status bar customisation ,disabling right click(5)
- **Domain Based Features**:Age of domain,DNS record,website traffic(7 features).

However to evaluate our results we also compared our results with sklearn so we had to transform our dataset to eliminate negative values and transform it so it can be used with sklearn also.



having_IP_A	URL_Length	Shortining_S	having_At_S	double_slash	Prefix_Suffix	having_Sub_	SSLfinal_Stat	Domain_regi	Favicon	port	HTTPS_toker	Request_URl	URL_of_Ancl	Links_in_tag:	SFH	Submitting_t	Abnormal_U	Redirect
2	1	1	1	2	2	2	2	2	1	1	2	1	2	1	2	2	2	0
1	1	1	1	1	2	0	1	2	1	1	2	1	0	2	2	1	1	0
1	0	1	1	1	2	2	2	2	1	1	2	1	0	2	2	2	2	0
1	0	1	1	1	2	2	2	1	1	1	2	2	0	0	2	1	1	0
1	0	2	1	1	2	1	1	2	1	1	1	1	0	0	2	1	1	0
2	0	2	1	2	2	1	1	2	1	1	2	1	0	0	2	2	2	0
1	0	2	1	1	2	2	2	1	1	1	1	2	2	0	2	2	2	0
1	0	1	1	1	2	2	2	1	1	1	2	2	0	2	2	1	1	0
1	0	2	1	1	2	1	1	2	1	1	2	1	0	1	2	1	1	0
1	1	2	1	1	2	2	1	2	1	1	1	1	0	1	2	1	1	0
1	1	1	1	1	2	0	1	1	1	1	1	2	0	0	2	2	2	0
1	1	2	1	1	2	1	2	2	1	1	1	1	2	2	2	2	2	0

(screenshot of the updated DataSet)

3. PROPOSED MODEL FRAMEWORK

Dataset was preprocessed and used for analysis, we trained and tested our classifiers with the given four algorithms, dataset was split into two sets training and test set where training was used to train our classifier and test set was used to test our classifier, we used split ratio of 70:30 to split our Dataset into train and test set respectively, a brief description about the algorithms that were used for classification are:

A. Decision Tree

Decision tree is the classification technique which divides the observations into branches to build a tree structure to enhance the accuracy of prediction. Tree is constructed by using top-down divide and conquer approach which is used in a recursive manner. The root item is selected based on the factor called Information gain i.e. the test attribute is selected based on the statistical/heuristic measure. The attribute with the highest information gain is selected as a root and is then partitioned further recursively.

This method has the following conditions:

1. All the samples for a given node should belong to the same class

2.There is no attribute left for further partitioning

3.There are no remaining samples

The most commonly used algorithms for splitting attributes are Entropy based i.e information gain (used in ID3, C4.S, CS). Entropy is calculated with the given formula:

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

B. **Random Forest**

Random forest is basically a combination of decision trees. Instead of simply averaging the prediction of trees, this model is based on the following factors because of which we call it a random forest. The factors are:

1. Random sampling of training data points while building trees

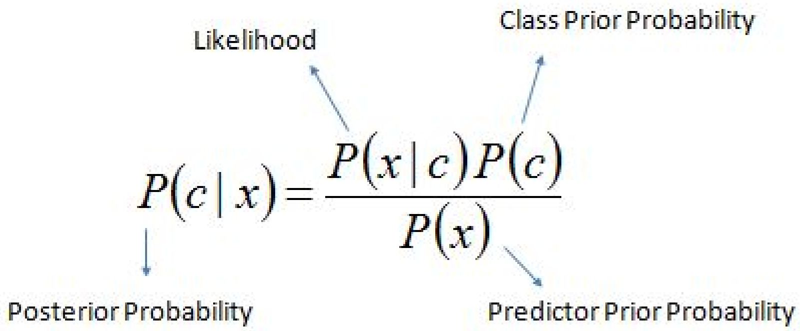
2. Random subsets of features considered while splitting nodes

When we train each tree in random forest, each tree learns from a random sample of data points. In a similar fashion, the other trees are trained and at the end the average of predictions is calculated to get the averaged prediction, This process is also termed as bagging.

Also, while splitting the individual nodes in each decision tree, we only consider the subset of features and again at the end, the average is calculated to get the prediction.

C. Naïve Bayesian Classifier

Naive Bayes classifier is the probabilistic machine learning model which is used for classification tasks [4]. The core part of this method is based on Bayes theorem. Bayes theorem states that we can find the probability of an event given that another event has already occurred i.e. if we have to calculate P(c|x) which means we have to find the probability of event c given event x. Here x is the evidence and c is the hypothesis. It is assumed that the predictors/features are independent from each other. The existence of a particular feature does not affect another feature due to which it is known as naive technique. The formula for calculating the probability using naive bayes method is as follows:


Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

D. **Logistic Regression**

Logistic Regression is a Machine Learning algorithm which is used for classification problems, it is a predictive analysis algorithm and based on the concept of probability. A Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

(Hypothesis limits cost function between 0 and)

$$f(x) = \frac{1}{1+e^{-(x)}}$$

(Sigmoid Function)

Cost function J represents optimization objective i.e.to minimize it so that we can develop an accurate model with minimum error.

$$J(\theta) = -\frac{1}{m}\sum\left[y^{(i)}\log(h\theta(x(i))) + \left(1-y^{(i)}\right)\log(1-h\theta(x(i)))\right]$$

(Cost Function)

To minimize the cost function gradient descent is run on each parameter such that:

$$\theta j := \theta j - \alpha\frac{\partial}{\partial\theta j}J(\theta)$$

4. EXPERIMENTS AND RESULTS

Performance measure: The performance is measured on the basis of accuracy which is calculated using confusion matrix which records output in terms of correctly (True positives, True negatives) and incorrectly classified instances (false positives and false negatives)

Predicted Values/Actual Values	1	0
1	True Positive	False Positive
0	False negative	True Negative

Precision can be defined as of all websites we predicted phishing websites what websites are actually phishing websites where as recall can be defined as of all websites in our data set which are actually phishing websites what fraction did me correctly detect as phishing websites and accuracy it is simply a ratio of correctly predicted observation to the total observations.

Accuracy = True positive + True Negative / (True Positive + False positive + True Negative + False negative)

Precision=True positive / (True Positive + False positive)

Recall = True Positive / (True positive + False negative)

4.1 EXPERIMENTAL FRAMEWORK

4.2 RESULTS

We implemented the algorithms on our own and also to evaluate our results we used scikit learn as well as weka also to evaluate our results and also to compare the accuracy of the results

Algorithm	Our Implementation	Using Sklearn
Decision Tree	**87.25%**	**96%**
Naive Bayes Classifier	**77.20%**	**78%**
Logistic Regression	**83.49%**	**90%**
Random Forest	**89.15%**	**95%**

4.3 CONTRIBUTIONS AND DISCUSSION

Dataset was preprocessed and transformed by Pulkit, we took 4 algorithms to be implemented on DataSet,as we had 4 members so each individual implemented one algorithm leading to mutual contribution

Work Done	Team member	Contributions
Decision Tree	**Ayush Arya**	**25%**
Naive Bayes Classifier	**Pulkit Wadhwa**	**25%**
Logistic Regression	**Munish Sehdev**	**25%**
Random Forest	**Gurvinder Singh Bhai Ka**	**25%**

5. CONCLUSION

On the basis of accuracy, we evaluated the results of different algorithms on our data set and we out certain patterns i.e. knowledge discovery in terms of what kind of websites are phishing websites on the basis of URL, page style security etc. and thus it would be easy to detect these websites and protect people from these websites in the real world.

Almost 80% of people in the world suffer from these attacks if we could recognise these websites prior it will be a huge success to counter these websites,people in various fields including information security and banking sector could be provided with a huge boost.

References

1. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5501434
2. https://dergipark.org.tr/en/download/article-file/333655
3. https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5
4. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
5. M. Kaytan and D. Hanbay, "The Determining with Artificial Neural Network Based Intelligent System Against The Attacks to The Internet Sites by Phishing Method", International Conference on Natural Science and Engineering, ICNASE'16, pp.3221-3226, 2016, Kilis 7 Aralık University, Kilis.
6. Y. Li, L. Yang and J. Ding, "A minimum enclosing ball-based support vector machine approach for detection of phishing websites",Optik, 127(1), pp.345-351, 2016.
7. https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76