SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING 15CS496L MAJOR PROJECT

AN EFFECTIVE MODEL OF TERMINATING PHISHING WEBSITES AND DETECTION BASED ON LOGISTIC REGRESSION

Student 1 Reg No:RA1611003010184
Student 2 Reg No:RA1611003010231
Batch ID:CSE02310184
Guide:Mrs.K.R.JANSI (Assitant Professor)

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be **University** *u/s 3 of UGC Act, 1956*

Table of contents

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Abstract

These days phishing is one of the greatest and the quickest developing risk in light of the fact that phishing assailants will get data which was entered by the client and phishing aggressors will utilize that data without the client information. At present personal information is the most important thing compared to money, So phishing website attackers are focusing on the personal information of the user and they taking advantage when the user enters there personal information in the phishing websites. So the main theme of these projects is to avoid misuse of personal information by phishing attackers. To overcome this problem machine learning algorithm (Logistic regression) is used and provided with the massive dataset. From that dataset, it trains the algorithm and helps in detecting the new web links which are fraudulent links. Attackers disguise their website as legitimate and try to get data from the user for which they make users visit a website and get the personal information that is needed. Before trying to enter into those websites it is important to check whether the given link is good or a phishing link. By checking the link we can save ourselves from the attackers and can keep our data safe

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Introduction

- The Internet has become an imperative foundation that carries extraordinary comfort to human culture.
- The Internet is additionally described by some unavoidable security issues, for example, phishing, malignant programming, and protection exposure, which have just carried genuine dangers to the economy of clients.
- The APWG describes phishing as a criminal instrument using both social structure and concentrated deception to take singular character data and budgetary record capabilities of buyers.
- Phishing websites have so many ways to get the client's personal information, Among those ways clicking weblink is one of the dangers and popular ways.
- In this phishing attack, the attacker sends an attractive link to the client through the mail or any other way. If you open that link and proceed to fill all the sensitive information then phishing attackers take advantage of the received information from a common man is used to taking money from the account or selling information to some third parties which can be used any were.

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

- Nowadays there is much software that detects fraudulent ones and blocks. But most people ignore the links whether it is good or bad, So to detect link's we are implementing a machine learning technique that finds out whether the link is good or bad.
- In our modern digital life, Phishing attacks are expanding rapidly because phishing attackers are taking advantage of technologies like email, mobiles, and web links.
- The daily clients may receive many links from email or any other way or even the client may click the link without knowledge about the link that type of links may be or bad, So the main idea of the project is to provide awareness to the client about the link by using the machine learning algorithm (logistic regression).

SRM INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

LITERATURE SURVEY

S.NO	TITLE	CONTENT	AUTHOR	YEAR
1	Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection	Phishing is a form of cyber attack that leverages social engineering approaches and other sophisticated techniques to harvest personal information from users of websites. The average annual growth rate of the number of unique phishing websites detected by the Anti Phishing Working Group is 36.29% for the past six years and 97.36% for the past two years. In the wake of this rise, alleviating phishing attack. s has received a growing interest from the cyber security community. Extensive research and development have been conducted to detect phishing attempts based on their unique content, network, and URL characteristics.	Zuochao Dou , Abdallah Khreishah	2013

SRM INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

LITERATURE SURVEY

S.NO	TITLE	CONTENT	AUTHOR	YEAR
2	Intelligent rule-based phishing websites classification	Phishing is described as the art of echoing a website of a creditable firm intending to grab user's private information such as usernames, passwords and social security number. Phishing websites comprise a variety of cues within its content-parts as well as the browser-based security indicators provided along with the website. Several solutions have been proposed to tackle phishing. Nevertheless, there is no single magic bullet that can solve this threat radically.	Fadi Thabtah Lee McCluskey	2014

SRM INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

LITERATURE SURVEY

S.NO	TITLE	CONTENT	AUTHOR	YEAR
3	Neural Markers of Cybersecurity: An fMRI Study of Phishing and Malware Warnings	The security of computer systems often relies upon decisions and actions of end users. In this paper, we set out to investigate users' susceptibility to cybercriminal attacks by concentrating at the most fundamental component governing user behavior-the human brain. We introduce a novel neuroscience-based study methodology to inform the design of user-centered security systems as it relates to cybercrime. In particular, we report on an functional magnetic resonance imaging study measuring users'	Nitesh Saxena Jose Maximo Omar	2016

security performance .

SRM INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

LITERATURE SURVEY

S.NO	TITLE	CONTENT	AUTHOR	YEAR
4	Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity	Social networks have become one of the most popular platforms for users to interact with each other. Given the huge amount of sensitive data available in social network platforms, user privacy protection on social networks has become one of the most urgent research issues. As a traditional information stealing technique, phishing attacks still work in their way to cause a lot of privacy violation incidents. In a Web-based phishing attack, an attacker sets up scam Web pages (pretending to be an important Website such as a social network portal) .	Wenqian Tian Zhenkai Liang	2017

Inference from the survey

- From the survey, we have seen that there are different types of methodologies to find Phishing websites.
- Most of them are to avoid such type of links by blocking them. Many have a different methodology and have different accuracy in it.
- The accuracy of finding those is nearly 95% in all these above methodologies. Even then attackers are finding new ways to fool people and escape these securities.
- By using a different approach and different methodology we can increase the accuracy in our cases.
- No one in the previous papers has used this algorithm called logistic regression by using this we can increase the accuracy and even by including the data set of large links

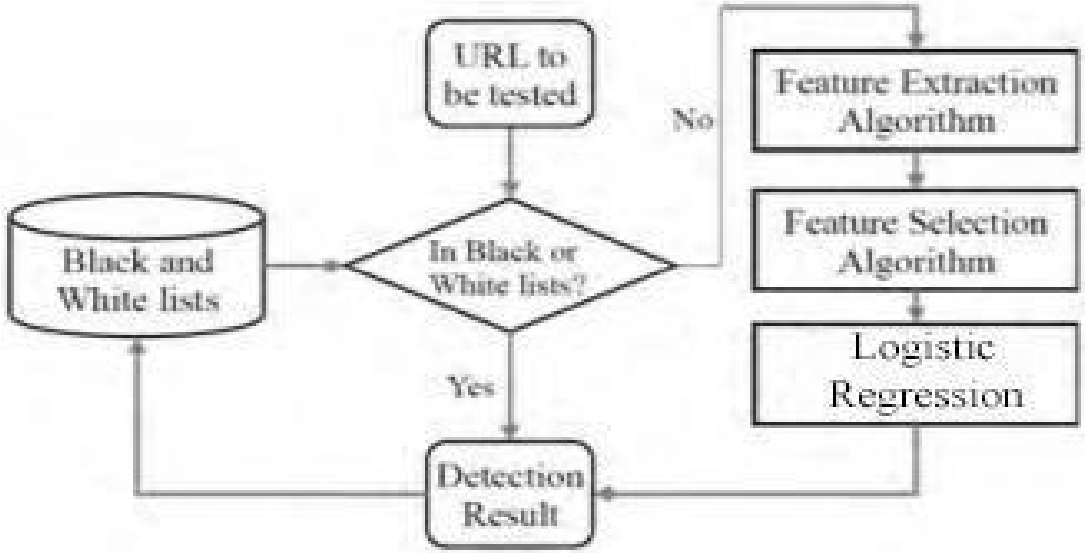
SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Objective of the Project

- The main objective of the project is to make the client find weather the link is spam or not and also the client will get warning information..
- In this project, Different methodologies are using to find out whether the given link is of a good website or a type of phishing website.
- If the phishing website is detected then the client will get the warning information about that website.
- These projects also focused on increasing the accuracy compared to previous projects to extract the data from the data set.

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be **University** *u/s 3 of UGC Act, 1956*

Architecture diagram



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Modules Description

- Data collection and pre processing
- Data cleaning and data transformation
- Data selection and data set
- Feature extraction
- Accuracy

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Data Collection and Pre-processing :

- Data Collection is one of the most important tasks in building a machine learning model.It is the gathering of task related information based on some targeted variables to analyse and produce some valuable outcome.However, some of the data may be noisy, i.e. may contain inaccurate values, incomplete values or incorrect values. Hence, it is must to process the data before analysing it and coming to the results. Data pre-processing can be done by data cleaning, data transformation, data selection.Data clean-ing: Fill in missing qualities, smooth uproarious information, recognize or expel excep-tions, and resolve.

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Data cleaning and Data transformation

- Data cleaning is the process in which we get noisy data. This noisy data in the sense which is not fully completed and has some missing values that will be written fully. This will be taken and the entire data will be read and will take the possibility of clean data so that the data can go to the next process Data transformation may include smoothing, aggregation, generalization, transformation which improves the quality of the data

Data selection and Data sets

- Data selection includes some methods or functions which allow us to select the useful data for our system. Data input: After finding the best algorithm, that algorithm is used to find the phishing websites. After that input is given to the algorithm and then it gives the output.
- A collection of data of a particular instance is stored in a particular sheet which is also known as a dataset and it is used to work on the machine learning that data set can be used for many purposes. Training Data set: A data set that To train our model, it feeds into the logistic regression algorithm. This is the data type used to provide an unbiased evaluation of the final that is completed and fit on the training data set that data is used for the learning purpose.

Feature extraction

In order to reach a website on the internet, the URL is used as a medium. Phishing links are generally disguised as legitimate by doing phishing attacks, the attacker can make a user click on their web links which are pretending to be the original ones. Thesephishing links have few identifiable features.These features are divide into three classes they are divide into three classes they are

- Address bar
- Abnormal Features
- Domain features

Address bar

- In this IP address special characters like (_,,%,,$,#) etc.. and the length of the input URLs are needed to specified to detect phishing attacks accurately. It also has a length
- of the Uniform resource locator (URL), Short URL existence. We also check the fea-tures like Domain registration length. We check all these features and see whether the given link contains all these or not if the criteria is satisfied then only it will go to the other one.

Abnormal Features

- Abnormal features can be checked from the requested URL and the link tags like

<META>,<Script>,<Link>, The S F H status information. Checking from the handler. Check Whether the web page having the link that is going to submit the data entered to any mail. Lastly, say it as an Abnormal URL or not.

Domain features

This section of the code generally checks the age of the domain because that phishing websites have a very short time. It checks the DNS record Web Traffic and page rank as the phishing pages have very few visitors it will have the index of 0 or 1 not more than that. It also checks the google index for that and the number of pages pointing to that page all this information of the features is saved in the Statistical report.

Accuracy

- Exactness is the proportion of the quantity of right forecasts to the all out number of information tests. Exactness functions admirably just if there are an equivalent num-ber of tests having a place with each class. For instance, consider that there are 98% examples of class An and 2% tests of class B in our preparation set.in our case, there are 4,00,000+ info tests and that is taken as contribution to prepare the calculation for highlight extraction this strategic relapse gives the exactness as 96%.

Novelty in Methodology

- In this project, the main goal is to increase the accuracy compared to the previous models. These can be possible by using the new algorithm from machine learning which is logistic regression.
- Logistic regression is a machine learning algorithm that is not used in any other project from these categories, so these make them unique from other models.
- In this project, The size of the data set is increased so much for more security from the phishing attackers. The size of the data set is 4L links.

Equations, derivation and Algorithms used

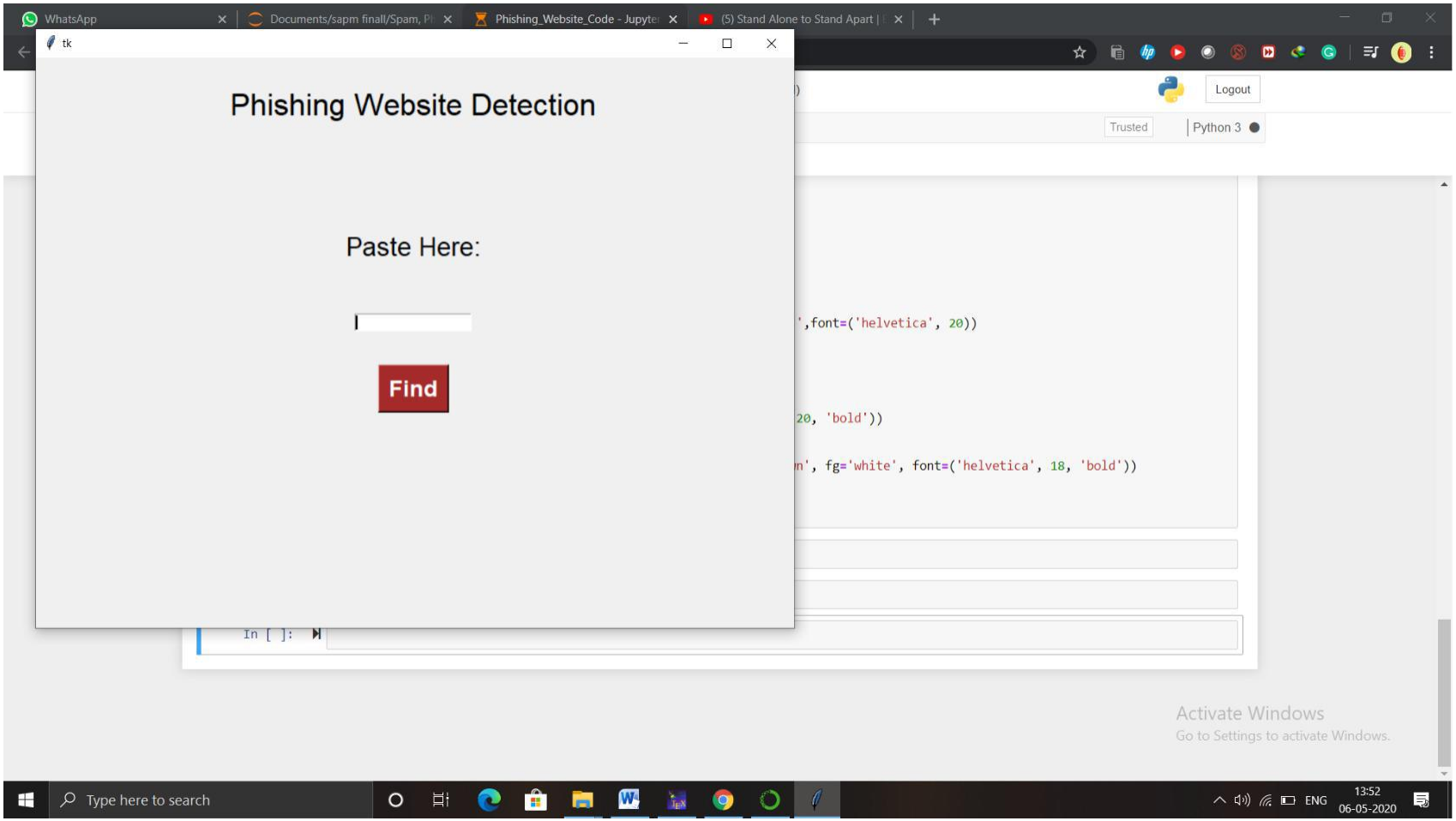
- Before getting to know about the logistic regeression we should know what is regression first.
- **Regression:**
- It is a predictive modelling technique it estimates the relationship between a dependent (target) and an independent variable (predector).
- **example:-**
- **Dependent**:-What is the sale on feb 14. Her sale is the dependent variable.
- **Independent**:-i)Number of product sale
- ii)quantity(predector).
- **Logistic Regression:-**
- Logistic regression produces results in a binary format which is used to predict the outcome of a categorical depedent variable. so the outcome should be discrete/categorical in 0 (or) 1,yes (or) no,true (or) false,good (or) bad,High(or)Low etc..

Equations, derivation and Algorithms used

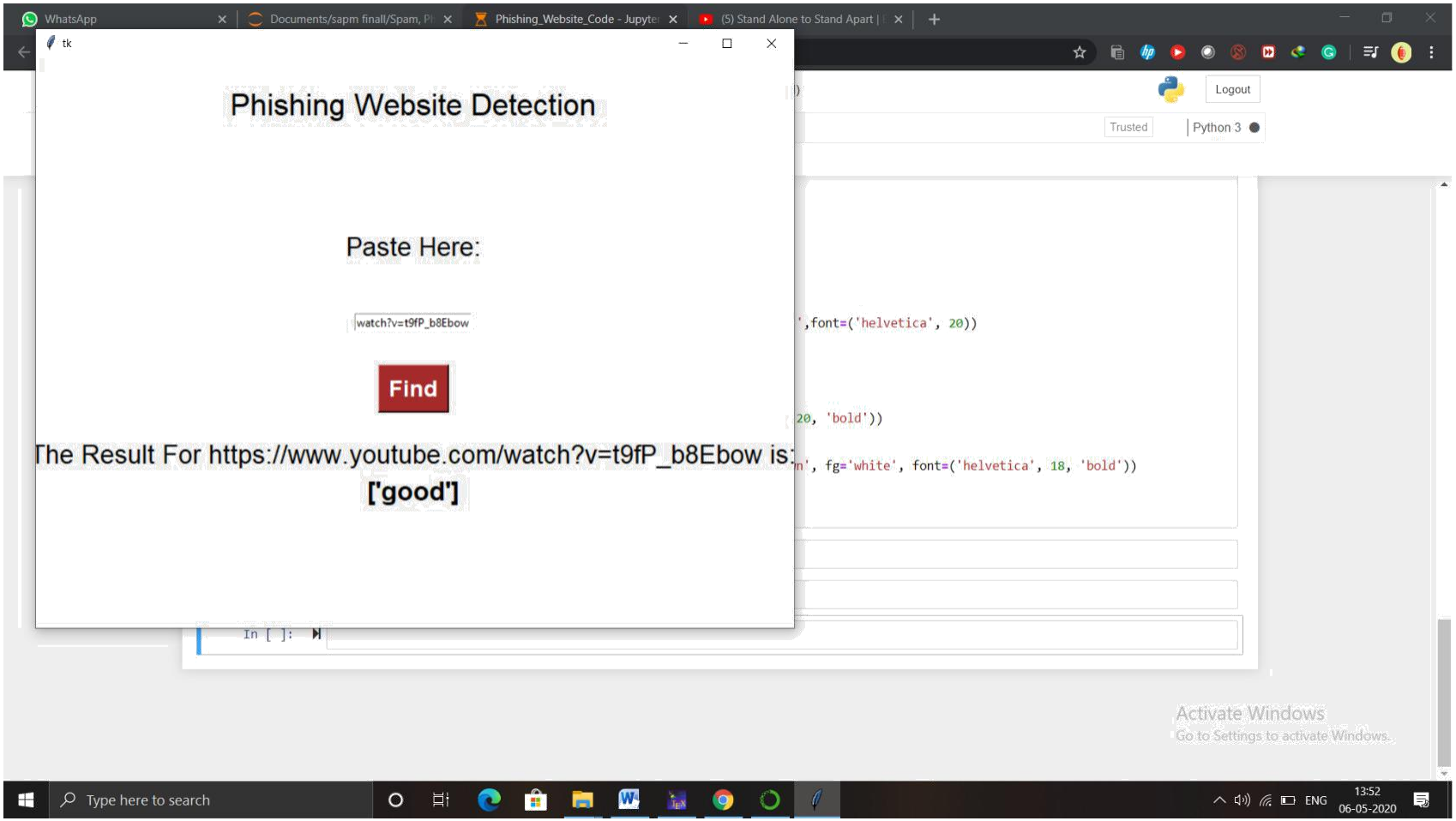
- Normally linear regression has a range from 0-∞and -∞ to +∞.
- In logistic regression is using sigmoids function so linear line ihas to be clipped to 0 and 1.By sigmoid with this the threshold value indicate the probability of 0or 1 or good or bad.
- **Equation:**
- These equation are derived from the Straight line equation.**y=m1x1+c range(-∞ to +∞)**
- for the logistic regression range 0-∞
- to convert it to 0-1 we use
- **equation:-**
- **y/(1-y) where y=0 then 0**
- y=1 the ∞
- for transforming it further to get between -∞ to +∞
- **log({y)/(1-y)}**
- final logitic regression equation

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be **University** *u/s 3 of UGC Act, 1956*

Screen shot of the project



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be **University** *u/s 3 of UGC Act, 1956*



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

Results and Discussions

- Phishing is an unspeakable risk in the web field. In this episode, the basic man inputs individual data to a bogus site which looks equivalent to a typical site. We had done a survey on phishing methods based on visual comparison. This provided a good understanding of the attack and many solutions. Various approaches have conversed in this paper for the detection of phishing; however, most of the methods still have boundaries like accuracy, failing to distinguish objects, and so forth. But In this paper Logistic regression technique finds the accuracy as it is an open challenge in this phishing field the accuracy obtained is 95% percent and it may further increase based on the research
- In this project we are using new algorithm called logistic recurrsion
- As of now the accuraacy is 96% but I can be increased in Based on that data set we can get the result used our decision tree and logistic regression algorithm to predict the result. Here we can also find out the accuracy rate of the algorithm. It will be helpful for finding phishing website whether it is good or badfuture.
- We are going to show that weather the given link is good or phishing link.

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

References

- [1] (2016). PhishMe Q1 2016 Malware Review. [Online]. Available: https://phishme.com/project/phishme-q1-2016-malware-review/
- [2] A. Belabed, E. Aimeur, and A. Chikh, ‘‘A personalized whitelist approach for phishing webpage detection,’’ in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012, pp. 249–254.
- [3] Y. Cao, W. Han, and Y. Le, ‘‘Anti-phishing based on automated individual white-list,’’ in Proc. 4th ACM Workshop Digit. Identity Manage., 2008, pp. 51–60.
- [4] T.-C. Chen, S. Dick, and J. Miller, ‘‘Detecting visually similar Web pages: Application to phishing detection,’’ ACM Trans. Internet Technol., vol. 10, no. 2, pp. 1–38, May 2010.

- [5] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell, ‘‘Clientside defense against Web-based identity theft,’’ in Proc. 11th Annu. Netw. Distrib. Syst. Security Symp. (NDSS), 2004, pp. 1–16
- [6] C. Inc. (Aug. 2016). Couldmark Toolbar. [Online]. Available: http://www.cloudmark.com/desktop/ie-toolbar
- [7] J. Corbetta, L. Invernizzi, C. Kruegel, and G. Vigna, ‘‘Eyes of a human, eyes of a program: Leveraging different views of the Web for analysis and detection,’’ in Proceedings of Research in Attacks, Intrusions and Defenses (RAID). Gothenburg, Sweden: Springer, 2014.
- [8] X. Deng, G. Huang, and A. Y. Fu, ‘‘An antiphishing strategy based on visual similarity assessment,’’ Internet Comput., vol. 10, no. 2, pp. 58–65, 2006.
- [9] Z. Dong, K. Kane, and L. J. Camp, ‘‘Phishing in smooth waters: The state of banking certificates in the US,’’ in Proc. Res. Conf. Commun., Inf. Internet.

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be **University** *u/s 3 of UGC Act, 1956*

Publication details