A STUDY ON REGION BASED ANALYSIS TO DETECT THE ORIGINALITY OF THE WEBSITES USING MACHINE LEARNING ALGORITHM

Mrs. S. DEEPIKA[1]
Assistant Professor, Department of B.Com Business Analytics
PSGR Krishnammal College for Women, Coimbatore, India.
deepika@psgrkcw.ac.in
S. INDRA[2]
UG Scholar, Department of Business Analytics
PSGR Krishnammal College for Women, Coimbatore, India.
indrasoundararajan39@gmail.com

ABSTRACT
To determine whether the websites are original or fake with the help of https anddomainreglen. To avoid the phishing website by analyzing and the count for original & fakewebsite is deducted from the dataset. Classifying the data into training set and test set in theNaive Bayes Classifier Algorithm and training the model. Fitting the training model intoprediction and then evaluating the prediction model. Accuracy is deducted from the Bayestheorem that isin percentageform.
Keywords: Original Website, Fake Website, Naive Bayes Classifier Algorithm, MachineLearning,Data Mining.

I. INTRODUCTION

Machine learning is a method of data analysis that automates analytical model building. It is abranch of artificial intelligence based on the idea that systems can learn from data, identifyingpatternsand makingthedecisions with minimal humanintervention[7].

Supervised learning is the most popular paradigm for machine learning.Whenfully-trained,thesupervisedlearningalgorithmwillbeabletoobserveanew, never-before-seenexampleandpredictagoodlabel for it. Supervised learning is often described as task-oriented because of this. It is highlyfocused on a singular task, feeding more and more examples to the algorithm until it canaccuratelyperform onthattask[8]

Unsupervised learning is very much the opposite of supervised learning. It features no labels. Instead, our algorithm would be fed a lot of data and given the tools to understand the properties of the data. From there, it can learn to group, cluster, and/or organize the data in away such that a human (or other intelligent algorithm) can come in and make sense of the newly organized data. Because unsupervised learning is based upon the data and its properties, we can say that unsupervised learning is data-driven. The outcomes from an unsupervised learning task are controlled by the data and the way it is formatted. Reinforcement learning is fairly different when compared to supervised and unsupervised learning. [8]

Clustering is a type of unsupervised machine learning algorithm. It is used to group data point shaving similar characteristics as clusters. Ideally, the data points in the same cluster should exhibit similar properties and the points in different clusters should be as dissimilar possible.[6]

Classification is a supervised learning method. Classification involves classifying the input data as one of the class labels from the output variable. Classification involves the prediction of the input variable based on the model building. Classification algorithms need the data to be split a straining and test data for predicting and evaluating the model. Classification algorithms deal with labeled data, Classification process involves two stages – Training and Testing. As classification deals with a greater number of stages, the complexity of the classification algorithms is higher than the clustering algorithms whose aim is only to group the data. [6]

II. LITERATURE OF REVIEW

By successfully exploiting human vulnerabilities, fake websites have emerged as a major source of online fraud. Fake websites continue to inflict exorbitant monetary losses and also have significant ramifications for online security. We explore the process by which salient performance-related elements could increase the reliance on protective tools and, thus, reduce the success rate of fake websiteseefficacy in dealing with threats, and reliance on such tools. The research method was a controlled lab experiment with a novel and extensive experimental design and protocol in two distinct domains: online pharmacies and banks. Furthermore, reported reliance on the detector showed a significant impact on the users' performance in terms of self-protection.Therefore, users' perceived response efficacy should be used as a critical metric to evaluate the design, assess the performance, and promote the use of fake-website detectors. We also found that cost of detect or error had profound impacts on threat perceptions. We discuss the significant theoretical and empirical implications of the findings. [4]

Online security is a major problem for financial institutions worldwide. Account hijacking and online fraud are on the rise. Financial losses in the banking industry due to attacks have been estimated in 2003 to be about US$1.2 billion in the US alone. Studies also indicate that security concerns are a major issue for an increasing number of consumers. The rapid growth in phishing attacks

threatens the future of online banking. In the absence of an adequate response, banks are likely to incur even greater costs and experience a significant decline in consumer trust. Essential among the recommendations is the need to involve the consumer in managing security concerns. Specific recommendations to help improve actual security and increase consumer trust in the system are proposed. [2]

Fake websites have become increasingly pervasive, generating billions of dollars in fraudulent revenue at the expense of unsuspecting Internet users. The design and appearance of these websites makes it difficult for users to manually identify them as fake. Automated detection system shave emerged as a mechanism for combating fake websites, however most are fairly simplistic in terms of their fraud cues and detection methods employed. Consequently, existing systems are susceptible to the myriad of obfuscation tactics used by fraudsters, resulting in highly ineffective fake website detection performance. In light of these deficiencies, we propose the development of a new class of fake website detection systems that are based on statistical learning theory (SLT). Using a design science approach, a prototype system was developed to demonstrate the potential utility of this class of systems. We conducted a series of experiments, comparing the proposed system against several existing fake website detection systems on a test bed encompassing 900 websites. [3].

By using the anonymous structure of the Internet, attackers set out new techniques, such as phishing, to deceive victims with the use of false websites to collect their sensitive information such as account IDs, usernames, passwords, etc. Understanding whether a webpage is legitimate or phishing is a very challenging problem, due to its semantics-based attack structure, which mainly exploits the computer users' vulnerabilities. Although software companies launch new anti-phishing products, which use blacklists, heuristics, visual and machine learning-based approaches, these products cannot prevent all of the phishing attacks, execution, detection of new websites, independence from third-party services and use offeature-richclassifiers.

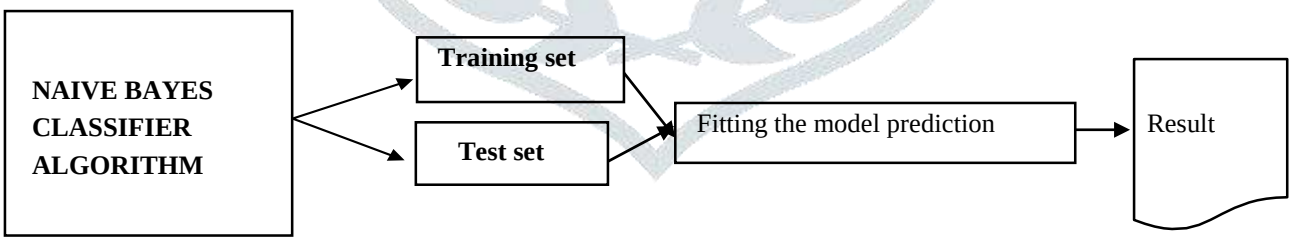
Formeasuringtheperformanceofthesystem,anewdatasetisconstructed, and the experimental results are tested on it. According to the experimental and comparative results from the implemented classification algorithms, Random Forest algorithm with only NLP based features gives the best performance with the 97.98% accuracy rate for detection of phishing URLs. [5]

Most of the anti-phishing solutions are having two major limitations; the first is the need of a fast access time for a real-time environment and the second is the need of high detection rate. Black-list-based solutions have the fast access time but they suffer from the low detection rate while other solutions like visual similarity and machine learning suffer from the fast access time. In this paper, we propose a novel approach to protect against phishing attacks using auto-updated white-list of legitimate sites accessed by the individual user. Our proposed approach has both fast access time and high detection rate. When users try to open a website which is not available in the white-list, the browser warns users not to disclose their sensitive information. Furthermore, our approach checks the legitimacy of a webpage using hyper link features. For this, hyper links from the source code of a webpage are extracted and apply to the proposed phishing detection algorithm. Our experimental results show that the proposed approach is very effective for protecting against phishing attacks as it has 86.02% true positive rate while less than 1.48 % false negative rate. Moreover, our proposed system is efficient to detect various other types of phishing attacks (i.e., Domain Name System (DNS) poisoning, embedded objects, zero-hour attack). [1]

III. METHODOLOGY

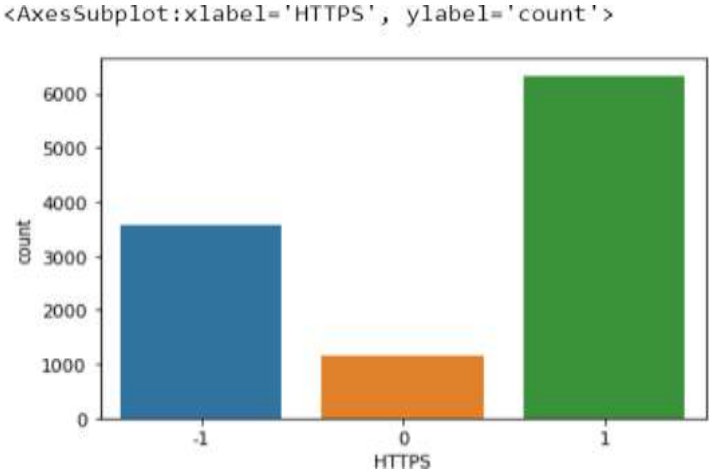
Naïve Bayes algorithms area set of supervised machine learning algorithms based on the Bayes probability theorem, which we'll discuss in this article. Naive Bayes algorithms assume that there's no correlation between features in a dataset used to train the model. [10]

WORKFLOW



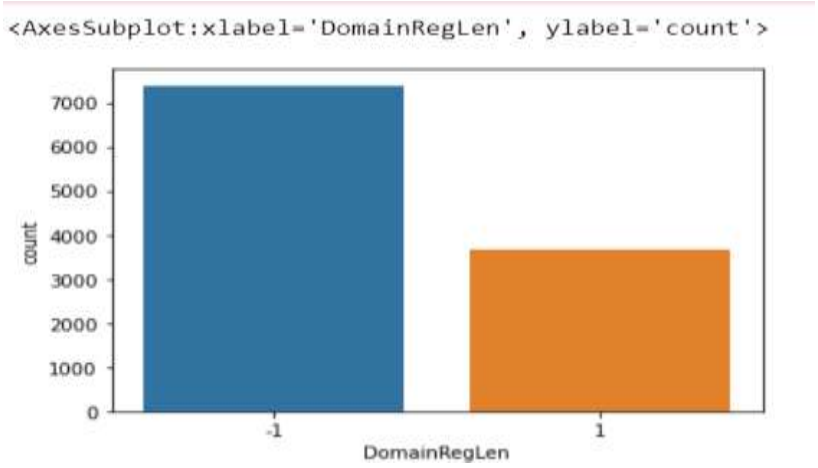
IV. RESULT

Fig–4.1



In the above figure HTTPS consist of- 1&0 as fake website and 1 as original website, it shows the count plot where the count for no. of observations in it are displayed.

Fig–4.



In the above figure DOMAINREGLEN consists of-1as fake website and1as original website. It shows the count plot where the count for no. of observations in it is displayed.

ACCURATE VALUE -

Accuracy = (y_test,y_pred)*100= 98.046309696

In the above figure data pre-processing is done, training the model, fitting the model into prediction and evaluating the model with the result as accuracy score is evaluated here as 98.04630969609262 in percentage as its proposed according to the Naive Bayes Classifier Algorithm.

IV CONCLUSION AND FURTHER WORK

CONCLUSION

The Website security fraud analysis with machine learning is based on original or fake website analysis and to find out the count and accuracy of original website by attributes. Using the Naïve Bayes Classifier Algorithm data is been predicted and model evaluation is done by fitting the model prediction into evaluation of the model and the result is deducted by the accuracy score in percentage.

FURTHER WORK

Naïve Bayes Algorithm accepts only categorical input variables, and compares better the logistic regression models; it can be further developed by bringing high the acceptance of numerical values and its performance to make the users much more flexible in handling Naïve Bayes Classifier algorithm.

REFERENCES

[1]. A novel approach to protect against phishing attacks at client side using auto-updated white-list Ankit Kumar Jain, Brij B Gupta EURASIP Journal on Information Security 2016(1), 1-11, 2016

[2]. Addressing consumers' concerns about online security: A conceptual and empirical analysis of banks' actions Dan Sarel, Howard Marmorstein Journal of Financial Services Marketing 11(2), 99-115, 2006

[3].Detecting fake websites: The contribution of statistical learning theory Ahmed Abbasi, Zhu Zhang, David Zimbra, Hsinchun Chen, Jay F Nunamaker Jr Mis Quarterly,435-461,2010

[4]. Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance Fatemeh Mariam Zahedi,Ahmed Abbasi, Yan Chen
Journal of the Association for InformationSystems16 (6),2,2015

[5]. Machine learning based phishing detection from URLs Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir,Banu Diri Expert Systems with Applications117, 345-357,2019

[6]. https://www.upgrad.com/blog/clustering-vs-classification/

[7]. https://www.sas.com/en_in/insights/analytics/machine-learning.html

[8].https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756fhttp://localhost:8888/tree

[9]. http://localhost:8888/edit/Desktop/jupyter/websitedataset1.csv

[10]. https://heartbeat.fritz.ai/naive-bayes-classifier-in-python-using-scikit-learn-13c4deb83bcf