Detection of Phishing Website Using Support Vector Machine and Light Gradient Boosting Machine Learning Algorithms

Mr. B.Narasimha Rao

Associate Professor ,Dept. Of Computer Science and Engineering

B. Varaprasad, B. Lokesh Varma, B. Srinivas, G. Roshan Surendra

B.Tech Students ,Dept. Of Computer Science and Engineering

Bonam Venkata Chalamayya Engineering College, Odalarevu.

ABSTRACT-- Phishing is still one of the most straightforward attacks a cybercriminal can launch in order to obtain sensitive information from their targets such as unique usernames and passwords, and even banking information. With the rate of such threats increasing, professionals in the cybersecurity field are focusing their resources on developing more efficient techniques for detecting phishing websites. This work studies the implementation of machine learning algorithms that are capable of recognizing phishing URLs by feature extraction and analyzing various differences from legit websites. More specifically, Decision Tree, Random Forest, and Support Vector Machine (SVM) algorithms are used for classification. In addition, the aim of the study is to improve detection performance by using LightGBM alongside SVM for the phishing detection task.

Keywords:* *SVM, Light GBM, phishing website.

1.INTRODUCTION

The internet has become an integral part of our everyday lives as it has changed the way we communicate, learn, conduct business, and shop. With the growth of the internet, almost every physical task nowadays can be done online. The internet offers an extensive repository of valuable data that can aid in the growth of an individual, organization, economy, or society. You can search and find any form of information, anytime, anywhere around the globe.

B. Narasimha Rao (Associate Professor)

Phishing is an act of email scamming wherein a fraudulent email is sent to users in an attempt to get the users to provide confidential information via deceitful websites or messages.

The focus of this paper will be on URL phishing, where we identify phishing URLs by extracting and analyzing features that differentiate malicious URLs from legitimate ones. The used algorithms in the detection process include Decision Trees, Random Forest, and Support Vector Machines (SVM) which are some of the Machine Learning algorithms that are used.

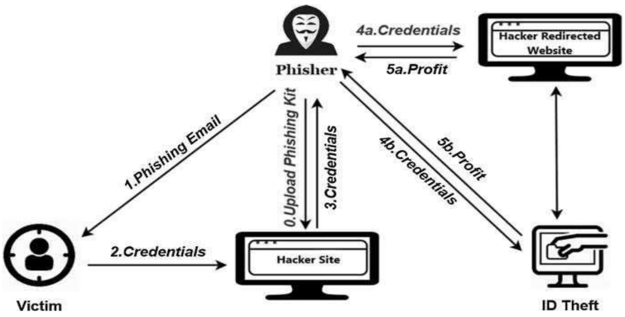


Figure 1: Process of Phising Attack

2.LITERATURE SURVEY

2.1 Phishing Attacks and Their Impacts

- Phishing websites are designed to deceptively solicit personal information from users.
- Phishing techniques have become more effective than traditional blacklisting approaches and detection methods.

2.2 Traditional rules for phishing detection by definition

- Blacklist-Based Detection is where a fake site refers to pre-identified rules of phishing behaviors, however, it often mistakes legitimate sites as phishing.

2.3 Machine Learning Based Phishing Detection

- Knowledge-Based Machine Learning models are used to classify domain names of webpages by attributes that are extracted from the pages.

3.Existing System

Phishing is a method of deception characterized by the sender masquerading as an established and trusted site of a third party typically within a communication that appeared to be legitimate. The bogus communication often contains a link or file or both you could click on to obtain sensitive information or download malicious code. For years, phishing attacks were done on a mass scale, spamming random or targeted users in the hopes that some small percentage of those users would click on the malicious web link or download themalicious file.

Numerous techniques have been explored to identify and detect such attacks, one of those techniques is machine learning. This technique incorporates several different machine learning algorithms by providing a machine learning model with URLs produced and sent to users. The machine learning model receives links which are processed and based on the model's algorithm it will predict if a user was sent a phishing attempt or a legitimate site. Numerous different machine learning algorithms can be utilized in this classification task including Decision Trees, Support Vector Machines (SVM), random forests, neural networks, XGBoost, etc.

The proposed methodology will specifically apply Random Forest and Decision Tree classifiers to identify phishing URLs.

4.Proposed System

Phishing is an increasingly important cybersecurity risk in which attackers send phishing messages designed to replicate messages from legitimate, trusted parties. These phishing messages often include malicious links or attachments that could compromise the user's personal information or infect the user's device with malware when accessed. In the past, phishing attacks were typically based on large-scale spam campaigns - randomly disseminating emails or messages to as many people as possible in the hopes of exploiting a few.

To mitigate these risks, a number of detection schemes have been developed, and machine learning schemes to detect phishing have had particular success. In a machine learning based approach, URLs that users receive are placed into trained models, which analyze and classify the URLs as phishing or legitimate. A number of different machine learning algorithms, such as Support Vector Machines (SVM), Neural Networks, Random Forests, Decision Trees, and XGBoost, have been used to improve the accuracy of these classifications

For this study, we are taking a different approach to detecting phishing URLs by applying Random Forest and Decision Tree classifiers.

4.1 METHODOLOGY

This study aimed to determine phishing websites using machine learning methods Support Vector Machine (SVM) and Light Gradient Boosting Machine (LGBM). The methods began with the extraction of relevant features from the website URL, and the features extracted were used to build models of classification. The models were tried using a good quality secondary dataset, which included legitimate URLs and also phishing URLs. Therefore, we applied a suite of evaluation methods to help determine how well the models performed.

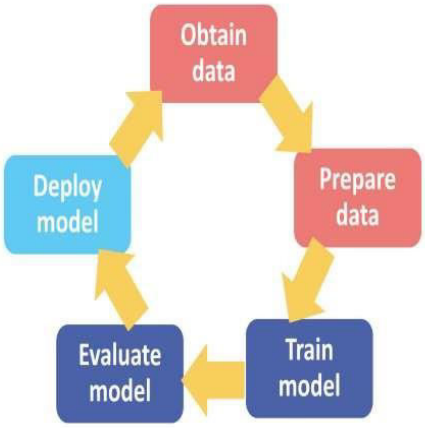


Figure 2: Process Diagram

4.1.1 Dataset Description

The dataset used for this study is the NSL-KDD phishing dataset, publicly available at: https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machinelearning.

The features are described in the previous subsection and supported using the primary data and analysis presented in the domain of study for two periods - 1st stage (13 January - 22 May 2015; 2nd stage (17 May - 19 June 2017). There were 10,000 records of websites were collected - 5,000 legitimate websites and 5,000 phishing websites. Each record included 48 features, which has been systematically extracted to reveal a number of behavioral and structural features

4.1.2 Modelling

After the data has been thoroughly preprocessed and ready to analyze, we can enter the modeling stage of the process by utilizing the Support Vector Machine and Light Gradient Boosting Machine algorithms. During the modeling stage, machine learning exercises as structured systems to explore the anticipated outputs based on the objectives of our task.

The SVM and LightGBM algorithm were selected due to their demonstrated effectiveness in classification tasks. The modeling stage trains the model(s) on the features extracted from the data set, optimizes hyperparameters, and evaluates performance using commonly applied measures

4.1.3 SVM

Support Vector Machine (SVM) is an important supervised machine learning algorithm for classifications and regressions tasks. SVM is one of the most used algorithms because of its superior prediction accuracy and was proposed by Vladimir Vapnik and the colleagues at AT&T Bell Laboratories. Statistical learning theory, particularly VC (Vapnik–Chervonenkis) theory, provided the original mathematical foundation for SVM, which since was first introduced in the early 1980s. It operates by identifying a non-probabilistic binary linear classifier; therefore, the model is learned through a set of labeled data points, with each data point misclassified to one of the two classes. The objective of SVM is to search for the optimal hyperplane that separates these classes and does this by maximizing the margin of separation. The support vectors used in SVM are the nearest points to the separating hyperplane and new data points into the feature space are classified based on which side of the separating hyperplane they appear in. While SVM is a deterministic classifier by nature, some SVM approaches authors will present based on fields are probabilistic classifier examples, which can be achieved through Platt scaling and others.

4.1.4 Light GBM Algorithm

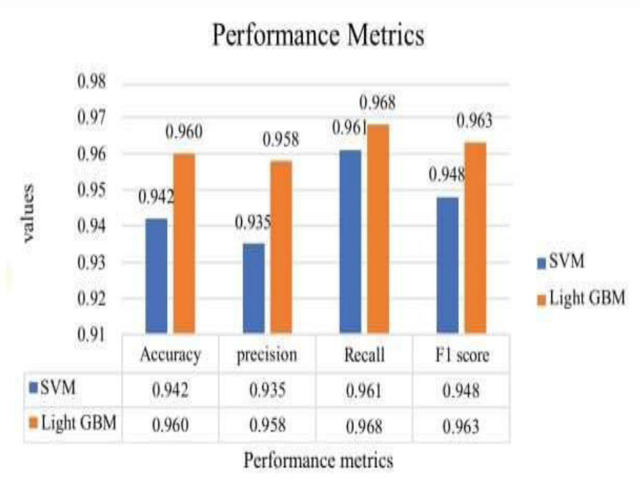
Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework designed and developed by Microsoft. It is designed to be a more efficient, scalable, and faster alternative to decision tree-based learning. It works similarly to other boosting models (e.g., XGBoost), by measuring the contribution of individual predictors, but uses many new methods to speed up the process and improve performance.

LightGBM uses a method called gradient based one side sampling (GOSS) to choose training -samples. It builds trees leaf-wise as opposed to level-wise. This means that it splits leaves that will reduce loss the most since it takes the leaves which have the highest gradient. By doing so, It is faster at making predictions and has more accurate models. It can also be used for distributed or decentralized training. This offers the ability to handle gigantic datasets without crashing or slowing down the process.

In addition, LightGBM also has L1 (Lasso) and L2 (Ridge) regularization to prevent over fitting. It utilizes multiple algorithms such as GBT as used in model trees, GBDT (gradient boosting decision tree), GBRT, GBM (gradient boosting model), MART (multiple additive regression trees), in addition to the old Random Forests (RF) model.

The innovations of LightGBM over standard boosting model are to optimize for sparse features, parallel training, bagging, early stopping, and multiple loss functions. It uses a bottom-up leaf-wise tree construction which is different from a level-wise methods as in the case of XGBoost, which affords speed and state-of-the art predictive power for a variety of machine learning predictions.

5.PERFORMANCE MATRICS

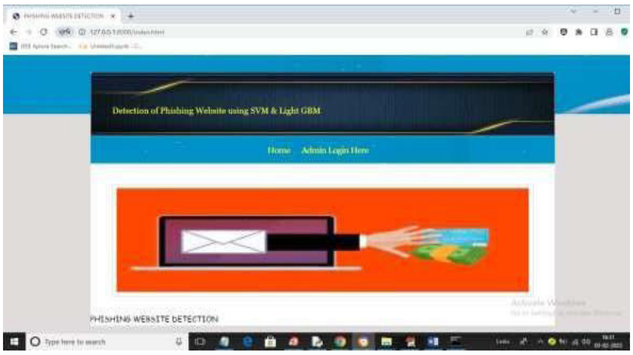


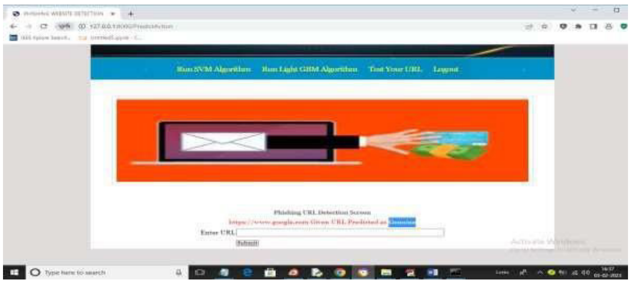
	Accuracy	precision	Recall	F1 score
■SVM	0.942	0.935	0.961	0.948
■Light GBM	0.960	0.958	0.968	0.963

Figure 3: Performance Metrics of model

6.RESULT

The chosen algorithms are applied on the dataset for investigating the extracted features associated with the identified phishing attacks. After training and testing the models, the accuracy reported by the system is the next evaluation metric. At this state of the project, it will provide insight into the characteristics of phishing attacks identified during training, and also validate model performance and feature selections in the proposed framework.


PHISHING WEBSITE DETECTION
Detection of Phishing Website using SVM & Light GBM
Home
Admin Login Here
PHISHING WEBSITE DETECTION
Type here to search



Phishing URL Detection Screen

https://www.google.com Given URL Predicted as Genuine

Enter URL

Submit

Figure 4 &5: Results of model

7.CONCLUSION

The goal of this research was to enhance detection of phishing websites through the use of machine learning techniques. The results of the experiments presented demonstrate that a model using the Random Forest algorithm was able to detect phishing urls with an accuracy of 97.14% via detection, with a lower false positive rate. Furthermore, the results indicate that the longer training times, and more data to train with, produced promising results.

Given that the results of the research show that implementing Random Forest would be a hybrid approach mixing traditional blacklist detection, these methods should enhance Accuracy and trustworthiness in phishing detection methods.