An Ensemble ML model in Detection of Phishing Websites

ABSTRACT

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing [1]. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

PROCESS

We will collect the dataset from open source platforms, extract the required features from the URL database, analyse and preprocess the data. Then we will be using machine learning algorithms with deep learning techniques like SVM, Random Forest Classifier and Autoencoder on the dataset. We will finally compare the different approaches and find out which one is the better alternative for this purpose.

```
┌─────────────────────────┐
│  DATASET CONTAINING     │
│                         │
│    ALL THE URLS         │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│      FEATURE            │
│                         │
│    EXTRACTION           │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│    NEW DATASET          │
│                         │
│    CREATION             │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│   0 – LEGITIMATE        │
│                         │
│   1 – PHISHING          │
│                         │
│   2 - SUSPICIOUS        │
└─────────────────────────┘
```