8-9-2021

The Classification of Phishing Websites using Supervised Data Mining Techniques

Justin Lichtfuss
Georgia State University, jlichtfuss1@student.gsu.edu

Frank Lee
Georgia State University, flee@gsu.edu

Trezha Berryman
Georgia State University, tberryman1@student.gsu.edu

The Classification of Phishing Websites using Supervised Data Mining Techniques

TREO Talk Paper

Justin Lichtfuss
Georgia State University
Jlichtfuss1@student.gsu.edu

Frank Lee
Georgia State University
flee@gsu.edu

Trezha Berryman
Georgia State University
tberryman1@student.gsu.edu

Abstract

Phishing attacks are on the rise, and the consequences for businesses are severer. The impact of a phishing attack not only causes financial loss but also triggers data breaches. The data breaches caused by phishing attacks often lead to reputational damage and business disruption. Therefore, detecting potential phishing attempts has received tremendous attention. The purpose of this study is to identify the feature predicting the presence of a phishing site by using the public phishing URL dataset. The dataset used in this study includes 87 predictor variables across three distinct feature groups, including 1) 56 URL-based features obtained by analyzing the text of URLs, 2) 24 Content-based features extracted by loading the web pages of URLs and analyzing their HTML contents, 3) and seven external features obtained by querying reference third party services and search engines. The top-7 most meaningful inputs from each feature group are selected and analyzed in three different supervised data mining techniques to determine which feature group produces the most robust model for classifying and detecting phishing websites. The result of this study shows that the inputs from the external features group consistently had the highest Accuracy, Specificity, Sensitivity, and Precision across all supervised data mining techniques. This study also finds that the model can be improved by using a combination of inputs from all three feature groups, including 3 URL-based features, 2 Content-based features, and 2 External features. The result of this study will help shape and strengthen security awareness training for organizations and be used as the foundation for building preventative tools for both individuals and companies against phishing attacks.