COS 781 Project Proposal Nathan Peter Opperman - u21553832

Classifying Phishing Websites Using Machine Learning Techniques

Phishing is the malicious practice where attackers impersonate organizations, such as banks or retailers, through emails or messages, in an attempt to trick individuals into revealing sensitive information. Phishing websites, which mimic legitimate sites, play a crucial role in these attacks by luring users to enter their private information. It is quite clear that phishing attacks pose a significant threat and so there is a need for effective detection methods to identify and mitigate these threats. With this project I intend to develop an ML model capable of accurately (and efficiently - important) detecting these phishing websites to enhance cybersecurity. Some important questions that I want to answer are: What are the key features that differentiate phishing websites from legitimate websites? How efficient should practical detection methods be? What approach will be best suited with efficiency and usability in mind?

Data

For this project I intend to utilize the publicly available Phishing Websites tabular dataset. The data was collected from PhishTank archive, MillerSmiles archive, Googleâ€™s searching operators. There are a total of 11055 instances with 30 attributes (both numeric and categorical) and a target attribute indicating whether it is a phishing website. Some key attributes include the use of URL shortening, URL length, the presence of double slashes, and others that may indicate a phishing website. The dataset luckily has no missing values and has a relatively balanced class distribution (55% positive and 44% negative).

Approach

In order to tackle the problem of phishing website classificatio I intend tol use a combination of exploratory data analysis, preprocessing, and various machine learning models. Since the dataset is balanced and has no missing values, I will focus on feature scaling and one hot encoding where necessary. The models I will use for comparison include XGBoost, Decision Trees, and SVM's. Performance will be evaluated using accuracy, precision, recall, F1-score as well as speed/efficiency. Overall, I aim to identify the model that balances accuracy, precision, and efficiency for practical use in phishing detection systems. I however suspect that XGBoost will give the best results, while decision trees will provide a degree of "explainability" which might be useful. I expect SVN to handle the binary classification well, particularly with optimal hyperparameters.

Success Criteria

Given the context of this problem an important aspect is the value of a True Positive vs a False Positive. Blocking harmless sites can lead to frustration and cause users to ignore or bypass the detection system, undermining its effectiveness. This is why precision and accuracy are crucial metrics for evaluating the success of the model. That way I can maximize protection without needlessly interrupting a potential user's experience. Currently there are many available baselines to compare our results with, these attempts include the use of ANN, Random Forrests, SVM as well as many others. This should give a good baseline from which to compare my results with.

End-of-Semester Deliverables: At the end of the semester I expect to have a greater understanding/insight of what the requirements are for a practical detection method and how well different methods are suited to this problem with a final model that can efficiently and effectively maximize protection without being overly protective.