Summary 18

Shaun Pritchard

Florida Atlantic University

CAP 6778

November 23, 2021

M. Khoshgoftaar

Feature Popularity Between Different Web Attacks with Supervised Feature Selection Rankers

The study evaluated feature popularity using ensemble feature selection techniques for three different types of web attacks that were part of the CSE-CIC-IDS2018 dataset: brute force attacks, SQL injection attacks, and XSS attacks. Feature popularity is a new concept, which uses ensemble feature sections to produce easier and simpler machine learning models by producing feature importance lists with the top 20 features compiled for the three web attacks using four supervised rankers to determine how the FSTs work with two classifiers. A combination of classifiers and FSTs consisted of Light LGB and XGBoost (XGB), while Random Forest (RF) and Cat Boost (CB) were used for FST only. In this study, AUC is applied to the learning measure performance and ROC is applied to the performance metrics.

The initial implementation involved first building a list of important features for each dataset followed by finding features that are shared among the FSTs for each dataset and then building a list of feature similarities among the FSTs and ds(a). This was followed by a full analysis implementing ensemble learning as proposed research methods

Even though they had been closely following web attack protection over the past few years, this research gained new insights into attack deception even after they applied these four FSTs to the top 20 features including Flow Bytes s, Flow Packet s, and Flow IAT Max. According to the Jaccard similarity frameworks, SQL Injection and XSS are the most similar feature subsets among the three web attacks. All three of these most popular features are based on time-related attributes. Each flow contains a set of bytes per second, a set of packets per second, and a maximum interval between two flows known as Flow IAT Max.

Having explored common features between different FSTs, this research can sometimes lead to the discovery of even better feature subsets by adjusting our popularity criteria to be more or less restrictive in two dimensions when building ensembles. They have demonstrated that they can evaluate classification performance on a variety of cyberattacks using different feature popularity thresholds.