Task 1: Data Selection and Preprocessing

Diana Pham

Progress Report:

The CICIDS-2017 dataset has millions of data points since it consists of eight traffic monitoring sessions that closely represent the current real-world network traffic, so the first task of my project focused on data selection and preprocessing for future feature selection and modeling tasks. Since I am recreating the results from the research paper, CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection, by Kurniabudi, D. Stiawan, et al, I selected 20% of the full dataset which resulted in about 600,000 rows. This step was done so that the classifiers could run within a reasonable amount of time in case any debugging steps were necessary.

For data preprocessing, I removed redundant features such as `Fwd Header Length` to reduce the number of features that will be processed in future tasks. In addition, the original labels were relabeled to the following updated class attacks:

Updated Class Label	**Original Class Label**
Normal	Benign
Bot	Bot
Brute Force	FTP-Patator, SSH-Patator
DoS/DdoS	DDoS, DoS, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed
Infiltration	Infiltration
Portscan	PortScan
Web Attack	Web Attack - Brute Force, Web Attack - Sql Injection, Web Attack - XSS

Lastly, the preprocessed dataset was split into train and test sets by randomly selecting 70% to be the training data and the remaining 30% as the test data.

Going forward with this project, the next task I will be working on will be implementing feature selection using information gain and creating subsets of features for models that will detect anomalies.