Final Project Proposal

CYBR 5320: Cybersecurity Network Analytics

For this project, I intend to work alone on an alternative project presented to me by Dr. Dave Eargle as I am a part of the MSBA - Security Analytics program. The goal of this project is to utilize the CTU-13 dataset to generate classification models so that I can predict whether or not a particular IP address is a part of a botnet. My final deliverable will outline the CRISP-DM format, including the following that will mirror my production process:

I. **Business Understanding**
   A. Business Objective Outline
   B. Situation Assessment
   C. Goal Definition
   D. Project Plan Outline

II. **Data Understanding**
   A. Data Collection Overview
   B. Data Description
   C. Data Exploration
   D. Data Quality Check

III. **Data Preparation**
   A. Data Selection
   B. Data Cleaning
   C. Ad Hoc Data Construction
   D. Ad Hoc Data Integration
   E. Data Formatting

IV. **Modeling**
   A. Technique Selection
   B. Data Splitting
   C. Model Generation
      1. Model 1
      2. Model 2
      3. Model 3
   D. Model Assessment

V. **Evaluation**
   A. Result Evaluation
   B. Process Review
   C. Deployment Preparation

VI. **Deployment**
   A. Final Report Preparation
   B. Project Review

While the deployment phase of CRISP-DM oftentimes includes further maintenance sections, this project is more isolated and does not mandate a maintenance framework. However, I do plan to provide a theoretical framework for deployment, as if this were meant to be a full, ongoing project. I will be tracking my progress on GitHub; the repository is linked below.

Repository Link: https://github.com/NeilCollinsMS/CTU-13-Classification

As for my first thoughts, I am expecting to utilize random forest classification models, from the Sklearn (scikit-learn) library in Python as well as basic logistic regressions, and based on my discoveries during data exploration, I may make use of Naive Bayes, Support Vector Machines, and/or other classification algorithms. Initial research suggests that random forest classification, support vector machines, and NB are the best algorithms for botnet classification.

This is important for the field of cybersecurity as identifying botnet activity solely based on IP address is largely beneficial and can be utilized both by websites looking to analyze their customer traffic while filtering out bot traffic, and by professionals aiming to stem the spreading of malware and the use of DDOS attacks.

Data Link: https://www.stratosphereips.org/datasets-ctu13