CTU-13 Classification Preliminary Results

***Abstract* – The goal of this project is to utilize classification models to detect whether an IP within the CTU-13 dataset college by CTU University in the Czech Republic in 2011. Thus far, the primary algorithm utilized is untuned Scikit-Learn's Random Forest Classification, though I intend to utilize Naïve Bayes in the future, as well as potential usage of XGBoost and Support Vector Machines. This abstract will be revised and added upon prior to final submission.**

I. Initial Introduction

The goal of this project is to utilize classification models to detect whether an IP within the CTU-13 dataset college by CTU University in the Czech Republic in 2011. Thus far, the primary algorithm utilized is untuned Scikit-Learn's Random Forest Classification, though I intend to utilize Naïve Bayes in the future, as well as potential usage of XGBoost and Support Vector Machines.

As of March 1st, 2021, 13 initial classifier models have been generated using an 80/20 train/test split and a maximum of 100 leaf nodes. There was one model built per scenario with minimal tuning, targeting a premade binary column entitled 'Malicious' based on IPs that were reported as infected by CTU University in the initial study. These models are not intended to be final, though they are being used to better understand each scenario so that insights can be transferred into aggregate data models. Per CTU University's suggestion, bidirectional net flows are being utilized for model building.

II. Initial Results

Initial insights suggest that accuracy will not be a feasible metric to grade models without significant tuning as the data is incredibly skewed (there are many more non-malicious net flows than there are malicious) and is likely leading to overfitting in the initial models, which will need to be dealt at some point as consistent accuracy scores above 0.99 do little to inform us about the models. Furthermore, the F1 scores are being skewed by the fact that the models lean towards false positives, likely another result of the skewed class priors. Moving forward, models will be tuned to put more weight on TPR over PPV when calculating F1 scores.

The positive outcome of this is that models that were being used to gain initial insights into the data and were never intended to be implemented for true classification will be far more useful than intended. Had it not been for these models, a variety of pitfalls would've been discovered after more time was spent fitting more expansive models.

III. Next Steps

The next steps in the project will be to aggregate the scenarios into larger groups in two different ways. The first aggregation method will be to combine all of the net flow data into one master data frame before performing a train/test split for general analysis, though this may be limited by the amount of RAM allotted to me in Google Colab (the IDE). If there are issues, I may pivot to utilizing the local machine or the class server to increase computing power. If all the above measures fail, slimming the data symmetrically for ease of computation will be explored. The second aggregation method will be to combine 12 scenarios into a single training data frame, and to test on the 13th scenario.

After reviewing the initial models, a few key concepts will be given additional consideration when tuning aggregate models:

A. Target Feature Skew (Altering F1 algorithm to focus more on TPR than PPV)

B. AUC ROC

C. Alternate Train/Test Splits (5 Fold Validation, etc…)

D. Optimzing Confusion Matrix Thresholds

IV. Repository

General progress logs and initial code can be found here: https://github.com/NeilCollinsMS/CTU-13-Classification

V. Related Works

Haghighat, M. H., & Li, J. (2018, November). *Edmund: Entropy based attack Detection and Mitigation engine Using Netflow Data*. Retrieved February 27, 2021, from https://www.researchgate.net/publication/329457489_Edmund_Entropy_based_attack_Detection_and_Mitigation_engine_Using_Netflow_Data.

Garcia, Sebastian. Malware Capture Facility Project. Retrieved from https://stratosphereips.org

Le, D., & Zincir-Heywood, N., & Heywood, M., *Data analytics on network traffic flows for botnet behaviour detection*. Retrieved February 27, 2021 from https://ieeexplore.ieee.org/document/7850078

Vishwakarma, A. (2020, May). *Network Traffic Based Botnet Detection Using Machine learning*. Retrieved March 1, 2021 from https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1917&context=etd_projects