Machine Learning Application in Cyber Security Network Anomaly Detection With Machine Learning

Adeoluwa Agbakosi[1, 2] Peculiar Abolade[1] Daniel[1]

1. Federal University of Agriculture, Abeokuta 2. LearnDev Foundation

INTRODUCTION

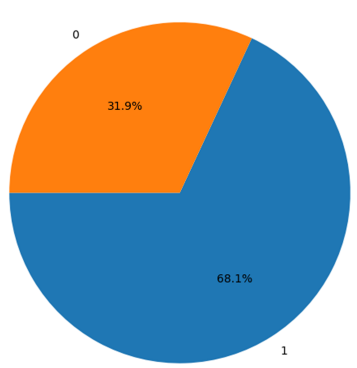
As the frequency of cyberthreats continues to escalate, the demand for robust and efficient security solutions becomes increasingly critical. In this study we explore the applications of traditional machine learning techniques in Network Anonaly Detection, aiming to enhance overall security.

Conventional rules based systems often struggle to detect sophisticated and evolving threats, motivating the adoption of machine learning techniques.

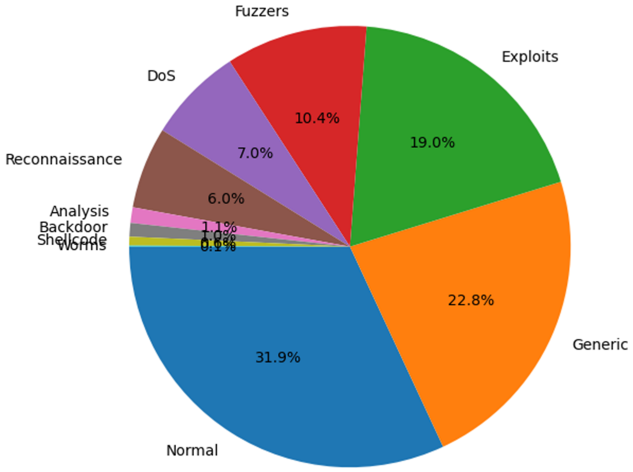
THE DATASET

The dataset used in this reaserarch is the UNSW-NB15 Dataset created by the University of New South Wales in 2015. It contains 2 milion plus network traffic records, which are divided into it's training and testing records. The dataset includes records of 9 diffrent types of atacks and their network properties.
Each record in the dataset contains 49 eatures that describe the traffic. The UNSW-NB15 dataset is a popular dataset for intrusion detection research. It is known for its diversity of attacks and its realistic representation ofmodern network traffic.
The labels of the dataset is divided into two and they are whether or not the traffic data is malicious (1) or benign(0). The pie chart below is a representation of the distribution of the dataset.

Distribution of Dataset Labels



Distribution of Attack Types

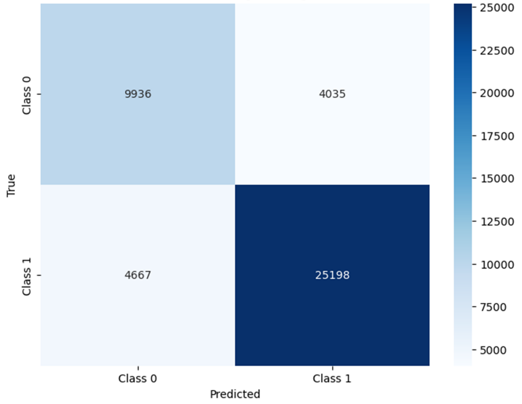


Futher analysis of the dataset can expose the behaviour of anomalous and benign traffic. It exposes patterns such as meanings to phenomenen like what high volumes of small data transfers mean, what sudden spikes in short lived connections and many more scenarios like that.

MODELS AND RESULTS

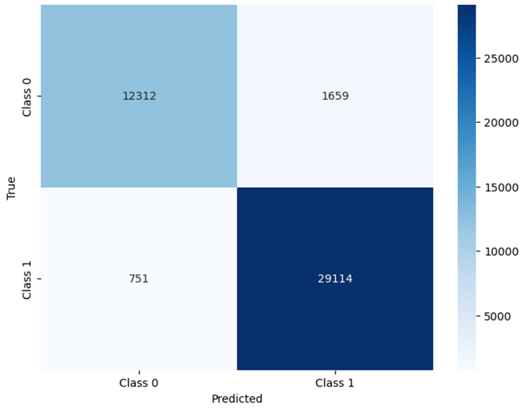
The Logistic Regression model performs with an accuracy of 0.80. When futher checks are done, it has an AUC for the precision-recall curve of 0.91 and the AUC for the ROC curve is 0.88

Confusion Matrix of the Logistic Regression Model



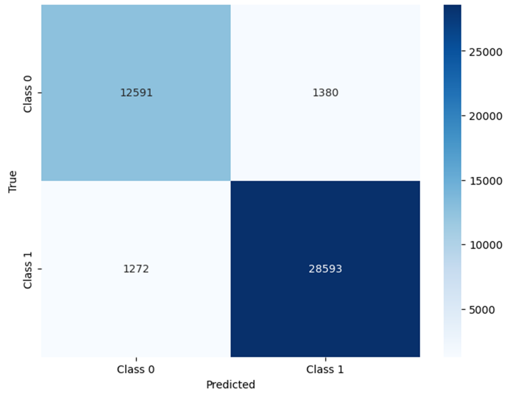
The CatBoost model performs with an accuracy of 0.94. When futher checks are done, it has an AUC for the precision-recall curve of 0.97 and the AUC for the ROC curve is 0.94

Confusion Matrix of the CatBoost Classifier



The Decision Trees model performs with an accuracy of 0.95. When futher checks are done, it has an AUC for the precision-recall curve of 1.00 and the AUC for the ROC curve is 0.99

Confusion Matrix of the Decision Trees Classification



Next Steps

- Train more models
- Test model performance in real life scenerio