Modeling Anomaly Detection in Network Traffic

Team CyberSleuth

Yusuf Afolabi, Joshua Adesina, & Oluwadamilola Salau

Project Overview

The goal of this project is to build a machine learning model that can detect anomalous behavior in network traffic. This model could be particularly useful in identifying potential security threats such as hacking attempts, malware infections, and other malicious activities.

Dataset

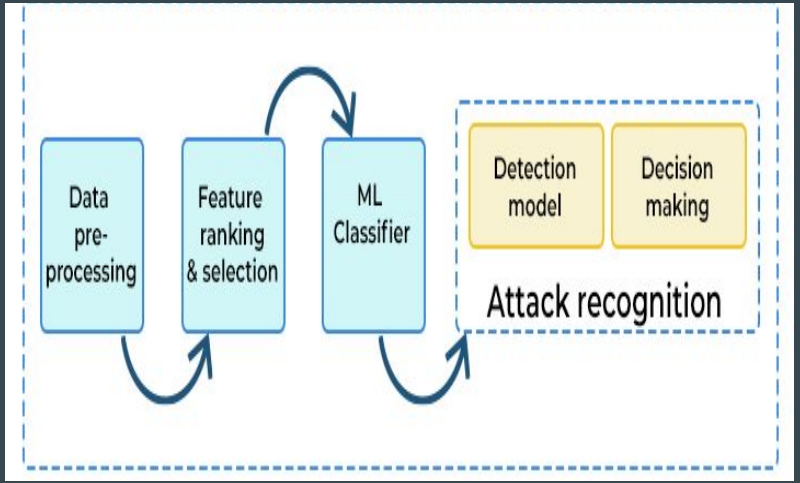
The dataset used is CICIDS2007 available on Kaggle (https://www.kaggle.com/datasets/cicdataset/cicids2017).

The dataset consists of eight files for a 5-day (Monday to Friday) data stream on a network with approximately 2.5 million network traffic records, including both benign and malicious traffic attacks, such as DoS, DDoS, port scanning, botnet, and infiltration attacks.

General Approach

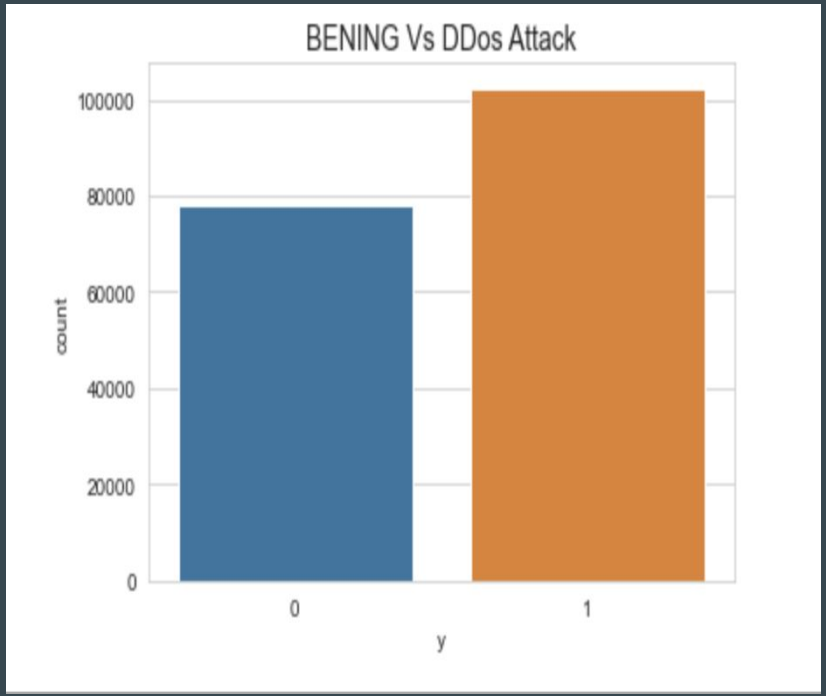
We consider two scenarios with balanced and imbalanced datasets: DDos & All attacks

- Data cleaning
- Train test split
- EDA
- Based model and metric
- Machine learning models
- Performance metric employed etc.
- Final model



Results for DDos Attack: balanced dataset

- Approach: train test split (20%) .
- Metric: accuracy_score.
- PCA decomposition for dimension reduction.
- Kfold validation set with k = 5.
- Logistic regression model achieved about 76% accuracy
- KNN model with tuning achieved about 99% accuracy.
- SVM with parameter tuning - 99%



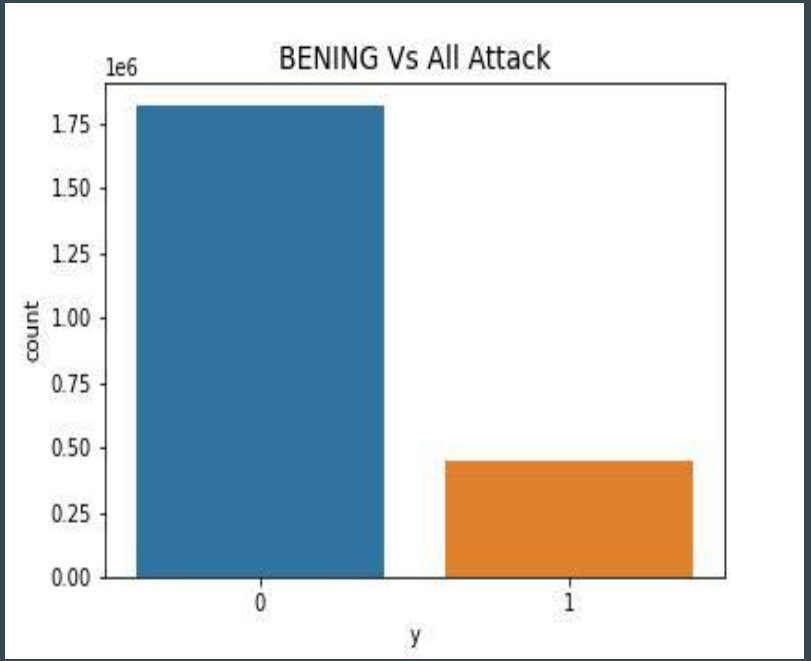
Results for DDos Attack: balanced dataset

- Random Forest with max depth 10 and 100 estimator performed best - via GridSearchCV method.
- Random Forest achieved about the same accuracy score on both the kfold validation set and the test set - 99%.
- The top 10 most important features

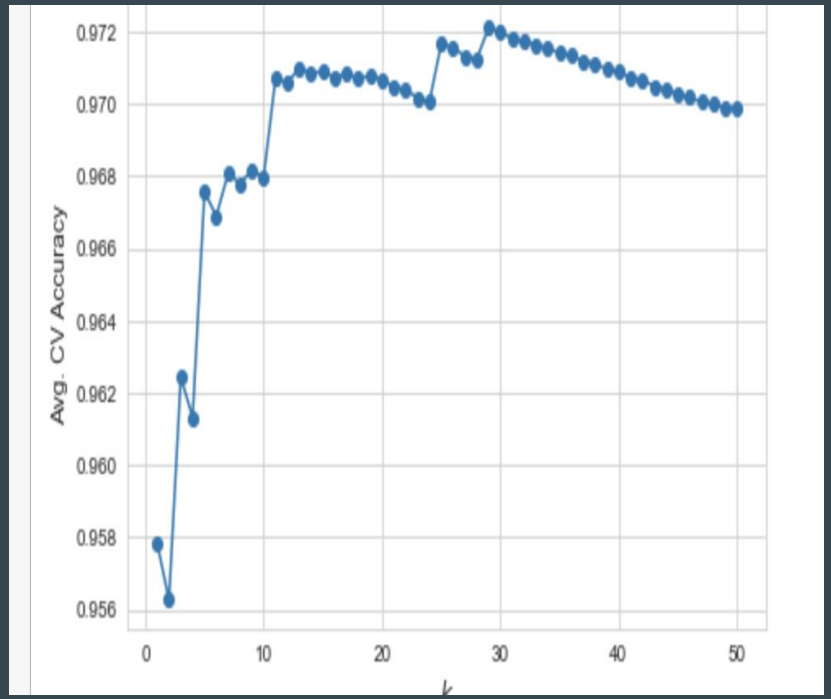
	feature_importance_score
Fwd Packet Length Max	0.126056
to scroll output; double click to hide	
Fwd Packet Length Mean	0.090432
Avg Fwd Segment Size	0.078203
Total Length of Fwd Packets	0.069720
act_data_pkt_fwd	0.055592
Fwd IAT Total	0.046403
Subflow Fwd Packets	0.046064
Subflow Fwd Bytes	0.044213
Fwd IAT Max	0.038387
Fwd IAT Std	0.037523

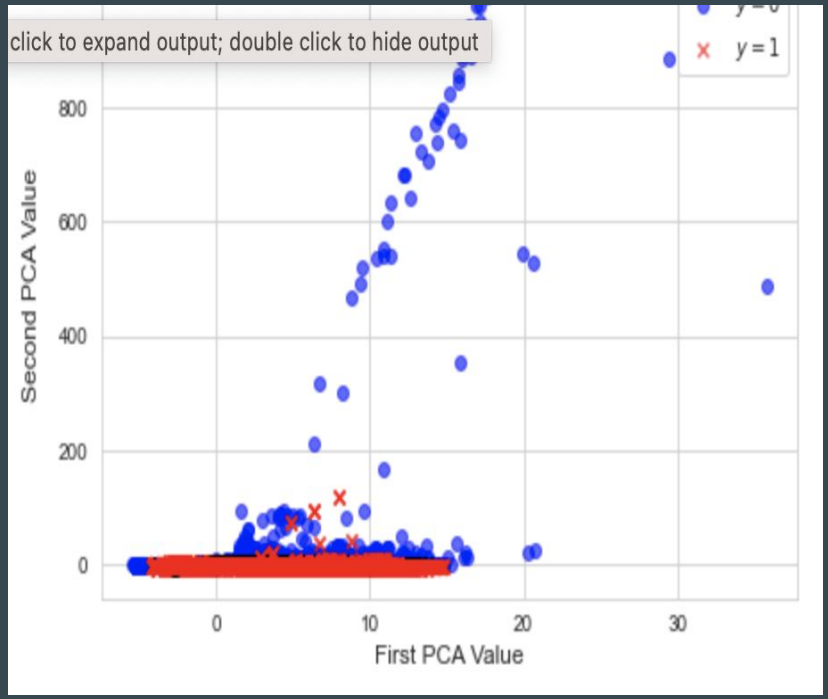
Results for All Attacks: highly imbalanced dataset

- Approach: resample (upsampling of the attack label)
- Metric: F1_score
- PCA decomposition for dimension reduction
- Kfold cross validation with k = 5
- We achieve a 97% accuracy with k = 30 for KNN model
- We tried other models as well but we are constraint by GPU and time.



Results for All Attacks: highly imbalanced dataset





Limitations

- The dataset used in this project was observed to be imbalanced, prompting the need for resampling to address the issue and enhance the dataset. However, it is important to note that this resampling process might have introduced data bias that was not specifically addressed within the scope of this project.
- We are constraint by the GPU availability and the time to try other models on ‘All attacks’ dataset.

Future Work

The primary objective of this project is to model network anomalies by comparing benign network behavior with attacks, without a specific emphasis on the individual types of attacks. As a result, future work will be dedicated to developing machine learning models that can effectively detect and classify specific types of network attacks. The focus will be on enhancing the understanding and prediction of the occurrence of distinct network attack types through advanced machine learning techniques.

References

1) CICataset. (2020, January 3). CICIDS2017. Kaggle. https://www.kaggle.com/datasets/cicdataset/cicids2017

2) Kostas, Kahraman. (2018). Anomaly Detection in Networks Using Machine Learning.