INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY, BANGALORE.

2021

A Project Report On

“UNSW-NB15 Dataset Intrusion Detection System”

Project Guide

Prof. Jyotsna Bapat

Report By

Shubham Jha (MT2020514)

Rayaan Ahmed Shariff (MT2020527)

Introduction

With the rise of Internet usage, it is very important to protect Networks. The most common risk to a network’s security is an intrusion such as brute force, denial of service or even an infiltration from within a network.

Evaluating NIDS(Network Intrusion Detection System)s using the existing benchmark data sets of KDD99 and NSLKDD does not reflect satisfactory results, due to three major issues: (1) their lack of modern low footprint attack styles, (2) their lack of modern normal traffic scenarios, and (3) a different distribution of training and testing sets. To address these issues, the UNSW-NB15 data set has recently been generated. This data set has nine types of the modern attacks fashions and new patterns of normal traffic, and it contains 49 attributes that comprise the flow based between hosts and the network packets inspection to discriminate between the observations, either normal or abnormal.

Certain ML techniques have been evaluated on the UNSW-NB15 dataset. The accuracy and false alarm rate of the techniques are assessed, and the results compared with previous research and our findings of the problem. The best machine learning techniques and flow identifiers of source/destination IP addresses and protocols can effectively and efficiently detect botnets and their origins as a Network forensic mechanism.

Proposed Solution

The IXIA PerfectStorm tool was utilized to generate a rich hybrid set of normal and abnormal modern network traffic. The IXIA tool proactively harvests and aggregates publicly known vulnerabilities and exposures relative to information systems security.

In order to conceptually understand the features and how they are related to our label of attack and non-attack category, network basics were seen. Performing a Network Flow Analysis, which involves capturing, collecting and logging network data, aggregating and analysis, greatly helps in building a better machine learning model. To analyse the data, deeper packet inspection is generally slower than flow-level analysis, which relies more on statistical properties of the network.

The vulnerabilities of Internet Protocols like HTTP and DNS are breached by attackers. DNS features and HTTP features form the basis of selecting important features, some of which are mentioned below:

- Source/Destination IP address i.e. srcip/dstip
- Source/Destination Port no. i.e. sport/dport
- Protocol i.e. proto
- Last time of connection i.e. ltime
- Domain name subject of the query
- List of resource description in answer of the query
- Length of query/answer
- Caching intervals of the answer
- Number of connections to the same source/destination in 100 records according to the last time i.e. ct_srv_src, ct_srv_dst
- Number of connections of the same source address/destination address and the destination port/ source port in 100 records according to the last time i.e. ct_src_dport_ltm, ct_dst_sport_ltm

These features are in sync with the important features for different models like XG Boost and LightGBM.

The dataset comprises 1,75,341 training examples and 82,332 testing cases for which label is to be predicted, both of them having 45 features. There are no null values in the dataset. In order to reduce the number of features, relations between different features were seen. Most correlated features are:

- sbytes, sloss
- dpkts, dbytes, dloss
- sttl, ct_state_ttl, label
- swin, dwin
- stime, dtime

- tcprtt, synack, ackdat
- ct_srv_src, ct_dst_src_ltm, ct_srv_dst
- ct_dst_ltm, ct_src_ltm, ct_src_dport_ltm, ct_dst_sport_ltm.

Now, there are 4 categorical features and the rest of them being integers. The range of some features is very high. Therefore, we standardize the features to a mean of 0 and standard deviation of 1 before fitting the machine learning models. After the process of Exploratory Data Analysis involving plotting countplot, pairplot,etc. , standardizing the features, dropping some features and using encoding, the dataset is ready to be fed to machine learning models. The machine learning models, used for the biased dataset containing almost 69% of the attack label in training, are XG Boost, Support Vector Machine, Logistic Regression and Light GBM.

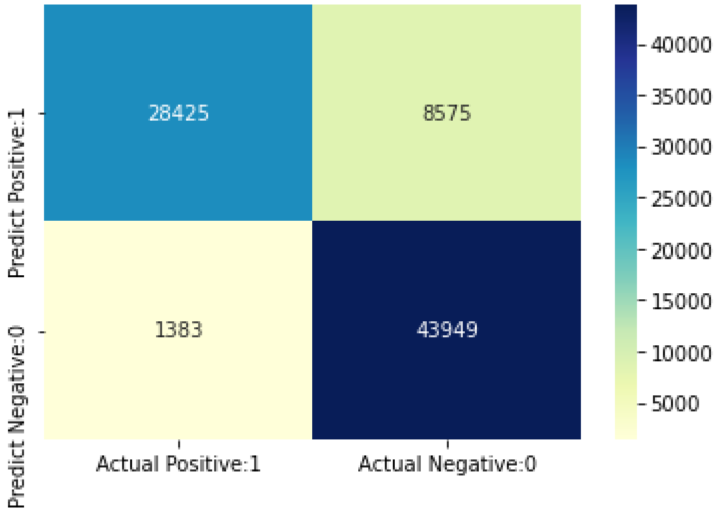
Results

The UNSW-NB 15 Dataset now needs to be used in training our machine learning models and we verify the result by checking the various metrics like F1 score and AUC score apart from the accuracy of our model as accuracy is not a good performance metric for an imbalanced dataset. Thus, various models were tried during the process to check which ones perform the best. A summary of different models along with their results is given below:

Model	Accuracy(%)	Precision	Recall	F1 score
XG Boost	87.33	0.89	0.87	0.87
Logistic Regression	81	0.85	0.81	0.80
SVM	81.57	0.86	0.82	0.81
Light GBM	87.7	0.89	0.88	0.87
Light GBM with Bayesian optimization	87.91	0.89	0.88	0.88

So, the best performing model was Light GBM with Bayesian optimization. **Bayesian Optimization** is a probabilistic model based approach for finding the minimum of any function that returns a real-value metric.Bayesian optimization utilizes the Bayesian technique of setting a prior over the objective function and combining it with evidence to get a posterior function.The prior belief is our belief in parameters before modeling process. The posterior belief is our belief in our parameters after observing the evidence. **LightGBM** is a gradient boosting framework that uses tree based learning algorithms. The advantages of LightGBM are faster training speed and higher accuracy, better accuracy, lower memory use,etc. LightGBM grows trees vertically rather than horizontally i.e. leaf wise growth.

Now, the figure below shows the times our LightGBM model got the correct results and wrong results using the confusion matrix.



Using the confusion matrix, we can calculate the false alarm rate as follows:

$$FPR = fp \div (fp + tn); FNR = fn \div (fn + tp)$$

$$False\ Alarm\ Rate = (FPR + FNR) \div 2$$

Using the above equations, we found the false alarm rate to be 0.13.

Also, the AUC score on the test data was found to be 0.9939 for this model.

Conclusion

For an Intrusion Detection System to work, we need to consider both the methodology for creating a comprehensive profile of the different patterns of an intrusion in the network as well as establishing an efficient and lightweight Decision Engine to the system that scales well with size and speed of traffic. This remained the main focus of the project, and in that direction, the UNSW-NB 15 dataset was used to train different machine learning models. The results considering the imbalanced nature of the dataset were studied and gives a direction to future works on this subject. This may lead to using machine learning models to achieve better results on metrics like False Alarm Rate, and F1 score. Also this problem was dealt with as a multivariate binary classification problem, with the two classes being attack and non-attack. But, a broader study could be done where the 9 categories of attack could be considered as a class in itself making it a multivariate multi-class classification problem, which will have its own challenge of generating a balanced dataset for each attack category for better results. As the internet of things architecture grow in size each day, the challenges of it being secure from intrusions like botnets, DDoS attacks, stealth attack, malware etc. is a big challenge as the attack pattern changes from time to time. With cross-dimensional study on cyber-security, internet of things and artificial intelligence, it is possible to make the system more attentive and responsive to these attacks.

References

[1] URL link to the official site for the dataset UNSW-NB15

https://research.unsw.edu.au/projects/unsw-nb15-data-set

[2] UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)

https://ieeexplore.ieee.org/abstract/document/7348942