Network Intrusion Detection System

Objective To build network intrusion detection system to detect anamolies and attacks in the network

Business Problem

- Binomial Classification: Activity is normal or attack
- Multinomial classification: Activity is normal or DOS or PROBE or R2L or U2R

Data Availability

- Basic features of each network connection (Duration, Protocol_type etc.)
- Content related features
- Time related traffic features
- Host based traffic features

Approach:

1 Importing and understanding the data

- Importing the Train and Test
- Seggregating the attack type variable (dependent variable)
  Based on the data dictionary Attack type is segreggated.

Attack Class	Attack Type
DoS	Back, Land, Neptune, Pod, Smurf,Teardrop,Apache2, Udpstorm, Processtable, Worm (10)
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint (6)
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named (16)
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps (7)

- 1. DOS: Denial of service is an attack category, which depletes the victim‟s resources thereby making it unable to handle legitimate requests –e.g. syn flooding. Relevant features: “source bytes” and “percentage of packets with errors”
- 2. Probing: Surveillance and other probing attack‟s objective is to gain information about the remote victim e.g. port scanning. Relevant features: “duration of connection” and “source bytes”
- 3. U2R: unauthorized access to local super user (root) privileges is an attack type, by which an attacker uses a normal account to login into a victim system and tries to gain root/administrator privileges by exploiting
- some vulnerability in the victim e.g. buffer overflow attacks. Relevant features: “number of file creations” and “number of shell prompts invoked,”
- 4. R2L: unauthorized access from a remote machine, the attacker intrudes into a remote machine and gains local access of the victim machine. E.g. password guessing Relevant features: Network level features – “duration of connection” and “service requested” and host level features - “number of failed login attempts”

2 Data audits

- Check for duplicate rows
- Check for missing values

3 Data preparation

- Encoding the attack type variable [Bi class] - Based on the data dictionary each attack type identified encoded as Normal or Anomaly

- Handling and one hot encoding of categorical variables [Bi class]

- Encoding the attack type variable [Multi class] - Based on the data dictionary each attack type identified encoded as Normal or the type of Anomaly
- Handling and one hot encoding of categorical variables [Multi class]

- Deriving X and Y of [Bi/Multi class]

4 Feature Selection

Considering only Xs (Common for both Bi and Multi class)

Since this type of feature reduction considers only independent variables it is common both the Bi amd Multi class problems

- VIF (Variance Inflation Factor) –
  VIF was performed on the X variables and it was found that 24 variables have VIF value below 5 which are suitable for model building.
- PCA (Principal Component Analysis)
  By PCA, 24 PCs were able to explain 80% of the variance of all Xs. Factor loading matrix was created with the 24 PCs and the variables having high correlation with each PCs were selected.

Considering X and Binomial Y

This section is only for the Binomial classification. Where Y = Normal or Anamoly

- Correlation
  Correlation matrix was created between X and Y. It was found that 12 variables were highly correlated with Y.
- F Regression
- RFE

The above methods were performed around 10 variables were selected from each method.

The selected variables were combined which forms the X.
The Train and Test split was made.
As the 1% and 0% was almost equal there was no need for Over/Under sampling.

5 Model Implementation + Hyper parameter tuning + Evaluation

5.1 Part I Binomial Classification (Y = Normal or Anamoly)

- Random Forest
  Best parameter : {'max_depth': 6, 'n_estimators': 50}
  Accuracy train : 0.99
  Accuracy test : 0.82

- XGBoost
  Best parameter : {'max_depth': 8, 'n_estimators': 50}
  Accuracy train : 0.99
  Accuracy test : 0.81

Conclusion on Binomial Classification: Both the models Random Forest and XGBoost was performing good. But the Random Forest model took only 1/5 th of the time for execution comparatively. *Random Forest model* can be chosen as the final model for predicting if there is any Anamoly in the Network

5.2 Part II Multinomial Classification (Y = normal or DOS or PROBE or R2L or U2R)

- Preparing Y multi for skmultilearn
  Since the skmultilearn algorithm requires the multi Y as dummified array it is one hot encoding was performed.

- Preparing Y multi for sklearn
  Here the Y is single dimensional array with encoded values for different types.

Feature Reduction considering both X and Y multi

- Decision Tree feature importance
  By performing the Decision Tree classification on the X and Y multi we abled to get features which have high relation with the Y multi.

Model Implementation

Scikit multi learn algorithms

- Binary Relevance
- Classifier chains

The above was not performing good for which the accuracy and other metrics poor.

Scikit multi learn algorithms

- Random Forest
  Best parameter : {'max_depth': 8, 'n_estimators': 50}
  Accuracy train : 0.99
  Accuracy test : 0.75

- XGBoost
  Best parameter : {'max_depth': 8, 'n_estimators': 50}
  Accuracy train : 0.99
  Accuracy test : 0.76

- Naive Bayes
  Accuracy train : 0.38
  Accuracy test : 0.27

Conclusion for Multi Classification : From all the above models Random Forest and XGBoost perform better than other models giving an *average test accuracy of 75%*. Since the Random Forest model is more time efficient, we will choose the *Random Forest* as the final model for this dataset.