**Project 2) -Case Study Networking Network Intrusion Detection**

**Objective** – the objective of theIntrusion Detection Systems (IDSs), as special-purpose devices to detect anomalies and attacks in the network

**Business Problem** –

Your task to build network intrusion detection system to detect anamolies and attacks in the

network. There are two problems.

1. Binomial Classification: Activity is normal or attack

2. Multinomial classification: Activity is normal or DOS or PROBE or R2L or U2R

Please note that, currently the dependent variable (target variable) is not defined explicitly.

However, you can use attack variable to define the target variable as required.

**Data Availability** –

This data is KDDCUP’99 data set, which is widely used as one of the few publicly available data sets for network-based anomaly detection systems.

**ASIC FEATURES OF EACH NETWORK CONNECTION VECTOR**

**1 Duration:** Length of time duration of the connection

**2 Protocol\_type:** Protocol used in the connection

**3 Service:**Destination network service used

**4 Flag:** Status of the connection –Normal or Error

**5 Src\_bytes:** Number of data bytes transferred from source to destination in single connection

**6 Dst\_bytes:** Number of data bytes transferred from destination to source in single connection

**7 Land:** if source and destination IP addresses and port numbers are equal then, this variable takes value 1

else 0

**8 Wrong\_fragment:** Total number of wrong fragments in this connection

**9 Urgent:**Number of urgent packets in this connection. Urgent packets are packets with the urgent bit

activated

**CONTENT RELATED FEATURES OF EACH NETWORK CONNECTION VECTOR**

**10 Hot:** Number of „hot‟ indicators in the content such as: entering a system directory, creating programs

and executing programs

**11 Num\_failed \_logins:** Count of failed login attempts

**12 Logged\_in Login Status**: 1 if successfully logged in; 0 otherwise

**13 Num\_compromised**:Number of ``compromised' ' conditions

**14 Root\_shell:**1 if root shell is obtained; 0 otherwise

**15 Su\_attempted:** 1 if ``su root'' command attempted or used; 0 otherwise

**16 Num\_root:** Number of ``root'' accesses or number of operations performed as a root in the connection

**17 Num\_file\_creations:**Number of file creation operations in the connection

**18 Num\_shells:**Number of shell prompts

**19 Num\_access\_files**:Number of operations on access control files

**20 Num\_outbound\_cmds:** Number of outbound commands in an ftp session

**21 Is\_hot\_login:** 1 if the login belongs to the ``hot'' list i.e., root or admin; else 0

**22 Is\_guest\_login:**1 if the login is a ``guest'' login; 0 otherwise

**TIME RELATED TRAFFIC FEATURES OF EACH NETWORK CONNECTION VECTOR**

**23 Count:** Number of connections to the same destination host as the current connection in the past two

seconds

**24 Srv\_count:** Number of connections to the same service (port number) as the current connection in the

past two seconds

**25 Serror\_rate:** The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the

connections aggregated in count (23)

**26 Srv\_serror\_rate:** The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among

the connections aggregated in srv\_count (24)

**27 Rerror\_rate:** The percentage of connections that have activated the flag (4) REJ, among the connections

aggregated in count (23)

**28 Srv\_rerror\_rate:** The percentage of connections that have activated the flag (4) REJ, among the

connections aggregated in srv\_count (24)

**29 Same\_srv\_rate:** The percentage of connections that were to the same service, among the connections

aggregated in count (23)

**30 Diff\_srv\_rate:** The percentage of connections that were to different services, among the connections

aggregated in count (23)

AnalytixLabs, Website: www.analytixlabs.co.in Email: info@analytixlabs.co.in phone: +91-88021-73069**31 Srv\_diff\_host\_ rate:** The percentage of connections that were to different destination machines among

the connections aggregated in srv\_count (24)

**HOST BASED TRAFFIC FEATURES IN A NETWORK CONNECTION VECTOR**

**32 Dst\_host\_count:** Number of connections having the same destination host IP address

**33 Dst\_host\_srv\_ count:** Number of connections having the same port number

**34 Dst\_host\_same \_srv\_rate:** The percentage of connections that were to the same service, among the

connections aggregated in dst\_host\_count (32)

**35 Dst\_host\_diff\_ srv\_rate:** The percentage of connections that were to different services, among the

connections aggregated in dst\_host\_count (32)

**36 Dst\_host\_same \_src\_port\_rate**: The percentage of connections that were to the same source port,

among the connections aggregated in dst\_host\_srv\_c ount (33)

**37 Dst\_host\_srv\_ diff\_host\_rate:** The percentage of connections that were to different destination

machines, among the connections aggregated in dst\_host\_srv\_count (33)

**38 Dst\_host\_serro r\_rate: The** percentage of connections that have activated the flag (4) s0, s1, s2 or s3,

among the connections aggregated in dst\_host\_count (32)

**39 Dst\_host\_srv\_s error\_rate:** The percent of connections that have activated the flag (4) s0, s1, s2 or s3,

among the connections aggregated in dst\_host\_srv\_c ount (33)

**40 Dst\_host\_rerro r\_rate**: The percentage of connections that have activated the flag (4) REJ, among the

connections aggregated in dst\_host\_count (32)

**41 Dst\_host\_srv\_r error\_rate:** The percentage of connections that have activated the flag (4) REJ, among

the connections aggregated in dst\_host\_srv\_c ount (33)

**Historical Data** –

* Using 125973 train data of the system and 22544 as test data to validate
* It had around 42 Features

**Tools Used –** Python (numpy, pandas, sklearn,scipy,matplotlib,seaborn,xgboost,statsmodels,etc.), Excel

**Techniques –**

* Used techniques like Logistic Regression,Random Forest Classifier and XGBClassifier
* Used **Classification** to find out if system is attacked or normal
* For dataset’s Graphical representation and insights used **Matplotlib and Seaborn**
* For exploratory analysis I used all **statistical methods like Univariate Analysis, Bivariate analysis, Hypothetical Testing, Correlation.**

**Metrices**-

* For classification I used AUC as a metrics, and it is best metric to accuracy. If AUC value is high our model accuracy is good.
* For traditional approach used AUC and GINI as a metrics and it is best metric to accuracy. If AUC and GINI value is high our model accuracy is good.

**Tuning Parameters –**

For this system intrusion detection problem we used without tuning parameters in all Models

**Validation –**

* It will train my model using train data set I will verify result against test data set. If we will get GINI value nearly then I will say my model is good model.

**Challenges** –

* Data Collection
* Feature Selection
* Identify the Positive Drivers