Predicting Cyber Attacks on Networks

Joshua Joseph

Problem Statement

- The number of cyber attacks on system networks have dramatically increased over the past several years. As a result, there is a growing need to better understand and mitigate these threats to business and consumers alike

Dataset

- The dataset simulates a military's network
- Simulated US Air Force LAN with various attacks
- To create the dataset an environment that acquired raw TCP/IP packets were used
- Consists of Training and Testing sets
- Date created 2018-10-09
- Derived from Kaggle
- Test set contains
  - 22,544 rows & 41 columns
- Training set contains
  - 25,192 rows & 42 columns
- Link: https://www.kaggle.com/sampadab17/network-intrusion-detection

Analyzing the data (Training Set)

General statistics of dataset

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromised	...	dst_host_count	dst_host_srv_count
count	25192.000000	2.519200e+04	2.519200e+04	25192.000000	25192.000000	25192.00000	25192.000000	25192.000000	25192.000000	25192.000000	...	25192.000000	25192.000000
mean	305.054104	2.433063e+04	3.491847e+03	0.000079	0.023738	0.00004	0.198039	0.001191	0.394768	0.227850	...	182.532074	115.063036
std	2686.555640	2.410805e+06	8.883072e+04	0.008910	0.260221	0.00630	2.154202	0.045418	0.488811	10.417352	...	98.993895	110.646850
min	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000
25%	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	...	84.000000	10.000000
50%	0.000000	4.400000e+01	0.000000e+00	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	...	255.000000	61.000000
75%	0.000000	2.790000e+02	5.302500e+02	0.000000	0.000000	0.00000	0.000000	0.000000	1.000000	0.000000	...	255.000000	255.000000
max	42862.000000	3.817091e+08	5.151385e+06	1.000000	3.000000	1.00000	77.000000	4.000000	1.000000	884.000000	...	255.000000	255.000000

8 rows × 38 columns

dst_host_same_srv_rate	dst_host_diff_srv_rate	dst_host_same_src_port_rate	dst_host_srv_diff_host_rate	dst_host_serror_rate	dst_host_srv_serror_rate	dst_host_rerror_rate	dst_host_srv_rerror_rate
25192.000000	25192.000000	25192.000000	25192.000000	25192.000000	25192.000000	25192.000000	25192.000000
0.519791	0.082539	0.147453	0.031844	0.285800	0.279846	0.117800	0.118769
0.448944	0.187191	0.308367	0.110575	0.445316	0.446075	0.305869	0.317333
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.050000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.510000	0.030000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.070000	0.060000	0.020000	1.000000	1.000000	0.000000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Unique values

duration	758
protocol_type	3
service	66
flag	11
src_bytes	1665
dst_bytes	3922
land	2
wrong_fragment	3
urgent	2
hot	22
num_failed_logins	5
logged_in	2
num_compromised	28
root_shell	2
su_attempted	3
num_root	28
num_file_creations	20
num_shells	2
num_access_files	7
num_outbound_cmds	1
is_host_login	1
is_guest_login	2
count	466
srv_count	414
serror_rate	70
srv_serror_rate	56
rerror_rate	72
srv_rerror_rate	42
same_srv_rate	97
diff_srv_rate	79
srv_diff_host_rate	57
dst_host_count	256
dst_host_srv_count	256
dst_host_same_srv_rate	101
dst_host_diff_srv_rate	101
dst_host_same_src_port_rate	101
dst_host_srv_diff_host_rate	63
dst_host_serror_rate	100
dst_host_srv_serror_rate	88
dst_host_rerror_rate	101
dst_host_srv_rerror_rate	100
class	2

dtype: int64

Analyzing the data (Testing Set)

General statistics of dataset

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_compromised	...	dst_host_count	dst_host_srv_count
count	22544.000000	2.254400e+04	2.254400e+04	22544.000000	22544.000000	22544.000000	22544.000000	22544.000000	22544.000000	22544.000000	...	22544.000000	22544.000000
mean	218.859076	1.039545e+04	2.056019e+03	0.000311	0.008428	0.000710	0.105394	0.021647	0.442202	0.119899	...	193.869411	140.750532
std	1407.176612	4.727864e+05	2.121930e+04	0.017619	0.142599	0.036473	0.928428	0.150328	0.496659	7.269597	...	94.035663	111.783972
min	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000
25%	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	121.000000	15.000000
50%	0.000000	5.400000e+01	4.600000e+01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	255.000000	168.000000
75%	0.000000	2.870000e+02	6.010000e+02	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	...	255.000000	255.000000
max	57715.000000	6.282565e+07	1.345927e+06	1.000000	3.000000	3.000000	101.000000	4.000000	1.000000	796.000000	...	255.000000	255.000000

8 rows × 38 columns

dst_host_same_srv_rate	dst_host_diff_srv_rate	dst_host_same_src_port_rate	dst_host_srv_diff_host_rate	dst_host_serror_rate	dst_host_srv_serror_rate	dst_host_rerror_rate	dst_host_srv_rerror_rate
22544.000000	22544.000000	22544.000000	22544.000000	22544.000000	22544.000000	22544.000000	22544.000000
0.608722	0.090540	0.132261	0.019638	0.097814	0.099426	0.233385	0.226683
0.435688	0.220717	0.306268	0.085394	0.273139	0.281866	0.387229	0.400875
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.070000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.920000	0.010000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.060000	0.030000	0.010000	0.000000	0.000000	0.360000	0.170000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Unique values

duration	624
protocol_type	3
service	64
flag	11
src_bytes	1149
dst_bytes	3650
land	2
wrong_fragment	3
urgent	4
hot	16
num_failed_logins	5
logged_in	2
num_compromised	23
root_shell	2
su_attempted	3
num_root	20
num_file_creations	9
num_shells	4
num_access_files	5
num_outbound_cmds	1
is_host_login	2
is_guest_login	2
count	495
srv_count	457
serror_rate	88
srv_serror_rate	82
rerror_rate	90
srv_rerror_rate	93
same_srv_rate	75
diff_srv_rate	99
srv_diff_host_rate	84
dst_host_count	256
dst_host_srv_count	256
dst_host_same_srv_rate	101
dst_host_diff_srv_rate	101
dst_host_same_src_port_rate	101
dst_host_srv_diff_host_rate	58
dst_host_serror_rate	99
dst_host_srv_serror_rate	101
dst_host_rerror_rate	101
dst_host_srv_rerror_rate	100

dtype: int64

Proposed Solution

- Analyze relationships between attributes to find positive relationships
- Sort and visualize most relevant attributes
- Possibly fit a machine learning model on training set to predict threats and then test the model's accuracy on the testing set