Network Intrusions

By Judah Drelich

Introduction

- Global cyber crime costs are expected to grow 15% per year for the next 5 years reaching $10.5 trillion by 2025

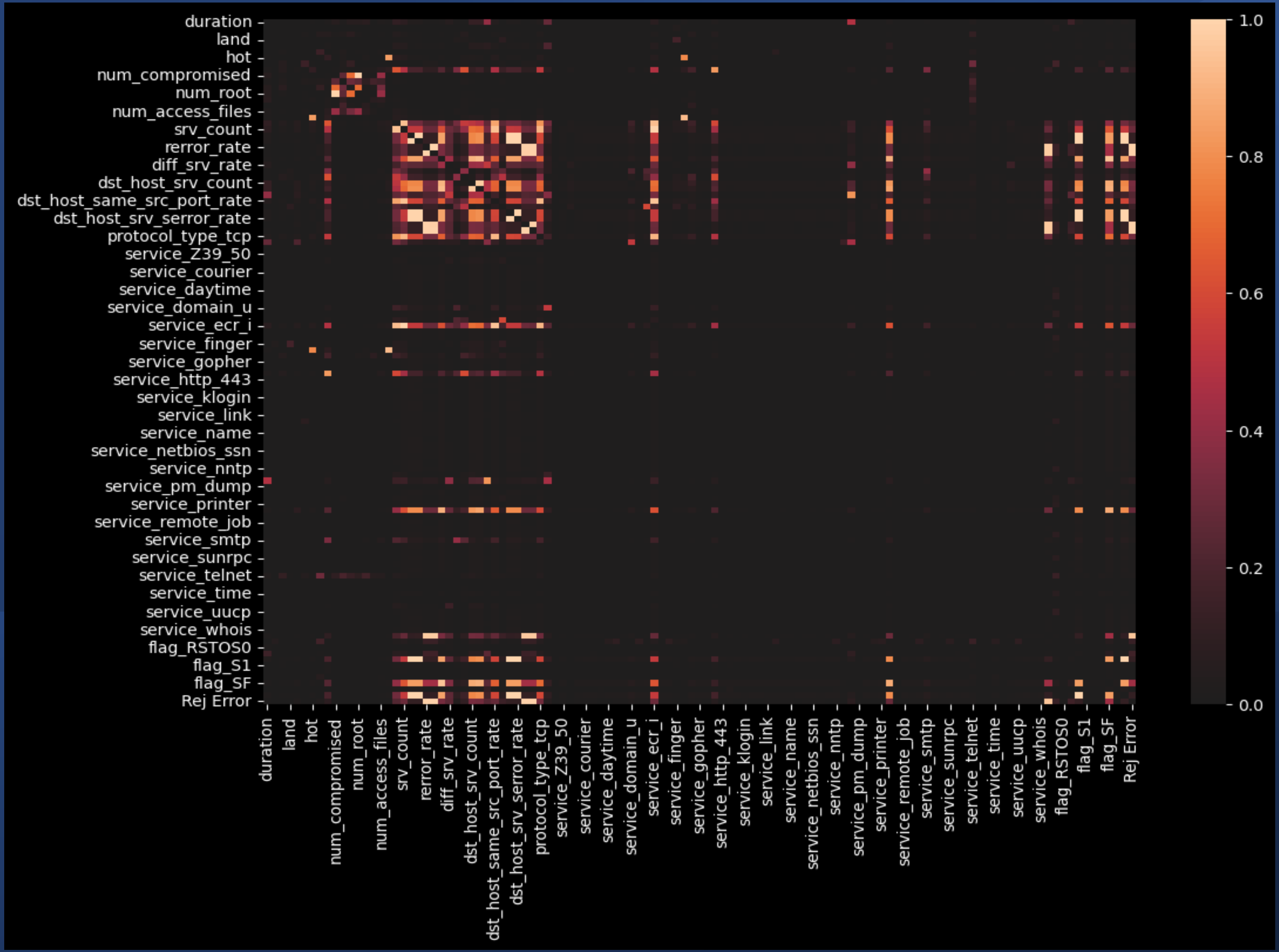
- I analyzed the 1998 DARPA Intrusion Detection Program's dataset that was managed and prepared by MIT's Lincoln Labs

- Training data has ~5 million observations, 41 features
- Test data has 300,000 observations, with the same features.

Problem Statement and Data

Data was taken from the 1999 KDD Cup
[KDD-CUP-99 Task Description (uci.edu)]

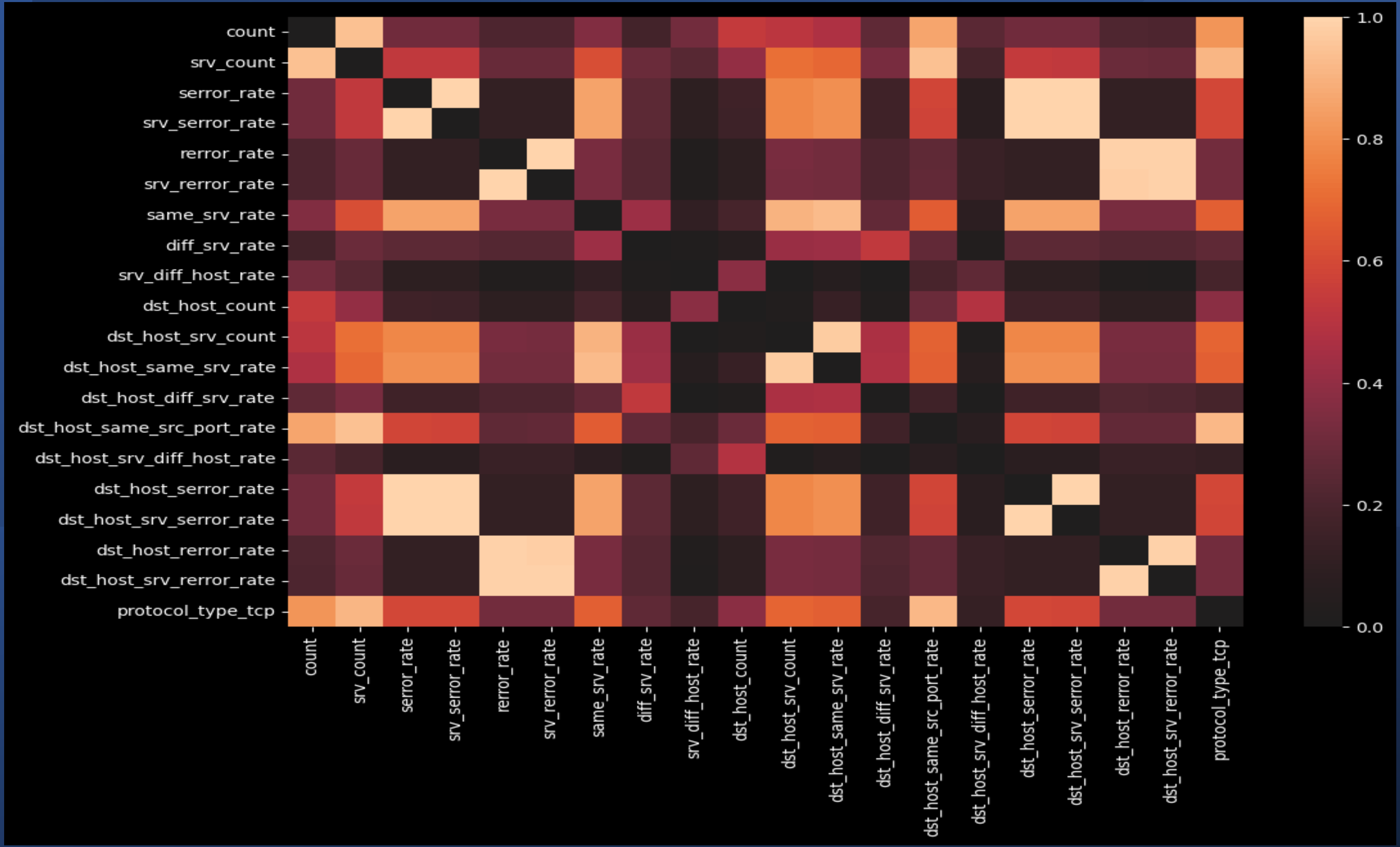
Data Wrangling



Multicollinearity

Initial heatmap for the correlations between the features

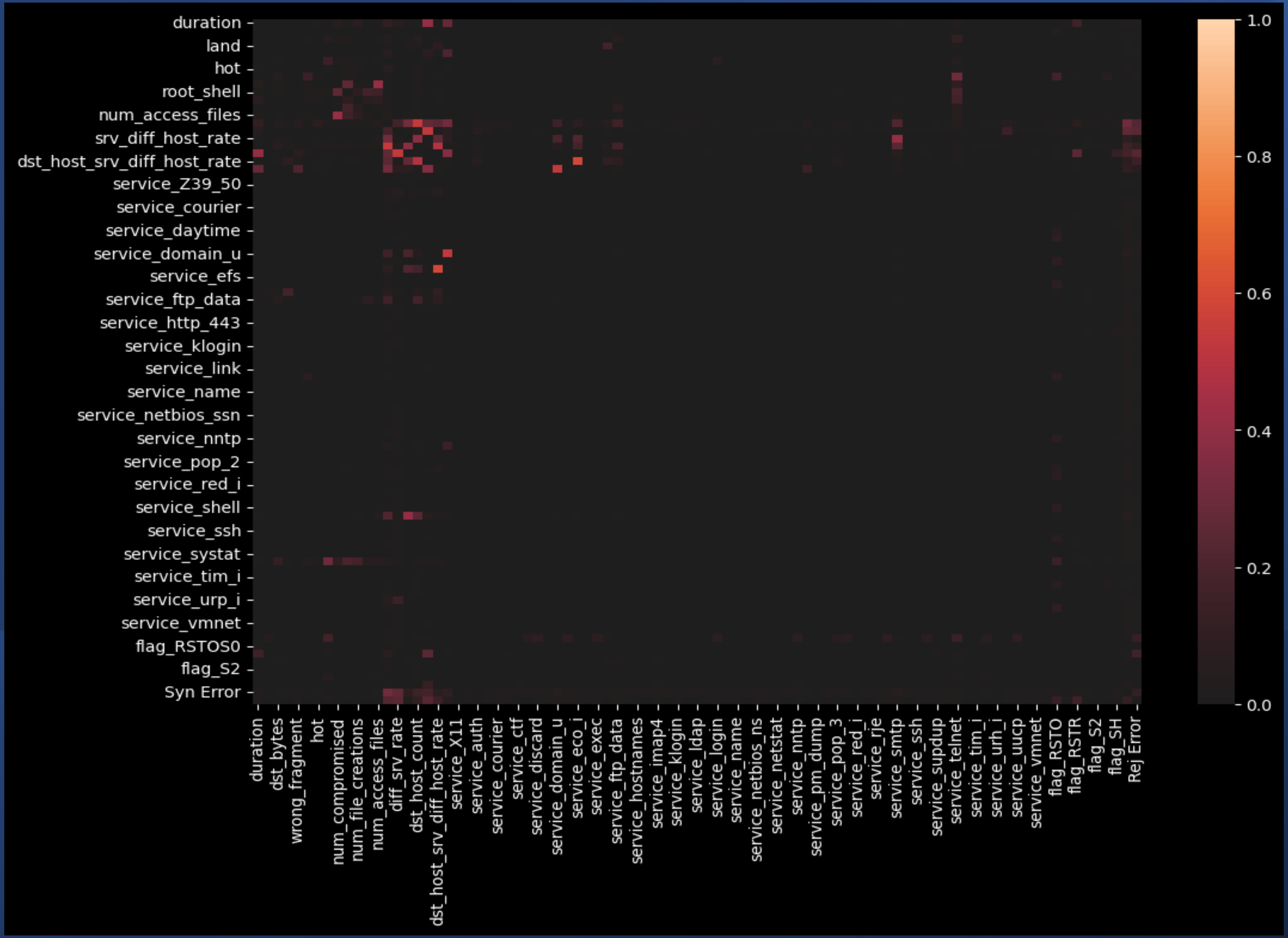
Data Wrangling



Multicollinearity

Heatmap that is focused on the most collinear features

Data Wrangling



Multicollinearity

Final heatmap after all the collinear columns have been dropped or combined.

Data Wrangling

- Lasso Regularization is a technique for variable selection that uses linear regression to evaluate the effect that each features has on a target variable.

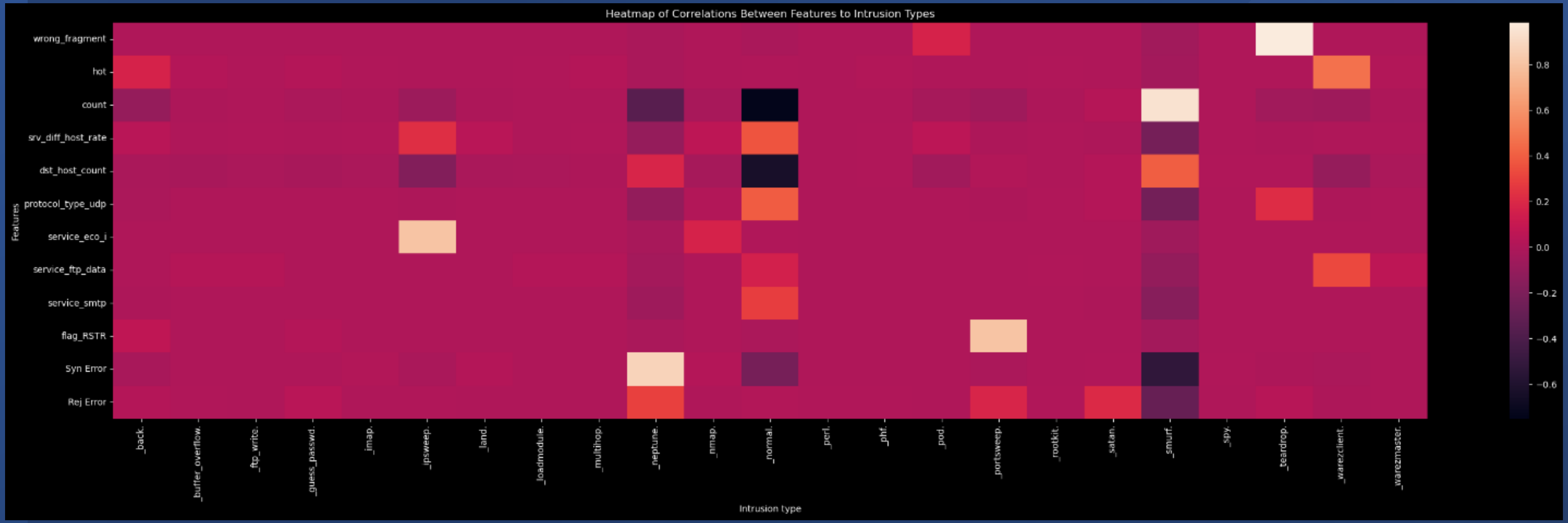
$$Loss(\beta_1, \ldots \beta_n) = SSD + \alpha \sum_{i=1}^{n} |\beta_i|$$

- I used Lasso Regularization to create to datasets
  - Small: 12 features, binary intrusion or not
  - Big: 61 features, multi class model, determines intrusion types

Lasso Regularization

Feature Selection

EDA

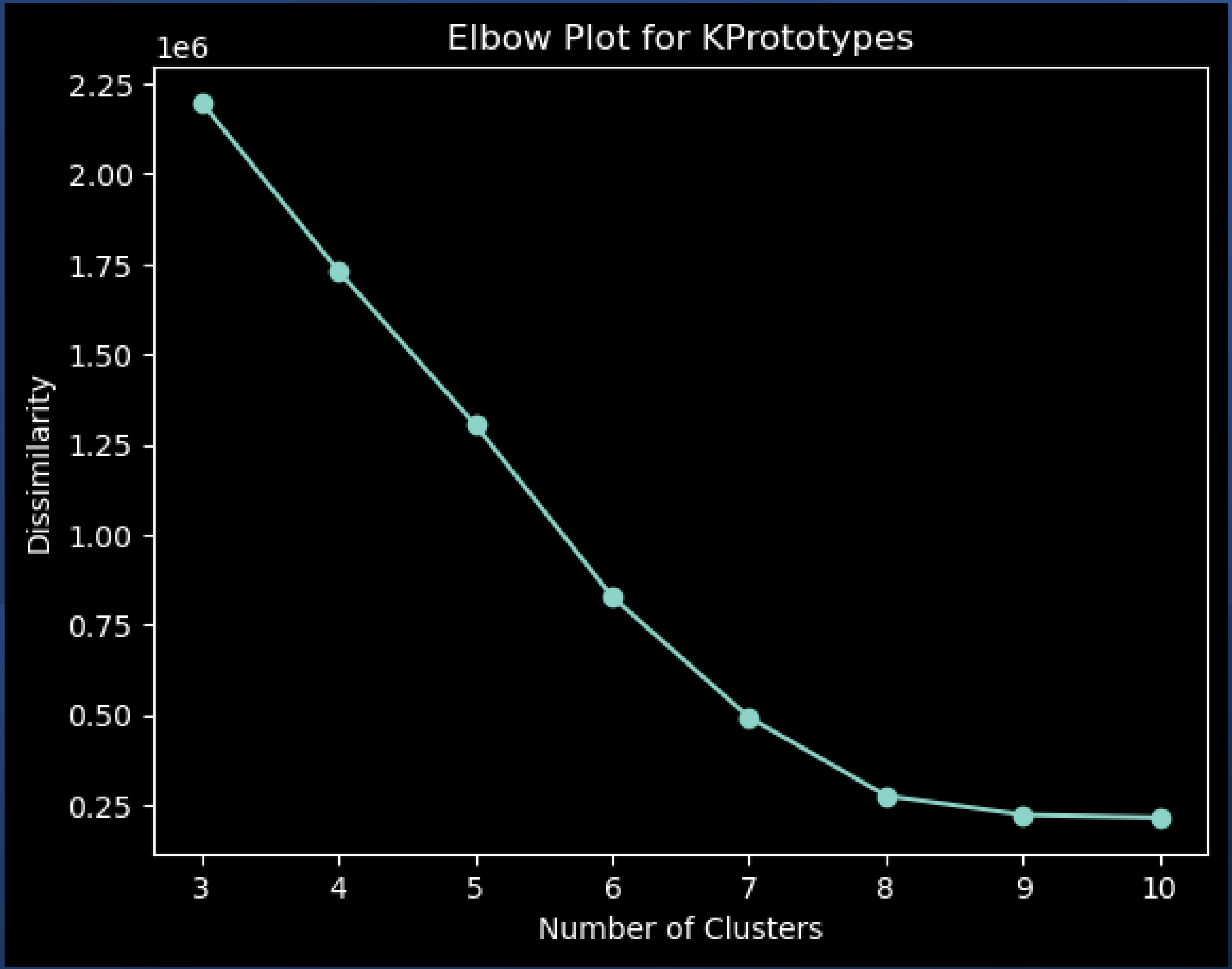


Correlation heatmap between features and intrusion types

Smurf.	56.8378%
Neptune.	21.6997%
Normal.	19.6909%

Intrusion Frequency

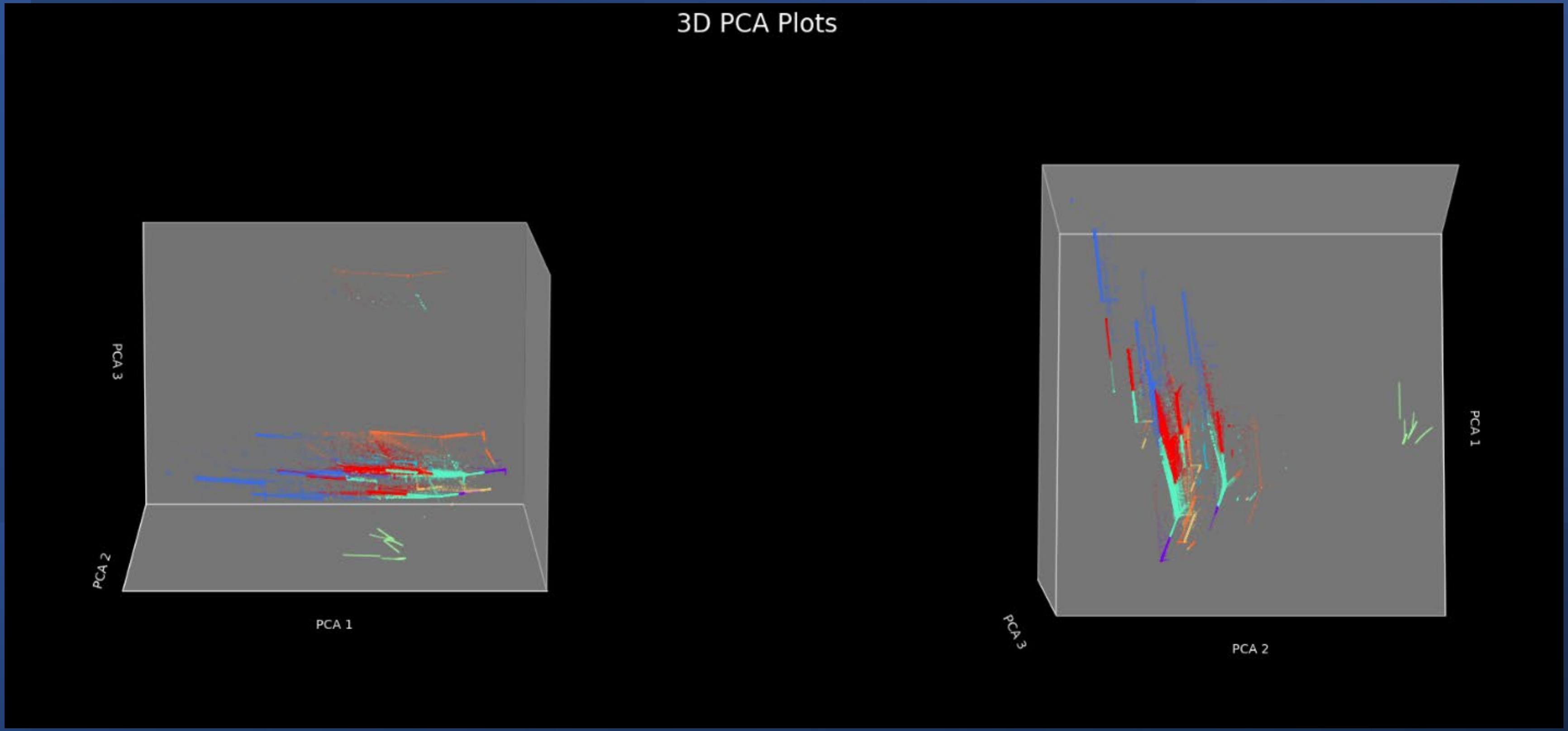
EDA



Clustering

Elbow Plot For K-Prototypes

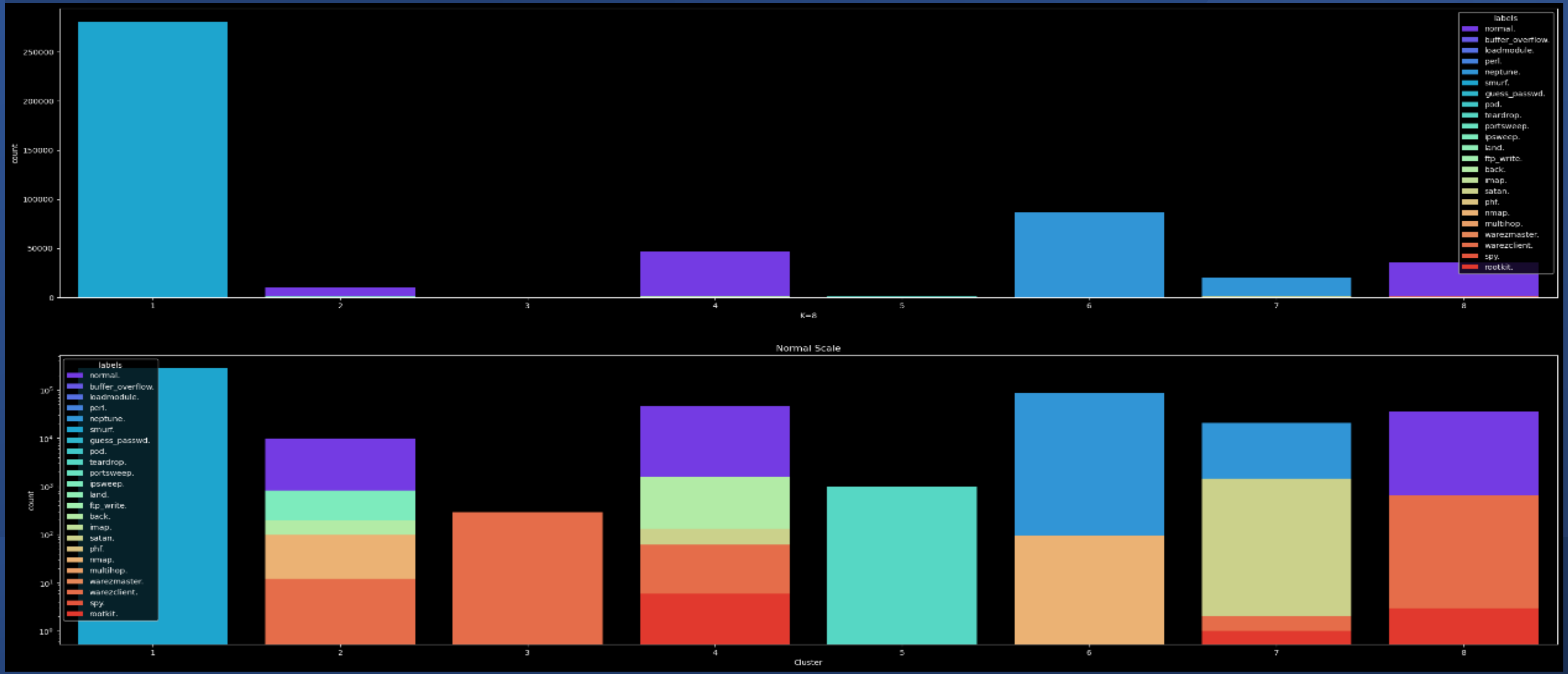
EDA



Principal Component Analysis (PCA)

3-D PCA for the dataset to try and visualize the structure of the data.

EDA



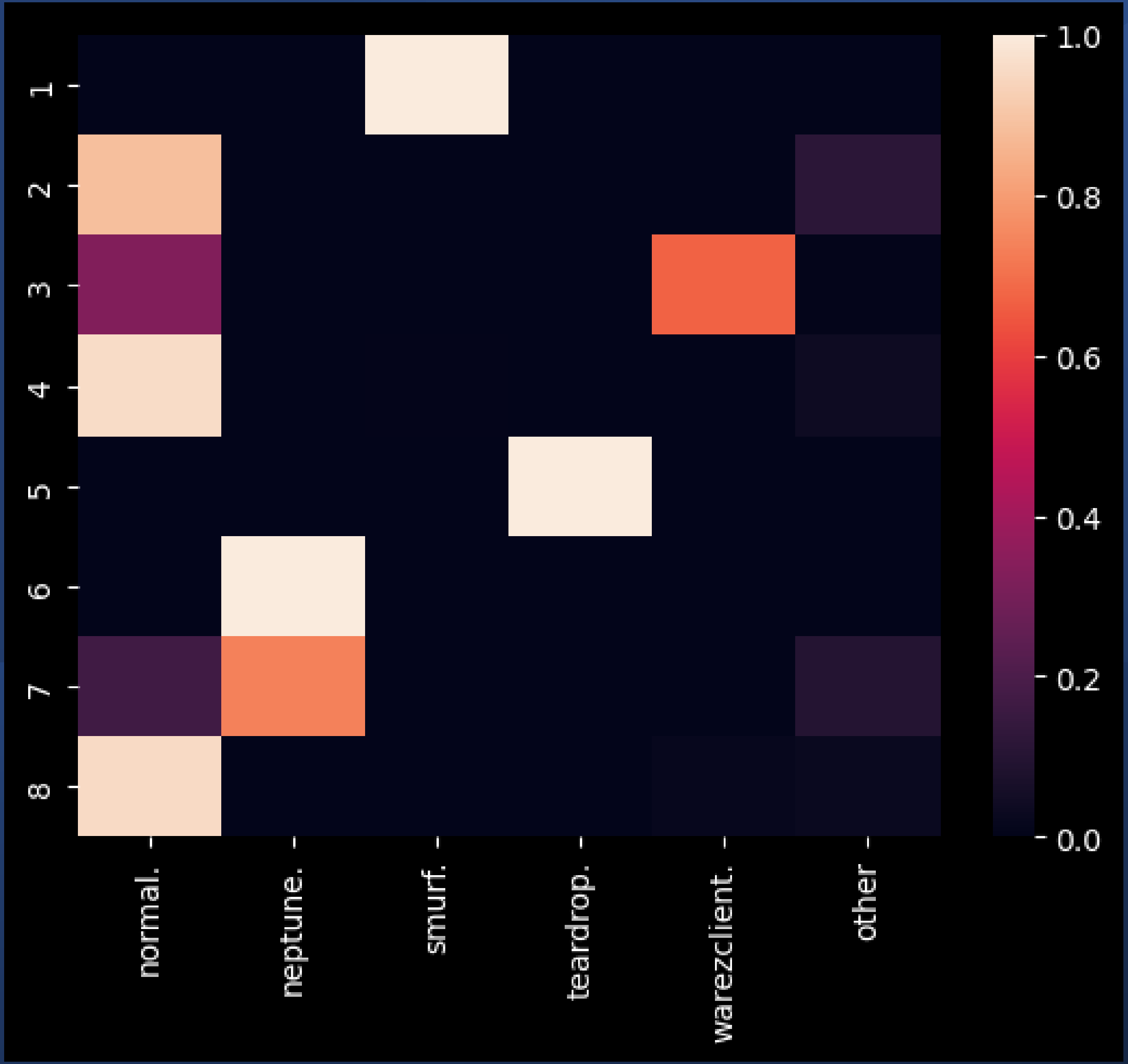
Histogram of the Clusters colored by intrusion type.

Cluster Analysis

Linear y-axis

Log y-axis

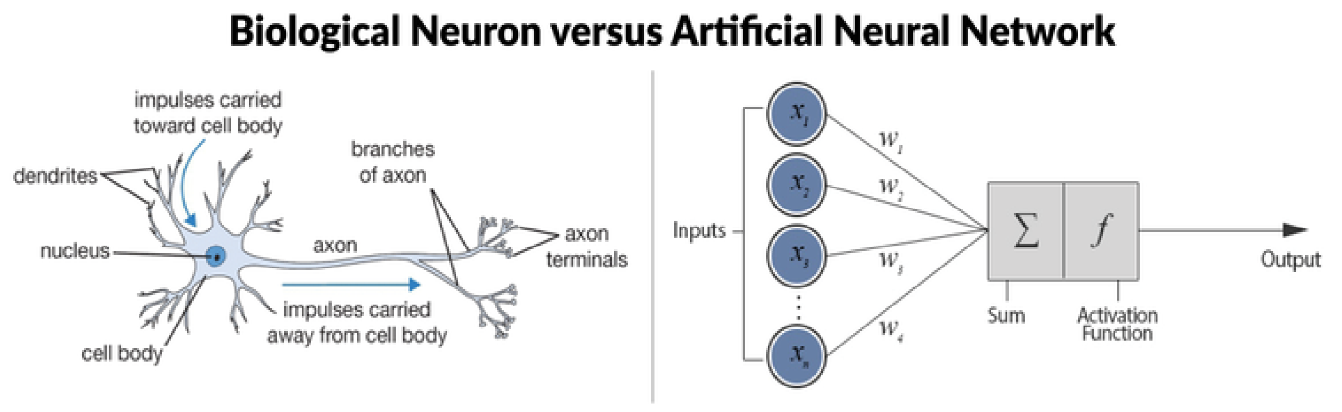
EDA



Cluster Analysis

Heatmap of the odds that a given intrusion will be found in each cluster. Rarer types are grouped as other.

Modelling



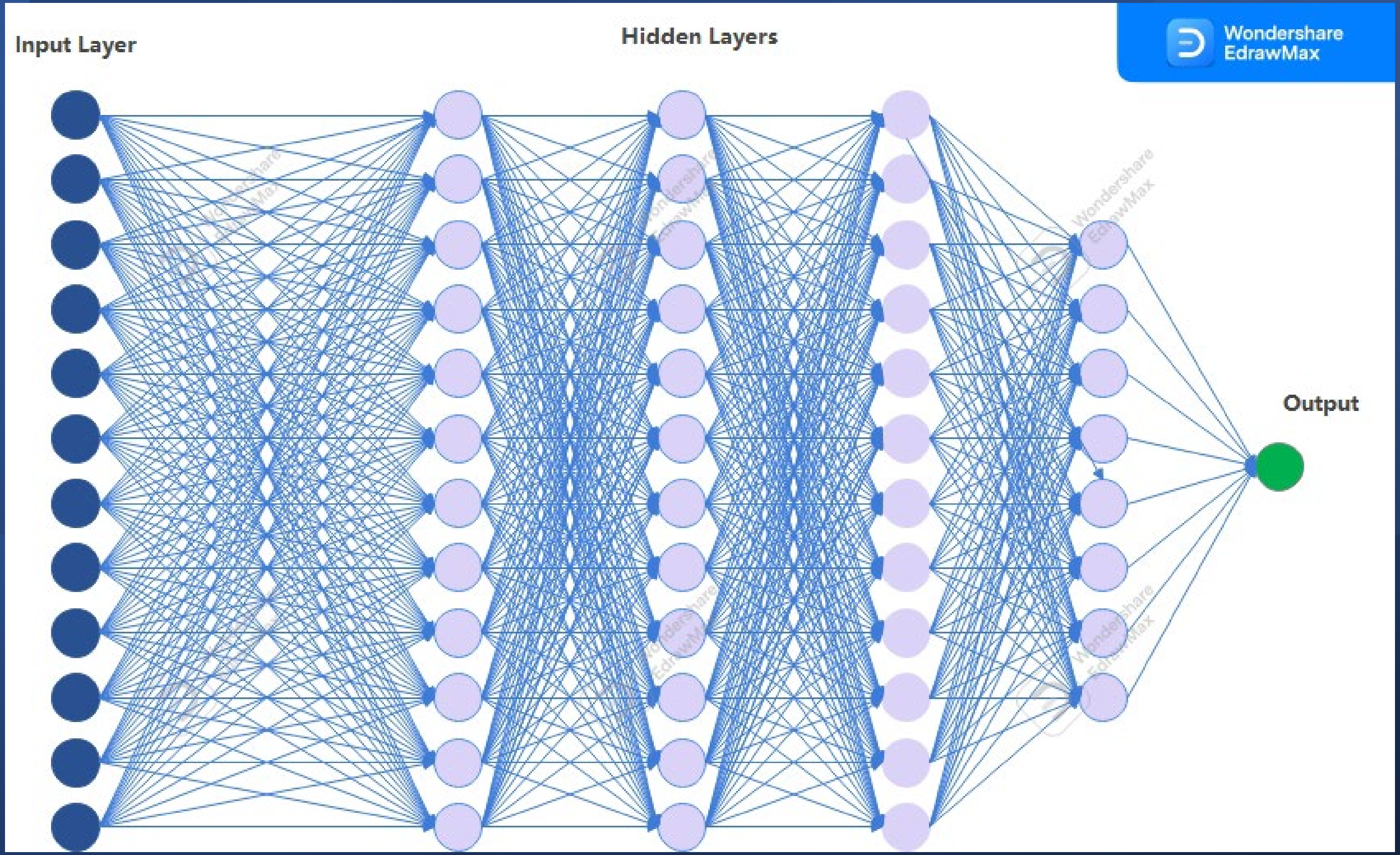
$$Output = f\left(\sum_1^n w_i x_i\right) + b$$

Neural Networks

Diagram of the similarity between biological neuron and a node in a neural network

Formula for a node in a neural network

Modelling



Neural Network

Graphical representation of the model that I used

Modelling

	precision	recall	f1-score	support
0	1.00	1.00	1.00	243103
1	1.00	1.00	1.00	981505
accuracy			1.00	1224608
macro avg	1.00	1.00	1.00	1224608
weighted avg	1.00	1.00	1.00	1224608

	precision	recall	f1-score	support
0	0.75	0.84	0.79	60592
1	0.96	0.93	0.94	250436
accuracy			0.91	311028
macro avg	0.85	0.88	0.87	311028
weighted avg	0.92	0.91	0.91	311028

Neural Network

Results for the smaller, binary model,

Cross Validational set

Official Test Set

Modelling

	Correct	Wrong	Net	% Correct
xsnoop.	4	0	4	100.0
sqlattack.	2	0	2	100.0
apache2.	790	4	786	99.5
processtable.	729	30	699	96.0
saint.	701	35	666	95.2
xlock.	4	5	-1	44.4
ps.	7	9	-2	43.8
named.	7	10	-3	41.2
xterm.	5	8	-3	38.5
sendmail.	5	12	-7	29.4
mscan.	189	864	-675	17.9
httptunnel.	3	155	-152	1.9
udpstorm.	0	2	-2	0.0
worm.	0	2	-2	0.0
snmpguess.	1	2405	-2404	0.0
mailbomb.	0	5000	-5000	0.0
snmpgetattack.	0	7741	-7741	0.0

No discernable features for the model to use

Neural Network

Results for the smaller dataset with a binary model by new intrusion types.

Modelling

Cross Validation

	precision	recall	f1-score	support
brute force	0.0000	0.0000	0.0000	20
files	0.0000	0.0000	0.0000	5
internal	0.9999	0.9997	0.9998	703030
none	0.9968	0.9991	0.9980	242880
pings	0.9995	0.9999	0.9997	268110
scripts	0.0000	0.0000	0.0000	6
sweeps	0.9899	0.9717	0.9807	10281
warez	0.0000	0.0000	0.0000	276
accuracy			0.9991	1224608
macro avg	0.4983	0.4963	0.4973	1224608
weighted avg	0.9989	0.9991	0.9990	1224608

Attack types that the model predicted: ['sweeps', 'internal', 'none', 'pings']

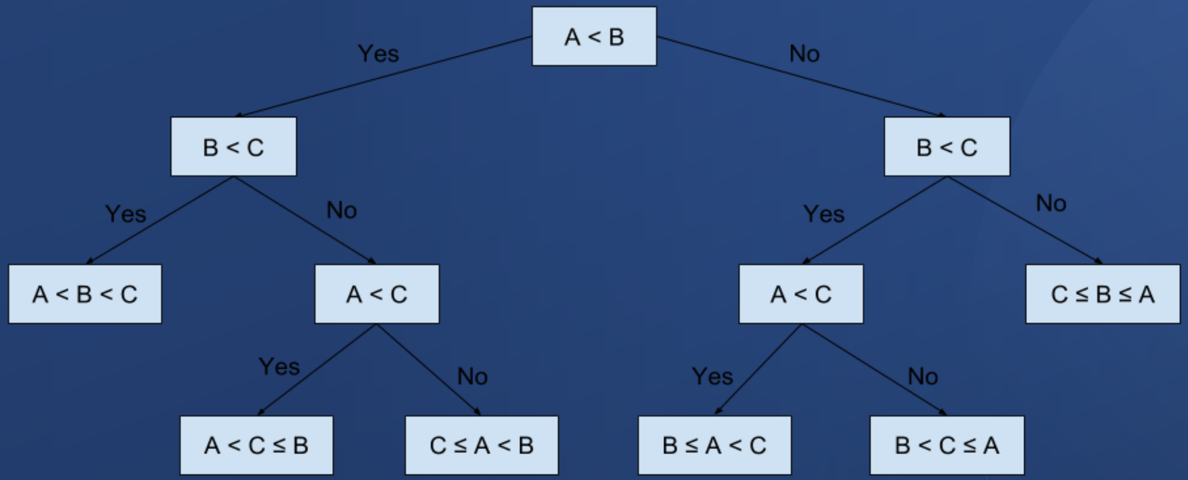
Test Set

	precision	recall	f1-score	support
brute force	0.0000	0.0000	0.0000	7574
files	0.0000	0.0000	0.0000	5
internal	0.9760	0.9960	0.9859	165217
none	0.7540	0.9017	0.8213	60592
pings	0.8832	0.9186	0.9005	63109
scripts	0.0000	0.0000	0.0000	8763
sweeps	0.8161	0.7499	0.7816	4166
warez	0.8998	0.2859	0.4339	1602
accuracy			0.9026	311028
macro avg	0.5411	0.4815	0.4904	311028
weighted avg	0.8601	0.9026	0.8791	311028

Neural Network

Results for the bigger dataset using a multi class classifier

Modelling



$$Gini = 1 - \sum_{i=1}^{n} p_i{}^2$$

Random Forest

Model of a single decision tree in a random forest

Purity measurement formula

Modelling

Random Forest

Small data, binary model
Cross Validitional Test Set

	precision	recall	f1-score	support
0	1.00	1.00	1.00	243103
1	1.00	1.00	1.00	981505
accuracy			1.00	1224608
macro avg	1.00	1.00	1.00	1224608
weighted avg	1.00	1.00	1.00	1224608

Small data, binary model
Official Test Set

	precision	recall	f1-score	support
0	0.70	0.85	0.76	60592
1	0.96	0.91	0.94	250436
accuracy			0.90	311028
macro avg	0.83	0.88	0.85	311028
weighted avg	0.91	0.90	0.90	311028

Modelling

	precision	recall	f1-score	support
brute force	0.03	0.25	0.05	7574
files	0.00	0.00	0.00	5
internal	1.00	0.00	0.00	165217
none	0.21	0.79	0.34	60592
pings	0.98	0.28	0.43	63109
scripts	0.00	0.00	0.00	8763
sweeps	0.78	0.72	0.75	4166
warez	1.00	0.13	0.23	1602
accuracy			0.23	311028
macro avg	0.50	0.27	0.23	311028
weighted avg	0.79	0.23	0.17	311028

'brute force', 'internal', 'none', 'pings', 'sweeps', 'warez'

Random Forest

Big Data, Multi Class Model Official Test Set

Categories that were predicted

Conclusion

Cons
"As a side note, the [Intrusion Detection Systems] IDS research community vehemently discourages the use the DARPA dataset (and the derived KDD Cup dataset) despite it's appealing availability."

- The background traffic generator is not publicly available so there is no way to evaluate the background traffic in the dataset.

Pros
"the non-availability of any other dataset that includes the complete network traffic"

This dataset is the perfect training dataset. It gives me a chance to hone my skills and techniques on a large amount of data in an important field while getting clear and interesting results.

Important Insights:

- **The predictions that the Multi Class Neural Network Model gave were significantly better than a trivial classifier**
- **Cluster Analysis gave a view into how features can affect different intrusion types.**