Detecting Attacks and Securing Networks

MS Business Analytics, Dominican University of California
Justin Brown, Rana Demirer, Jackie Ocaña, Kohsuke Uchimura, Rahmat Ullah
August 12, 2023

Kohsuke Uchimura

Technical Leader-1

Justin Brown

Presentation Leader-1

GOLDEN GATE
Keepers

Jackie Ocaña

Project Leader

Rana Demirer

Technical Leader-2

Rahmat Ullah

Presentation Leader-2

Agenda

01 **Cybersecurity**
What is it? How does this affect me?

02 **Our Solution**
Present our solution to current problems.

03 **Data Breakdown**
Describe our dataset.

04 **Model Building & Predicting**
Building our solution to the problem.

05 **Insights & Questions**
Summarize our content.

CYBERSECURITY IS AN ART

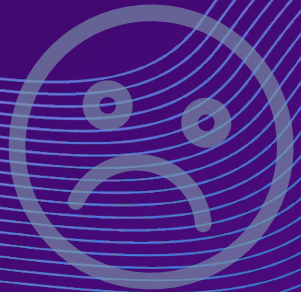
Confidentiality

Integrity

Availability

$11,000,000,000,000

Expected Loss Due to Cybercrime in 2023



Source: World Economic Forum

Sectors Most Affected by Cyberattacks

Small Businesses

Healthcare

Government

Finance

Education

Utility Companies

3.4

million

Job Openings

in cybersecurity

By 2025, lack of talent or human failure will be responsible for over half of significant cybersecurity incidents.

Source: NIST

www.kahoot.it

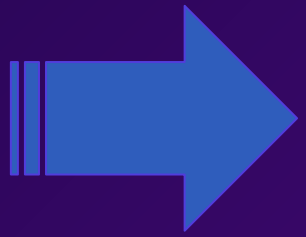
Our Solution to the Problem: Building an IDS System

What is an IDS?

- Intrusion Detection System

- Monitoring network traffic and detecting anomalous activity

- Automated defense tools against network attacks.



Limitations

Datasets are outdated, lack traffic diversity, insufficient attack coverage, and inadequate anonymization.

Cleaning Our Data

Data Cleaning

Eliminates erroneous or inconsistent data points for improved data quality.

Data Transformation

Ensures consistent and comparable format for all features.

Feature Engineering

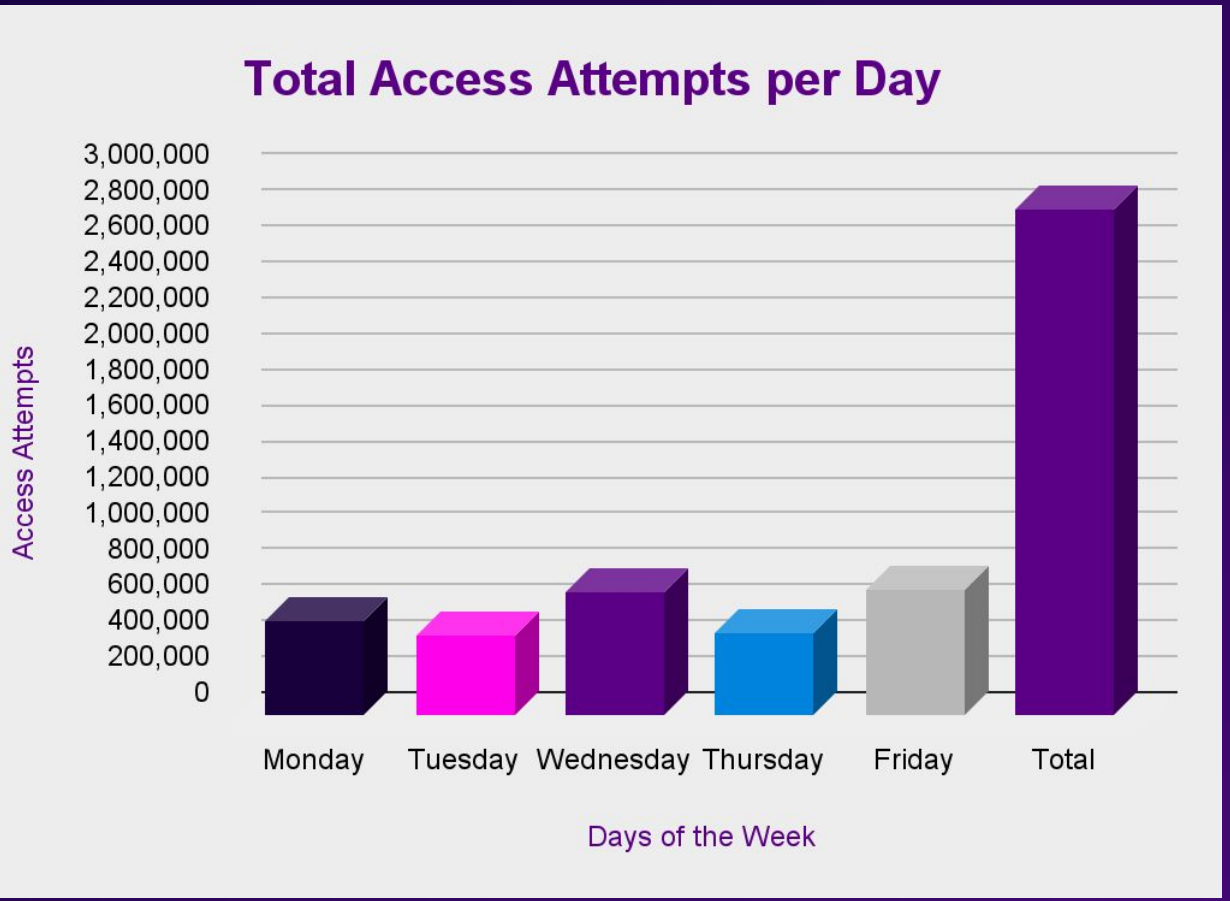
Extracts relevant features to enhance dataset's informativeness.

Preprocessing

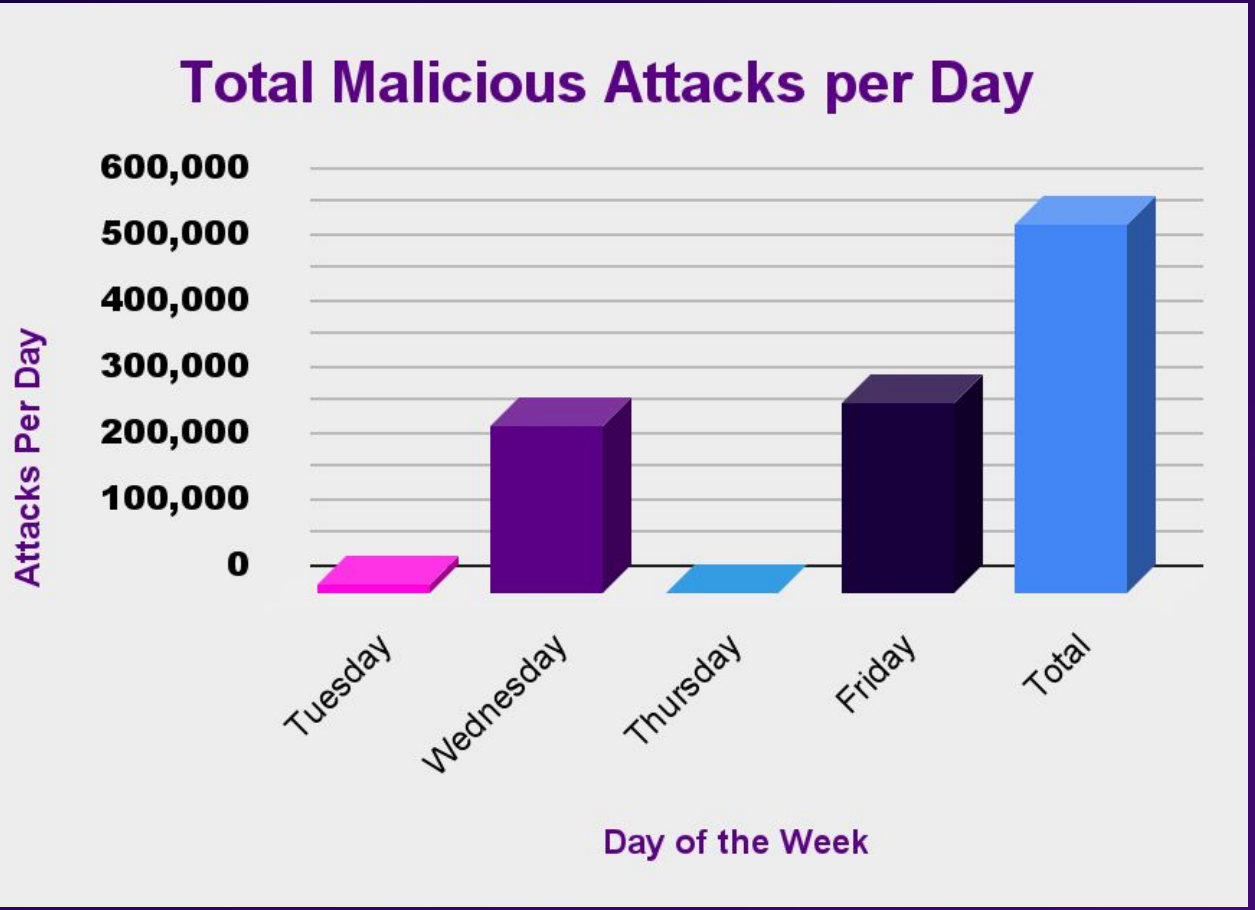
Forms a robust foundation for intrusion detection and network security analysis.

Exploring Our Data

2,830,743	Samples (rows) identified	0	Missing Values
8	8 CSV files, combined	14	Attack types identified
2017	Year University of New Brunswick compiled	19.69%	What percent were attacks?
80	Features available	40.6 GB	Total file size



Friday Access Attempts: 703,245



Friday Attack Attempts: 288,923

Most Common Attack Types

PortScan

Scans for open ports to identify points of entry.

DoS Hulk

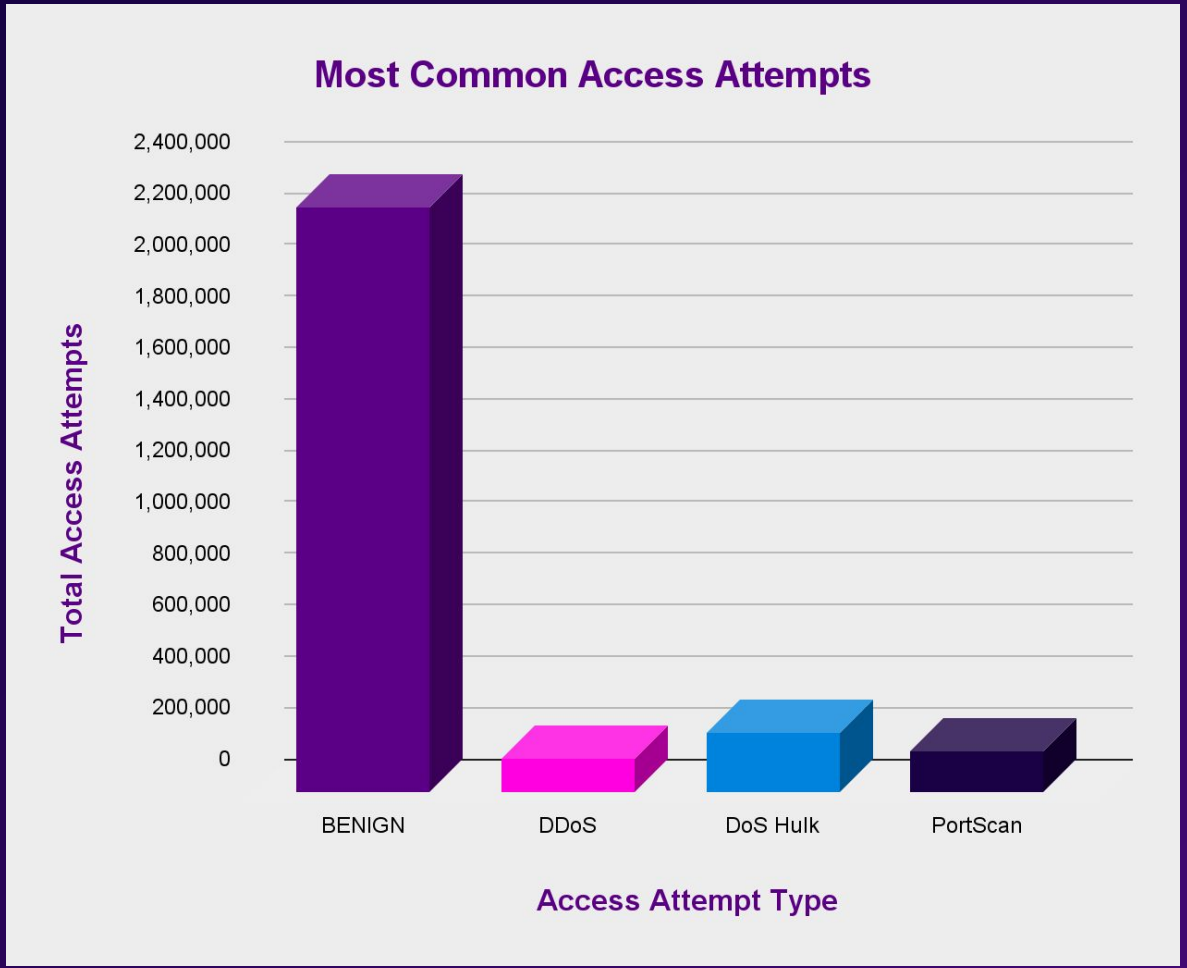
Sends large amount of TCP SYN packets to overwhelm the system.

DDoS

Distributed denial of service, floods network with traffic. Many subtypes.

Benign

Normal non-malicious traffic.



Total DDoS, DoS Hulk, PortScan Attempts 518,030

Victim & Attacker Network Information

Attackers network:

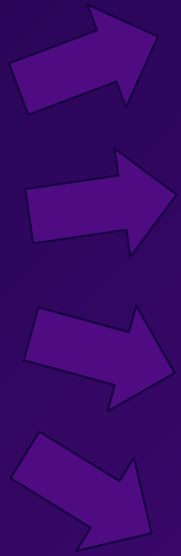
- Kail: 205.174.165.73
- Windows: 205.174.165.69(70, or 71)

Victim network:

- Web server 16 public: 192.168.10.50 and more
- Ubuntu server 12 public: 192.168,10.51 and more
- Ubuntu 14.4 32B: 192.168.10.17
- Windows 7 pro 64B: 192.168.10.8
- MAC: 192.168.10.25

Procedure for Modeling

- Encode Label(Benign, Dos, etc) to numeric
- Split feature and target into train data and test data
- Standardize features



Decision Tree Model

K Nearest Neighbor Model

Support Vector Machine Model

XGBoost Model

Label Encode

```
Encoded data:  [ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14]
Decoded_data:  ['BENIGN' 'Bot' 'DDoS' 'DoS GoldenEye' 'DoS Hulk' 'DoS Slowhttptest'
 'DoS slowloris' 'FTP-Patator' 'Heartbleed' 'Infiltration' 'PortScan'
 'SSH-Patator' 'Web Attack \x96 Brute Force' 'Web Attack \x96 Sql Injection'
 'Web Attack \x96 XSS']
```

Modeling

Classification Model (Use whole data)

1. Decision Tree model

   Accuracy Score: 0.996
2. KNN model

   Can’t classified correctly
3. SVM model

   Too expensive to use whole data
4. XGBoost

   Accuracy Score: 0.898

Problem using whole data

- Some model can’t shows prediction correctly since dataset is huge
- Even decision tree model shows high accuracy
- some malicious access cannot be identified correctly by this model

	precision	recall	f1-score	support
0	1.00	1.00	1.00	682132
1	0.97	0.69	0.81	597
2	1.00	1.00	1.00	38471
3	0.96	0.99	0.97	3095
4	0.99	1.00	0.99	69061
5	0.99	0.94	0.96	1697
6	0.92	0.94	0.93	1717
7	1.00	1.00	1.00	2408
8	0.75	0.60	0.67	5
9	0.00	0.00	0.00	13
10	0.99	1.00	1.00	47542
11	1.00	0.99	1.00	1770
12	0.53	0.11	0.18	494
13	0.00	0.00	0.00	4
14	0.91	0.05	0.09	217
accuracy			1.00	849223
macro avg	0.80	0.69	0.71	849223
weighted avg	1.00	1.00	1.00	849223

Problem using whole data

- For example, labels 9, 12, 13 and 14 was very low accuracy about both precision and recall score.
- 9 = Infiltration, 12 = Brute Force, 13 = SQL Injection, 14 = XSS
- To improve this, we tried to pick the sample from whole data and train model again.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	682132
1	0.97	0.69	0.81	597
2	1.00	1.00	1.00	38471
3	0.96	0.99	0.97	3095
4	0.99	1.00	0.99	69061
5	0.99	0.94	0.96	1697
6	0.92	0.94	0.93	1717
7	1.00	1.00	1.00	2408
8	0.75	0.60	0.67	5
9	0.00	0.00	0.00	13
10	0.99	1.00	1.00	47542
11	1.00	0.99	1.00	1770
12	0.53	0.11	0.18	494
13	0.00	0.00	0.00	4
14	0.91	0.05	0.09	217
accuracy			1.00	849223
macro avg	0.80	0.69	0.71	849223
weighted avg	1.00	1.00	1.00	849223

Confusion Matrix

Predicted

TRUE

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	681136	12	41	80	506	13	17	1	1	4	296	0	25	0	0
1	183	414	0	0	0	0	0	0	0	0	0	0	0	0	0
2	21	0	38445	0	5	0	0	0	0	0	0	0	0	0	0
3	0	0	0	3073	14	1	7	0	0	0	0	0	0	0	0
4	79	0	0	37	68894	0	51	0	0	0	0	0	0	0	0
5	28	0	0	0	1	1596	71	0	0	0	0	0	1	0	0
6	52	0	0	3	14	6	1622	0	0	0	0	0	20	0	0
7	4	0	0	0	0	0	0	2404	0	0	0	0	0	0	0
8	2	0	0	0	0	0	0	0	3	0	0	0	0	0	0
9	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	71	0	0	0	17	0	0	0	0	0	47453	0	1	0	0
11	10	0	0	0	0	0	0	1	0	0	0	1759	0	0	0
12	422	0	0	17	0	0	0	0	0	0	0	0	54	0	1
13	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0
14	206	0	0	0	0	0	0	0	0	0	0	0	1	0	10

Usage of Model

We have explored and assessed various models to determine the best fit and accuracy for our prediction task, our dataset is now trained and prepared for making predictions on unseen data.

By applying this trained model to new, previously unseen data, now we can confidently predict outcomes with an impressive accuracy rate of approximately 98%.

This means that our model has effectively learned from the training data and has developed a strong ability to generalize its predictions to new and unfamiliar instances. The process of selecting the optimal model, training it, and achieving such high accuracy on unseen data demonstrates the robustness and reliability of our predictive system.

```
[ ]:
prediction_data = prediction_data[important_features]
PredictNewTraffic(prediction_data,0)
```

Insights and Recommendations

Insights and Recommendations (cont.)

Up-to-date Datasets	Advanced Anonymization	Feature Engineering
Ensemble Learning	Sample Selection	Real-time Monitoring


THANK YOU
GOLDEN GATE
Keepers

References

IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB. (n.d.). https://www.unb.ca/cic/datasets/ids-2017.html

National Institute of Standards and Technology. (2023, June). *Cybersecurity Workforce Demand*. https://www.nist.gov/document/workforcedemandonepager.

https://www.nist.gov/document/workforcedemandonepager

US warns of huge cyber-espionage campaign, and other cybersecurity news to know this month. (2023, June 16). World Economic Forum.

https://www.weforum.org/agenda/2023/06/us-china-cyber-espionage-campaign-cybersecurity-news/

Western Governors University. (2021, August 3). 6 Industries most vulnerable to cyber attacks. *Western Governors University*.

https://www.wgu.edu/blog/6-industries-most-vulnerable-cyber-attacks2108.html#close

What is an Intrusion Detection System (IDS)? Definition & Types | Fortinet. (n.d.). Fortinet. https://www.fortinet.com/resources/cyberglossary/intrusion-detection-system

What is Cybersecurity? | CISA. (2021, February 1). Cybersecurity and Infrastructure Security Agency CISA. https://www.cisa.gov/news-events/news/what-cybersecurity

GitHub code. https://github.com/KHUC1998/Practicum-capstone_code/blob/main/MSBA5510_0712.ipynb