REVA UNIVERSITY

Botnet Detection in Network Traffic Based on GBM

Kiran Muloor
Shashidhara GM
Somesh Sahu
Sandeep Shyam Bajaj

- Business Problem/ Understanding
- Literature Review
- Data Pipeline
- Data Understanding
- Modeling
- Conclusion

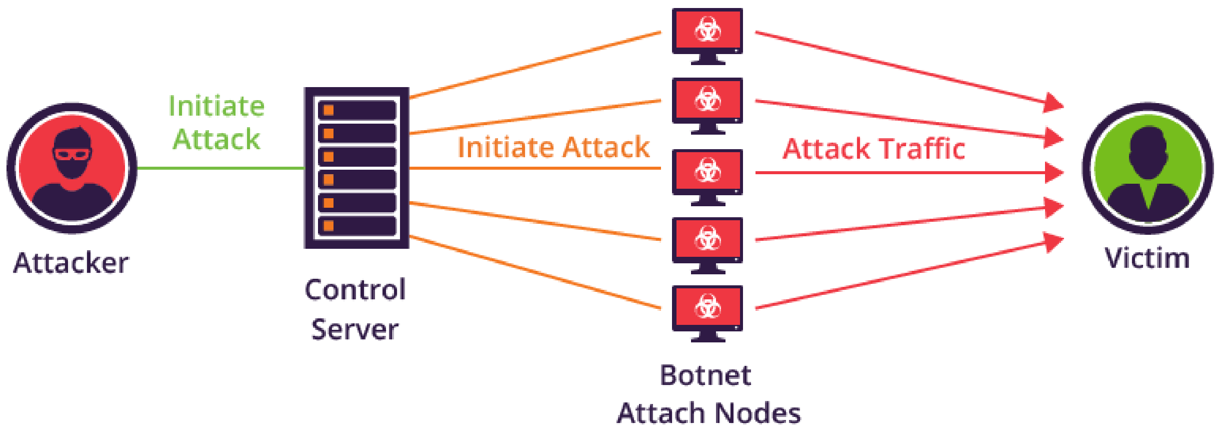


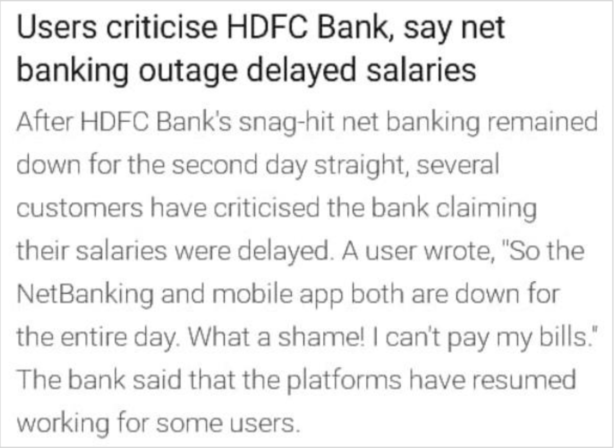
Image Source: https://www.imperva.com/blog/how-to-identify-a-mirai-style-ddos-attack/

Service provider deliverables are impacted due to Intermittent Network down or slowness causing,

- Service failure
- Financial loses
- Reputation

Different attacks been identified

3ve—2018
3ve botnet gave rise to three different yet interconnected sub-operations, each of which was able to evade investigation after perpetrating ad fraud skillfully. Google, White Ops, and other tech companies together coordinated to shut down 3ve’s operations. It infected around 1.7 million computers and a large number of servers that could generate fake traffic with bots.

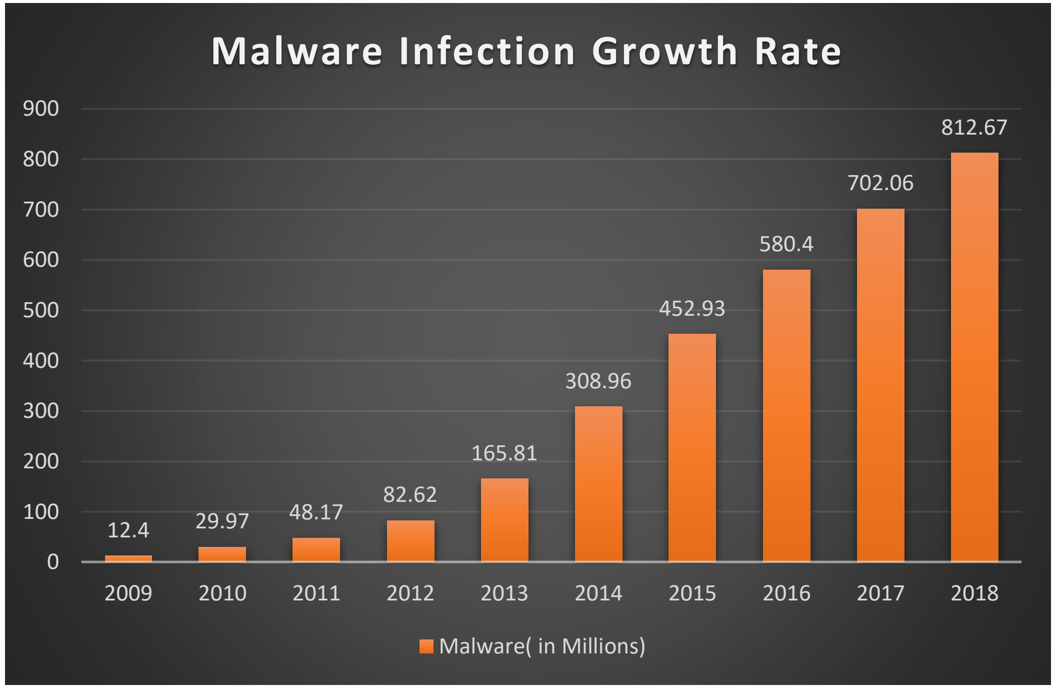

Users criticise HDFC Bank, say net banking outage delayed salaries

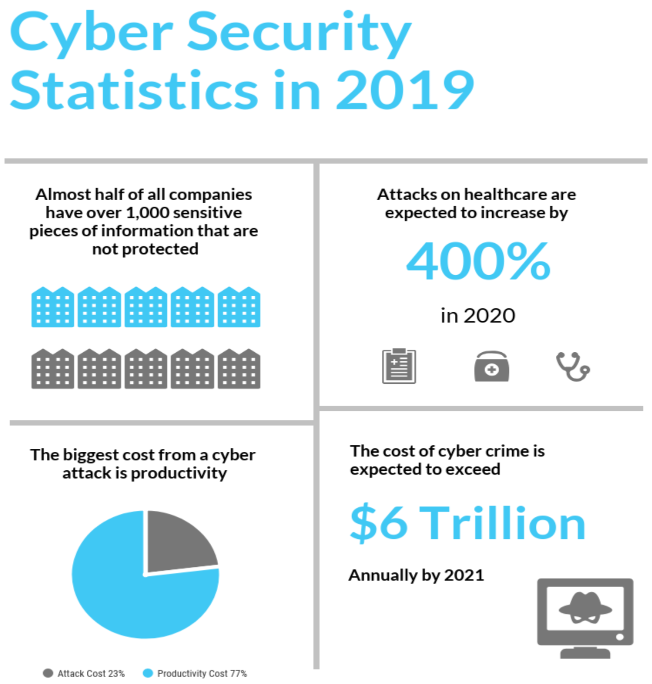
After HDFC Bank's snag-hit net banking remained down for the second day straight, several customers have criticised the bank claiming their salaries were delayed. A user wrote, "So the NetBanking and mobile app both are down for the entire day. What a shame! I can't pay my bills." The bank said that the platforms have resumed working for some users.

Mirai—2016
Mirai infects digital smart devices that run on ARC processors and turns them into a botnet, which is often used to launch DDoS attacks. If the default name and password of the device is not changed then, Mirai can log into the device and infect it. In 2016, the authors of Mirai software launched a DDoS attack on a website that belonged to the security service providing company.

Article Source:

- https://blog.eccouncil.org/9-of-the-biggest-botnet-attacks-of-the-21st-century/
- https://timesofindia.indiatimes.com/business/india-business/hdfc-bank-online-snag-persists-for-second-day/articleshow/72356772.cms




Cyber Security Statistics in 2019

Almost half of all companies have over 1,000 sensitive pieces of information that are not protected

Attacks on healthcare are expected to increase by **400%** in 2020

The biggest cost from a cyber attack is productivity

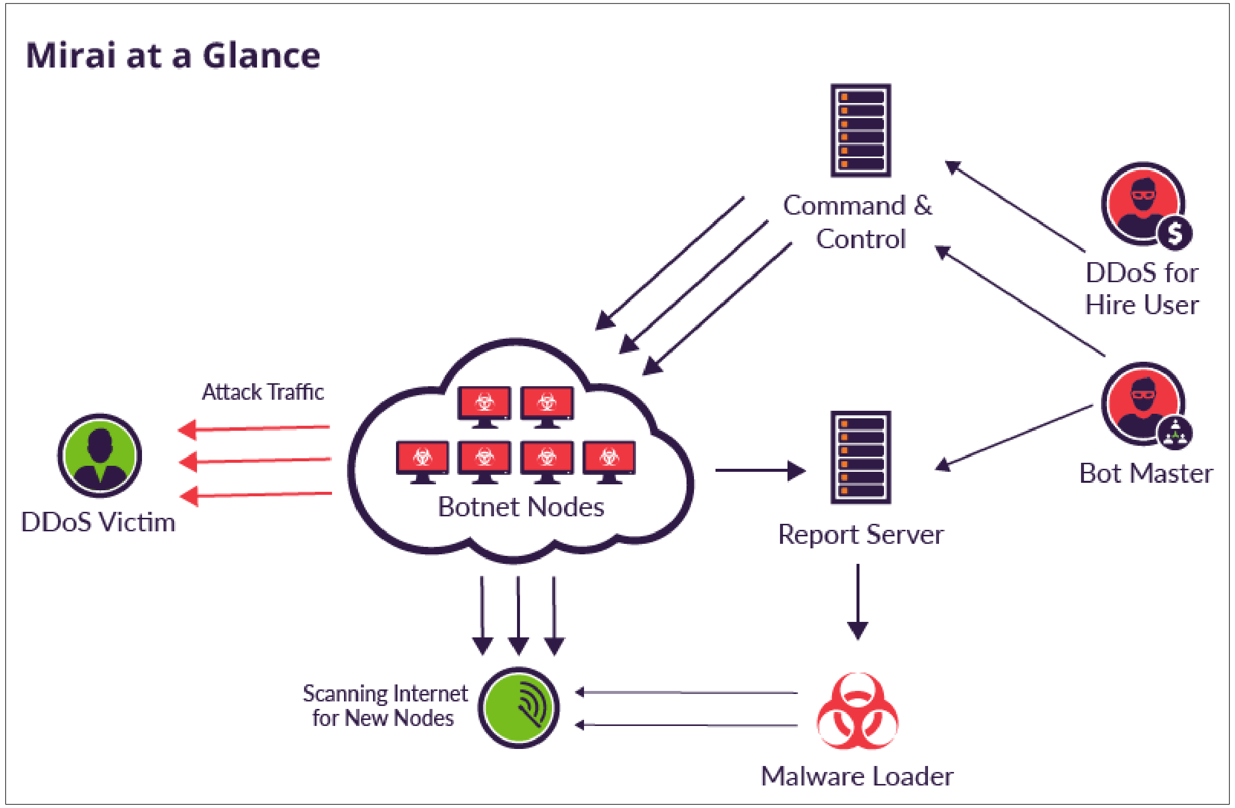
The cost of cyber crime is expected to exceed **$6 Trillion** Annually by 2021

- Malware rose 79% from 2017
- In 2018, 90% of financial institutions reported being targeted by malware
- 92% of malware is delivered by email.
- New malware variants for mobile increased by 54% in 2018
- Botnets were shifted from Windows platforms towards Linux and IoT platforms, leading to the fast decline of older Windows-based families and the thriving of new IoT-based ones.

Image Source:
- https://purplesec.us/resources/cyber-security-statistics/
- https://www.gigabitmagazine.com/telecoms/ensuring-network-security-5g-era

Botnet attack amasses a large number of compromised hosts sending useless packets to jam its services

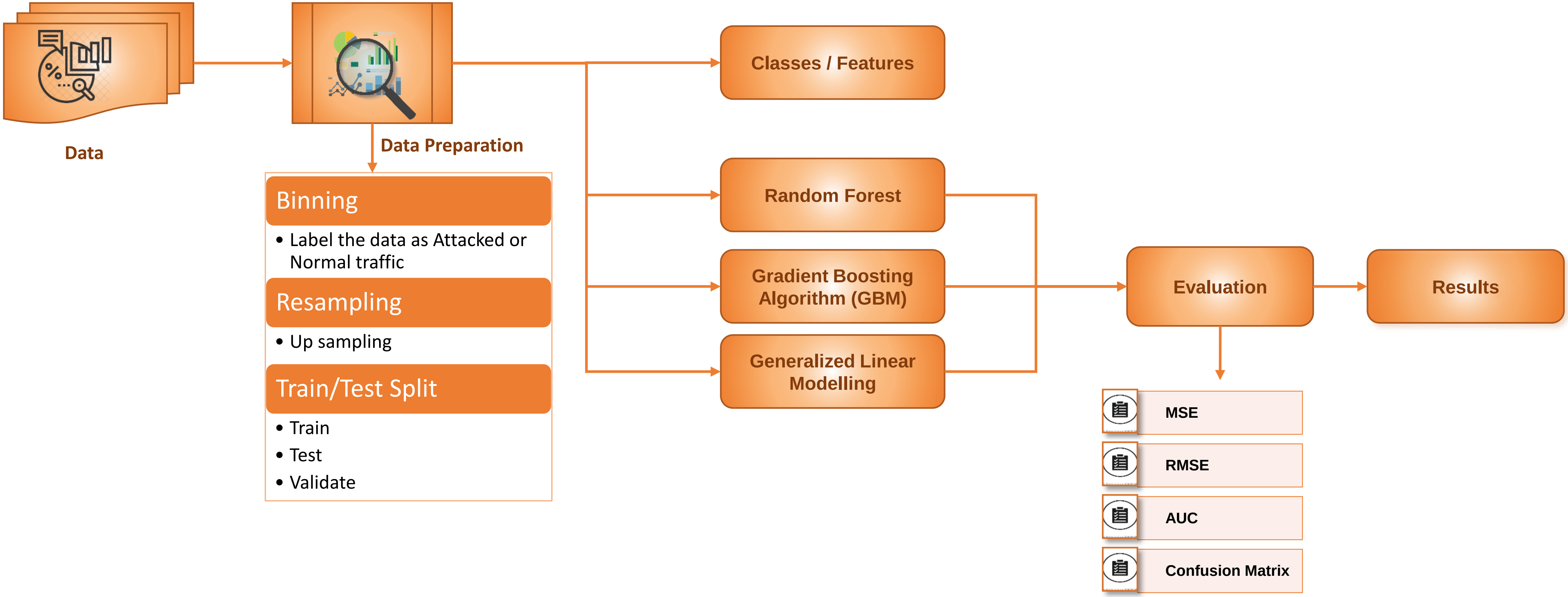
- **Breach of availability**
  Unauthorized destruction of data
- **Theft of service**
  Unauthorized use of resources
- **Denial of service (DOS)**
  Prevention of legitimate use
  Bots are typically controlled remotely



Research Paper	Author	Machine Learning Models	Results
Botnet Detection Based On Machine Learning Techniques Using DNS Query Data	Hoang et.al (2018)	Random Forest	90.80%
Automated Botnet Traffic Detection via Machine Learning	Wai, F. K. et.al (2018)	Support Vector Machine , Random Forest etc.	90.8% (**Average Score**)
An Adaptive Multi-Layer Botnet Detection Technique Using Machine Learning Classifiers	Khan, R. U et.al (2019)	Decision Trees	98.7%

Research Paper Gap

- Above Research papers have analyzed botnet through various classification ML models and has given good results
- Lesser work done on H2O Gradient Boosting Machine (GBM)
- Scope to work on H2O GBM detecting botnet



Data

Data Preparation

Binning

- Label the data as Attacked or Normal traffic

Resampling

- Up sampling

Train/Test Split

- Train
- Test
- Validate

Classes / Features

Random Forest

Gradient Boosting Algorithm (GBM)

Generalized Linear Modelling

Evaluation

Results

MSE

RMSE

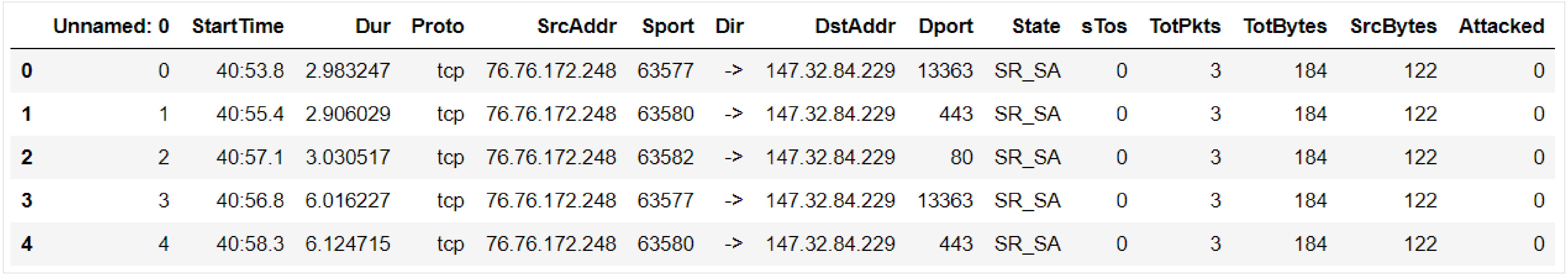
AUC

Confusion Matrix

Data

- Total number of Records – 1.8M
- Features - 16
- Data has been labelled

Sample Data



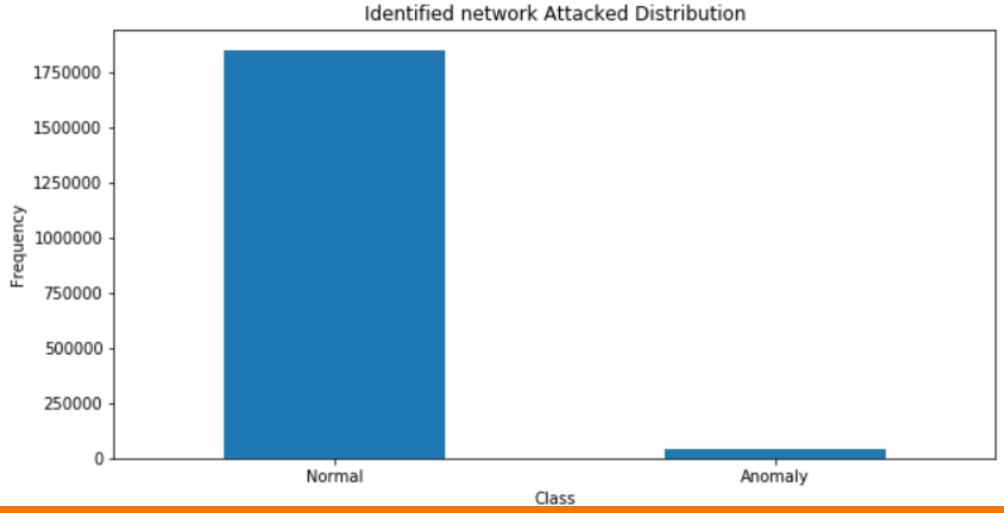
	Unnamed: 0	StartTime	Dur	Proto	SrcAddr	Sport	Dir	DstAddr	Dport	State	sTos	TotPkts	TotBytes	SrcBytes	Attacked
0	0	40:53.8	2.983247	tcp	76.76.172.248	63577	->	147.32.84.229	13363	SR_SA	0	3	184	122	0
1	1	40:55.4	2.906029	tcp	76.76.172.248	63580	->	147.32.84.229	443	SR_SA	0	3	184	122	0
2	2	40:57.1	3.030517	tcp	76.76.172.248	63582	->	147.32.84.229	80	SR_SA	0	3	184	122	0
3	3	40:56.8	6.016227	tcp	76.76.172.248	63577	->	147.32.84.229	13363	SR_SA	0	3	184	122	0
4	4	40:58.3	6.124715	tcp	76.76.172.248	63580	->	147.32.84.229	443	SR_SA	0	3	184	122	0

Data Source: Collected from End Points [Firewall, switch etc.]

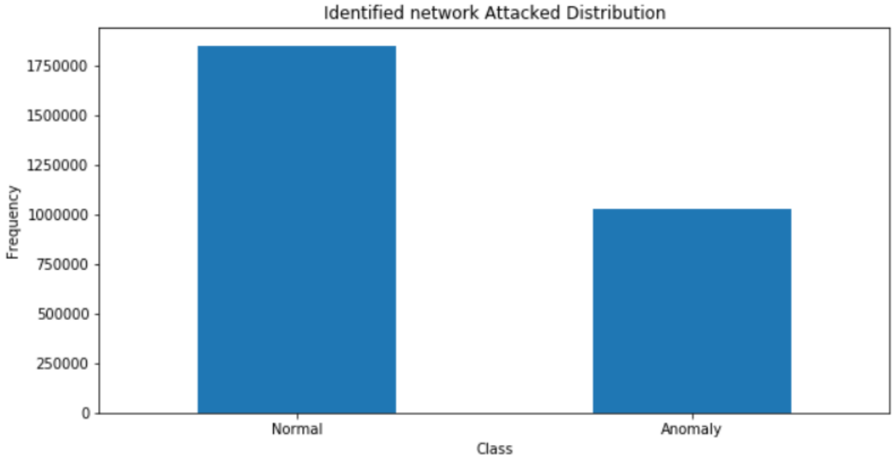
Initial Data

- Labelled data
- Anomaly – 40003 (2% of total data)

Before Up sampling



After Up sampling



Data Up-sampling

- Overcome the problem of overfitting
- Minority -class increased to 55%

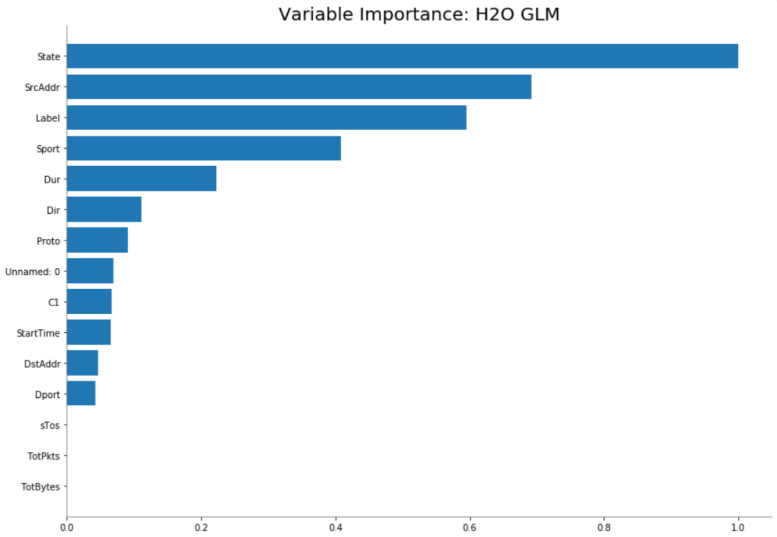
WHY Gradient Boosting Algorithm:

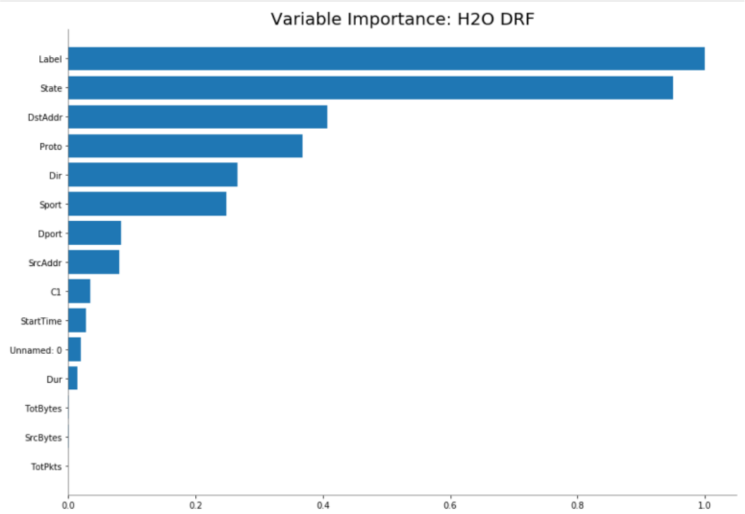
- Gradient boosting is a machine learning technique for regression and classification problems
- GBM produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
- It build model in stage-wise fashion and it generalizes them by allowing optimization of arbitrary differentiable loss function

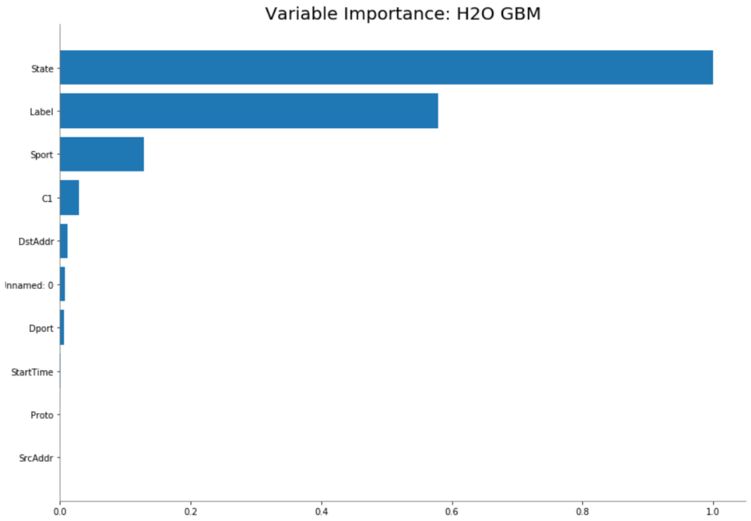
WHY H2O.ai

- It streamlines the process for development into an intuitive workflow
- Trains models faster than popular packages like sci-kit learn
- Delivers a fast and accessible ML platform for large datasets that is equipped with user-friendly and high-performing tools.

Variable importance for the different models







Our experiment is focused on GBM

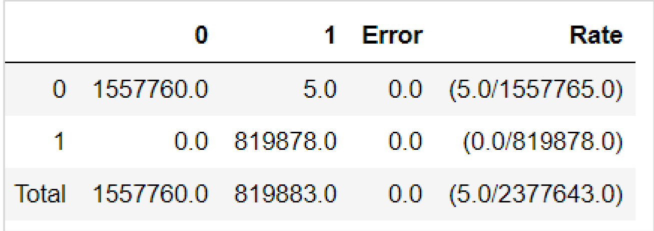
Experiment 1

[Train/Test Split]

Train: 80%

Validate: 20%

- **Mean Square Error:** 1.3916525385408573e-05
- **Root Mean Square Error (RMSE) :** 0.003730485944941835
- **Accuracy:** 0.9999999996781965



	0	1	Error	Rate
0	1557760.0	5.0	0.0	(5.0/1557765.0)
1	0.0	819878.0	0.0	(0.0/819878.0)
Total	1557760.0	819883.0	0.0	(5.0/2377643.0)

Experiment 2

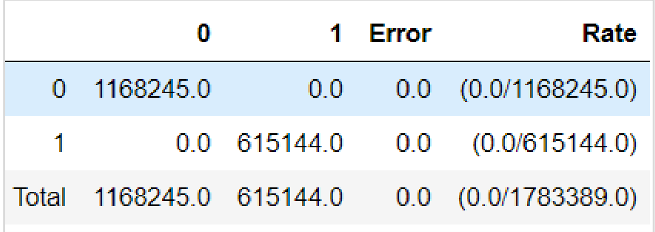
[Train/Test Split]

Train: 70%

Test: 15%

Validate: 15%

- **Mean Square Error:** 2.9326425712147635e-09
- **Root Mean Square Error (RMSE) :** 5.415387863500419e-05
- **Accuracy:** 1.0



	0	1	Error	Rate
0	1168245.0	0.0	0.0	(0.0/1168245.0)
1	0.0	615144.0	0.0	(0.0/615144.0)
Total	1168245.0	615144.0	0.0	(0.0/1783389.0)

Experiment 3

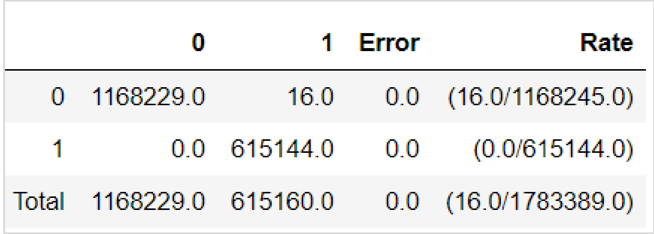
[Train/Test Split]

Train: 60%

Test: 20%

Validate: 20%

- **Mean Square Error:** 2.958186564514117e-05
- **Root Mean Square Error (RMSE) :** 0.005438921367802735
- **Accuracy:** 0.9999999983733137



	0	1	Error	Rate
0	1168229.0	16.0	0.0	(16.0/1168245.0)
1	0.0	615144.0	0.0	(0.0/615144.0)
Total	1168229.0	615160.0	0.0	(16.0/1783389.0)

- True Positives and True Negatives have been precisely classified with small error in Experiment 1 and 3.
- The precise classification of True positives and True negatives also mirrors with high AUC.
- The approach used in this study helps Cybersecurity teams to detect Botnet attacks proactively, increase network uptime and minimize the business impact.
- Based on the results using H2O GBM demonstrates high AUC ranging from 0.9999 to 1.0

Recommendation:

- We recommend using **Experiment 2**

Future Studies:

- Further we would like to expand our study in predict different types of Botnet attacks and work on productizing solution.

Thank you

Q & A