Network Intrusion Detection -

Classification of network traffic

Melissa Ng

Background

With increasing cases of cyber attacks, it is important to identity irregular or abnormal network activities.

Medibank Becomes Latest Target of Cyber Attack in Australia

- Health insurer detected unusual activity on its network
- No evidence any sensitive, customer data was accessed

By Keira Wright
13 October 2022 at 09:58 GMT+8

Jonathan Greig
October 11, 2022

Government
Nation-state
News

Coverage of Killnet DDoS attacks plays into attackers' hands, experts say

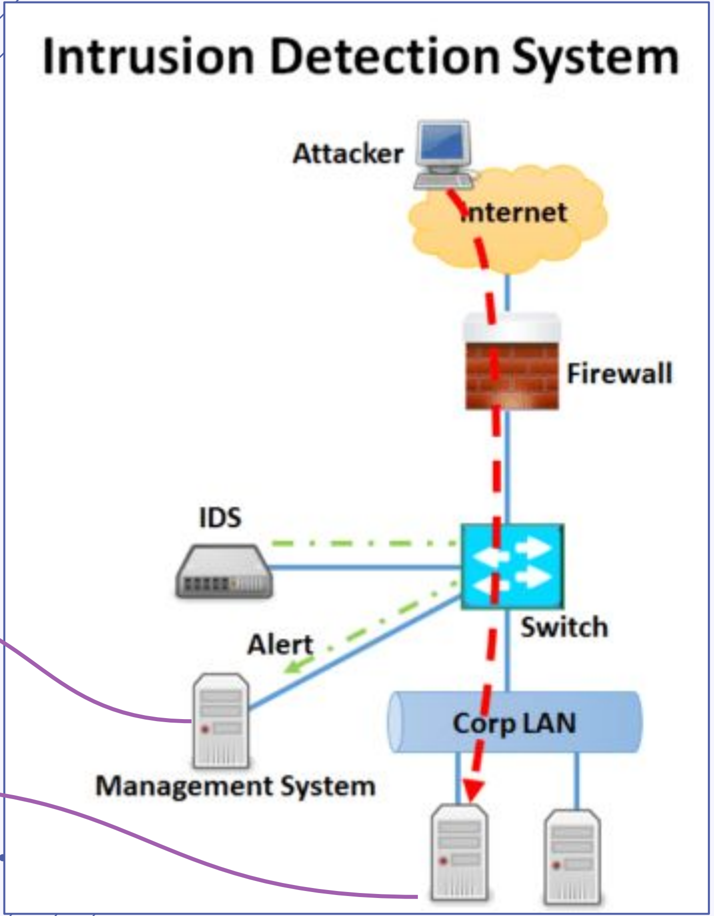
A notorious pro-Russian hacking group drew headlines on Monday after launching distributed denial-of-service (DDoS) attacks on the websites of airports in at least 24 different states and threatening more operations against U.S. entities.

Background

A network-based intrusion detection system (NIDS) is place to **collect** and **analyse** network traffic and **report** any behaviour that **falls outside normal activity**.

However, the known challenges of NIDS are

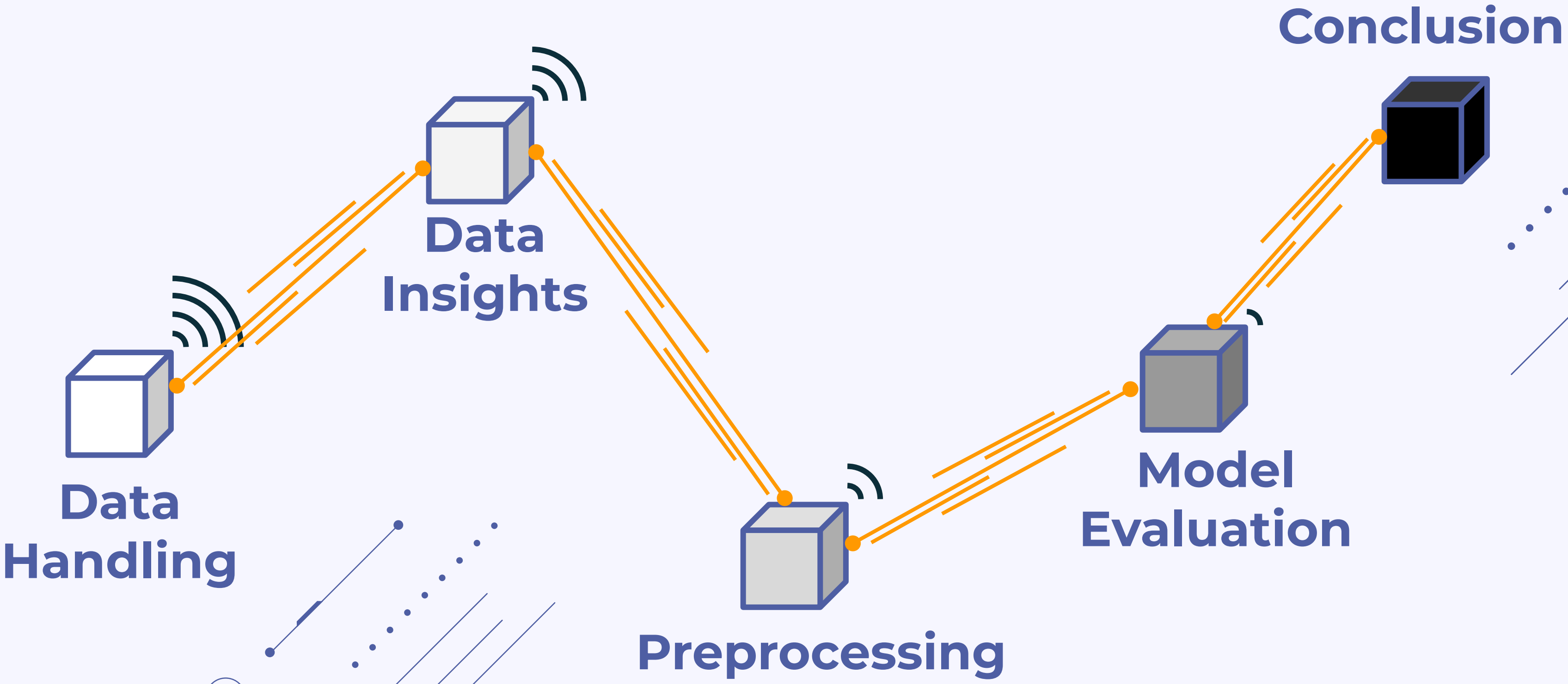
- High false alarm (*high false positive*) -> operation overhead
- Low detection rate (*high false negative*) -> prolonged attack go undetected



Problem Statement

1. Improve intrusion detection rate by **reducing false negatives**
2. Reduce operational overhead by **reducing false alarm rate**

Checkpoints



01

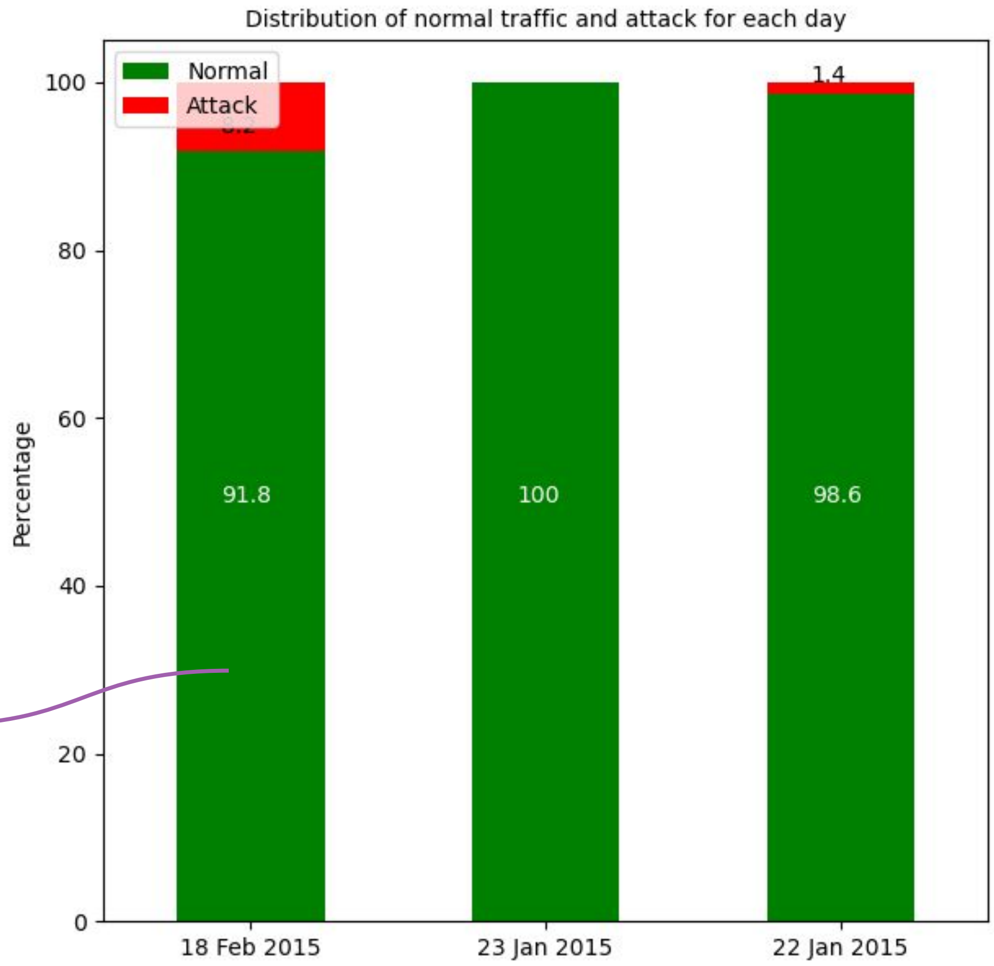
Data Handling

Dataset

(from University of New South Wales, Canberra, Australia)

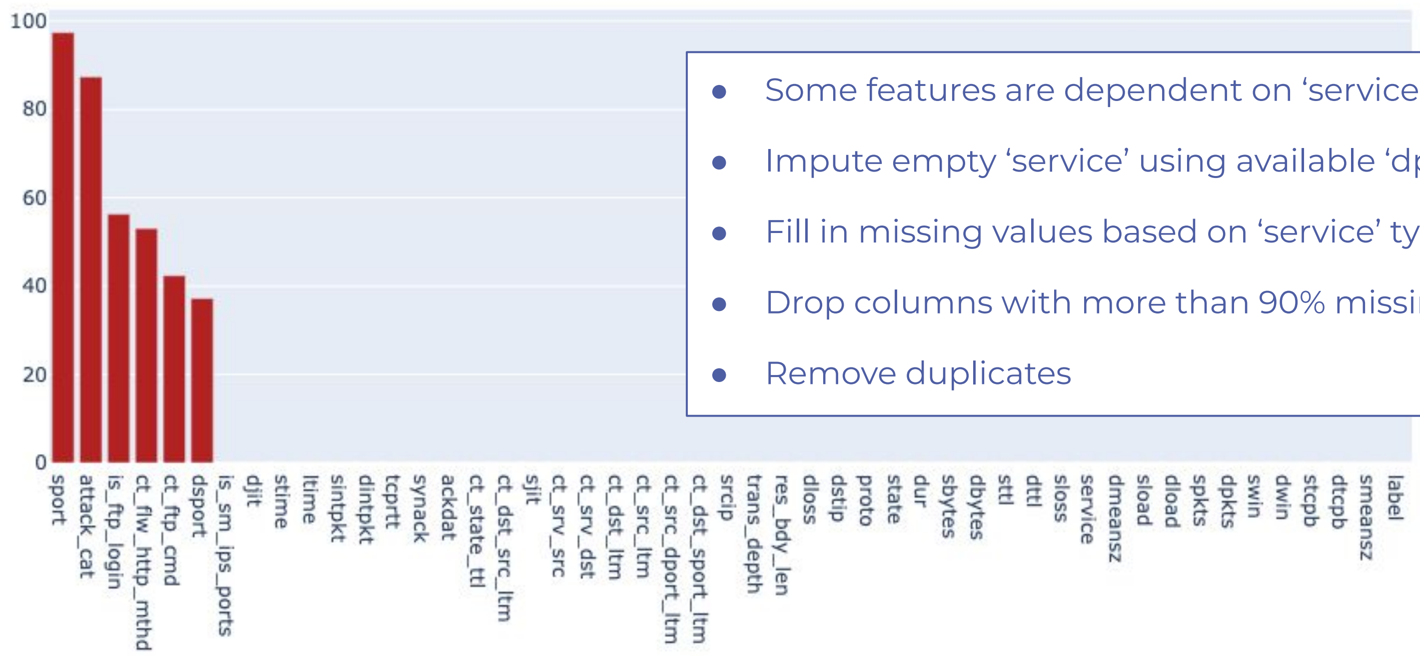
The dataset is made up of 2.5 million records consisting a hybrid of **real** modern **normal** activities and **synthetic** contemporary **attack** behaviours generated over 3 days.

Use a subset from 18 Feb 2015 due to the large dataset and is the least imbalance



Data Cleaning (based on best knowledge)

Percentage of missing values for each feature

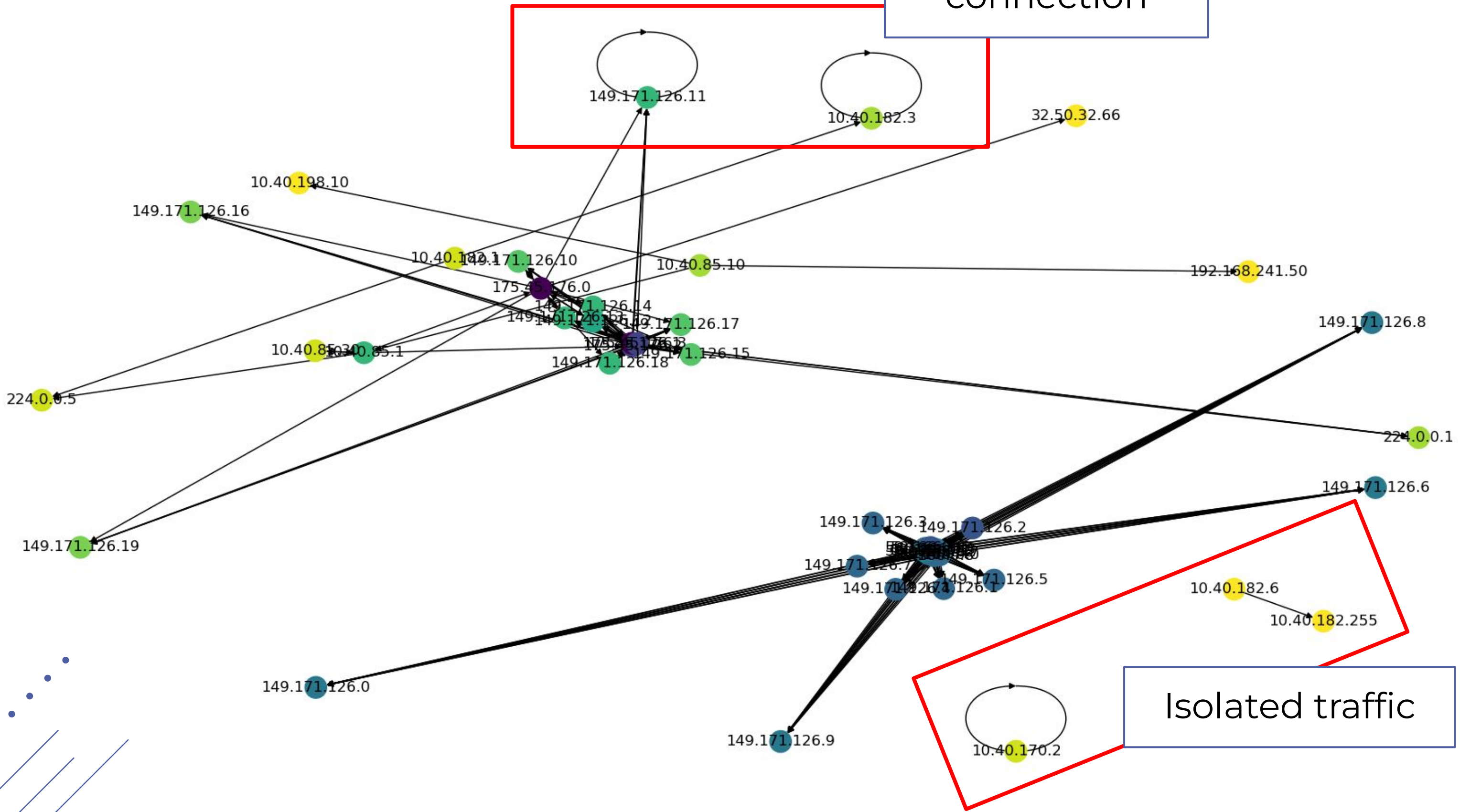


- Some features are dependent on ‘service’ type
- Impute empty ‘service’ using available ‘dport’ number
- Fill in missing values based on ‘service’ type
- Drop columns with more than 90% missing values
- Remove duplicates

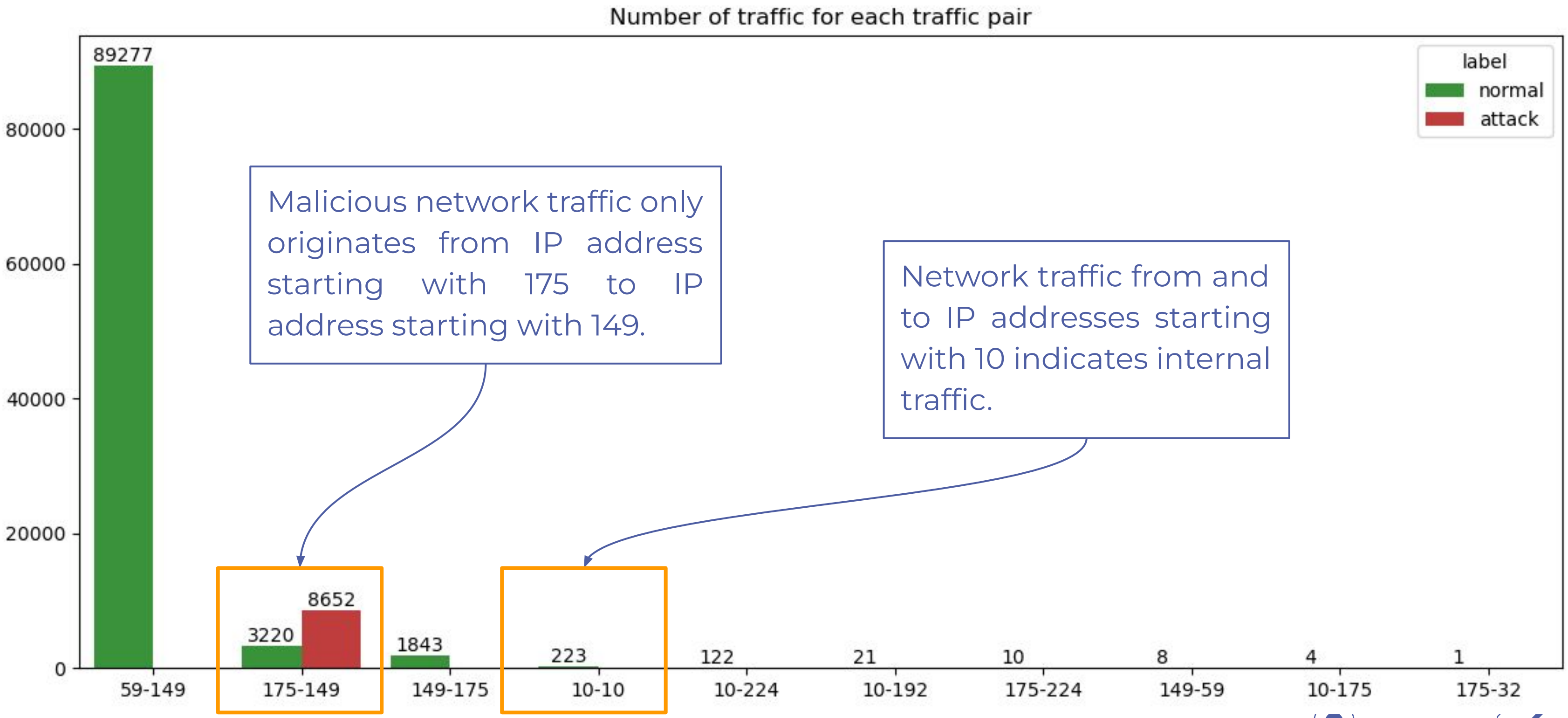
02

Data Insights

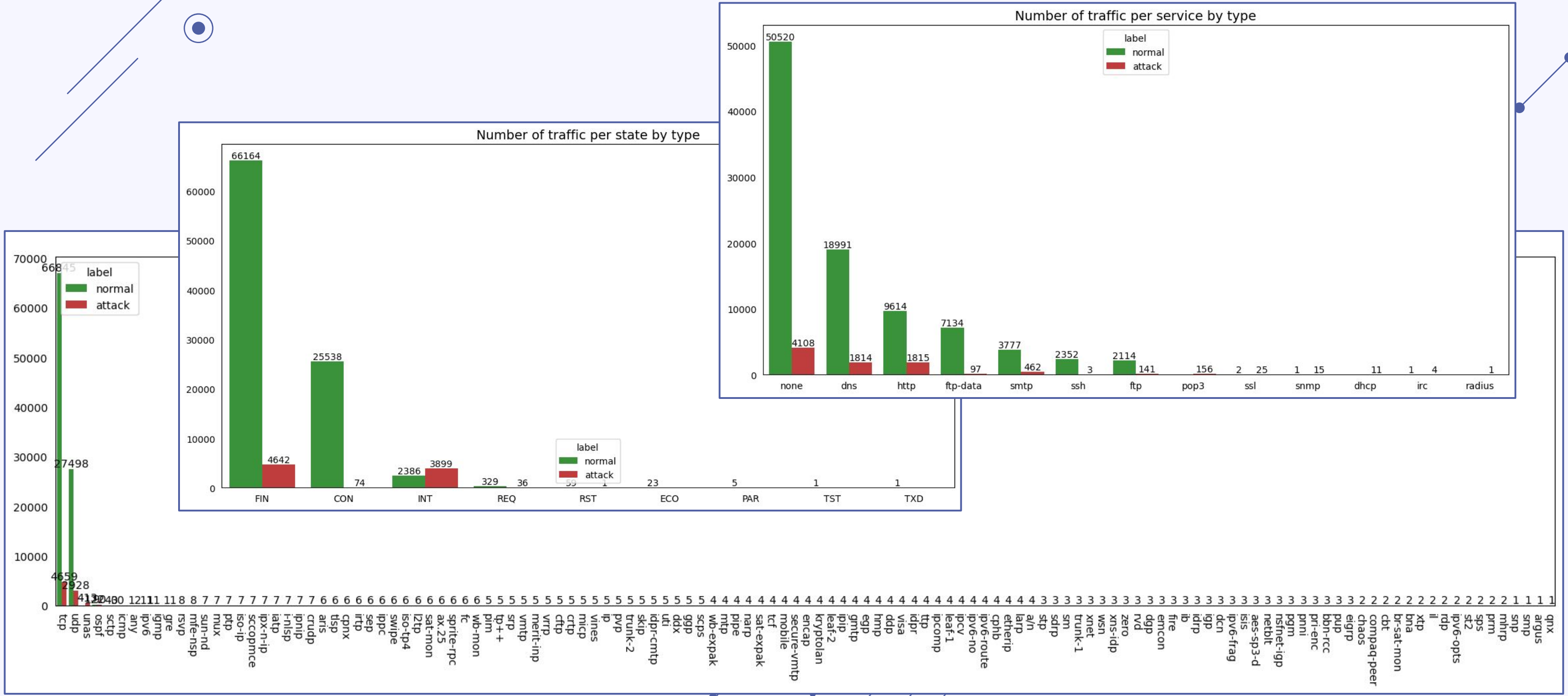
Traffic Behaviour



Traffic Pairs



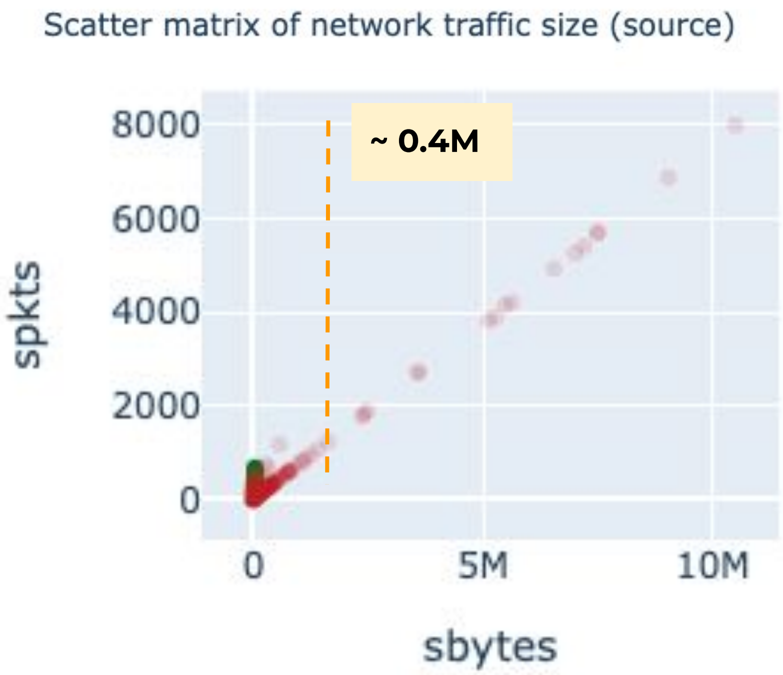
Protocol, State, Service

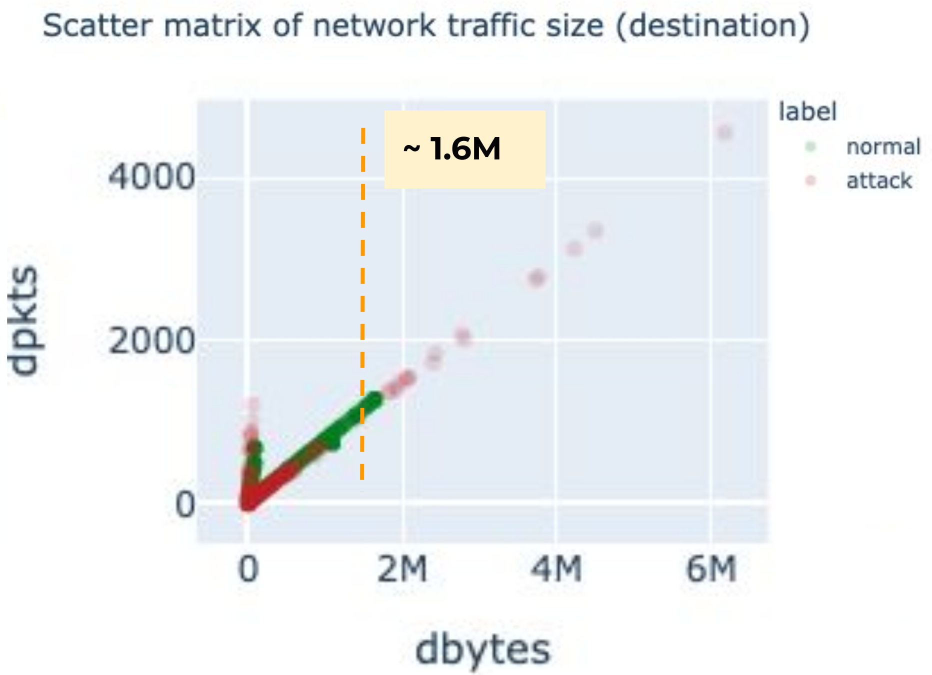


Protocol, State, Service

Transaction Size

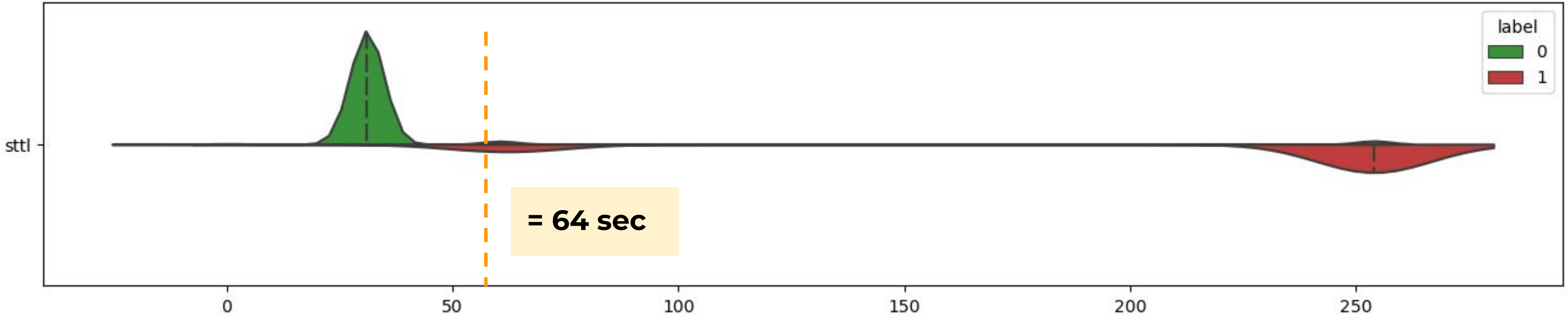
- Very strong positive linear relationship
- Attack tend to have larger transaction size for source and destination

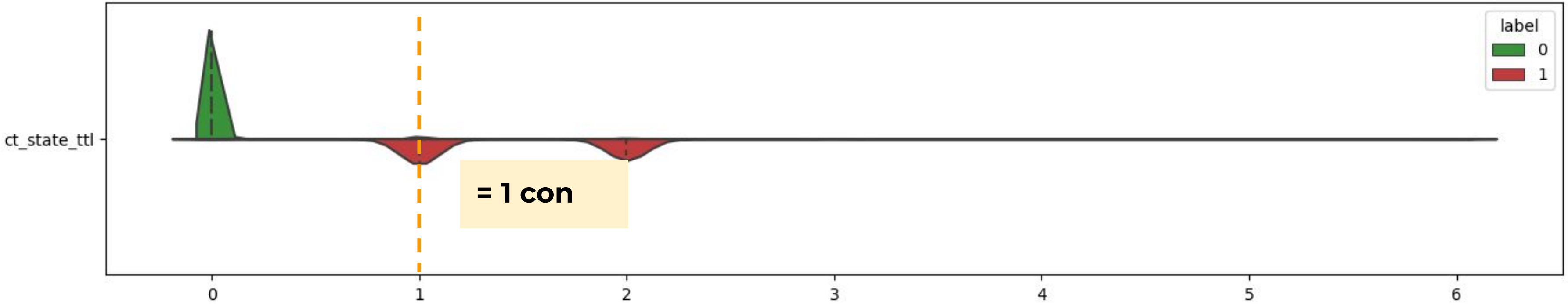




Time to live value

- Attack packets generally exist in network for a longer period before being discarded

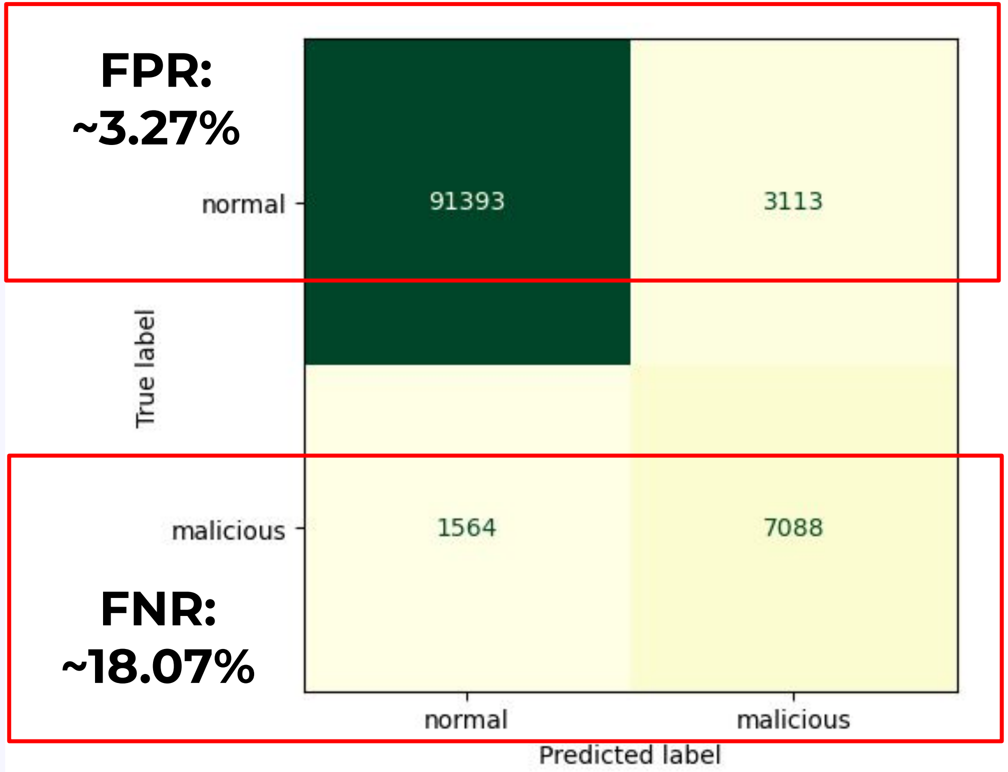




Baseline (rule-based filtering)

Classify traffic as attack using the following conditions:

- **dbytes > 1641360**
- **sbytes > 407794**
- **sttl > 64**
- **ct_state_ttl > 1**
- **service == other**
- **protocol == other**



03

Preprocessing

Preprocessing

Multicollinearity

- Remove highly correlated features using VIF

Prevent data leakages

- Remove IP addresses and attack types to prevent model from self-classifying

Preprocessing

- One-hot encode categorical features (service, protocol, state)
- Apply MinMaxScaler() since the value spread is wide for most of the features

Severely imbalance data

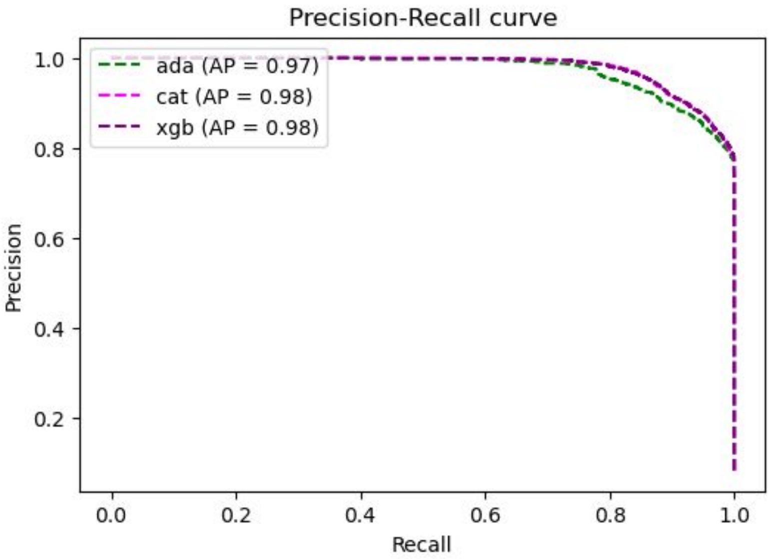
- Use algorithms that take care of class-weights

04

Model Evaluation

Model Evaluation

	model	time	train_f1	test_f1	gen_f1	train_precision	test_precision	train_recall	test_recall
0	ada	10min 19s	0.902	0.901	0.111	0.880	0.878	0.926	0.924
1	cat	7min 11s	0.919	0.893	2.829	0.850	0.816	1.000	0.985
2	xgb	9min 47s	0.921	0.908	1.412	0.932	0.914	0.910	0.903



- F1 scores are quite similar
- Generalise well
- Catboost model has the highest recall score and also runs the fastest

Problem Statement

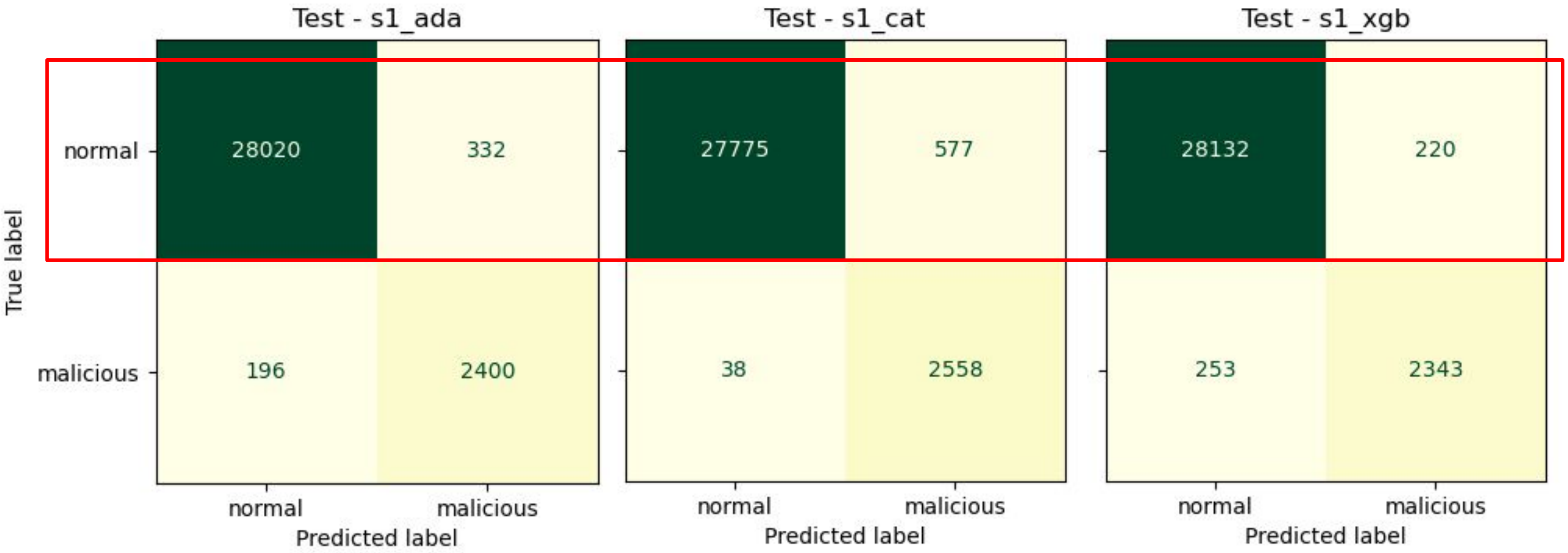
1. Improve intrusion detection rate by **reducing false negatives**
2. Reduce operational overhead by **reducing false alarm rate**

False Negative Rate

Model	AdaBoost	CatBoost	XGBoost
FNR	7.55%	1.46%	9.75%

False Positive Rate

Model	AdaBoost	CatBoost	XGBoost
FPR	1.17%	2.04%	0.78%



Test Set

Previously.....

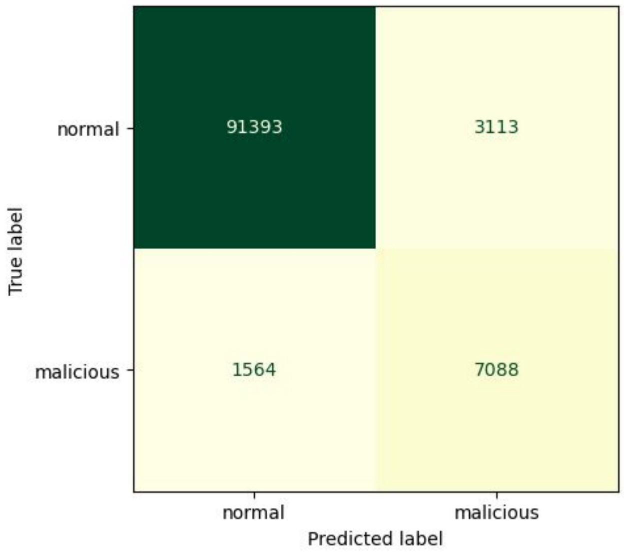
- Subset (10%) of 18 Feb 15 data was used to identify best model
- Remaining data (90%) is unseen to simulate production data

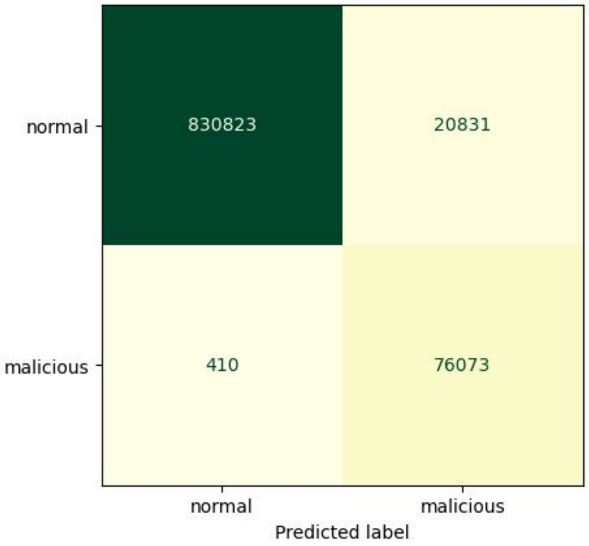
Predict

Fit the model on the train data (without train_test_split)

Test Set

	Without Model	With CatBoost Model	Improvement
FNR	18.07%	0.54%	- 17.53%
FPR	3.27%	2.45%	- 0.82%





06

Conclusion

Future Improvements

The results are very positive for a severely imbalanced dataset. This could be due to the lack of variation in the malicious network traffic.

- Use dataset with **more variety** as attacks would not originate from or target specific IP addresses in real life cases.
- Use dataset with a **longer time period** to train the model for better representation.
- Consider **time series split** and check if model generalises well with future data.

Thank You

Special thanks to
Justin, Stephen, Jeryll
and **classmates** for the
enjoyable journey