Y-DATA
SCHOOL OF DATA SCIENCE

Prediction of security policies for network DD imperva S

Team Introduction

Students	Itay Koren Ludmila Galkovskaya
Mentor	Gleb Ivashkevich (Datarythmics)
Industry partner	Johnathan Azaria (Imperva)

Company Introduction

IMPERVA

Solutions that protect users against cyber attacks

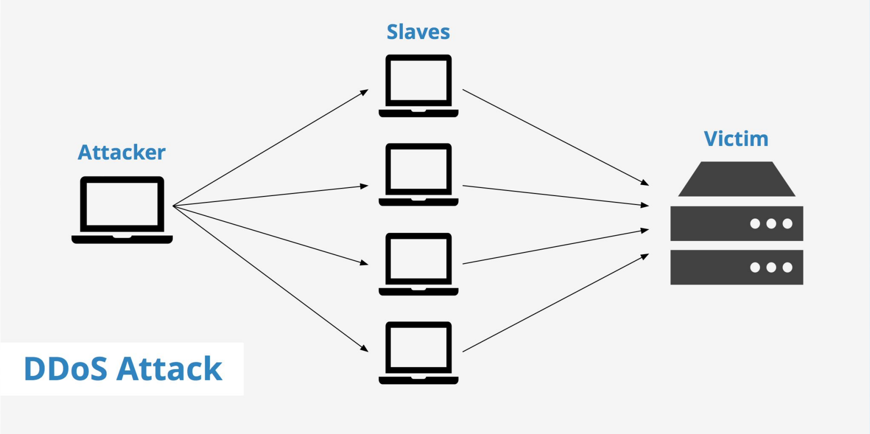
- Application Security
- Data Security
- Network Security
- Application Performance

6000+ Customers

Understanding our domain

What is DDoS?

A distributed denial of service attack, overwhelming a server, service, or network with more data than it can handle.



IP & IP Range

Start IP: 192.168.0. 1

End IP: 192.168.0. 254

« « « «

Vectors of attacks:

TCP

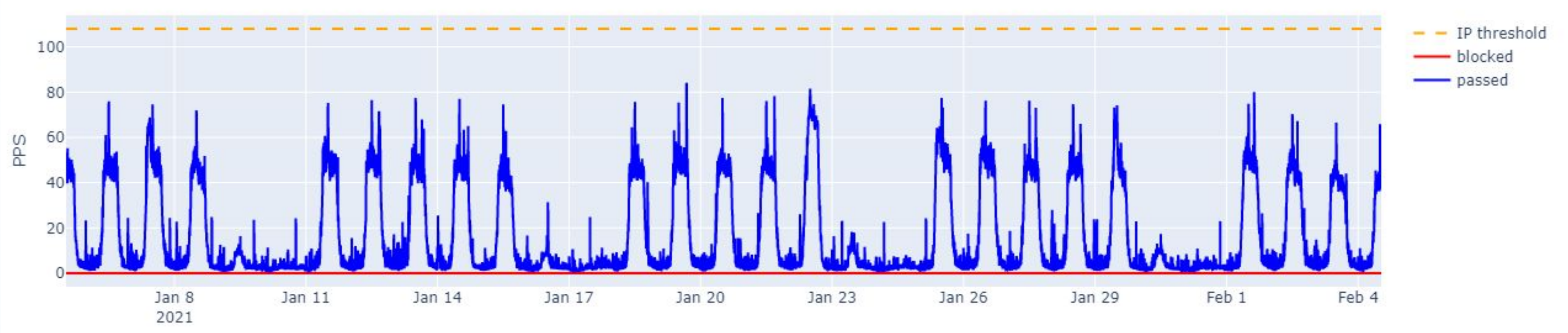
UDP

The largest attack mitigated by Imperva ~ 1,37 Tbs

Understanding our domain

What is a Network security policy?

It is a set of thresholds, each of them activates a set of rules to handle the traffic.



Problem statement

Why do we have to mitigate DDOS attacks?	To ensure business continuity, guarantee uptime and no performance impact

$300K is the average cost of 1 hour of downtime

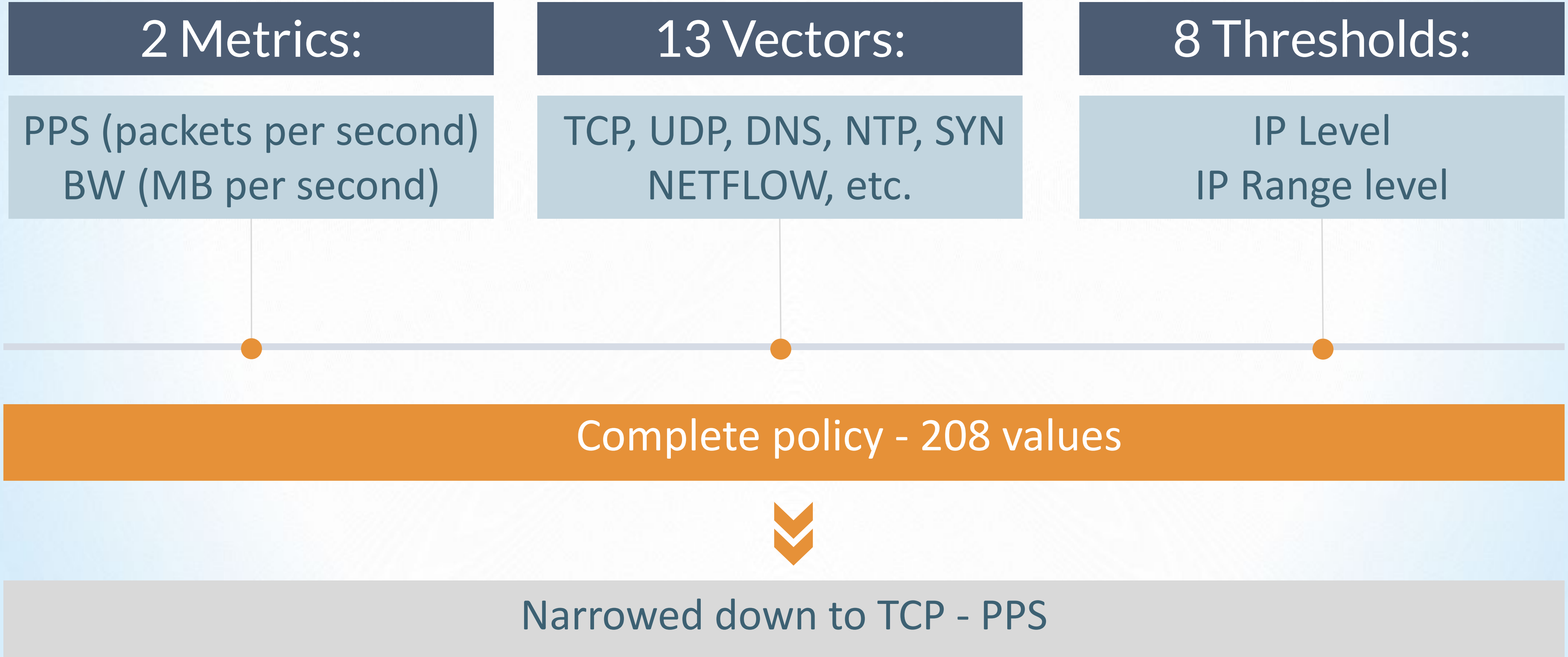
Why do we have to update security policies?	“Normal” amount of traffic changes over time
Why do we have to automate creating policies?	It is a human and time consuming task, people make mistakes

Manual creation doesn’t scale anymore

Project goals

Main	Create a model for complete network security policy
Project impact	Enable Imperva to set up-to-date, more accurate and personalized security policies

Data overview. Target



2 Metrics:

PPS (packets per second)
BW (MB per second)

13 Vectors:

TCP, UDP, DNS, NTP, SYN
NETFLOW, etc.

8 Thresholds:

IP Level
IP Range level

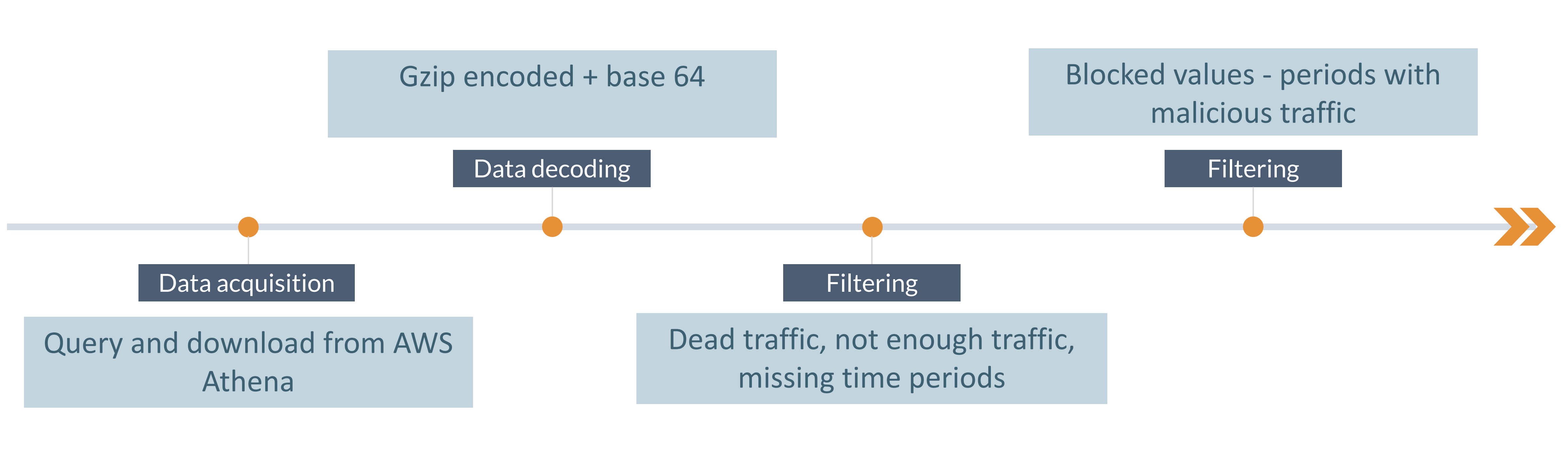
Complete policy - 208 values

Narrowed down to TCP - PPS

Data overview. Input

Data format	Parquet files		Encoded data
Size	133 000 samples		71 GB
Time series sample	1 minute intervals	max mean	30 days per sample
Metadata	Comments		Reasons

Data overview. Preprocessing



Data acquisition

Query and download from AWS Athena

Data decoding

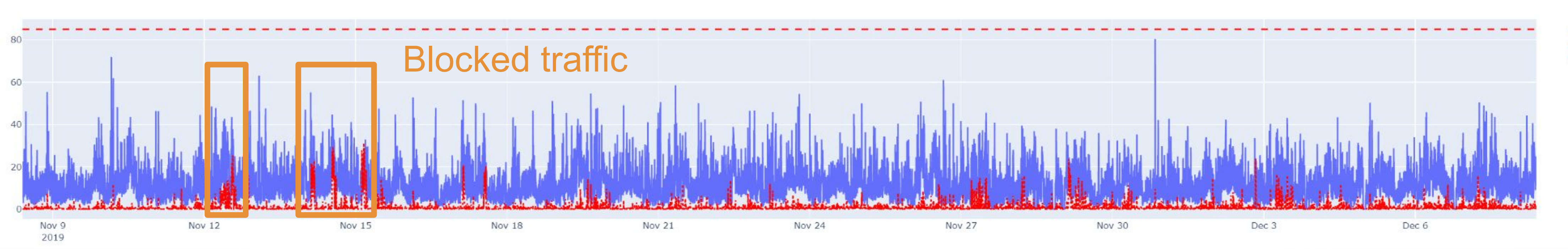
Gzip encoded + base 64

Filtering

Dead traffic, not enough traffic, missing time periods

Filtering

Blocked values - periods with malicious traffic



Blocked traffic

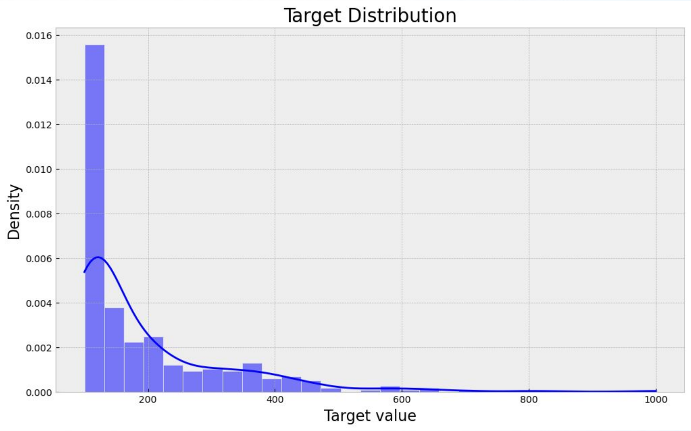
Our approach. Filtering and transformation

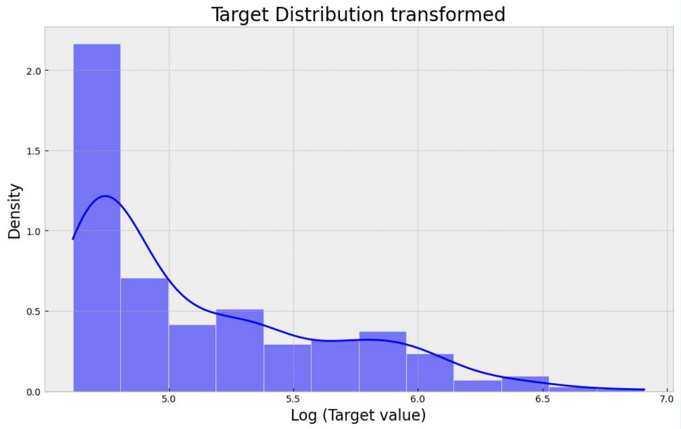
Filtering samples by:

- Target values - constant/ default values, particular ranges
- Comments/ reasons
- Irrelevant policies/ edge cases

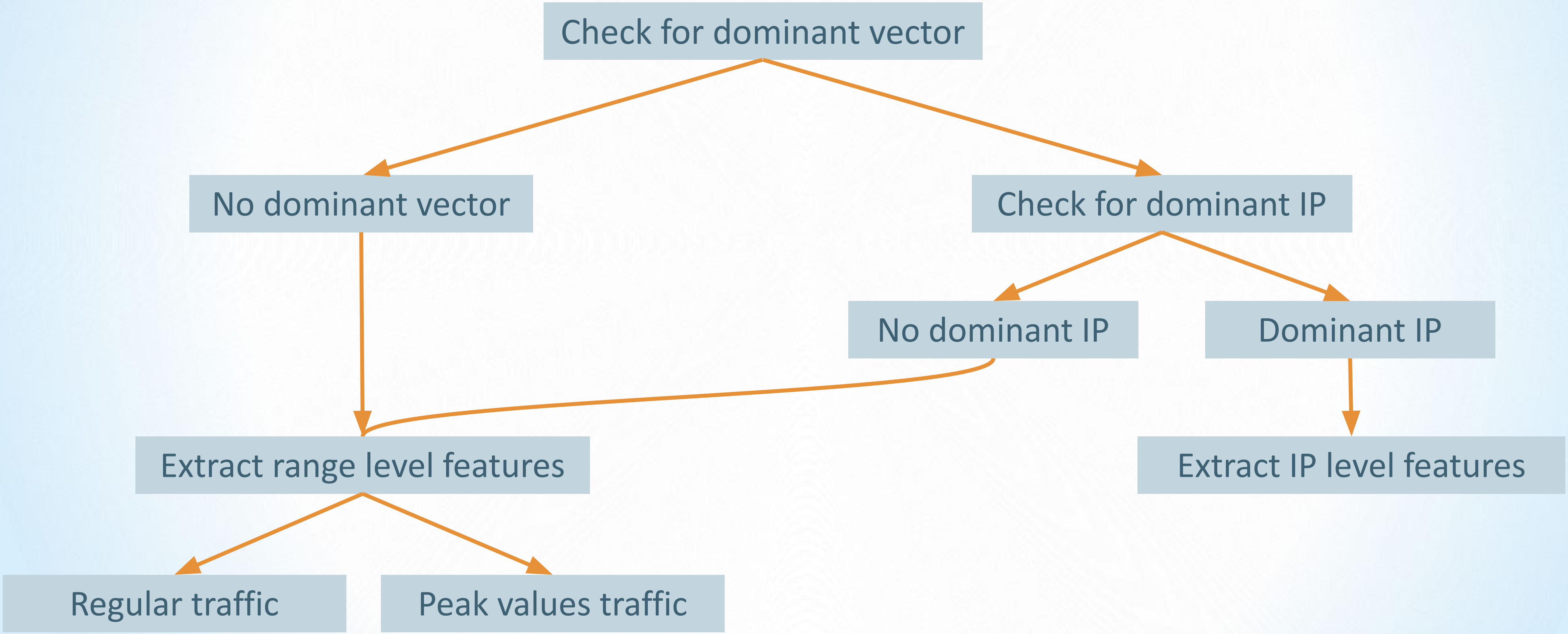
Transformation:

- Target distribution has long and light tail - used log-transform





Our approach. Feature extraction



Check for dominant vector

No dominant vector

Check for dominant IP

No dominant IP

Dominant IP

Extract range level features

Extract IP level features

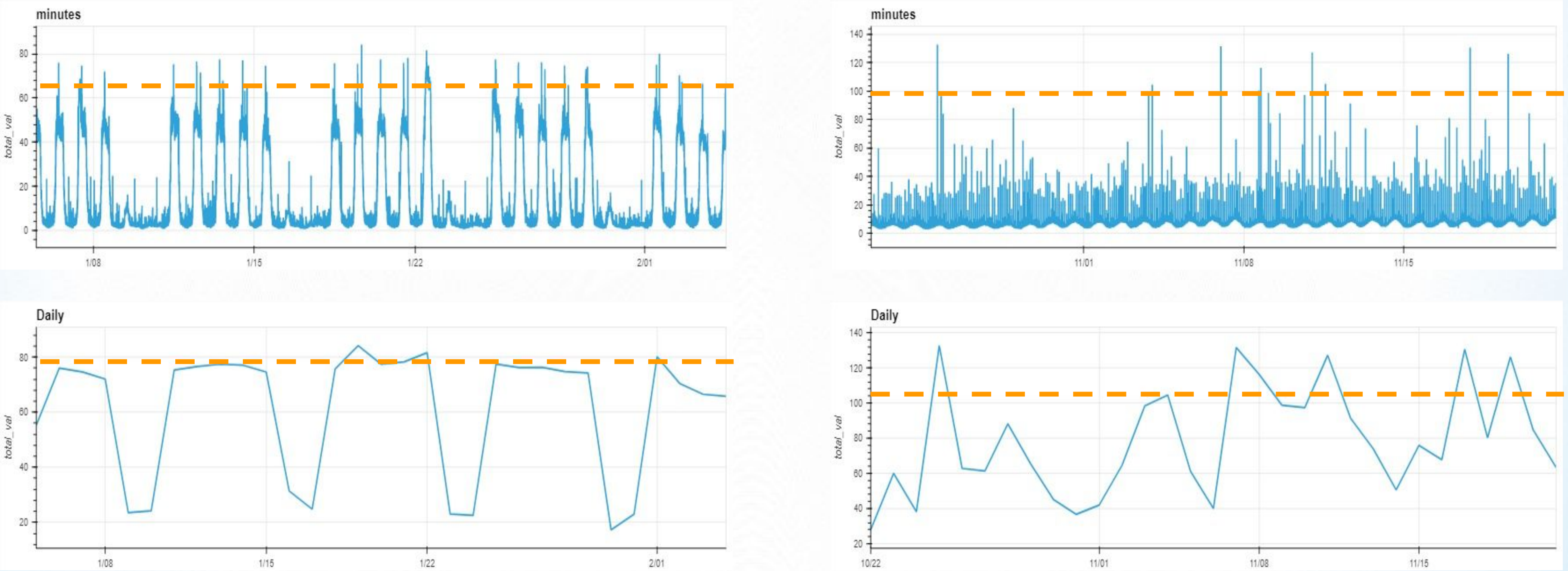
Regular traffic

Peak values traffic

Our approach. Feature extraction

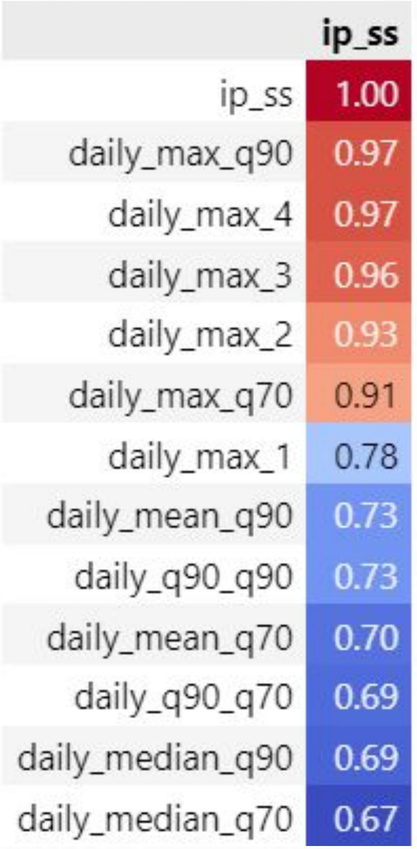
Patterns of traffic:

“Regular” “Irregular”



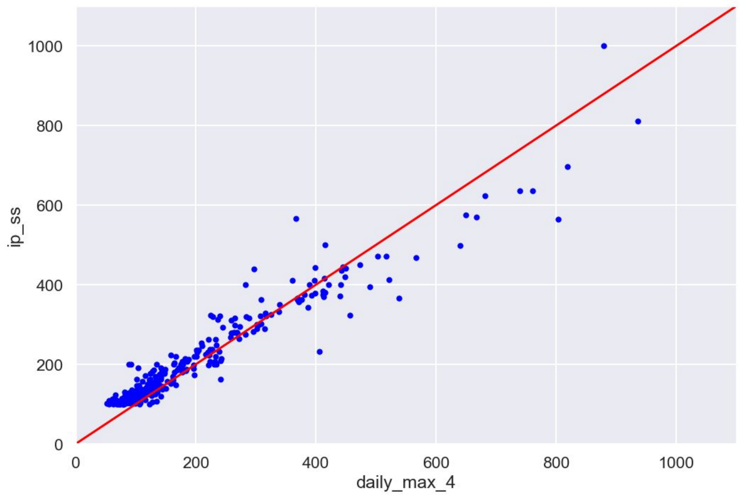
Daily Max 90th quantile - - - - - -

Our approach. Features

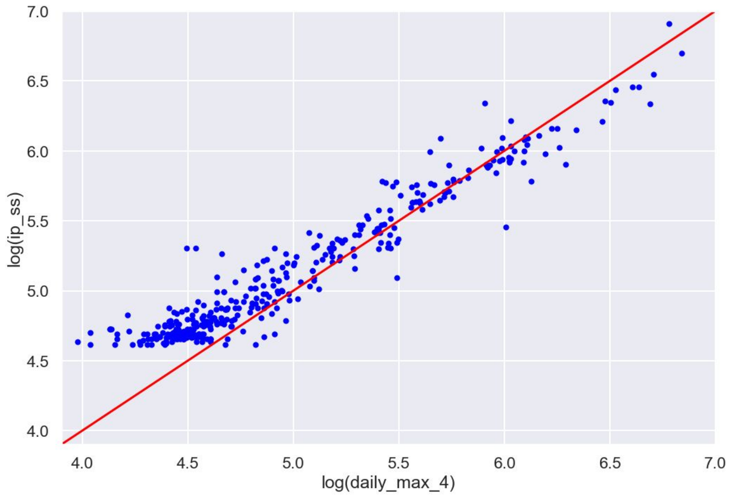


	ip_ss
ip_ss	1.00
daily_max_q90	0.97
daily_max_4	0.97
daily_max_3	0.96
daily_max_2	0.93
daily_max_q70	0.91
daily_max_1	0.78
daily_mean_q90	0.73
daily_q90_q90	0.73
daily_mean_q70	0.70
daily_q90_q70	0.69
daily_median_q90	0.69
daily_median_q70	0.67

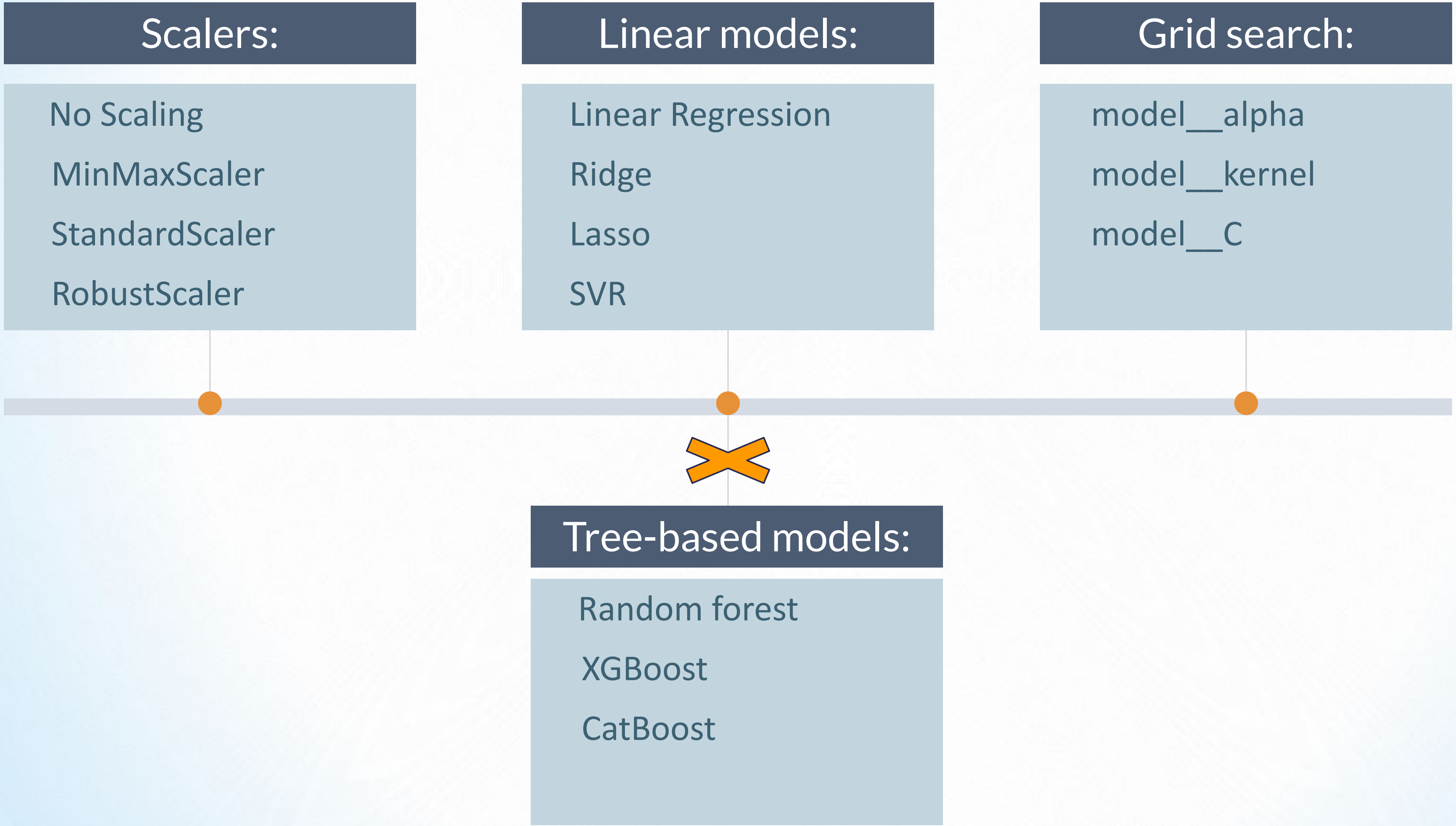
Initial



Log-transformed



Our approach. Training



Scalers:

No Scaling

MinMaxScaler

StandardScaler

RobustScaler

Linear models:

Linear Regression

Ridge

Lasso

SVR

Grid search:

model__alpha

model__kernel

model__C

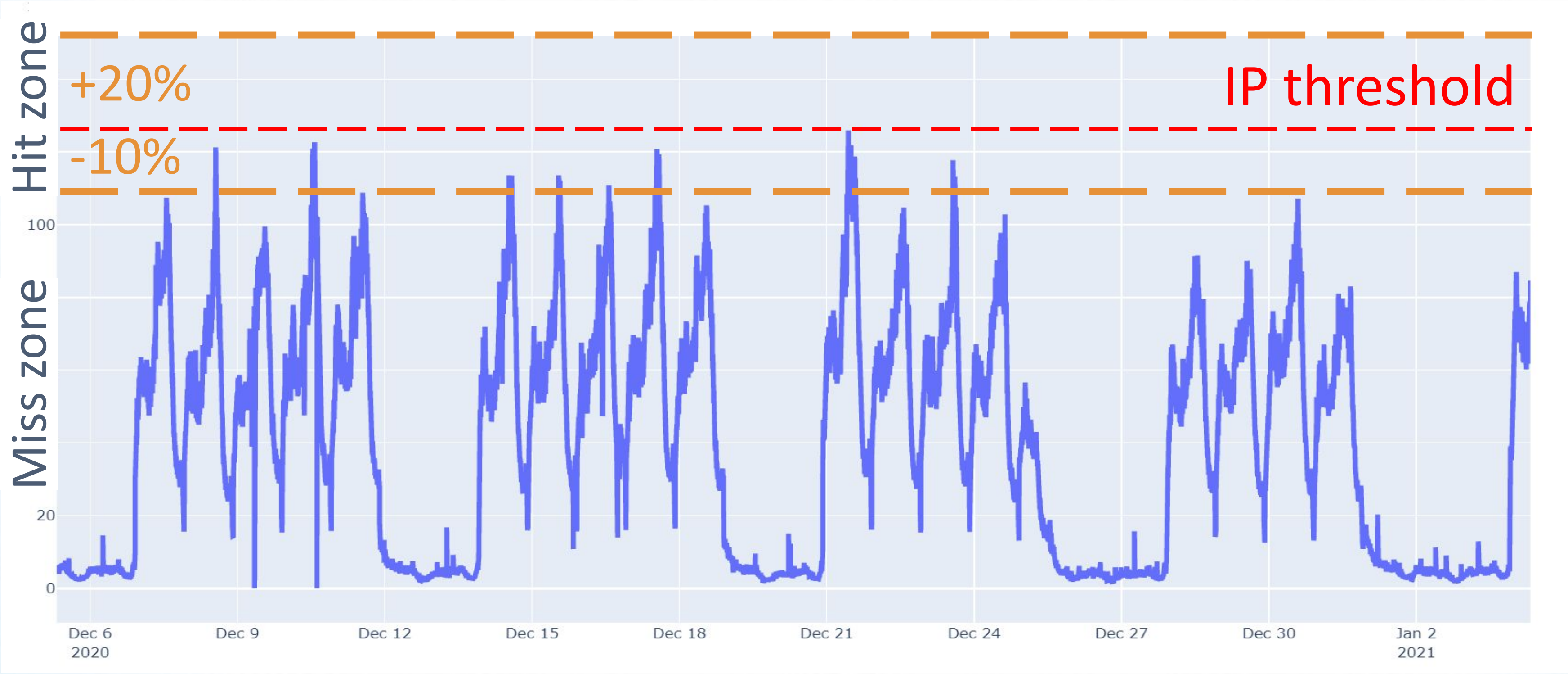
Tree-based models:

Random forest

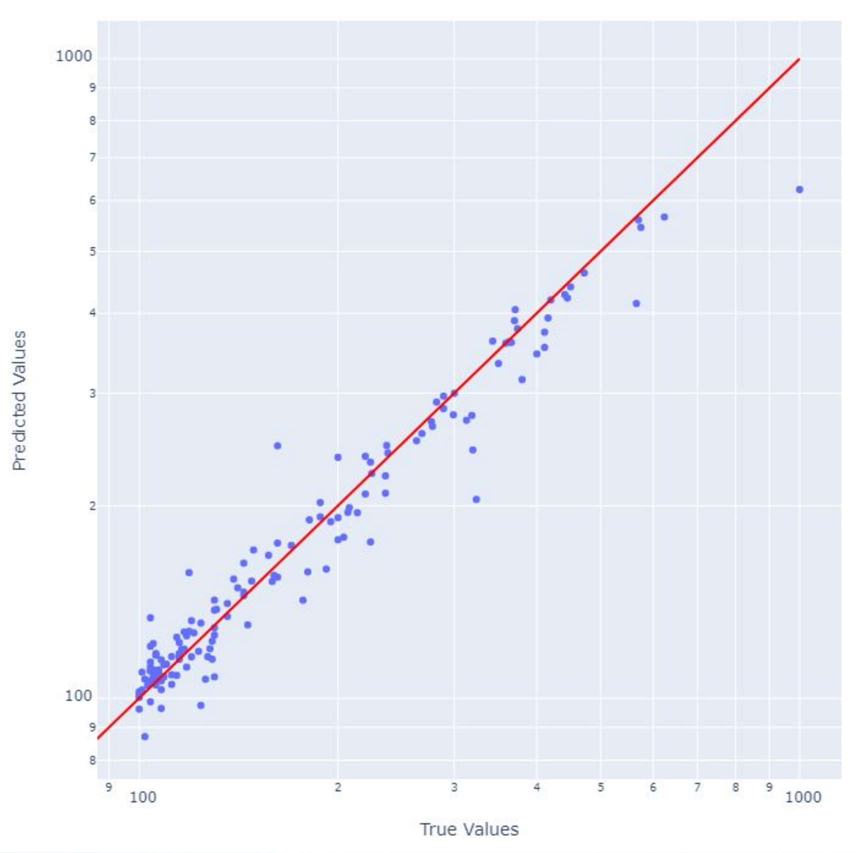
XGBoost

CatBoost

Metrics and evaluation



Project results



MinMaxScaler + SVR + symetric threshold
accuracy: 0.83 ± 0.05

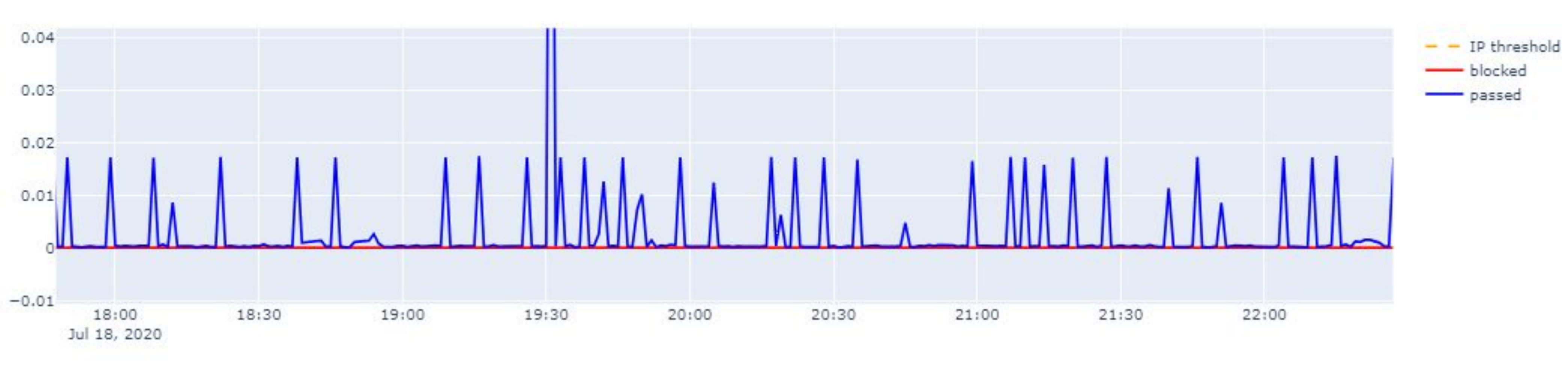
MinMaxScaler + SVR + asymmetric threshold
accuracy: 0.85 ± 0.03

Expected accuracy was ~0.9

Future work

Advanced goal:

Detect port scanning operations and handle them, so the model for security policies won't be affected



THANK YOU!

СПАСИБО!

TODA RABA!