Anomaly detection in a time series

EDAN70 -Project in Computer Science

2023-4

LUND UNIVERSITY

Supervisor: Marcus Klang
Professor: Jacek Malek

Presentation Domingo Carvajal

Content

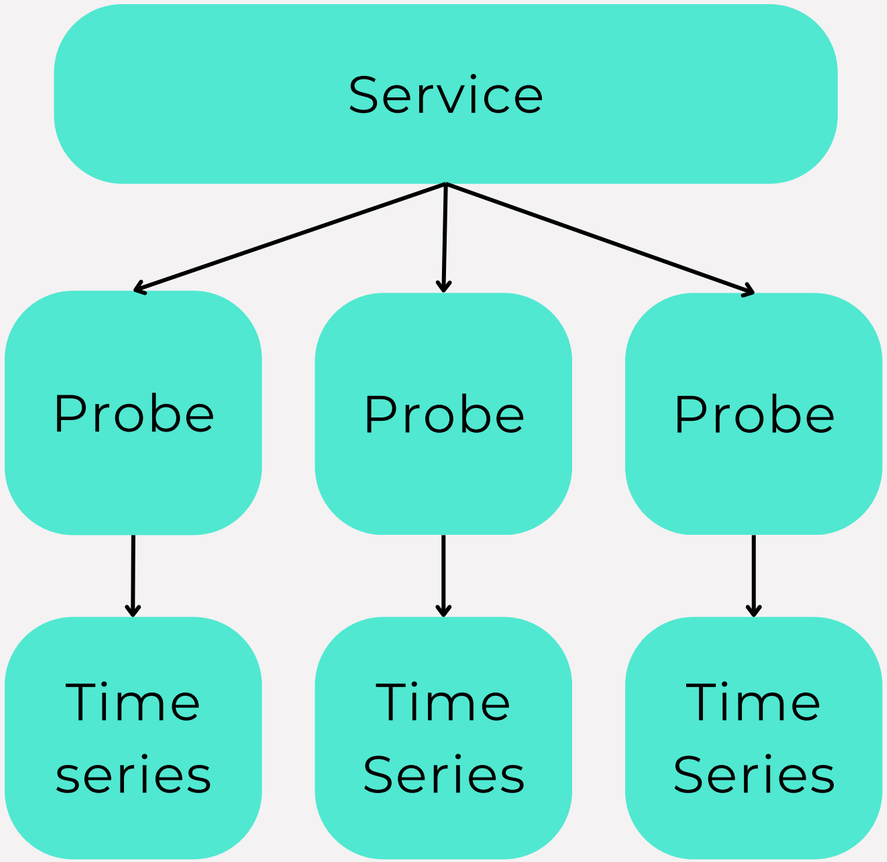
1. Approach
    - Data
    - Problem
2. Implementation
    - Preprocessing the data
    - Solution
3. Results
4. Model Validation
5. Conclusion

LUND UNIVERSITY

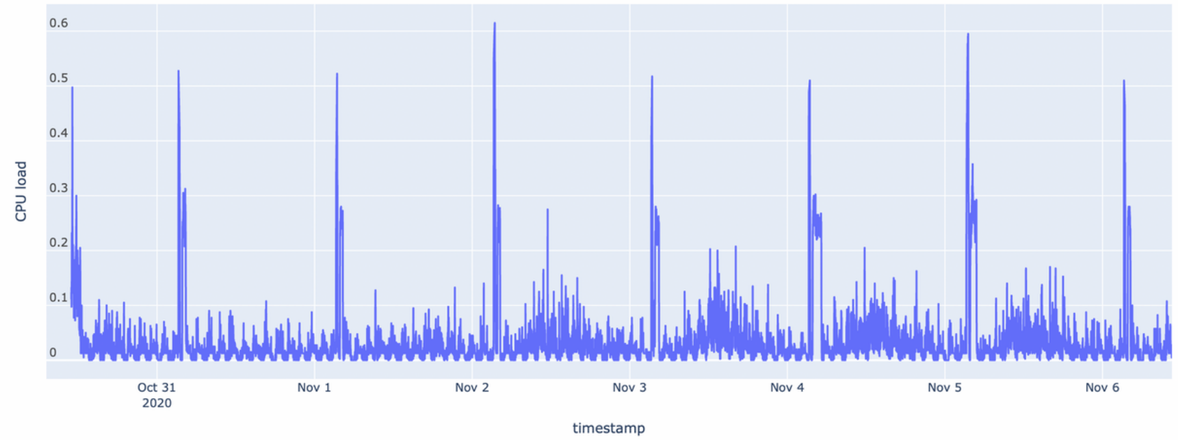
1.Approach

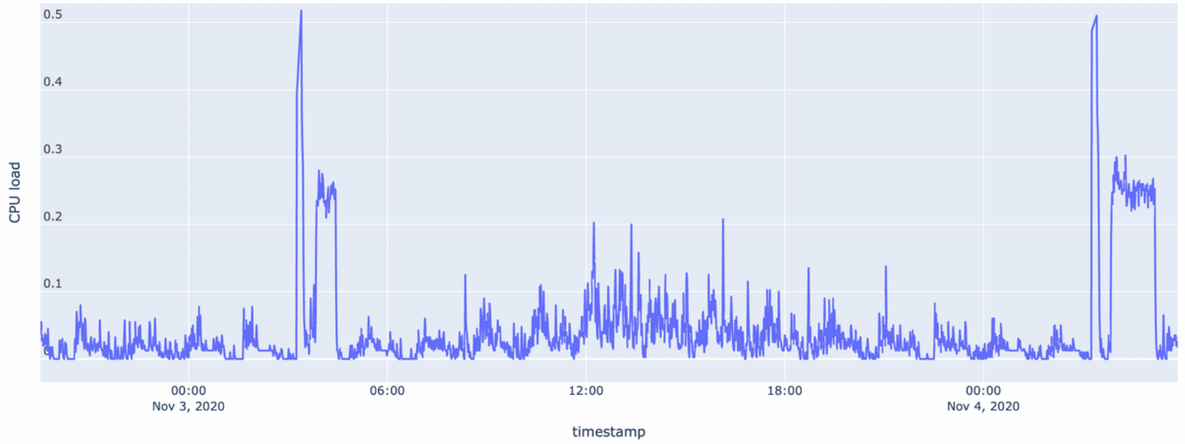
Dataset

Department of Computer Science Servers

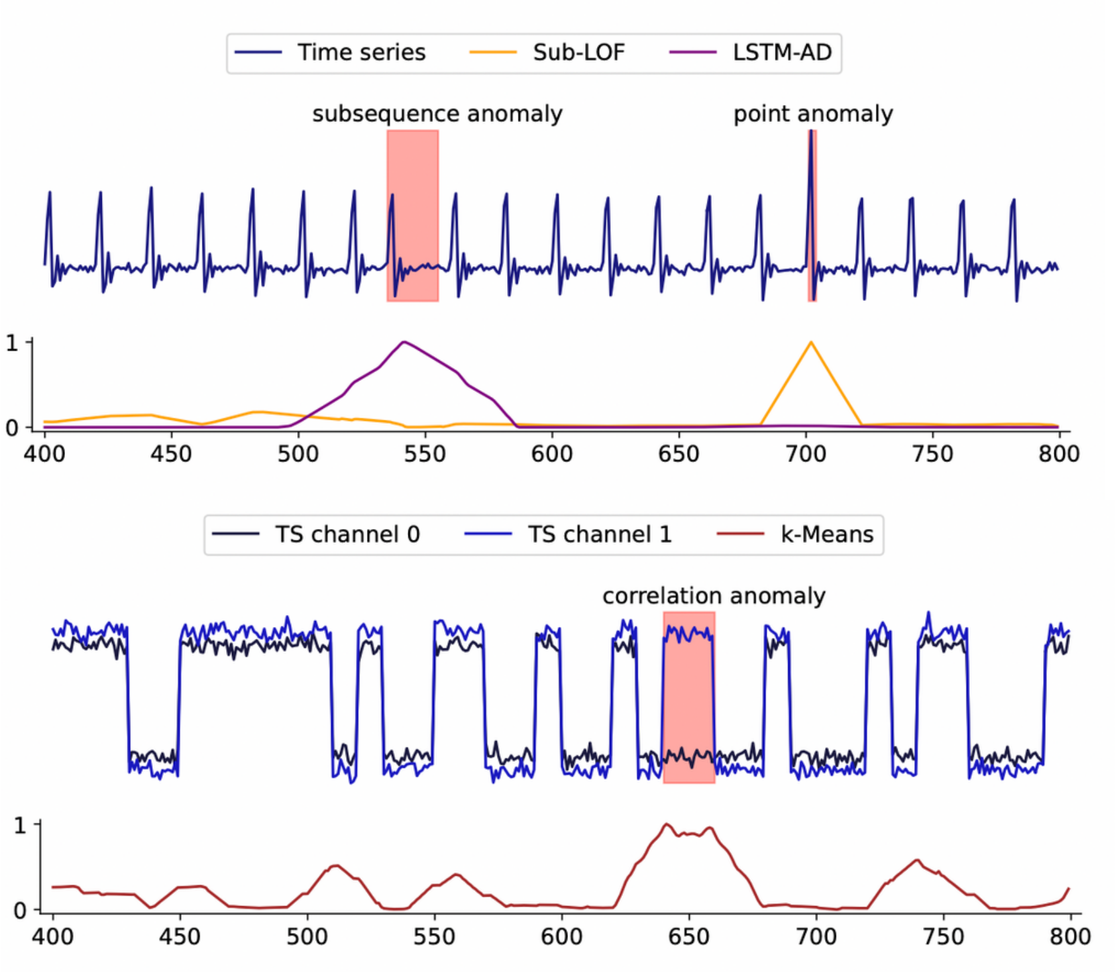


Timeseries





Anomaly Detection



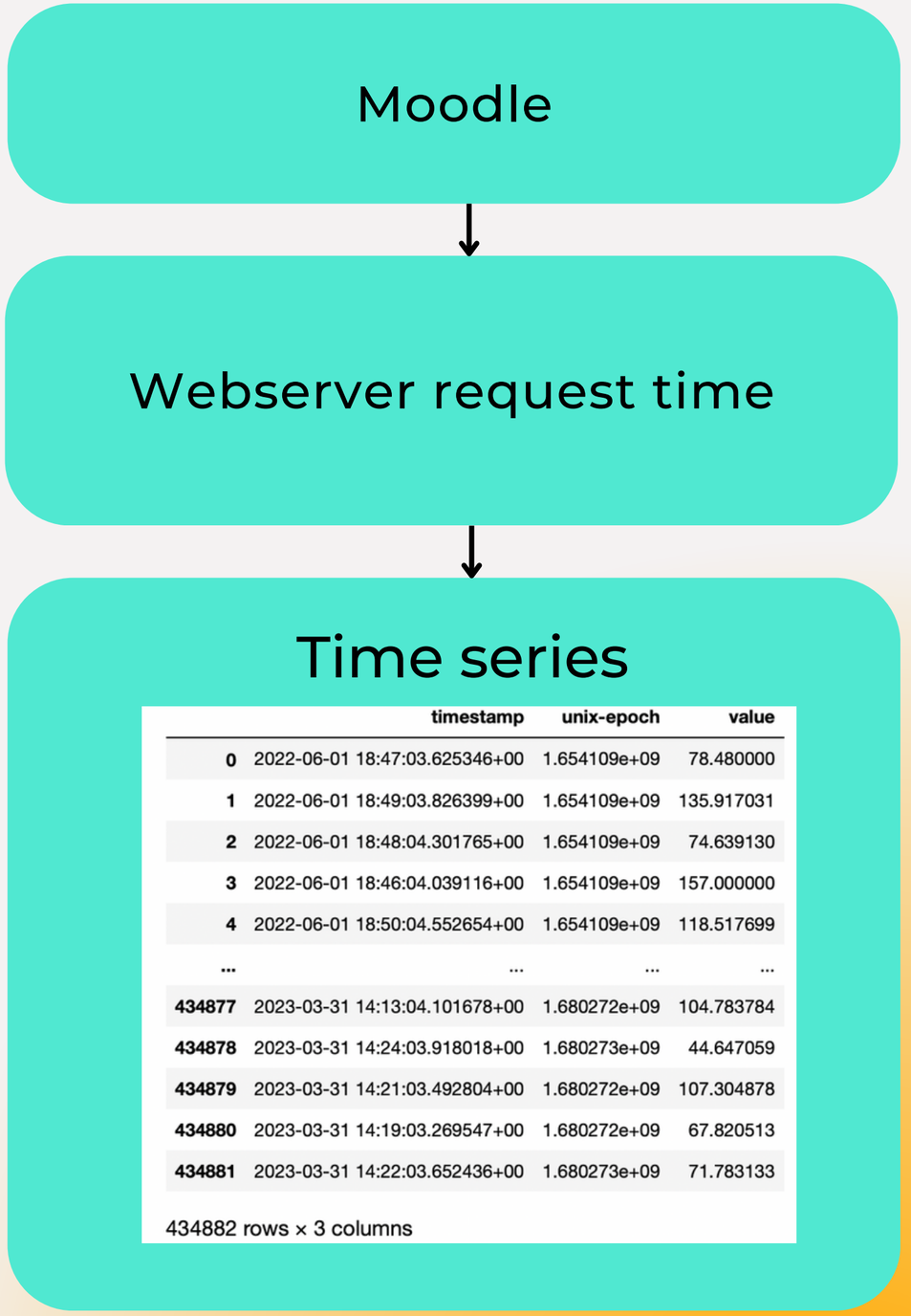
Anomaly Types

1. Subsequence
2. Point
3. Correlation

Choosing the time series

Over 1000 time series
to choose from

- Service: **Moodle**
- Probe: **Webserver request time**
- In miliseconds
- From June 2022 to March 2023



	timestamp	unix-epoch	value
0	2022-06-01 18:47:03.625346+00	1.654109e+09	78.480000
1	2022-06-01 18:49:03.826399+00	1.654109e+09	135.917031
2	2022-06-01 18:48:04.301765+00	1.654109e+09	74.639130
3	2022-06-01 18:46:04.039116+00	1.654109e+09	157.000000
4	2022-06-01 18:50:04.552654+00	1.654109e+09	118.517699
...	...	...	...
434877	2023-03-31 14:13:04.101678+00	1.680272e+09	104.783784
434878	2023-03-31 14:24:03.918018+00	1.680273e+09	44.647059
434879	2023-03-31 14:21:03.492804+00	1.680272e+09	107.304878
434880	2023-03-31 14:19:03.269547+00	1.680272e+09	67.820513
434881	2023-03-31 14:22:03.652436+00	1.680273e+09	71.783133

434882 rows × 3 columns

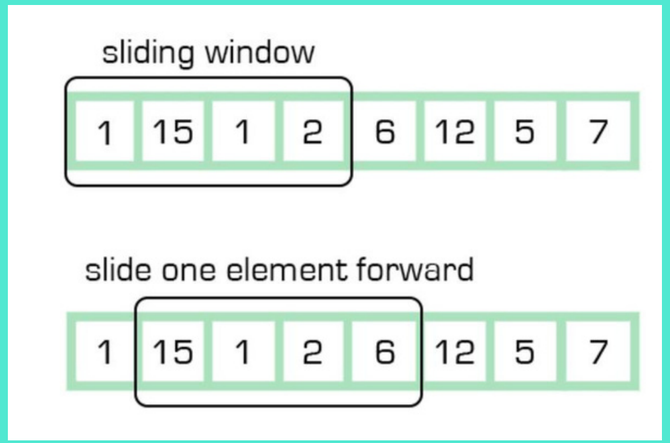
Preprocessing the data

	datetime	value
0	2022-06-01 16:21:00	73.074380
1	2022-06-01 16:22:00	68.822115
2	2022-06-01 16:23:00	70.881279
3	2022-06-01 16:24:00	88.367742
4	2022-06-01 16:25:00	69.131068
...	...	...
436199	2023-03-31 14:20:00	107.853403
436200	2023-03-31 14:21:00	107.304878
436201	2023-03-31 14:22:00	71.783133
436202	2023-03-31 14:23:00	76.673267
436203	2023-03-31 14:24:00	44.647059

Resampling

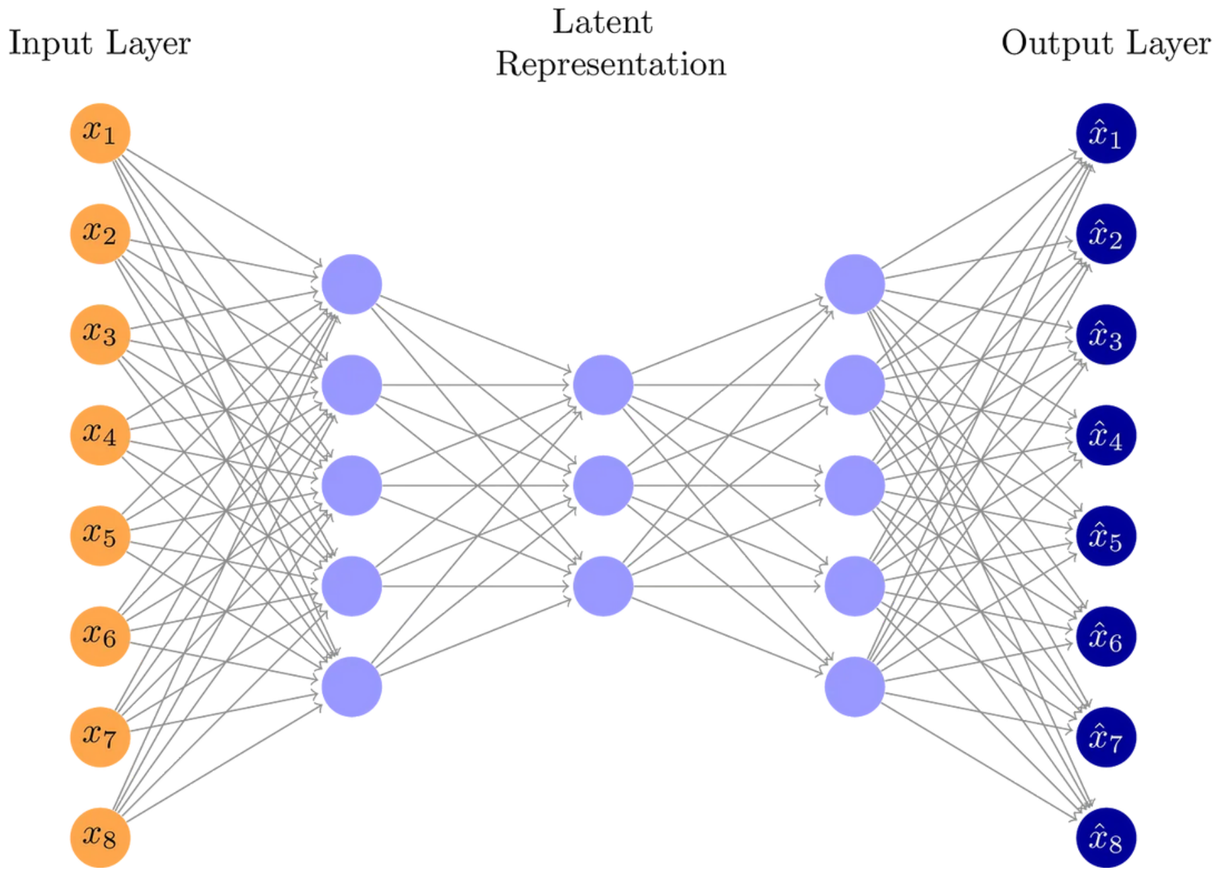
- Hours that had less than **30** values were not considered
- Interpolation if there were missing values

Filtering & Interpolating



Sliding Window

Auto Encoder

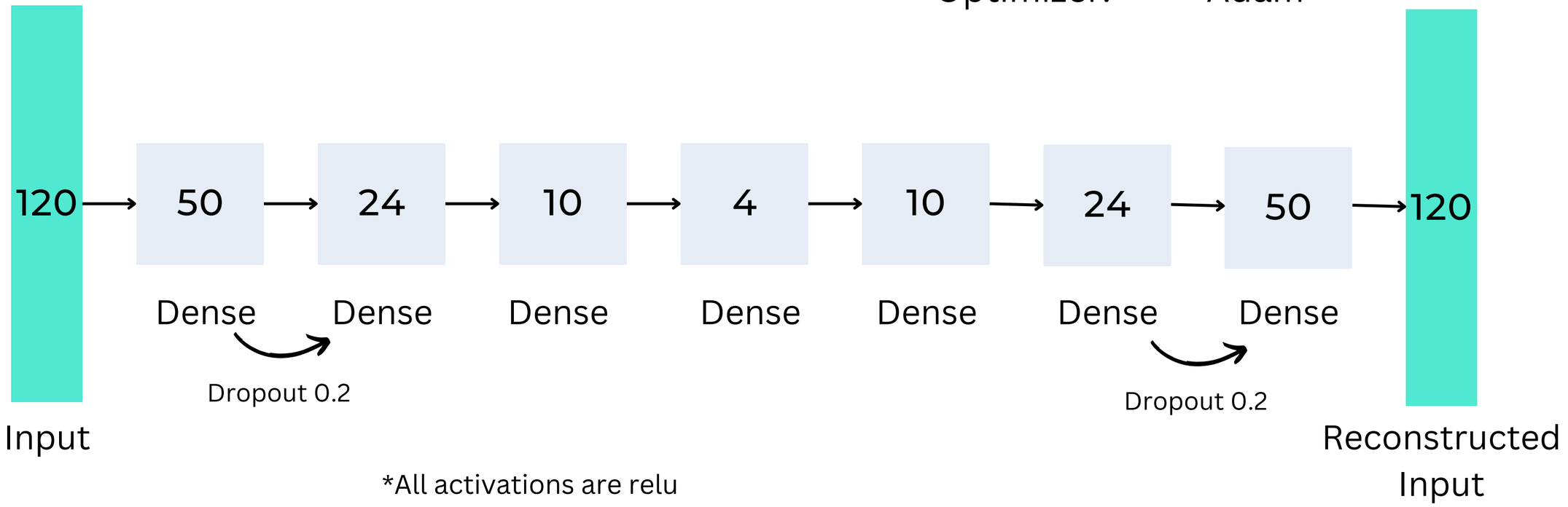


- Viable Unsupervised Learning
- Reconstructs Patterns with key features
- Anomalies will differ from the predicted pattern

Architecture

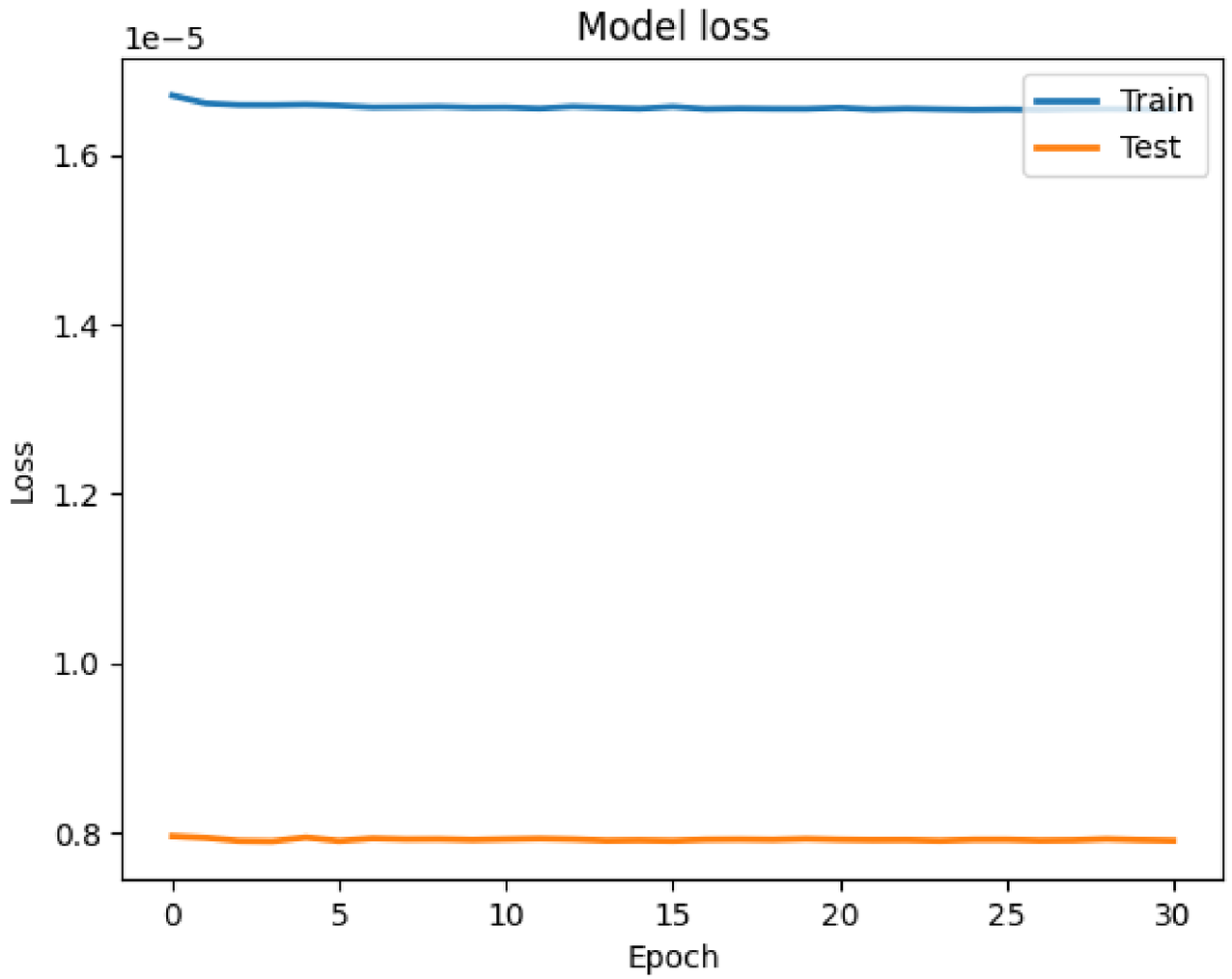
Tensorflow

Epochs:	50
Learning Rate:	1e-7
Loss:	MSE
Optimizer:	Adam



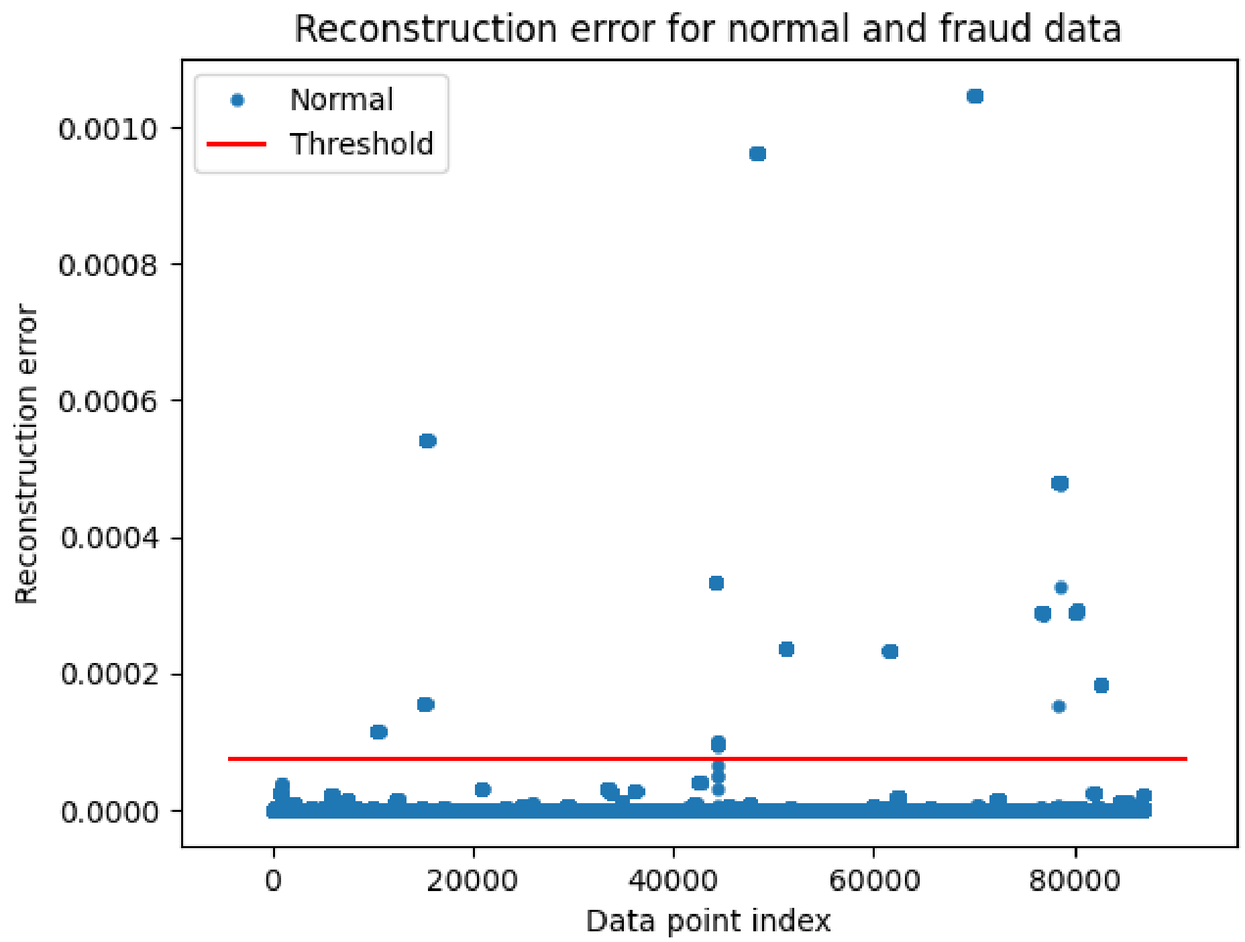
*All activations are relu

Loss History



3. Results

Prediction

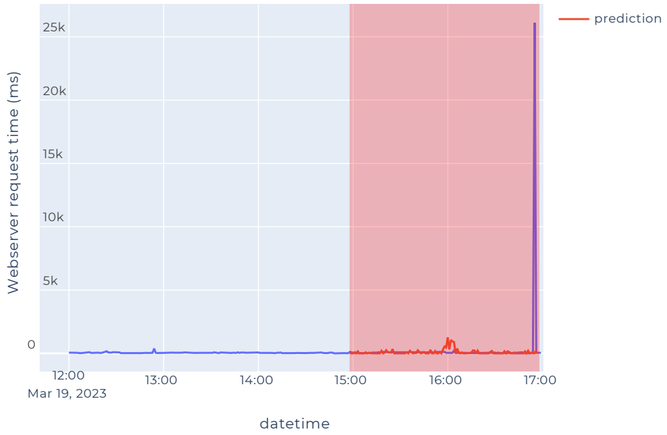


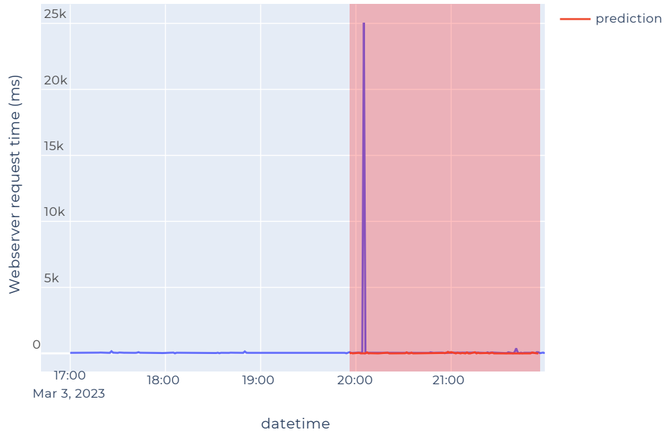
Threshold is done by a percentile value of the error in the train data.

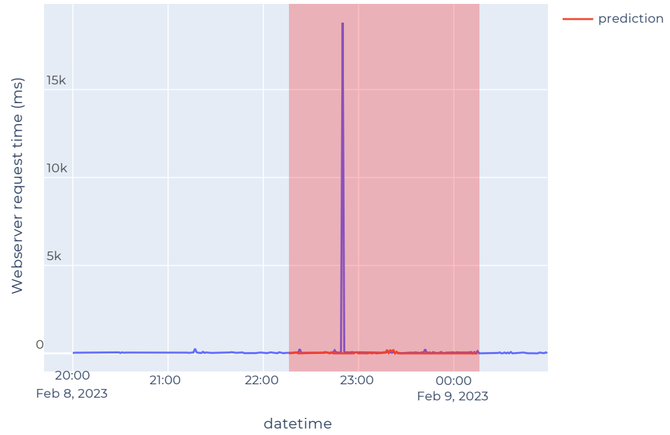
If the error in the prediction is bigger than the threshold then it is counted as an anomaly.

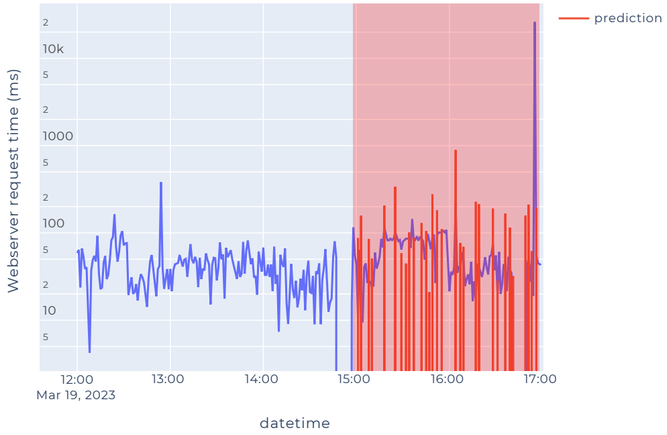
In this case the percentile used was 99%

Anomalies

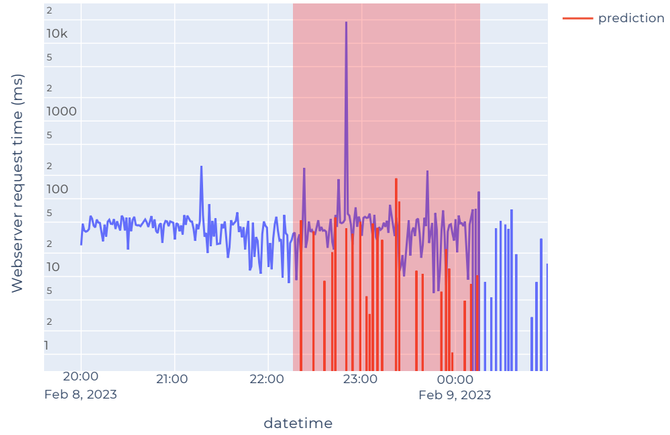

prediction
Webserver request time (ms)
25k
20k
15k
10k
5k
0
12:00
Mar 19, 2023
13:00
14:00
15:00
16:00
17:00
datetime


prediction
Webserver request time (ms)
25k
20k
15k
10k
5k
0
17:00
Mar 3, 2023
18:00
19:00
20:00
21:00
datetime

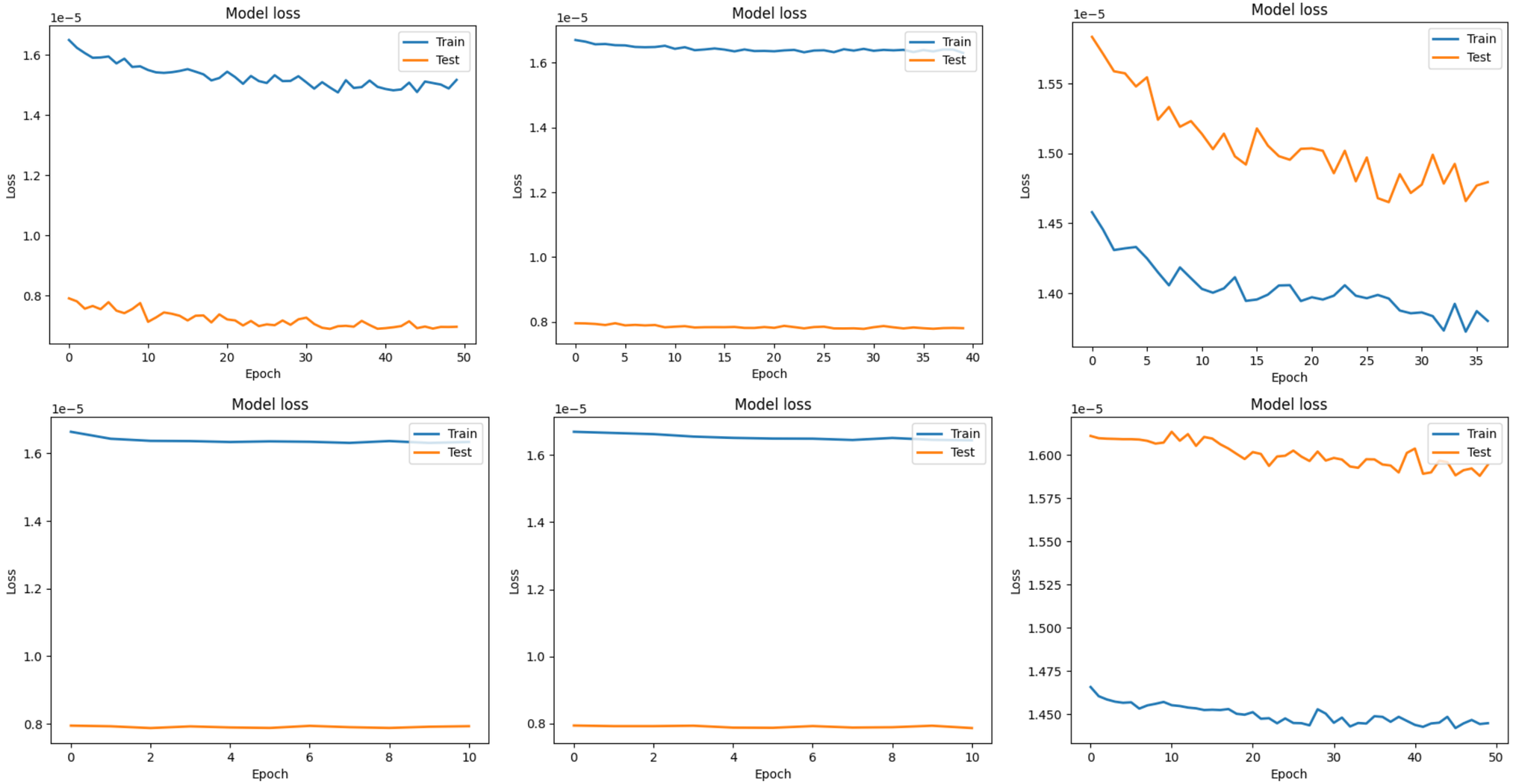

prediction
Webserver request time (ms)
15k
10k
5k
0
20:00
Feb 8, 2023
21:00
22:00
23:00
00:00
Feb 9, 2023
datetime


prediction
Webserver request time (ms)
2
10k
5
2
1000
5
2
100
5
2
10
5
12:00
Mar 19, 2023
13:00
14:00
15:00
16:00
17:00
datetime

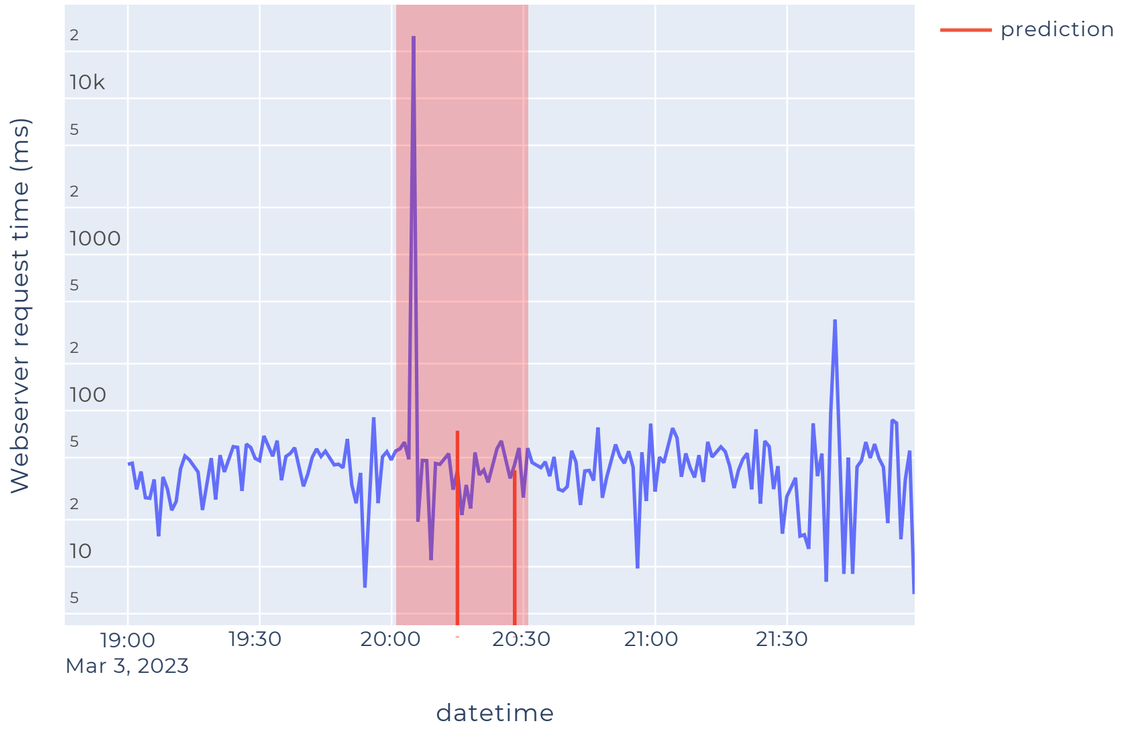

prediction
Webserver request time (ms)
2
10k
5
2
1000
5
2
100
5
2
10
5
2
1
5
17:00
Mar 3, 2023
18:00
19:00
20:00
21:00
datetime

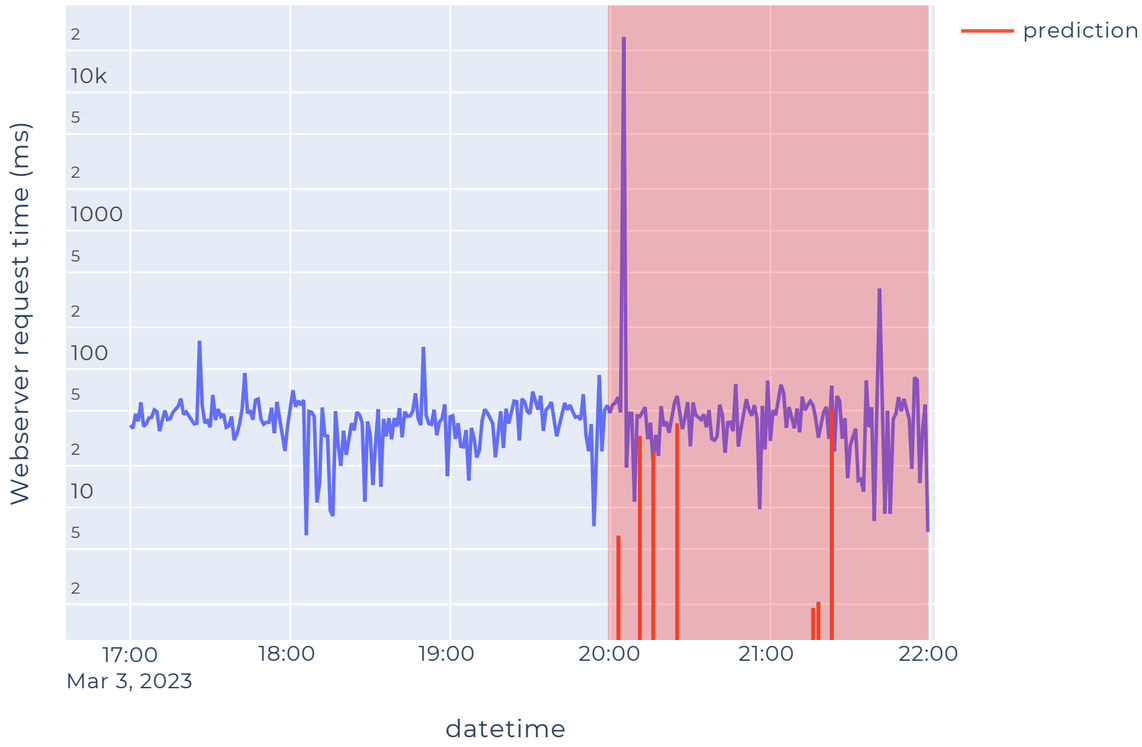

prediction
Webserver request time (ms)
2
10k
5
2
1000
5
2
100
5
2
10
5
2
1
20:00
Feb 8, 2023
21:00
22:00
23:00
00:00
Feb 9, 2023
datetime

Toggling Hyperparameters

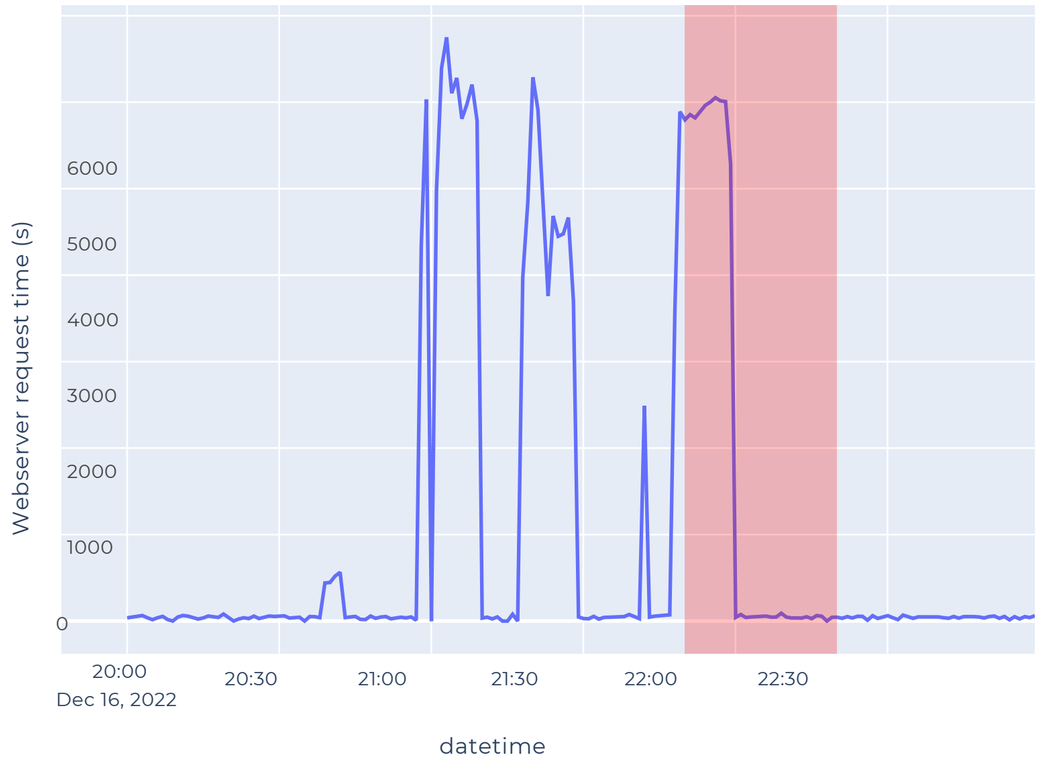


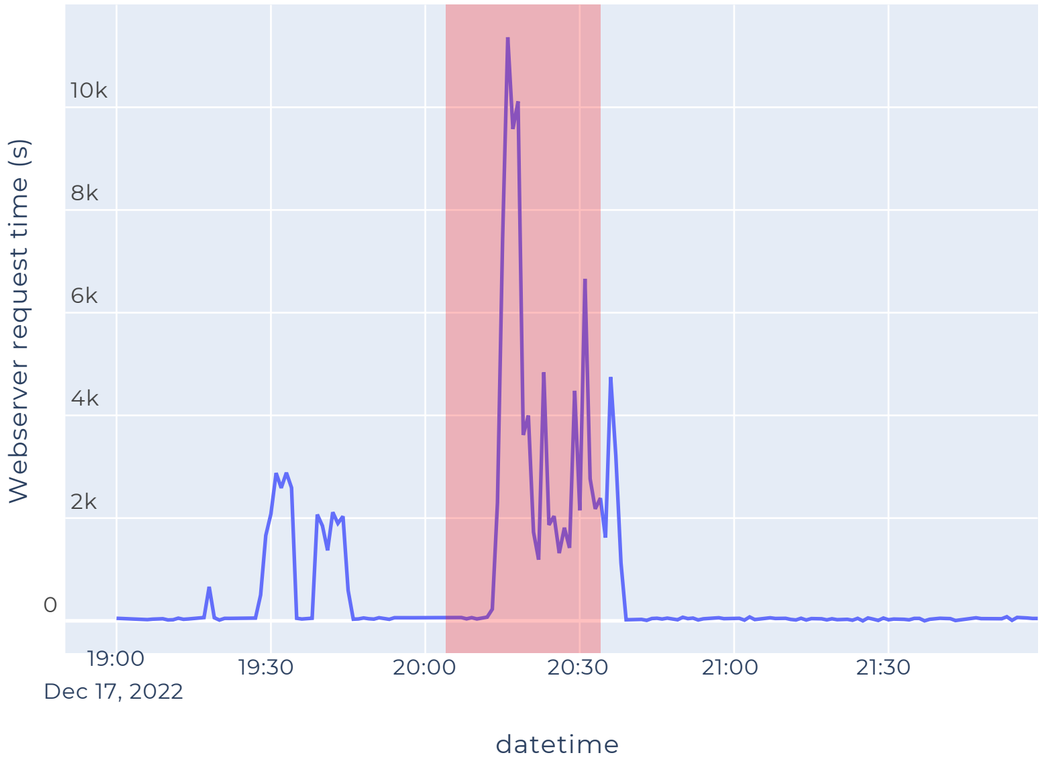
Results of previous models





Notable Anomalies



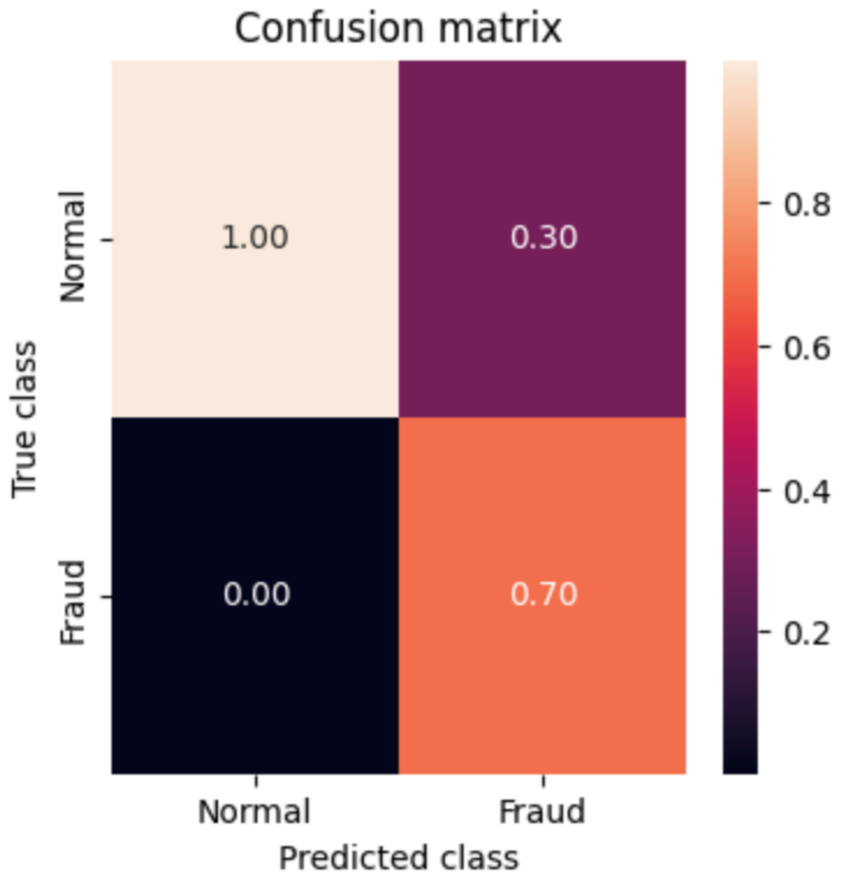


4. Model Validation

Credit Card Fraud Detection

Kaggle

- Labeled dataset
- Adjusted autoencoder to fit the data
- Adjusted threshold for a favorable f1 score
- The metrics obtained from the prediction model show high reliability.



Accuracy: 0.99840244373
Recall: 0.22330097087
Precision: 0.67647058823
F1: 0.33576642336

Conclusion

- The model is useful to find anomalies
- The model has a bias for point anomalies
- Other implementations, like LSTM could work better but time is an issue.
- Trying models with multiple time series as the input could show different results.

LUND UNIVERSITY