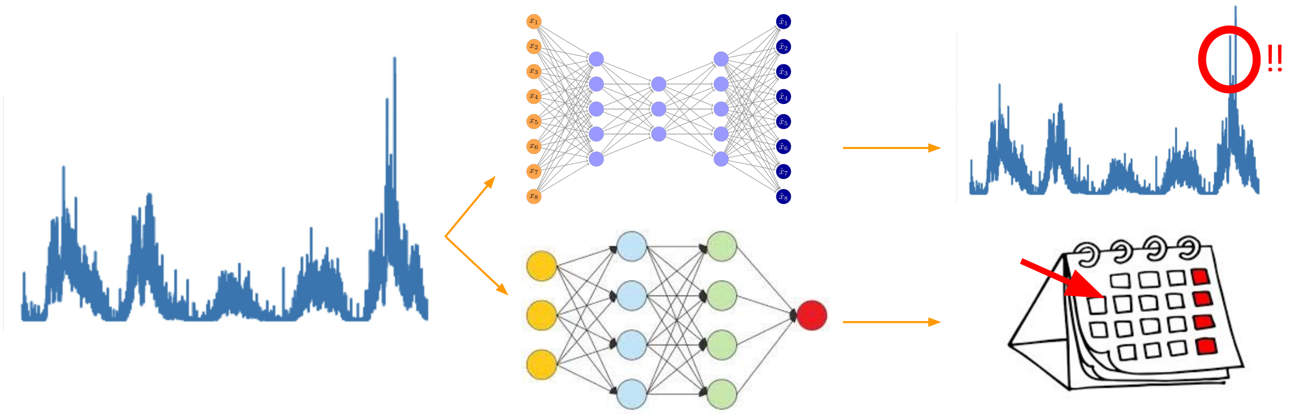
DAY PREDICTION AND ANOMALY DETECTION



Sofía Boselli and Esther Colmenar

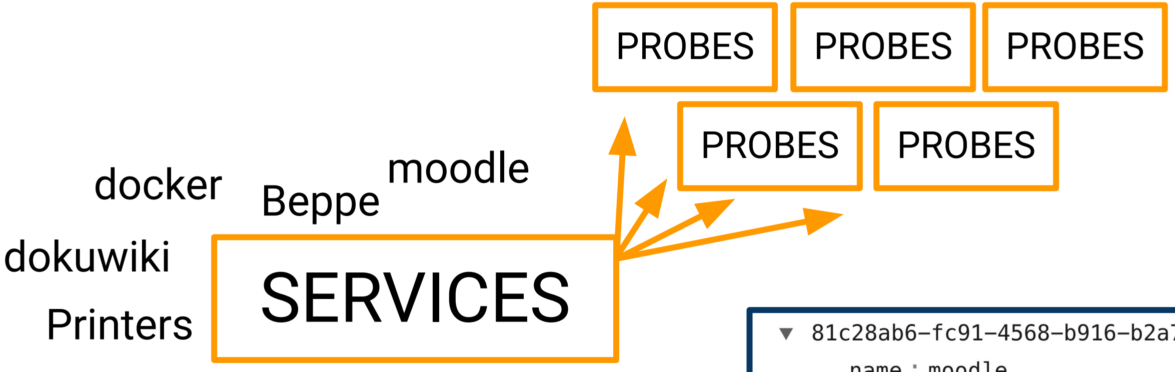
Supervisor: Marcus Klang
Project in Computer Science
LTH - Lund University

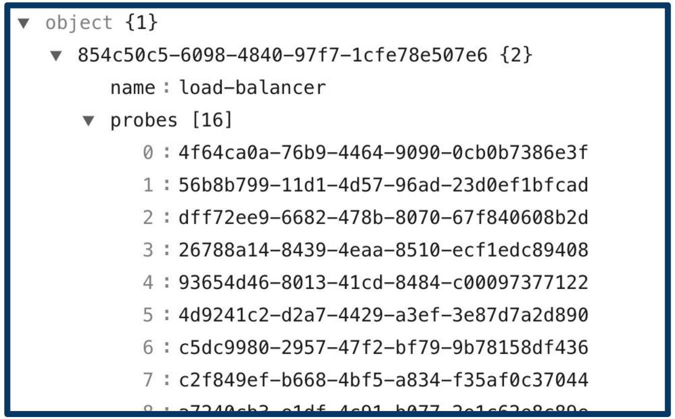
BRIEF INTRODUCTION

In this project a big amount of data from the department of computer science servers was provided and our job was to implement machine learning algorithms to achieve any goals of our choosing. In this project we focus on:

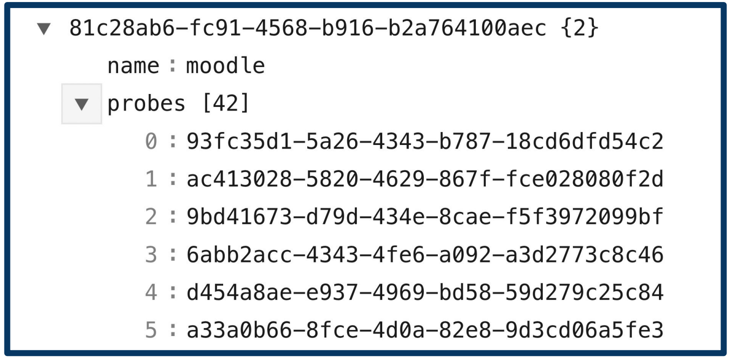
- **Day Prediction**
- **Anomaly Detection**

DATASET



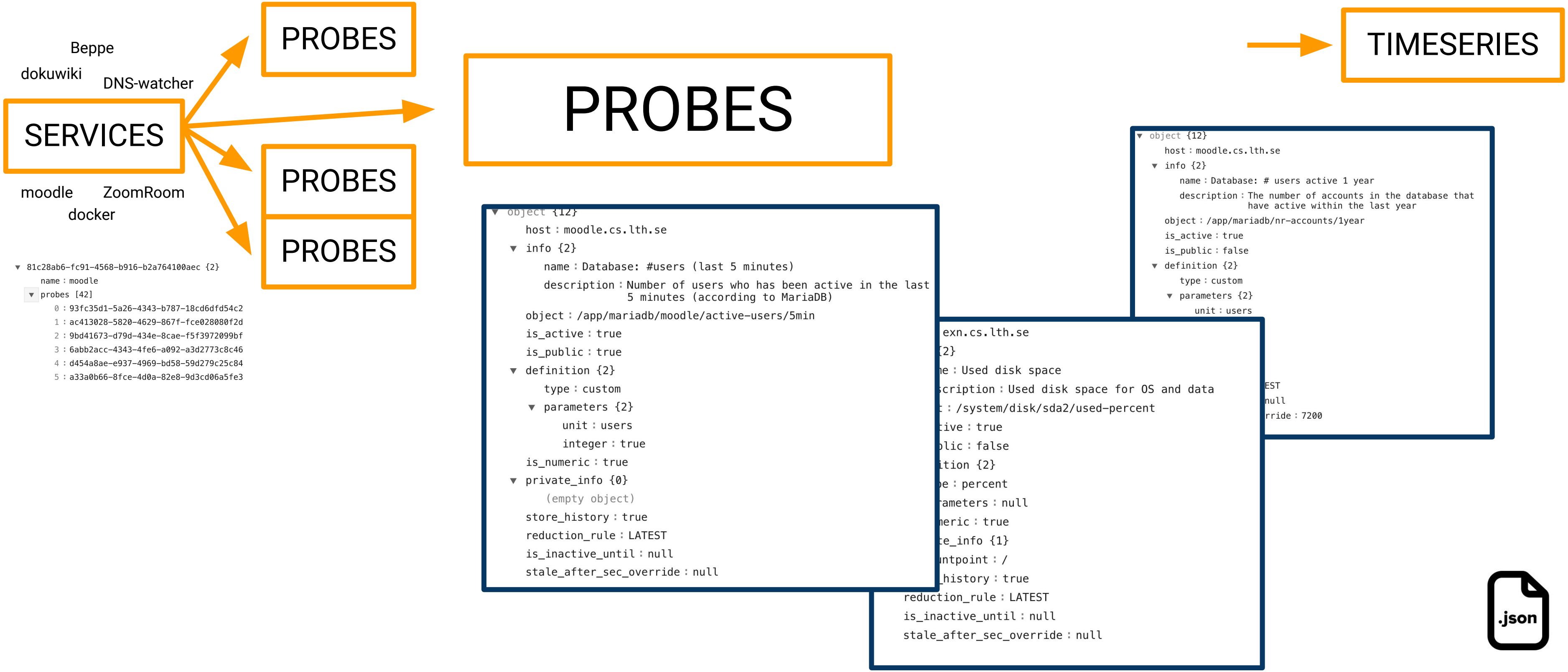


```
▼ object {1}
   ▼ 854c50c5-6098-4840-97f7-1cfe78e507e6 {2}
         name : load-balancer
      ▼ probes [16]
            0 : 4f64ca0a-76b9-4464-9090-0cb0b7386e3f
            1 : 56b8b799-11d1-4d57-96ad-23d0ef1bfcad
            2 : dff72ee9-6682-478b-8070-67f840608b2d
            3 : 26788a14-8439-4eaa-8510-ecf1edc89408
            4 : 93654d46-8013-41cd-8484-c00097377122
            5 : 4d9241c2-d2a7-4429-a3ef-3e87d7a2d890
            6 : c5dc9980-2957-47f2-bf79-9b78158df436
            7 : c2f849ef-b668-4bf5-a834-f35af0c37044
            8 : a7240cb3-e1df-4c01-b077-2e1c63e8c89e
```



```
▼ 81c28ab6-fc91-4568-b916-b2a764100aec {2}
      name : moodle
   ▼ probes [42]
         0 : 93fc35d1-5a26-4343-b787-18cd6dfd54c2
         1 : ac413028-5820-4629-867f-fce028080f2d
         2 : 9bd41673-d79d-434e-8cae-f5f3972099bf
         3 : 6abb2acc-4343-4fe6-a092-a3d2773c8c46
         4 : d454a8ae-e937-4969-bd58-59d279c25c84
         5 : a33a0b66-8fce-4d0a-82e8-9d3cd06a5fe3
```

DATASET



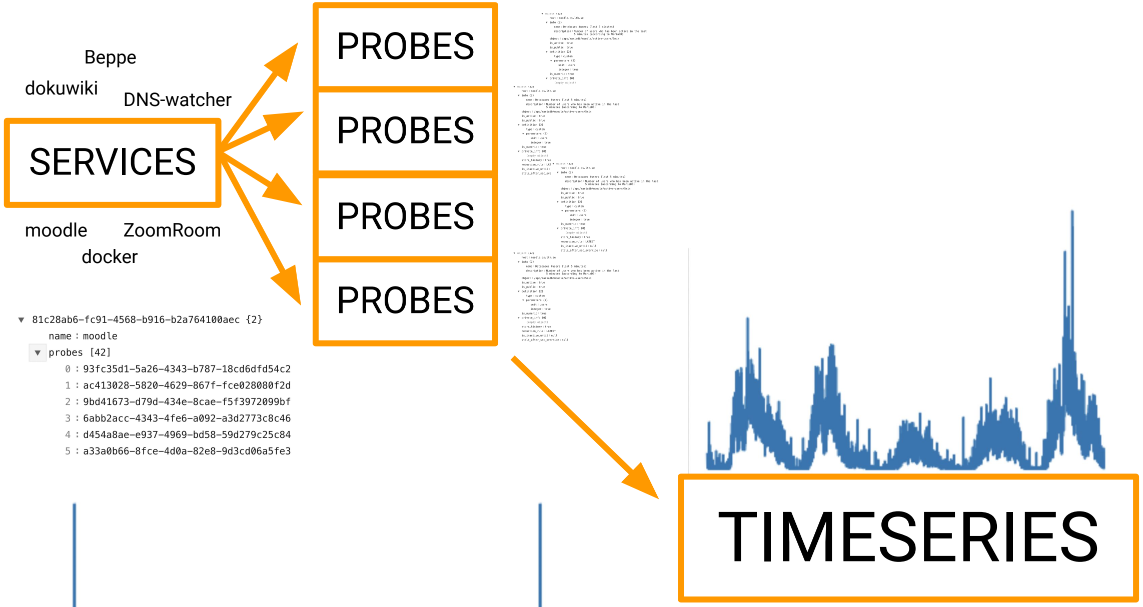
```
▼ 81c28ab6-fc91-4568-b916-b2a764100aec {2}
     name : moodle
   ▼ probes [42]
         0 : 93fc35d1-5a26-4343-b787-18cd6dfd54c2
         1 : ac413028-5820-4629-867f-fce028080f2d
         2 : 9bd41673-d79d-434e-8cae-f5f3972099bf
         3 : 6abb2acc-4343-4fe6-a092-a3d2773c8c46
         4 : d454a8ae-e937-4969-bd58-59d279c25c84
         5 : a33a0b66-8fce-4d0a-82e8-9d3cd06a5fe3
```

```
▼ object {12}
     host : moodle.cs.lth.se
   ▼ info {2}
        name : Database: #users (last 5 minutes)
        description : Number of users who has been active in the last
                          5 minutes (according to MariaDB)
     object : /app/mariadb/moodle/active-users/5min
     is_active : true
     is_public : true
   ▼ definition {2}
        type : custom
      ▼ parameters {2}
           unit : users
           integer : true
     is_numeric : true
   ▼ private_info {0}
         (empty object)
     store_history : true
     reduction_rule : LATEST
     is_inactive_until : null
     stale_after_sec_override : null
```

```
 exn.cs.lth.se
{2}
e : Used disk space
cription : Used disk space for OS and data
: /system/disk/sda2/used-percent
ive : true
lic : false
ition {2}
e : percent
ameters : null
eric : true
e_info {1}
ntpoint : /
history : true
reduction_rule : LATEST
is_inactive_until : null
stale_after_sec_override : null
```

```
▼ object {12}
     host : moodle.cs.lth.se
   ▼ info {2}
        name : Database: # users active 1 year
        description : The number of accounts in the database that
                          have active within the last year
     object : /app/mariadb/nr-accounts/1year
     is_active : true
     is_public : false
   ▼ definition {2}
        type : custom
      ▼ parameters {2}
           unit : users
EST
null
rride : 7200
```

DATASET



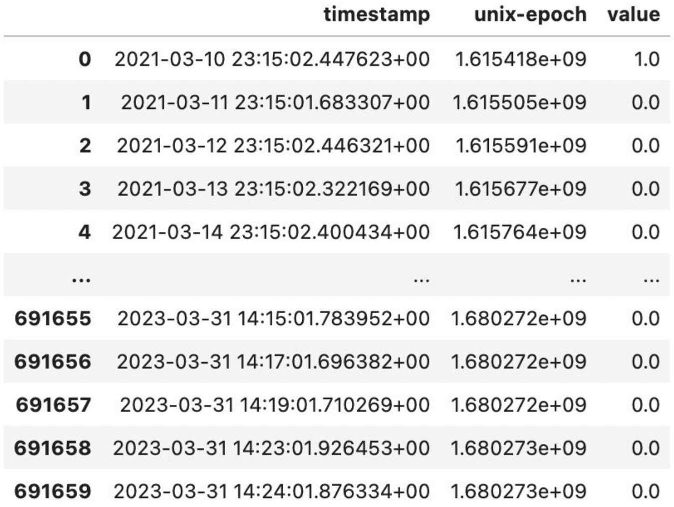
```
▼ 81c28ab6-fc91-4568-b916-b2a764100aec {2}
    name : moodle
  ▼ probes [42]
       0 : 93fc35d1-5a26-4343-b787-18cd6dfd54c2
       1 : ac413028-5820-4629-867f-fce028080f2d
       2 : 9bd41673-d79d-434e-8cae-f5f3972099bf
       3 : 6abb2acc-4343-4fe6-a092-a3d2773c8c46
       4 : d454a8ae-e937-4969-bd58-59d279c25c84
       5 : a33a0b66-8fce-4d0a-82e8-9d3cd06a5fe3
```

```
▼ object {12}
    host : moodle.cs.lth.se
  ▼ info {2}
      name : Database: #users (last 5 minutes)
      description : Number of users who has been active in the last
                    5 minutes (according to MariaDB)
    object : /app/mariadb/moodle/active-users/5min
    is_active : true
    is_public : true
  ▼ definition {2}
      type : custom
    ▼ parameters {2}
        unit : users
        integer : true
    is_numeric : true
  ▼ private_info {0}
      (empty object)
```

```
▼ object {12}
    host : moodle.cs.lth.se
  ▼ info {2}
      name : Database: #users (last 5 minutes)
      description : Number of users who has been active in the last
                    5 minutes (according to MariaDB)
    object : /app/mariadb/moodle/active-users/5min
    is_active : true
    is_public : true
  ▼ definition {2}
      type : custom
    ▼ parameters {2}
        unit : users
        integer : true
    is_numeric : true
  ▼ private_info {0}
      (empty object)
    store_history : true
    reduction_rule : LAT
    is_inactive_until :
    stale_after_sec_ove
```

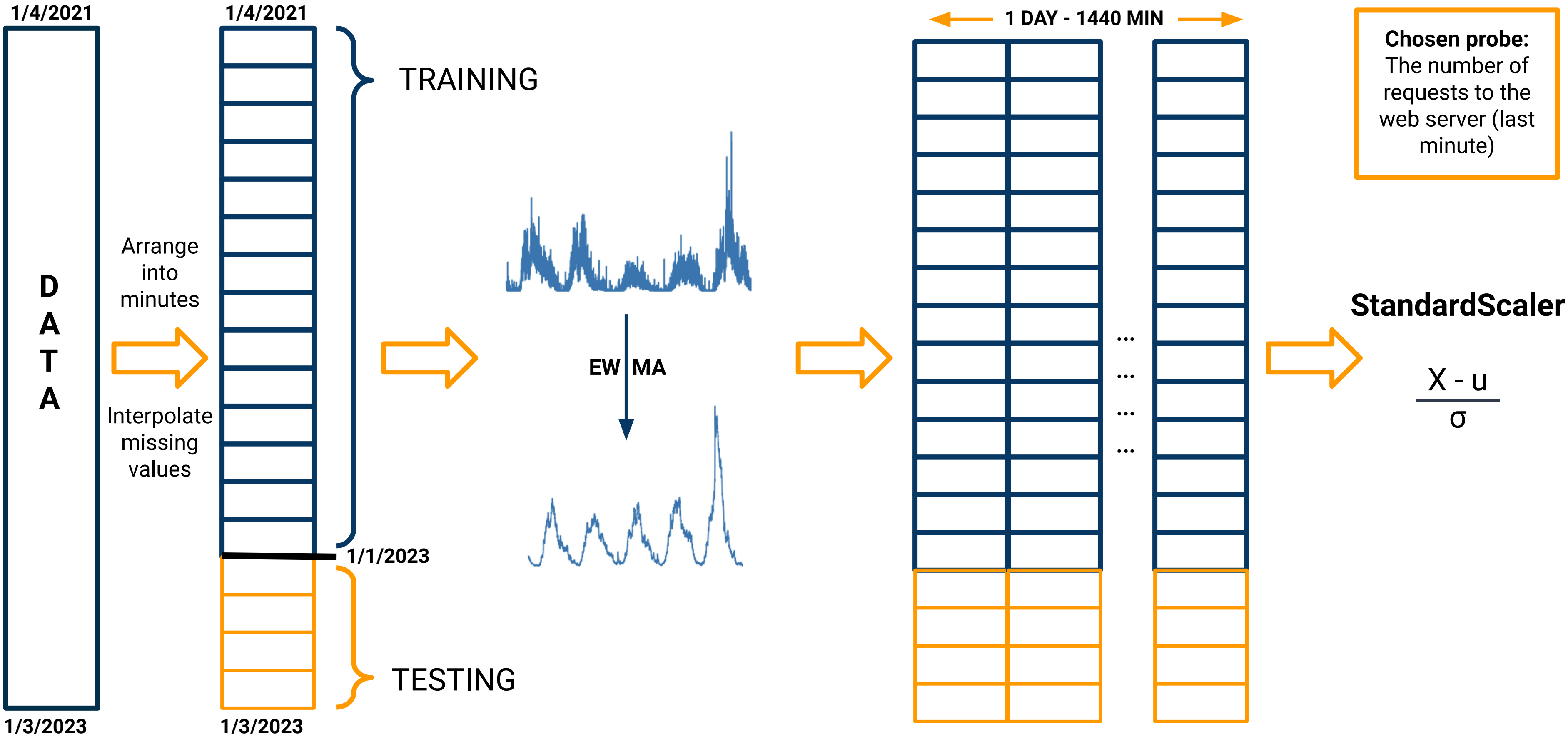
```
▼ object {12}
    host : moodle.cs.lth.se
  ▼ info {2}
      name : Database: #users (last 5 minutes)
      description : Number of users who has been active in the last
                    5 minutes (according to MariaDB)
    object : /app/mariadb/moodle/active-users/5min
    is_active : true
    is_public : true
  ▼ definition {2}
      type : custom
    ▼ parameters {2}
        unit : users
        integer : true
    is_numeric : true
  ▼ private_info {0}
      (empty object)
    store_history : true
    reduction_rule : LATEST
    is_inactive_until : null
    stale_after_sec_override : null
```

```
▼ object {12}
    host : moodle.cs.lth.se
  ▼ info {2}
      name : Database: #users (last 5 minutes)
      description : Number of users who has been active in the last
                    5 minutes (according to MariaDB)
    object : /app/mariadb/moodle/active-users/5min
    is_active : true
    is_public : true
  ▼ definition {2}
      type : custom
    ▼ parameters {2}
        unit : users
        integer : true
    is_numeric : true
  ▼ private_info {0}
      (empty object)
    store_history : true
    reduction_rule : LATEST
    is_inactive_until : null
    stale_after_sec_override : null
```



	timestamp	unix-epoch	value
0	2021-03-10 23:15:02.447623+00	1.615418e+09	1.0
1	2021-03-11 23:15:01.683307+00	1.615505e+09	0.0
2	2021-03-12 23:15:02.446321+00	1.615591e+09	0.0
3	2021-03-13 23:15:02.322169+00	1.615677e+09	0.0
4	2021-03-14 23:15:02.400434+00	1.615764e+09	0.0
...	...	...	...
691655	2023-03-31 14:15:01.783952+00	1.680272e+09	0.0
691656	2023-03-31 14:17:01.696382+00	1.680272e+09	0.0
691657	2023-03-31 14:19:01.710269+00	1.680272e+09	0.0
691658	2023-03-31 14:23:01.926453+00	1.680273e+09	0.0
691659	2023-03-31 14:24:01.876334+00	1.680273e+09	0.0

DATA PRE-PROCESSING



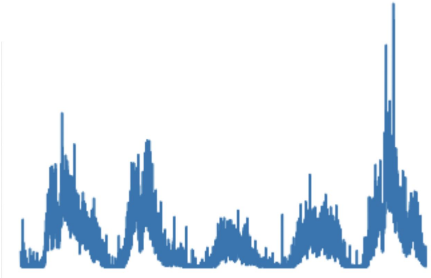
Chosen probe: The number of requests to the web server (last minute)

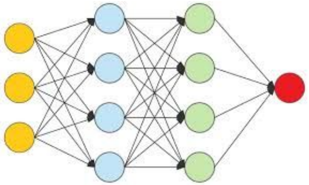
StandardScaler

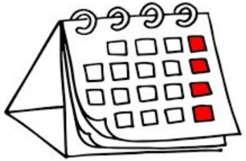
$$\frac{X - u}{\sigma}$$

DAY PREDICTION

Identify which day of the week it is by observing the number of requests to the moodle web server in the last minute.







Maybe something simple will work...

KNN

- Neighbours: 3, 20

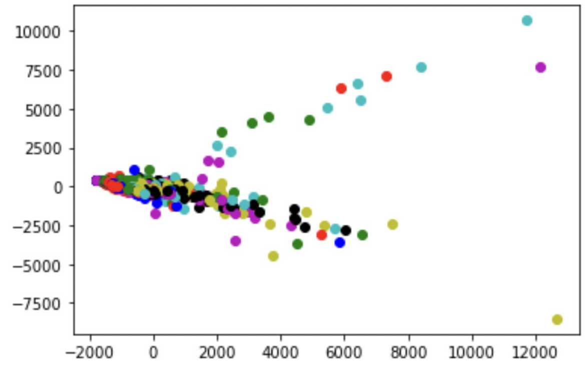
LOGISTIC REGRESSION

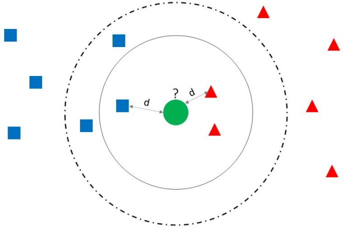
- max_iter=1000
- C = 0.01

Or Not...

Separable in 2D?

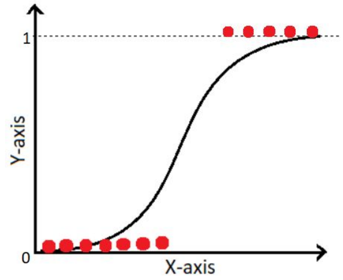
PCA





Training Score: 39%
Testing Score: 0.97%

Training Score: 54%
Testing Score: 10%



Neural Network

Python Library: Pytorch

Dataset: Custom dataset that provides items with corresponding labels

DataLoader: Provides elements from Dataset randomly and in batches of 32.

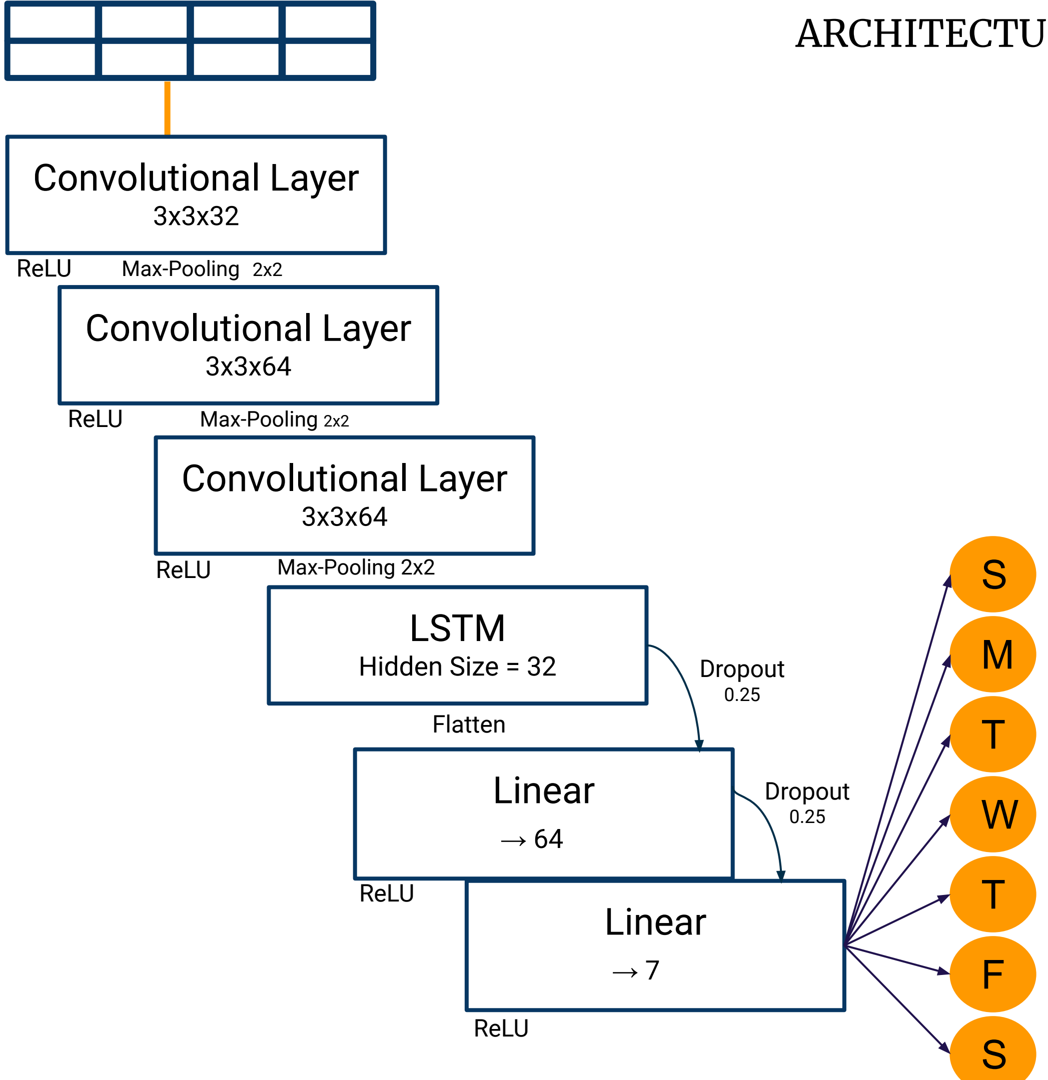
Criterion: Cross Entropy Loss

Optimization: ADAM

Epochs: 100

Learning Rate: 1e-4

ARCHITECTURE:



Convolutional Layer
3x3x32

ReLU Max-Pooling 2x2

Convolutional Layer
3x3x64

ReLU Max-Pooling 2x2

Convolutional Layer
3x3x64

ReLU Max-Pooling 2x2

LSTM
Hidden Size = 32

Dropout
0.25

Flatten

Linear
→ 64

ReLU

Dropout
0.25

Linear
→ 7

ReLU

S
M
T
W
T
F
S

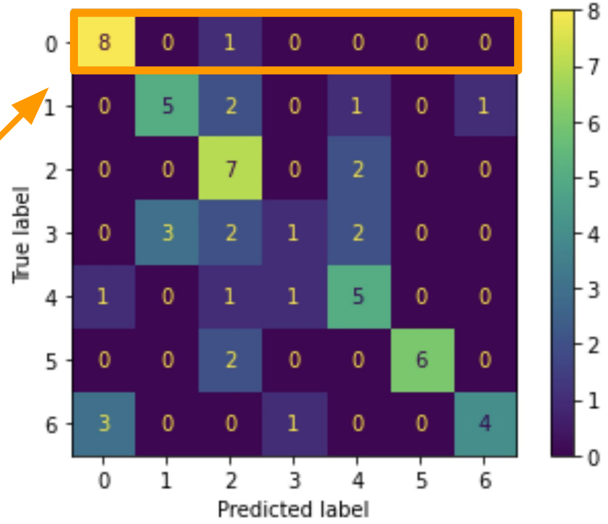
Day prediction

Class	Recall (%)	Precision (%)
0. Sunday	88.9	66.7
1. Monday	55.5	62.5
2. Tuesday	77.8	46.7
3. Wednesday	12.5	33.3
4. Thursday	62.5	50.0
5. Friday	75.0	100
6. Saturday	50.0	80.0

Accuracy: 61%

F1 : 0.59

Sundays are clearly distinguishable



	0	1	2	3	4	5	6
0	8	0	1	0	0	0	0
1	0	5	2	0	1	0	1
2	0	0	7	0	2	0	0
3	0	3	2	1	2	0	0
4	1	0	1	1	5	0	0
5	0	0	2	0	0	6	0
6	3	0	0	1	0	0	4

RESULTS

Day prediction

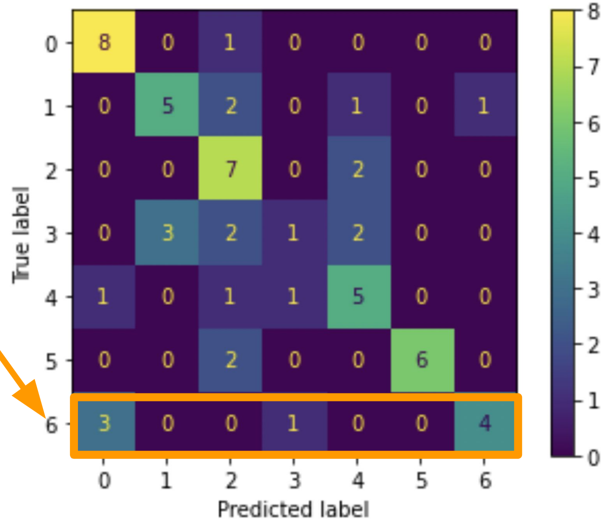
Class	Recall (%)	Precision (%)
0. Sunday	88.9	66.7
1. Monday	55.5	62.5
2. Tuesday	77.8	46.7
3. Wednesday	12.5	33.3
4. Thursday	62.5	50.0
5. Friday	75.0	100
6. Saturday	50.0	80.0

But Saturdays are very similar…

Sundays are clearly distinguishable

Accuracy: 61%

F1 : 0.59



True label \ Predicted label	0	1	2	3	4	5	6
0	8	0	1	0	0	0	0
1	0	5	2	0	1	0	1
2	0	0	7	0	2	0	0
3	0	3	2	1	2	0	0
4	1	0	1	1	5	0	0
5	0	0	2	0	0	6	0
6	3	0	0	1	0	0	4

RESULTS

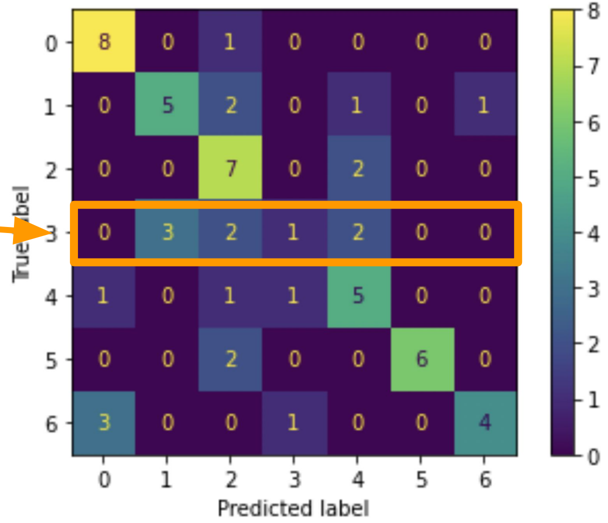
Day prediction

Class	Recall (%)	Precision (%)
0. Sunday	88.9	66.7
1. Monday	55.5	62.5
2. Tuesday	77.8	46.7
3. Wednesday	12.5	33.3
4. Thursday	62.5	50.0
5. Friday	75.0	100
6. Saturday	50.0	80.0

Week days are harder, especially in the middle

Sundays are clearly distinguishable

Accuracy: 61%
F1 : 0.59



	0	1	2	3	4	5	6
0	8	0	1	0	0	0	0
1	0	5	2	0	1	0	1
2	0	0	7	0	2	0	0
3	0	3	2	1	2	0	0
4	1	0	1	1	5	0	0
5	0	0	2	0	0	6	0
6	3	0	0	1	0	0	4

But saturdays are very similar…

RESULTS

Day prediction

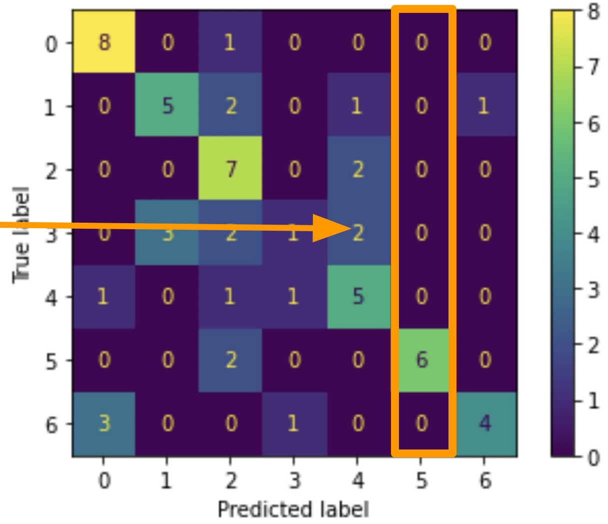
Class	Recall (%)	Precision (%)
0. Sunday	88.9	66.7
1. Monday	55.5	62.5
2. Tuesday	77.8	46.7
3. Wednesday	12.5	33.3
4. Thursday	62.5	50.0
5. Friday	75.0	100
6. Saturday	50.0	80.0

Accuracy: 61%

F1 : 0.59

Sundays are clearly distinguishable

Fridays have a very distinguished class

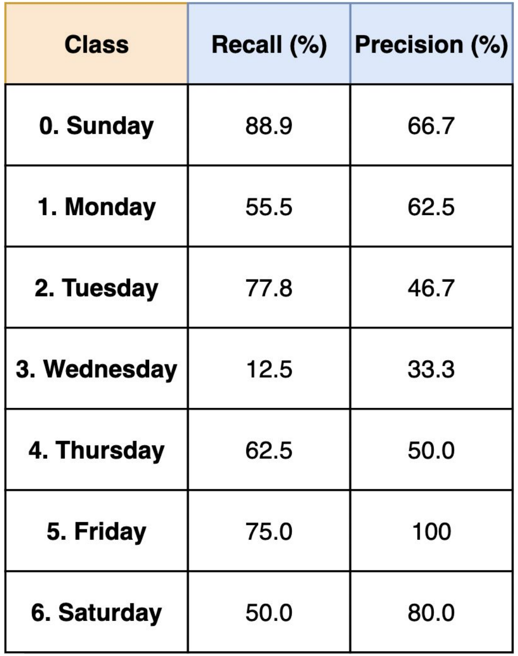


True label \ Predicted label	0	1	2	3	4	5	6
0	8	0	1	0	0	0	0
1	0	5	2	0	1	0	1
2	0	0	7	0	2	0	0
3	0	3	2	1	2	0	0
4	1	0	1	1	5	0	0
5	0	0	2	0	0	6	0
6	3	0	0	1	0	0	4

Week days are harder, especially in the middle

But saturdays are very similar...

RESULTS



Class	Recall (%)	Precision (%)
0. Sunday	88.9	66.7
1. Monday	55.5	62.5
2. Tuesday	77.8	46.7
3. Wednesday	12.5	33.3
4. Thursday	62.5	50.0
5. Friday	75.0	100
6. Saturday	50.0	80.0

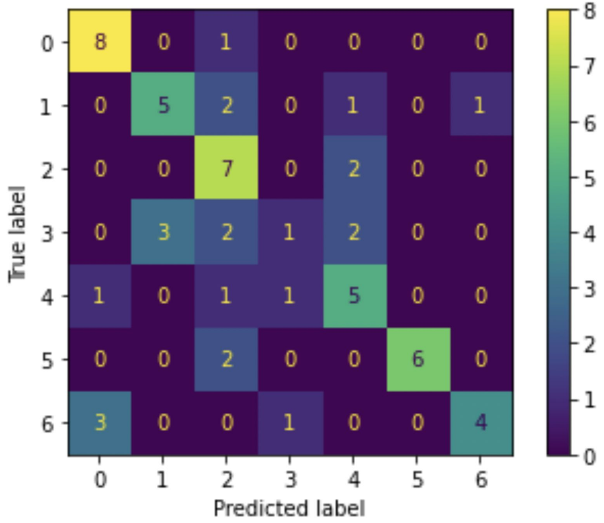
Fridays have a very distinguished class

Sundays are clearly distinguishable

Accuracy: 61%

F1 : 0.59

There clearly exists some kind of pattern…



True label \ Predicted label	0	1	2	3	4	5	6
0	8	0	1	0	0	0	0
1	0	5	2	0	1	0	1
2	0	0	7	0	2	0	0
3	0	3	2	1	2	0	0
4	1	0	1	1	5	0	0
5	0	0	2	0	0	6	0
6	3	0	0	1	0	0	4

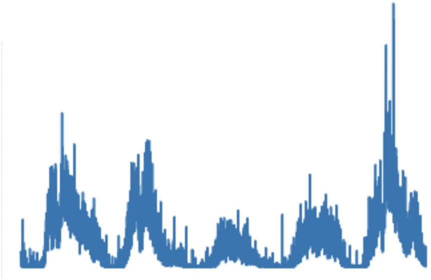
Week days are harder, especially in the middle

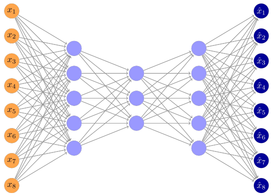
But saturdays are very similar…

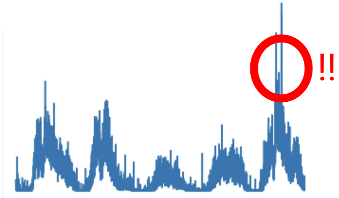
RESULTS

ANOMALY DETECTION

Identify possible anomalies in the number timeseries of requests to the moodle web server in the last minute.







AutoEncoder

Python Library: Pytorch

Dataset: Custom dataset that provides days

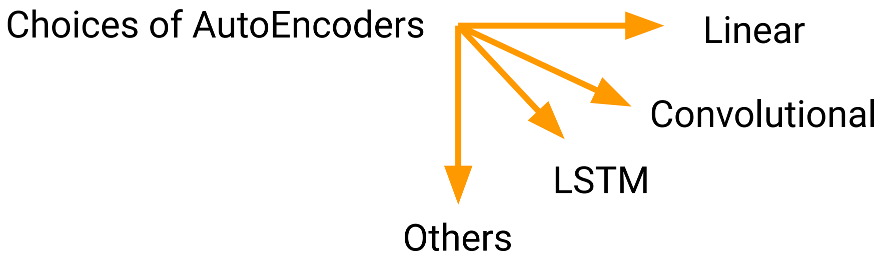
DataLoader: Provides elements from Dataset randomly and in batches of 32.

Criterion: MSE Loss

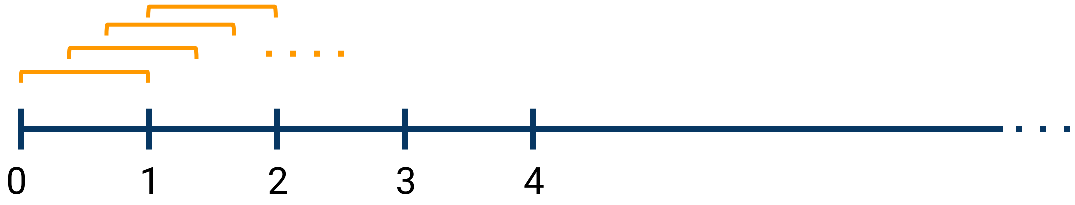
Optimization: ADAM

Epochs: 100

Learning Rate: 1e-4

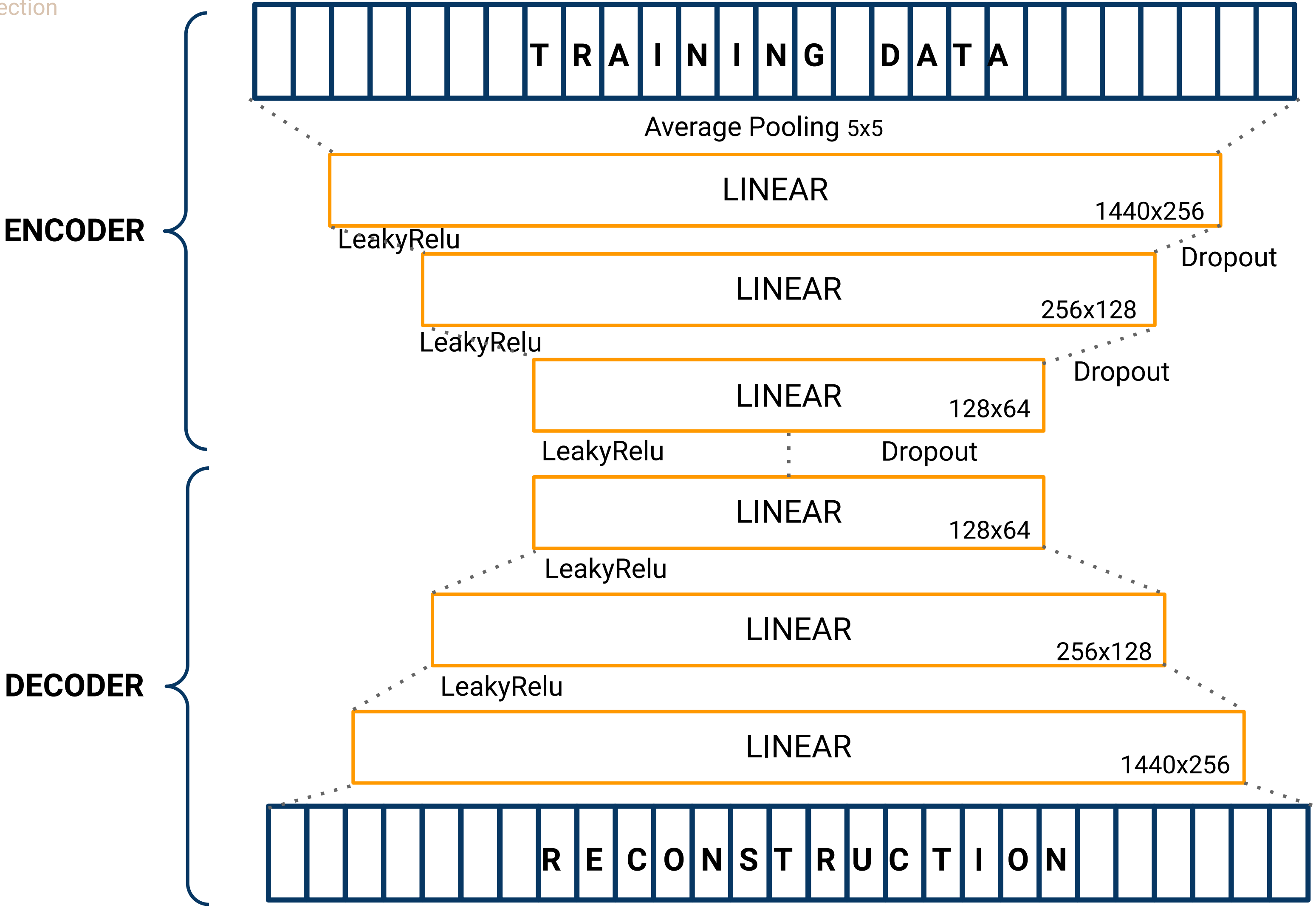


Some Further Processing:



Expanding the data set by overlapping data.

Let’s try Linear…



TRAINING DATA

Average Pooling 5x5

ENCODER

LINEAR 1440x256

LeakyRelu

Dropout

LINEAR 256x128

LeakyRelu

Dropout

LINEAR 128x64

LeakyRelu

Dropout

DECODER

LINEAR 128x64

LeakyRelu

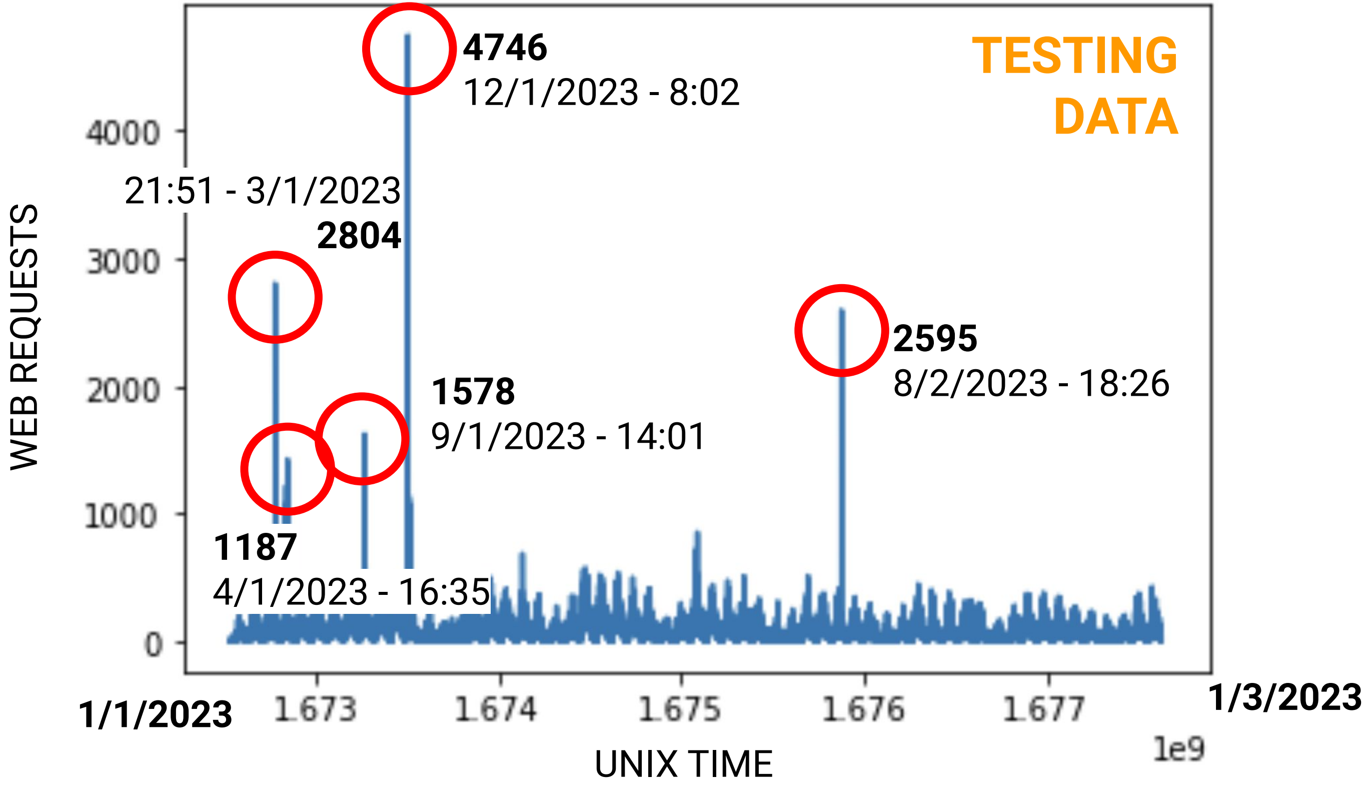
LINEAR 256x128

LeakyRelu

LINEAR 1440x256

RECONSTRUCTION

Identified Anomalies



TESTING DATA

4746
12/1/2023 - 8:02

21:51 - 3/1/2023
2804

1578
9/1/2023 - 14:01

2595
8/2/2023 - 18:26

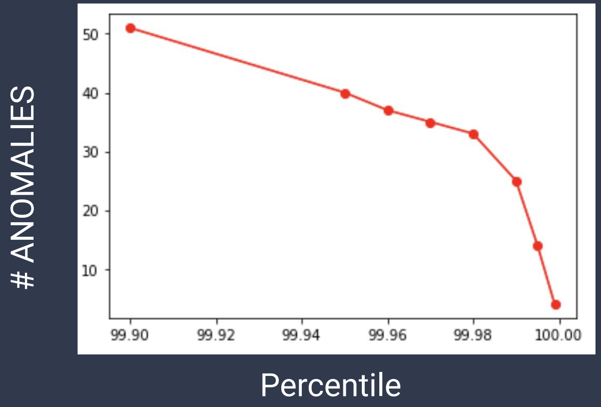
1187
4/1/2023 - 16:35

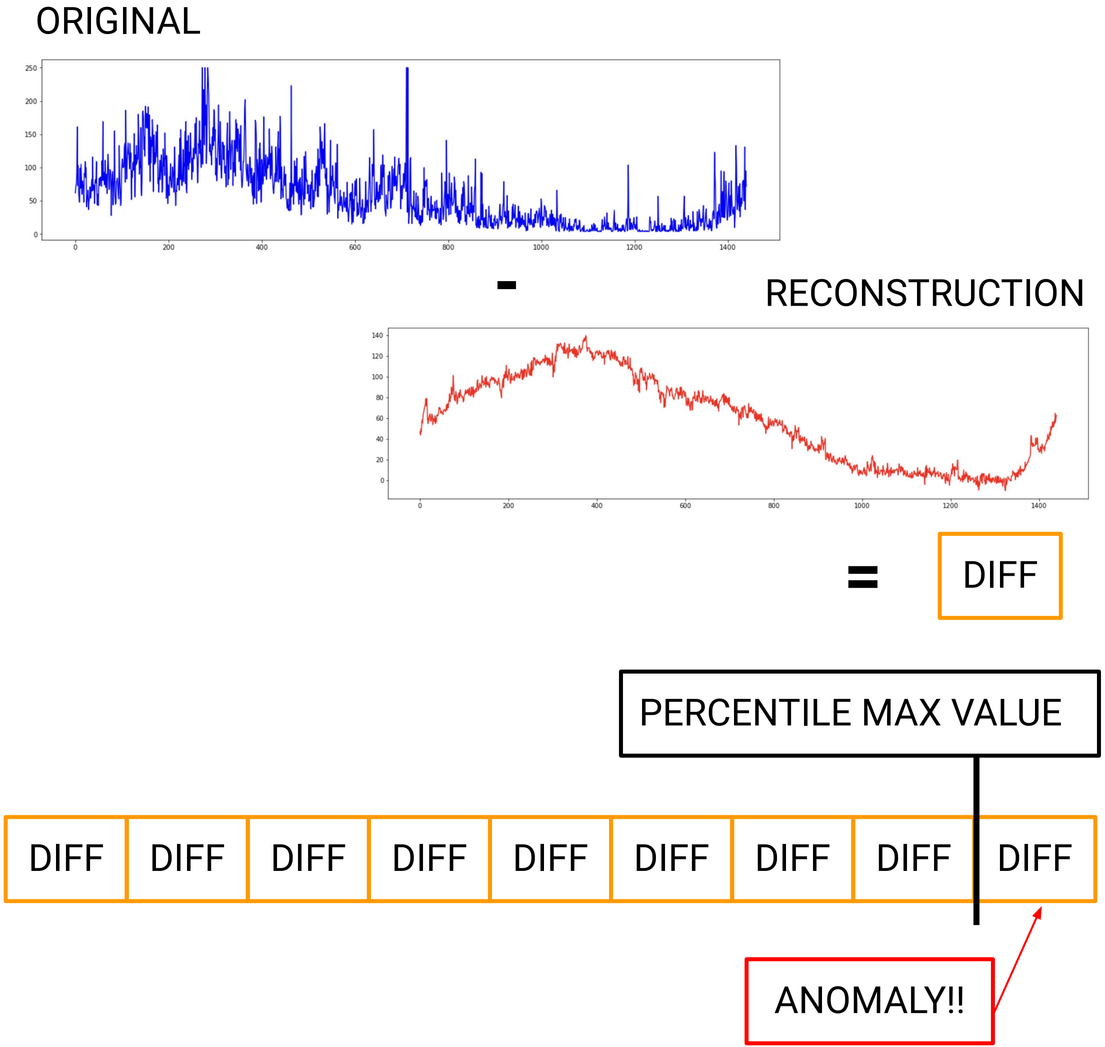
1/1/2023

1/3/2023

How Do We Know?

- Percentile becomes a parameter to adjust
- 99.99 percentile ~ 3σ,4σ = 1 in 370 - 1 in 15787





ORIGINAL

-

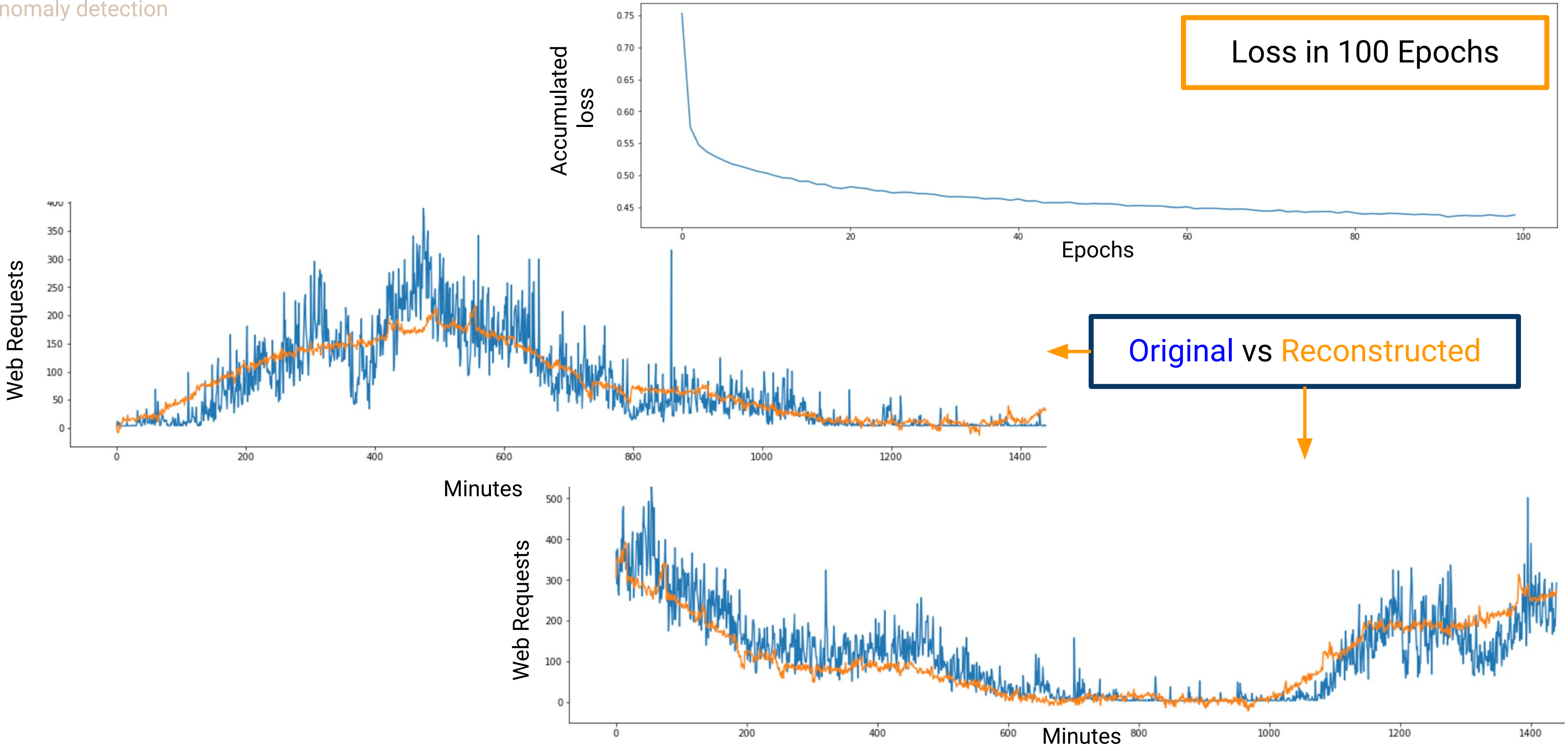
RECONSTRUCTION

= DIFF

PERCENTILE MAX VALUE

DIFF DIFF DIFF DIFF DIFF DIFF DIFF DIFF DIFF

ANOMALY!!



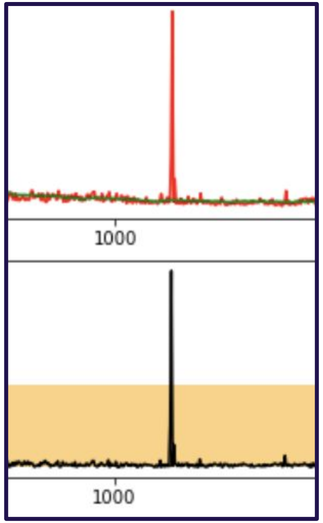
Loss in 100 Epochs

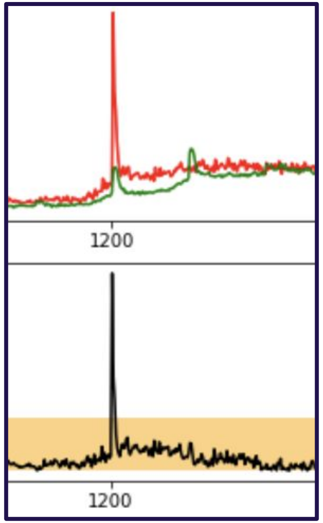
Original vs Reconstructed

RESULTS

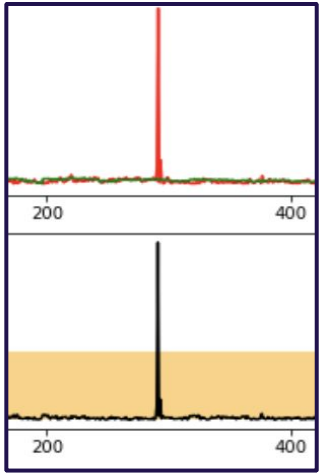
FOUND ANOMALIES

Bigger than the percentile, so Anomaly!!!



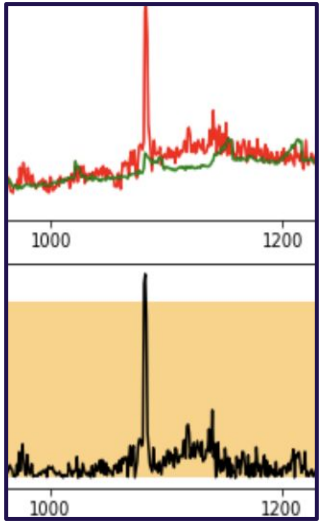


4 anomalies found



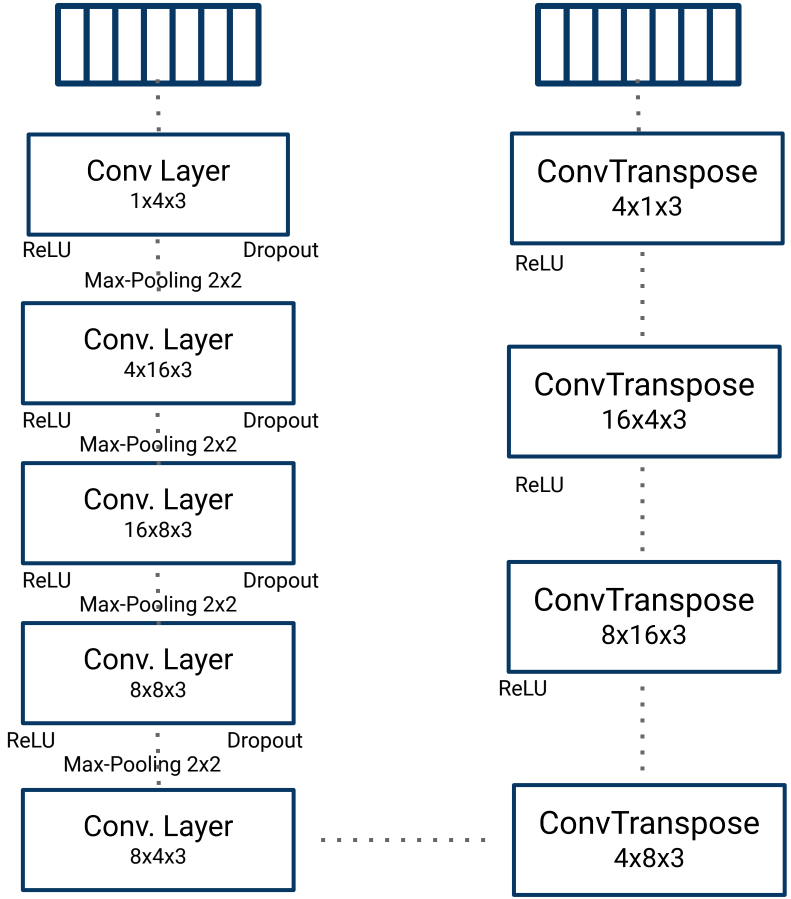
- Original
- Reconstruction
- Difference

Percentile zone



Exploring Other Options

CNN



Conv Layer
1x4x3
ReLU
Dropout
Max-Pooling 2x2

Conv. Layer
4x16x3
ReLU
Dropout
Max-Pooling 2x2

Conv. Layer
16x8x3
ReLU
Dropout
Max-Pooling 2x2

Conv. Layer
8x8x3
ReLU
Dropout
Max-Pooling 2x2

Conv. Layer
8x4x3

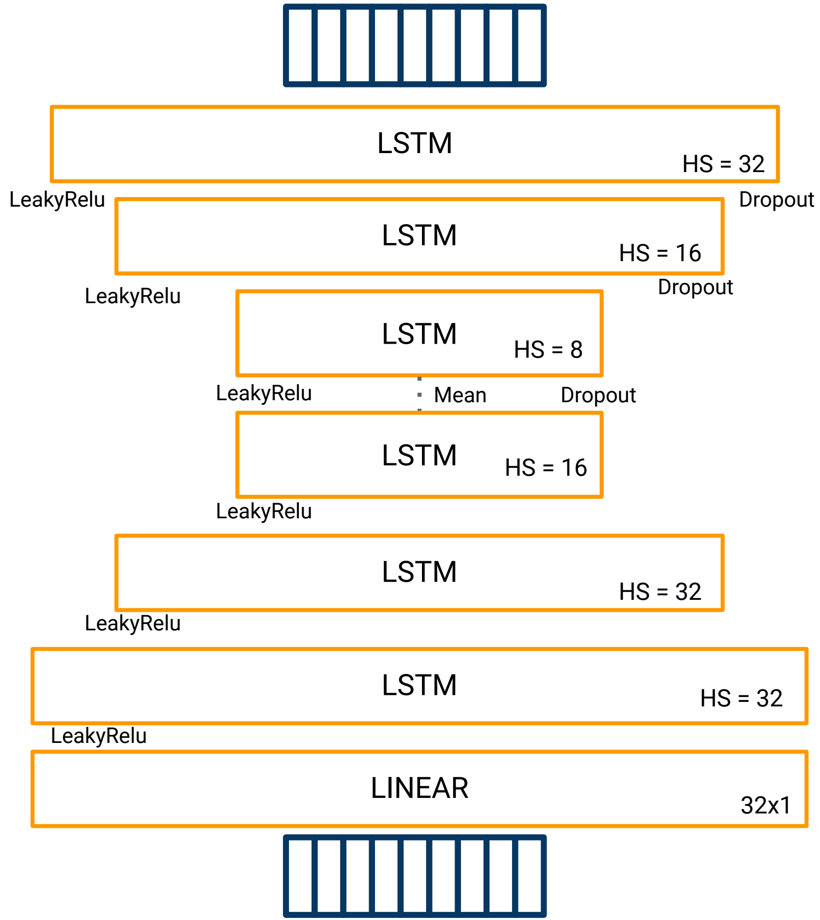
ConvTranspose
4x1x3
ReLU

ConvTranspose
16x4x3
ReLU

ConvTranspose
8x16x3
ReLU

ConvTranspose
4x8x3

LSTM



LSTM
HS = 32
LeakyRelu
Dropout

LSTM
HS = 16
LeakyRelu
Dropout

LSTM
HS = 8
LeakyRelu
Mean
Dropout

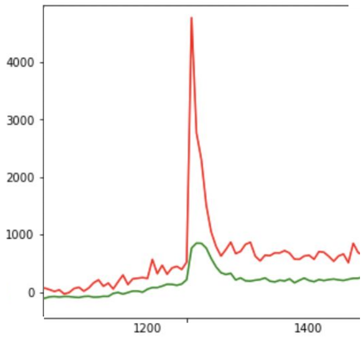
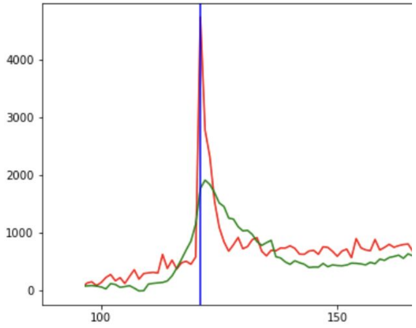
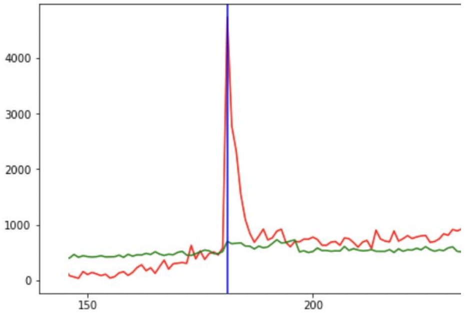
LSTM
HS = 16
LeakyRelu

LSTM
HS = 32
LeakyRelu

LSTM
HS = 32
LeakyRelu

LINEAR
32x1

Exploring Other Options

	LINEAR	CNN	LSTM
Percentile - Threshold	99.97% - 1292.0	99.97% - 986.8	99.97% - 1294.4
Average Training Error per Minute	19.7	17.9	18.6
Average Testing Error per Minute	27.5	24.2	25.6
Average Testing MSE per Window	35.47e5	47.57e4	54.37e4
Training Time	~10 min	~20 min	~9 hrs (!!)
Anomalies Found	4	3	4
An example of an anomaly			

OTHER DATASETS: OmniAnomaly

train test_label test

This dataset contains 38 features.

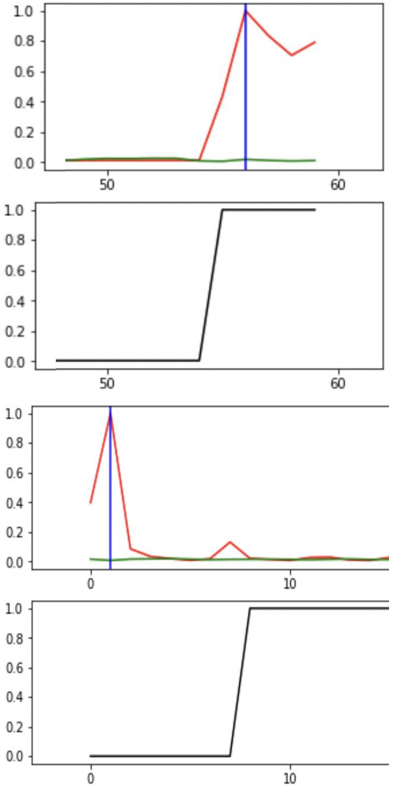
For each time series:

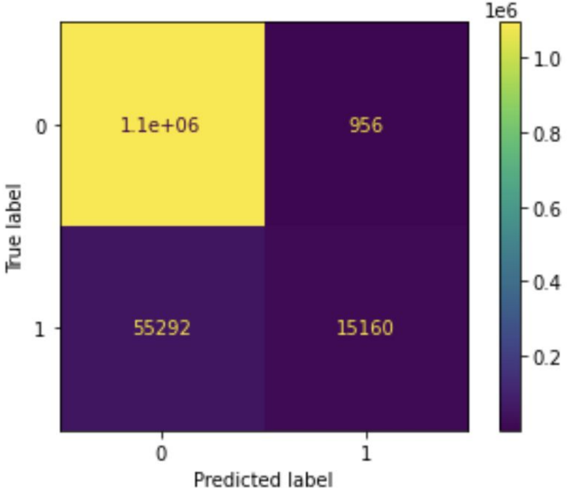
Grouping ⟶60

Overlapping ⟶ 10

1 data point is of size: 38x60

LINEAR AUTOENCODER





We can only predict “spike” anomalies ☹

Conclusions

- There exist a pattern for predicting days
- Anomalies were found with relatively simple method
- Difficulty of overfitting
- Difficulty of finding right hyperparameters

Future Work

- Tuning Hyperparameters to achieve better results
- Attempt prediction with encoded data
- Encode more than one moodle feature