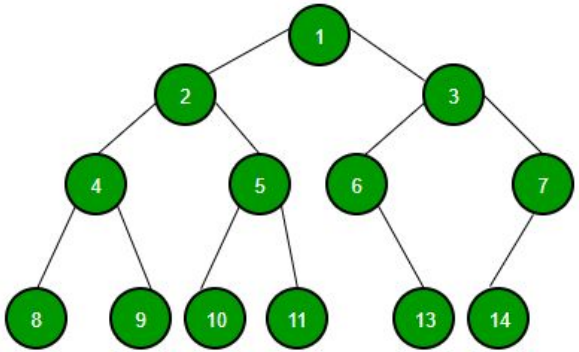
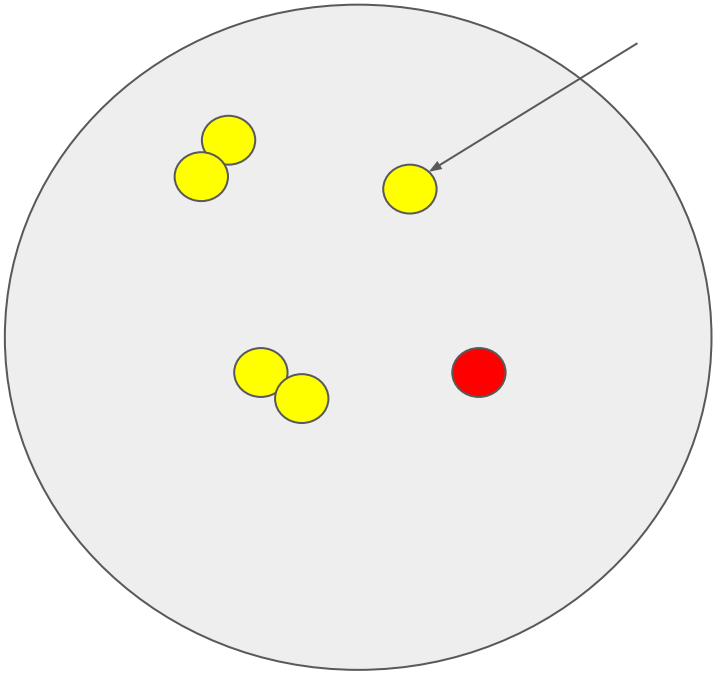
Mouse Embryos
And some very noisy data

Emil Schou Martiny

Data description

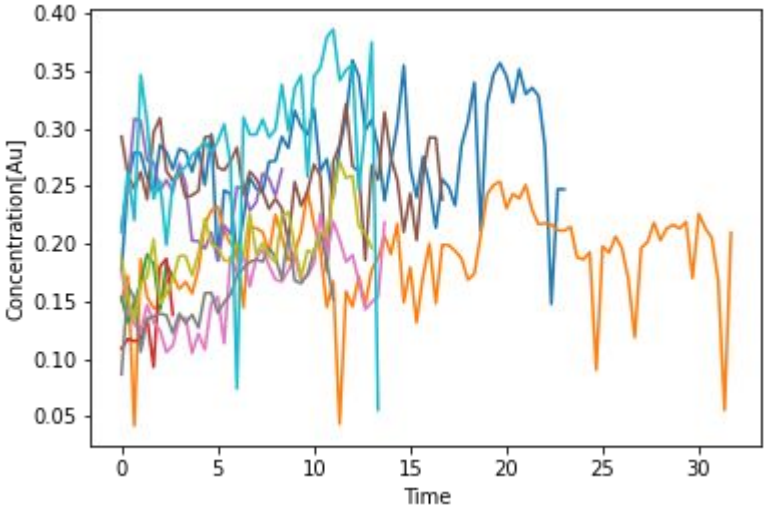
Data structure is a tree



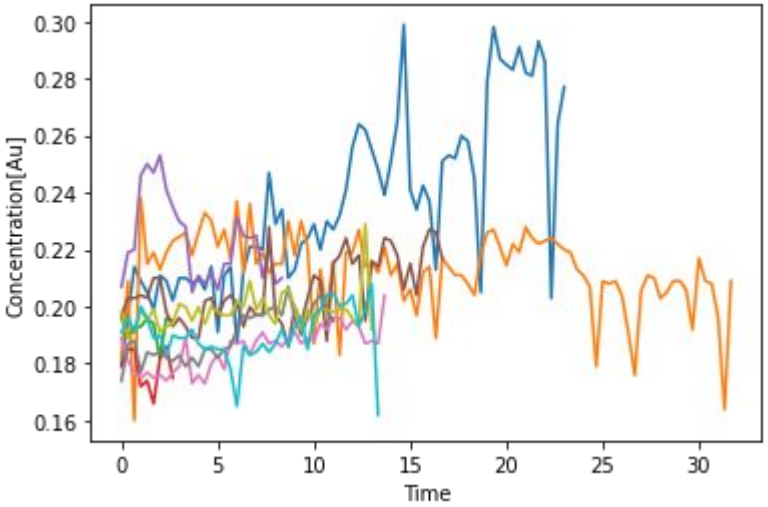


Spliced Fluorescent protein unto two genes

Two different proteins



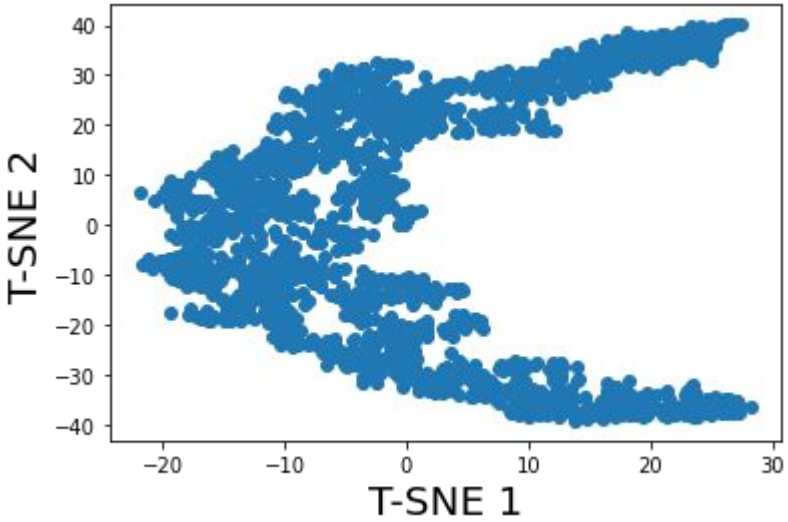
HEX



H2B

Super noisy datasets
Varying length
1500 series/cells

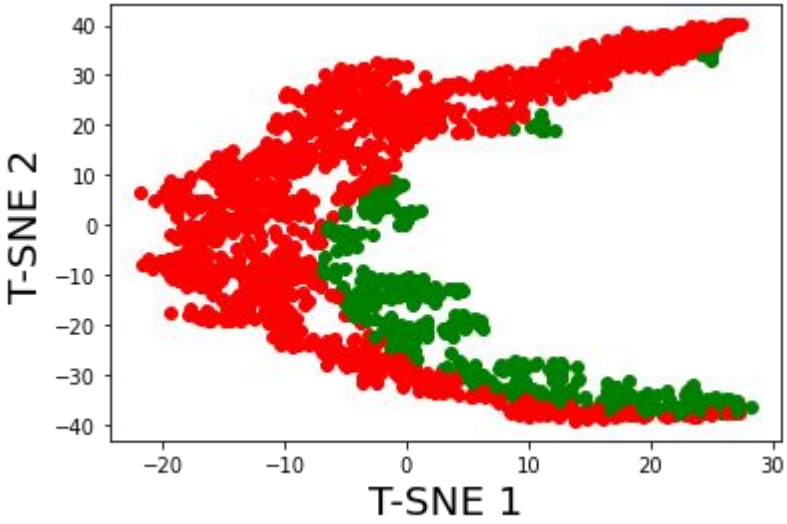
T-SNE



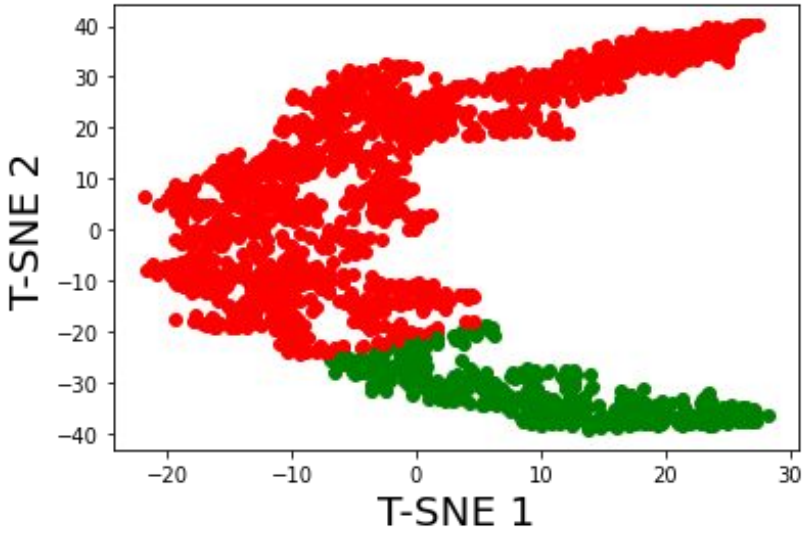
T-sne was were the project started
This structure is roughly theorized to be:
If one protein is high, the other is not.

T-SNE

Top 25% of protein 1



Top 25% of protein 2

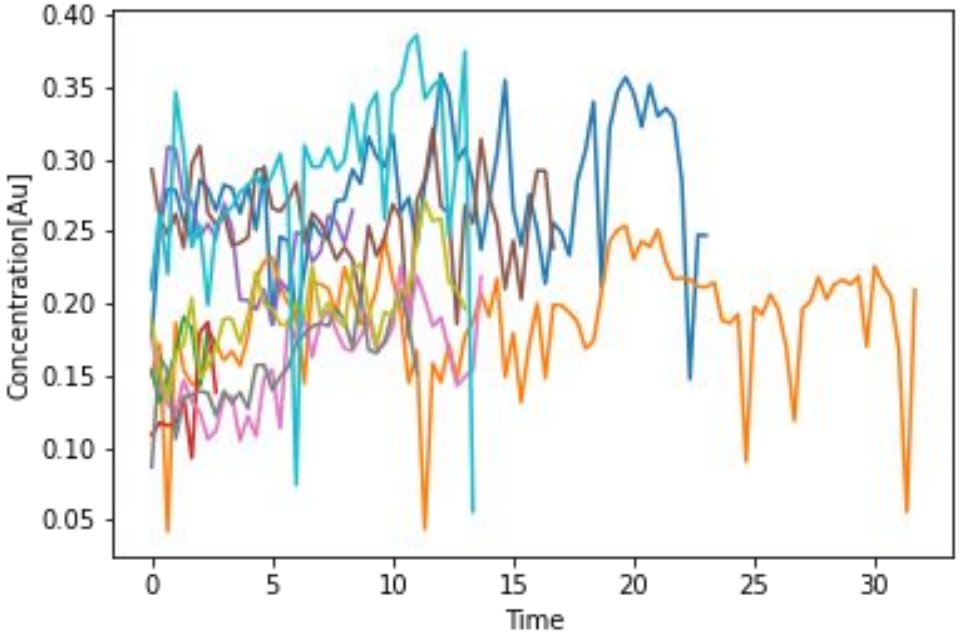


The quest for a meaningful label

In vivo vs. in vitro gives problems in defining differentiating

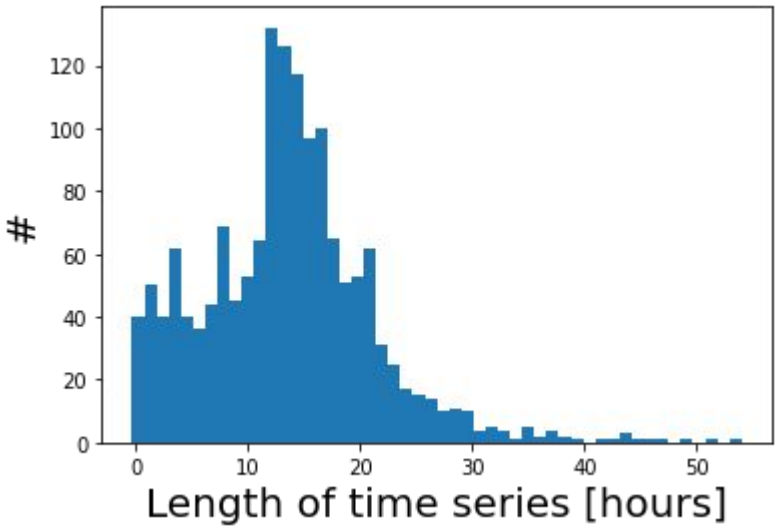
Possible labels

- Average
- End point
- Frequency
- Goes up in the end
- Standard deviation of data series

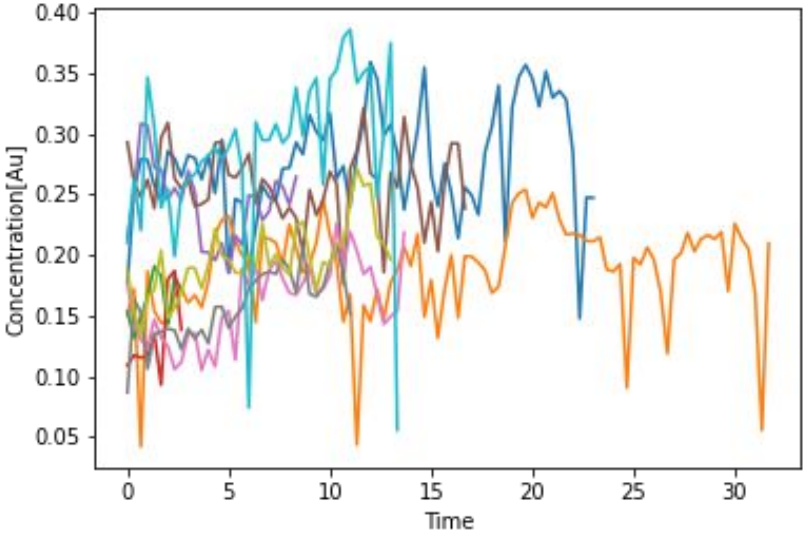


Data preparation

Padding



When to Normalize?
Before or after padding?



Initial models

LSTM with 1-3 layers

GRU

LSTM? (CNN, GRU, bidirectional etc)

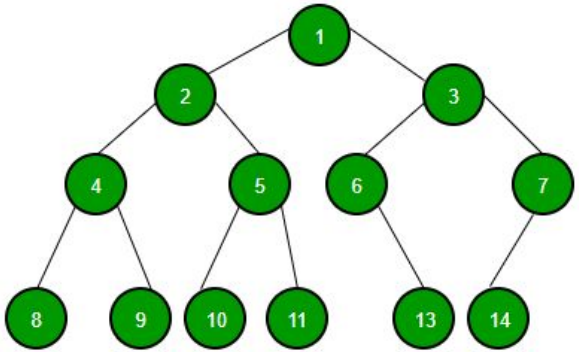
LGboost

- Average
- End point
- Frequency
- Goes up in the end
- Standard deviation of data series

Best result! Random Guessing.
There is no signal (I can Find)

Daughter cells

Can we predict them?



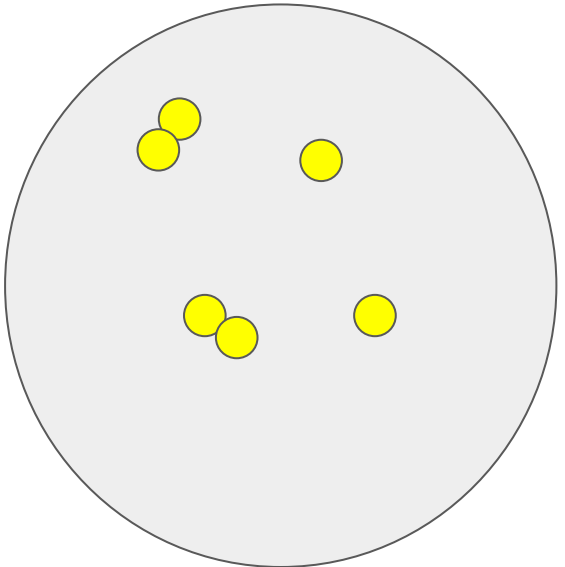
- Average
- End point
- Frequency
- Goes up in the end
- Standard deviation of data series

Nope

Reflections on no results

Machine learning has some limits, it is not magic

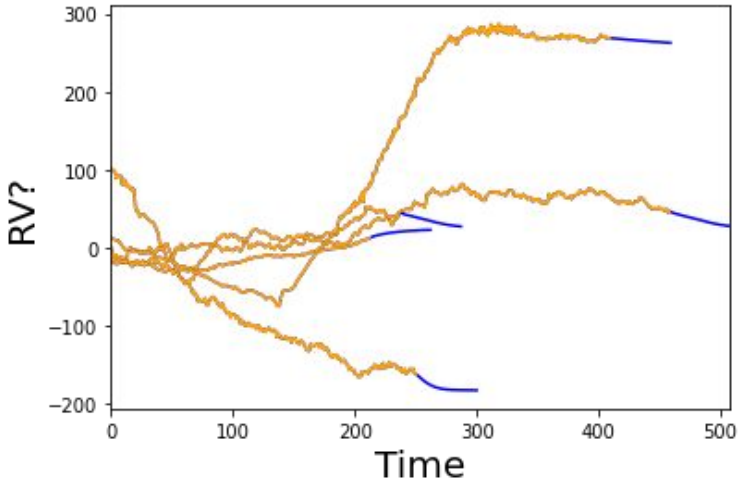
Maybe there is no signal since it is an in vitro experiment



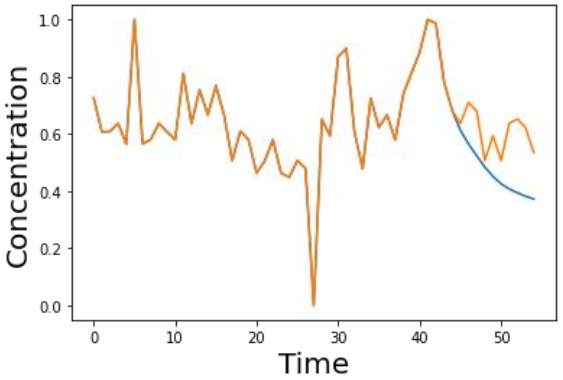
Time series predictions

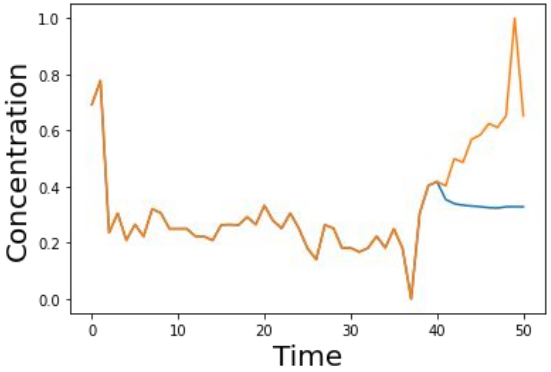
Two layer LSTM

Some astro data



My crappy protein data



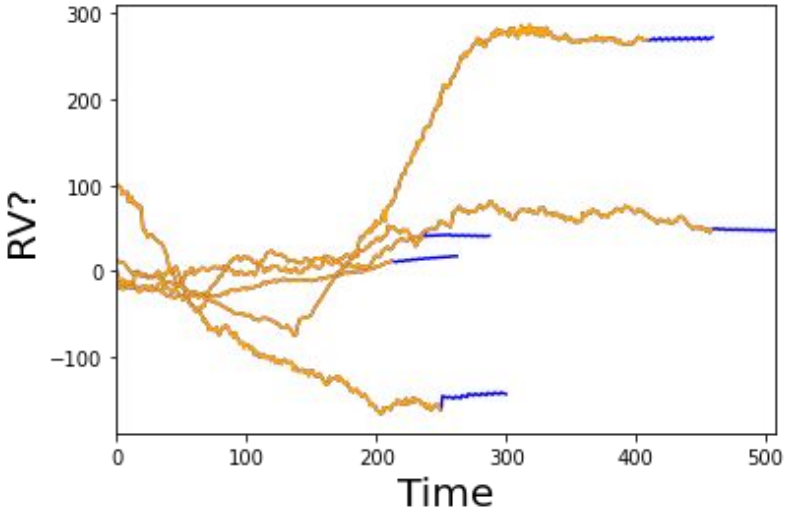


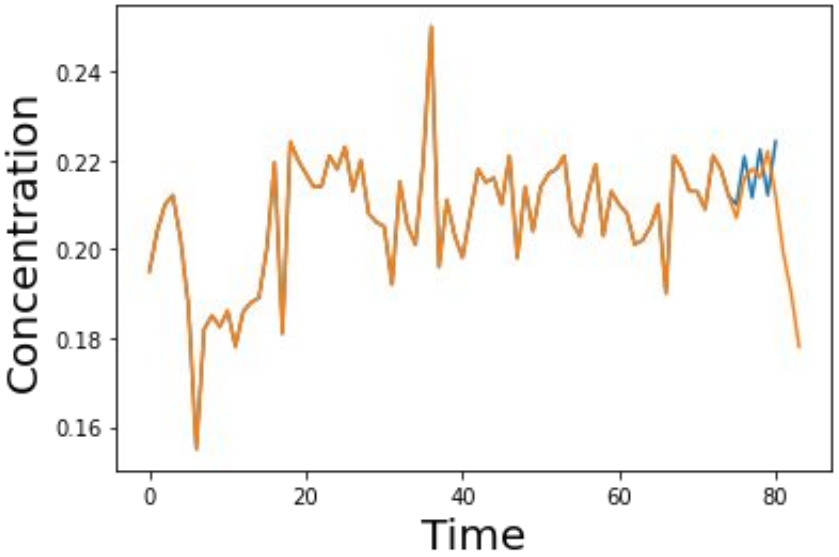
Problems

Noisy

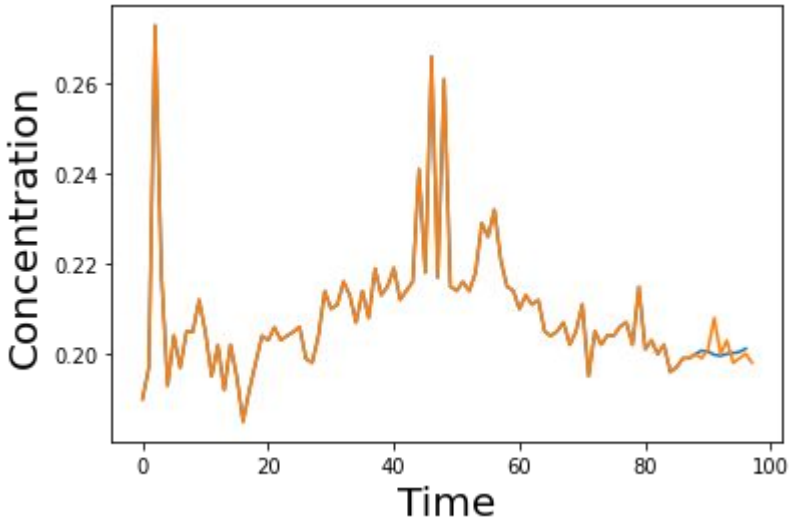
oscillating around zero, so fewer trends to catch up on

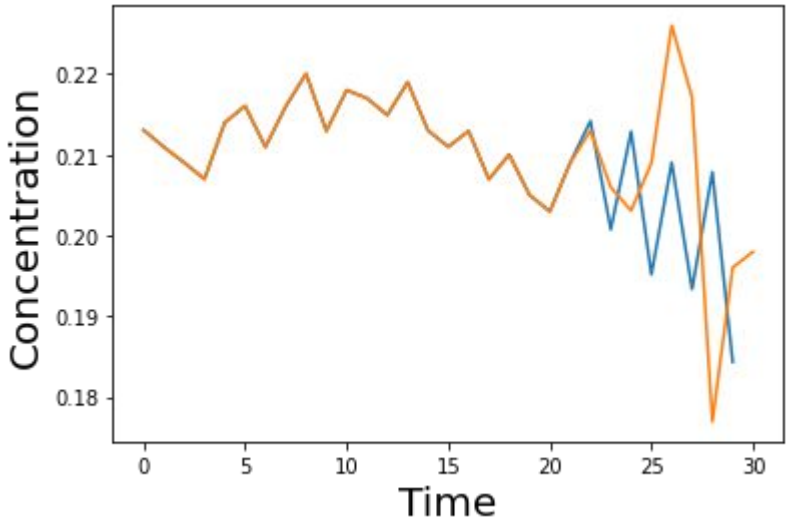
Some improvements





More examples





Questions

Appendix

Packages: LightBGM, Keras Tensorflow, SKlearn

I looped over different models with my different inputs and labels

Multi-layer LSTM with or without dropouts, GRU

Histogram of labels

Green and red bars, shows cutoff