Machine Learning on the OMXC25 Stocks

Bastian Bakkensen
Carlos Fernando Duarte Faurby
Daniel Hans Munk
Rasmus Nielsen

Date: 15-06-2022

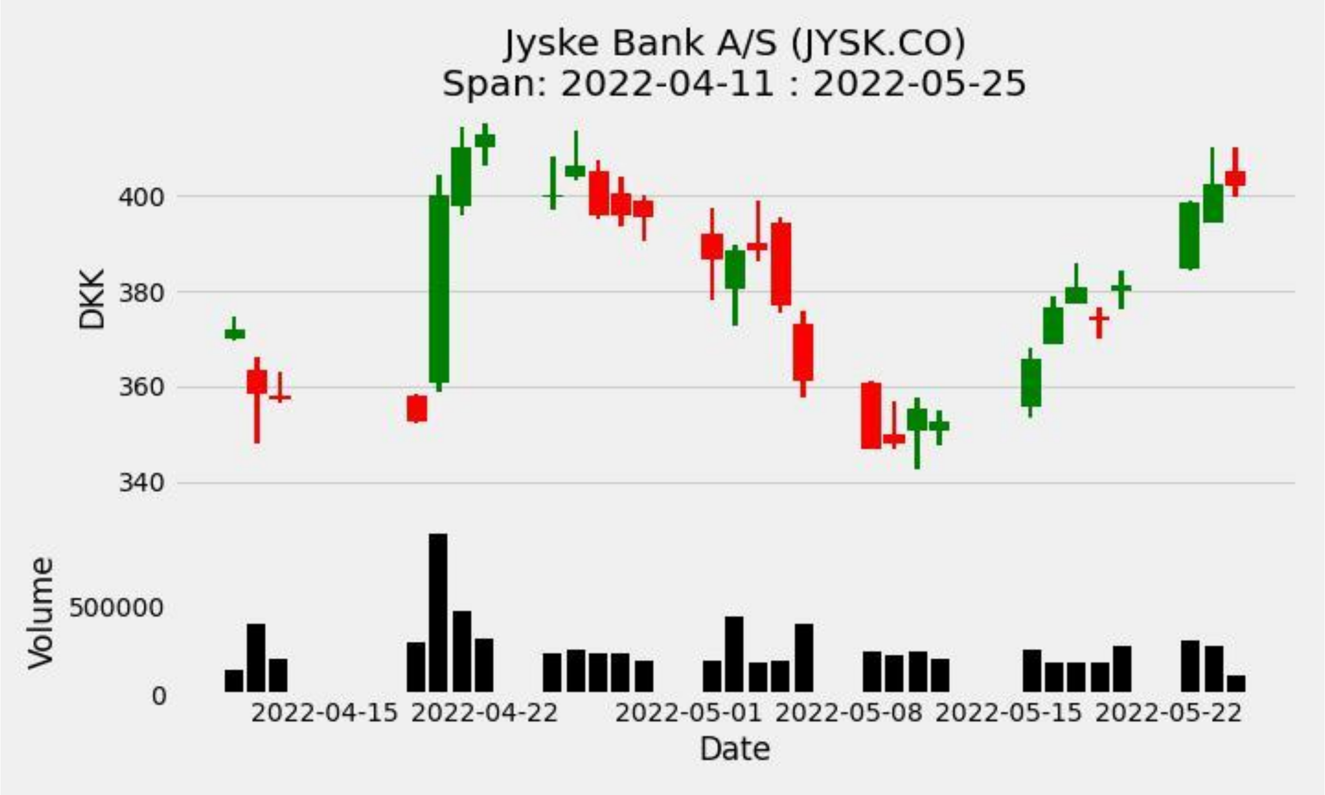
UNIVERSITY OF COPENHAGEN

Objective:

With ~20 years of stock data from the OMX Copenhagen 25 index, what is the predictive ability of modern Machine Learning?

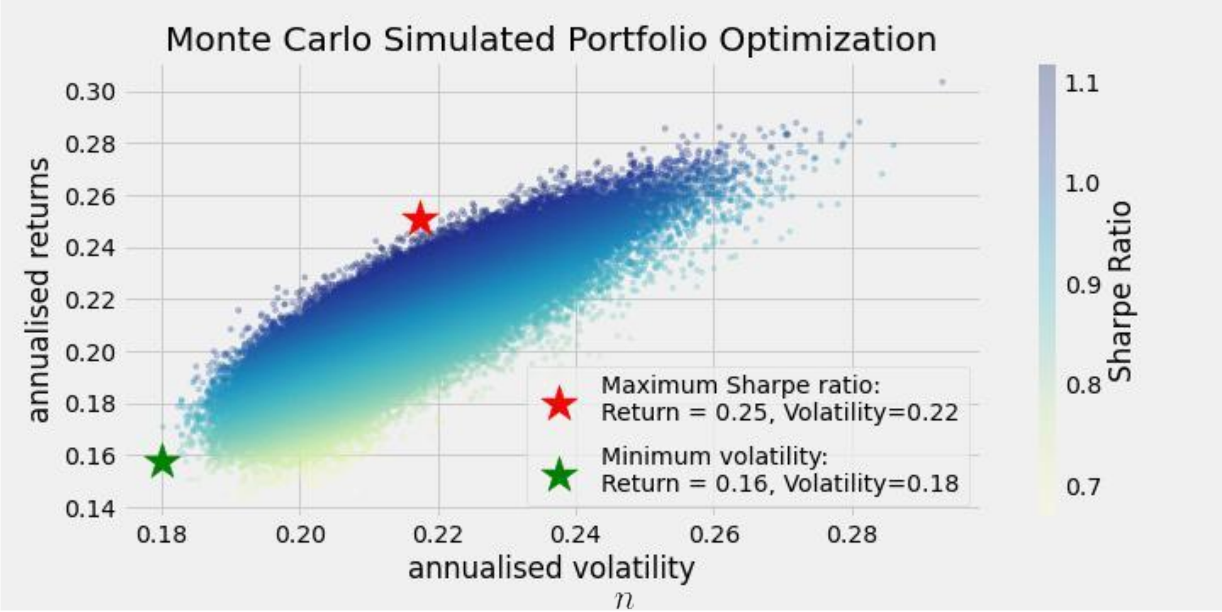
Data - the ‘ohlcv’ format

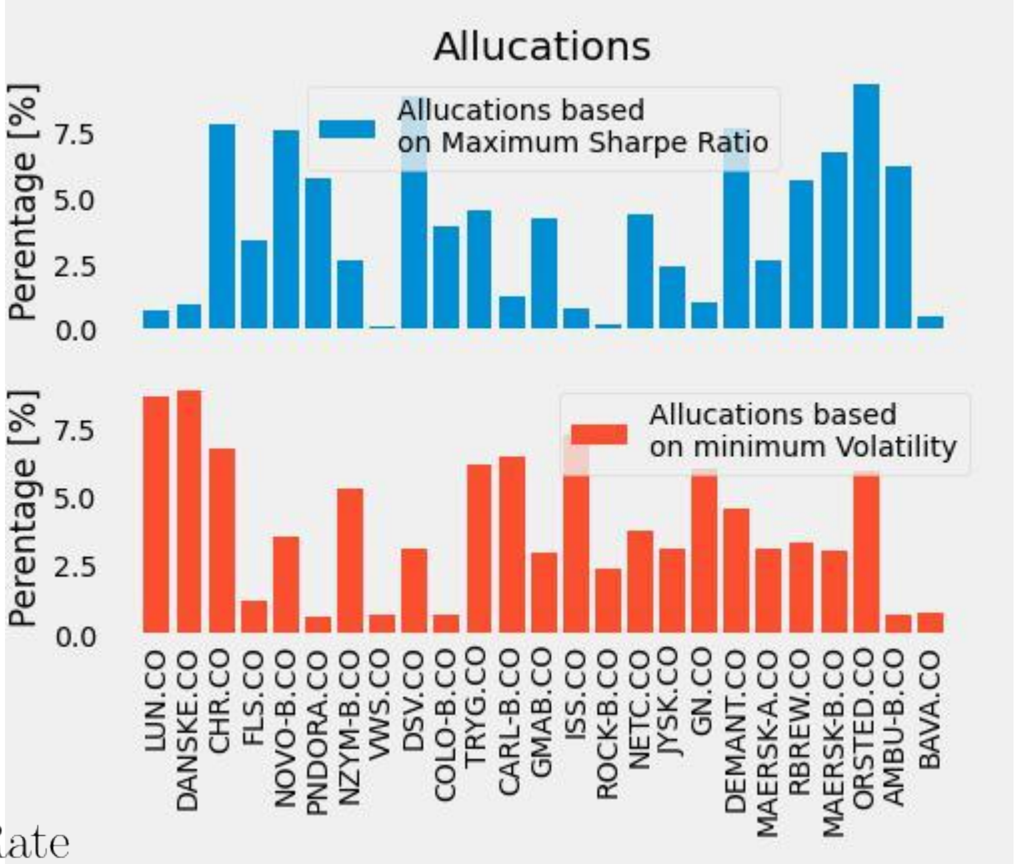
Ørsted
DSV
Carlsberg
Novo Nordisk
Chr. Hansen Holding
Tryg
Novozymes
FLSmidth & Co.
Genmab
Rockwoll
Royal Unibrew
Coloplast
Vestas Wind Systems
Pandora
Ambu
A.P. Møller-Mærsk (A)
A.P. Møller-Mærsk (B)
Demant
ISS
Danske Bank
GN Store Nordic
Netcompany Group
Bavarian Nordic
Jyske Bank
Lundbeck



In hindsight it is easy!

Modern Portfolio Theory





$$Mean(\text{Portfolio Returns}) = \sum_{i=1}^{n} w_i r_i$$

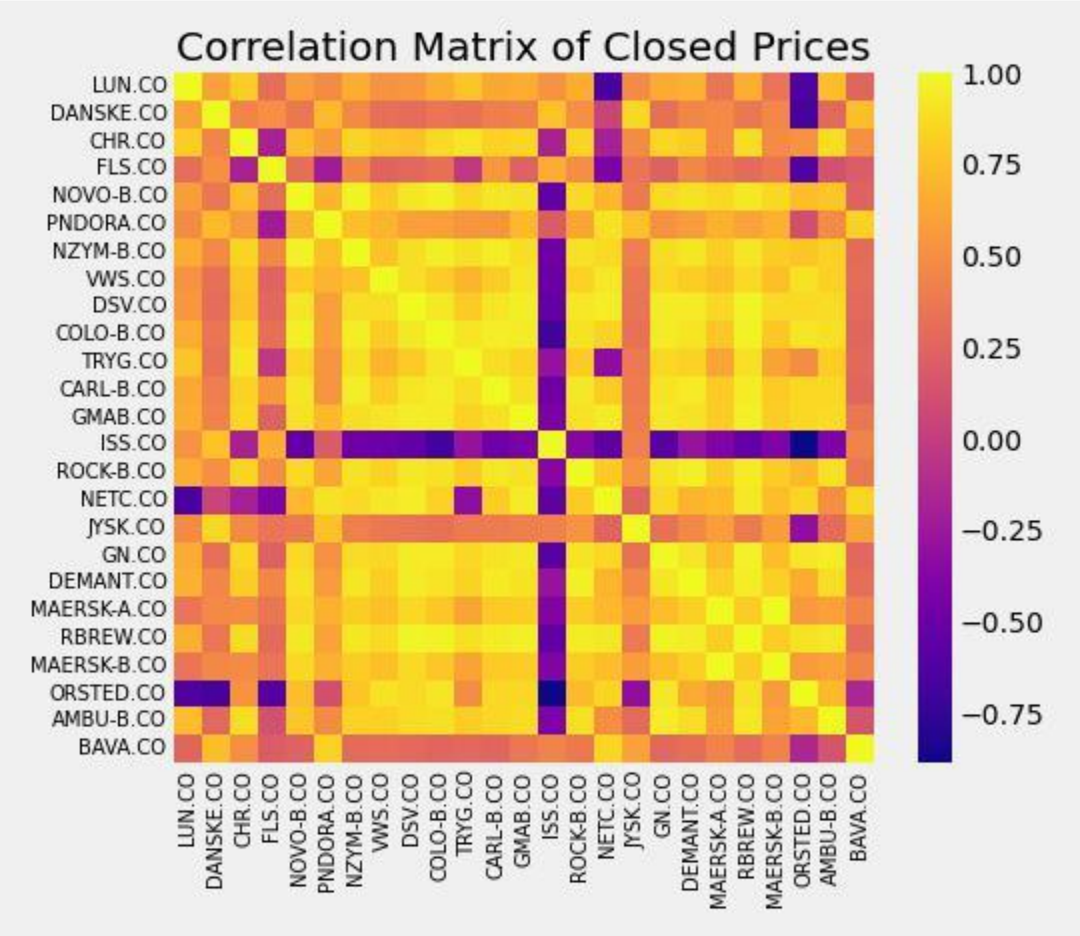
$$Var(\text{Portfolio Returns}) = \mathbf{w}^T \cdot (\mathbf{cov} \cdot \mathbf{w})$$

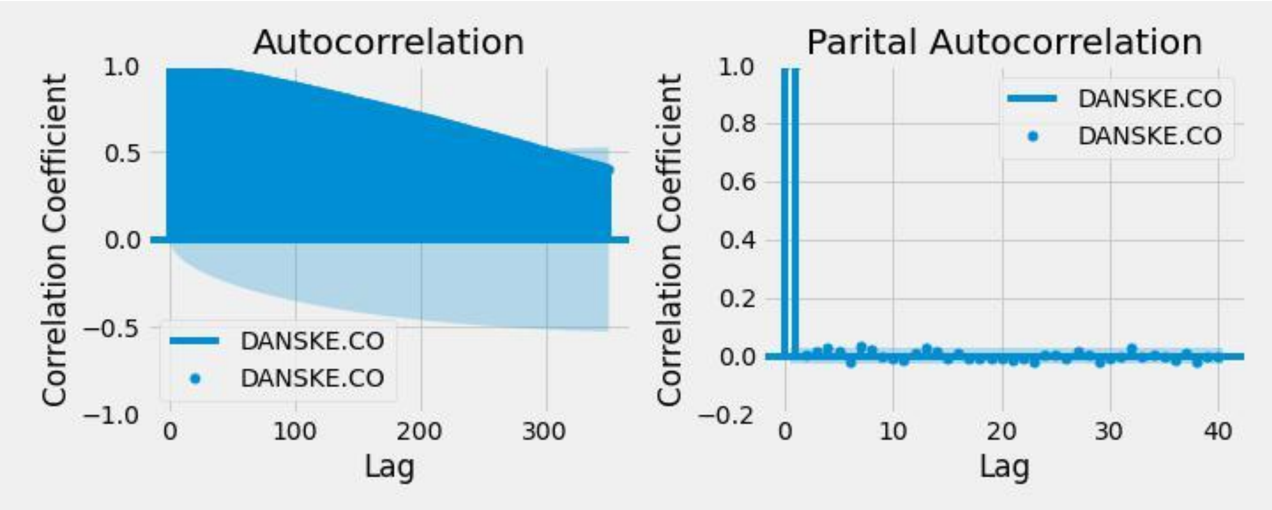
$$\text{Sharpe Ratio} = \frac{\text{Annual } Mean(\text{Portfolio Returns}) - \text{Risk Free Rate}}{\text{Annual } STD(\text{Portfolio Returns})}$$

$$\text{Minimum volatility} = argmin\{\text{Annual } STD(\text{Portfolio Returns})\}$$

$$\text{Maximum Sharpe Ratio} = argmax\{\text{Sharpe Ratio}\}$$

Correlations and Auto-Correlations

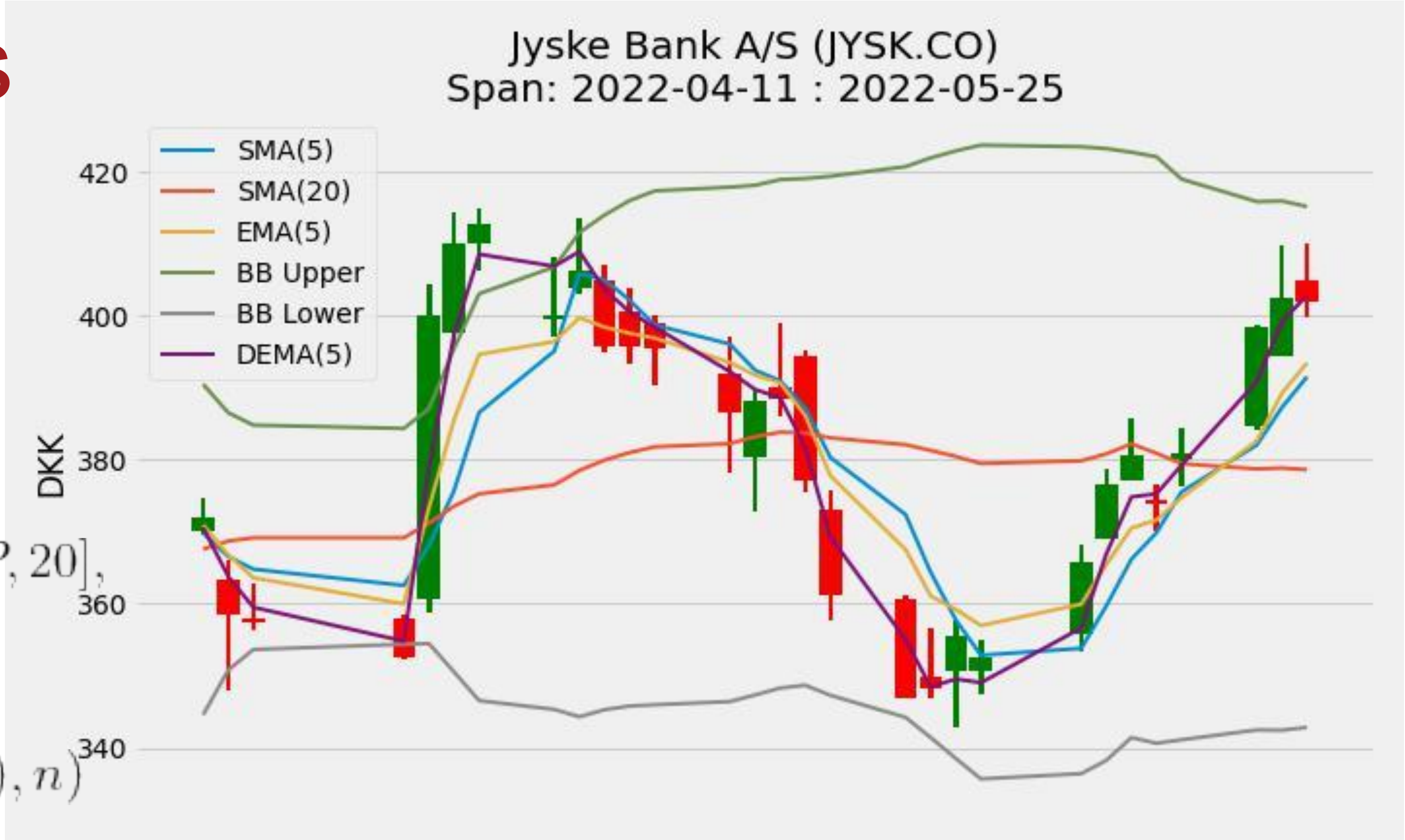




Autocorrelation:

No predictive ability after day 300

Technical Indicators



$$\mathrm{SMA}(n) = \frac{A_1 + A_2 + \cdots + A_n}{n}$$

$$\text{Bollinger Bands} = SMA(TP, 20) \pm 2\sigma[TP, 20],$$
$$TP = \frac{High + Low + Close}{3}$$

$$\mathrm{DEMA}(n) = 2\mathrm{EMA}(n) - \mathrm{EMA}(\mathrm{EMA}(n), n)$$

$$\mathrm{EMA}(n) : \mathrm{EMA}_{today} = \mathrm{Close}_{today} \times \left(\frac{2}{1+n}\right) + \mathrm{EMA}_{yesterday} \times \left(1 - \frac{2}{1+n}\right)$$

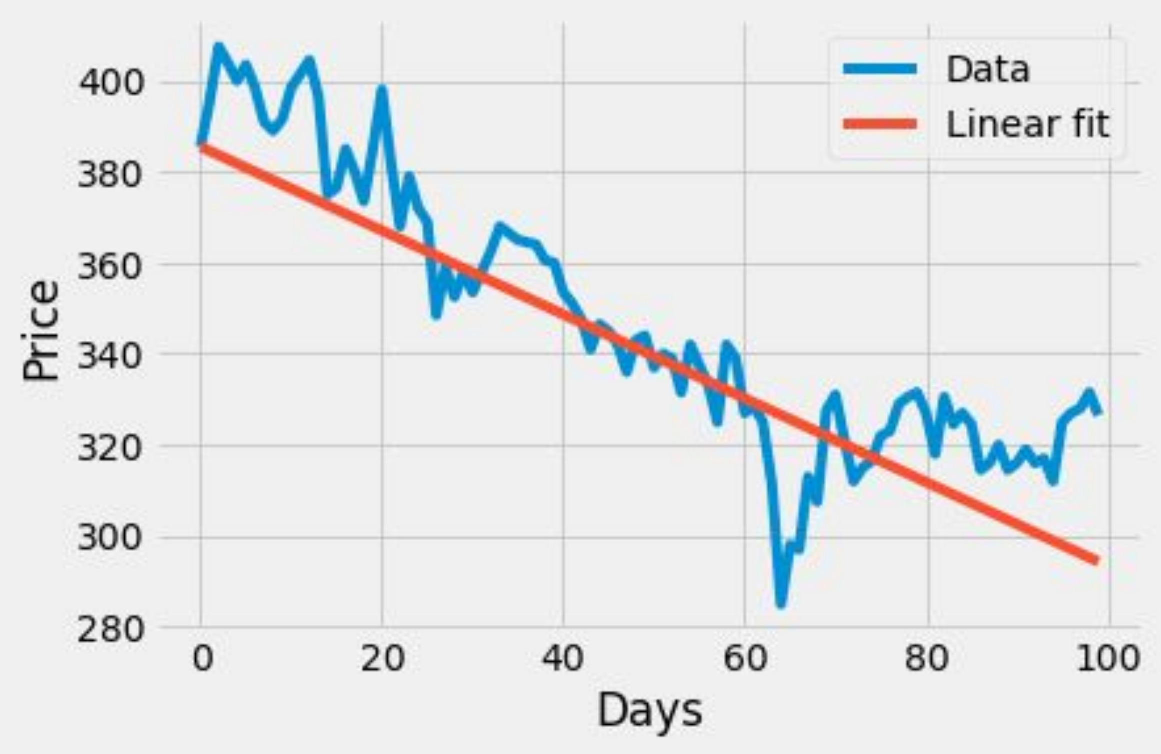
Full List of Technical Indicators

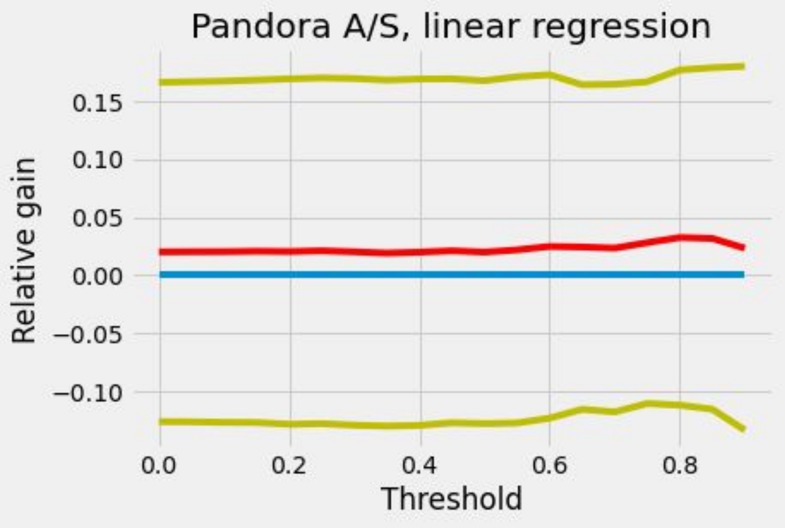
Acceleration Bands (ABANDS)
Accumulation/Distribution (AD)
Average Directional Movement (ADX)
Adaptive Moving Average (AMA)
Absolute Price Oscillator (APO)
Aroon (AR) Aroon Oscillator (ARO)
Average True Range (ATR)
Volume on the Ask (AVOL)
Volume on the Bid and Ask (BAVOL)
Bollinger Band (BBANDS)
Band Width (BW)
Commodity Channel Index (CCI)
Chande Momentum Oscillator (CMO)
Double Exponential Moving Average (DEMA)
Directional Movement Indicators (DMI)
Exponential Moving Average (EMA)
Fill Indicator (FILL)
Ichimoku (ICH)
Keltner Channel (KC)
Linear Regression (LR)
Linear Regression Angle (LRA)

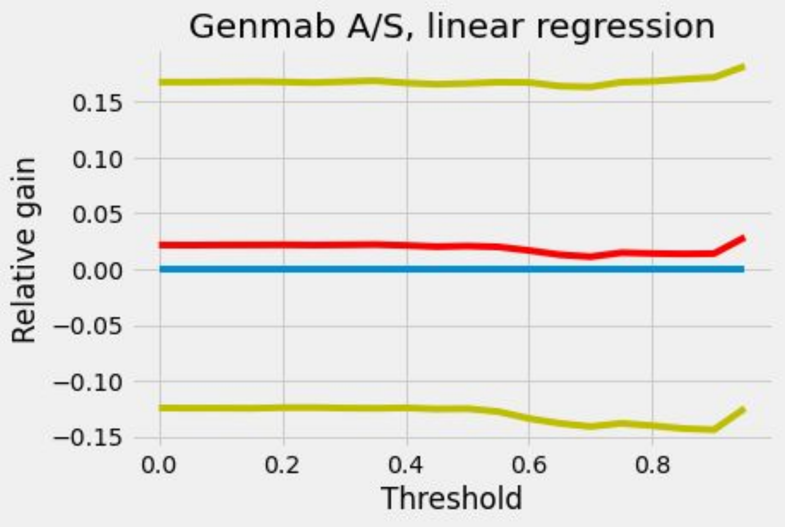
Linear Regression Intercept (LRI)
Linear Regression Slope (LRM)
Moving Average Convergence Divergence (MACD)
Max (MAX)
Money Flow Index (MFI)
Midpoint (MIDPNT)
Midprice (MIDPRI)
Min (MIN)
MinMax (MINMAX)
Momentum (MOM)
Normalized Average True Range (NATR)
On Balance Volume (OBV)
Price Channel (PC)
Percent Price Oscillator (PPO)
Price Volume Trend (PVT)
Rate of Change (ROC)
Rate of Change (ROC100)
Rate of Change (ROCP)
Rate of Change (ROCR)
Relative Strength Indicator (RSI)
Session Volume (S_VOL)

Parabolic Sar (SAR)
Simple Moving Average (SMA)
Standard Deviation (STDDEV)
Stochastic (STOCH)
Stochastic Fast (StochF)
T3 (T3)
Triple Exponential Moving Average (TEMA)
Triangular Moving Average (TRIMA)
Triple Exponential Moving Average Oscillator (TRIX)
Time Series Forecast (TSF)
TT Cumulative Vol Delta (TT CVD)
Ultimate Oscillator (ULTOSC)
Volume At Price (VAP)
Volume (VOLUME)
Volume Delta (Vol ∆)
Volume Weighted Average Price (VWAP) Williams % R (WillR)
Weighted Moving Average (WMA)
Welles Wilder's Smoothing Average (WWS)

Linear regression

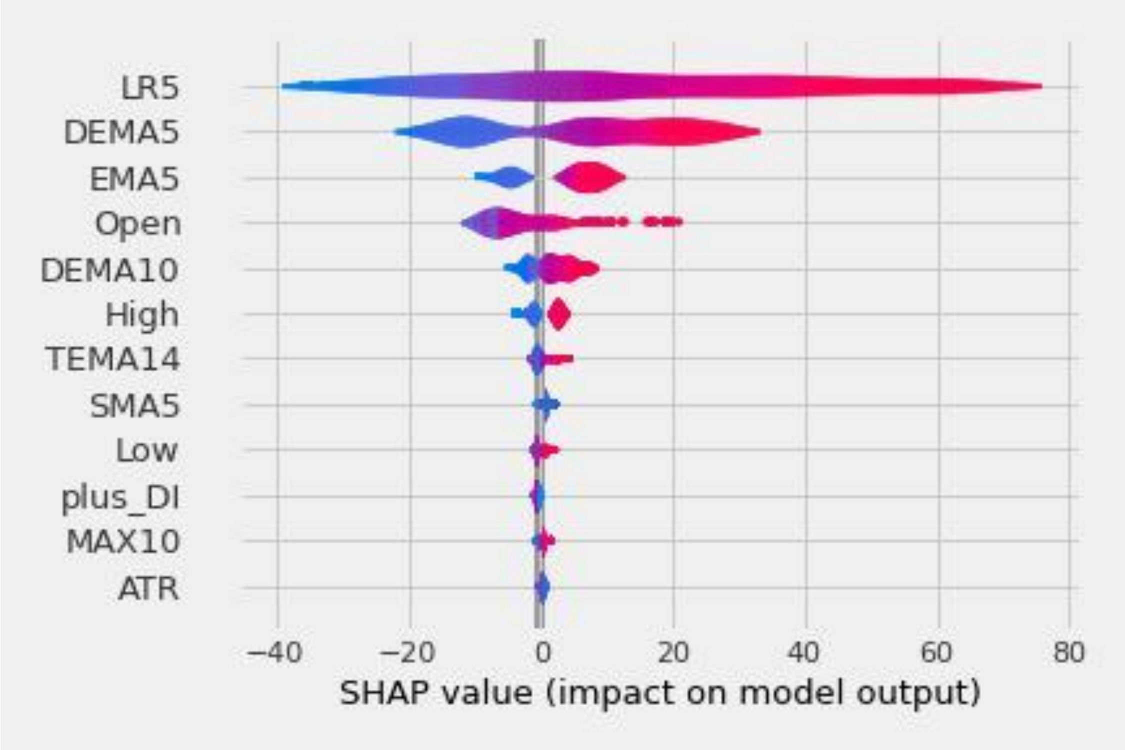




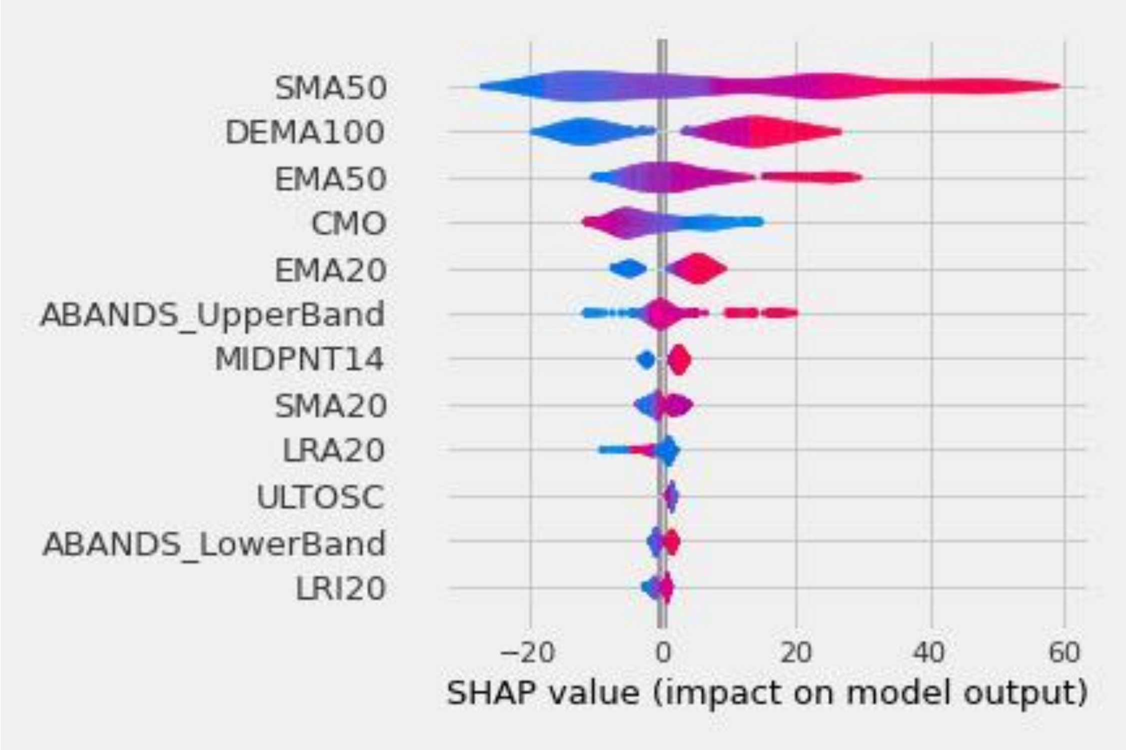


SHAP with LightGBM (Technical Indicators)

Predicting 1 day ahead (Jysk)

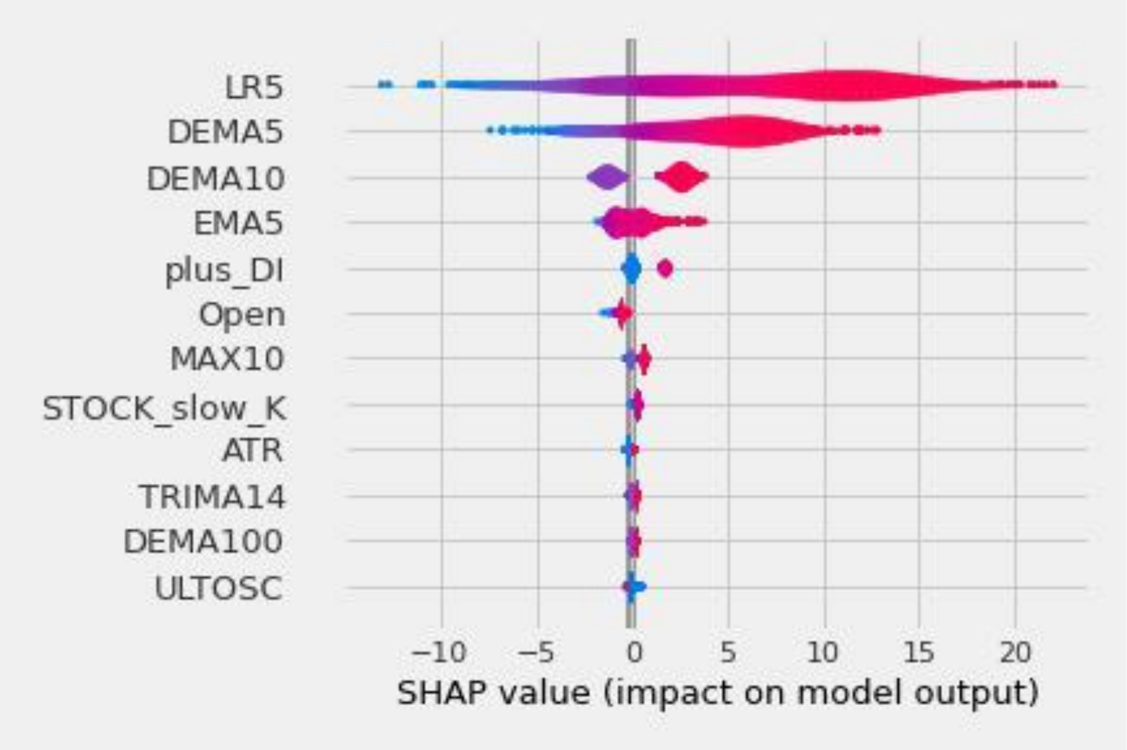


Predicting 20 day ahead (Jysk)

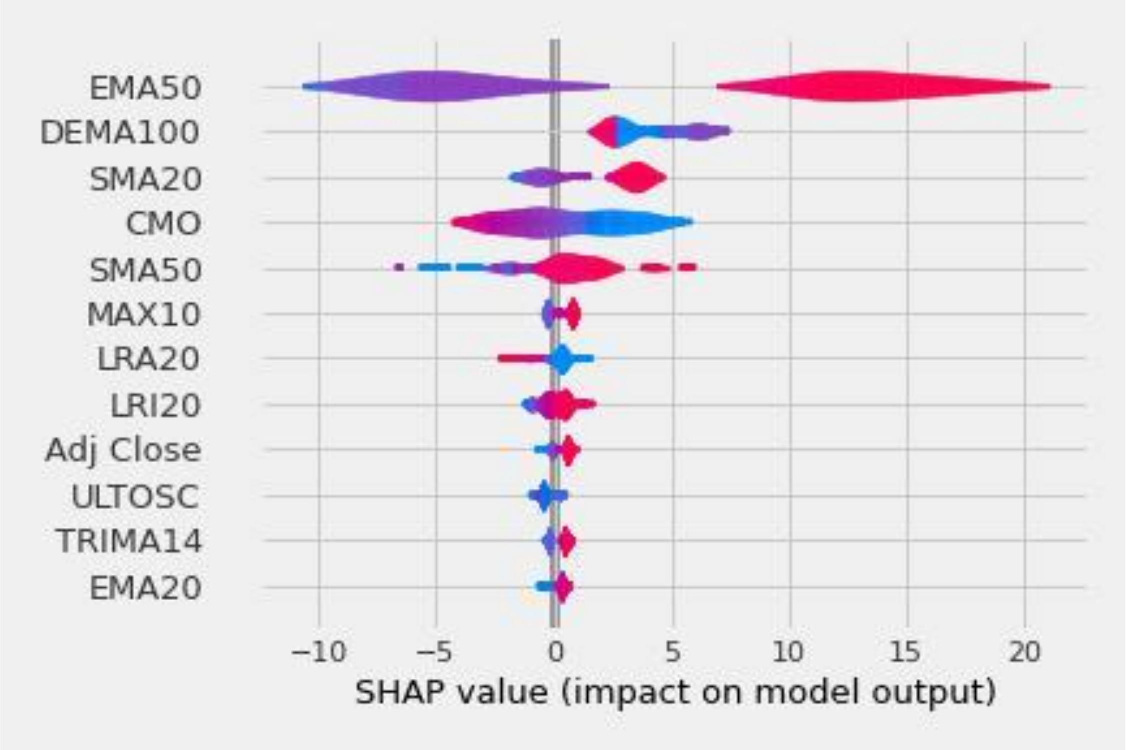


SHAP, Danske Bank

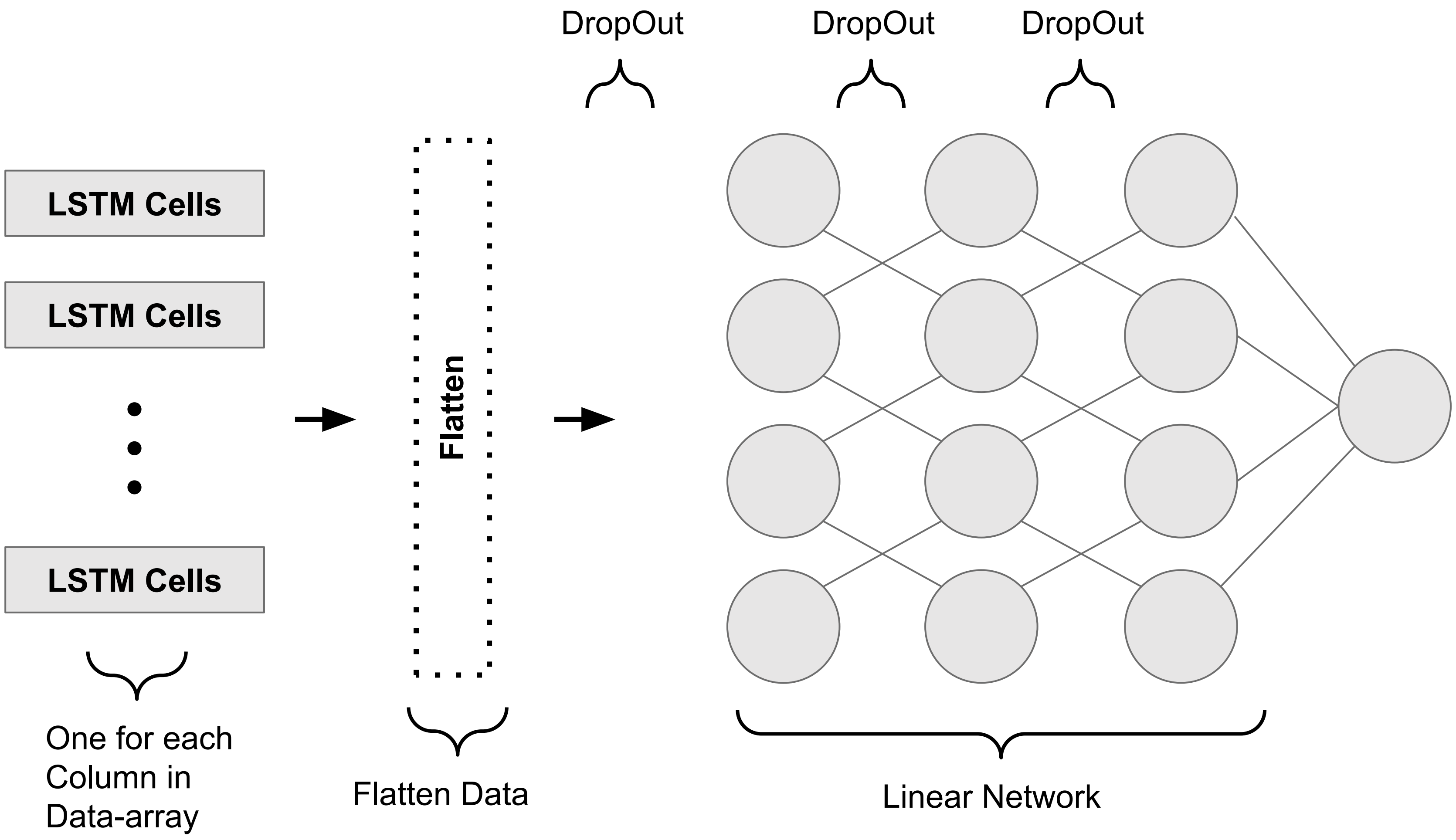
Predicting 1 day ahead (Danske Bank)



Predicting 20 day ahead (Danske Bank)



PyTorch LSTM Network



Params:

- LogLoss
- Cross-validation
- Shuffle
- Batches

Output:
Probability of price increasing in m days

Metrics

Testing Metrics

- LogLoss
- AUC
- Accuracy
- MaxGain
- MaxGain Probability

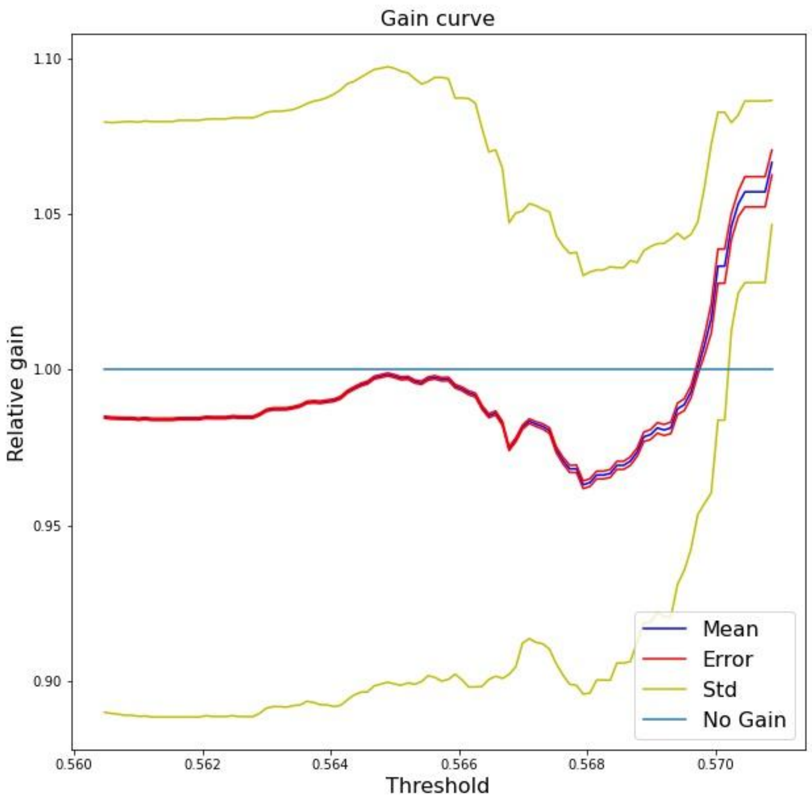
$$\mathrm{Gain}(t) = \mathrm{mean}(\mathrm{Return}\{P(\mathrm{UpTick}) > t\})$$

$$\mathrm{MaxGain} = \mathrm{Max}\{\mathrm{Gain}(t), t\}$$

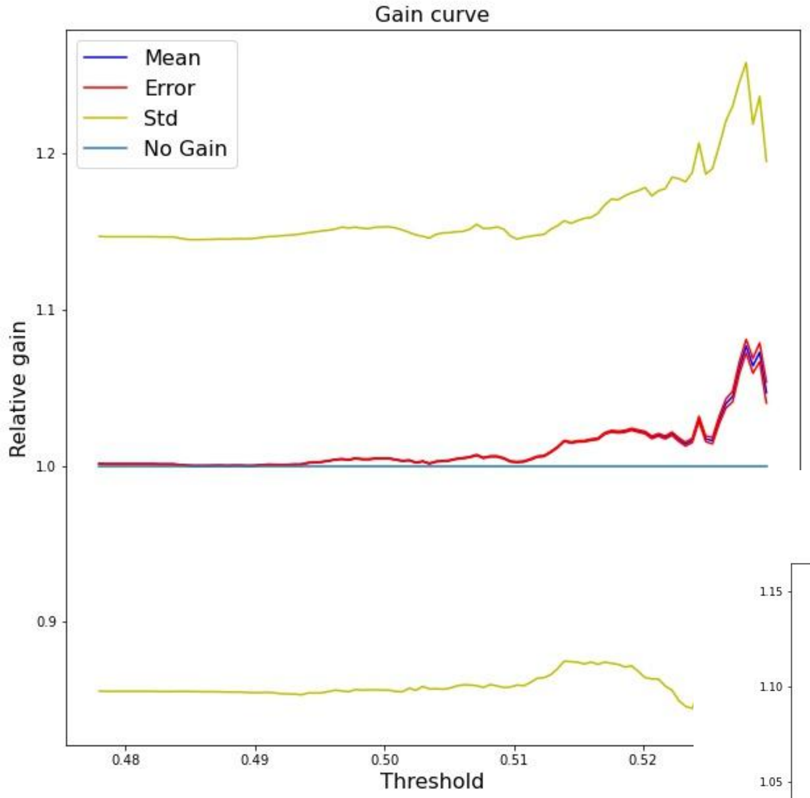
$$\mathrm{MaxGainProbability} = \mathrm{Max}\left\{\frac{\mathrm{Gain}(t) - 1}{STD(\mathrm{Gain}(t))}, t\right\}$$

Gain curves

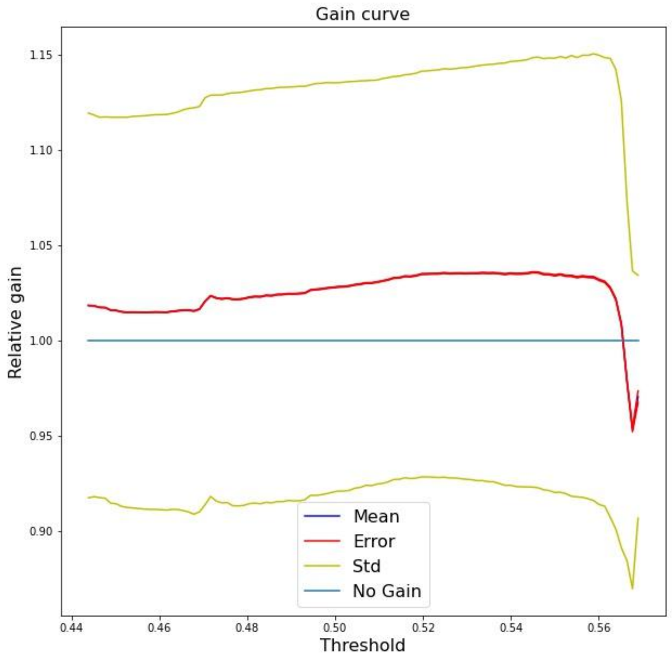
Ørsted A/S



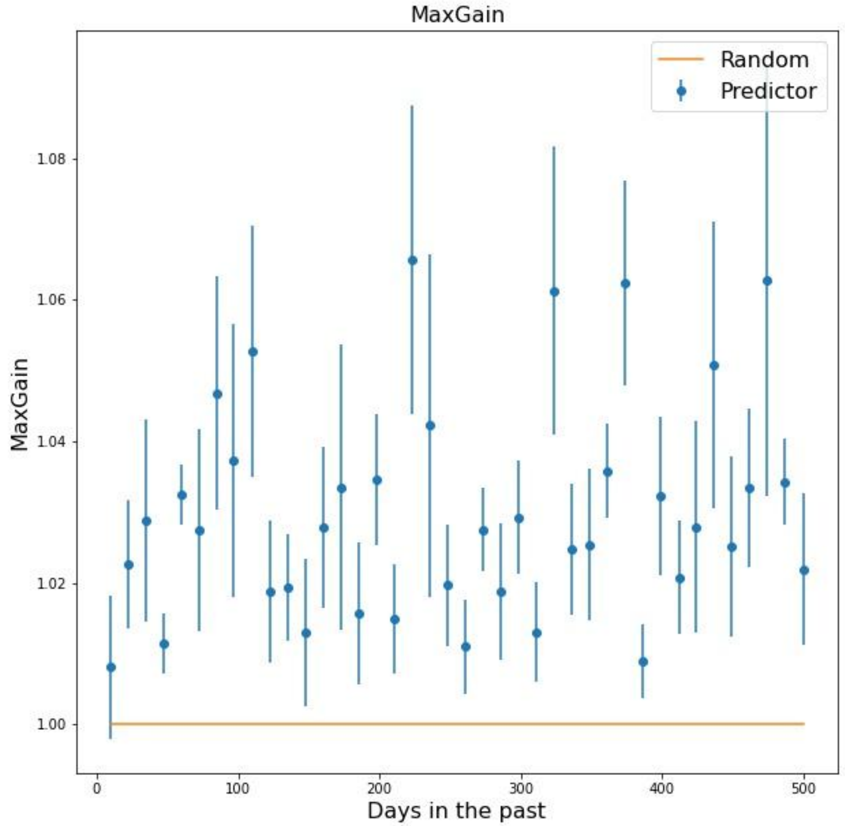
Bavarian Nordic A/S

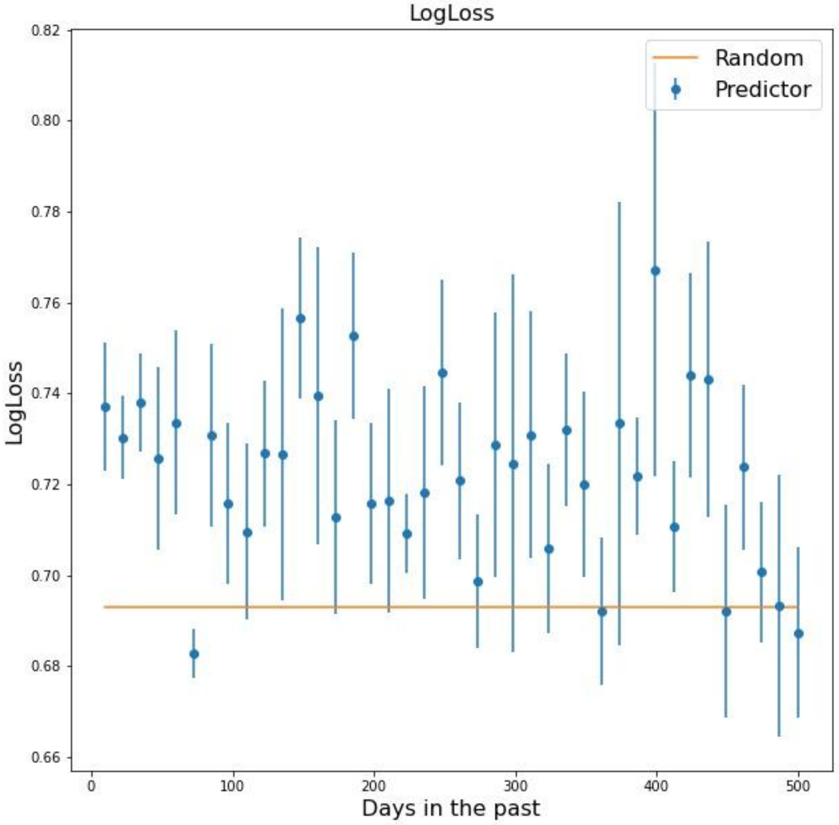


Genmap A/S



Parameter optimization: Days in the past

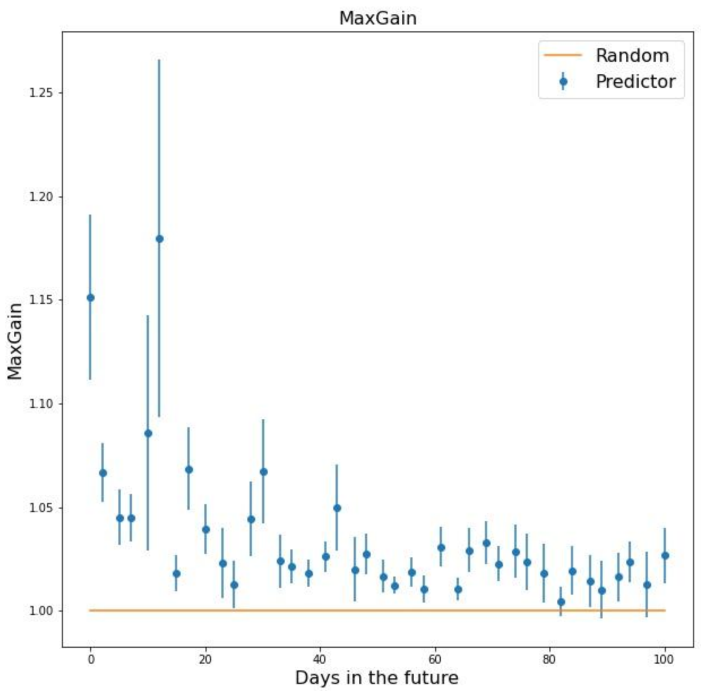


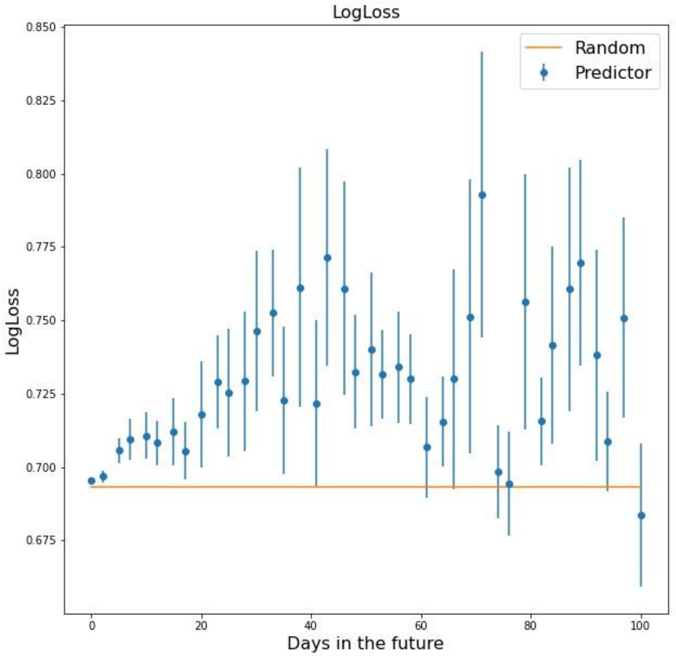


Nr. of days in the past to base prediction on

Used 5-fold cross validation to get errors

Parameter optimization: Days in the future



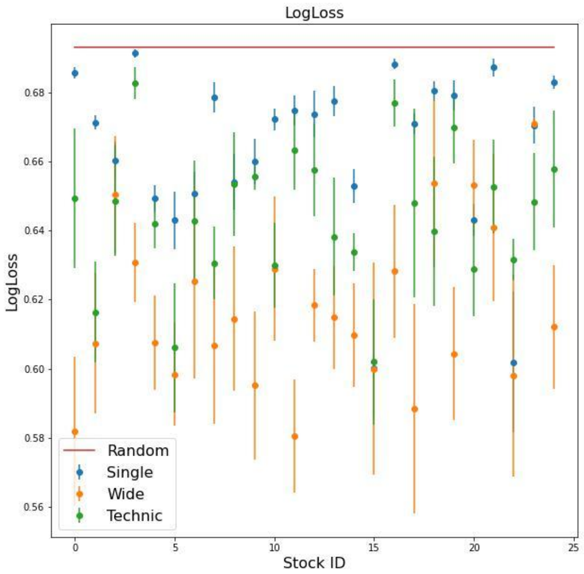


Nr. of days into the future to predict

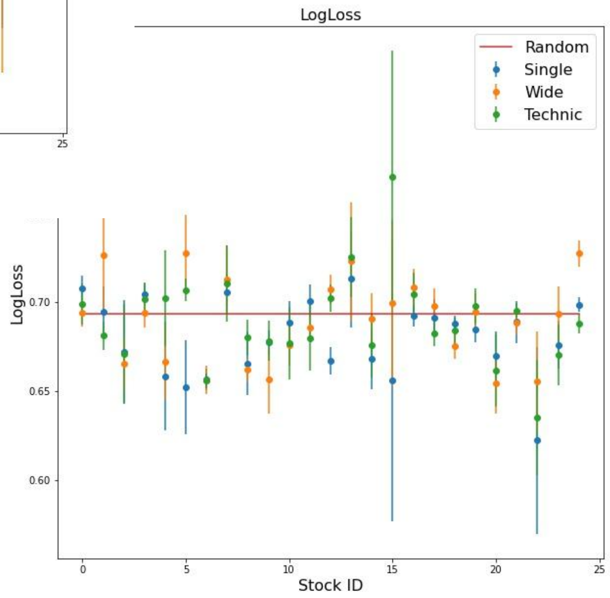
Used 5-fold cross validation to get errors

Data types

Shuffled data



Ordered data



Wide is best for shuffled data

Essentially the same for ordered data, technical indicators are a bit better

Different data for prediction:

- Single: Stock data for the stock to predict
- Wide: Stock data for all stocks to predict one of them
- Technic: Stock data and technical indicators for the stock to predict

Failed attempts

Training on all stocks

idea:

- More data to learn trends from
- Better chance of recognizing critical events

Reality:

- Stocks are too different

Using technical indicators for all stocks to predict one stock value

Idea:

- Using all stocks worked well
- Technical indicators are useful

Reality:

- Too much data (68*25*20=34000 inputs to linear network): Unable to train

Evaluation

Model	Performance	Training
LSTM network	Good performance	Slow and sometimes unreliable
Linear regression	Slightly better than random	Super fast
Support vector machine	Random	Fast
EchoState network	Bad (did not work)	Fast

More things to be done..

- SHAP values on PyTorch
- Get errors on model via bootstrapping
- Get optimal threshold for all stocks
- Optimize Hyperparameters
- Incorporate Stock-Portfolio Allocations on the go
- Shadow-Trading with optimal model

Appendix

Details and workflow of the projects

The following were done to obtain the objective: Make an ML algorithm to get as much return on the OMXC25 Stocks as possible:

We tried the following:

Making an LSTM-NN which could predict if future stocks would increase in value. Here we first found the number of days into the future that we wanted to predict as well as how many days in the past should be used for each prediction. Then we used the standard OHLCV data for each stock, compared that to using all data (WIDE) from all of the OMXC25 index. We then calculated +60 Technical Indicators which we also used in addition.

The idea was that if the model is sure (above som threshold) that a stock will go up, then buy and sell later. This buy-hold strategy was our way of checking how much “money” we have earned.

There are loads of studies on stock market behavior, and using Modern Portfolio Theory (MPT) we cross-referenced what we found with the ML to see if our machine learning algorithm actually made sense.

This was the case since MPT told us, that all stocks in the OMXC25 index were correlated (except ISS) and that made sense when the WIDE dataset were giving better result then just the pure OHLCV data. In addition the MPT told us that Ørsted is a good choice for good returns, both with the SHARPE-ratio strategy and with the minimum volatility strategy. This was true when looking at the result for MaxGain graph for Ørsted which was really good.

We also tried to look at a “simple” linear fit-model where we used a linear fit to predict future values.

The Dataset were really large when adding technical indicators and it was hard to tell which of them were the best. THere we used a lightGBM model to see the best performing technical indicators. However we wanted to do it on PyTorch, but this wouldn’t work.

Another thing we have worked on is to incorporate Echo-States into the algorithm, however here we has a lot of problems and it was hard to get any good predictions.

Roc- and gain curves

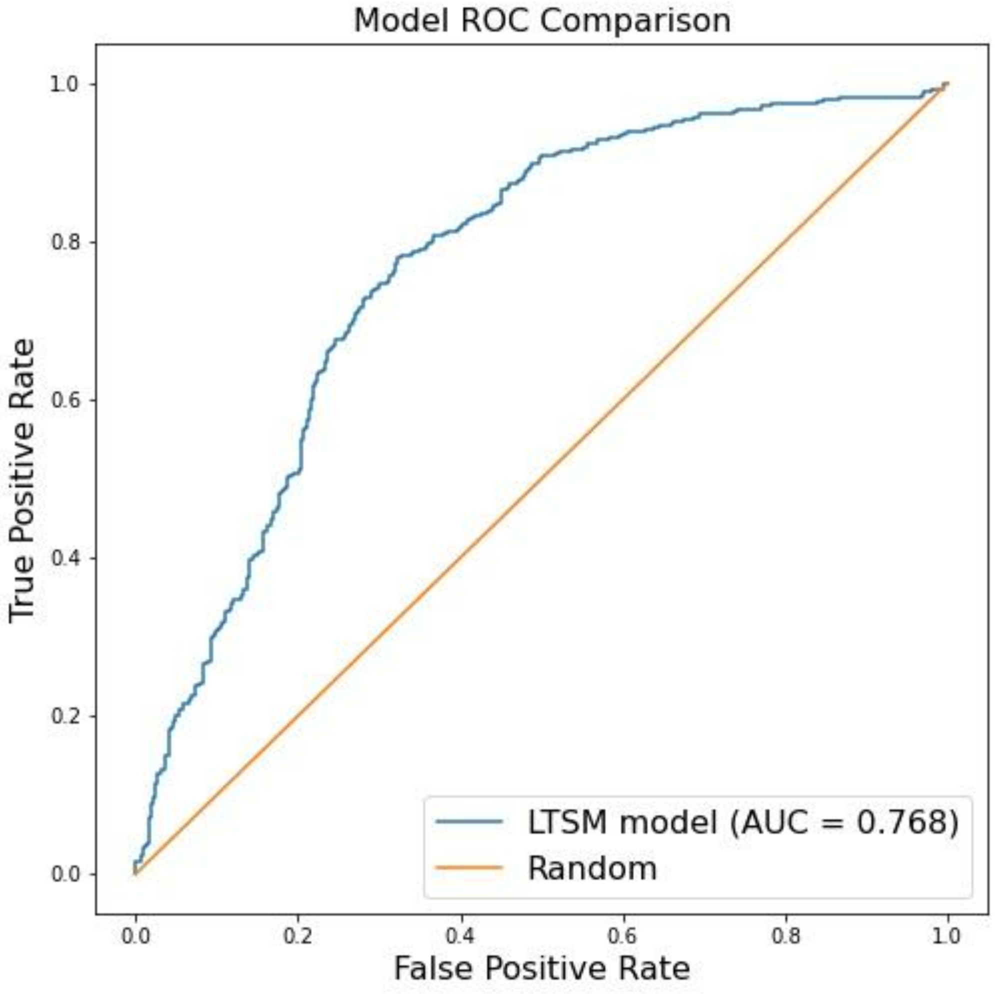
Left: Shuffled data

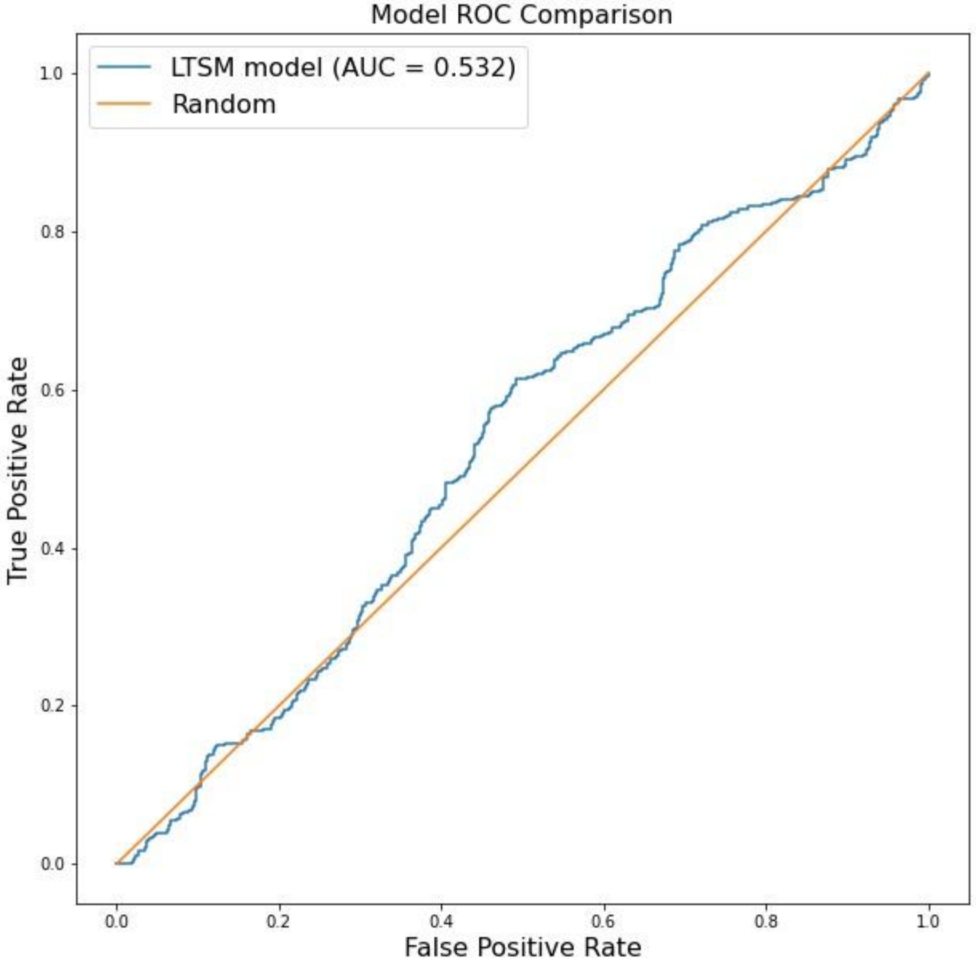
- Training is done on the entire dataset for random points used for testing

Right: Ordered data

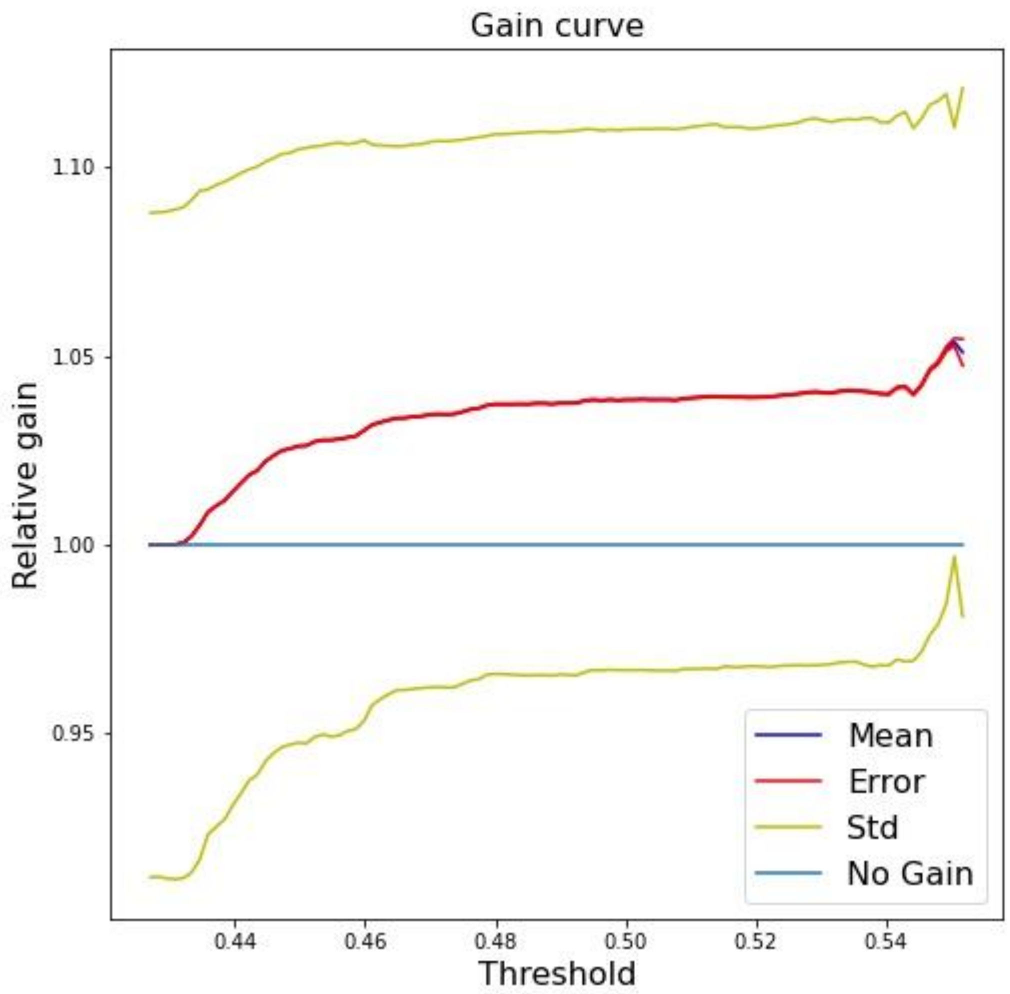
- Training is done up until some time, testing is done on the rest of the dataset

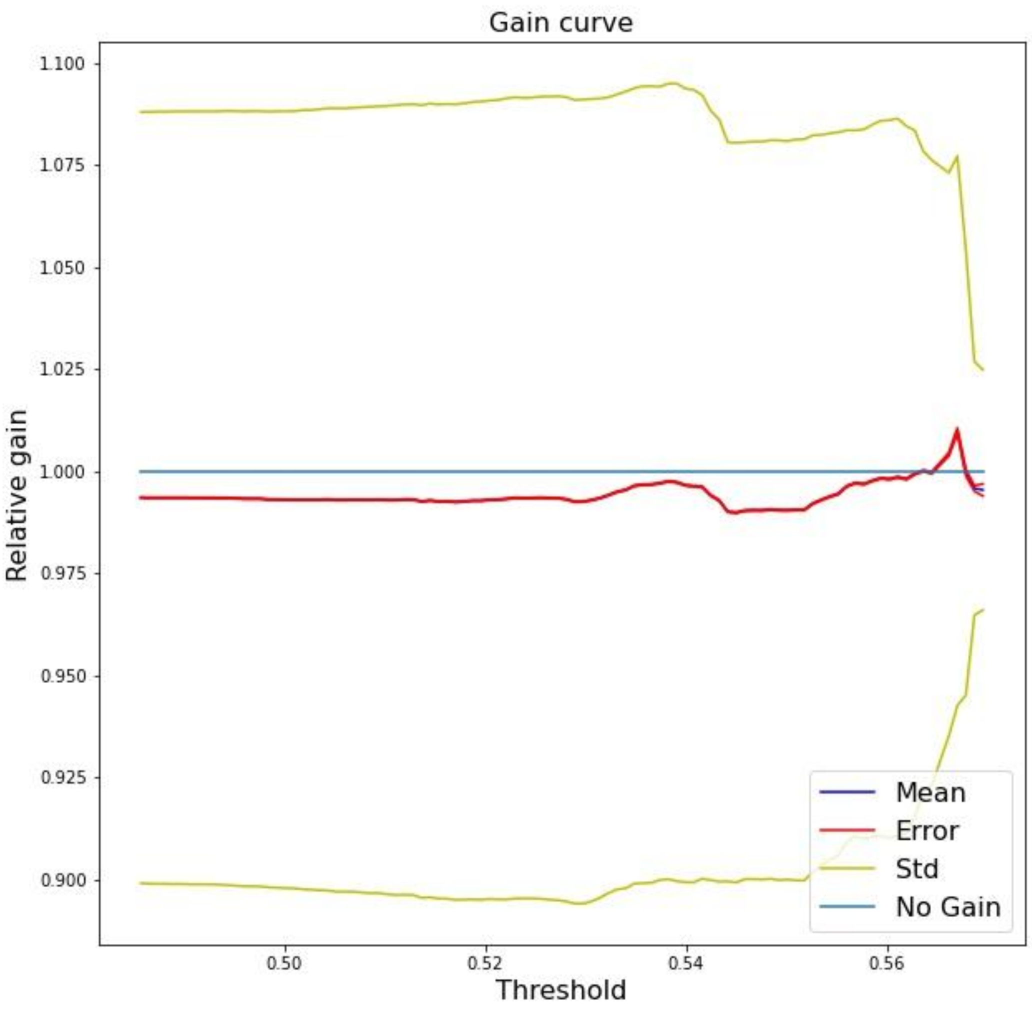
H. Lundbeck A/S



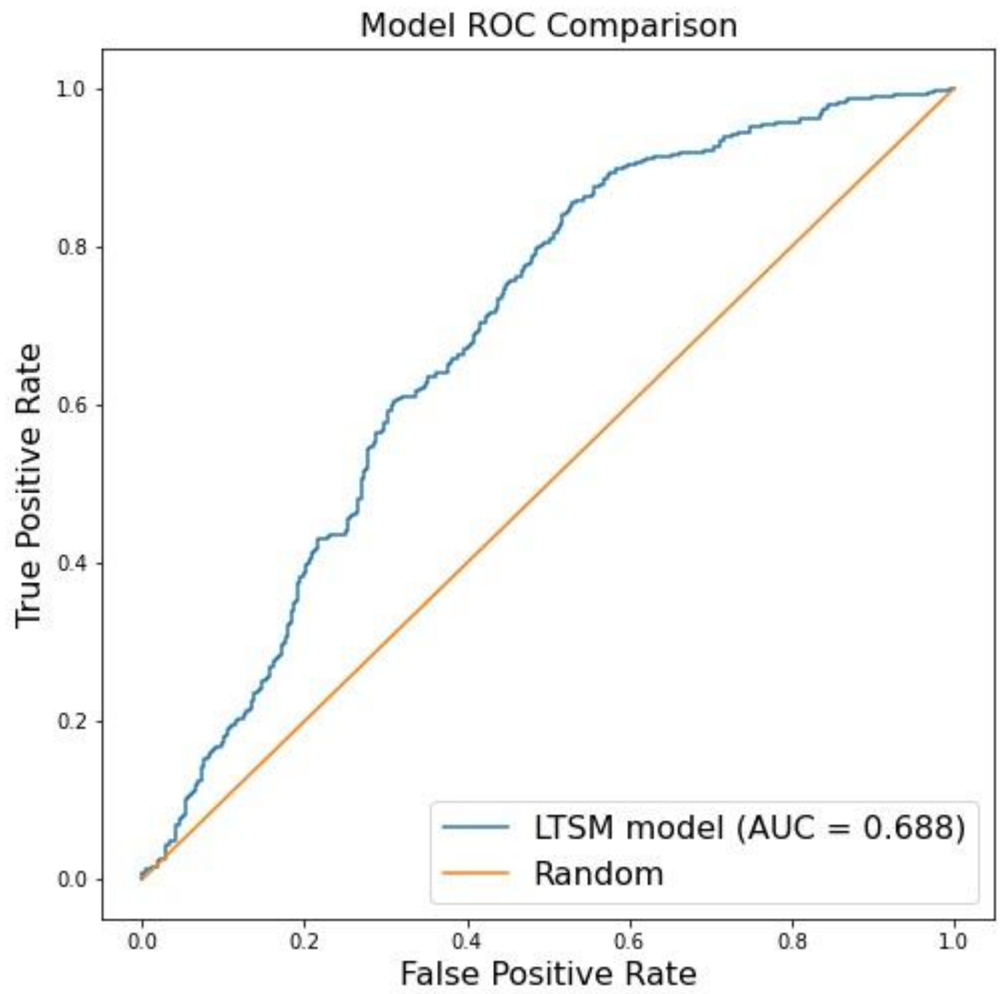


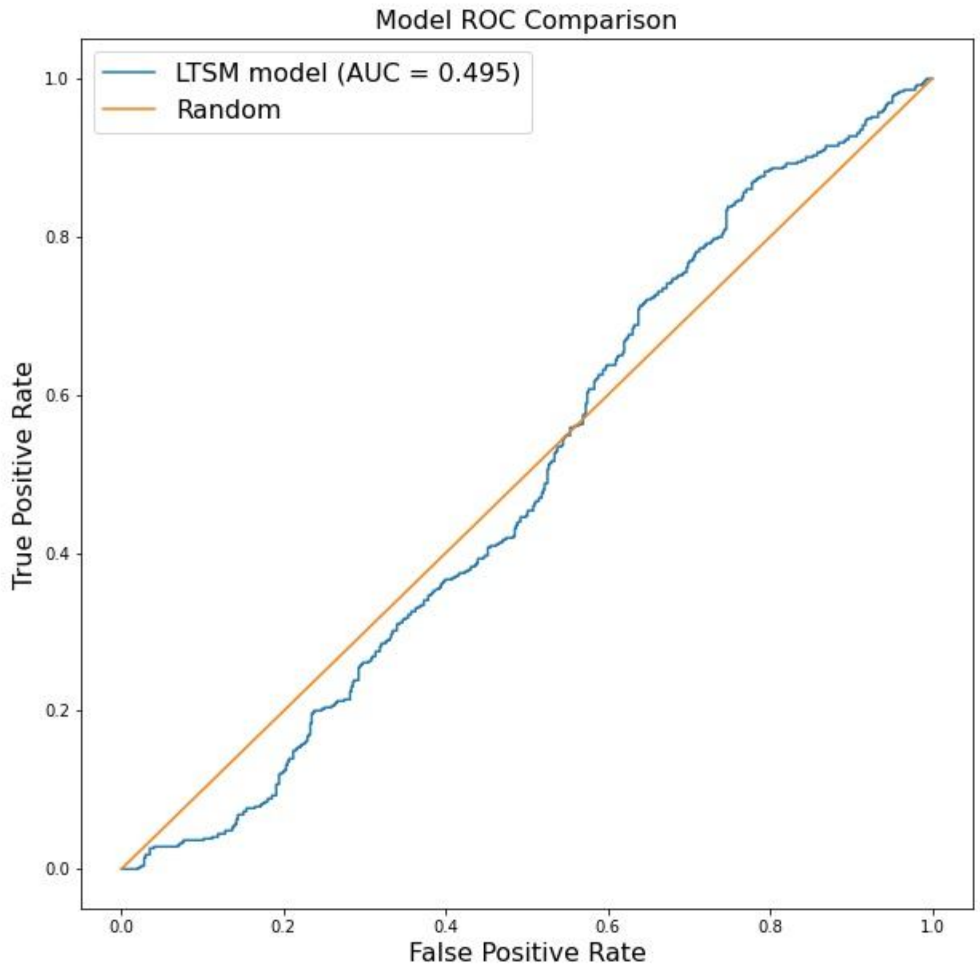
H. Lundbeck A/S



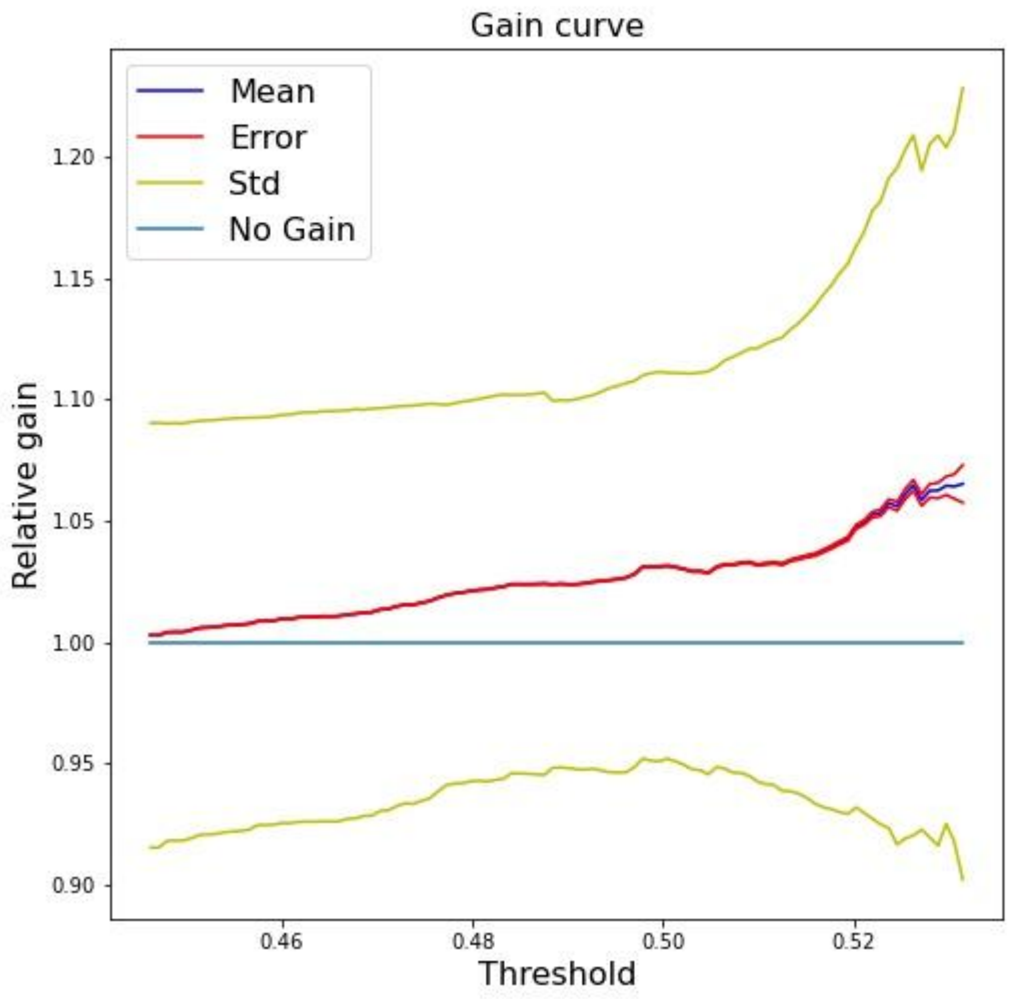


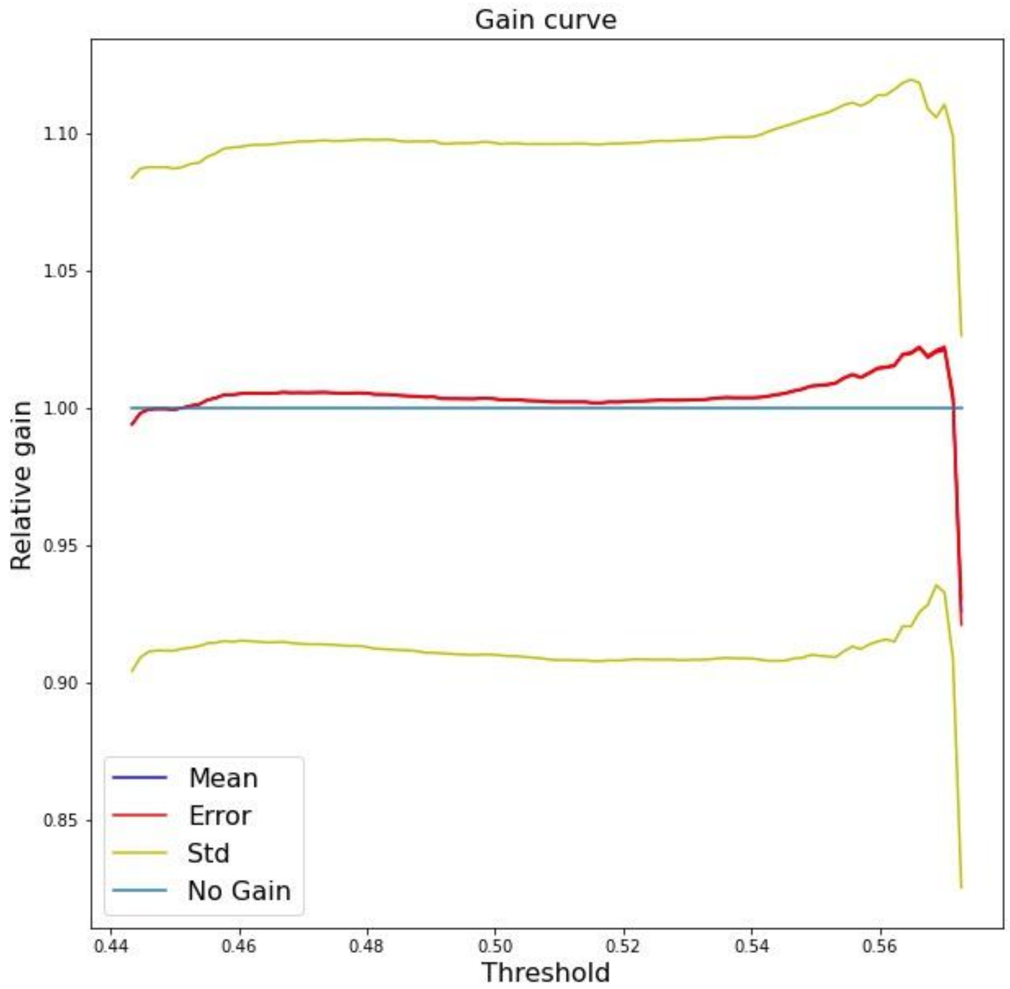
Danske Bank A/S



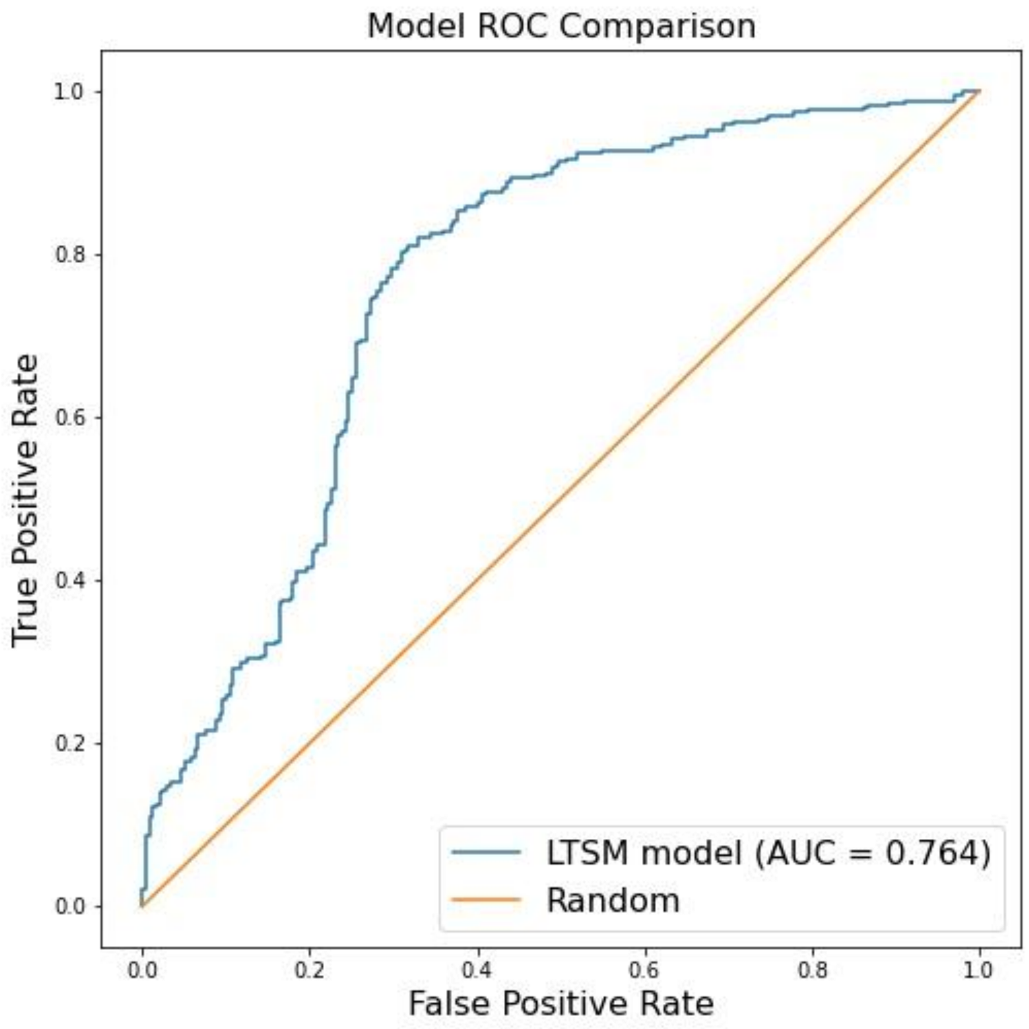


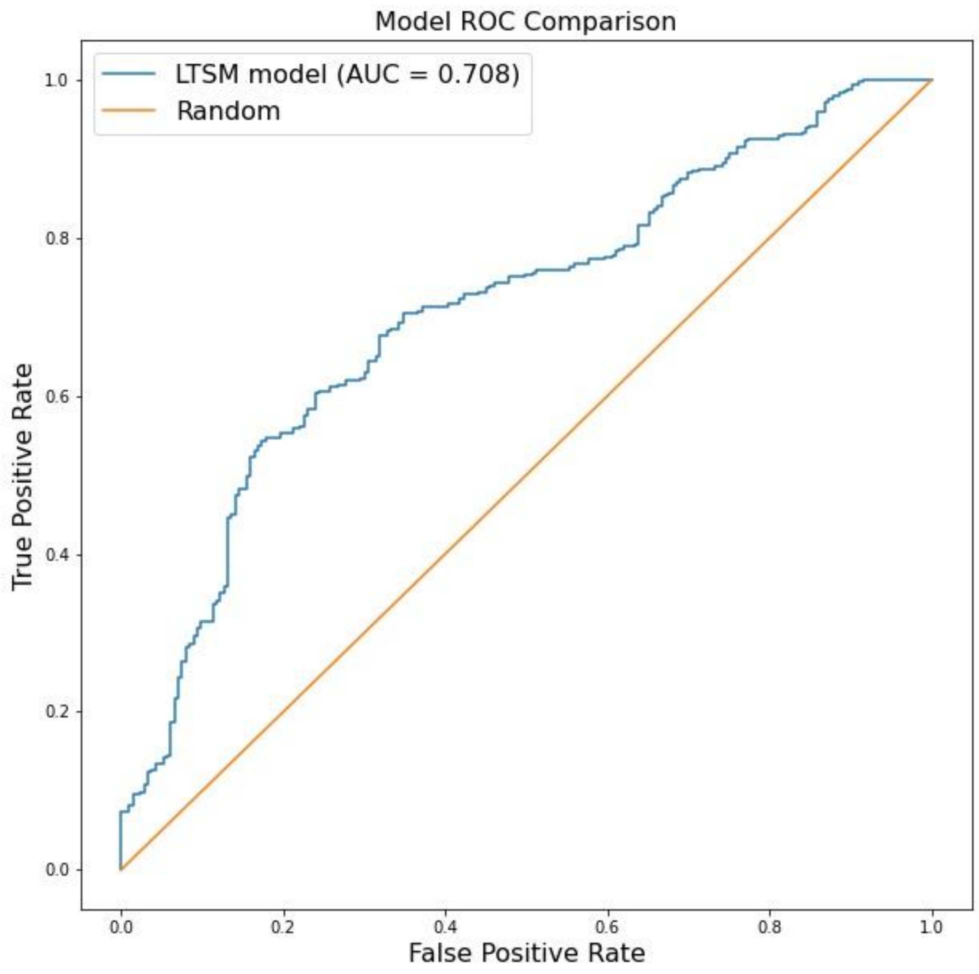
Danske Bank A/S



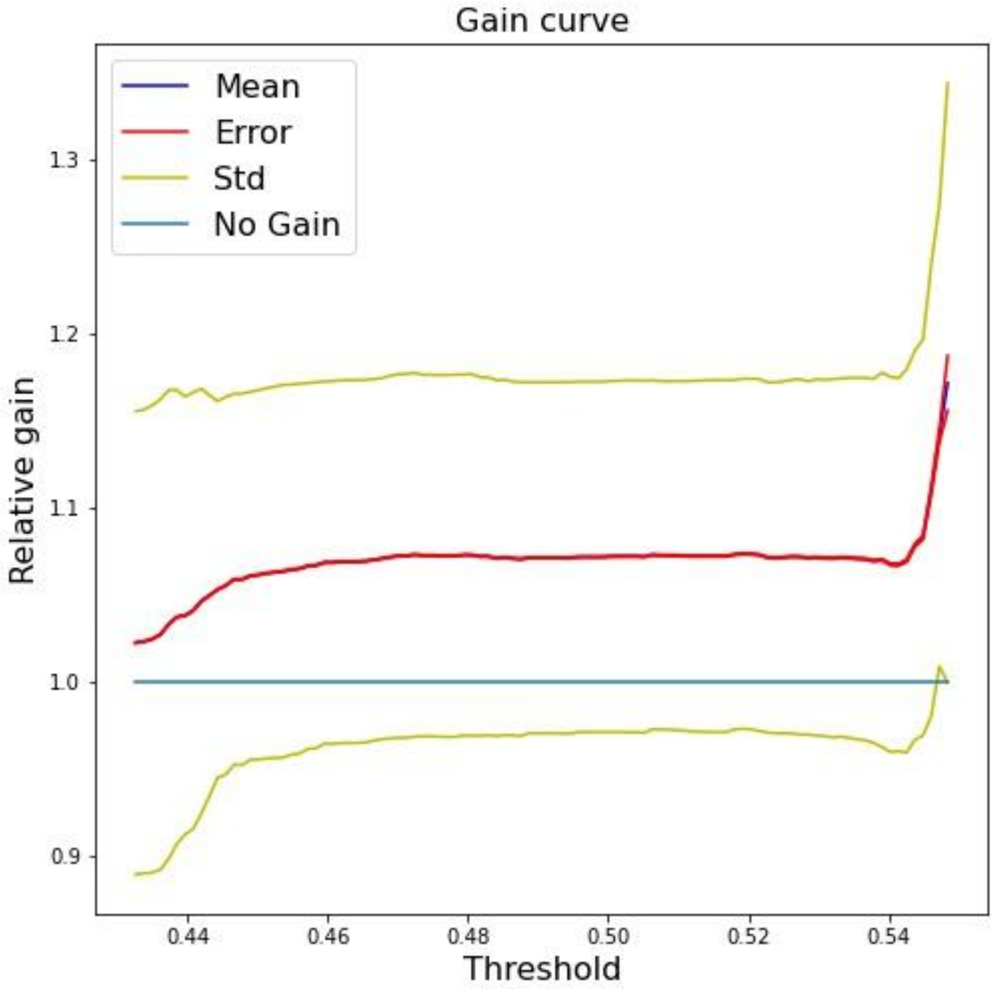


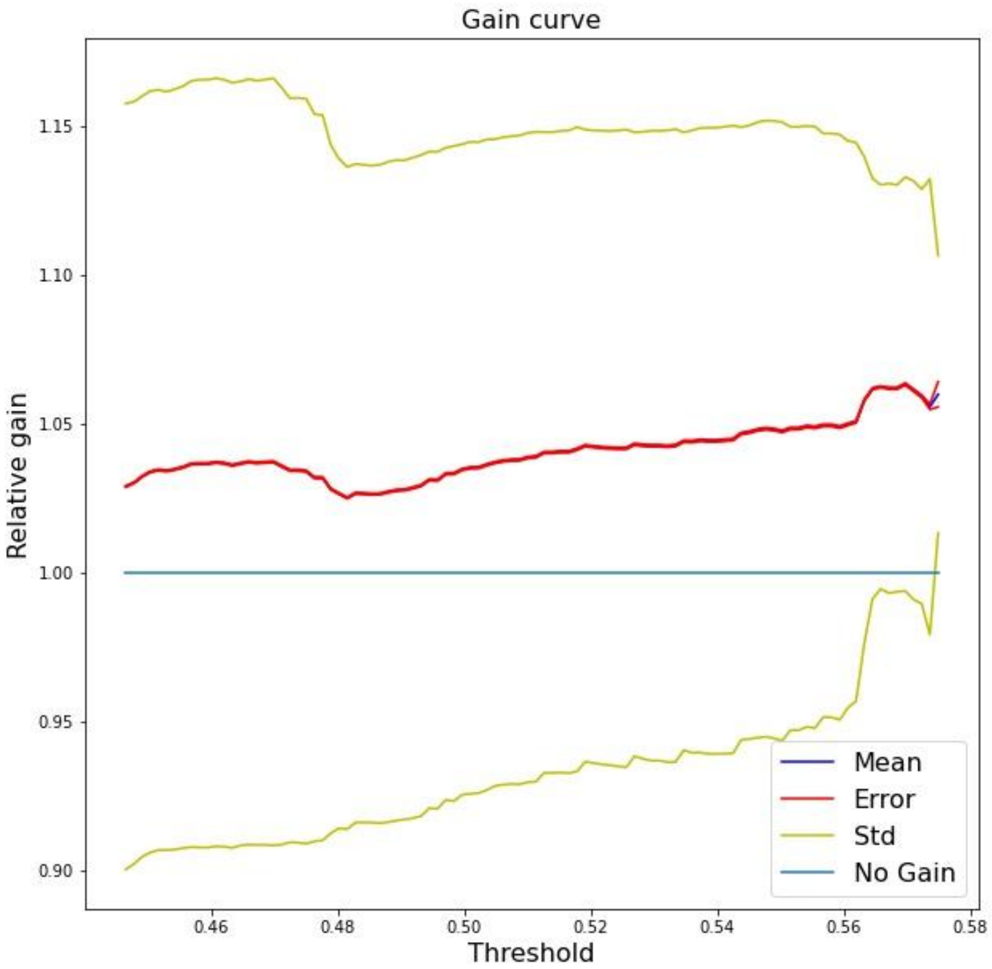
Pandora A/S


Model ROC Comparison
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
LTSM model (AUC = 0.764)
Random

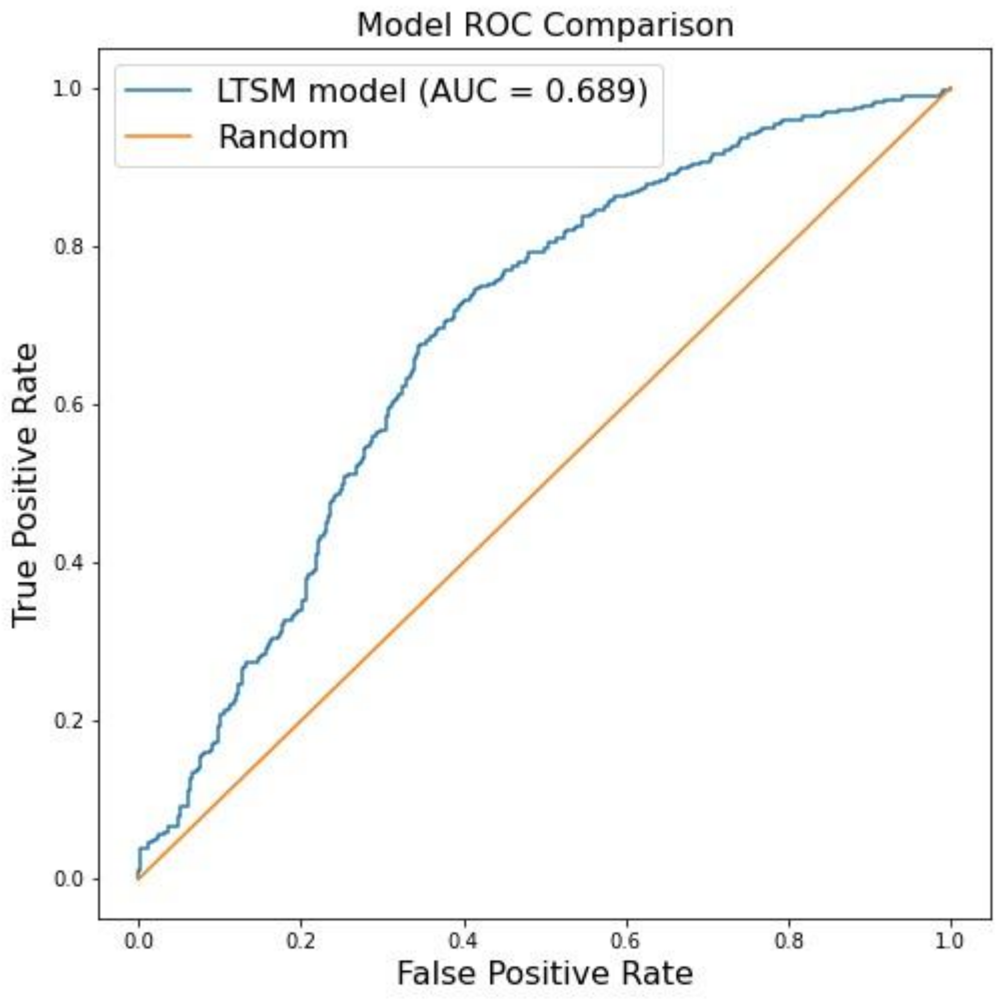

Model ROC Comparison
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
LTSM model (AUC = 0.708)
Random

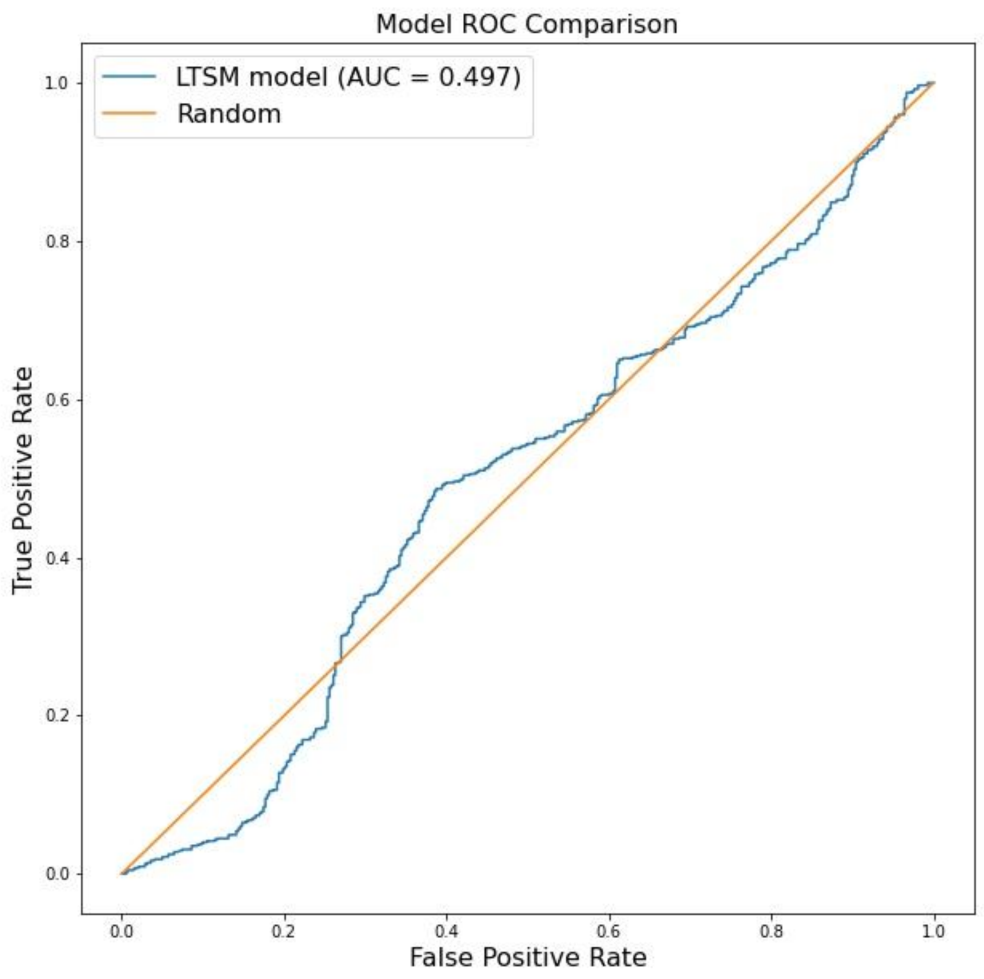
Pandora A/S


Gain curve
Mean
Error
Std
No Gain
Relative gain
Threshold
1.3
1.2
1.1
1.0
0.9
0.44
0.46
0.48
0.50
0.52
0.54

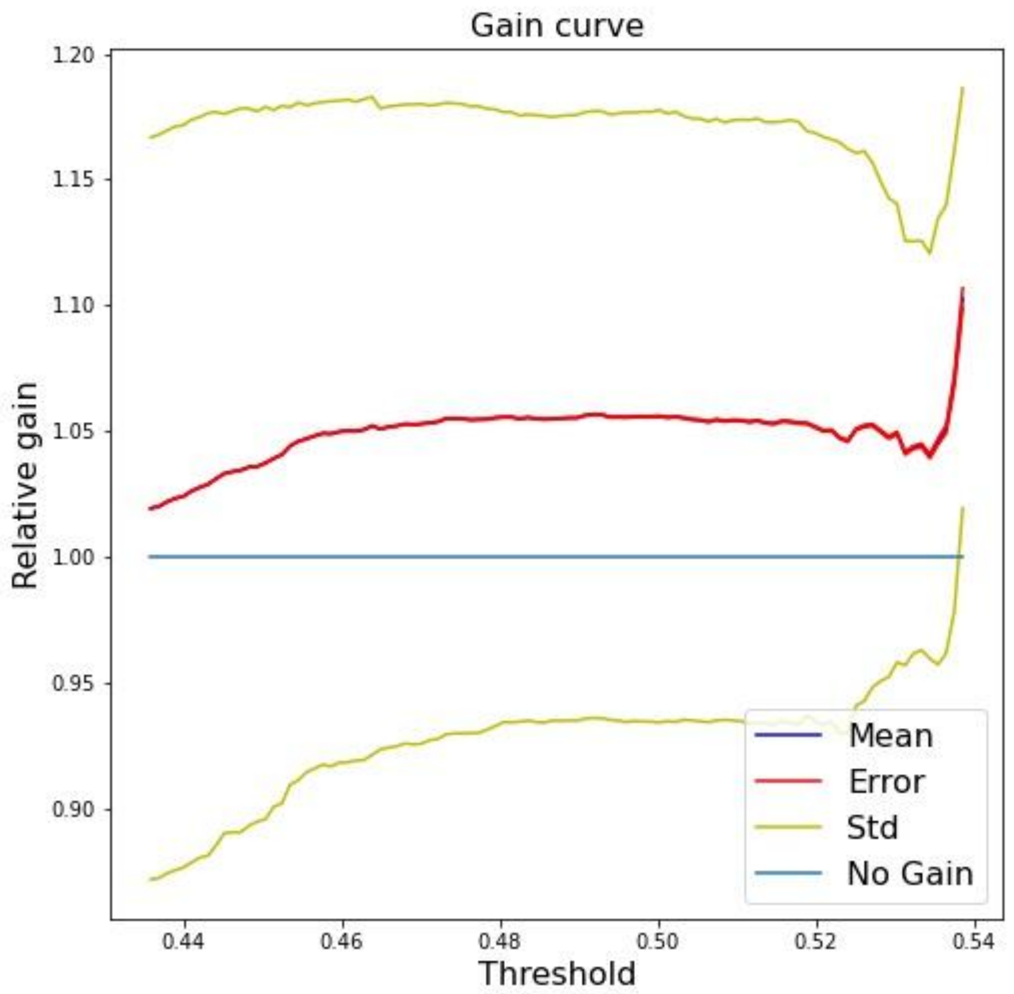

Gain curve
Mean
Error
Std
No Gain
Relative gain
Threshold
1.15
1.10
1.05
1.00
0.95
0.90
0.46
0.48
0.50
0.52
0.54
0.56
0.58

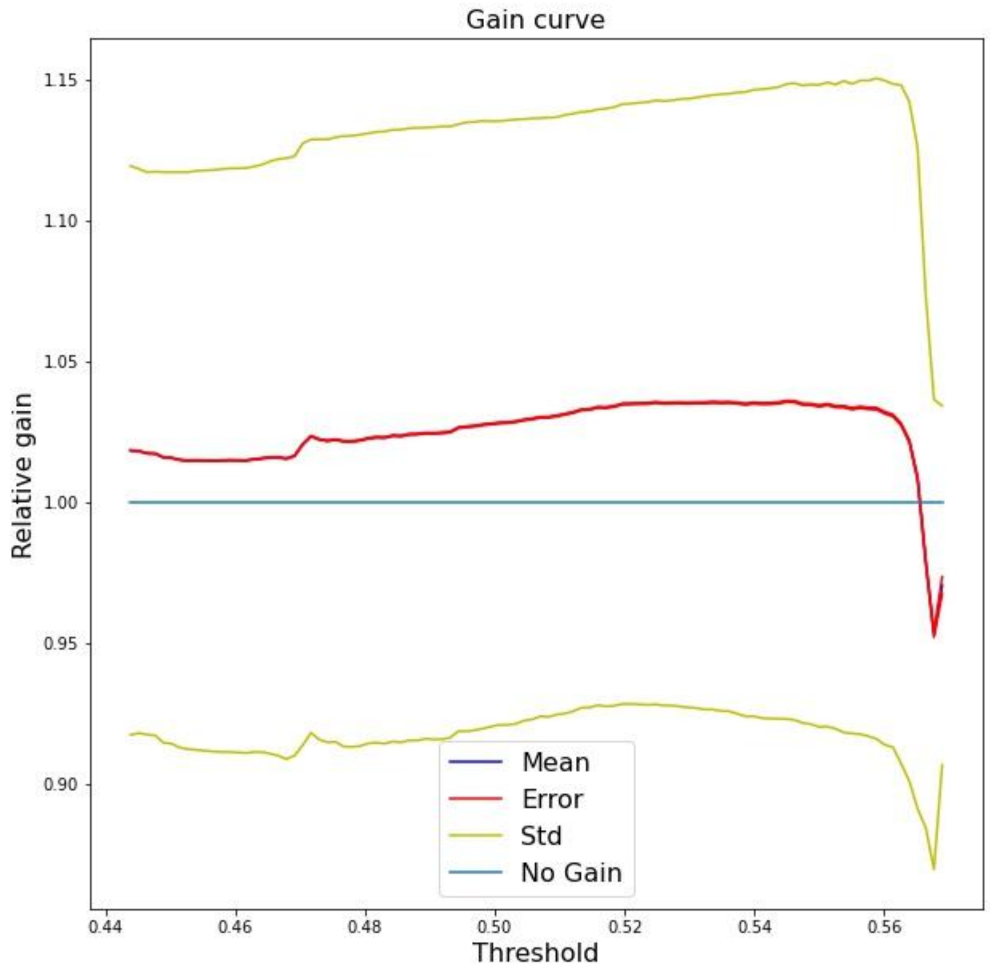
Genmab A/S


Model ROC Comparison
LTSM model (AUC = 0.689)
Random
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

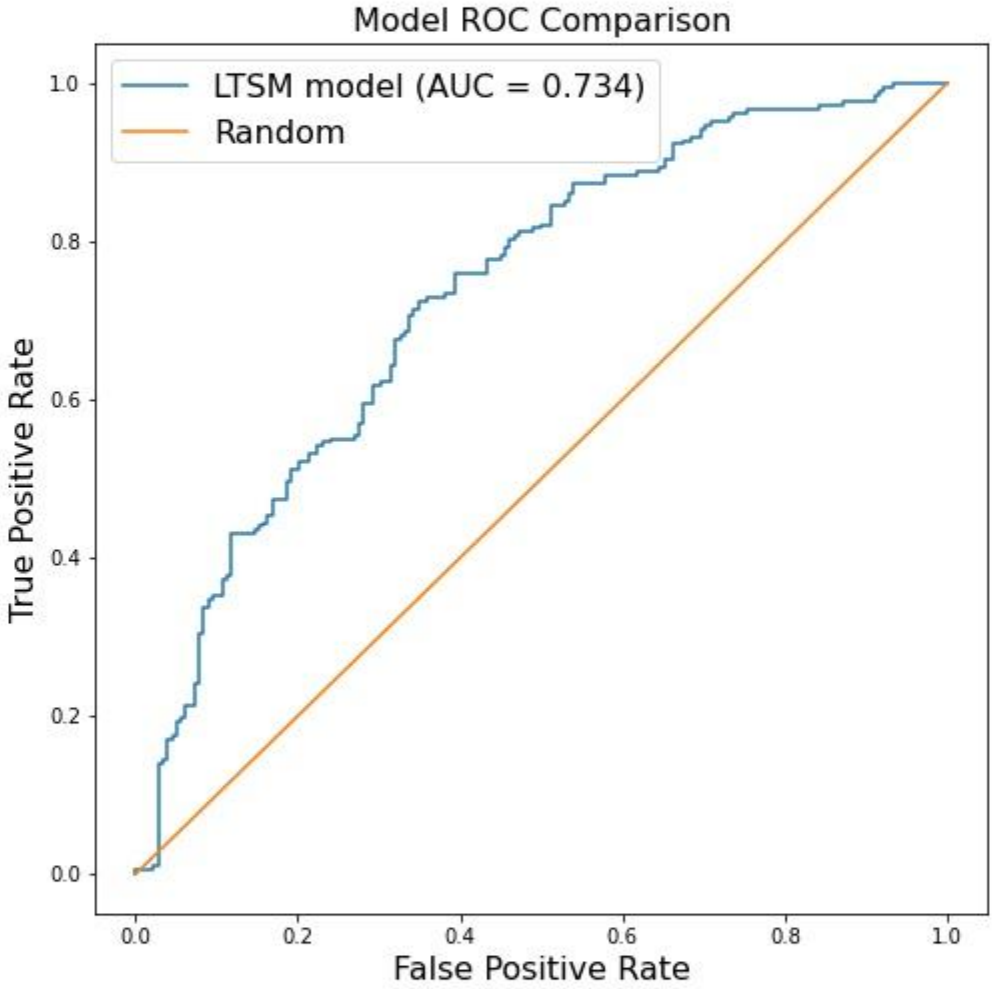

Model ROC Comparison
LTSM model (AUC = 0.497)
Random
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

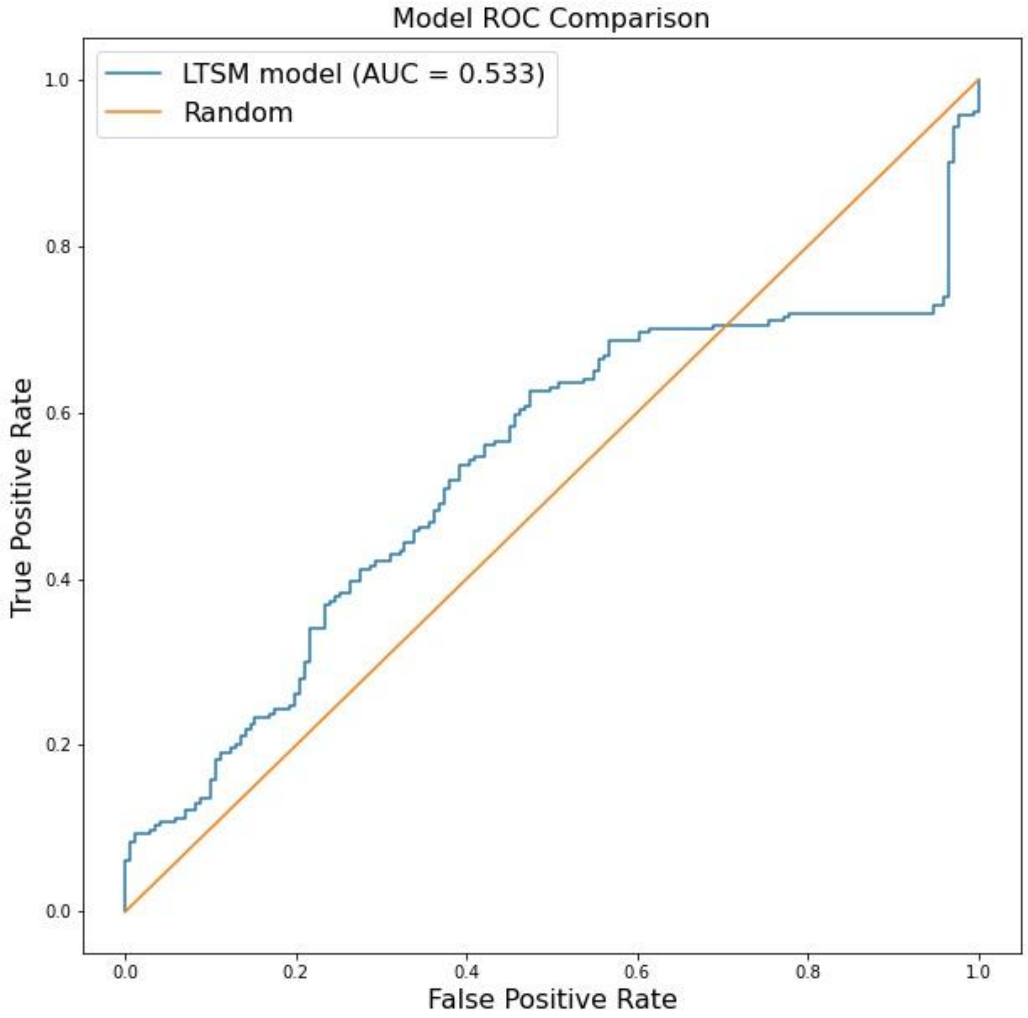
Genmab A/S


Gain curve
1.20
1.15
1.10
1.05
1.00
0.95
0.90
Relative gain
0.44
0.46
0.48
0.50
0.52
0.54
Threshold
Mean
Error
Std
No Gain

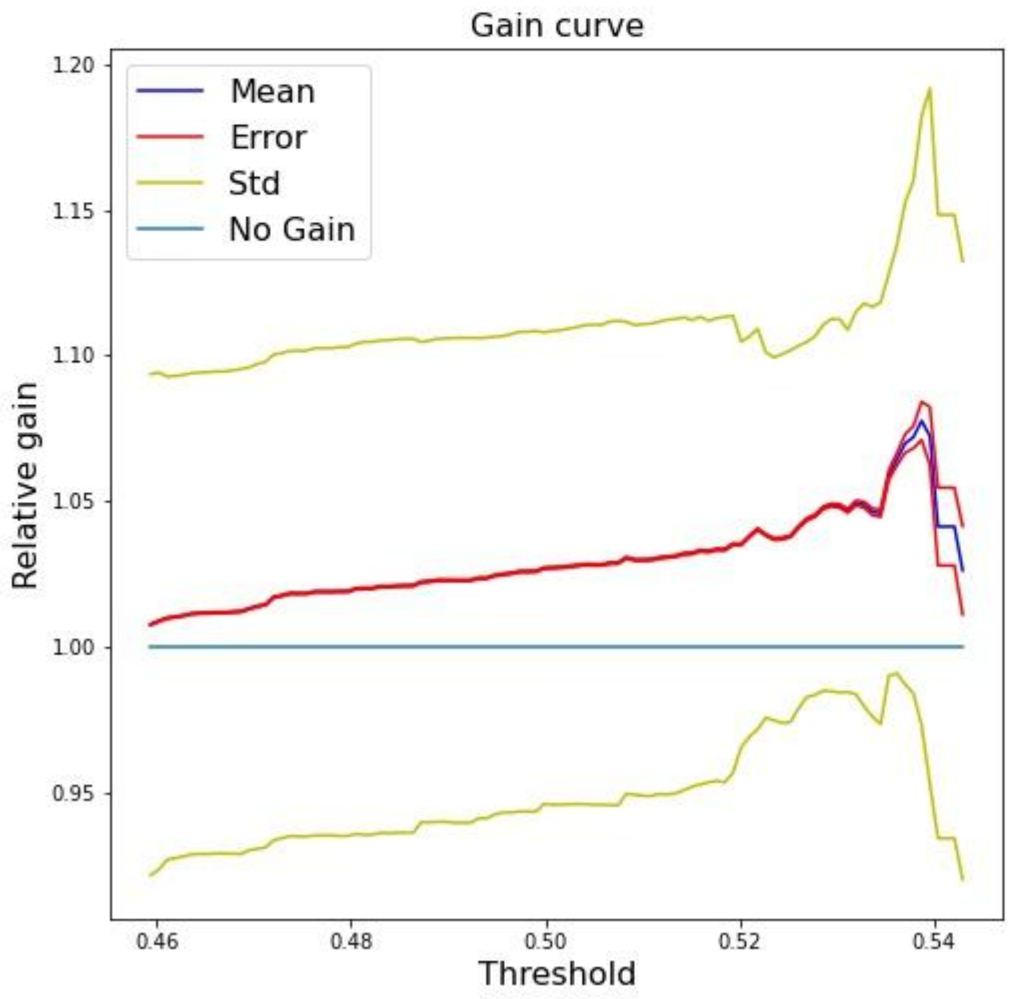

Gain curve
1.15
1.10
1.05
1.00
0.95
0.90
Relative gain
0.44
0.46
0.48
0.50
0.52
0.54
0.56
Threshold
Mean
Error
Std
No Gain

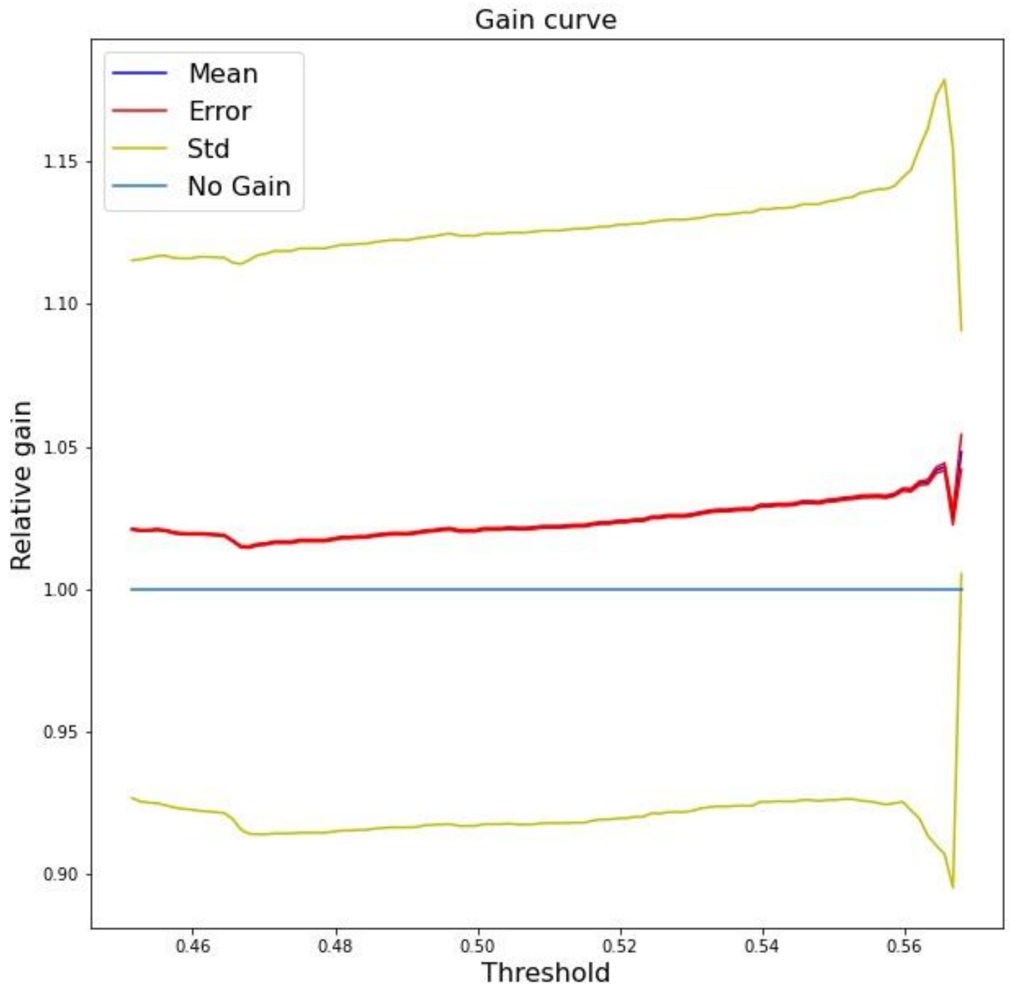
ISS A/S


Model ROC Comparison
LTSM model (AUC = 0.734)
Random
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

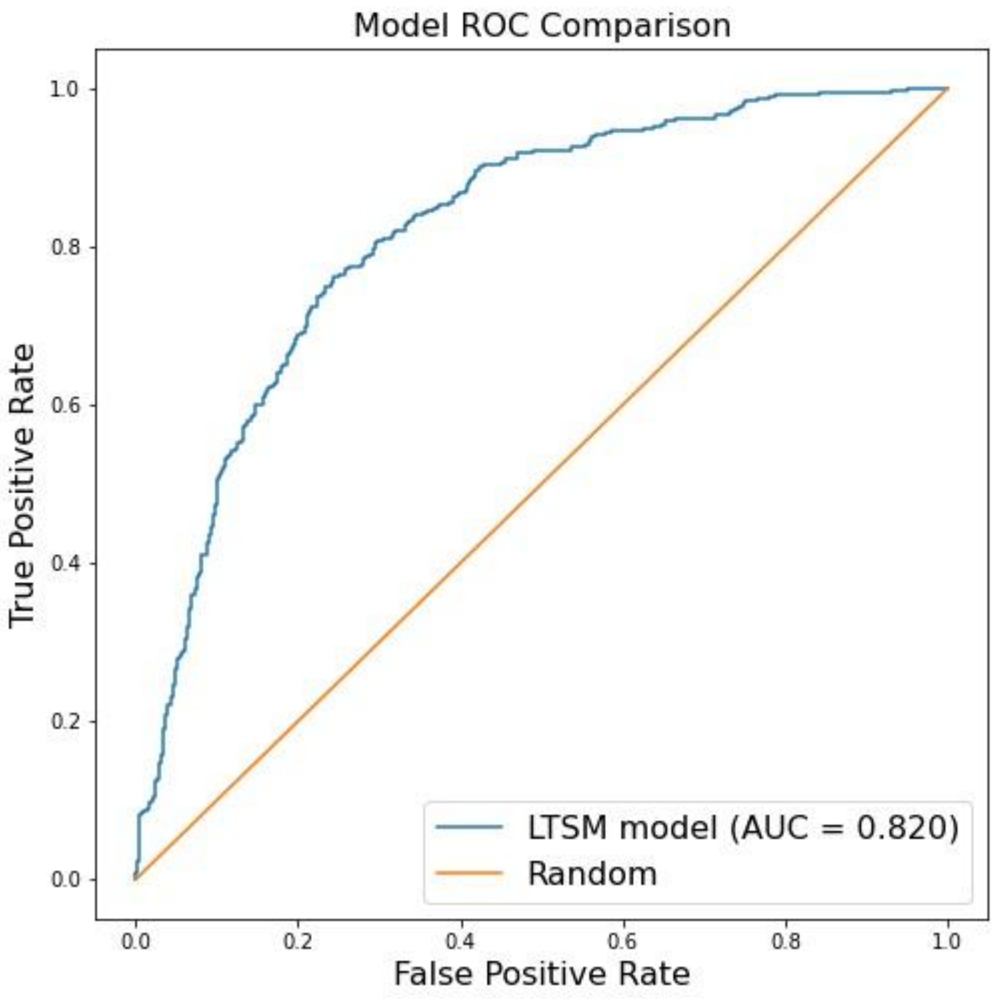

Model ROC Comparison
LTSM model (AUC = 0.533)
Random
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

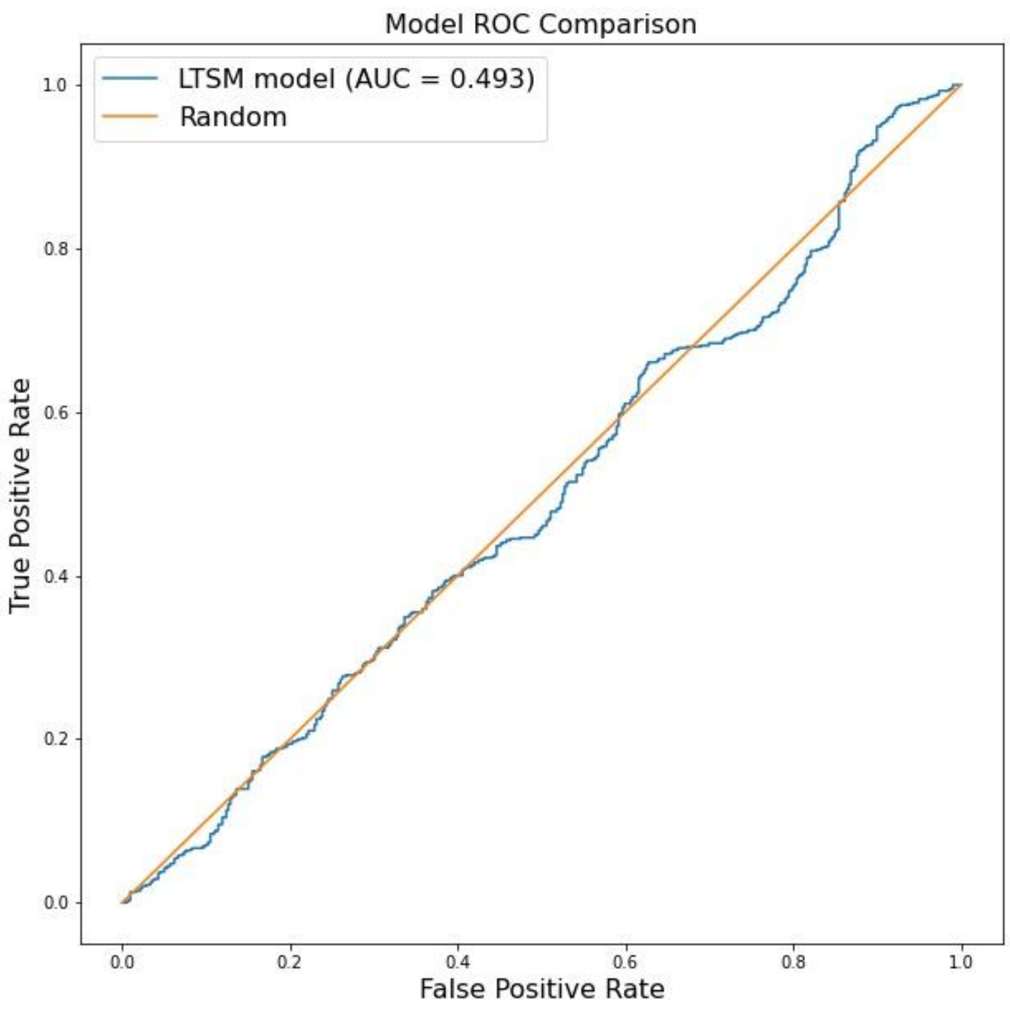
ISS A/S


Gain curve
Mean
Error
Std
No Gain
1.20
1.15
1.10
1.05
1.00
0.95
Relative gain
0.46
0.48
0.50
0.52
0.54
Threshold

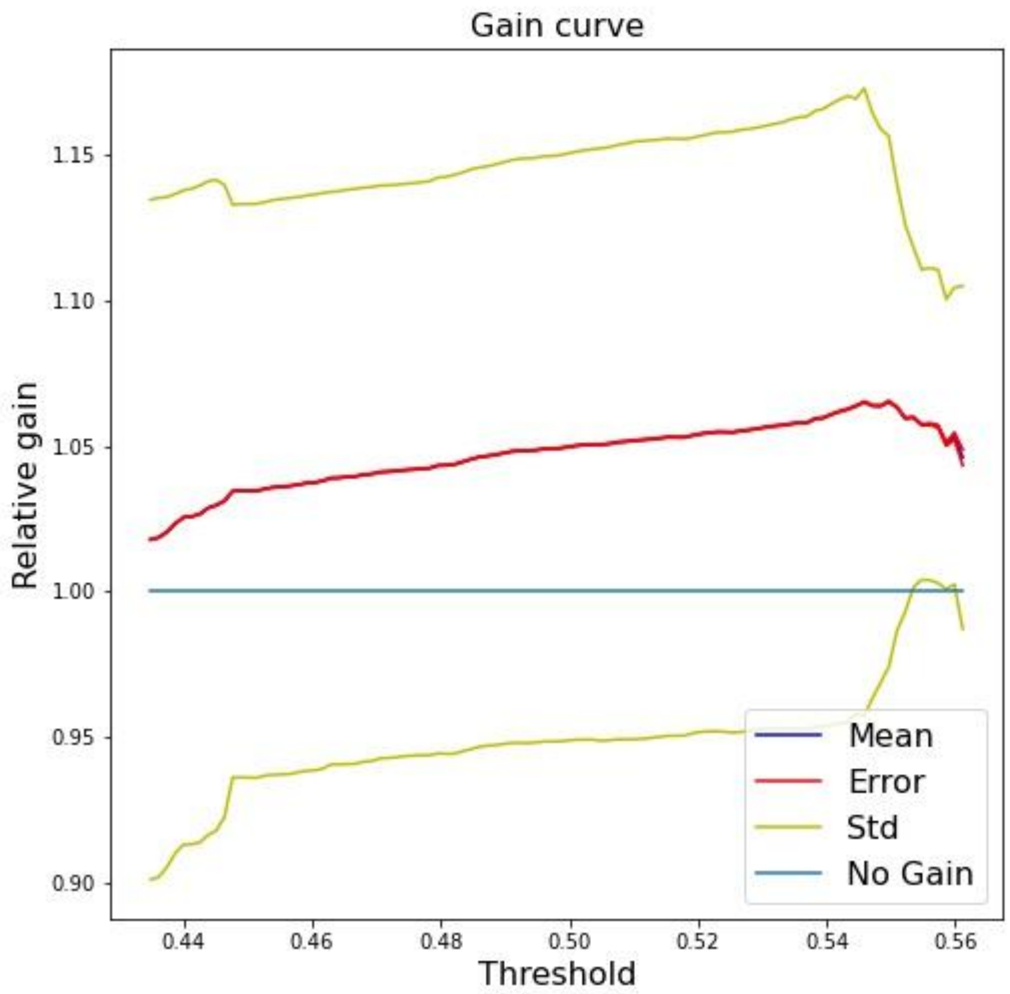

Gain curve
Mean
Error
Std
No Gain
1.15
1.10
1.05
1.00
0.95
0.90
Relative gain
0.46
0.48
0.50
0.52
0.54
0.56
Threshold

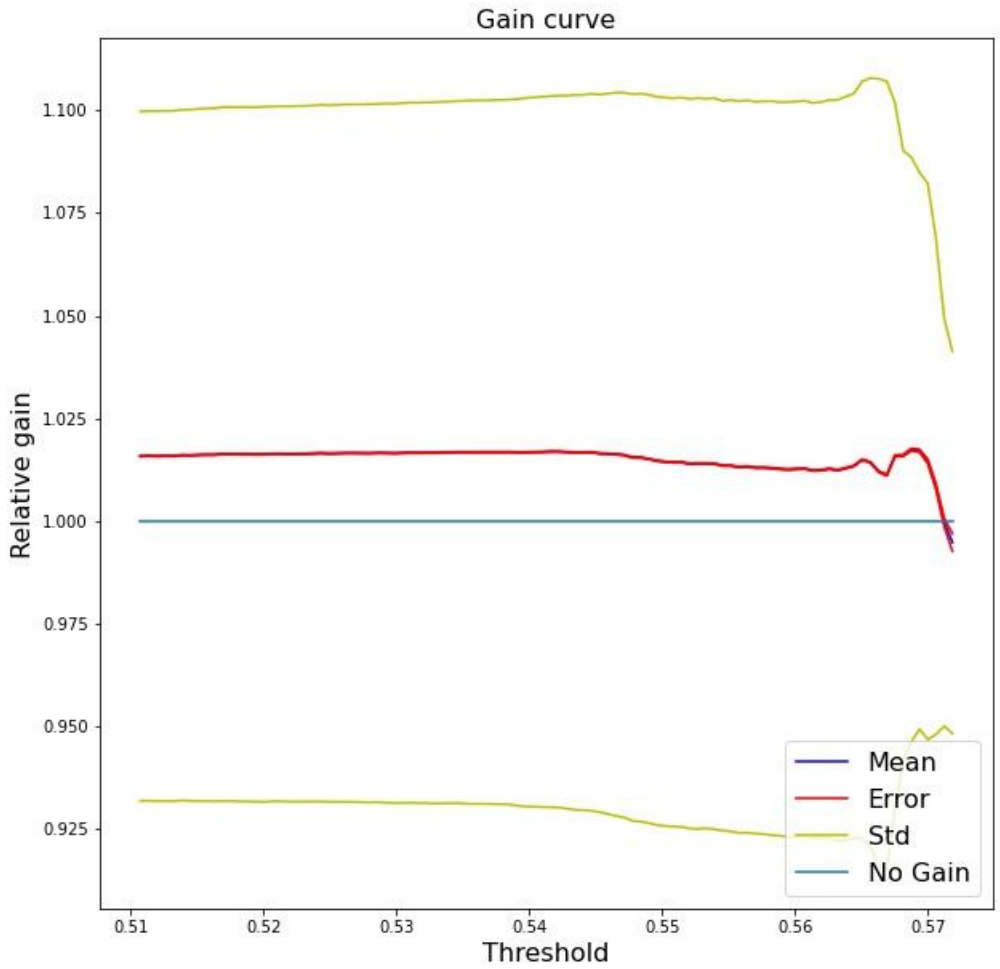
Royal Unibrew A/S



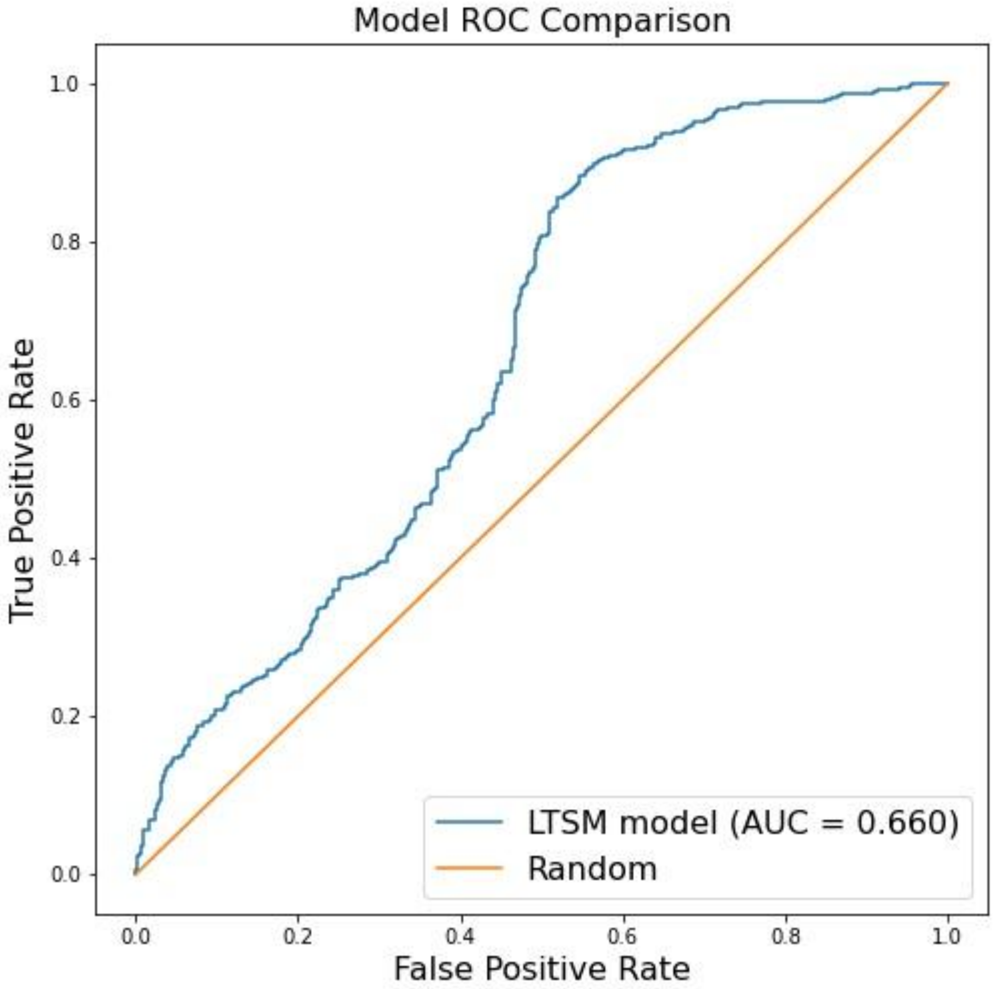


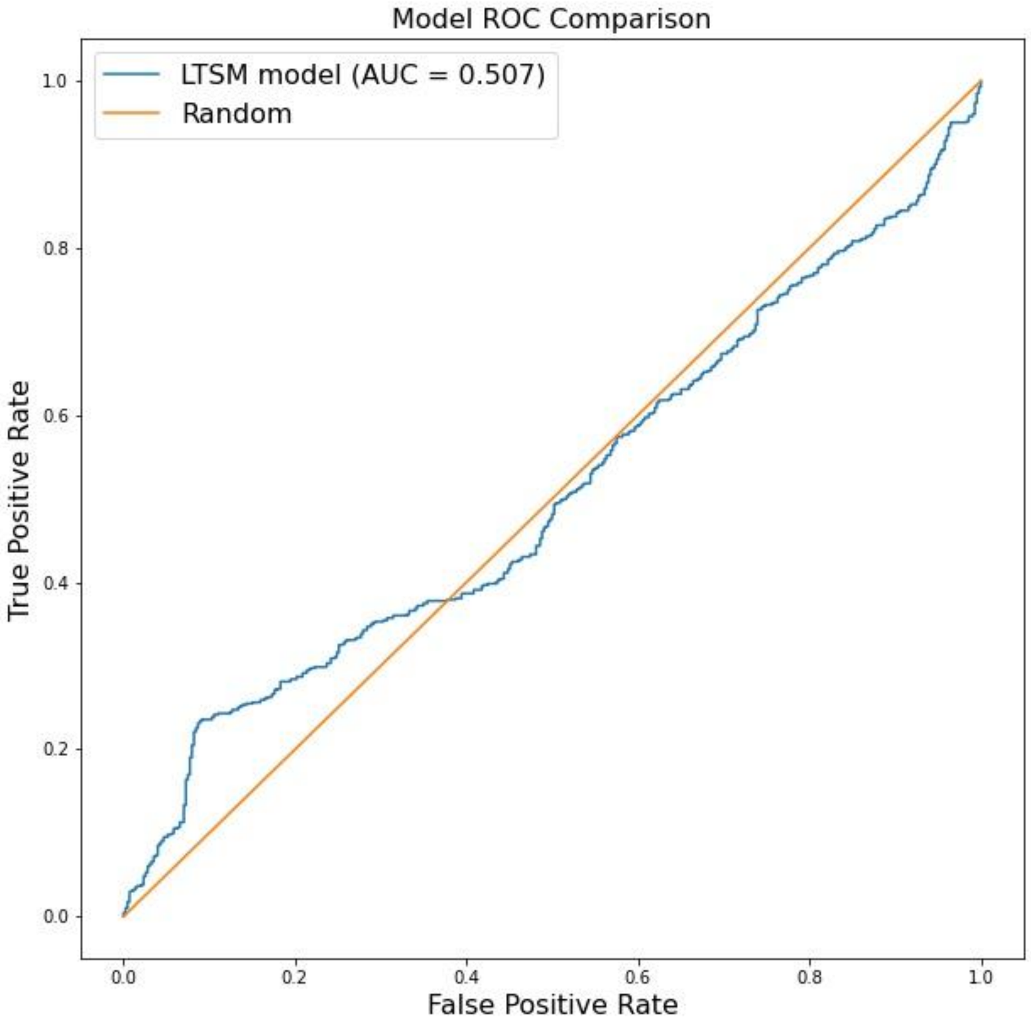
Royal Unibrew A/S



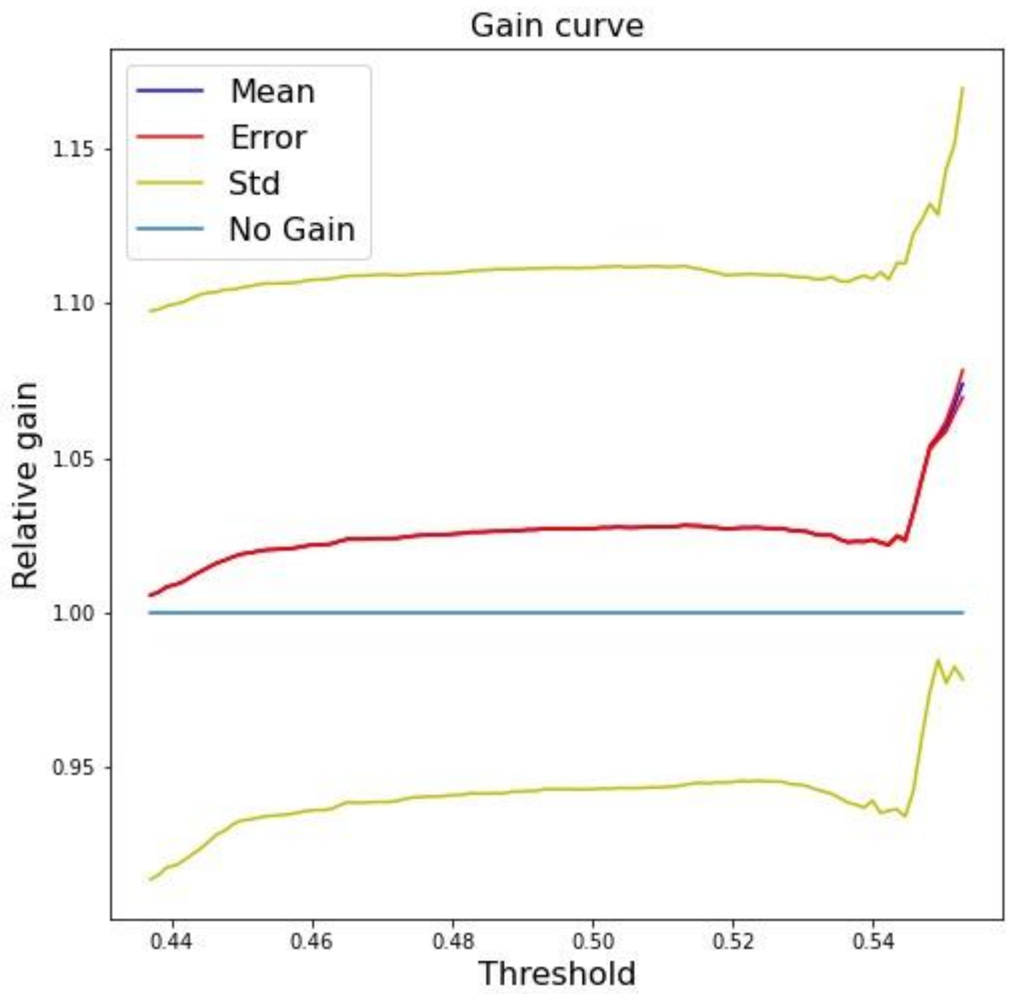


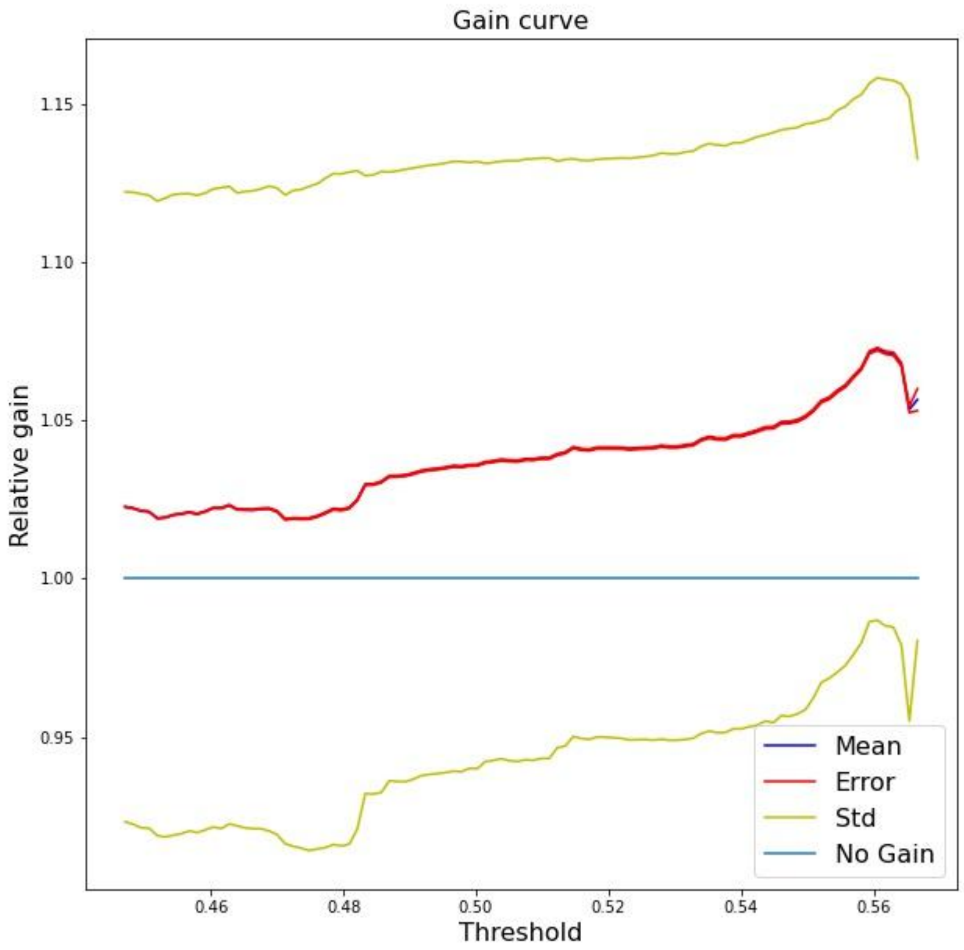
A.P. Møller - Mærsk A/S



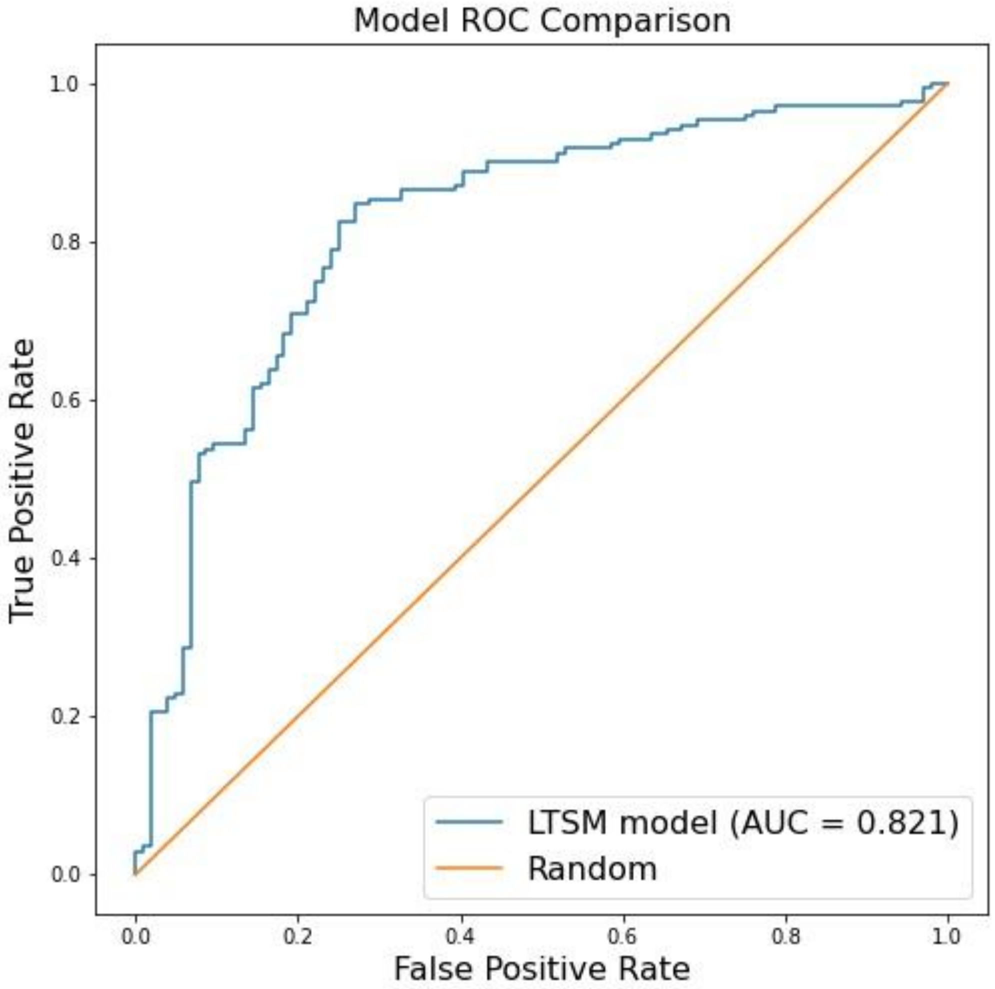


A.P. Møller - Mærsk A/S

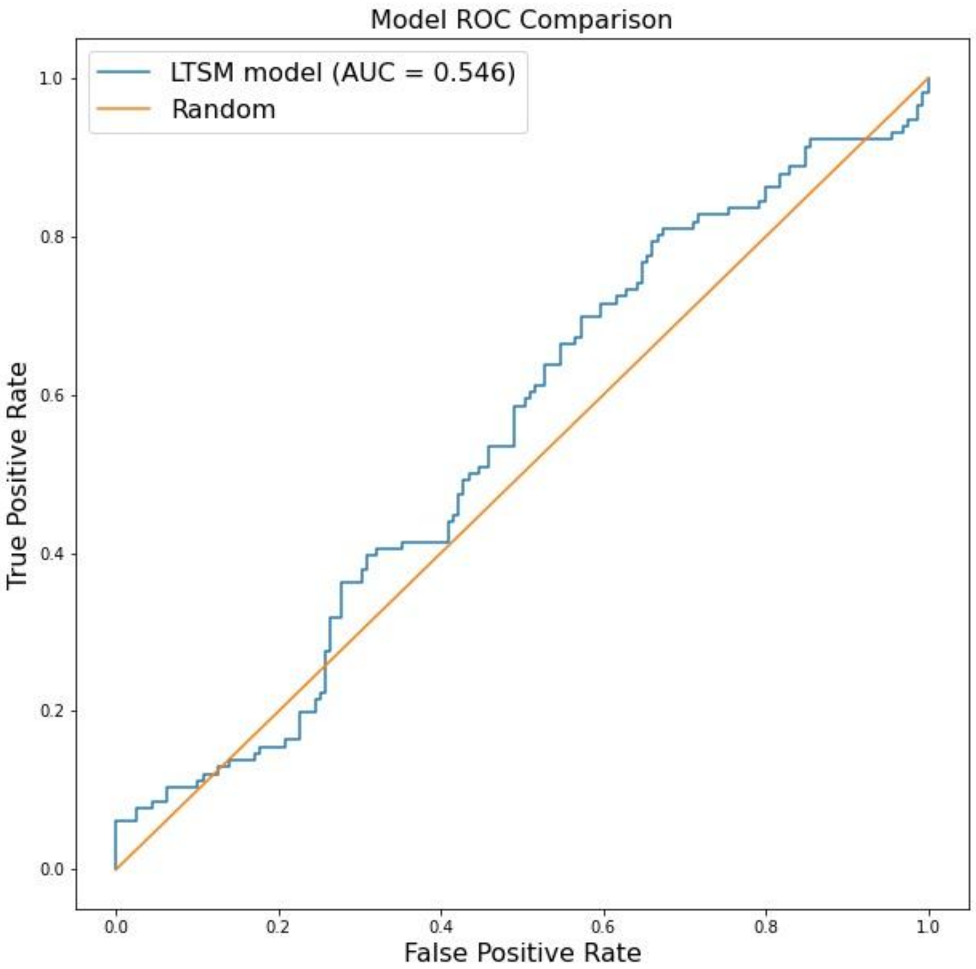




Ørsted A/S

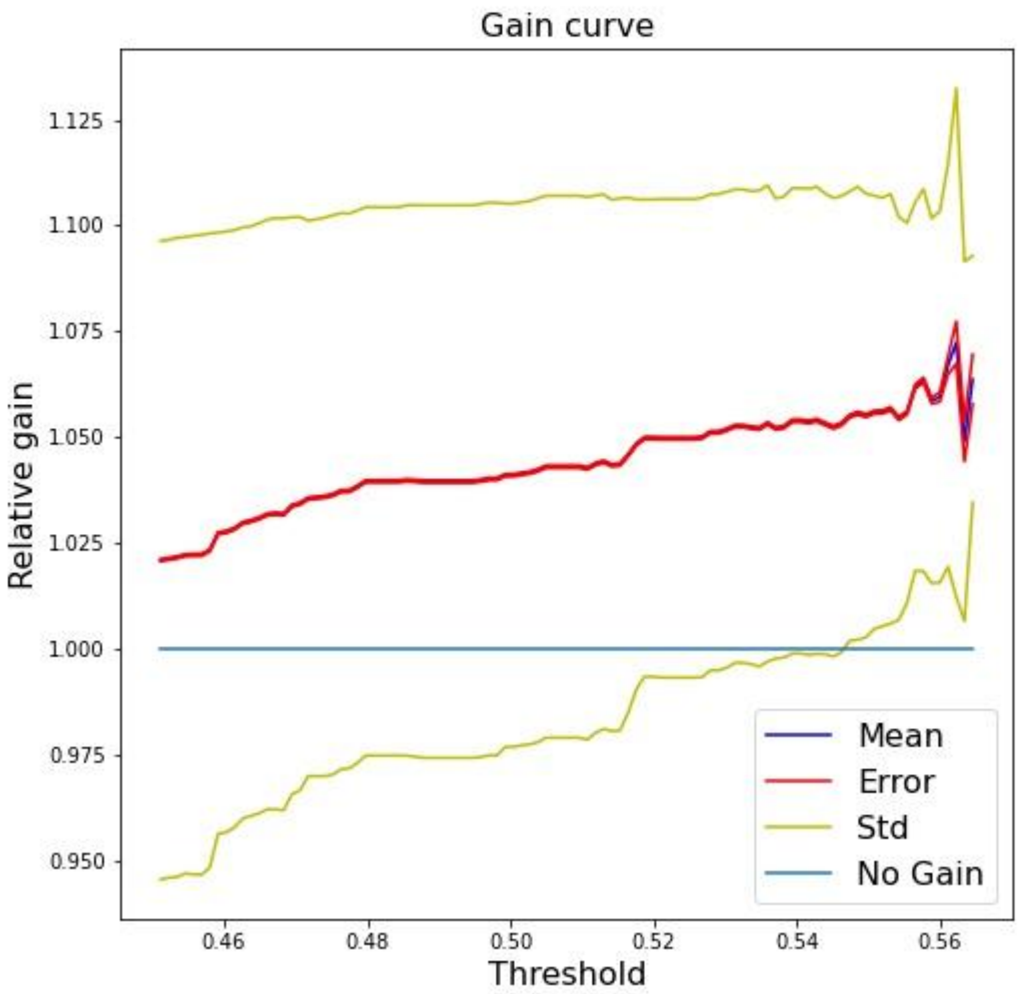


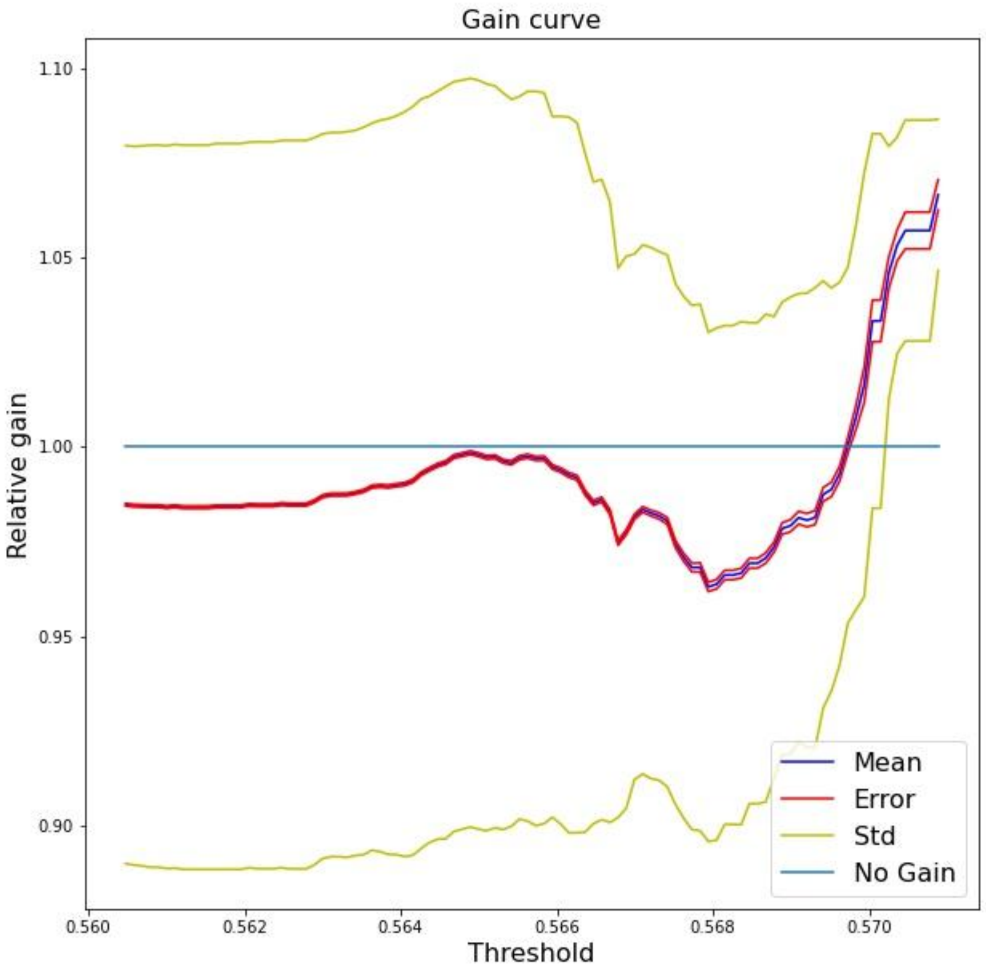
Model ROC Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LTSM model (AUC = 0.821)
Random



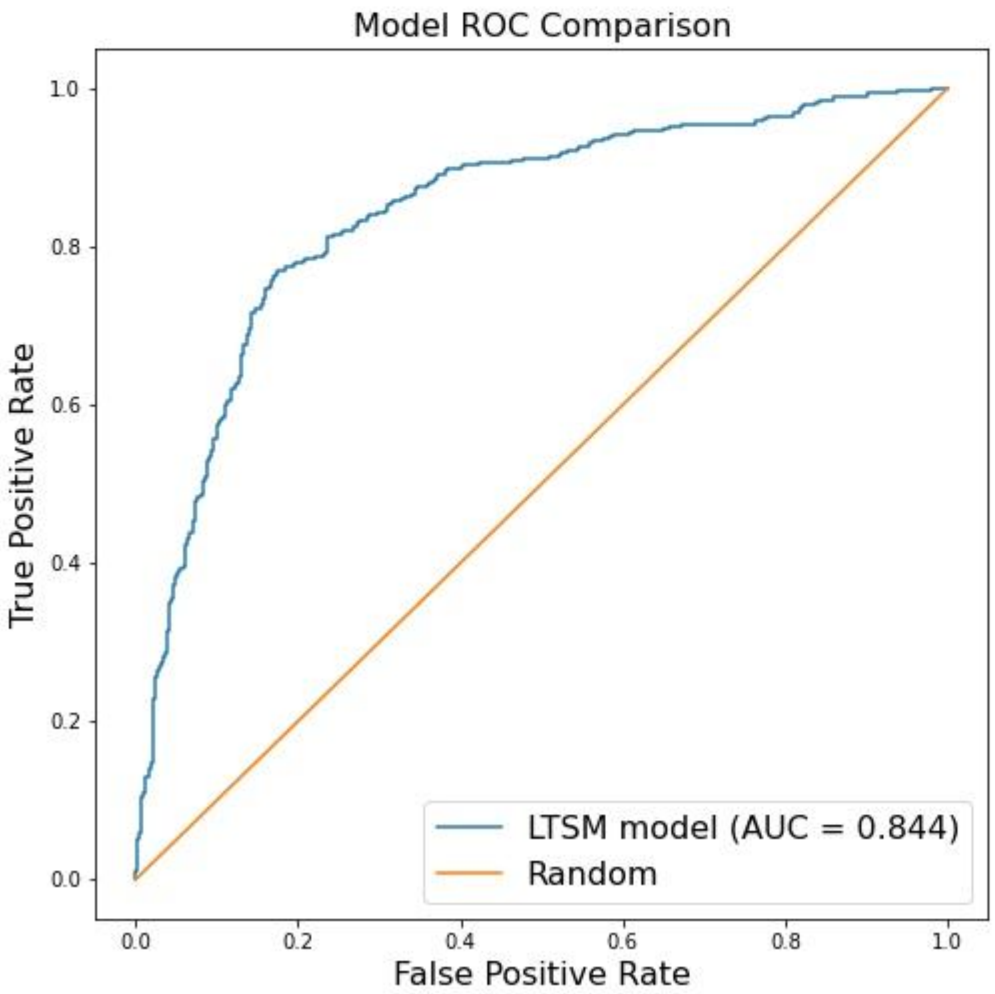
Model ROC Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LTSM model (AUC = 0.546)
Random

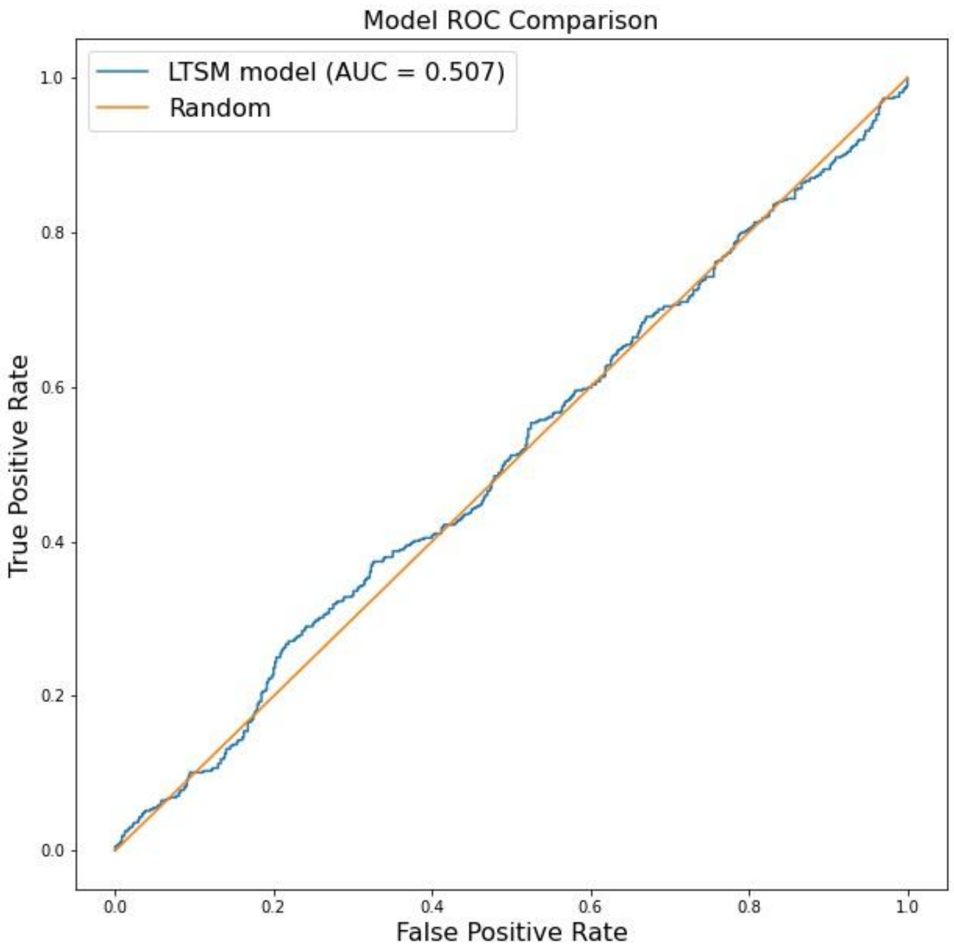
Ørsted A/S


Gain curve
1.125
1.100
1.075
1.050
1.025
1.000
0.975
0.950
Relative gain
0.46
0.48
0.50
0.52
0.54
0.56
Threshold
Mean
Error
Std
No Gain

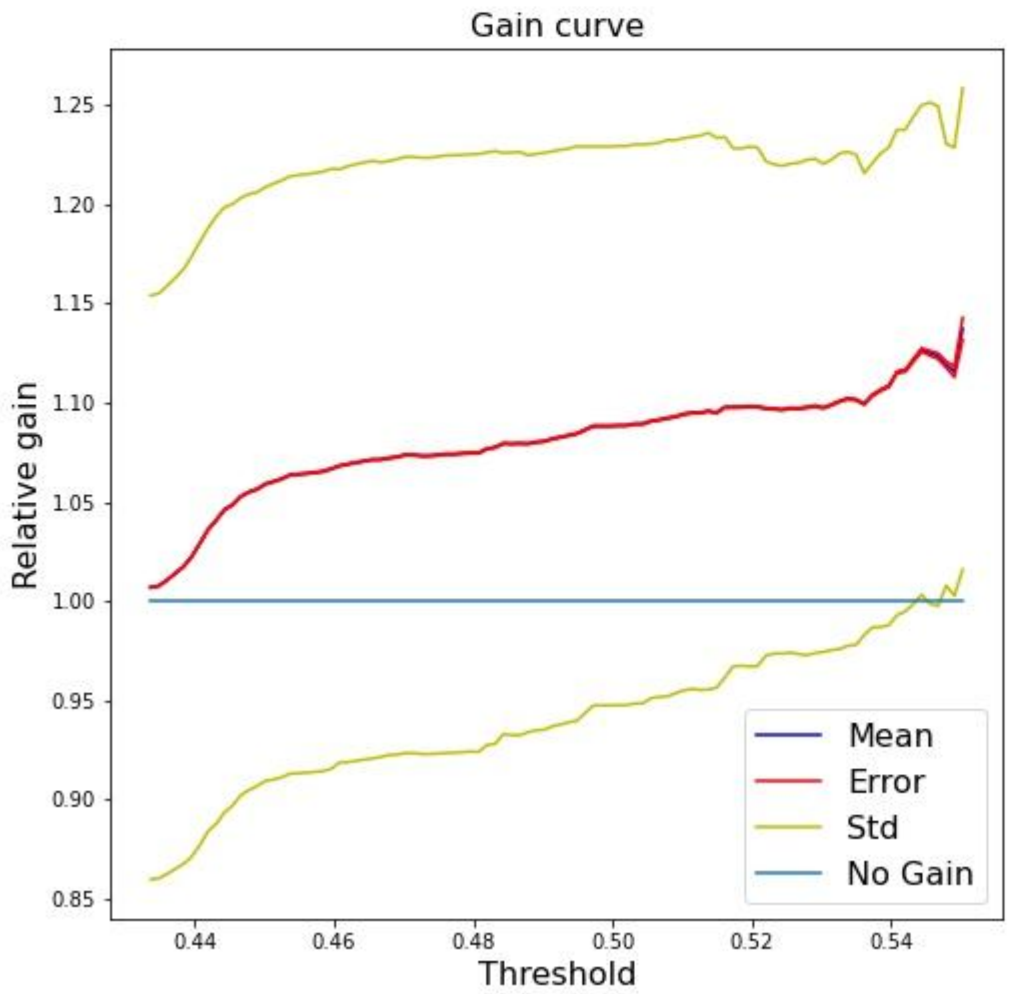

Gain curve
1.10
1.05
1.00
0.95
0.90
Relative gain
0.560
0.562
0.564
0.566
0.568
0.570
Threshold
Mean
Error
Std
No Gain

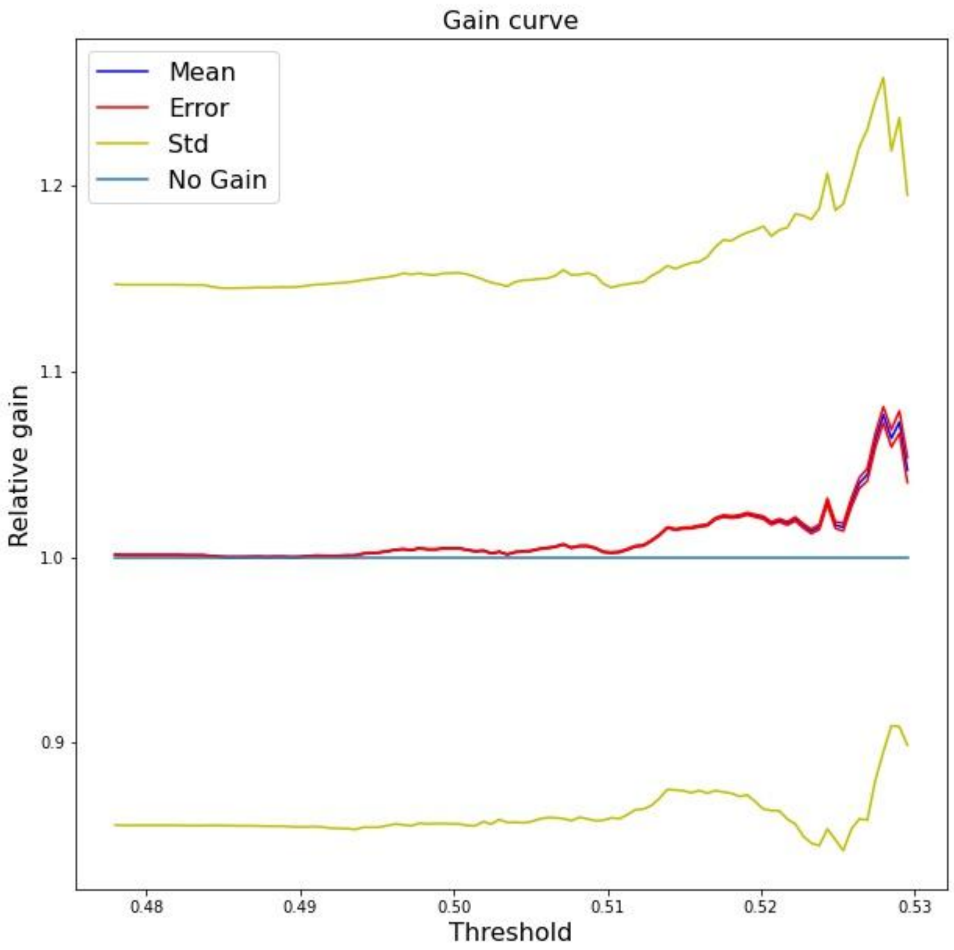
Bavarian Nordic A/S



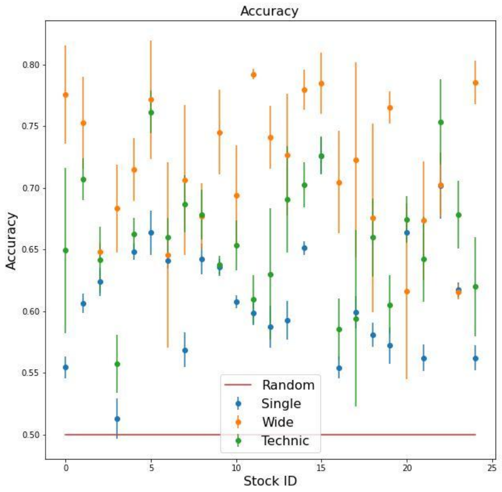


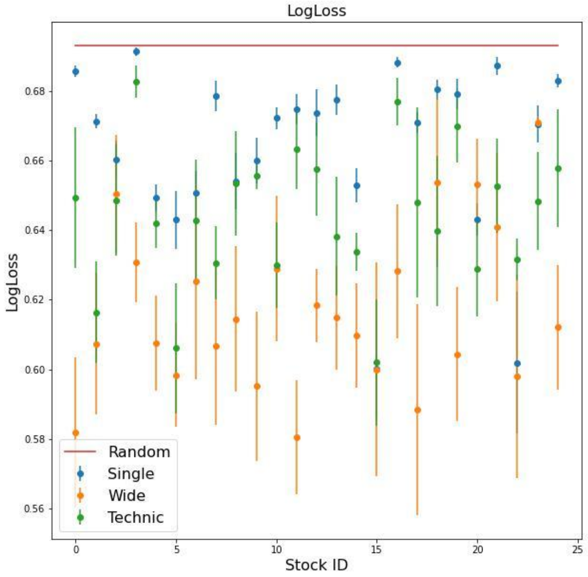
Bavarian Nordic A/S

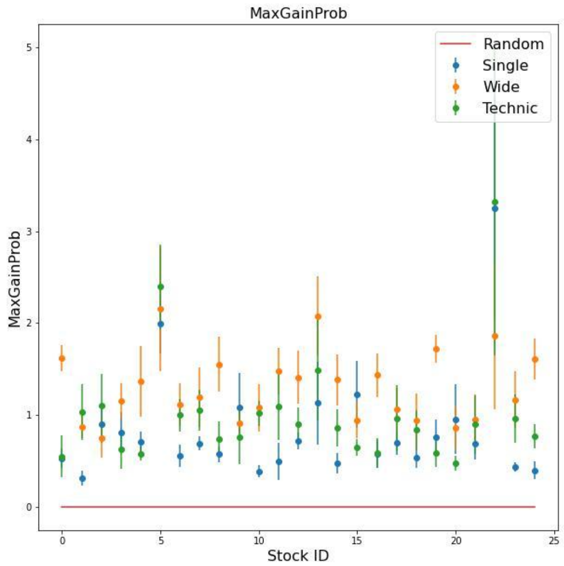


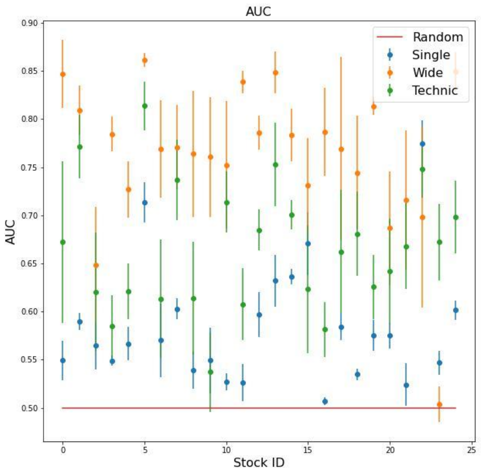


Shuffled LSTM evaluation



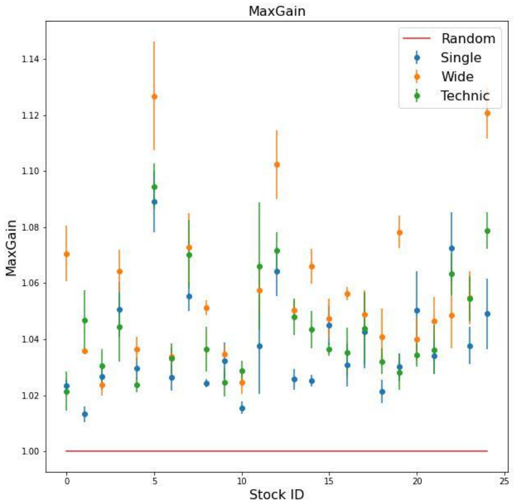




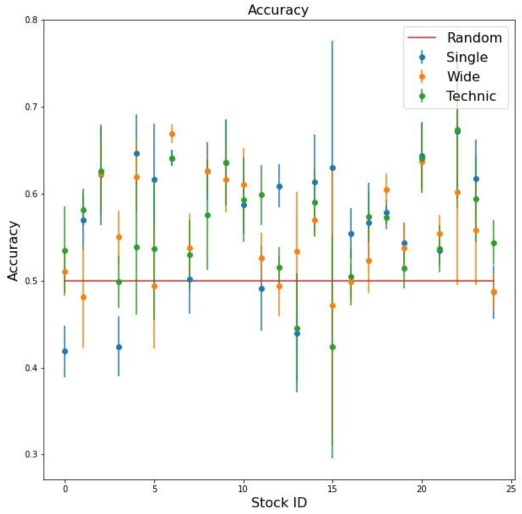


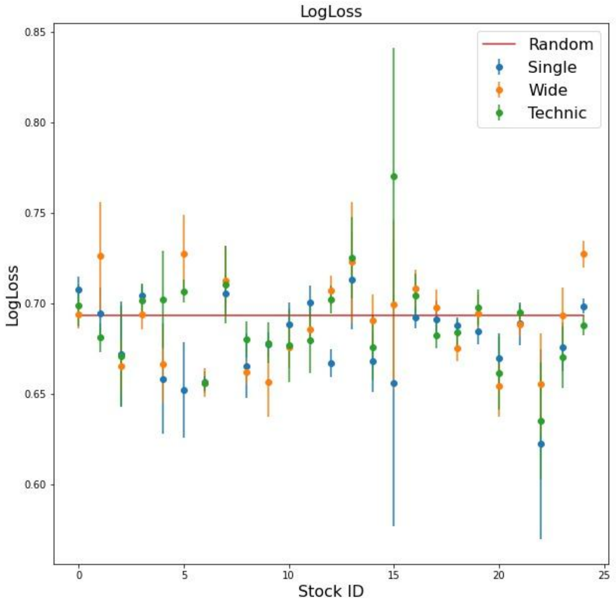
5-fold cross validation to get errors
Lines indicate a completely random model

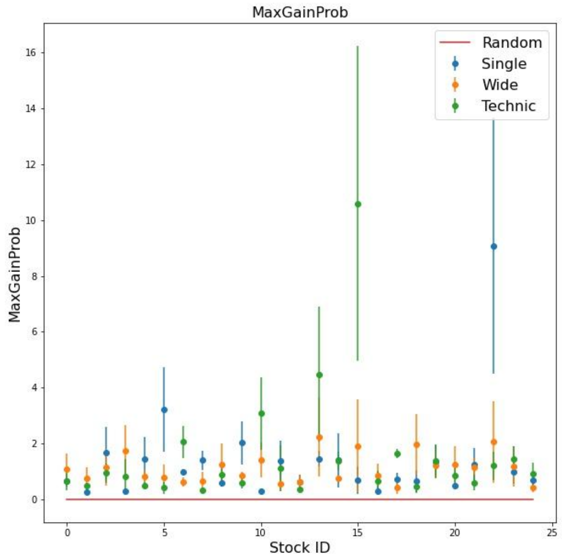
Wide most is best, followed by technical indicators and single mode is worst

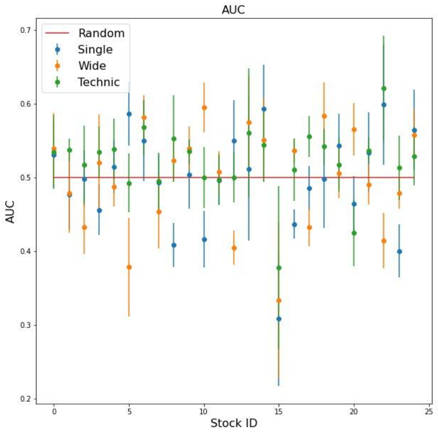


Ordered LSTM evaluation



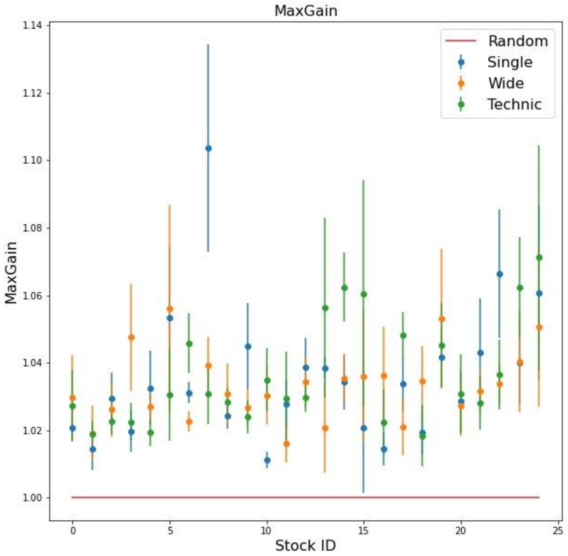






5-fold cross validation to get errors
Lines indicate a completely random model

Wide most is best, followed by technical indicators and single mode is worst



Support vector machine