


Predicting Singapore index using machine learning



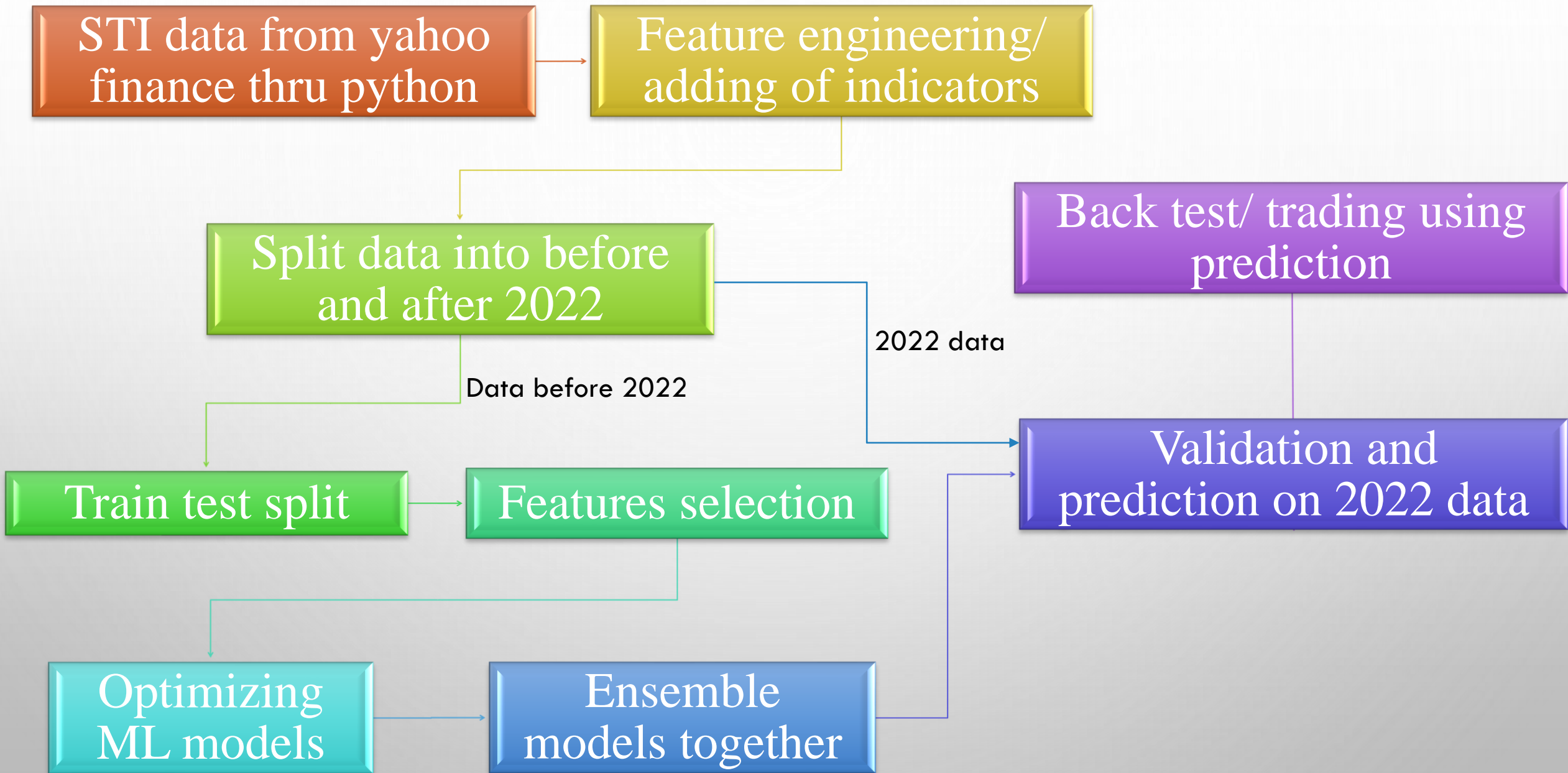
Problem statement

Financial institutes offering STI can only project returns of around 5% per annum, not sufficient to beat current inflation of 6.744%*

- Market has been volatile for recent years due to covid, war and inflations
- Using machine learning models to predict STI next movement and have better performance than financial institutes
- Market index is a measurement of a certain number of top companies' stocks and calculated into a number
- Example of market indices:
 - Russell
 - Standard & poor (S&P)
 - Nasdaq
 - Nikkei
 - Straits time index (STI)

 +35.1% 31-Dec-2019: 2.08 31-Dec-2020: 2.81	 +19.9% 31-Dec-2019: 16.20 31-Dec-2020: 19.43	 +15.5% 31-Dec-2019: 1.74 31-Dec-2020: 2.01	 +12.9% 31-Dec-2019: 4.12 31-Dec-2020: 4.65	 +11.2% 31-Dec-2019: 2.60 31-Dec-2020: 2.89
 +4.7% 31-Dec-2019: 8.86 31-Dec-2020: 9.28	 +0.7% 31-Dec-2019: 55.60 31-Dec-2020: 55.88	 +0.3% 31-Dec-2019: 2.97 31-Dec-2020: 2.98	 -3.0% 31-Dec-2019: 3.94 31-Dec-2020: 3.82	 -3.2% 31-Dec-2019: 25.88 31-Dec-2020: 25.04
 -7.3% 31-Dec-2019: 8.32 31-Dec-2020: 7.71	 -7.6% 31-Dec-2019: 0.92 31-Dec-2020: 0.85	 -8.4% 31-Dec-2019: 10.98 31-Dec-2020: 10.06	 -10.9% 31-Dec-2019: 2.39 31-Dec-2020: 2.13	 -12.2% 31-Dec-2019: 2.46 31-Dec-2020: 2.16
 -12.5% 31-Dec-2019: 3.75 31-Dec-2020: 3.28	 -14.5% 31-Dec-2019: 26.41 31-Dec-2020: 22.88	 -14.7% 31-Dec-2019: 1.12 31-Dec-2020: 0.96	 -17.4% 31-Dec-2019: 0.89 31-Dec-2020: 0.74	 -18.8% 31-Dec-2019: 30.65 31-Dec-2020: 24.88
 -20.5% 31-Dec-2019: 6.77 31-Dec-2020: 5.38	 -21.3% 31-Dec-2019: 5.06 31-Dec-2020: 3.98	 -25.3% 31-Dec-2019: 2.29 31-Dec-2020: 1.71	 -27.0% 31-Dec-2019: 5.71 31-Dec-2020: 4.17	 -27.2% 31-Dec-2019: 10.95 31-Dec-2020: 7.97
 -28.2% 31-Dec-2019: 5.75 31-Dec-2020: 4.13	 -29.8% 31-Dec-2019: 2.38 31-Dec-2020: 1.67	 -31.5% 31-Dec-2019: 3.37 31-Dec-2020: 2.31	 -35.1% 31-Dec-2019: 30.10 31-Dec-2020: 19.57	 -52.7% 31-Dec-2019: 9.04 31-Dec-2020: 4.28

*From <https://www.rateinflation.com/inflation-rate/singapore-inflation-rate/>



Machine learning to predict STI

- Combining technical analysis with supervised classification machine learning models
 - In finance, technical analysis is an analysis methodology for analyzing and forecasting the direction of prices through the study of past market data
 - Common indicators used for technical analysis
 - ✓ Exponential moving averages (EMA)
 - ✓ Moving average convergence divergence (MACD)
 - ✓ Stochastic oscillator
 - ✓ Force index



Supervised machine learning classification

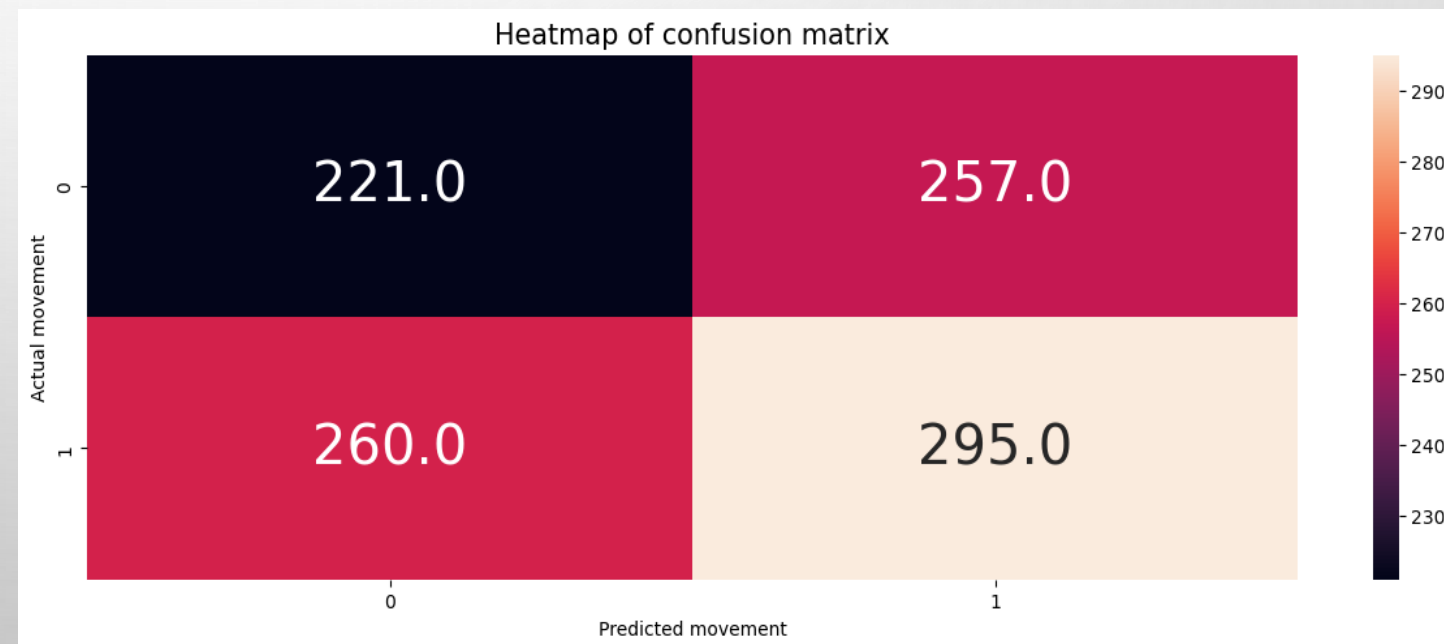
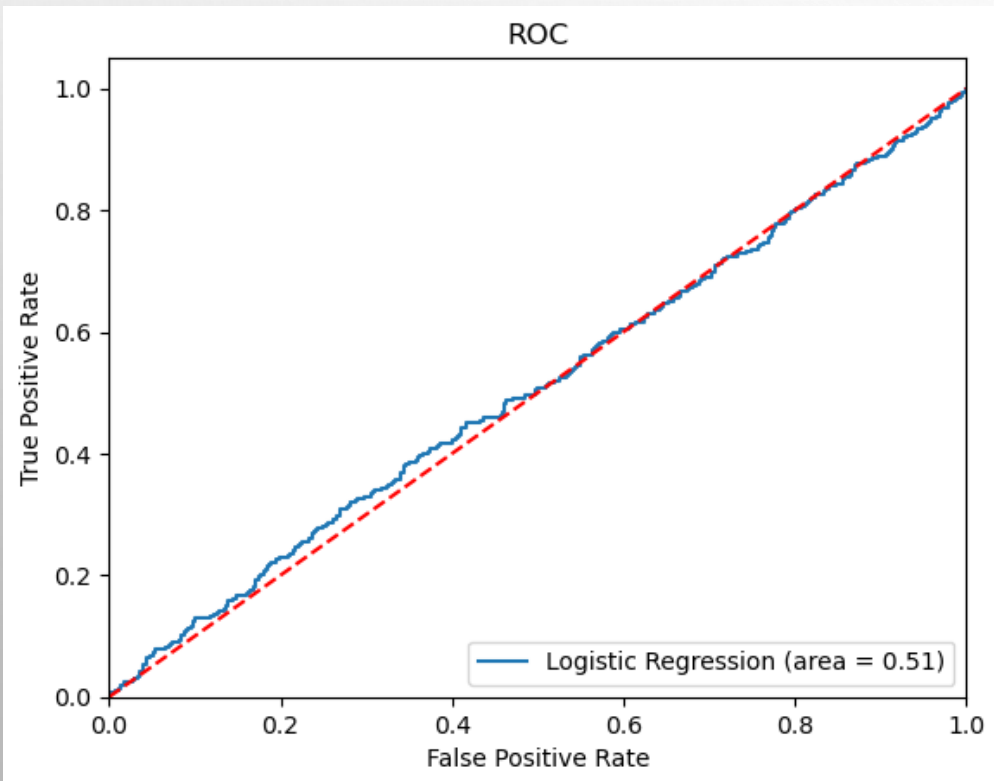
- Using feature engineering to convert indicators into numbers, supervised classification can be done to predict next day market closing price movement
 - If closing price for tomorrow is above today price, will be classify as “1” and if lower will be classify as “0”
 - Daily change in the indicators are added as features

Date	Open	High	Low	Close	Next_day_movement	ema_period_2	ema_period_5	ema_period_5_diff	force_index_5	force_index_5_diff
2022-11-11	3213.350098	3238.070068	3211.199951	3228.330078	1.0	3208.238390	3177.223570	25.553254	1.042484e+10	7.773506e+09
2022-11-14	3234.510010	3279.100098	3226.830078	3260.800049	1.0	3243.279496	3205.082396	27.858826	1.154840e+10	1.123564e+09
2022-11-15	3271.350098	3285.719971	3261.959961	3275.280029	0.0	3264.613185	3228.481607	23.399211	9.181710e+09	-2.366695e+09
2022-11-16	3274.850098	3282.909912	3258.489990	3266.169922	1.0	3265.651010	3241.044379	12.562772	5.087568e+09	-4.094142e+09
2022-11-17	3274.669922	3298.100098	3263.679932	3286.040039	0.0	3279.243696	3256.042932	14.998553	5.220141e+09	1.325730e+08
2022-11-18	3289.330078	3308.300049	3265.530029	3272.229980	0.0	3274.567886	3261.438615	5.395683	2.285863e+09	-2.934278e+09
2022-11-21	3264.649902	3281.600098	3237.350098	3250.620117	1.0	3258.602707	3257.832449	-3.606166	-3.231694e+08	-2.609032e+09
2022-11-22	3261.310059	3277.350098	3259.560059	3259.560059	0.0	3259.240941	3258.408319	0.575870	3.841184e+08	7.072877e+08
2022-11-23	3278.750000	3285.479980	3249.520020	3255.989990	0.0	3257.073641	3257.602209	-0.806110	3.776109e+06	-3.803423e+08
2022-11-24	3268.679932	3273.429932	3250.149902	3252.879883	0.0	3254.277802	3256.028100	-1.574109	-2.051484e+08	-2.089245e+08
2022-11-25	3249.530029	3254.409912	3233.189941	3244.550049	0.0	3247.792633	3252.202083	-3.826017	-5.672639e+08	-3.621155e+08
2022-11-28	3241.689941	3248.689941	3221.889893	3240.060059	1.0	3242.637583	3248.154742	-4.047342	-7.287961e+08	-1.615322e+08
2022-11-29	3235.340088	3277.010010	3232.379883	3276.360107	1.0	3265.119266	3257.556530	9.401789	2.680756e+09	3.409552e+09
2022-11-30	3279.860107	3290.560059	3270.030029	3290.489990	1.0	3282.033082	3268.534350	10.977820	3.956572e+09	1.275816e+09
2022-12-01	3306.489990	3313.800049	3288.459961	3292.729980	0.0	3289.164348	3276.599560	8.065210	2.845439e+09	-1.111133e+09

Problem with using closing price as classification

- Logistic regression model was used as a baseline model
- Results from logistic regression model shows that model was unable to learn any patterns
- From AUC, even by increasing the threshold the model will not improve further

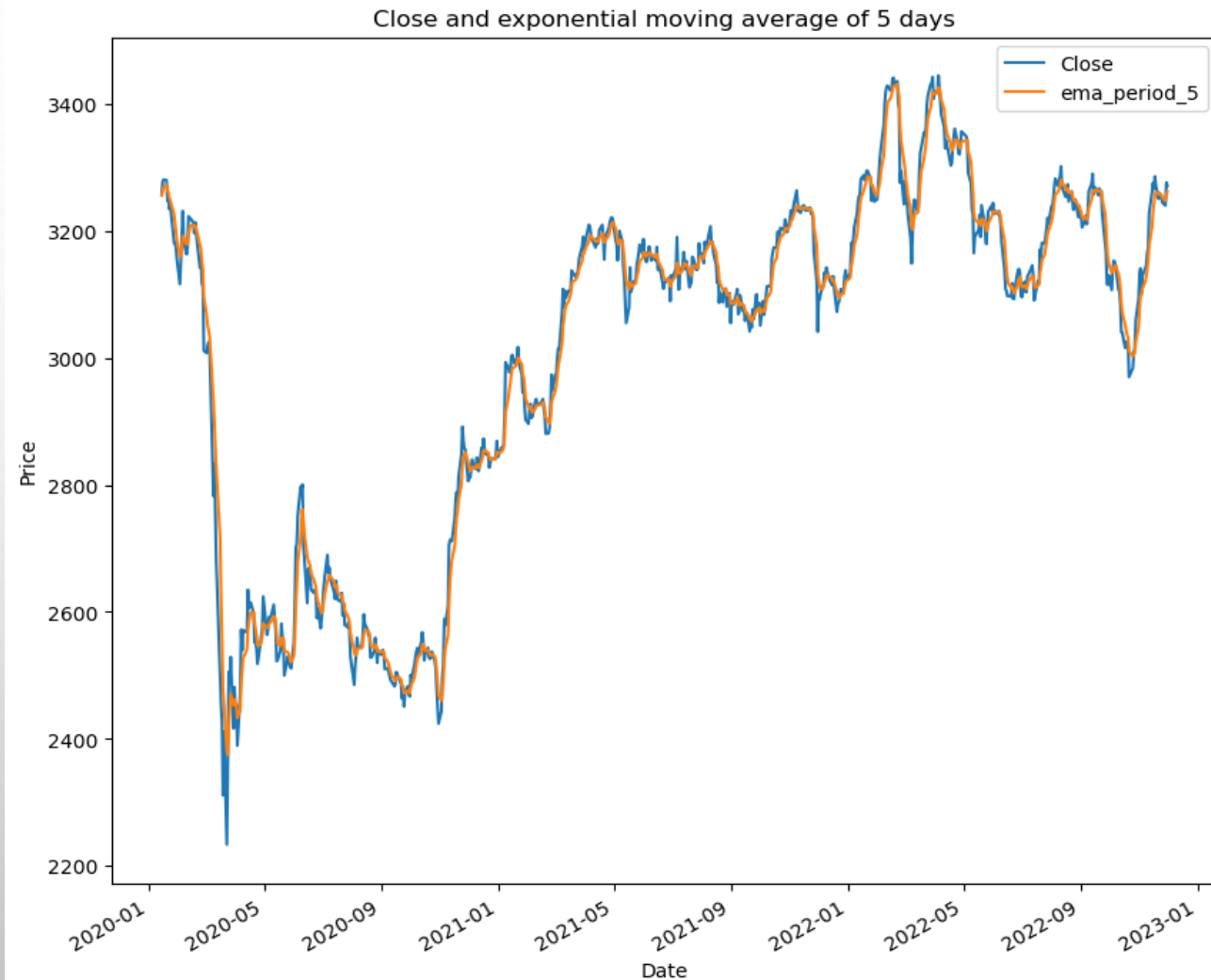
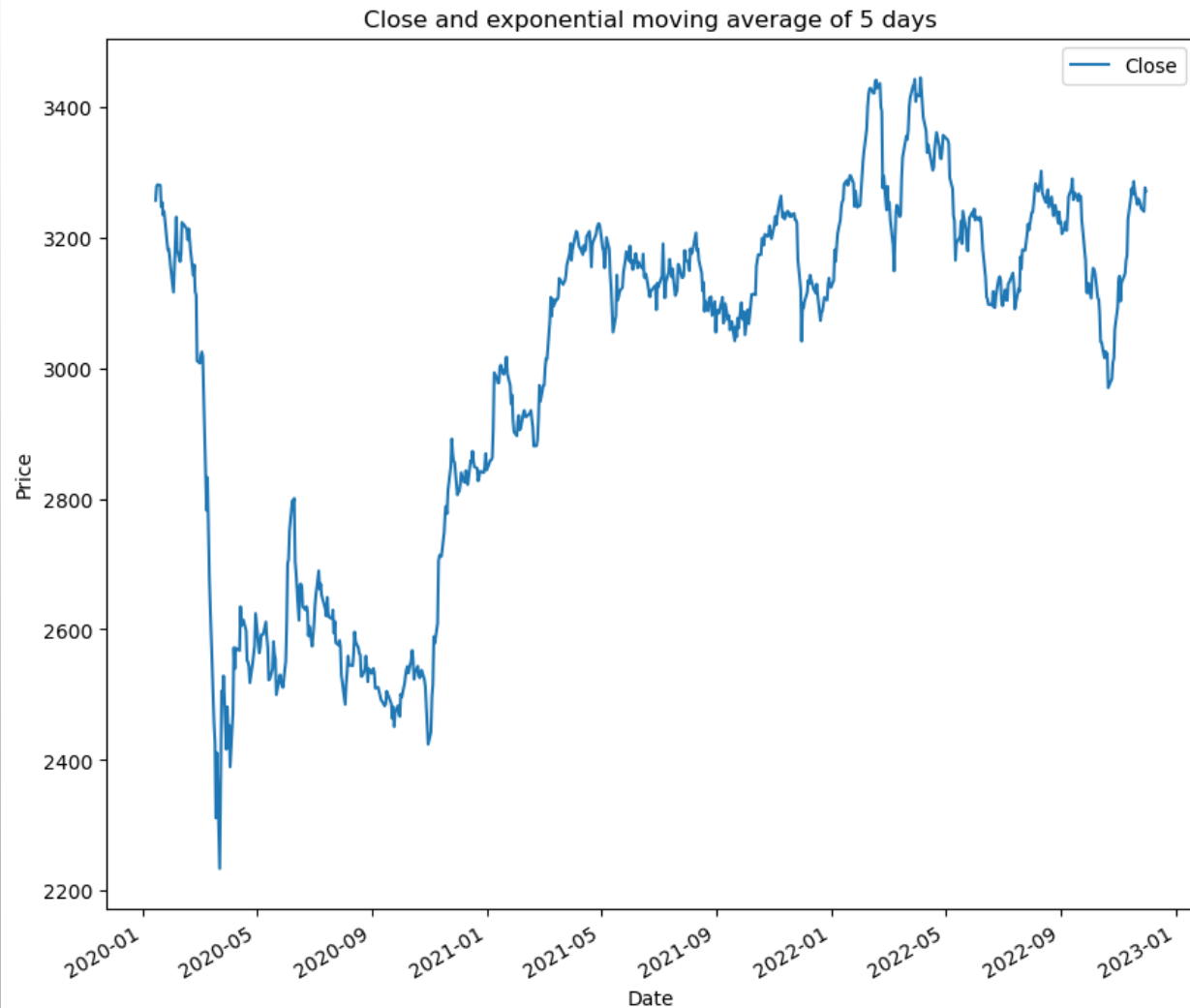
Metrics used	Score
Accuracy	0.4995
Precision	0.5344
Log loss	0.6954
AUC	0.51



Using exponential moving average as classification

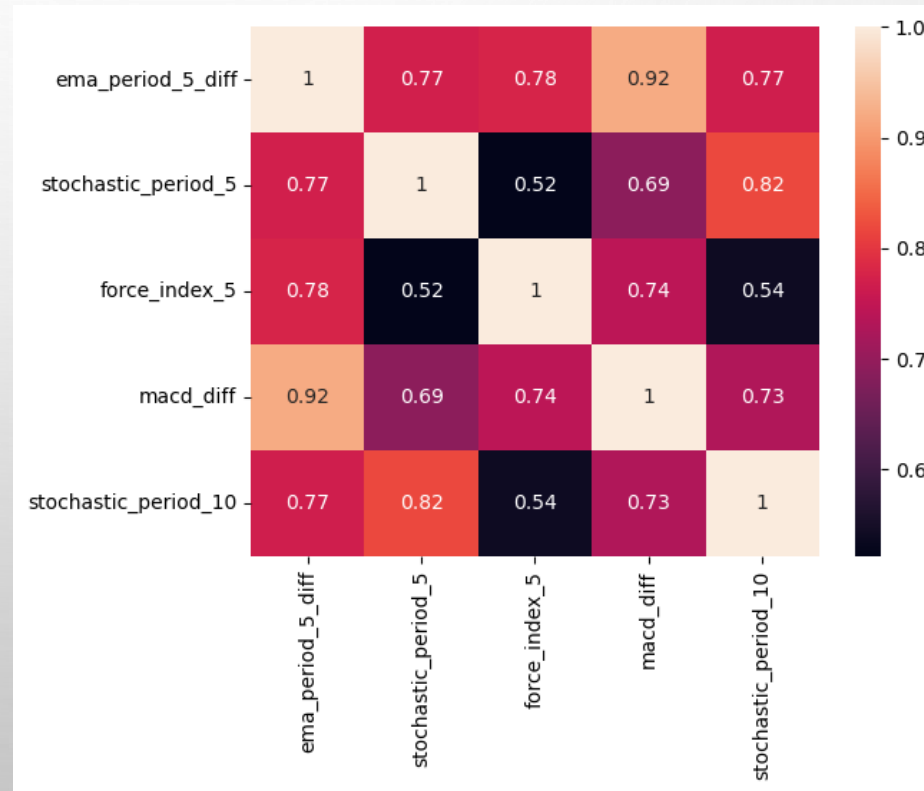
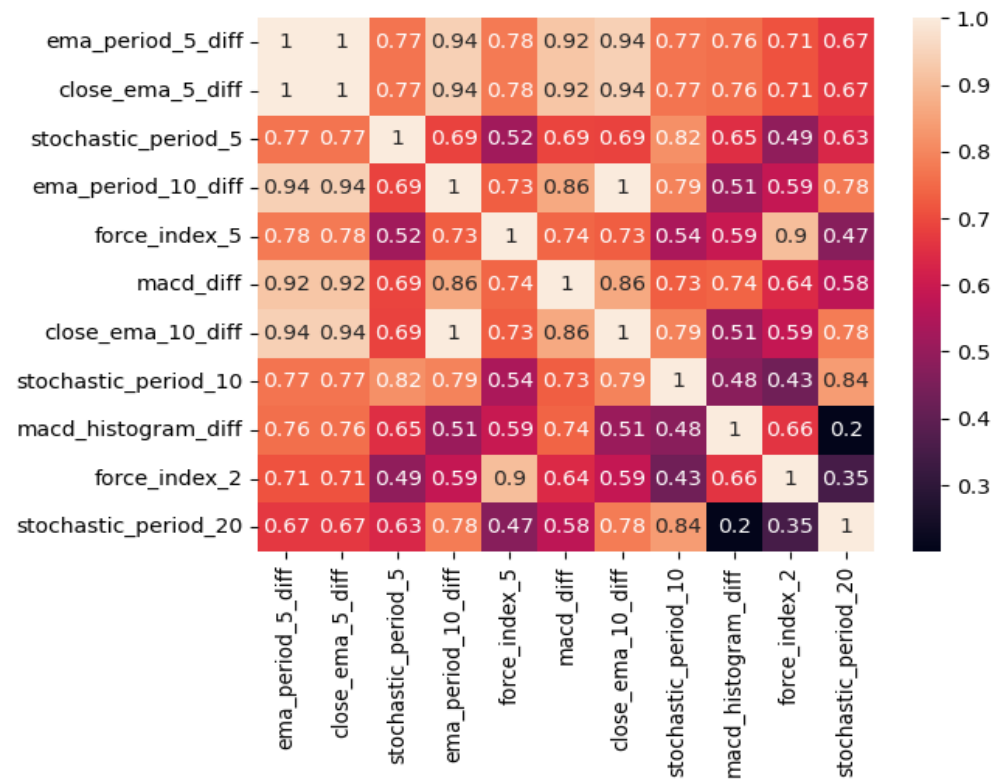
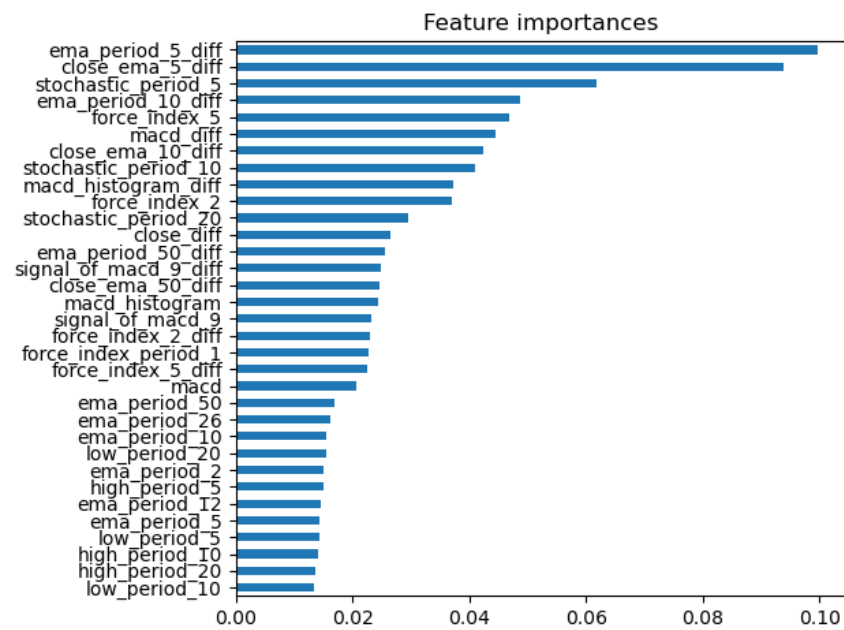
- Closing price has too much volatility and inconsistency
- Price need to be smooth out to be have a better prediction
- EMA of 5 days is used as a proxy to predict closing price movement

$$EMA_{Today} = (Value_{Today} \times (\frac{Smoothing}{1 + Days})) + EMA_{Yesterday} \times (1 - (\frac{Smoothing}{1 + Days}))$$



Features selection

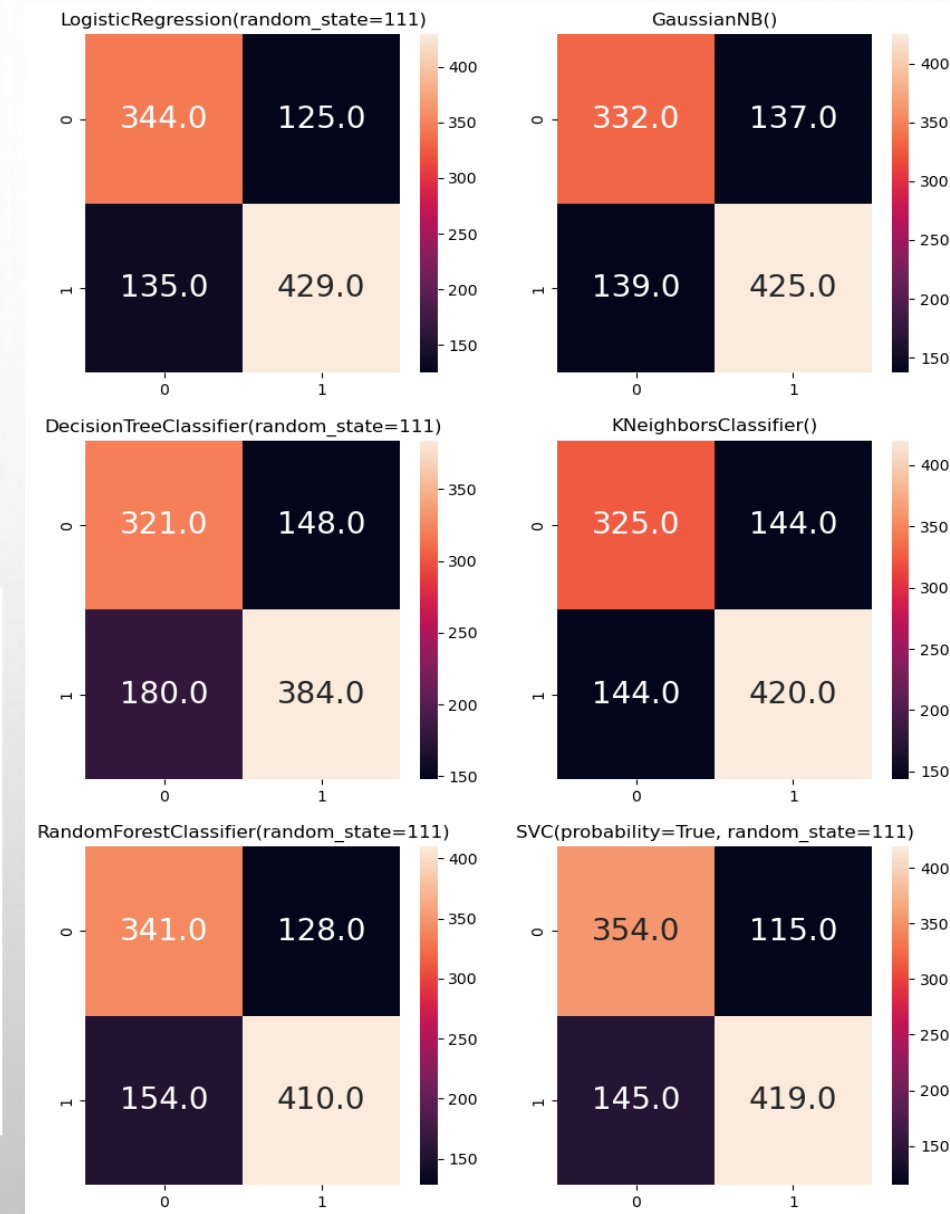
- Feature importance from random forest classifier to choose features machine learning
- Top 11 features from feature importance are chosen to perform correlation to further filter out features
- At least one feature from each indicator is retained
- 5 features used to test on classification models



Confusion matrix and results of different classification model

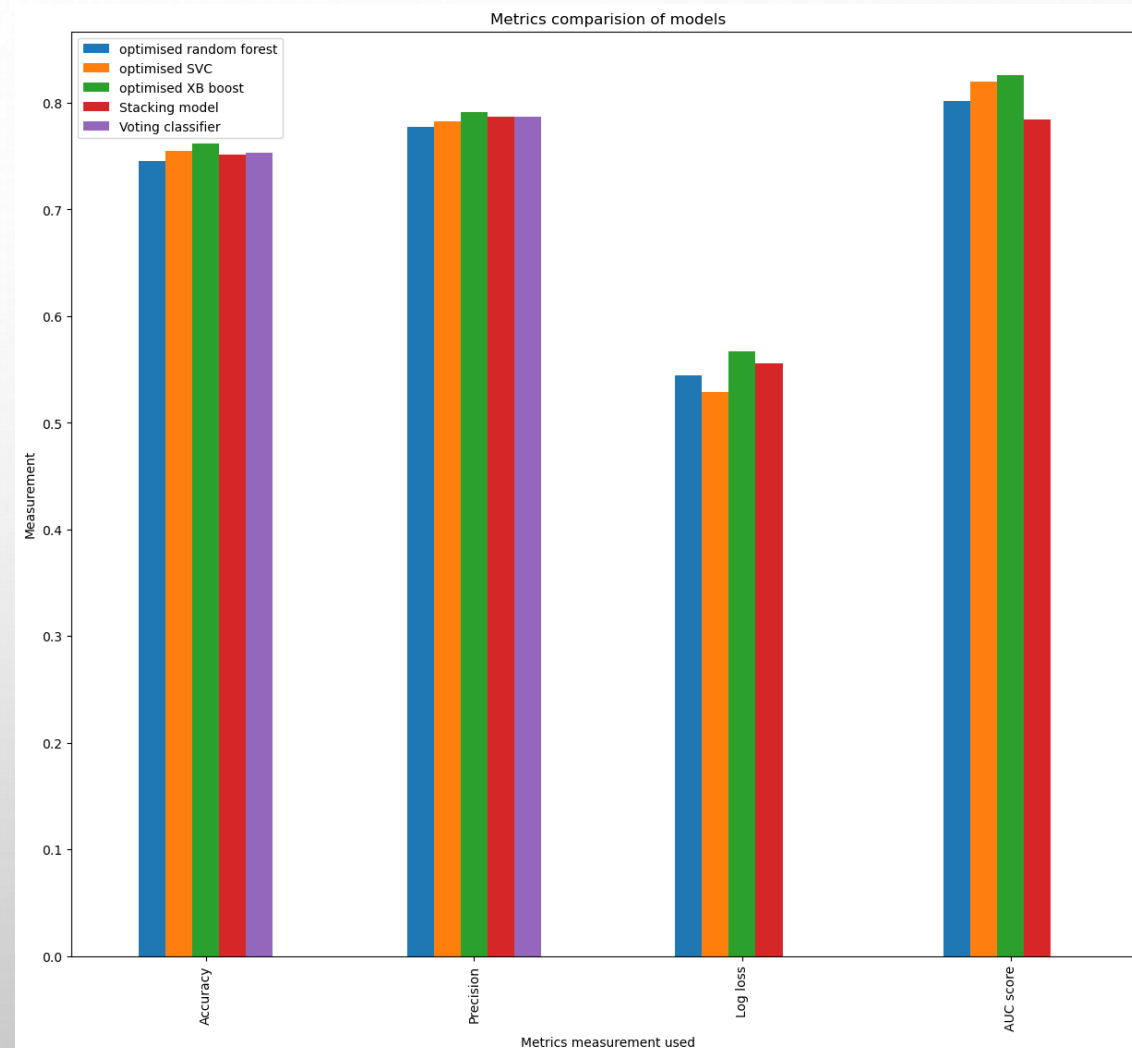
- Six different classification machine learning models (un-optimized) are tested on train data set
- Models are compared based on their accuracy, precision, log loss and AUC scores
- Logistic regression model being used as a baseline
- Random forest and SVC are further optimized on their hyperparameter
- XG boost is also used as a replacement for decision tree

	Accuracy	Precision	Log loss	AUC
Logistic Regression	0.7483	0.7744	0.5082	0.8283
GaussianNB	0.7328	0.7562	0.9428	0.8199
Decision Tree Classifier	0.6825	0.7218	10.966	0.6826
K Neighbors Classifier	0.7212	0.7447	2.3260	0.7750
Random Forest Classifier	0.7270	0.7621	0.7252	0.8087
C-Support Vector Classification	0.7483	0.7846	0.5443	0.7969



Confusion matrix and results of optimized classification model

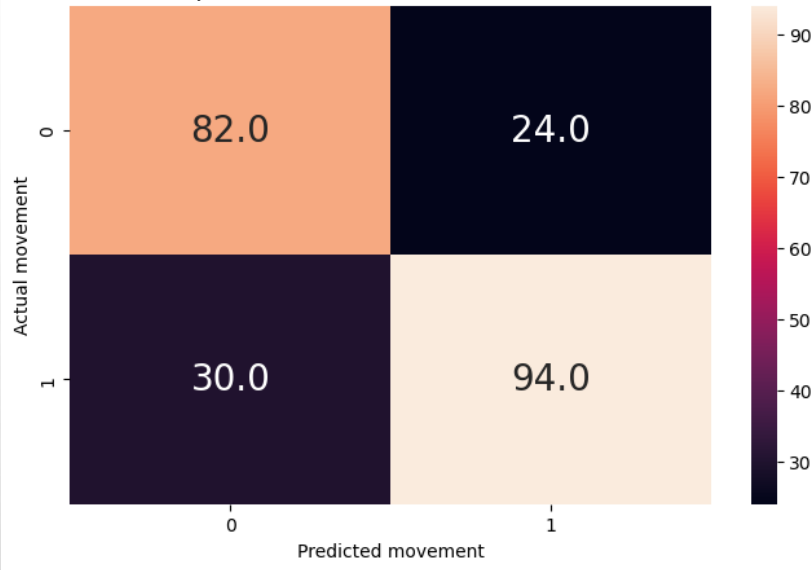
- Slight improvement on the scores of all three optimized models
- Optimized models with logistic regression are ensemble using ensemble models: stacking and voting models
- XB boost perform the best in terms of metric score
- Difference among the models are not significant to have XG boost as the final model
- Stacking model will be used as final model to avoid bias



	Optimized random forest	Optimized SVC	Optimized XB boost	Stacking model	Voting classifier
Accuracy	0.745402	0.755082	0.761859	0.751210	0.753146
Precision	0.777164	0.782214	0.791209	0.786916	0.786642
Log loss	0.5444	0.5288	0.5668	0.5557	NaN
AUC score	0.802020	0.819795	0.825716	0.784728	NaN

Validation and back test of final model using 2022 data

Heatmap of validation data confusion matrix



- 2022 data which is unseen by model yet, use to validate final model
- Result similar to test metrics results
- Proceed to perform back test with the following strategy and assumption
 - Start of with no position in market, will use the first prediction (0/1) to buy or sell
 - With the sell or buy, check next day prediction whether is it the same as currently position
 - Hold position if prediction and current position are the same
 - If next prediction is not in favor of current position, exit current position and take up the prediction direction
 - Assuming buy and sell price is exactly at closing price of the prediction appear

Metrics used	SVC score
Accuracy	0.7652
Precision	0.7966
Log loss	0.5357
AUC score	0.8058

Date	Close	Predict
2022-01-20	3294.820068	1.0
2022-01-21	3294.860107	1.0
2022-01-24	3283.350098	0.0
2022-01-25	3247.760010	0.0
2022-01-26	3271.570068	1.0
2022-01-27	3260.030029	0.0
2022-01-28	3246.330078	0.0
2022-01-31	3249.590088	0.0
2022-02-03	3315.989990	1.0
2022-02-04	3331.409912	1.0

- ← Up prediction , since starting with no position so buy at 3294.82
- ← Up prediction, hold position since already have a long position
- ← Down prediction, sell twice at 3283.35, one to close long position and one to have a short position
- ← Down prediction, hold position since already have a short position
- ← Up prediction, buy twice at 3271.57, one to close short position and one to have a long position
- ← Down prediction, sell twice at 3260.03, one to close long position and one to have a short position

Trading actions and results of final model predictions

- 50 number of trades done
- Profit from long position is \$450
- Profit from short position is \$308
- Total profit is \$758
- Returns of 23.4% based on 2022 average price of 3233.94

Backtesting of 2022 data





Thank you

Q & A

<https://github.com/MikoPoh/STI-prediction>