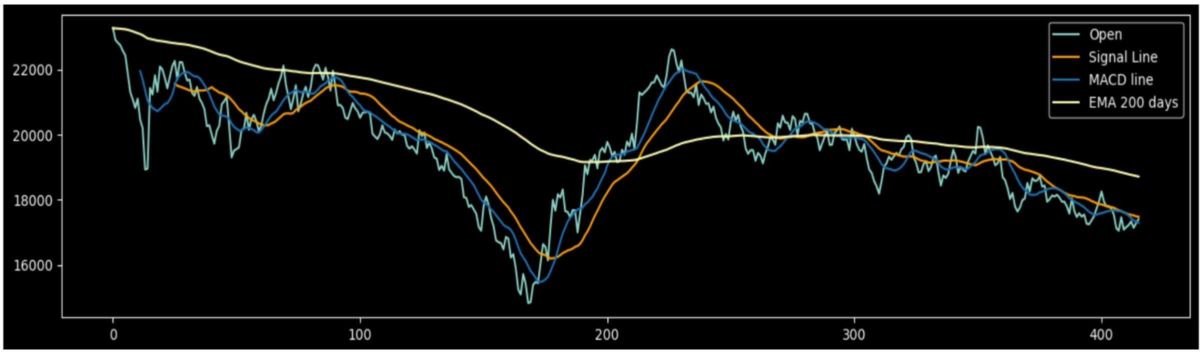
1. Introduction

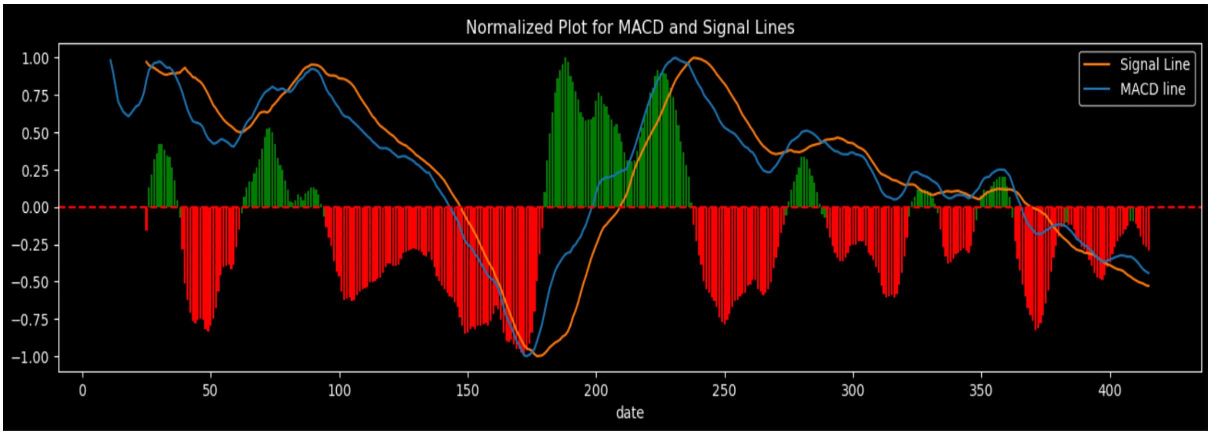
The Hang Seng Index (HSI) has served as an impactful indicator to indicate the Hong Kong stock market since the 1960s. It is a market capitalization-weighted index of the constituent stocks from different categories such as real estate and finance.

2. Elementary Data Analysis & Insights

We employed a data analysis of the given excel file (i.e. HSI.xlsx). This excel file contains the High, Low, Open, Close, Up votes, and Down votes regarding HSI. For the Up or Down votes, they simply refer to the prediction vote or sentiment from the social media users before the market opens. First and foremost, we imported the xlsx file into a Jupyter notebook and investigated the information of the dataset. We found out that there were missing data entries in both Up/Down votes and one incorrect datetime. To deal with the NaN in these columns, one alternative approach is to replace the NaN values by the mean. Besides, one minor problem of the Up/Down votes columns was that the sum of Up and Down votes was not equal to one. This minor problem was fixed subsequently. Since the HSI data is very noisy, it is better to obtain intuitions from the moving average. Therefore, we plotted the exponential moving average (EMA) in order to see the trend of data. Here, we plotted the 12 days (MACD line), 26 days (signal line), and 200 days moving average of HSI Open prices.



This plot is impactful in technical analysis as the cross of signal line and MACD line indicates the buy and sell time. If we are interested in using Moving Average Convergence Divergence (MACD) as our strategy, we should focus on the below plot which shows the difference between MACD and Signal line.



Here we normalize the data bounded by -1 and 1 for simplicity. We can see that there is an upward momentum when the MACD line crosses the Signal line and is higher than the Signal line subsequently. On the contrary, when the signal line crosses the MACD line in the lower region and the signal line is lower the MACD line later, it indicates downward momentum. Simply relying on the crossover between MACD and Signal lines is not enough since there are false signals, and we eliminate these false signals from our trading strategy if we insist on using MACD.

Besides, we can obtain some insights from the difference between MACD and signal lines, and we simply called this line the "Difference line". From the MACD plot, it suggests that the nature of the "Difference line" is mean-reverting, meaning that the "Difference line" fluctuates around some values. The mathematical description of a mean-reverting price series is that the change of the price series in the next moment is proportional to the difference between current and the mean price. Though we claim that the "Difference line" is mean-reverting, we did not carry out a rigorous proof of its nature. To prove the mean-revering nature of a "Difference line", we implement the Augmented Dickey-Fuller (ADF) test. The ADF statistics and 95% confidence interval are -3.933 and -2.869, where the p-value of this test is 0.0018. The ADF statistics highly suggest that the "Difference line" is indeed a mean-reverting series. Though we have shown that "Difference line" is mean-reverting, we cannot use the "Difference line" to implement cointegration or pair trade since we cannot long or short the MACD and signal lines. Therefore, the mean-reverting strategy fails in using MACD.

Lastly, we can look at one statistical indicator of the "Difference line" which is the half-life. We can simply imagine that the half-life of a mean-reverting series is the time window to revert. Therefore, the half life indeed provides us insight on being long or short the HSI futures. By assuming "Difference line" obeying Ornstein-Uhlenbeck formula, we can find out its half life by leveraging linear regression. We compute the half-life of "Difference line" is roughly 73 days. The calculated half-life is suggesting. If we compare the time window between two crosses of MACD and signal line, the time difference is roughly close to 70 days.

3. Trading Strategies

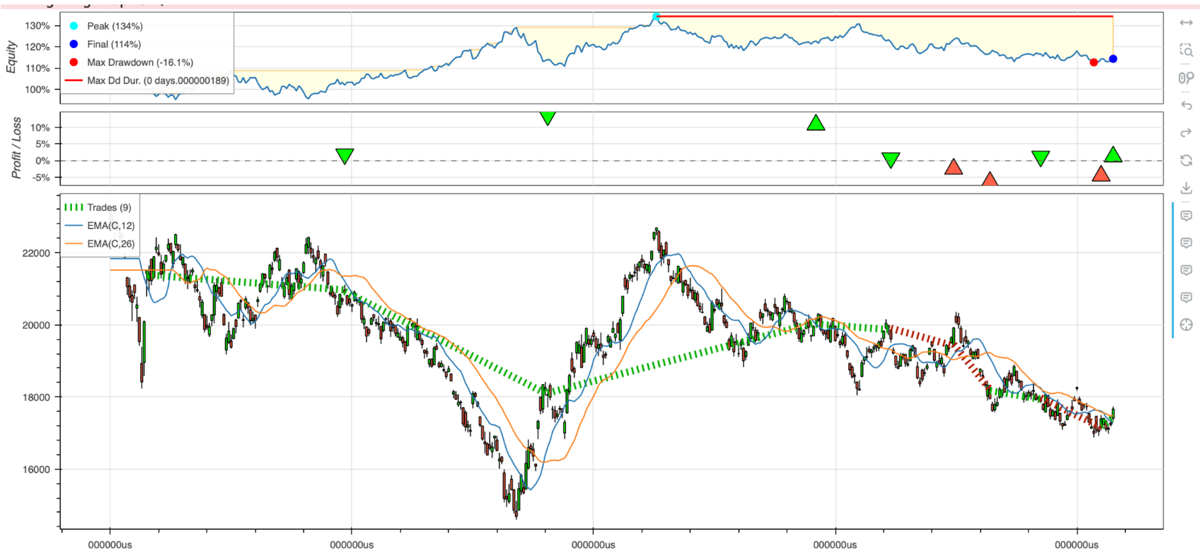
3.1 Moving Average Convergence Divergence (MACD)

3.2 Machine Learning for Up Votes / Sentiment Prediction

We attempted to predict Up votes using machine learning techniques. We simply map the problem to binary classification where Up Votes = 1 and Down votes = 0. We hope to predict the sentiment such that we can have a better idea on when we should long or short. Besides, we implement features engineering to build more features from Open, Close, High, and Low from the data. For instance, we compute the MACD and signal line of these fields and their combinations. To employ binary classification, we first split the data into training and test set where we set the split ratio to 0.8. Since the range of stock prices are at 10^4 order while the target variable is bounded by [0,1]. We had to transform our data such that they were also bounded by [0,1]. The last step is to call the Logistic Regression module from Scikit Learn library in Python and train the model. The accuracy and precision were 0.6385 and 0.7213, respectively. This was expected since the size of the dataset was not large enough to train the model, resulting in poor predictive power. Therefore, we must avoid using machine learning to predict sentiment from social media users under the limitation of dataset size.

4. Backtesting

To implement backtesting, we leveraged Backtesting.py module, which was an open-source library to do the work. We implemented the idea of MACD and utilize the Up votes data as the conditions to buy or sell. The below plots showed the result of our backtesting.



For the statistic of the backtesting, it was summarized as below

```
Exposure Time [%]                   96.394231
Equity Final [$]                  26182.47494
Equity Peak [$]                   30757.02854
Return [%]                          14.326164
Buy & Hold Return [%]              -22.869359
Return (Ann.) [%]                         0.0
Volatility (Ann.) [%]                     NaN
Sharpe Ratio                              NaN
Sortino Ratio                             NaN
Calmar Ratio                              0.0
Max. Drawdown [%]                  -16.108881
Avg. Drawdown [%]                   -3.876672
Max. Drawdown Duration     0 days 00:00:00....
Avg. Drawdown Duration     0 days 00:00:00....
# Trades                                    9
Win Rate [%]                        66.666667
Best Trade [%]                      13.642441
Worst Trade [%]                     -6.415771
Avg. Trade [%]                       1.625902
Max. Trade Duration        0 days 00:00:00....
Avg. Trade Duration        0 days 00:00:00....
Profit Factor                         2.21773
Expectancy [%]                       1.811243
SQN                                  0.841044
```

The details of the implementation is in the MACD_backtest.ipynb file.

5. Conclusion

In this report, we summarized the result of elementary data analysis of HSI data. From the HSI data, we could construct MACD and signal lines. By comparing them, we obtain a mean-reverting line which we simply called the “Difference line”. We proved that the “Difference line” was indeed mean-reverting and we calculated its half-life, which is 73 days. This half-life matched our intuition from the plot. In Backtesting, we implemented the idea of MACD and utilized the ‘Up votes’ data. By doing so, we can have arbitrage from HSI and it may serve as an indicator to buy or sell HSI futures.