RMS Project – Final Interim Report

Student name / Student no.: Wang Yucheng/56198686

Project Title: Daily Hang Seng Index Trend Prediction applying Deep Learning

Supervisor: **Jacky Keung**

1 Introduction and Topic of Research

The Hang Seng Index is a free-floating-adjusted market-capitalization-weighted stock-market index which is the main indicator of the Hong Kong stock market. Nowadays, the prediction of the Hang Seng Index along with other stock markets has attracted a lot of attention in both industrial and research fields. The stock price prediction plays an important role in trading strategies for the speculators to identify the right moment to buy and sell a stock. The non-stationary characteristics of the stock market contributes to the difficulty predicting the index trend. The samples are filled with noises and uncertainty. It is because that the performance of the stock market is not only related to the previous stock data but also closely related to the policies and current affairs such as macroeconomy. Therefore, a precise prediction of the stock market is considered impossible by the researchers, let alone purely use the historical market data. However, this challenge of prediction on the stock on a complex time series attract many researchers to work on it.

This study focuses on predicting the daily Hang Seng Index Trend applying machine learning, deep learning to be specific. Basically, the prediction of the stock index is a kind of time sequential type deep learning. We use the previous several days of data to predict the data of the next day. Traditionally, recurrent neural network (RNN) is used in this model. LSTM (Long-short-term network) is a model fit this problem very well. However, it still has lots of drawbacks in the actual operation. This study aims at improving the model and perform the analysis on the Hang Seng Index.

2 Related Work and Background

2.1 Stock Analysis Background

Generally speaking, the traders will use the fundamental analysis and technical analysis to predict the trend of the stock market. The fundamental way inspects different companies by analyzing the company parameters and the related news. It is concerned more with the company rather than the actual stock. The analysts make their decisions based on the past performance of the company. The technical analysis uses the price history to perform the prediction. It deals with the stock price based on the past patterns of the stock. When applying Machine Learning on stock data, we are actually using technical analysis to see whether the algorithm we use can reveal the underlying patterns in the stock time series. The same way can be used to forecast the performances of the company. To reach the most successful stock prediction systems, a hybrid analysis model will be involved by using both fundamental and technical analysis.

2.2 Traditionally used models

Traditional methods of analyzing the stock market such as linear regression, moving average were commonly used as it is easy to be accepted by the public. Some machine learning algorithms like logistic regression, random forest and k-nearest neighbors are also utilized to extract the relationship between the price trend with other features. Through these analyses, we can inspect the changing trend of the stock price once some specific pattern happens. Meanwhile, Support vector machine (SVM) and artificial neural networks are also adopted to conduct some financial research due to their outstanding performance in mapping features. However, the uncertainty of the stock market and the limited quality of the features creates obstacles for the relationship mapping. Meanwhile, overfitting is also a main challenge that has not been tackled.

At the initial stage of the study, I tried to train some basic machine learning model on the dataset to investigate the initial result. Several traditional machine learning model, linear regression, K nearest neighbors, are tested and the predicted value is far away from what the real value is. Since these models lacks complete feature engineering and consideration of the continuous floatation of the trend, the features they selected are restricted. Some research focuses on the regression method like sigmoid regression, polynomial regression, linear regression. All these methods perform not well in this problem. Regarding this problem, models based on the time series forecasting are more frequently used. Some research some basic time series machine learning model like auto-ARIMA and Prophet which is designed by Facebook. Though they can give a correct trend, it is not helpful for us when predicting the exact value day by day.

The structure of the neural network has great impact on the feature extraction process. Existing works have tried on different type of neural networks for the financial time series modeling. The deep learning methodologies stand out as they provide a group of units for designing the neural network structures as the feature varies.

Most of the research choses LSTM, which is a kind of recurrent neural network to solve this problem. Compared with several models mentioned before, this model performs much better in predicting the trend in the long term. However, after scaling the time range to 100 days, we find that the predicted values always have one day delay compared to the true value. In this case, the model is not effective when evaluating the stock index in the real life.

2.3 Technical Indicators

The technical analysis with the support of statistics is essential in this study. In this problem, a few technical indicators are commonly used to investigate this problem. Following are some indicators that are normally used:

Moving Average (MA): The average of the past n values till today.
Exponential Moving Average (EMA): Gives more weightage to the most recent values while not discarding the older observation entirely.
Rate of Change (ROC): The ratio of the current price to the price n quotes earlier. (n is generally 5 to 10 days).
Relative Strength Index (RSI): Measures the relative size of recent upward trends against the size of downward trends within the specified time interval (usually 9-14 days).
Daily return: Measures the dollar change in a stock's price as a percentage of the previous day's closing price.

Generally, these indicators are extracted from the historical price data to describe some specific patterns. Since these mathematical expressions are based on some presumed patterns, some information may be lost in this process.

3 Problem Formulation and Solution

3.1 Problem Formulation

The problem focuses on the modification of RNN models on the Hang Seng Index prediction and the exploration of other approaches to help quantitative investments.

3.2 Proposed method

3.2.1 LSTM

One of the core characteristics of RNN is that it can link the previous message to the current task. For example, it can use part of the previous videos to help itself understand what the current part is showing. To achieve this, many factors are depended. Sometime, the current task only depends on the nearest information. In this case, RNN can directly learn the previous message if the interval is relatively small. However, more complicated circumstances occasionally occur. The position of relevant information might be far away from the position we are now predicting. Unfortunately, when the interval keeps increasing, RNN loses the capability to link to the information which is that far. LSTM is designed to solve this problem.

LSTM is a special model of RNN (Recurrent neural network). It can study the information which is needed in the long term. Remembering the long-term message is the basic ability of LSTM practically. All RNNs has a kind forms consisted of repetitive neural network modules. In the standard RNN, the repetitive module has only one simple structure. Similarly, LSTM also has such kind of repetitive neural network module structure. But compared with the single neural network layer, LSTM has four layers interacting with each other in a special way.

The key of LSTM is the cell state. The cell is passed on the chain with a few linear interactions, which makes it easy to keep the information stably passed. There is a ‘gate’ served to add or remove information to or from the cell state. The ‘gate’ involves a sigmoid neural network layer and a pointwise multiplication.

The ‘Forget gate’ is to decide what information is not needed and can be discarded. This is the first step in LSTM model. The next step is to determine what kind of new information is needed to be stored in the cell state. Two parts are involved: The sigmoid layer is called ‘input gate layer’, determining which value is to be reset. Then, a ‘tanh layer’ creates a new candidate vector which will be added to the state.

In the next step, the old cell state will be updated to the new cell state. We multiply the old state by f_t, forgetting the things we decided to forget earlier. Then we add $i_t * \tilde{C}_t$, which is a new candidate values, scaled by how much we decided to update each state value.

Finally, we should decide which value we are going to output. This is based on our cell state, which is a filtered version. We run a sigmoid layer to decide which part of the cell state will be outputted. Next, after using tanh to process the cell state, we multiply the result with the output of the sigmoid gate and output the result.

3.2.2 GRU

Gate Recurrent Unit is also one kind of RNN. It is a variant of LSTM with easier structure and better performance. Just like LSTM, it is also designed to fix problems like long term memory and gradient in the backpropagation.

GRU’s structure of input and output is just like the normal RNN. The hidden state passed from previous node with the related information and the current input will lead to the output of this hidden layer and the hidden state that is passed to the next node.

There are two gate state generated from the previous hidden state and the current input: reset gate and update gate. The data from the hidden state will reset through the reset gate and will be spliced with the current input. Then the model will use a tanh function to scale the data to a range from -1 to 1. This process is akin to the selective memory part in LSTM.

The core stage of GRU is called the memory update. The model uses the update gate to remember or forget some memories. The gate signal is number in the range of 0 to 1. The closer the signal is to one, the more data is "memorized"; The closer you get to zero, the more data will be forgotten.

The peculiarity of GRU is that it uses only one gate to perform the forget and selective memory simultaneously. Instead, LSTM uses several gates.

3.2.3 Bidirectional RNN

Bidirectional recurrent neural network (BRNN) was raised by Mike Schuster in 1997. The core idea of BRNN is to split the state neurons of the traditional RNN to two parts, one is responsible for the positive time direction (forward states) and another one is responsible for the negative time direction (backward states). The output of the forward states will not connect to the input of the backward states. The structure is just the same as the traditional RNN if the backward layer does not exist.

The structure provides each single point in the input series of output layer with the complete context both in the past and in the future. The following picture is a bidirectional recurrent neural network along with the time series. The particular weight will be reused many times. The weight corresponds to the input to the forward and backward hidden layer, the hidden layer to itself and the hidden layer to the output layer. It is worth mentioning that there is no information flow between the backward layer and the forward layer.

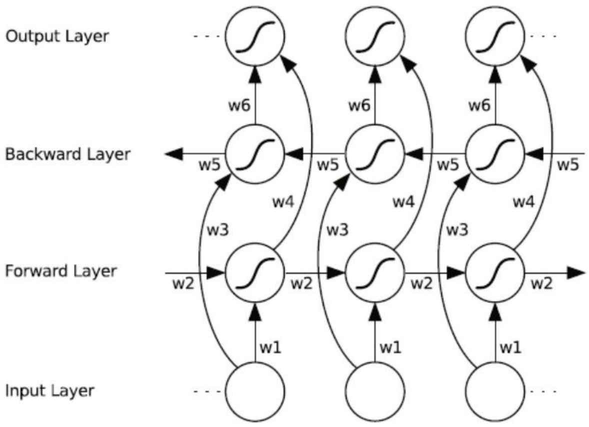


Figure 1 BRNN Structure

3.2.4 Sequence to Sequence Learning

Sequence to Sequence Learn was first put forward by Bengio in 2014. It is an algorithm that modifies RNN encoder-decoder algorithm. The encoder-decoder structure is consisted of two parts. The 'encoder' part is to encode the input information and transform it to a form of matrix. The 'decoder' part is to decode the matrix and restore to the output series.

The sequence-to-sequence learning is a structure with multiple layers. It can be combined with the RNN structure like LSTM and GRU. Each structure contains several layers of encoders and decoders. The output of the encoder will be abandoned, and we only need to restore the hidden state as the input of the next encoder. Comparatively, the output of the decoder will be restored as an input.

The sequence-to-Sequence model is frequently used in the machine translation. Theoretically, any supervised problem that is from sequence to sequence can use this model.

4 Implementation, Experiments, Analysis

4.1 Data Collection

After setting the time period covering the recent ten years, I downloaded the data from yahoo finance at first. There are only 7 columns in the dataset: Date, opening price, high price, low price, closing price, adjusted closing price and volume. In the real experiment, we need far more than these seven parameters to reach the best performance. However, we will only use the previous closing price to train the model and predict the closing price on the targeted date. The size of the dataset is 2466. We divide it into the training set and validation set. 80% of the data will be

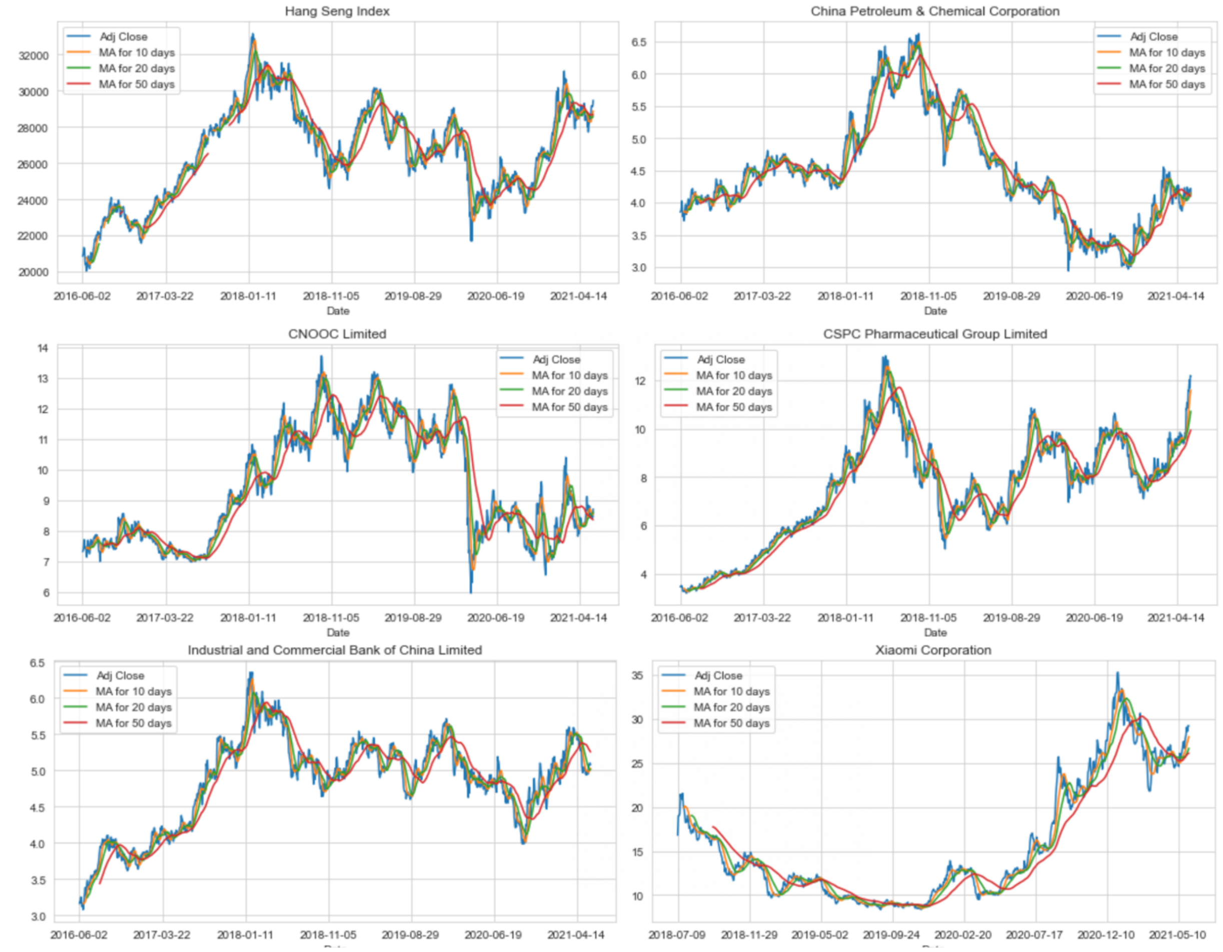


Figure 2 Moving Average of the Selected Stock Price

used as the train set and the remaining 20% will be regarded as a test set.

4.2 Analysis on Main Components

The first step of the whole study is to conduct some analysis on Hang Seng Index and some main components of the Hang Seng Index. In the real implementation, I choose five main components of Hang Seng Index having largest volume: Xiaomi Corporation (1810.HK), Industrial and Commercial Bank of China Limited (1398.HK), CNOOC Limited (0882.HK), China Petroleum & Chemical Corporation (0386.HK), CSPC Pharmaceutical Group Limited (1093.HK).

The analysis involves the present of some technical indicators and the correlation between each company.

4.2.1 Analysis on Moving Average

Moving Average refers to the average of the past n values till today. In the analysis, I choose 10, 20, 50 days to analysis the moving average respectively. The intend of the MA is to mitigate the impacts of random, short-term fluctuations on the stock price in a certain time period.

4.2.2 Analysis on Daily Return

To analyze the risk of a specific stock, daily return is frequently used. It measures the dollar change in a stock's price as a percentage of the previous day's closing price. In order to do so, we need to analyze daily changes of the stock, instead of the value itself. In this part, the study research on the daily return of these companies in the recent one year by using the line chart.

As shown in the graph, the percent change of Hang Seng Index is the smallest because of its great volume, which is commonsensible. By analyzing the line chart, we can also see that the company like CNOOC has a varying daily return that is not stable enough. In most of the days, the daily return is restricted in the range from -0.05 to 0.05. However, the daily return once reached plus/minus 0.15 in some days. By contrast, Xiaomi’s daily return is uniformly distributed in the range from -0.10 to 0.10.

Though we can draw some conclusions from the line chart, it is not intuitive enough for the readers. Here, we use the histogram to further analyze the frequency, visualizing the distribution of the daily return in different intervals.

As illustrated in the graph, the daily return of all the companies selected conforms to the normal distribution. It can be clearly seen that

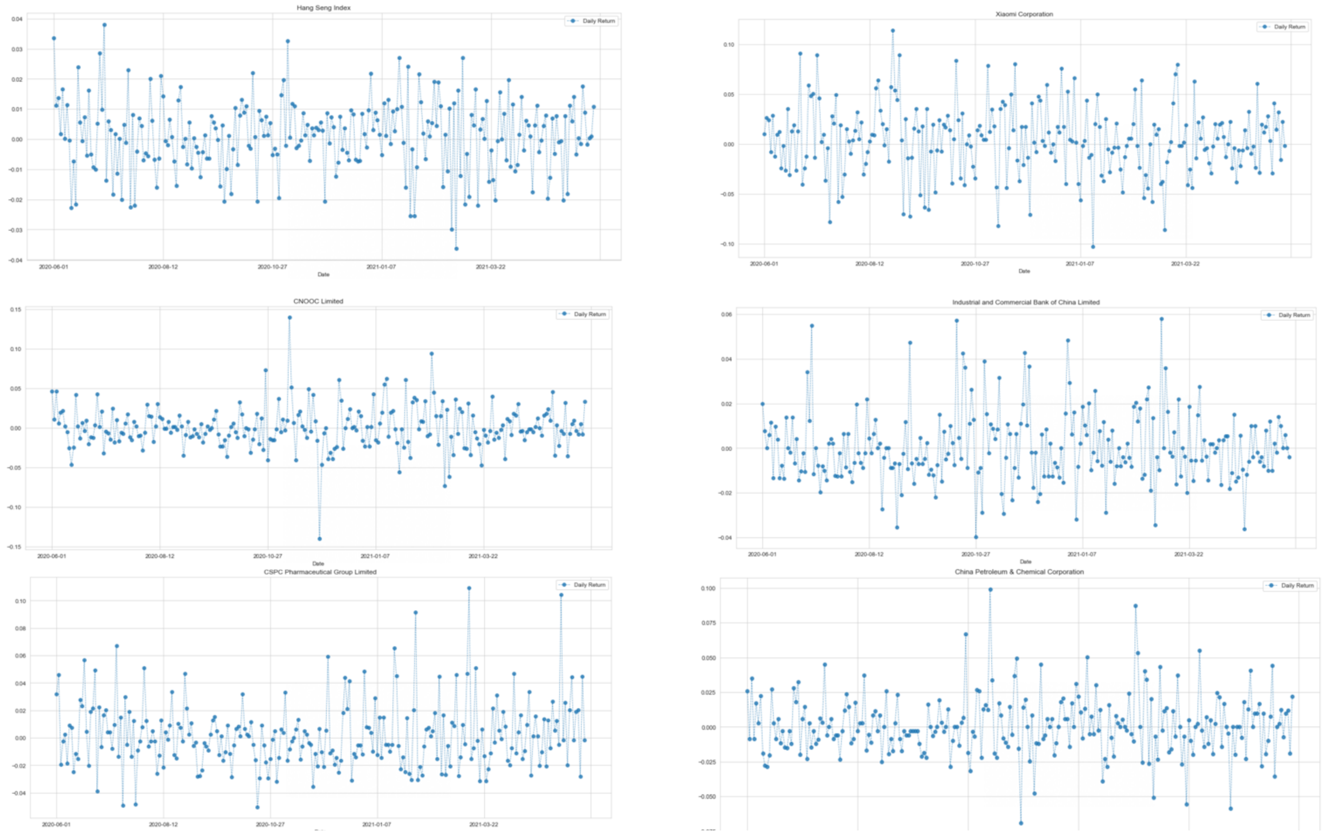
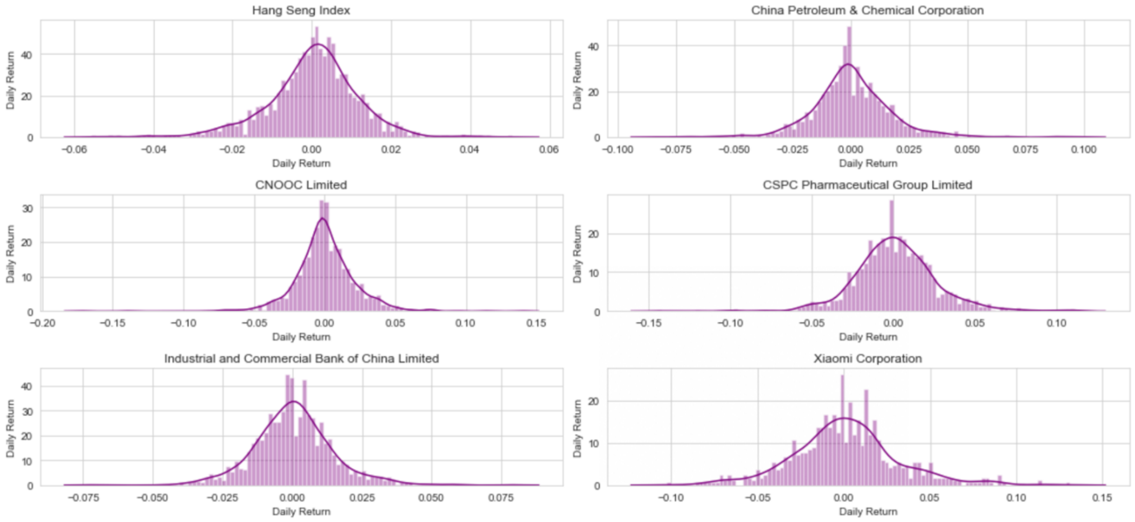


Figure 3 Return Value of the Selected Stock Price


Hang Seng Index

China Petroleum & Chemical Corporation

CNOOC Limited

CSPC Pharmaceutical Group Limited

Industrial and Commercial Bank of China Limited

Xiaomi Corporation

Figure 4 Distribution of the Return Value

the standard deviation of the daily return of Industrial and Commercial Bank of China Limited is the smallest which means the distribution is the most concentrated. Correspondingly, the stock price of this company is the most stable one with little fluctuation and risk. By contrast, the standard deviation of Xiaomi is much larger than that of Commercial Bank of China Limited.

4.2.3 Correlation between stock prices

This part analyzes the correlation between the closing price of each of the companies. I use seaborn and pandas to repeat the comparison analysis for each combination of these 6 stocks.

The relationship between CNOOC Limited (0883.HK) and China Petroleum & Chemical Corporation (0386.HK) is typical example showing that these two stocks are positively correlated with each other. As illustrated in the graph, a linear relationship between the two

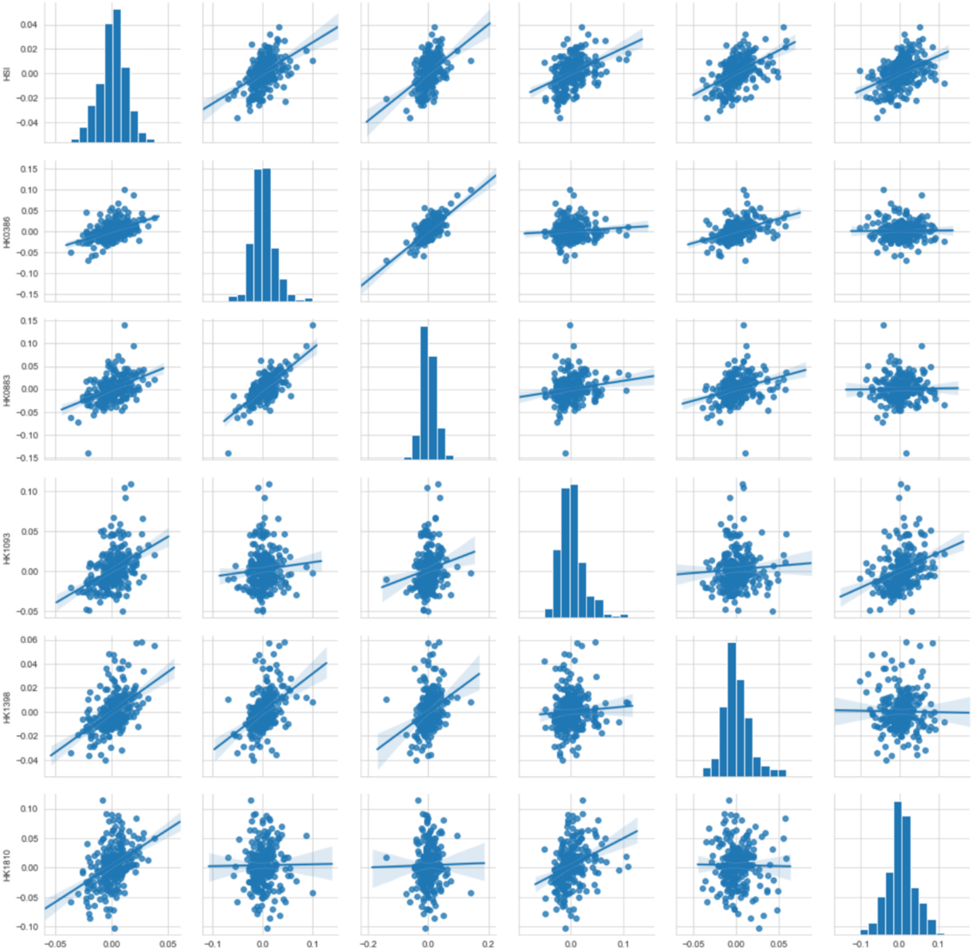


Figure 5 Correlation between the Selected Stock Price

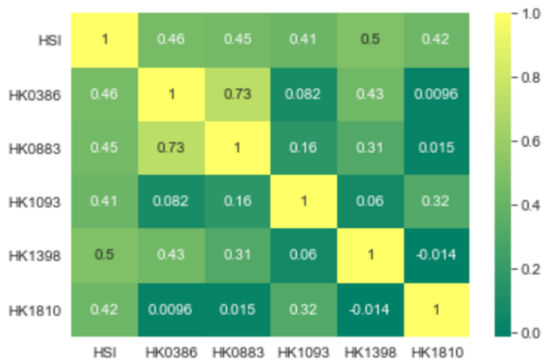


Figure 6 Correlation of the Selected Stock Price (Heatmap)

company's daily return occurs. Almost all of the points lie in the first quartile and the third quartile. To investigate the reason behind the linear relationship, we can easily find that these two companies both are closely related to the petroleum.

Another interesting conclusion is related to Xiaomi Corporation (1810.HK). It is not hard to see that the daily return of Xiaomi has no linear relationship with any other stock.

I also use the heat map to visualize the same correlation. It is even more visualizable since we can directly get the numerical value for the correlation between the return values of the stock. Here we can verify our observation through the scatter plot. CNOOC Limited (0883.HK) and China Petroleum & Chemical Corporation (0386.HK) has the correlation. The correlation coefficient between Xiaomi Corporation and other companies are close to 0 which means there does not exist linear relationship between them. From the heat map, we can also find that the correlation coefficient between Heng Seng Index and its components is around 0.5.

4.3 Basic Model Rework

After collecting the data, I redid the experiment using several models which have been discussed in the related work. These results can be regarded as a reference standard to see how the algorithms improves the result afterwards.

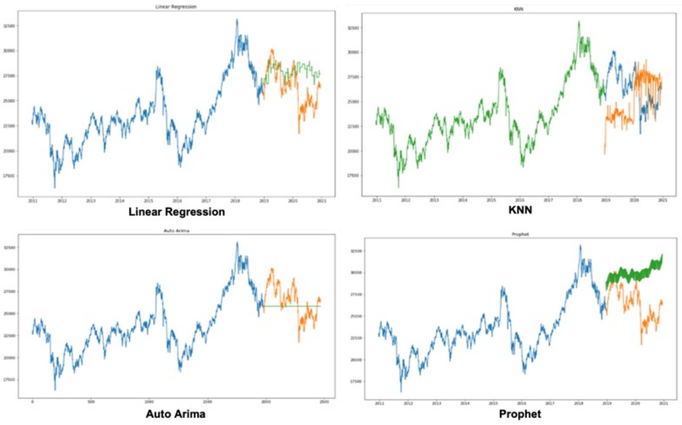


Figure 7 Performance of the Basic Model

With default settings, these models perform terribly as expectation. Feature engineering is an important step before training these models. Without enough useful indicators and features, the model can only explore the results by using some basic data without too much information.

4.4 Deep Learning model Improvement

In the real implementation, I use in total six models to analyze the performance and the improvement of the result. The results of LSTM and GRU are used as baselines. In detail, I apply bidirectional RNN and sequence-to-sequence RNN on LSTM and GRU respectively to analyze the improvement of the structure.

The training set contains the data in the past 10 years. Meanwhile, a test size of 30 days is set to test the result. For each model, I repeated the process for 5 times to generate 5 different predictions. Every simulation contains 100 epochs. The layer size is set as 128 and the learning rate is set as 0.01. The value takes the performance of the model into consideration. Higher learning rate results in less epoch to reach the convergence. However, the performance of the model is restricted.

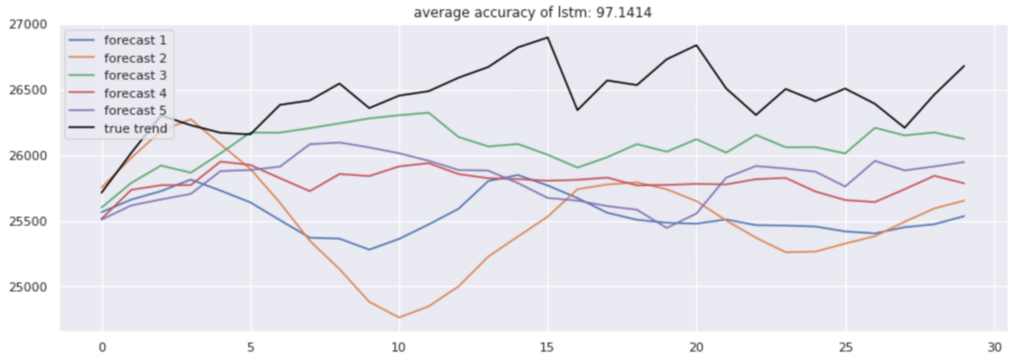


Figure 8 Performance of Long-short Term Network (LSTM)

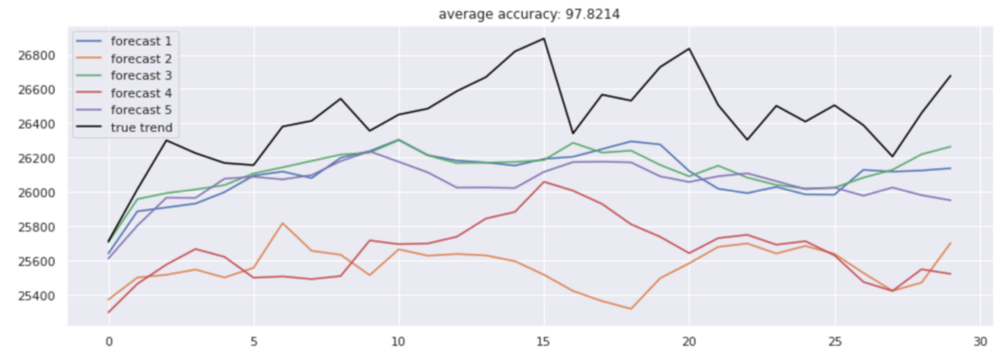


Figure 9 Performance of Bidirectional LSTM

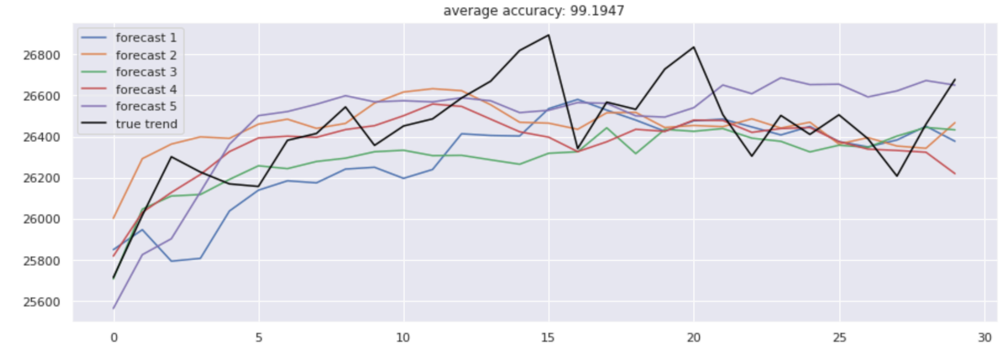


Figure 10 Performance of Sequence-to-sequence LSTM

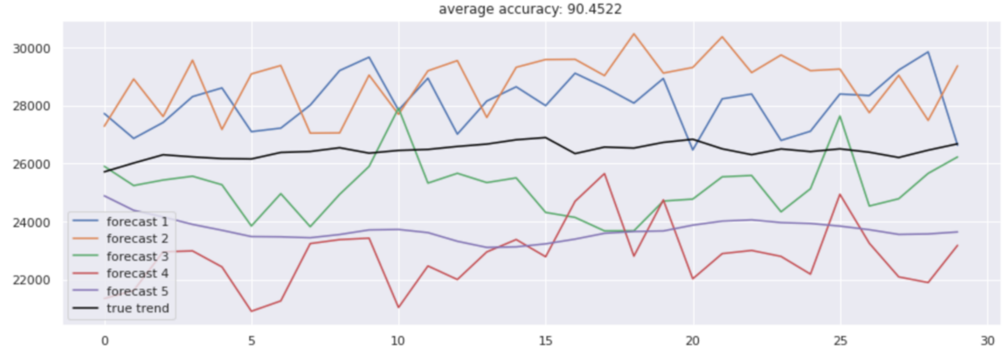


Figure 11 Performance of Gate Recurrent Unit (GRU)

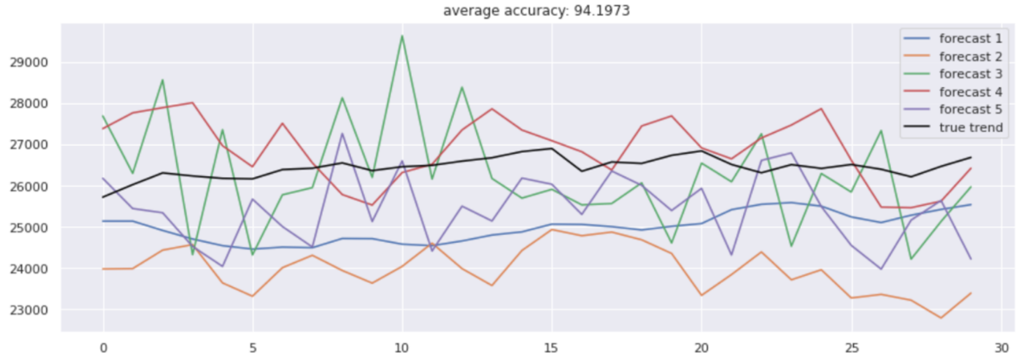


Figure 12 Performance of Bidirectional GRU

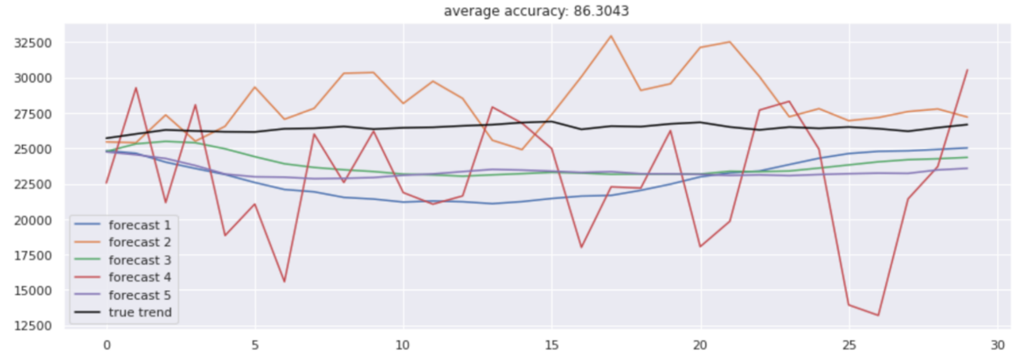
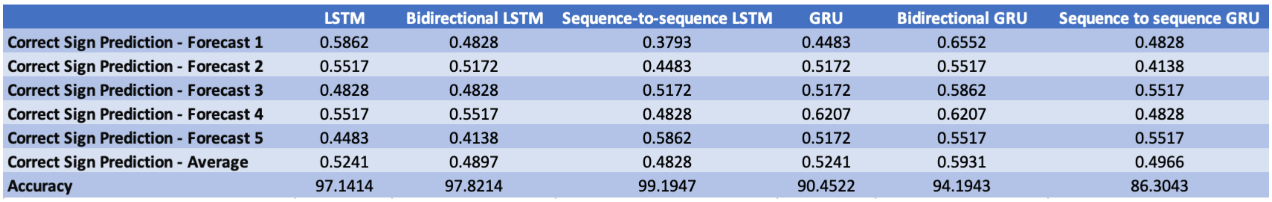


Figure 13 Performance of Sequence-to-sequence GRU



	LSTM	Bidirectional LSTM	Sequence-to-sequence LSTM	GRU	Bidirectional GRU	Sequence to sequence GRU
Correct Sign Prediction - Forecast 1	0.5862	0.4828	0.3793	0.4483	0.6552	0.4828
Correct Sign Prediction - Forecast 2	0.5517	0.5172	0.4483	0.5172	0.5517	0.4138
Correct Sign Prediction - Forecast 3	0.4828	0.4828	0.5172	0.5172	0.5862	0.5517
Correct Sign Prediction - Forecast 4	0.5517	0.5517	0.4828	0.6207	0.6207	0.4828
Correct Sign Prediction - Forecast 5	0.4483	0.4138	0.5862	0.5172	0.5517	0.5517
Correct Sign Prediction - Average	0.5241	0.4897	0.4828	0.5241	0.5931	0.4966
Accuracy	97.1414	97.8214	99.1947	90.4522	94.1943	86.3043

Figure 14 Testing Results

4.5 Results Discussion

The model can be evaluated by two criteria. The first one is the accuracy. Literally, it calculates the root mean square of distance between the real value and the predicted value divided by the real value. It is a common way to evaluate the model. As illustrated in the graph, LSTM algorithm with its variant performs much better than that of GRU, which means the predicted value of LSTM algorithm is far closer than that of GRU. Among them, the sequence-to-sequence LSTM performs better.

Another important criterion is the correct sign prediction. It records the correct rate of predicting the rise and fall of stocks over 30 days. For most of the models, the correct rate is about 0.5 and no improvement is shown by modified the model. Bidirectional GRU, with the average correct rating around 0.6, has a higher rating than any other model by almost 10 per cent.

5 Conclusion and reflection on experience gained through participation in RMS

5.1 Research Conclusion

The study focuses on the improvement of the deep learning model on predicting the stock market. I apply the bidirectional structure and sequence-to-sequence structure on the basic Long-short-term network and Gate Recurrent Unit. Some results improve a little bit, but we cannot actually predict the stock price simply through these models. The prediction on the stock market involves some more specific indicators and need to combine the model with the influence of current affairs. The next step of the research will take the emotional analysis into analysis by scratching the related news on daily basis.

5.2 Reflection

The research mentoring scheme helps me get to learn about the whole process before publishing a paper. By looking into several related papers, I have learnt about the structure and the format of a paper, which is essential that introduces me into the research world. Meanwhile, I have to look up many unapprehensive concepts and search for it to understand the thesis. Through this process, the knowledge builds up a whole picture about machine learning, deep learning and stock prediction for me. I really appreciate the chance which guides me to the research.

6 References

[1]. Long, Wen, Lu, Zhichen, & Cui, Lingxiao. (2019). Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Systems, 164, 163–173. https://doi.org/10.1016/j.knosys.2018.10.034

[2]. Singh, R., Srivastava, S. Stock prediction using deep learning. Multimed Tools Appl 76, 18569–18584 (2017). https://doi.org/10.1007/s11042-016-4159-7

[3]. Chung, Junyoung, Gulcehre, Caglar, Cho, KyungHyun, & Bengio, Yoshua. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

[4]. Schuster, M, & Paliwal, K.K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681. https://doi.org/10.1109/78.650093

[5]. Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, & Bengio, Yoshua. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.

[6]. Rao, Polamuri Subba, Srinivas, K, & Mohan, A. Krishna. (2020). A Survey on Stock Market Prediction Using Machine Learning Techniques. In *ICDSMLA 2019* (pp. 923–931). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-1420-3_101

[7]. Joo, I.-T., & Choi, S.-H. (2018). A stock price prediction model based on a bidirectional LSTM circulatory neural network. The Journal of the Korea Institute of Information, Electronics and Communication Technology, 11 (2), 204–208. https://doi.org/10.17661/JKIIECT.2018.11.2.204

[8]. Althelaya, Khaled A, El-Alfy, El-Sayed M, & Mohammed, Salahadin. (2018). Evaluation of bidirectional LSTM for short-and long-term stock market prediction. *2018 9th International Conference on Information and Communication Systems (ICICS)*, 151–156. IEEE. https://doi.org/10.1109/IACS.2018.8355458

[9]. S. Mootha, S. Sridhar, R. Seetharaman and S. Chitrakala, "Stock Price Prediction using Bi-Directional LSTM based Sequence to Sequence Modeling and Multitask Learning," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, 2020, pp. 0078-0086, doi: 10.1109/UEMCON51285.2020.9298066.