University of Sunderland

Faculty of Technology
Department of Computer Science

PROM02 – MSc Dissertation
MSc Data Science

Student Name: Jackie Chu Lai Ting
Supervisor Name: Marta Rudina
Second Marker Name: Ming Jiang

A stock forecast model for swing trading recommendation in the Hong Kong stock market

July 2023

Declaration

I declare the following:

(1) that the material contained in this dissertation is the end result of my own work and that due acknowledgement has been given in the bibliography and references to **ALL** sources be they printed, electronic or personal.

(2) the Word Count of this Dissertation is 13768.

(3) that unless this dissertation has been confirmed as confidential, I agree to an entire electronic copy or sections of the dissertation to being placed on the eLearning Portal, if deemed appropriate, to allow future students the opportunity to see examples of past dissertations. I understand that if displayed on the eLearning Portal it would be made available for no longer than five years and that students would be able to print off copies or download.

(4) I agree to my dissertation being submitted to a plagiarism detection service, where it will be stored in a database and compared against work submitted from this or any other Department or from other institutions using the service.

In the event of the service detecting a high degree of similarity between content within the service this will be reported back to my supervisor and second marker, who may decide to undertake further investigation that may ultimately lead to disciplinary actions, should instances of plagiarism be detected.

(5) I have read the University of Sunderland Policy Statement on Ethics in Research and I confirm that ethical issues have been considered, evaluated and appropriately addressed in this research.

SIGNED: 

DATE: **6Jul2023**

Abstract

Investors are eager to find effective ways to predict the stock market for successful investments. This project aimed to develop a machine learning forecasting model to optimize swing trading investment strategy in the Hong Kong stock market by providing visual recommendations for future actions. The study began with a literature review to define objectives, followed by building logistic regression and LSTM models for trend and price prediction on four selected stocks in the Hong Kong stock markets. Historical stock data were collected from Yahoo Finance, and exploratory data analysis was conducted using visualization methods. The logistic regression model using PCA feature set outperformed the entire available feature set for most of the selected stock, while the LSTM model using 60-days sliding adjusted close price performed better with lower RMSE, MAE, and MAPE. The best-performing algorithms were applied to real-time data, and the results showed high accuracies up to 71% for logistic regression model and normalized RMSE kept at 5% below or the lowest for HSI of only 1% respectively for LSTM model. The models were robust and effective, as illustrated by different visuals. Future work includes extending the study to more stocks, enhancing the hyperparameter tuning process, and combining trend and price prediction to provide more informed investment recommendations. Overall, this study provides an end-to-end approach to machine learning forecasting models in stock prediction, with satisfactory results.

Contents

1 Introduction

1.1 Background

Many investors perceive stock forecasting as a crystal ball. With the advancement of technologies, billions of sophisticated algorithms backed by super computers enable the forecasting models process ever-larger amounts of data, this increases the certainty and confidence during investment decision making. Machine learning forecasting is therefore being on spotlight for investors to leverage the forecasting result to grasp the special pattern of the stock market and maximize their opportunities to win.

From the source of Hong Kong Securities and Futures Commission [1: HKSFC], Hong Kong stock market was one of the top 10 in the world and top 5 in Asia, its market capitalization up to USD 4,567 billion as end of Dec 2022. The Hong Kong stock market is fully regulated with huge trading volume characteristics, high liquidity, transparency and independent judiciary, machine learning algorithms could be applied for pattern recognition that can be used as the recommendation tools to maximize the opportunities to win during the investment.

There were a lot of research focused on other stock markets, for instance, Ravikumar and Saraf, 2020 [2] attempted on US stock price and trend prediction by various classification and regression algorithms. On the other hand, Yadav, Jha and Sharan, 2020 [3] studied the LSTM modelling on Indian stocks. This paper would like to study and explore machine learning model to predict the stock price for the Hong Kong stock market for swing trading recommendations.

1.2 Aims

It is a hot topic to apply machine learning algorithms onto stock investment. Machine learning refers to techniques that can identify patterns and make predictions, broadly classified into supervised learning, unsupervised learning and reinforcement learning. Nasteski, V. (2017) [4] opined that supervised learning nowadays has the most use cases. Supervised means classes are predetermined and finite, usually from historical information or human inputs. Depends on the nature of predictors, it divides into classification and regression models. Classification models map the input to a predefined categories while regression models keen on finding the model to derive the value in the domain. Tons of algorithms built from various mathematical concepts. Decision trees, support vector machines, logistic regressions are some of the common algorithms under supervised learning area.

In recent years, artificial neural networks (ANN) get the most attentions when doing machine learning due to the improvement of computing power. Huang, Y (2009) [5] studied the development of ANN and described different ANN methodologies based on the idea of human nervous systems in human brain to solve complex problems by computers. The development of ANN extended into many areas including stock market predictions.

This project aims to discuss the landscape of Hong Kong stock market, critically reviews on various machine learning algorithms and practically develop machine learning forecasting model to assist optimization of swing trading investment strategy in the Hong Kong stock market by providing a visual result as a recommendation for further action. This can then serve as a decision recommender by employing machine learning algorithms on the stock trend and price of the next day.

1.3 Objectives

To facilitate the development of machine learning forecasting model, six objectives have been set as below,

- Background research by evaluating different investment strategies and studying modern approaches and methods used for stock forecasting
- Literature review to critically evaluate available forecasting algorithms and its application
- Study the Hong Kong stock market landscape by reviewing economic situation and performance technical analysis
- Research data collection process by Yahoo Finance
- Identification and discussion of relevant professional, ethical, social and legal issues
- Build, train, deploy and visualize forecasting model(s) for evaluating the one with the highest prediction accuracy and efficiency to assist optimization of swing trading in the Hong Kong stock market.

1.4 Research Approach

As the aim is to develop a model that can be used effectively in Hong Kong stock market. This project takes the approach of performing a thorough literature review to understand the investment strategies available. Then learn from previous research, on types of stock predictions, features considered, as well as machine learning algorithms available and their performance correspondingly. The literature review process serves as an input to set the foundation to a practical experiment design on the Hong Kong stock market. Hang Seng Index (HSI) and 3 of the highest volume stocks which are Alibaba, Tencent, Meituan were selected.

Data collection through Yahoo Finance, data preprocessing, model building and evaluation in order to identify the best model amongst to be adopted in responding to the aim.

1.5 Structure of the Report

This paper begins with section 2 of a thorough literature review. Types of investment strategies to be selected, stock prediction approaches and its applications for stock forecasting, and software to be used in this project are the three big area to be focused. In particular to the stock prediction approaches and its application, there are critically discussions on the methodology of each studies including prediction target, features used and the outcome. This forms the basis of the practical experiment in the later sections.

Section 3 talks about the landscape of the Hong Kong stock market, data sources and illustrates the exploratory data analysis of the HSI and the 3 selected stocks.

Starting from section 4 is the practical experiment, including the project management plan, experiment design, data collection and preprocessing, as well as model building and evaluation. Following by section 5 depicting the experiment result and the evaluation.

The paper ends with the conclusion and evaluation of the practical experiment, as well as the recommendation on future works could be done and personal reflection at section 6. There are in ***bolded italic fonts*** in line as the interim key points.

1.6 Social, Ethical, Professional and Legal Considerations

In today's digital world, data is a new type of oil and it is important to maintain good use of data in various aspects including social, ethical, legal and professional considerations. According to Mageswaran et al, 2018 [6], *"Machine Learning is an emerging field which has created a significant positive and undesirable impact to many industries"*. This section would like to discuss the definition by each area and the impacts to each area particularly affected by this project.

Social – Traditionally investors made decision by actively looking at the financial reports and the technical indicators to find the best performing stocks to maximize their opportunities on investment returns. With the rapid machine learning algorithm developments, investors mostly make investment decision through the assistance of those model results. This makes the decision making process faster and interfered by the model recommendation. The good side is that this expands the stock market as more people can participate into the stock market with lesser knowledge by the help of model recommendations. On the contrary, in case the model has been

widely adopted by a lot of investors, the investment decisions will be fed in to as the data inflow then form the cycle of looping back to the model and eventually disrupted the stock market. Market sentiment is one of the area that greatly impact by this kind of "social decision". However, this project is an academic study without using market sentiment during model development therefore have not much impact to the society with this scale.

Ethical – Ethics is concerned with what is good for individuals and society as well as moral philosophy. Under the data science context, ethics focused on treating individual fairly, conveying into the protection of personal information, for instance, age, gender, religion, races… This project uses data from stock market and are publicly available therefore no ethical issues in this context.

Professional – Professional refers to the ability to interpret and practice according to specific professional area. The professional in this project is the model building process using quality data and adequate methodology. This project adopted data collected from Yahoo Finance which guarantees the data quality. Data cleansing and verification during exploratory analysis were performed to ensure again the data were suitable for model building. On the other hand, data models were built by iterative evaluation process, by undergoing train / test process and comparing to a baseline model to ensure the best among results were selected and adopted.

Legal – In a nutshell, legal refers to a set of rules that need to be adhered or otherwise punishable by law to regulate behaviour by the society. In the context of machine learning projects, legal concerns mostly related to how the data and information being used and whether the software and applications used are legally consented. Data privacy part has been discussed in the ethical session and here focus on the software and applications used. In this project, the data source comes from Yahoo Finance web site. It allows data download as well as using API like Python yfinance library. There is a disclaimer clause in yfinance that *"It is an open-source tool that uses Yahoo's publicly available APIs, and is intended for research and educational purposes"*. Therefore, there is no legal concern as this project is for academic purpose only.

2 Literature Review

This section consolidates the related works of this project. It divides into three areas:

1. Investment strategies
2. Modern approaches and its applications for stock forecasting
3. Software to be used for this project

2.1 Investment strategies

Investment strategies refer to a plan designed by an investor for achieving their investment goals. This is based on their existing capital, risk appetite and their personal circumstances, including age, source of income, living style, etc. All the above factors constitute investment attitude and formulated respective unique investment strategies, vary from conservative to highly aggressive. Conservative investment strategies are safe investments that normally achieved stable returns with lower risks in a longer investment period. Conversely, aggressive investment strategies employed by risk lovers or those having a goal of generating maximum returns and normally in a shorter investment period.

Different investment strategies corresponding to the trading methods by different trading frequencies, referring to day trading, swing trading and long-term trading. Graham, P. ed., (2009) [7] analysed stock market investment strategies and quoted *"Day trading is the strategy of buying and trading a stock within the same day. It has a notorious reputation for huge returns, as well as huge losses."* It leverages technical analysis and instant market data to catch the short-term opportunities and make quick decisions during the day. Minutes or hourly information are the primary consideration. In summary, it poses highest risk among the three methods.

In addition, Graham continued *"swing trading is the strategy of trading at the peaks of price oscillation over a period of a few days or weeks"*, requiring an investor with more understanding on the company direction in order to identify the entry and exit points by recent fluctuations activities. Usually, market indicators trends such as MACD and RSI are used where these information nowadays are handily available daily information at market websites.

On the other hand, long-term trading focused on the fundamental factors of the company, requiring investors to have in-depth analysis on company background and its potential to growth in few months and longer time span. Investors take the advantages of compound interest and benefit from long term capital appreciation in the stock market. This kind of investment

mitigates the risk of short-term market fluctuations however take long time for investment growth.

Chandra, A. (2008) [8] researched from psychological perspective, found that *"investors do not always act rationally during investment decision making while they usually suffer from several psychological and emotional biases."*

Leveraging model recommendations can help to minimize the psychological and emotional biases. In this case, swing trading is suitable trading method to proceed with machine learning model development. Swing trading involves a trading period of a few days or weeks, providing enough time for investors to consider the recommended decisions from the model and order their trade instruction. Additionally, with the advancement of technology and computer capabilities, daily stock data is readily available from the internet. Moreover, historical stock oscillation information of companies is suitable to determine decisions in swing trading. ***By utilizing a machine learning model for investment recommendations, the model can provide optimized swing trading recommendations*** while minimizing psychological and emotional biases.

2.2 Modern approaches and its applications for stock forecasting

There are many ways for stock analysis. Fundamental analysis and technical analysis are the most commonly used traditional methods. Whilst the development of advanced technology, machine learning algorithms are applied to stock predictions attempting to improve market forecasting when compared with traditional approaches for better decisioning.

Prediction purposes and data available are the key factors to decide which algorithms to be employed. For investors with purposes of getting signals about when to buy or sell stocks, price movement prediction with classifiers buy, sell or hold would be the options. This type is classification problem. On the other hand, stock price prediction gives not only a direction but also the predicted stock price that can assist investors to make decisions about their overall investment strategies. This type is regression problem.

Ravikumar and Saraf, 2020 [2] studied the stock prices prediction using machine learning algorithms in both regression and classification methodologies.

Regression models	Classification models
1. Simple Linear Regression	1. Support Vector Mahine
2. Polynomial Regression	2. K-Nearest Neighbors (KNN)
3. Support Vector Regression	3. Logistic Regression
4. Decision Tree Regression	4. Naïve Bayes
5. Random Forest Regression	5. Decision Tree Classification
	6. Random Forest Classification

Table 1: List of machine learning algorithms in regression and classification models

Apple stock dataset was used for model building and algorithm accuracy comparison. Learning and analysis performed and logistic regression model resulted the maximum mean accuracy of 68.622% amongst the aforementioned model techniques. Future work of neural network techniques can be applied and time series methods can be further explored. Moreover, social media information like twitter can also be considered as another type of data sources for incorporate the market sentiment into the prediction.

The rest of this section studies related works, critically evaluates on different approaches and their applications. It is used for formulation of the directions for a practical experiment of this project.

Logistic regression

Logistic regression is one of the most common types of classification algorithms. It aims to find the best-fitting model to describe the relationship between predictor variables and the binary response variable. Its equation is to predict the probability of event occurring by using sigmoid function, $P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n+\varepsilon_i)}}$

This is the graphical presentation of the sigmoid function.

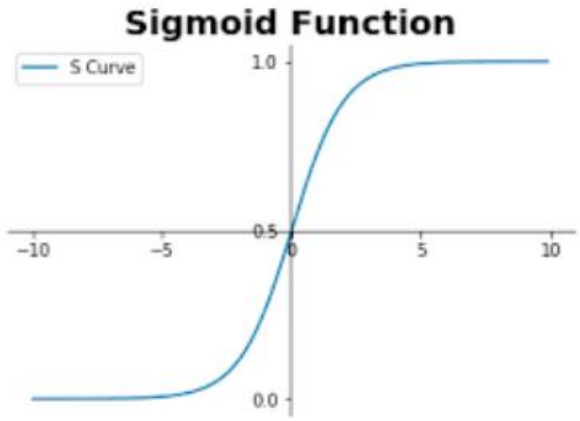


Figure 1: Graphical presentation of sigmoid function

In the context of stock prediction, the predicted targets classified buy, sell or hold, then calculates the probability of each class and the predicted class is the one with highest probability.

Smita, 2021 [9] performed a thorough review on logistic regression model. The paper explained *"regression analysis is one of the most useful and the most frequently used statistical methods."* Logistic regression variated by extending similar idea where dependent variable Y is categorical. Performances of company of S&P BSE 30 have been studied and performed logistic regression model building step by step in detail. Smita commented *"The logistic regression model has overcome the tedious, expensive and time-consuming process of traditional techniques of predictions. This model can be a stepping stone for future prediction technologies".*

Upadhyay, Bandyopadhyay and Dutta, 2012 [10] adopted multinominal logistic regression to predict and categorize the result into GOOD, AVERAGE and POOR, by comparing to the return versus market. The model used seven financial ratios as the dependent variables with the prediction rate of 56.8%, compared to the ordinary random rate of 33.3%, the model equation was a lot higher accuracy and concluded the model can uplift investor's stock price forecasting ability. Upadhyay, Bandyopadhyay and Dutta checked and confirmed the prediction capability, 101 test samples have been considered and result of validation showed the independent data set and proved the model was valid. The study could be further work on a longer data period and consider additional dimensions such as macro-economic factors to see if the accuracy could be further uplift.

Rout, 2020 [11] discussed the advantages and disadvantages of logistic regression. The most attractive advantage is that logistic regression is easy to implement and interpret, efficient to train with no assumptions requirement on distributions of classes in feature space, as well as accurate performance resulted when the dataset is in linear relationship. Moreover, it is easy to extend to multinomial regression with good accuracy for many simple datasets especially to those can be linearly separable. Nevertheless, it is also its greatest limitation as it requires a linear relationship between the dependent and independent variables whilst the data may be non-linear in real world examples.

In summary with the short development time and easy understandable logic***, logistic regression will be included in the practical experiment for trend prediction*.**

KNN

KNN is a simple and fast algorithm based on the idea of assigning same class of the most similar samples in the k nearest data points of its nearest neighbours due to its assumption that similar

things exist in close proximity. Mathematically it calculates the distance between points and identifying the shortest distances to assign the class to. The common distance function namely Euclidean distance with the below function,

$$d(x,y) = \|x-y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The model building process anchored at setting the value of K. Harrison, 2018 [12] described how to choose the right value of K by *"running the KNN algorithm several times with different values of K and choose the K that reduces the number of errors encounter while maintaining the algorithm's ability to accurately make predictions when it's given data it hasn't seen before."* Moreover, K usually set to be an odd number to ensure a tiebreaker.

The key advantage of KNN can make predictions without training, therefore KNN is good to used in real-time or streaming data problem which can be easily updated with new data to facilitate predictions to be updated from time to time. However, KNN is time consuming while sample sizes and number of features increases due to its complex searching mechanisms on getting the nearest neighbours for each sample. Normalization process and missing value handling are also important during KNN model building as they can be greatly affecting the model accuracy.

Alkhatib et al., 2013 [13] adopted KNN algorithms for stock prediction. Around 200 samples of each of five companies had been randomly selected for model training. Closing price was found to be the main factors that affects the prediction process. Results shown KNN with k=5 was an efficient prediction algorithm that was stable and robust with small error ratios. The charts plotted between actual and predicted values also shown that the prediction results were close to actual prices. These proved the model could help decision making. Although it is theoretically correct, the implementation and the information systems technology were the key concerns to make it practical and usable in the real situation.

Latha et al., 2022 [14] proposed KNN to predict the stock movement with the predicted value was either HIGH or LOW. Data of date, open, close, high, low and a derived attribute of change over time have been used for model building. Confusion matrix was used for model evaluation. The accuracy then calculated by the true prediction (either positive or negative) over total number of values. Accuracy of each model with different values of K was calculated and plotted. It is found that the highest accuracy resulted at K = 6 for accuracy of 59%. It is believed that more other parameters and details should be considered and KNN cannot incorporate the time series

nature only with the changes over time attributes. Other algorithms such as neural network techniques could be considered as the next step.

Manimegalai et al., 2022 [15] performed stock market trend prediction by ensemble approach. Combining conventional KNN and bagging classifier to form the bagging-based KNN. Manimegalai et al. elaborated *"the bagging approach uses numerous versions of a training set. By sampling with replacement, each version of the training sample can be created. Voting is used to aggregate the model results into a single output. Bagging can be used to assemble KNNs"*. Various algorithms included random forest, SVM, KNN, KNNB, logistic regression applied to four companies which KNNB resulted the highest accuracy of 94.67%. The paper demonstrated the best model found from various approaches as well as hybrid or ensemble approaches could be a consideration of enhancing from conventional models.

The reviews of using KNN worked well in both price trend and price predictions. Nevertheless, the complex searching mechanisms on getting the nearest neighbours for each sample greatly affect the training time and limited the number of dimensions to be used. Stock price was temporal in nature with many different factors affecting the upcoming trends, *KNN* ***may not be practicable as a good stock prediction tools*****.**

Random forest

Random forest is an ensemble method that the collection of decision tree models forms a more accurate model. Gini impurity is the common method to determine the best split that result in the most homogenous trees, calculate by

$$Gini\ Impurity = 1 - \sum_{i=1}^{n} p_i^2$$

The lower the Gini impurity, the more homogenous the trees will be. In stock prediction, random forest can be used to predict future stock price movements. It can handle large amounts of data irrespective numeric or categorical data. Moreover, it will not impact by data scaling and therefore easier data preprocessing progress.

Khaidem, Saha and Dey, 2016 [16] cited *"Intrinsic volatility in stock market across the globe makes the task of prediction challenging. Forecasting and diffusion modeling, although effective can't be the panacea to the diverse range of problems encountered in prediction, short-term or otherwise."* The study predicted stock price movement by random forest algorithms. It described in detail on data preprocessing process and leveraged technical indicators, which were

time-series nature and were used for checking bearish or bullish signals. The paper described to use those technical indicators as the features for model building. It included relative strength index (RSI), stochastic oscillator, William %R, moving average convergence divergence (MACD), price rate of change (PROC), on balance volume (OBV). Accuracy, precision, recall and specificity were the parameters to evaluate the model robustness. Next 1 month, 2 months and 3 months prediction model for those parameters have been calculated for tested stocks data. The results shown that the model was outperformed than other algorithms discussed in the paper. Future enhancement could try a shorter time window prediction and exploration of the application of deep learning practices such as weight coefficient on large directed and layered graph.

Yin et al., 2021 [17] employed random forest algorithms for stock trend prediction. Twenty technical indicators and feature importance analysis was performed and selected ten as input feature for model training. Original random forest, optimized random forest (by random parameter search) and LGBM model (light gradient boosting model) were trained and results shown that optimized random forest model was more suitable for medium and long-term prediction. Different expanding directions could be considered further such as expansion to stock index instead of listed stock companies, by other novel ensemble learning models, etc.

Argade et al., 2022 [18] did a review on machine learning algorithms in stock market prediction. The paper studied and compared a few methodologies including random forest, support vector machine (SVM) and long short-term memory neural network (LSTM). The application with advantages and limitations of each algorithm were discussed in real-world problems. The paper pointed out that SVM does not perform well when data set has more noise. LSTM is capable to predict future values however is prone to overfitting. Yet, random forest can reduce overfitting problem however required high computational power and resources.

In summary, random forest can predict stock movements in medium to long-term with high accuracy and the model is robust. Technical indicators which are time-series in nature, are considered as the feature input to the model**.** ***As high computational power and resources is needed during model training, LSTM will be explored in the coming section*****.**

LSTM

One of the state-of-the-art deep forecasting models is long short-term memory neural networks (LSTM). LSTMs are extension of recurrent neural networks (RNN), under the artificial neural networks (ANN) family. According to Turing.com, n.d. [19], *"RNNs have a recurrent connection*

in which the output is transmitted back to the RNN neuron rather than only passing it to the next node. Each node in the RNN model functions as a memory cell, continuing calculation and operation implementation. If the network's forecast is inaccurate, the system self-learns and performs backpropagation toward the correct prediction." Below illustrates an RNN block.

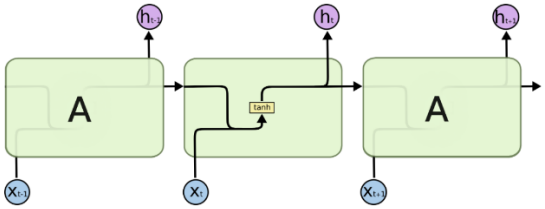
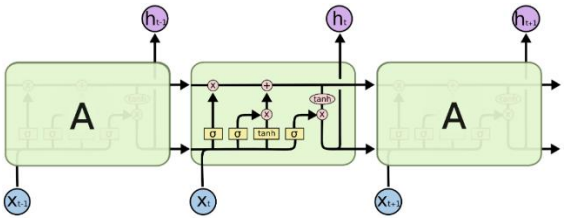
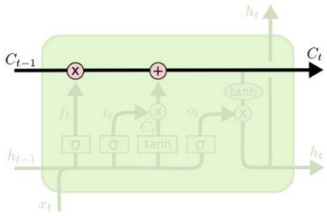


Figure 2: RNN block

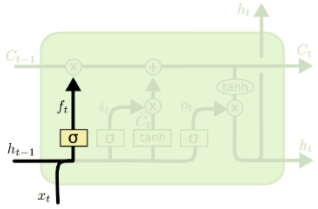
LSTMs in a nutshell, use special type of memory cell and gates, that controlled by activation functions. These gates enable the network to choose what to store and what to forget to facilitate the handling of long-term dependencies in sequential data well and resolve the vanishing gradient problem in RNN. They are good at handling complex time-series forecasting such as stock prediction.

With references to DagsHub Blog, 2023 [20] and ssla.co.uk, n.d. [21], below elaborates the detail of LSTM architecture.



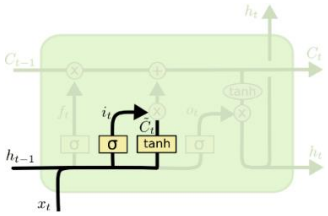


Cell state uses for information transfer throughout the entire model by sigmoid activation function



Forget gate $f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f)$

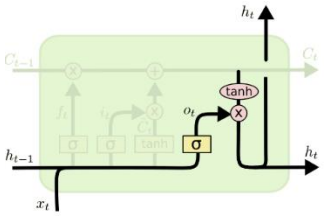
If output value of sigmoid activation function is closer to 0, information will be forgotten where closer to 1 means retain



Input gate $i_t = \sigma(W_i \,.\, [h_{t-1}, x_t] + b_i)$

$$C_t = tanh\,(W_C \,.\, [h_{t-1}, x_t] + b_C)$$

It decides the importance of information to update cell state



Output gate $O_t = \sigma(W_o\,[h_{t-1}, x_t] + b_o)$

$$h_t = o_t * tanh\,(C_t)$$

It regulates the present of hidden state (h_t), that contains information of previous inputs for prediction.

Figure 3: LSTM block and its different states

Number of nodes and hidden layers, number of units in dense layer, number of epochs, batch size, activation functions to be used are examples of hyperparameters that required during LSTM model building.

In one of previous study, Chimmula and Zhang, 2020 [22] *"adopted LSTM for Covid-19 transmission prediction in Canada, aimed to prevent infections and eliminated the pandemic. RMSE used as the measures of the trained model."* The model performed well by limited input data. It proved LSTM is efficient algorithms applying to temporal data.

Vijh et al., 2020 [23] studied the effectiveness between random forest and ANN on stock closing price prediction by five companies of 10 years data. Six derived variables including daily high-low difference, open-close difference, and moving averages by various time periods have been used for model training. Result has been evaluated by RMSE and MAPE with all five stocks have lower RMSE and MAPE by ANN models. ANN proved to be a better technique. For further study, various derived features could be considered and different variants of ANN to fit the time-series nature of stock data.

Thakkar and Chaudhari, 2021 [24] performed a comprehensive review on deep neural network (DNN) algorithms. Two main branches of DNN, which are convolutional neural networks (CNN) and RNNs, together with other DNN have been studied. An overview of each algorithm have been gone through for a conceptual understanding. A literature review summarized the advantages and disadvantages of various methods on recent applications in the stock market (2017-2020), with variates by type of input data, features and target output, and the training and testing details in various researched articles. The study experimented on nine deep learning-based NN models. To facilitate the demonstration of the prediction capabilities of different

DNNs, a common network architecture was adopted. The result indicated that DQN had the highest performance which LSTM demonstrated the capabilities of handling longer dependencies. Yet, the analysis proved the significance of model enhancements and revealed that setting different hyperparameters resulted significant impact.

Chen, Zhou and Dai, 2015 [25] did a case study on stock returns prediction of China stock market by LSTM. Ten features including high, low, open, close, volume of Shanghai and Shenzhen data were used for prediction of stock return categories. The model outperformed random prediction method yet the accuracy was not satisfied, with ranged 15.6% to 26.2%. It was a good demonstration on a preliminary setting of experimental design but believed a lot more following work could be done. Firstly, China stock market is mechanical therefore much more unpredictable. Secondly, more derived variables may be considered during feature selection such as moving averages and other technical indicators. Moreover, other international market indexes, news and sentiments could be considered as features during model training.

Li, Wu and Wang, 2020 [26] combined technical indicators and news sentiments for Hong Kong stock market prediction. 12 HSI constituents in four industries (Commerce, finance, properties and utilities) have been selected for experimenting LSTM with news sentiment prediction model. Results shown that model including news sentiment performed better than only use either technical indicators or news sentiments. The study also analyzed and indicated finance domain-specific news sentiment models was the best among the four selected.

Sentiment analysis is another big topic and it is still in developing stage when applying to prediction model. Balaji, Paul and Saravanan, 2017 [27] and Wankhade, Rao and Kulkarni, 2022 [28] studied on sentiment analysis and posed the challenges when using sentiment analysis depending on the sentiment dictionaries to be used, language and phases to be analysed and the diversity requirement of information to be collected. Considering the maturity of sentiment dictionaries for Hong Kong financial markets, it may not be a robust model to adopt sentiment in Hong Kong stock market predictions.

Istiake Sunny, Maswood and Alharbi, 2020 [29] applied LSTM and bi-directional LSTM model to google stock. The paper BiLSTM, a variant of LSTM which improved the accuracy by utilizing all previous information from both directions however LSTM model took lesser time to predict the data. On the other hand, tunning of hyperparameters resulted higher effectiveness of the model.

Overall, LSTM related papers have been reviewed and criticized at different aspects. LSTM is identified to be one of the most powerful algorithms for temporal and fluctuated data like stock series data. Above literature reviews proved that it could model both price trend and price in an accurate and robust way. It can be variated by incorporating sentiment information and enhanced to bi-directional learning. Temporal of stock prices or adding technical indicators could be variables considered for model training. The disadvantage of it is computing resources dependent that implies high implementation costs.

This paper will deep-dive on LSTM by a practical experiment on applying to Hong Kong stock market and analyse the effectiveness of it.

2.3 Software to be used for this project

From previous literature reviews, it confines the research approach of building logistic regression and LSTM model for trend and price prediction on swing trading. To practice it, suitable software will be used. This section describes features and merits of different software and those will be adopted for this project finally.

R

R language is a statistical software that has evolved from S language since 1990s. R is specially designed for statistical analysis. It is an open-source programming language which possesses a robust ecosystem for machine learning and data mining techniques. It is capable to process massive datasets and have a lot of libraries available for data exploration.

R is a powerful tool that combines the capabilities of complex statistical analysis, machine learning model building and data visualization. Its vast ecosystem of useful libraries makes it an attractive choice for data scientist. Data scientists can perform end-to-end machine learning projects with R in one go. In addition to its support for procedural and object-oriented programming, R can also be integrated with other languages like Java or Python directly during application implementation. This makes it a versatile tool for end-to-end machine learning projects. One of the key advantages of R is its extensive collection of user-created packages that extends R in various aspects. Caret package is one of the famous user-created examples. It consolidates a lot of data model types for classification and regression training.

Angadi and Kulkarni, 2015 [30] studied on stock market prediction using auto-regressive integrated moving average (ARIMA) model by R forecast package. Angadi and Kulkarni opined that *"The R dialect is generally utilized among analysts and data excavators for statistical*

programming and data analysis. Amid the most recent decade, the energy originating from both the scholarly world and industry has lifted the R programming dialect to turn into the most essential tool for computational insights, perception and data science." The study used auto.arima(), a very useful function in R for ARIMA modelling and visualization. The result was reasonably well in short-term prediction.

Sen, 2016 [31] employed R for totally eight methods including multivariate regression, decision tree, bagging, boosting, random forest, artificial neural networks, etc. to build a robust forecasting framework for stock movement prediction. Libraries of glm2, KNN in class, tree, bagging in ipred, bossting in adabag, randomforest, neuralnet, ksvm in kernlab were used. Model results gave accurate prediction on stock movement in a short-term forecasting.

Python

Python is another popular programming language apart from R for data scientists. Information from 6sense.com, n.d. [32], Python contributed 0.95% of market share, ranked the 4th in programming-language market, just behind HTML, PHP and NodeJS. Python is a programming language for general usage with a large and comprehensive standard library.

Python is easy to use with the simplified syntax. It is flexible and compatible with almost any kind of environment. Furthermore, the Python community supports the continuous development of libraries and frameworks. This helps the development of big data, machine learning and cloud computing skill uplift and assists automation and reengineering of organization processes.

Sharma, Modak and Sridar, 2019 [33] leveraged Yahoo finance Python API to pull Google stock data for analysis and stock market price prediction. Keras Python library was used for building LSTM networks. They collected historical information of closing price and trading volume for to stock prices prediction. On the other hand, Matplotlib was used for exploratory visualization. Matplotlib is a comprehensive library for creating various type of visualizations in Python. It allows static, animated and interactive visualizations. Sharma, Modak and Sridar leveraged it for initial graphing of the dataset. Possible future work could be utilizing more on Python's visualization capabilities for plotting different trends and charts for a thorough data understanding and model evaluation and selection.

ProjectPro, 2023 [34] compared between Python and R for data science project. ProjectPro pointed out that Python is *"a general-purpose language, can be used for many different things, such as data science, web development, gaming, and more. Whereas, R is limited to statistics and analysis."* Moreover, Python is better in terms of readability and simplicity in programming

syntax. Python's reliable performance make it popular for data scientists therefore increase its ability to build quality projects.

Ozgur et al., 2021 [35] compared between Python and R and concluded that Python is the best language to learn as it is easy to use open source coding that supported by large community.

Anaconda

Anaconda is a popular open-source distribution founded in 2012. It integrates with Jupyter Notebook, which is one of the most popular interactive development environments for data science and machine learning programming. It has a large community of users and contributors, implies a lot of support for troubleshooting and learning. Furthermore, Anaconda has a package manager and environment manager that housekeeping package versioning, as well as pre-installed common packages like NumPy, Pandas and matplotlib for easy to start with.

Summarizing the above, this project will use Python through Anaconda Jupyter notebook for a graphical interface for coding and debugging.

In this literature research, it performed studies to identify the differences between investment strategies. Modern approaches for stock predictions by different software were critically evaluated. It is observed **that** ***the practical experiment can build logistic regression and LSTM model by the use of Python through Anaconda Jupyter notebook for a graphical interface for coding and debugging.*** Sentiment information may easily lead to bias and the immaturity of sentiment dictionaries for Hong Kong financial markets drove the decision of not putting it in this practical experiment. ***Different mix of historical stock information as the set of features will be experimented*** to identify the highest prediction accuracy and efficiency model to assist optimization of swing trading in the Hong Kong stock market.

3 Background Research

3.1 Hong Kong Stock Market Landscape

The Hong Kong stock market has a long and rich history that dates back to the mid-19th century. In 1861, the first formal securities market in Hong Kong was established with the founding of the Association of Stockbrokers in Hong Kong. At the time, the market was primarily focused on trading shares in foreign companies, particularly those based in the United Kingdom. In 1969, the Hang Seng index was introduced. In the 1980s and 1990s, the market saw a surge in activity as Hong Kong emerged as a major financial hub in Asia.

Before 1997, the Hong Kong stock market was known for its stability and growth. The market was dominated by blue-chip companies such as HSBC, Jardine Matheson, and Swire Pacific, which provided steady returns to investors. The market was also characterized by high liquidity and strong regulatory oversight, which helped to prevent market manipulation and fraud.

In 1997, Hong Kong was reunited with China after more than a century of British colonial rule. The Hong Kong Stock Market continued to thrive under Chinese sovereignty, with the introduction of new regulations and reforms aimed at increasing transparency and improving investor confidence.

One of the biggest changes in the post-1997 Hong Kong stock market was the rise of mainland Chinese companies. China's economy grew rapidly, many Chinese companies began to list on the Hong Kong stock exchange, providing investors with exposure to China's booming economy. This influx of Chinese companies also helped to diversify the market and reduce its reliance on a few large companies.

However, the post-1997 Hong Kong stock market also faced challenges. One of the biggest challenges was the 2008 global financial crisis, which had a significant impact on the market. The market also faced criticism for its lack of transparency and for allowing insider trading and other manipulative practices. Another key challenge was the political unrest and tensions between Hong Kong and mainland China. These challenges have led to increased volatility in the market and uncertainty for investors. However, the market has continued to innovate and adapt, with new listings, new financial products, and new regulations aimed at improving transparency and preventing market manipulation.

Today, the Hong Kong stock market is one of the largest and most active in the world, with a market capitalization of US$4.567 billion [1: HKSFC]. It is home to a wide range of companies, including many of the largest and most well-known corporations in Asia, and is an important hub for international investment and trading.

Overall, the Hong Kong stock market has undergone significant changes over the past several decades, exploratory and technical analysis is performed in the following section to deep dive in long, medium and recent time span of the stock market.

Technical analysis encompasses a wide spectrum of techniques for analysing past price movements. It is believed that patterns that can be depicted from past price movements will infer probable future trends for recommending key buying and selling opportunities, which in short, the history repeats itself.

Below are a few typical patterns indicating the changing of investor's expectations. [36: Mai, n.d.]

Rounding tops and bottoms: A rounding top discerns the expectations of gradually change from bullish to bearish, which a rounding bottom is the turning signal from bearish to bullish.

Double tops and bottoms: A double tops occur when the stock price reaches a high point twice and is unable to breakthrough that level, revealing a possible reversing trend. Conversely for seeing double bottoms patterns for a potential uptrend may occur. These patterns are confirmed on a selling or buying signal when the price breaks below the neckline of the double top or above the neckline of the double bottom.

Head-and-shoulders: This pattern constitutes by a top, a fall, a higher top, then another decline, a move back to the first top and then the bearish formed. The first peak and last peak are known as the "shoulders" while the highest peak is known as "head". The pattern form due to the price cannot breakthrough its highs and reveal a potential bearish market. The reverse of head-and-shoulder conversely meaning a bullish starts.

Apart from the chart pattern, technical analysis has technical indicators. They are used together with the price charts for identifying trends, determining the entry and exit points, confirming price action, as well as setting up stop-loss level to manage risk. Below introduced some common technical indicators.

Moving average (MA): MA is the most popular technical indicator to smooth out price data to make it easier to interpret. The indicator calculates the average price of a stock over a specified

period, say 10, 20, 50, 200 days and then compare then. In general, when a short-term moving average crosses above a long-term one, it typically reveals a buy signal and vice versa.

Moving Average Convergence/Divergence (MACD): It derives from the difference between two exponential moving averages (EMA), usually 12-day and 26-day EMAs. A 9-day average, called the "signal" line, is drawn on the top of MACD. A crossover of the MACD line above the signal line interprets as a bullish signal, indicating that the trend may be shifting upwards. On the contrary, when the MACD line crosses below the signal line, it interprets as a bearish signal, indicating that the trend may be shifting downwards. The MACD also serves as an overbought or oversold indicator. For instance, if the MACD increases sharply, the stock price might be overextending and is likely to revert to more reasonable levels.

Relative Strength Index (RSI): RSI = 100-(100/(1+RS)) where RS is the average gain of up periods / average loss of down periods. It aims to get the internal strength of a stock over a given time period, usually 14 days. An RSI ranged 0-100. Its value above 70 is considered overbought and below 30 is considered oversold.

3.2 Data Source

Selected stocks were Hang Seng Index (^HSI), Alibaba (9988.HK), Tencent (0700.HK) and Meituan (3690.HK) to be deep dived and modelled. These selected stocks were impactful to Hong Kong stock market. Briefly highlighted the background of each of them.

Hang Seng Index (HSI) is a widely recognized market indicator which derived by the weighted of market capitalization of the largest listed companies on the Hong Kong Stock Exchange. Its fluctuations based on various factors, including economic conditions, geopolitical events and company-specific news.

Alibaba Group Holding Limited (9988.HK) was listed on Hong Kong Stock Exchange (HKEX) in November 2019. Alibaba is a Chinese multinational technology company specializing in e-commerce, retail, internet and technology. It is one of the world's largest e-commerce companies, and owns the largest e-commerce platform in China, including Taobao and Tmall. It also operates Alibaba Cloud, a leading cloud computing service provider in China, together with other fintech investment such as Alipay, Ant Group, etc.

Tencent Holding Limited (0700.HK), similar to Alibaba, is another Chinese internet giant. However Tencent focuses on social median and mobile gaming, with its WeChat social media

platform being the flagship product. It also keens on innovative technology development in AI-powered tools for healthcare such as WeDoctor.

Meituan,(3690.HK) is also a Chinese technology company, but operates in the online-to-offline (O2O) space, primarily in the food delivery, hotel booking, and travel sectors.

All the mentioned 3 companies are not only Chinese technology giants, expanded their global footprints and played a major role in the global economy, but also have significant impact to Hong Kong stock market. They added up representing more than 23% of the Hang Seng Index.

3.3 Exploratory Data Analysis

Exploratory data analysis is carried out to understand thoroughly on the data. This section performed summary statistics, visualization data analysis and technical analysis of the studied stocks.

By analysing historical market data, it gives valuable insights into how stocks have performed and this helps to make more informed and profitable investment decisions. In this section, 5+ years till few months for long, medium and recent time span will be explored on the stock trend.

The analysis started with the pre global financial crisis of 2007 and the China economic took off period. Summary statistics gave an overall understanding of the data by using <dataframe>.describe() in pandas library, with reference to figure 4 of maximum available data from 1Jan2007 to 31Mar2023 have been extracted and summarized.

	9988HKOpen	9988HKHigh	9988HKLow	9988HKClose	9988HKAdjClose	9988HKVolume	3690HKOpen	3690HKHigh	3690HKLow	3690HKClose	3690HKAdjClose	3690HKVolume
count	825	825	825	825	825	825	1,115	1,115	1,115	1,115	1,115	1,115
mean	169	171	167	169	169	38,856,106	170	174	166	170	170	25,998,119
std	66	67	66	66	66	25,858,558	91	93	89	91	91	19,149,379
min	61	65	60	61	61	-	41	43	40	41	41	-
25%	103	105	102	104	104	21,245,436	92	94	90	91	91	14,391,622
50%	184	186	178	183	183	31,609,172	166	171	161	166	166	21,861,746
75%	221	223	219	221	221	48,345,392	240	247	236	242	242	31,640,643
max	304	309	304	307	307	187,288,574	455	460	443	451	451	242,615,551

	0700HKOpen	0700HKHigh	0700HKLow	0700HKClose	0700HKAdjClose	0700HKVolume	HSIOpen	HSIHigh	HSILow	HSIClose	HSIAdjClose	HSIVolume
count	4,008	4,008	4,008	4,008	4,008	4,008	4,008	4,008	4,008	4,008	4,008	4,008
mean	193	196	191	193	190	22,708,682	23,255	23,399	23,075	23,240	23,240	1,932,254,972
std	181	184	179	181	179	14,547,261	3,688	3,686	3,680	3,682	3,682	776,518,248
min	5	5	2	5	5	-	11,155	11,747	10,676	11,016	11,016	-
25%	34	35	34	34	33	14,580,029	20,864	21,004	20,727	20,865	20,865	1,431,896,250
50%	131	133	130	131	128	19,340,028	23,046	23,153	22,881	23,029	23,029	1,763,589,900
75%	343	346	338	342	337	26,635,054	25,720	25,847	25,550	25,701	25,701	2,226,688,725
max	767	776	757	767	755	308,436,765	33,335	33,484	32,897	33,154	33,154	9,799,120,000

Figure 4: Summary statistics

It is observed that Alibaba have 825 records, due to the listed on HKEX in November 2019, so its performance in the Hong Kong stock market can only be available from that point onwards. Simple statistics summary provides mean and standard deviations of each selected stock. The means of the above illustrated the averages of each column, gave a general idea of how much the stock on average across the defined timeframe. Average of HSI over the past 16 years was

closed at 23,240 whereas the adjusted close prices were similar at 170 to 191 for the 3 selected stocks with Tencent was slightly higher. Average volumes were also available here.

Standard deviation, also known as volatility that measured how much the stock's price has varied from its average price over the defined time period. The larger the volatility, the more fluctuations of the stocks, revealed a higher risk. In the above statistics, adjusted close price of Tencent had the highest standard deviation, 180 among the selected 3 stocks, followed by Meituan and Alibaba with 91 and 66 respectively. All those selected stocks were China giant companies that kept rising since listed, with the larger spread of minimum and maximum price, it explained the huge difference of standard deviations.

While summary statistics provided a concise summary of the data of the defined time period, it hasn't covered on the shape and distributions of the data, by different time periods. Therefore, trend analysis has been performed to examine the patterns and relationships associated with economics events.

Stock price trends with their respective moving averages comparing with its volume have been plot on the same axis by matplotlib. While there was curious on repeatedly generated the charts by different stocks and period. A function "plot_stock_trend" was written to ease the programming operation. Detail codes attached in appendix separately.

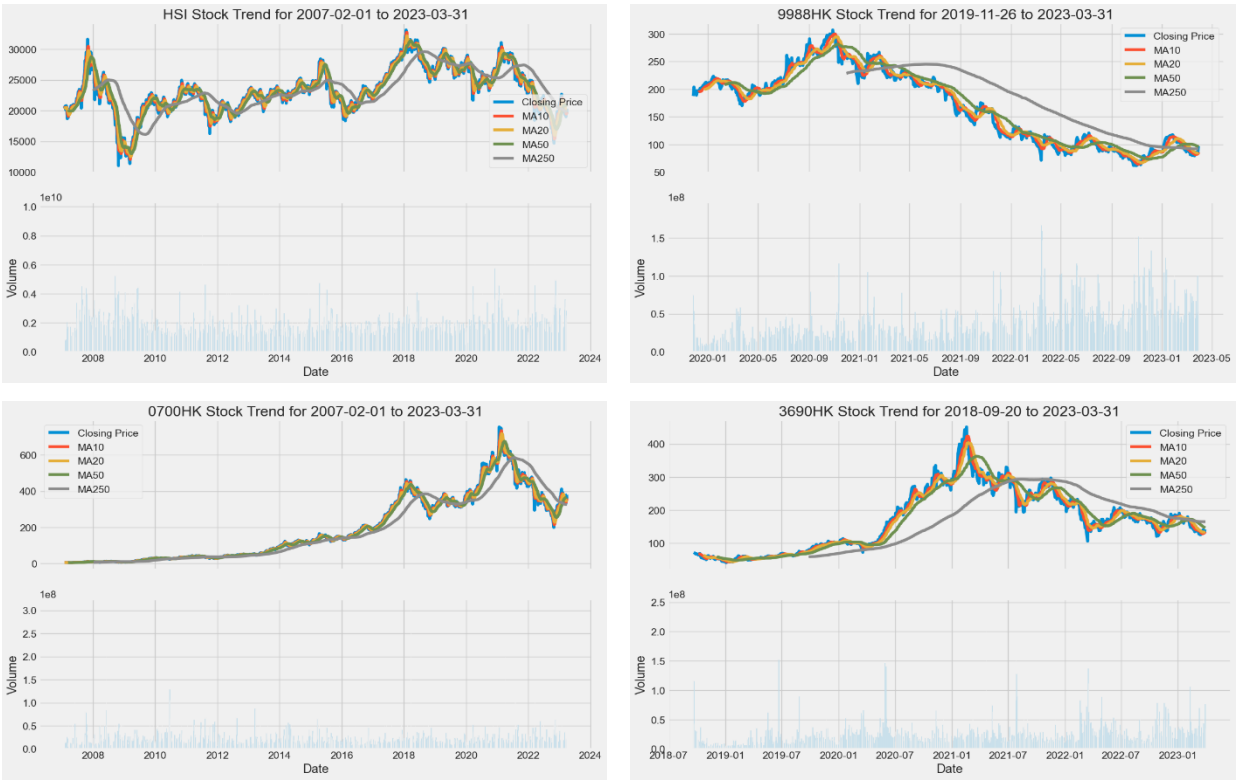


Figure 5: Long term price trend with moving averages and volume

Figure 5 showed the price trend from 2007 to 2023, it provided a long-term view of how the prices have changed over time. For HSI, there was a few waves of ups and downs, with a slightly upward trend between 20000 to 27000. It aligned with the important events aforementioned such as the global financial crisis in 2008-2009, the recent US-China trade war and Covid pandemic during 2018-2019 and 2020-2022 respectively. Referring to the price trend of Alibaba, since it began its international expansion as early as in 1999 and listed in US well before in Hong Kong. It is observed a different shape as the other two given Alibaba has faced to regulatory scrutiny specific to her. Tencent and Meituan observed a significant climbing trend at first until 2021 then following the macro-economic trending down due to US-China trade relationship tension and various global economy factors such as inflation concerns and Covid pandemic.

Zooming into some recent trends can provide a more short-term view of how the price has changed lately. For instance, 2-year time span can be used to analyse the fluctuations during the downtrend period and to identify reference record highs / lows during investment decisioning. While here focus on swing trading decision making, six-months period shown in figure 6 are more appropriate for decision analysis.

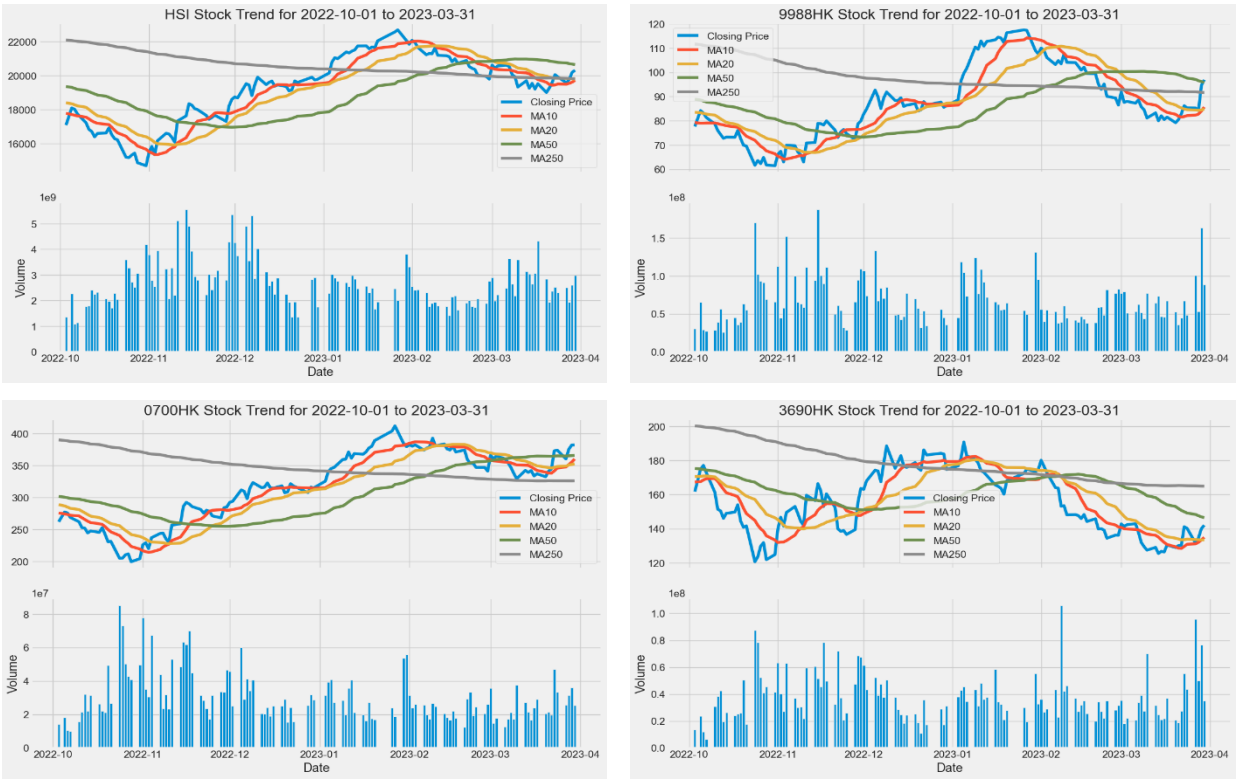


Figure 6: Short term price trend with moving averages and volume

In general, all the selected stocks saw a rebounded during Nov 2022. The price or short-term moving average observed to breakthrough a longer term moving averages. According to

Mitchell C., 2019 [37], it is a confirmed increasing trend with this pattern. The reason of increasing because the macroeconomic factor related to China reopening, lockdown was lifted and China economics resumed. However, lately in Mar2023, affected by 2023 banking crisis, three small- to mid-size US banks failed. This triggered the shrink in global stock markets, as well as Hong Kong stock market. However, according to South China Morning Post, 2023[38] quoted from Hong Kong SFC, *"the US banks' collapse and Credit Suisse takeover are 'low-exposure' events that spare the world from a repeat of 2008."*

Histograms of daily returns have been plotted to examine the distribution of daily return over past six-months period crossing the rising moments and past three-months period during the banking crisis period in figure 7.

It was noticed that they are all in bell shape, in which the return of HSI was in a very balanced shape centered at 0 with small oscillations between +/- 5%. The other 3 stocks were in general have a higher risk but positive return in conclusion in particular Alibaba which had a few noticeable spikes on the left side of the bell that there was some concentrations of returns in that range however the appearance of longer right tail revealed a positive return.

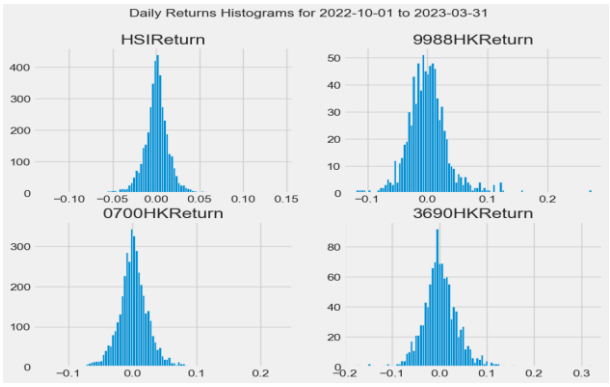


Figure 7: Histograms of daily returns

Apart from individual stock analysis, PairGrid and correlation matrix have performed to examine the bivariate distribution on daily returns by seaborn library. It is an extension of matplotlib that provides built-in visualization functions by a simple function like sns.PairGrid and sns.heatmap with the output in figure 8.

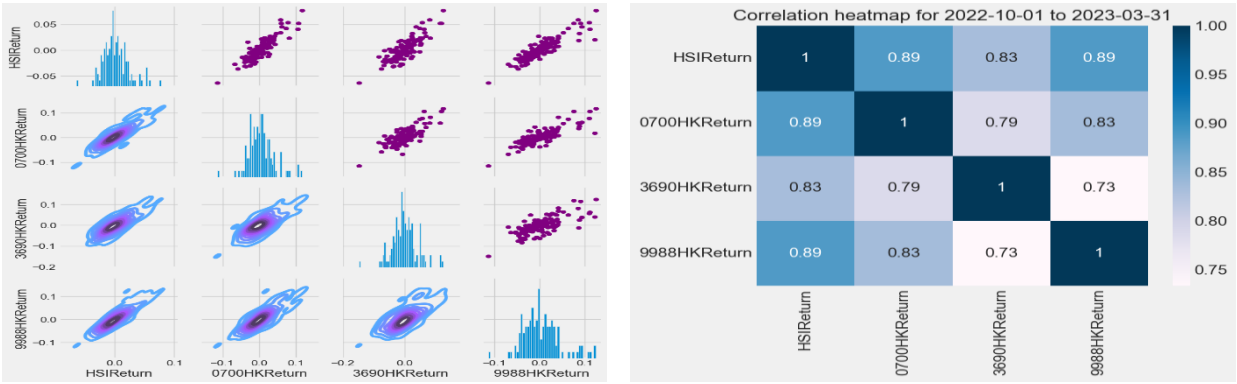


Figure 8: Bivariate plots and correlation matrix

No doubt with the fact that the three selected stocks were representing a large portion of Hang Seng Index therefore it is observed a strong linear relationship, illustrated in PairGrid on first column or first row for kdeplot and scatter plot respectively. Correlation coefficients were between 0.83 to 8.89, indicating the strong positive relationship. Although all three companies were Chinese giant technology companies, it is observed Meituan has a smaller relationship with other two.

Figure 9 denoted the risk and return analysis by a two-dimension plots on the mean and standard deviation of daily returns over past six months. Meituan was at top left quadrant, revealed that it was at high risk but negative return, that investors exposed to significant losses. In contrary, Alibaba and Tencent although posed a risk above 3%, they have positive expected return. HSI was a combination of different stock performance, resulted lower return with lower risk. It was curious the latest change of risk and return and the same plot performed using latest 3 months data only. It is observed the result was very similar with the overall risks of stocks was slightly lowered. It revealed a quiet market and lack of momentum to drive a confirmed bearish or bullish market direction.

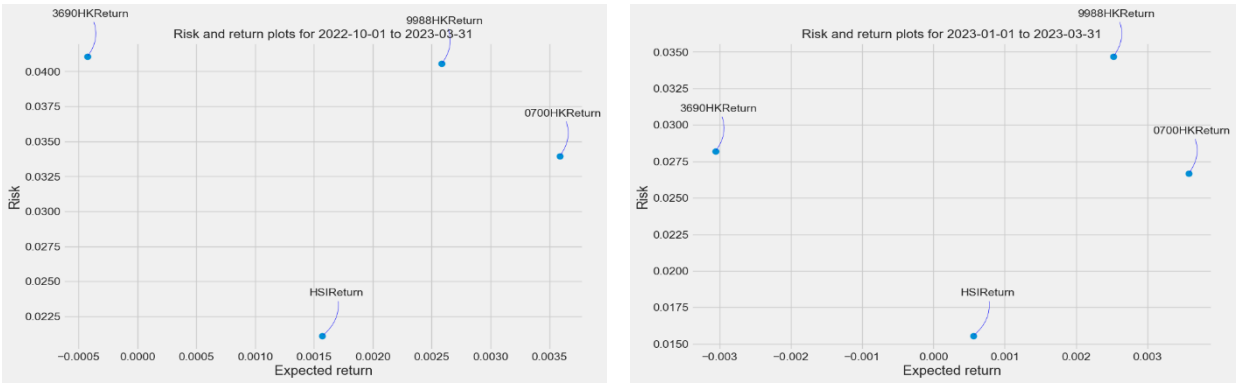


Figure 9: Risk and return analysis

The background landscape study helped understand the economic environment and technical analysis gave a holistic view of long, medium and in particular the lately situation of selected stocks. In summary, the market was volatile affected by various global factors and company specific events.

4 Practical Experiment

4.1 Project Management

Key sections of this project were problem identification, literature review, background research, technical analysis and practical experiment. Total project duration had set to be 10 months and effective project management process was vital to the success deliverables on time. Lucid Content Team, 2019 [39] opined *"the Agile development method has been the methodology of choice for today's product development team"*. While the project has clear objective and steps to achieve, the simple and easily understand project management process of waterfall was still beneficial. Therefore, this project managed by agile-waterfall hybrid approach. The waterfall approach is a sequential approach that each stage is completed before moving on to the next stage. This approach is useful when the project requirements are well-defined and the scope of the project is fixed. The waterfall approach was used for the initial stages of the project, such as problem identification, data collection, data preprocessing… These stages were completed in a sequential manner, with each stage being completed before moving on to the next stage.

However, within the model development stage during practical experiment, agile was adopted. The model development stage involved building and training the machine learning models, as well as evaluating their performance. These stages required more flexibility and adaptability, as the results from interim steps could inform subsequent steps and they were iterative and collaborative. The agile approach adopted to break down the model development stage into a few sprints with specific set of objectives and deliverables. It was better to handle more uncertainty and allowed for a more efficient and effective project management process, while still ensuring that the project objectives were met.

Wilson, 2003 [40] stated *"Gantt charts provide a quick and easily understood means for describing project."* A Gantt chart is a type of bar chart that illustrates a project schedule, showing the start and end dates of each task or activity. It provides a visual representation of the project timeline, duration and dependencies between tasks. It helps to understand the progress easily and enable easy identification of potential issues during project management. Gantt chart was used for project progress tracking to facilitate the display of important information during project management.

4.2 Experiment Design

To assist optimization of swing trading investment strategy in the Hong Kong stock market, logistic regression and LSTM model for trend and price prediction were developed. Figure 10 illustrated the experiment design work flow.

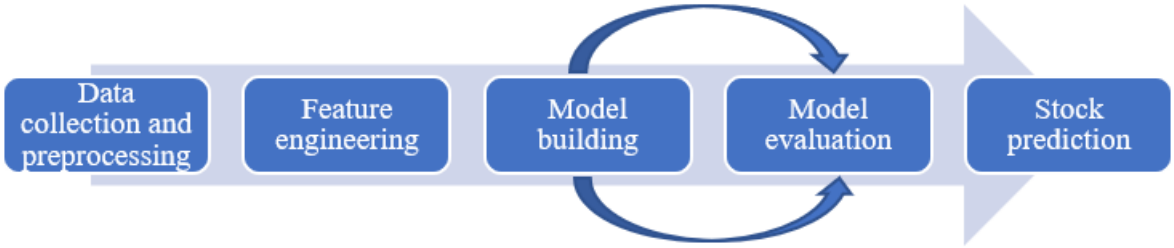


Figure 10: Methodology

It involved several critical steps, including data collection and preprocessing, feature engineering, model building, training, evaluation and selection. After training and evaluation, the models were selected based on their performance. Lastly, the selected models were used to make real data predictions. The predictions were then critically evaluated to determine the effectiveness of the different algorithms for stock predictions. The evaluation included an analysis of the strengths and weakness of each model, and recommendations for future research. The result visualized by appropriate graphs and charts to make them more understandable and appealing to readers.

4.3 Data Collection and Preprocessing

Data collection is the first step in any machine learning project. To obtain historical stock information for model development, yfinance Python library is adopted in this project. yfinance is an API that connects Yahoo Finance data to Python, which uses web scraping techniques to extract data from Yahoo Finance website and then formats the data into a pandas DataFrame. Ticker symbol and timeframe are the required input parameters, then a two-level-column dataframe are formed with information including ticker symbol, date, open, close, high, low, volume on daily basis for the date from 1Jan2007 to 31Mar2023, starting from pre global financial crisis of 2007 and the China economic took off period.

After downloading the data, two-level columns resulted in the downloaded dataframe with company name and the downloaded features of each ticker. Some preprocessing steps were performed to flatten the column name and standardize a naming convention with date as the index to facilitate subsequent model building steps.

On the other hand, missing values existed in any of the columns were checked. Handling missing values is an important step to maximize the use of data during model building. Imputation, deletion, interpolation and deep learning-based imputation are some common approaches. In this project, the missing values were mainly due to data unavailable before company listed. Therefore, deletion of dates before company listed was processed to maintain a valid data frame for model building. In particular for HSI, information in a few trading dates was missing. Considering the temporal nature, close values were probably most suitable to infer the missed value [41: KDnuggets. (n.d.)] and value of previous day had been imputed to the missed values.

4.4 Feature Engineering

Feature engineering is the process of creating or transforming features to improve the performance of machine learning models. It aimed for improving prediction performance, providing faster and more cost-effective predictors and better understanding of the underlying process that generated the data. New variables including moving averages, RSI and MACD have been derived for technical analysis as well as putting them into model building. All the variables were in different scaled that required to align to the same scale. Min-max scaling normalization process adopted to calibrate the feature range between 0 to 1. It preserved the relative distances to improve the performance of models like logistics regression and neural networks. [42: Adeyemo and Wimmer, 2018]

Cited from Waqar et al., 2017 [43], *"principal component analysis (PCA) can help to improve the predictive performance of machine learning methods while reducing the redundancy among the data."* Putting excessive features may complicate the model building steps or create overfitting problem, especially in the base line logistic regression model. PCA has performed for dimension reduction to facilitate effective model building.

The features put into PCA were adjusted close, open, high, low, volume, moving averages, RSI, MACD. Firstly, each feature went through min-max scaling to have a mean of 0 and a standard deviation of 1. This is important because PCA is sensitive to the scale of the features. Standardization ensured that all variables have equal importance. Secondly, computation of covariance, eigenvectors and eigenvalues were processed by the function of PCA(n_components=0.95). This is a Python script from the scikit-learn library. The n_components parameter specified the number of principal components to extract from the data. 0.95 indicated the PCA result can retain enough principal components to explain 95% of the

variance in the data. It calculated the loading of each feature as the ground to shortlist the features for model building.

Table 2 summarized the heavy loading features of first principal components of each stock as a starting of logistic regression model building. Open, high, low, moving averages (10-days, 20-days, 50-days and 250-days) were features with relatively higher loadings, indicating that these features were important for the component. This component can be interpreted as a measure of the overall level of the stock prices. And as an extended step, those features have been put into LSTM model to see if a larger set of features can improve the model accuracy.

		Explained	Principal components											
Stock	PCA	variance ratios	Adjusted Close	Volume	Open	High	Low	MA10	MA20	MA50	MA250	RSI	MACD	MACD Signal	MACD Histogram
HSI	1	67.42%	-0.33	0.01	-0.33	-0.34	-0.33	-0.36	-0.37	-0.38	-0.37	-0.03	-0.03	-0.04	0.01
HSI	2	19.27%	-0.09	0.03	-0.08	-0.08	-0.09	-0.04	0.03	0.13	0.31	-0.61	-0.48	-0.45	-0.24
HSI	3	7.49%	-0.05	0.04	-0.06	-0.05	-0.06	-0.12	-0.15	-0.01	0.51	0.47	-0.15	-0.38	0.55
HSI	4	3.14%	0.10	-0.14	0.10	0.10	0.11	0.06	0.09	0.25	-0.70	0.12	-0.32	-0.45	0.24
Alibaba	1	78.78%	0.33	-0.09	0.34	0.34	0.34	0.35	0.35	0.36	0.40	0.00	-0.06	-0.07	0.01
Alibaba	2	12.45%	0.11	0.03	0.11	0.11	0.11	0.07	0.00	-0.07	-0.20	0.45	0.59	0.53	0.25
Alibaba	3	4.63%	0.01	-0.07	0.02	0.01	0.02	0.13	0.15	0.01	-0.23	-0.48	0.12	0.48	-0.65
Tencent	1	88.42%	0.33	0.00	0.33	0.33	0.33	0.34	0.35	0.37	0.41	-0.03	0.003	0.005	-0.003
Tencent	2	0.07%	-0.07	0.04	-0.07	-0.07	-0.07	-0.04	0.00	0.05	0.15	-0.84	-0.36	-0.32	-0.17
Meituan	1	68.34%	0.32	-0.01	0.32	0.32	0.32	0.35	0.37	0.41	0.40	-0.04	0.003	0.01	-0.01
Meituan	2	19.71%	0.16	-0.05	0.15	0.15	0.15	0.12	0.05	-0.09	-0.51	0.43	0.45	0.47	0.11
Meituan	3	6.51%	0.02	0.00	0.04	0.03	0.02	0.12	0.20	0.14	-0.59	-0.57	-0.14	0.06	-0.47
Meituan	4	3.19%	0.08	-0.01	0.06	0.07	0.07	-0.01	0.03	0.29	-0.47	0.32	-0.37	-0.58	0.32

Table 2: PCA analysis results

4.5 Model Building and Evaluation

To perform the experiment, logistic regression and LSTM models were developed according to the goals defined in each sprint of the experiment design.

First sprint focused on logistic regression model training:

- Features - Logistic regression model of each stock trained, by putting different features into the model for comparison, and select the best feature set, taking into consideration of PCA analysis result.
- Data period used - As the recommendation focused on swing trade, which the trading period normally from few days to few weeks. Maximum two years of information captured the recent waves that helped the model to focus on recent market conditions.
- Target setting - Logistics regression is a classification algorithm and the target set to be binary indicator of increase or decrease from previous daily adjusted close. If price decreased, then it set to "True" and price increase was set to "False".
- Train / test split - References to the train / test split arrangement that Li, Wu and Wang, 2020 employed [26], the time window approach adopted for train / test split around 80% / 20% were used due to the temporal nature of the data.

- Performance metrics - To measure the effectiveness of the model, confusion matrix was the evaluation metric.

  Confusion matrix is a table summarizes the model performance by comparing the actual and predicted classes. Accuracy, precision, recall and specificity can then be derived.

		Predicted	
		FALSE	TRUE
Actual	FALSE	TN	FP
	TRUE	FN	TP

  - Accuracy measures model correctness: TN / (TN + TP + FN + FP)
  - Precision measures how many positives among predicted positives: TP / (TP + FP)
  - Recall measures how many positives among actual positives: TP / (TP + FN)
  - Specificity measures how many negatives among actual negatives TN / (TN + FP)

  In this project, the model result of predicting increase or decrease (ie. False or True) were equally important as long as it predicted correctly to assist informed investment decision. Accuracy was the key metrics to be examined.

<u>Second sprint focused on LSTM model training:</u>

- Features - LSTM model of each stock trained, 60-days of historical adjusted close series constructed the feature set.
- Data period used - LSTM allowed the model to capture more complex patterns and dependencies in the time series data. The forget gate mechanism in LSTM can help prevent the model from being overwhelmed by irrelevant or noisy information in the input sequence, which can make it more robust and accurate. The longest data period available was used.
- Target setting - LSTM able to predict the trend or the price. Next day adjusted close set to be the target. This gave more information with both the price and directions would be available.
- Train / test split - Same as logistic regression model, roughly 80% / 20% train / test split ratio were used.
- Performance metrics – Referenced to Ahuja et al., 2023 [44], error measuring techniques have been adopted for regression model evaluations.
  - Mean squared error (MSE) measures the differences between actual and predicted values on average. This metric is sensitive to outliers but the unit of the result has no direct meaning as it has been squared

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{zi} + z'i)^2$$

- Root mean squared error (RMSE), similar to MSE but square rooted on top of MSE to align the unit to original data. This is the most commonly used metrics for regression problems.

  $RMSE = \sqrt{(\frac{1}{N}\sum_{i=1}^{N}(\mathrm{yi} + \dot{\mathrm{y}}i)^2)}$

- Mean absolute error (MAE) measures the absolute differences instead of squared the average differences. It does not square the errors and therefore it is not as sensitive to outliers as RMSE

  $MAE = \frac{1}{N}\sum_{i=1}^{N}|\mathrm{zi} + z'i|$

- Mean absolute percentage error (MAPE) measures the absolute percentage differences where focuses on the relative size of errors.

  $MAPE = \frac{1}{N}\sum_{i=1}^{N}|\frac{\mathrm{zi} + z'i}{\mathrm{zi}}|$

  As errors represented the risks of adopting the model for assisting investment decisioning, the magnitude of errors was really a concern, RMSE would be the most important metrics for model evaluation, supported by MAE and MAPE.

- Hyperparameter used - Hyperparameter set on two LSTM layer with 30 and 15 neurons with one dense layer of single output. Epoch set to be 100 with batch size was 50 adopted as the first universal hyperparameter setting for model evaluation.

<u>Third sprint analysed further improvement on LSTM model by different feature sets and hyperparameters</u>:

During the third sprint of this experiment, it strived for further improvements to the LSTM model by expanding the feature set. The goal was to investigate whether adding more variables would result in better model performance. The variables that were selected from PCA analysis were included in the LSTM training to assess if the model performance would outperform the model that only used the historical adjusted close series from sprint 2. The performance measures were then compared to identify the better performing set of LSTM models. Once the better feature set have been identified, hyperparameters were studied to fine tune the model accuracies.

While the error measures within the same stock could compare due to same magnitude, it was difficult to evaluate if the error measures were smaller enough to say the model was good or not in general or compared across different stocks. Pointed out by Otto, S. 2019 [45], *"the coefficient of variation can be represented by the variance of each group standardized by its group mean."* The normalized RMSEs then represented how much variations (risk) if using the model that

drove for model improvement depended on risk acceptance. On the same time, facilitated the comparison of model goodness between models across different stocks. Therefore normalised RMSE was used as the key component of judging a model with suitable hyperparameters when the figure lower than 5%, a risk controlled to be lower than the average return of around 10% from stock market in general.

4.6 Stock Prediction

The selected models, logistic regression and LSTM, were applied to real-time data to evaluate their effectiveness in predicting stock trend and prices. The measurement of accuracy and errors of the models were the key performance measures used to assess their performance.

The real-time data between 1Apr2023 – 30Apr2023 was pre-processed using the same steps as in the model development stage. The pre-processed data was then fed into the models to generate predictions. The results of the stock predictions were compared to the actual stock prices to evaluate the model's effectiveness. Results were plotted in charts to facilitate easy understanding of the performance of the models that ease for comparison and insight identification.

5 Analysis of Results

5.1 Model Results

This section presents the results of the machine learning models developed. The objective of the project was to develop models for predicting stock prices using historical stock price data and financial indicators. Two machine learning models were developed and evaluated: logistic regression and LSTM.

The logistic regression model was used for trend prediction, and the LSTM model was developed as an advanced machine learning technique for price prediction. Both models were trained and tested using historical stock price data and financial indicators. The performance of the models was evaluated using various evaluation metrics, including accuracy, precision, recall, specificity and error metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). The results of the model development and evaluation process are presented in this section. The performance of each model is evaluated and compared to determine which model is more effective in predicting stock prices. The section also includes a discussion of the key findings and implications of the results.

Logistic Regression

In order to build a logistic regression model for predicting the stock trend, defined into increase or decrease compared with previous day adjusted close price, two sets of features were used. The first set included all variable features in the dataset, while the second set consisted of the features recommended by the PCA result. In theory, PCA helped on further reducing the dimensionality while performance improved or maintained. It aimed to see if it practically applied.

Both models were trained and tested on the same two year span, the performance of each model was evaluated.

Figure 11 illustrated the accuracies from each of the feature sets for different stocks. It was observed that PCA feature sets of HSI and Tencent resulted a higher accuracy than the entire feature sets with all variables. Alibaba with the same accuracy for both models, therefore, the one by PCA feature set adopted based on selecting the simplest model with same performance.

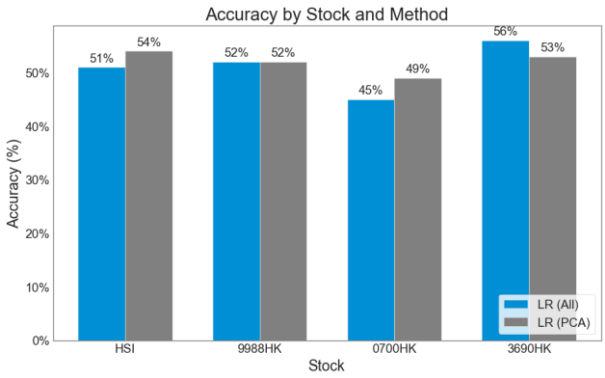


Figure 11: Accuracy of Logistic regression model

However, for Meituan, was in another way round with higher accuracy resulted from the entire variable set. It may due to the first PCA result from Meituan was not at the confirmed high side. Explained ratio was 68%.

The models with higher accuracy of each stock were selected for real time data predictions.

LSTM

LSTM is widely used for stock prediction due to its ability to model complex temporal relationships and handle long-term dependencies in sequential financial data. Similar to logistic regression, two feature set of LSTM model, one using the 60 consecutive daily sequential adjusted close data; with another one by the PCA feature sets, to predict the target of next day adjusted close price.

To compare the performance of using different feature sets, same set of hyperparameters have been adopted. It has been set to two LSTM layers with one dense layer for batch size to 50, and number of epoch to 100.

The models were trained and tested by 80 / 20 split, the performance of each model was evaluated.

	HSI		Alibaba (9988)		Tencent (0700)		Meituan (3690)	
	Adj. Close	PCA	Adj. Close	PCA	Adj. Close	PCA	Adj. Close	PCA
RMSE	373	396	7	23	16	19	9	10
MAE	282	299	6	22	12	15	8	9
MAPE	121%	132%	677%	2171%	329%	398%	477%	562%

Table 3: LSTM model performance metrics

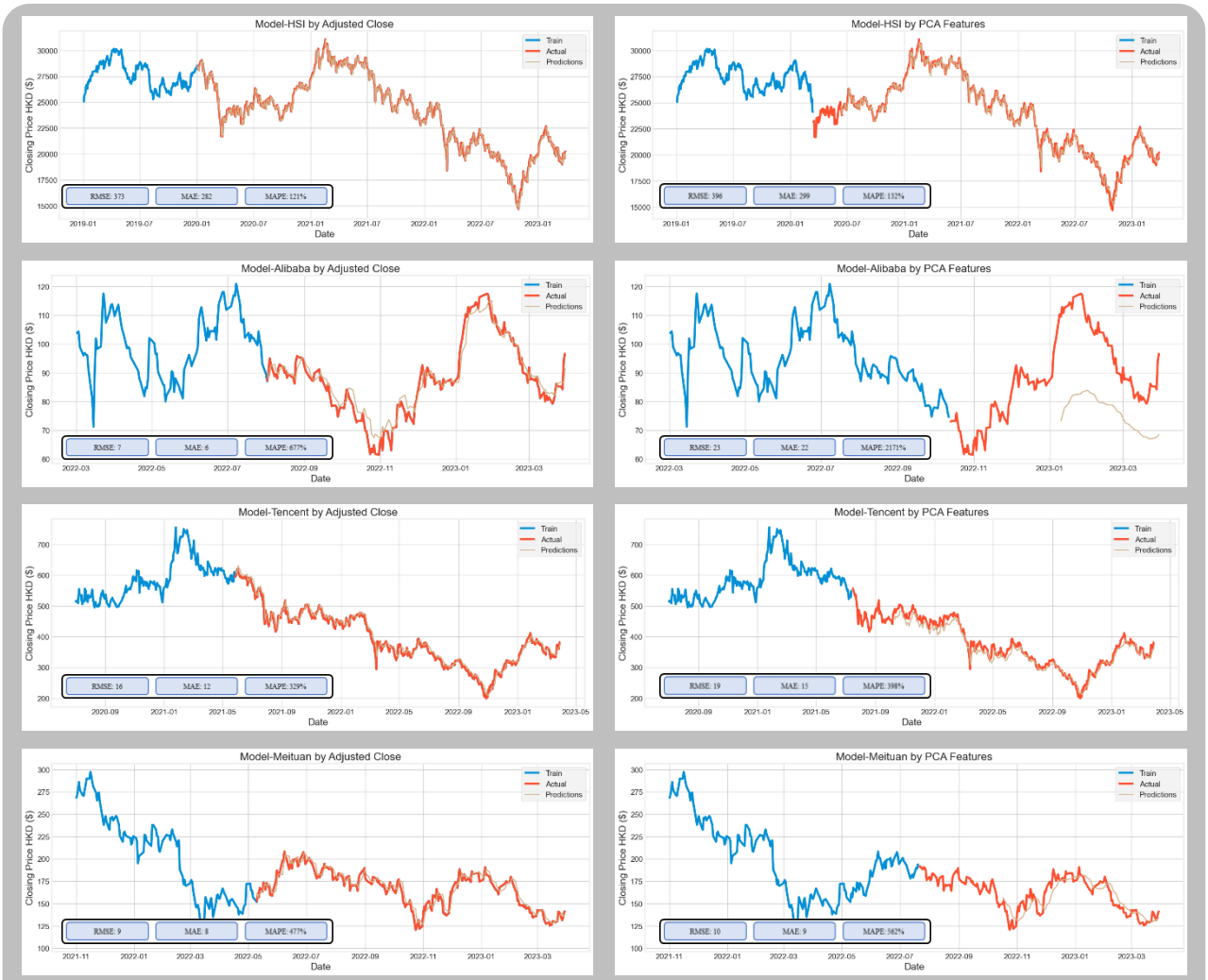


Figure 12: LSTM model result

Table 3 summarized the performance metrics. It denoted that all error metrics of RMSE, MAE and MAPE were smaller for models trained by adjusted close price performed in all the selected stocks. These have been further visualized the prediction results in figure 12, to see how the model performed between actual and predicted values. In each chart, blue lines represented the actual data for model training, red lines were the actual data during testing period while the tan lines were the predicted values came from model fitting. Charts on left-hand-side were models trained by adjusted close price only whereas the right-hand-side were model trained by PCA feature sets. The closer between red and tan lines, the higher predictive power of the model. It was observed that models trained from adjusted close price only were all outperformed with smaller gaps between actual and predicted.

The normalized RMSE over its average adjusted close prices in past 6 months of the testing period were examined in Table 4. It was denoted that the model built for HSI have the lowest percentage value, translating to the predicted value was around 2% deviations to the average close price on average.

	RMSE	Avg. Close in Recent 6 months	Normalized RMSE
HSI	373	19,167	1.9%
Alibaba	7	87	7.9%
Tencent	16	316	5.2%
Meituan	9	155	6.0%

Table 4: Normalized RMSE for selected models

On the contrary, Alibaba have the largest normalized RMSE which the prediction varied from recent average close price by around 8%, which was not a very ideal while using this for assisting investment decisions. Cited from Harsmi, 2020 [46], tuning hyperparameters can help achieve better performance. In this case, further model fine tuning was done to the selected feature set. *The purpose was to tune the model to have less than 5% normalized RMSE*.

The model tuning focused on four key hyperparameters with its usage:

Number of layers and neurons in each layers determine the complexity of the model for learning complex patterns of the data where number of layers more focus on learning hierarchical representations.

Batch size determines the number of samples that are processed at once during training and number of epochs determines the number of times the model iterates over the entire training datasets. It is used to compute the gradient of the loss function and update the model's weights to facilitate model performance improvements.

To fine tune the hyperparameters, manual search approach has adopted. The idea was to tune the hyperparameter one by one, with the value adjusted around a specific value, compare the RMSE to hint the direction of next trial until the desirable normalized RMSE resulted. In the above scenario, Alibaba, Tencent and Meituan were the target to have hyperparameters fine tune illustrated in table 5 below.

Alibaba	Number of layers	Number of Neurons	Batch size	Number of epoch	RMSE	Normalized RMSE
0	2	30 / 15	50	100	7	7.9%
1	2	60 / 15	50	100	8	8.8%
2	2	60 / 15	50	150	7	7.9%
3	2	60 / 15	100	150	12	13.5%
4	2	60 / 15	30	150	4	4.2%
5	3	60 / 30 / 15	30	150	15	16.8%

Tencent	Number of layers	Number of Neurons	Batch size	Number of epoch	RMSE	Normalized RMSE
0	2	30 / 15	50	100	16	5.2%
1	2	60 / 15	50	100	15	4.7%

Meituan	Number of layers	Number of Neurons	Batch size	Number of epoch	RMSE	Normalized RMSE
0	2	30 / 15	50	100	9	6.0%
1	2	60 / 15	50	100	8	5.0%
2	2	60 / 15	50	150	10	6.3%
3	2	90 / 15	50	150	7	4.3%

Table 5: Hyperparameter tunning result

5.2 Real-time Results

Real-time prediction is a crucial aspect of stock market analysis and trading. In today's fast-paced world, traders need to make quick decisions based on the latest market trends and price movements. This is where real-time prediction comes in handy. With the best performing models selected in the previous section for each stock, the next step is to apply these models to real-time data to reinforce the robustness and demonstrate their capability to predict future outcomes accurately and efficiently.

To achieve this, the April 2023 data with 17 trade days was used as the real-time data. Similar data download and preprocessing steps applied to the new data, ensuring that the data was consistent and in the same format as the training data. Once the data was preprocessed, the model.predict() function were applied to the new data frames, and performance metrics were derived to evaluate the model's robustness and performance.

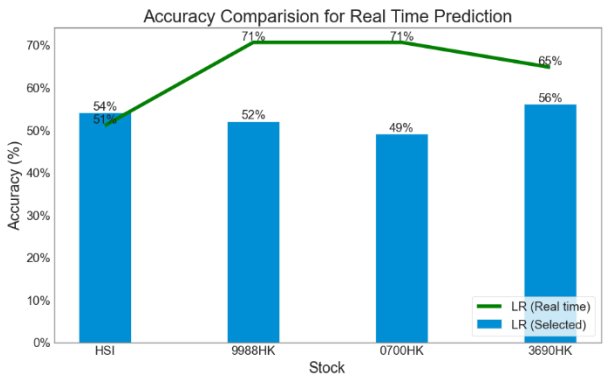


Figure 13: Accuracy comparison for real time prediction of logistic regression model

First of all, logistic regression for trend prediction have been applied, accuracies of each stock by the selected model shown in figure 13. The blue columns represented the accuracies of selected models for each stock where the green lines indicating the same accuracy results from real time prediction compared to the real time actual data. The line and bar chart depicted that most of the stock accuracies with the real time predictions were maintained or even better than the accuracies during model training.

On the other hand, LSTM for price prediction also performed, due to the randomized model intialization and stochastic nature of optimization, the LSTM model result varied slightly when processed multiple times. Therefore the LSTM models with identified hyperparameters were saved by model.save() function provided by keras library. This helped to ensure the same trained model applied to the real time data by load_model() function.

	HSI		Alibaba (9988)		Tencent (0700)		Meituan (3690)	
	Selected	Real time	Selected	Real time	Selected	Real time	Selected	Real time
RMSE	373	182	4	4	15	12	7	4
MAE	282	150	3	2	12	9	5	4
MAPE	1	1	3	2	3	3	3	3
Normalized RMSE	1.9%	1.0%	4.2%	4.1%	4.7%	3.7%	4.3%	3%

Table 6: LSTM real time predictions performance metrics

Table 6 tabulated the error metrics with the normalized RMSE comparing between selected model test set performance and the real time prediction performance. The error metrics of the real time predictions had a further improvement from the model test result. In particular, the normalized RMSE were maintained around 4% or below, it revealed the variances (risk) were less than 4% to the average price of the stocks when using the model for price prediction during investment decisioning.

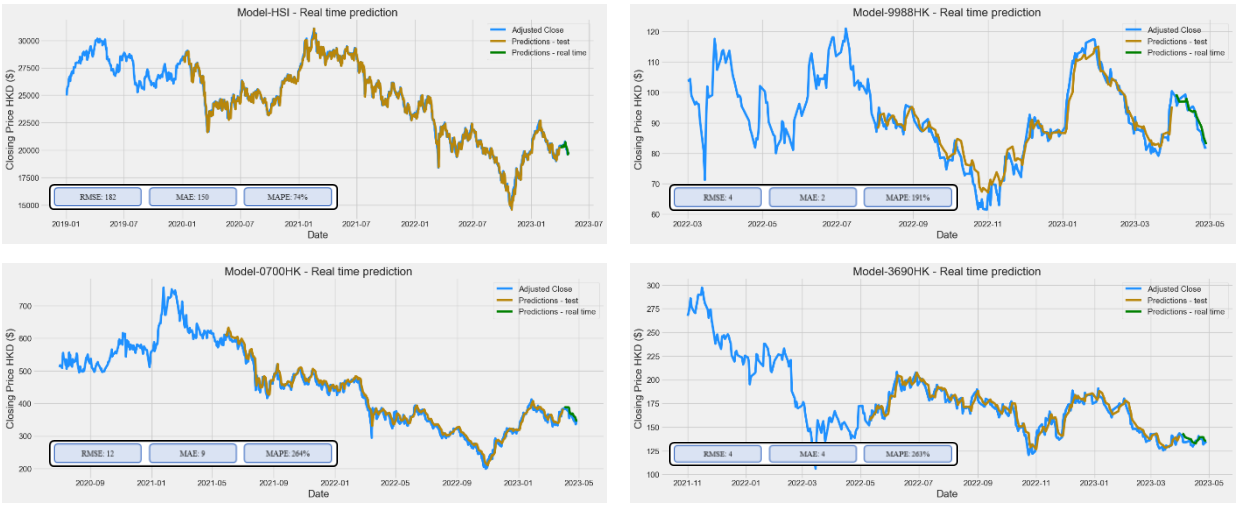


Figure 14: LSTM real time prediction result

Figure 14 plotted the entire period, crossed the model training, testing and real time prediction time span. Blue lines were the actual adjusted close prices, tan lines were the predicted values during testing period while the green lines were the real time prediction period. It is observed that both tan lines and green lines were almost overlapped with the blue actual prices, confirming the predictions were satisfactory.

6 Project Evaluation and Reflection

6.1 Conclusion of Project

The real time prediction results were highly encouraging, as they successfully achieved the aim of assisting in the optimization of swing trading investment strategies for selected stocks in the Hong Kong stock market. These results serve as an effective recommendation for further action.

Investors are eager to find effective ways to predict the stock market for successful investments. This project expanded from its original aim and broke down into various objectives including research on available solutions and practical experiments. A thorough literature review was performed to critically evaluate modern approaches to stock predictions by different software.

The literature review consolidated the framework of the practical experiment, which involved building logistic regression and LSTM models using Python through Anaconda Jupyter Notebook. Historical stock data obtained from Yahoo Finance was subjected to data cleansing and feature engineering, with technical analysis performed to understand the stock market landscape. Principal component analysis was studied for dimension reduction. Two feature sets of each algorithm were trained and evaluated. The algorithm with the highest prediction accuracy and efficiency was selected based on corresponding evaluation metrics with various charts created to articulate the result clearly.

The selected models were then applied to real time data to simulate the real-life situation. The same evaluation metrics were used for comparisons. The results maintained high accuracies up to 71% for logistic regression model and normalized RMSE kept at 5% below or the lowest for HSI of only 1% respectively. It was concluded that the models were robust and effective.

Furthermore, social, ethical, professional and legal considerations have been examined. The social impact of machine learning on investment decisions, ethical considerations in data science, the professional model building process and legal issues related to data collection method and usage. In short, there were no significant ethical or legal issues, while social and professional issues were well controlled under the academic purpose context.

Overall, the project has been completed satisfactorily, providing valuable insights into the use of machine learning for stock prediction. The project's thorough approach to data preprocessing, feature engineering, and model selection were a valuable resource that could look for further improvement in the future.

6.2 Recommendation on Future Works

In terms of future work, this project opens up several possibilities for further research and improvement. One potential avenue for future work is to expand the scope of the project to include more stocks in the practical research. This would allow for a more comprehensive evaluation of the predictive models across a broader range of stocks and could potentially extend to assist investment portfolio management.

Another area for improvement is in hyperparameter tuning. While the manual search approach used in this project was effective, more advanced searching approaches could be implemented to further improve the performance of the models in a systematic way. Methods such as grid search, Bayesian optimization, or genetic algorithms could be explored to identify the optimal hyperparameters for the models.

Additionally, this project focused primarily on predicting either the trend or the price of the stocks. However, future work could explore the possibility of combining trend and price predictions to provide a more comprehensive investment recommendation. By using both trend and price predictions, investors could be provided with a clearer picture of the market and be able to make more informed investment recommendation. This could be achieved by implementing a buy/sell signal strategy at different price reference points.

Overall, there are several avenues for future work that could build upon the results and insights gained from this project. By expanding the scope of the research, improving hyperparameter tuning, and exploring new strategies for investment recommendation, further advancements can be made in the field of stock prediction using machine learning algorithms.

6.3 Personal Reflection

This project was an incredibly valuable experience for me, as it allowed me to leverage my pre-existing knowledge and extend it into new areas of data science and machine learning. I began by setting clear objectives for the project, which helped me to stay focused and motivated throughout the process. Additionally, I found that breaking down the learning process into smaller, manageable tasks allowed me to pick up new skills in Python, model training, and literature review more effectively.

One of the most challenging aspects of the project was conducting a thorough literature review to identify the most relevant research and techniques in the field of stock prediction using machine learning algorithms. At first, I found it challenging to know where to start, but I quickly learned that breaking down the research into smaller subtopics made it easier to manage. Through my

research, I was able to formulate my research questions and develop a better understanding of the research area.

Additionally, I had to learn how to use Python and machine learning libraries such as Scikit-Learn and TensorFlow, which were essential in the implementation of the predictive models. This hands-on experience allowed me to develop a deeper understanding of these tools and how to apply them to real-world problems.

In the project, I also had to learn feature engineering and model training techniques, such as principal component analysis and hyperparameter tuning. These techniques were essential in preparing the data for machine learning models, and I gained valuable experience in applying them to real-world data.

Overall, the project was a challenging but rewarding experience that allowed me to grow significantly in the data science field. I gained hands-on experience in literature review skills, Python coding, and developed a deeper understanding of feature engineering and model training techniques. I am excited to continue developing my skills and knowledge in these areas and to apply them to future projects.

Appendix A. Reference List

[1] www.sfc.hk. (n.d.). *Statistics*. [online] Available at: https://www.sfc.hk/en/Published-resources/Statistics [Accessed 24 Jan. 2023].

[2] Ravikumar, S. and Saraf, P. (2020). Prediction of Stock Prices using Machine Learning (Regression, Classification) *Algorithms. 2020 International Conference for Emerging Technology (INCET)*. doi:https://doi.org/10.1109/incet49848.2020.9154061. [Accessed 18 Feb. 2023].

[3] Yadav, A., Jha, C.K. and Sharan, A. (2020). Optimizing LSTM for time series prediction in Indian stock market. *Procedia Computer Science*, 167, pp.2091–2100. doi:https://doi.org/10.1016/j.procs.2020.03.257. [Accessed 18 Feb. 2023].

[4] Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, 4, pp.51–62. doi:10.20544/horizons.b.04.1.17.p05. [Accessed 24 Jan. 2023].

[5] Huang, Y. (2009). Advances in Artificial Neural Networks – Methodological Development and Application. *Algorithms*, [online] 2(3), pp.973–1007. doi:https:/doi.org/10.3390/algor2030973. [Accessed 24 Jan. 2023].

[6] Mageswaran, G., Nagappan, S.D., Hamzah, N. and Brohi, S.N. (2018). Machine Learning: An Ethical, Social & Political Perspective. *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. doi:https://doi.org/10.1109/icaccaf.2018.8776702. [Accessed 24 Jan. 2023].

[7] Graham, P. ed., (2009). *Analysis of Stock Market Investment Strategies*. [online] digital.wpi.edu. Available at: https://digital.wpi.edu/pdfviewer/5h73pw33q. [Accessed 18 Feb. 2023].

[8] Chandra, A. (2008). *Decision Making in the Stock Market: Incorporating Psychology with Finance*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1501721. [Accessed 18 Feb. 2023].

[9] Smita, M. ed., (2021). *Logistic Regression Model - A Review*. [online] Available at: https://ijisrt.com/assets/upload/files/IJISRT21MAY1050.pdf [Accessed 20 Feb. 2023].

[10] Upadhyay, A., Bandyopadhyay, G. and Dutta, A. (2012). Forecasting Stock Performance in Indian Market using Multinomial Logistic Regression. *ProQuest*, [online] pp.16–39. Available at: https://www.proquest.com/docview/1025743869/FED33880413F41E3PQ/7?accountid=14154 [Accessed 20 Feb. 2023].

[11] Rout, A.R. (2020). *Advantages and Disadvantages of Logistic Regression*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/. [Accessed 20 Feb. 2023].

[12] Harrison, O. (2018). *Machine Learning Basics with the K-Nearest Neighbors Algorithm.* [online] Medium. Available at: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761. [Accessed 25 Feb. 2023].

[13] Alkhatib K., Najadat H., Hmeidi I. and Shatnawi M., 2013. *Stock price prediction using k-nearest neighbor (kNN) algorithm.* International Journal of Business, Humanities and Technology 3.3 (2013): 32-44. https://www.ijbhtnet.com/journals/Vol_3_No_3_March_2013/4.pdf [Accessed 25 Feb. 2023].

[14] Latha, R.S., Sreekanth, G.R., Suganthe, R.C., Geetha, M., Selvaraj, R.E., Balaji, S., Harini, K.R. and Ponnusamy, P.P. (2022). Stock Movement Prediction using KNN Machine Learning Algorithm. *2022 International Conference on Computer Communication and Informatics (ICCCI)*. doi:https://doi.org/10.1109/iccci54379.2022.9740781. [Accessed 25 Feb. 2023].

[15] Manimegalai, T., Manju, J., Rubiston, M.M., Vidhyashree, B. and Prabu, R.Thandaiah. (2022). *Prediction of OPTIMIZED Stock Market Trends using Hybrid Approach Based on KNN and Bagging Classifier (KNNB).* [online] IEEE Xplore. doi: https://doi.org/10.1109/CSNT54456.2022.9787638. [Accessed 25 Feb. 2023].

[16] Khaidem, L., Saha, S. and Dey, S. (2016). *Predicting the direction of stock market prices using random forest.* Applied Mathematical Finance, [online] 00(00), pp.1–20. Available at: https://arxiv.org/pdf/1605.00003.pdf [Accessed 25 Feb. 2023].

[17] Yin, L., Li, B., Li, P. and Zhang, R. eds., (2021). *Research on stock trend prediction method based on optimized random forest*. [online] CAAI Transactions on Intelligence Technology. Available at: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/cit2.12067 [Accessed 25 Feb. 2023].

[18] Argade, S., Chothe, P., Gawande, A., Joshi, S. and ` (2022). *Machine Learning in Stock Market Prediction: A Review*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4128716 [Accessed 9 Apr. 2023].

[19] www.turing.com. (n.d.). *Recurrent Neural Networks and LSTM: Overview and Uses | Turing*. [online] Available at: https://www.turing.com/kb/recurrent-neural-networks-and-lstm.

[20] DagsHub Blog. (2023). *RNN, LSTM, and Bidirectional LSTM: Complete Guide | DagsHub*. [online] Available at: https://dagshub.com/blog/rnn-lstm-bidirectional-lstm/ [Accessed 9 Apr. 2023].

[21] ssla.co.uk, (n.d.). *what is long short term memory? And how long is memory? | ssla.co.uk*. [online] Available at: https://www.ssla.co.uk/long-short-term-memory/. [Accessed 9 Apr. 2023].

[22] Chimmula, V.K.R. and Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, 135, p.109864. doi:https://doi.org/10.1016/j.chaos.2020.109864. [Accessed 10 Apr. 2023].

[23] Vijh, M., Chandola, D., Tikkiwal, V.A. and Kumar, A. (2020). *Stock Closing Price Prediction using Machine Learning Techniques.* Procedia Computer Science, 167, pp.599–606. doi:https://doi.org/10.1016/j.procs.2020.03.326 [Accessed 10 Apr. 2023].

[24] Thakkar, A. and Chaudhari, K. (2021). A Comprehensive Survey on Deep Neural Networks for Stock Market: The Need, Challenges, and Future Directions. *Expert Systems with Applications*, p.114800. doi:https://doi.org/10.1016/j.eswa.2021.114800. [Accessed 10 Apr. 2023].

[25] Chen, K., Zhou, Y. and Dai, F. (2015). A LSTM-based method for stock returns prediction: A case study of China stock market. *2015 IEEE International Conference on Big Data (Big Data)*. [online] doi:https://doi.org/10.1109/bigdata.2015.7364089. [Accessed 22 Apr. 2023].

[26] Li, X., Wu, P. and Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management*, p.102212. doi:https://doi.org/10.1016/j.ipm.2020.102212. [Accessed 22 Apr. 2023].

[27] Balaji, S.N., Paul, P.V. and Saravanan, R. (2017). *Survey on sentiment analysis based stock prediction using big data analytics*. [online] IEEE Xplore. doi:https://doi.org/10.1109/IPACT.2017.8244943. [Accessed 22 Apr. 2023].

[28] Wankhade, M., Rao, A.C.S. and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*. doi:https://doi.org/10.1007/s10462-022-10144-1. [Accessed 22 Apr. 2023].

[29] Istiake Sunny, Md.A., Maswood, M.M.S. and Alharbi, A.G. (2020). Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. doi:https://doi.org/10.1109/niles50944.2020.9257950. [Accessed 22 Apr. 2023].

[30] Angadi, M.C. and Kulkarni, A.P. eds., (2015). *Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R.* [online] Available at: https://www.researchgate.net/profile/Mahantesh-Angadi-4/publication/286890497_Time_Series_Data_Analysis_for_Stock_Market_Prediction_using_Data_Mining_Techniques_with_R/links/566eecb508ae4d4dc8f7f7bb/Time-Series-Data-Analysis-for-Stock-Market-Prediction-using-Data-Mining-Techniques-with-R.pdf [Accessed 6 May 2023].

[31] Sen, J. ed., (2016). *Stock Price Prediction Using Machine Learning and Deep Learning Frameworks.* [online] Academia. Available at: https://www.researchgate.net/publication/328873189_Stock_Price_Prediction_Using_Machine_Learning_and_Deep_Learning_Frameworks. [Accessed 6 May 2023].

[32] 6sense.com. (n.d.). *Python - Market Share, Competitor Insights in Programming Languages*. [online] Available at: https://6sense.com/tech/programming-language/python-market-share [Accessed 6 May 2023].

[33] Sharma, A., Modak, S. and Sridar, E. eds., (2019). *Data Visualization and Stock Market and Prediction.* [online] Available at: https://www.irjet.net/archives/V6/i9/IRJET-V6I9318.pdf [Accessed 6 May 2023].

[34] ProjectPro. (2023). *Data Science Programming: Python vs R*. [online] Available at: https://www.projectpro.io/article/data-science-programming-python-vs-r/128. [Accessed 6 May 2023].

[35] Ozgur, C., Colliau, T., Rogers, G. and Hughes, Z. (2021). MatLab vs. Python vs. R. *Journal of Data Science*, 15(3), pp.355–372. doi:https://doi.org/10.6339/jds.201707_15(3).0001. [Accessed 6 May 2023].

[36] Mai, A. (n.d.). *Technical Analysis from A to Z Technical Analysis from A to Z. www.academia.edu*. [online] Available at: https://www.academia.edu/34831627/Technical_Analysis_from_A_to_Z_Technical_Analysis_from_A_to_Z [Accessed 9 May 2023].

[37] Mitchell, C. (2019). *How to Use a Moving Average to Buy Stocks*. [online] Investopedia. Available at: https://www.investopedia.com/articles/active-trading/052014/how-use-moving-average-buy-stocks.asp. [Accessed 9 May 2023].

[38] South China Morning Post. (2023). *2023 will not turn into a repeat of 2008 crisis, Hong Kong SFC says*. [online] Available at: https://www.scmp.com/business/banking-finance/article/3214114/credit-suisse-opens-usual-hong-kong-after-us325-billion-takeover-ubs-sign-europes-banking-rout-under [Accessed 9 May 2023].

[39] Lucid Content Team (2019). *Agile-Waterfall Hybrid: Is It Right for Your Team? | Lucidchart Blog*. [online] Lucidchart.com. Available at: https://www.lucidchart.com/blog/is-agile-waterfall-hybrid-right-for-your-team. [Accessed 18 May 2023].

[40] Wilson, J.M. (2003). *Gantt charts: A centenary appreciation. European Journal of Operational Research*, 149(2), pp.430–437. doi:https:/doi.org/10.1016/s0377-2217(02)00769-5. [Accessed 18 May 2023]

[41] KDnuggets. (n.d.). *Missing Value Imputation - A Review*. [online] Available at: https://www.kdnuggets.com/2020/09/missing-value-imputation-review.html. [Accessed 18 May 2023]

[42] Adeyemo, A. and Wimmer, H. (2018). *Effects of Normalization Techniques on Logistic Regression in Data Science*. [online] Available at: https://proc.conisar.org/2018/pdf/4813.pdf [Accessed 18 May 2023].

[43] Waqar, M., Dawood, H., Guo, P., Shahnawaz, M.B. and Ghazanfar, M.A. (2017). Prediction of Stock Market by Principal Component Analysis. *2017 13th International Conference on Computational Intelligence and Security (CIS)*. [online] doi:https://doi.org/10.1109/cis.2017.00139. [Accessed 18 May 2023]

[44] Ahuja, R., Kumar, Y., Goyal, S., Kaur, S., Sachdeva, R.K. and Solanki, V. (2023). *Stock Price Prediction By Applying Machine Learning Techniques*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ESCI56872.2023.10099614. [Access 19 Jun 2023]

[45] Otto, S. (2019). *How to normalize the RMSE*. [online] ww.marinedatascience.co. Available at: https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/. [Accessed 18 Jun 2023]

[46] Hashmi, F. (2020). *How to find the best hyperparameters using Manual Search in Python*. [online] Thinking Neuron. Available at: https://thinkingneuron.com/how-to-find-the-best-hyperparameters-using-manual-search-in-python/ [Accessed 20 Jun. 2023].

Appendix B. Meeting Log

University of Sunderland

Meeting Log

Date:	7Feb2023	Time:	HKT 6:00-7:00pm
Called By:	Marta	Attendees:	Marta / Jackie

Hours spent on project since last meeting:	130

Brief description of work since last meeting

- Drafted planning review for a review discussion during the meeting
- Started study background research and get the alignment of the direction on topic, specified on the topic at swing trading recommendation

Issues identified

- Admin issue regarding ethics form, solved the logistics during the discussion
- Clarified research references

Issues Agreed tasks for next meeting

- To discuss the literature review contents

Next meeting

Date:	30 Mar 2023
Time:	HKT 1:00-2:00pm

University of Sunderland

Meeting Log

Date:	30Mar2023	Time:	HKT 1:00-2:00pm
Called By:	Marta	Attendees:	Marta / Jackie

Hours spent on project since last meeting:	250

Brief description of work since last meeting

- Drafted dissertation outline for a review discussion during the meeting
- Performed literature review
- Researched relevant professional, ethical, social and legal issue
- Learning python and trying some data cleansing and baseline model codings

Issues identified

- Marta updated her approval right at ethics form has been set, Jackie to follow up with King on the approval status
- Discussed and identified 3 more meetings in the rest period. End Apr, End May and Mid Jun

Issues Agreed tasks for next meeting

- Review literature review
- Update practical research progress

Next meeting

Date:	27 Apr 2023
Time:	HKT 12:30-1:30pm

University of Sunderland

Meeting Log

Date:	27Apr2023	Time:	HKT 12:30-13:30pm
Called By:	Jackie	Attendees:	Marta / Jackie

Hours spent on project since last meeting:	380

Brief description of work since last meeting

- Write the introduction and background for the dissertation
- Consolidating literature review
- Started coding for exploratory data analysis

Issues identified

- Jackie updated King has approval the ethics form
- Marta feedback on the introduction write-ups, elaborated about good citation with examples
- Jackie updated the literature review results and practical research at exploratory technical analysis
- Discussed the sentiment analysis from social media, Jackie did some studies and will discuss in dissertation on the immaturity of sentiment analysis in Hong Kong stock market that concluded not to use sentiment information for model building
- Jackie asked the viva session arrangement and would like to seek for the possibility of viva session to be held on or before 14Jul2023

Issues Agreed tasks for next meeting

- Review Marta's comment and revisit on introduction part
- Update practical research progress
- Marta to advise if there's update on the viva session arrangement

Next meeting

Date:	9Jun2023
Time:	HKT 6:30-7:30pm

University of Sunderland

Meeting Log

Date:	12Jun2023	Time:	HKT 6:30-7:30pm
Called By:	Jackie	Attendees:	Marta / Jackie

Hours spent on project since last meeting:	580

Brief description of work since last meeting

- Write the literature review and technical analysis for the dissertation
- At model evaluation stage of practical research

Issues identified

- Jackie updated the follow up with Vicky from student support that Viva session will be arranged in 11-12Jul2023
- Marta feedback on practical research result. She suggested to include all trials of logistic regression and LSTM model result, for trend and price prediction respectively in the dissertation. Therefore, slightly expand the objective of developing machine learning models instead of only LSTM model
- Jackie sought advice on the handling on similarity from Turnitin report. The similarity of first draft mainly from student papers due to declarations and those quoted sentences. Marta will check with university.
- Jackie asked how to attach proposal in the dissertation as it will be uploaded in PDF format. Marta advised to put in appendix
- Discussed the key topics to include in viva presentation

Issues Agreed tasks for next meeting

- Jackie to submit the completed dissertation and viva session materials by 23Jun
- Marta to provide comment in the week of 26Jun for final touch up

Next meeting

Date:	4Jul2023
Time:	HKT 5:00-5:30

University of Sunderland

Meeting Log

Date:	4Jul2023	Time:	HKT 5:00-5:30pm
Called By:	Jackie	Attendees:	Marta / Jackie

Hours spent on project since last meeting:	600

Brief description of work since last meeting

Finalized the dissertation and viva presentation

Issues identified

- Discuss the final review comments

Issues Agreed tasks for next meeting

- Review and follow up with the final comments

Next meeting

Date:	NA
Time:	NA