RO5

Web Application for Real-time Forex Rate and Stock Price Prediction

by
Ho Cheuk Yin and Wong Hoi Ming
RO5

Advised by
Prof. David ROSSITER

Submitted in partial fulfillment
of the requirements for COMP 4981
in the
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology 2017-2018
Date of submission: 19 April 2018

Abstract

In this project several supervised machine learning models had been implemented and test in order to perform the weekly / monthly price predictions and the monthly/quarterly trend prediction on 10 most traded foreign currencies (forex) and 10 most traded Hong Kong stocks. For weekly/monthly price prediction, it was discovered that in general simpler models like linear regression tend to have much better performance compared to the complex models like recurrent neural networks. For monthly/quarterly predictions, For monthly/quarterly trend prediction, the random forest model tends to perform satisfactory with accuracy much higher than random guessing. The prediction results of the models were illustrated in a web application constructed with responsive and intuitive web user interface design, which enable easy and user-friendly interpretation of the prediction results

Table of Contents

1. Introduction

1.1 Overview

With the advancement in globalization and international trade in recent decades, the importance of foreign currency exchange (forex) rate has grown greater and greater in most countries over the years. The fluctuation of forex rates nowadays can have a huge impacts toward different groups of stakeholders in society, including governments, companies and households, in cost and decision-making aspects.

Although stock markets are smaller than the global forex market. They too have a huge impact on people's decisions. Thus, both are closely followed on a 24/7 basis. And due to the great potential for financial loss and gain, the financial technology (fintech) industry has also greatly expanded to analyze financial trends and provide financial services. As a result, methodologies have been developed to effectively predict the likely future trends of forex and stock prices so as to help people estimate costs, manage risk and make informed decisions.

In this project, we are developing a web application for real-time forex and stock price prediction. We implemented and tested several different approaches, including statistical model for time series forecasting, supervised machine learning algorithms and an artificial neural network, comparing their accuracy in both regression prediction and trend forecasting for different time periods, Then, we selected one for long-term trend prediction and another for short-term average price prediction. The web application illustrates the results of forecasts to assist investors and business people in making business and financial decisions, and it also includes the probability of prediction confidence, prediction error metrics and some ensemble/voting prediction results.

1.2 Objectives

Currently, there are numerous approaches to making forex rate and stock price predictions. The goal of this project is to integrate the strengths of several methods to predict the top 10 most-traded forex pairs(e.g. EUR/USD, USD/JPY, etc.) and the top 10 most-traded stocks on the Hong Kong Stock Exchange (TENCENT, CCB, etc.) and also to develop a web application to allow people to access real-time information about these predictions. To achieve this goal, we have been working on the following objectives:

1. Build a data scraper to continually grab real-time data about the top 10 most-traded forex pairs and the top 10 most-traded Hong Kong stocks.
2. Build an online database to store all the data.
3. Build an online system to predict the long-term (i.e. monthly and quarterly) rise/fall trend and short-term(i.e. daily and weekly) average rates of the top 10 forex pairs and the prices of the top 10 local stocks, making predictions based on several indicators from statistics and technical analysis (i.e. simple moving averages, relative strength index, moving average convergence/divergence and etc.)
4. Build a user-friendly web application that allows users to view prediction results and visualize the trends via charts.
5. Include the probability of prediction confidence for long-term trend prediction.
6. Include prediction error metrics for short-term average price prediction.
7. Provide an ensemble/voting prediction result for long-term trend prediction.

To accomplish these objectives, we have been utilizing several open-source libraries for data scraping, technical analysis indicators, supervised machine learning models, recurrent neural networks, web app construction and data visualization.

1.3 Literature Survey

We did a survey on various methods to predict forex rates and stock prices. We also looked at ways to calculate the probability of prediction confidence, different prediction error metrics and ways to combine different methods to produce ensemble / voting results.

1.3.1 Technical Analysis Indicators

The indicators in technical analysis could be referred as the metrics derived from the general price information and activities in stocks or financial assets [1]. Most of them were developed for analyzing the general price movement and trend directions based on some empirical equations or statistical concepts.

1.3.1.1 Simple Moving Average

The Simple Moving Average (SMA) is one of the most popular and widely used technical indicator in trend determination and price smoothing [2]. By continuing taking the average of price in specified time frame, SMA help filtered out the noise and fluctuation of price and generate a trend-line based on the daily SMA.

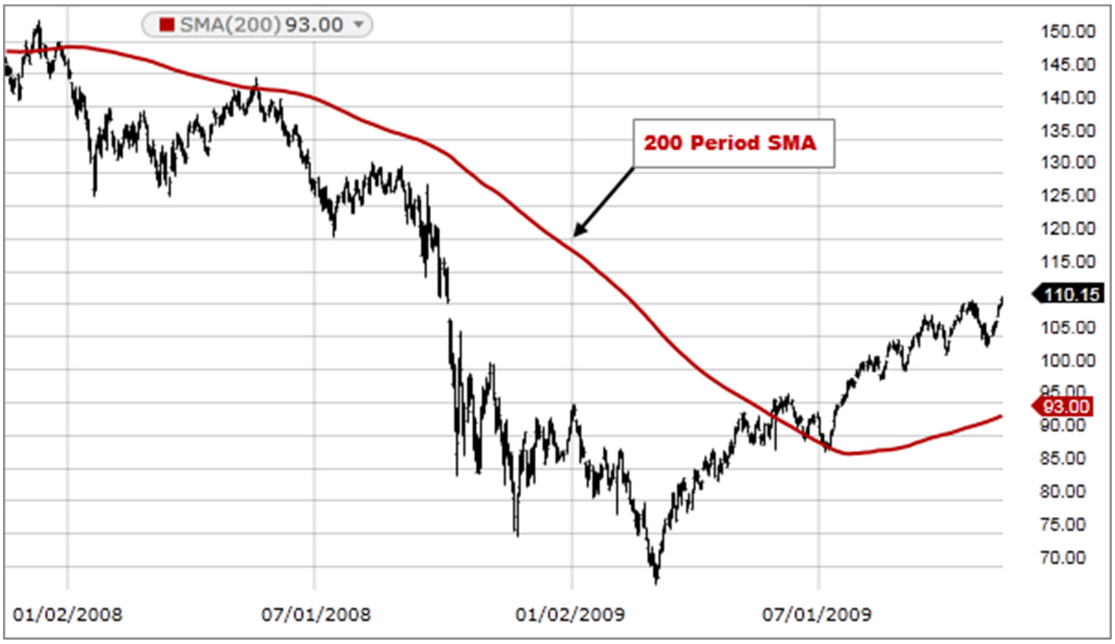
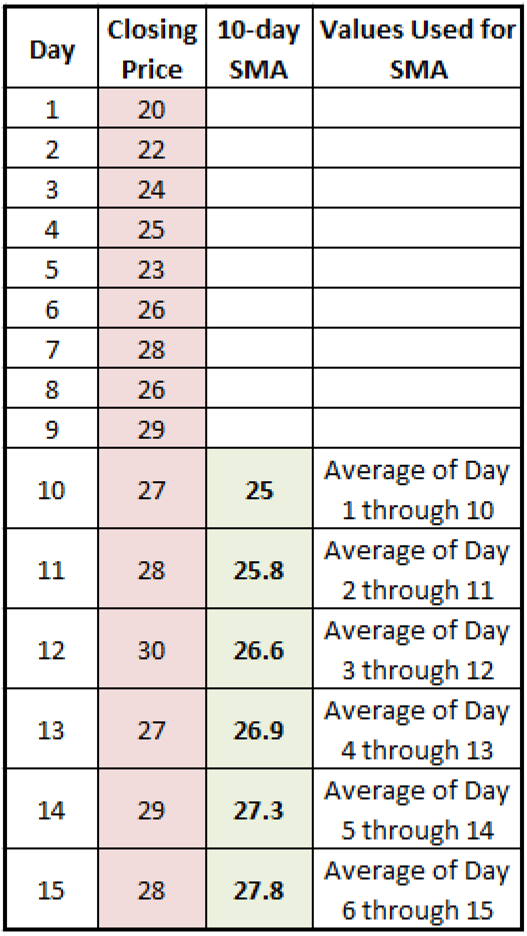


Figure 1: An graphical illustration of a 200-Period SMA, where the red line refers to the SMA, black curve refers to the original prices.

Formula 1: The formula of calculating SMA

$$\mathrm{SMA} = \frac{P_t + P_{t-1} + P_{t-2} \ + \cdots + P_{t-n}}{n}, n = specified \ \ time \ \ period$$



Day	Closing Price	10-day SMA	Values Used for SMA
1	20		
2	22		
3	24		
4	25		
5	23		
6	26		
7	28		
8	26		
9	29		
10	27	**25**	Average of Day 1 through 10
11	28	**25.8**	Average of Day 2 through 11
12	30	**26.6**	Average of Day 3 through 12
13	27	**26.9**	Average of Day 4 through 13
14	29	**27.3**	Average of Day 5 through 14
15	28	**27.8**	Average of Day 6 through 15

Figure 2: An illustration of SMA calculation for n = 10 days.

1.3.1.2 Relative Strength Index

The Relative Strength Index (RSI) was developed as a momentum indicator (i.e the indicator for measuring the strength of the rise/fall motion) to analyze whether the financial assets, including forex and stock, are currently under the overbought or oversold by referring to formula 2:

Formula 2: The formula of calculating RSI

$$\mathrm{RSI} = \ \ 100 - \frac{100}{(1 + \mathrm{RS})}$$

$$\mathrm{RS} = \frac{\mathrm{Average \ \ gain \ \ of \ \ up \ \ periods \ \ in \ \ specified \ \ time \ \ frame}}{\mathrm{Average \ \ loss \ \ of \ \ down \ \ periods \ \ in \ \ specified \ \ time \ \ frame}}$$

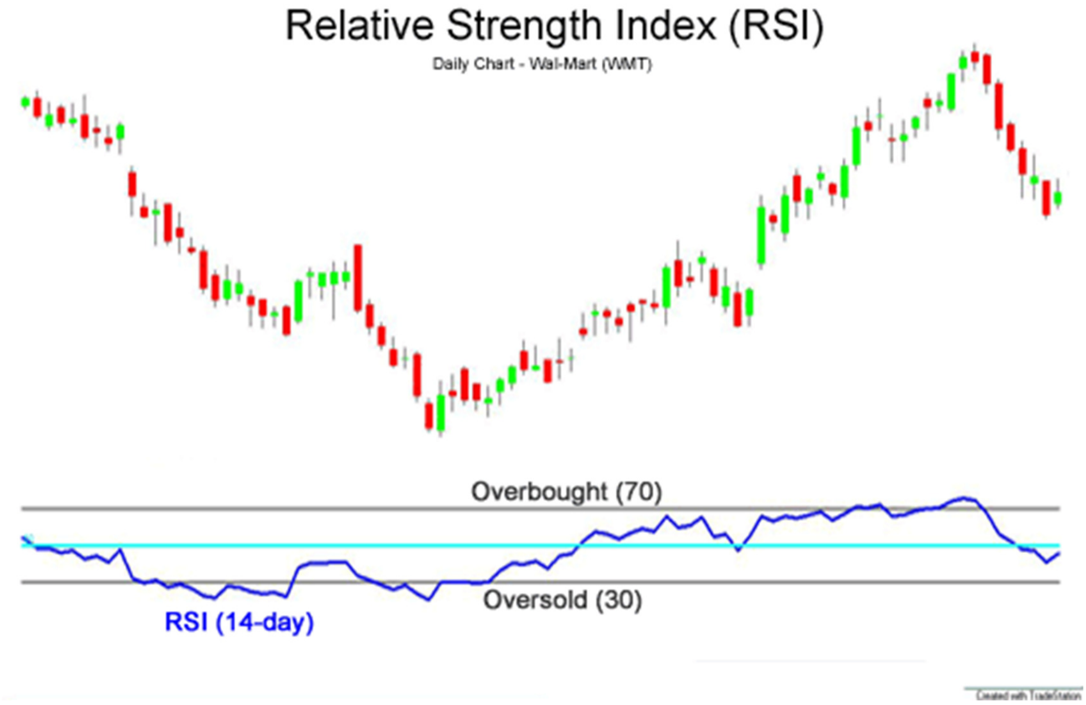


Figure 3: A graphical illustration of a 14-Period RSI.

The RS metric was referred as the relative strength of the up or down motion of the price in the specified time frame. Empirically, the time frame is specified as 14 consecutive trading sessions (i.e. 14 consecutive business days for daily prices).According to the definition of RSI, the value is within the range of 0 to 100. In general, when RSI exceeds 70 (below 30), the financial asset it refers to is regarded as under the overbought (oversold) condition and a reverse in price trend is expected to occur in the future[3].

1.3.1.2 Moving Average Convergence Divergence

Moving Average Convergence Divergence(MACD) is another momentum indicator determine the trend of the financial assets by calculating the differences between two Exponential Moving Average(EMA) with different time periods. Apart from the MACD indicators, a EMA of the MACD is also calculated as a signal line to help for determining the trend and the signal of possible trend reversal.

Formula 3: The formula of calculating MACD and the signal line.

Where t3 refers to the time period for the EMA.

$$\mathrm{MACD} = \mathrm{EMA}(n_2) - EMA(n_1)$$

$$\text{Signal line} = \text{EMA}(\text{MACD}, t_3)$$

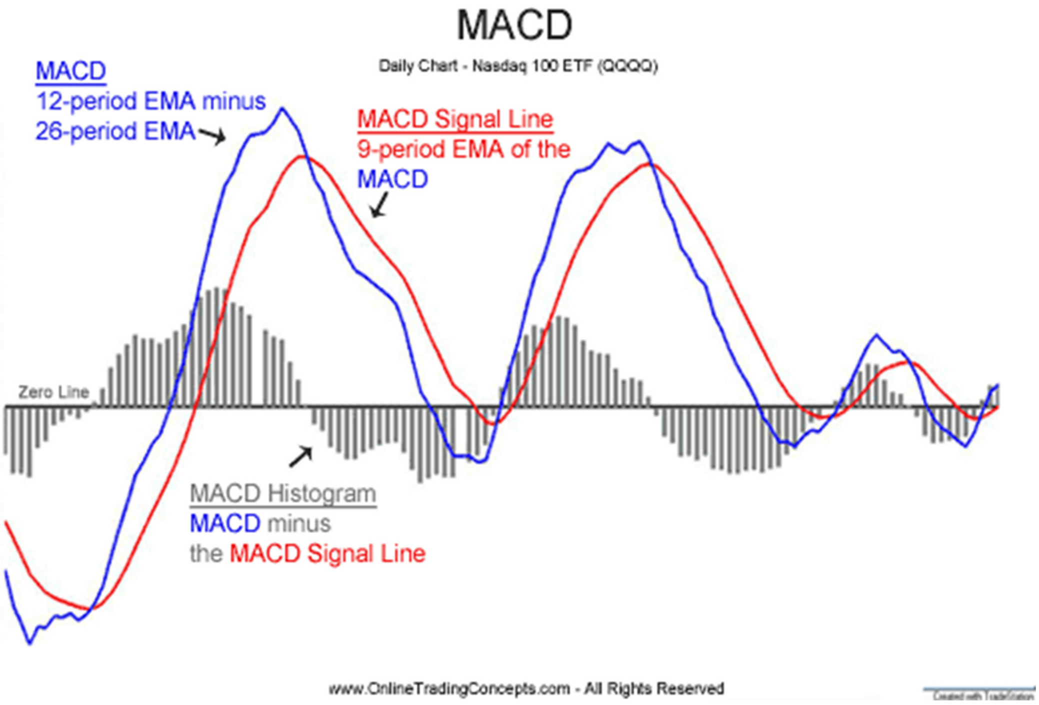


Figure 4: An graphical illustration of MACD.

By convention, n_2 and n_1 are set as 12 and 26 respectively, and t_3 is set as 9 [4]. It was generally interpreted that trend-reversal is likely to occur when crossover exists between the MACD and the signal line.

1.3.2 Price Prediction Models

There are in general two main approaches to produce price prediction results, the classic statistical models, including the Time series models, and the supervised machine learning models based on some features from datasets which are correlated to the general price trend in a specific time frame.

For Time series models , one of the popular example is the Autoregressive integrated moving average (ARIMA) model, which is a statistical model make use of time series data to predict future value of the series based on the differences and the moving average of the time series[5].

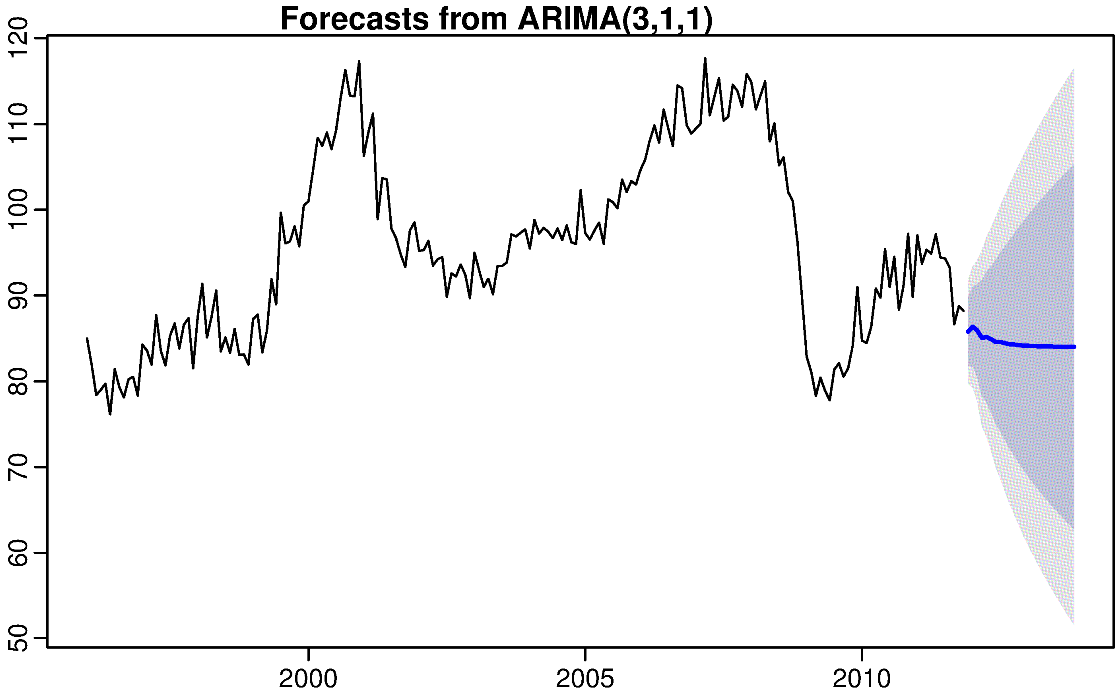


Figure 5: An example of a regression prediction result generated by ARIMA model,

The blue curve refers to the predicted value of the series and the shaded area refers to the confidence interval of the predicted value.

For the supervised machine learning models, they in general perform classification and regression predictions based on some given set of features, including prices, statistical indicators and other factors may contribute to the motion of price trend.

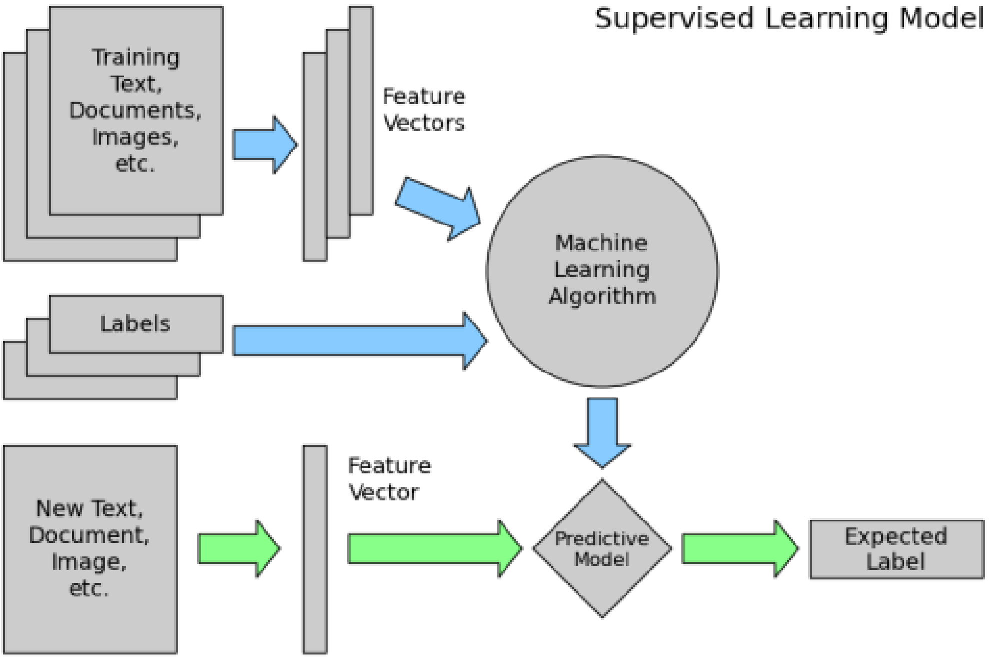

Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Figure 6: The general workflow of training a supervised machine learning model.

Apart from traditional supervised machine learning algorithms like random forests, k-nearest neighboring and linear regressions, recently a brand new model called recurrent neural network is developed based on the artificial neural networks. The architecture of a typical recurrent neuron can be shown in the following figure:

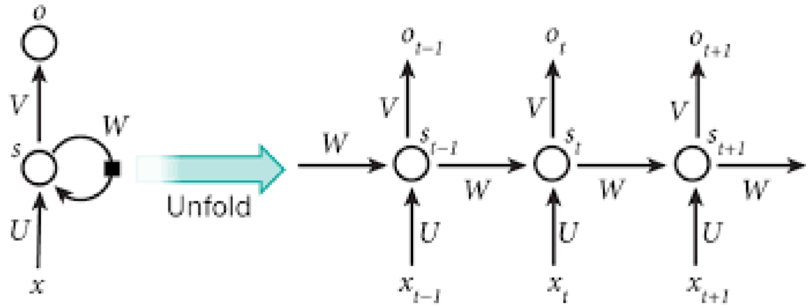


Figure 7: The typical architecture of a recurrent neuron, which the neuron tends to recurrent itself and be able to unfold with timesteps for different inputs and outputs.

The recurrent neural networks were originally developed for solving tasks in Natural Language Processing(NLP) aspect in machine learning, in which NLP tasks often involve ordering of inputs and different ordering of inputs sentences can have significant influence in affecting the accuracy of the prediction. For example, with the sentence “I drink milk”, the ordering of words (I, drink and milk) in the sentence is important in determine the meaning of the complete sentence, which the sentence “I drink milk” and “milk drink I” invoke complete different meaning. In traditional machine learning algorithms, it tends to ignore the ordering of the features in the dataset and thus produce unsatisfactory results to the tasks emphasize the ordering of features (including NLP tasks and time series prediction tasks). With recurrent neural networks, they tends to treat inputs in different time steps as in sequences, which data with same features but different order are now considered as different cases, which improves the prediction accuracy significantly for NLP tasks in recent years. As time series predictions, including stocks and financial assets prices forecasting tasks, share similar characteristics with NLP tasks in terms of sequential inputs with time, the model is expected to produce satisfactory results and be able to improve the prediction accuracy in the related prediction problems.

2. Methodology

2.1 Design

2.1.1System Overview

Our online prediction system can generally be divided into three main components: the web user interface for illustrating the prediction results of specified forex currency pairs or Hong Kong stock by the users and process the user requests to the web server and database in the back-end, the web server for handling user's requests (for example, to view the monthly trend prediction results of USD/JPY forex currency pairs), produce predictions of specified financial assets based on the supervised machine learning models developed and a database for storing the daily price data fetched from internet data sources and the prediction results produced by the web server for users queries and performance evaluation purposes.

The general workflow of the system can be shown in the following figure:

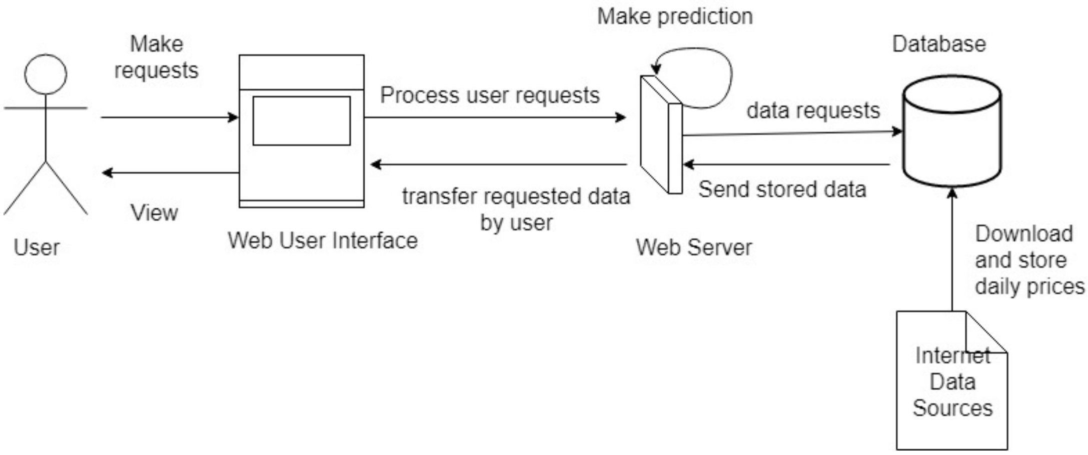


Figure 8: The general workflow of the online prediction system

2.1.2Defining prediction goals and scope

In order to simplify the process of performance evaluation and perform predictions in effective way, we decided to provide two main type of predictions to the users. The short-term average price prediction and the long-term price trend predictions.

For short-term price prediction, it can be further divided into “weekly” and “monthly” price predictions. The “weekly” price prediction is defined to provide the daily close prices prediction for every business days within one week ahead (which is referred as “weekly” price prediction). For example, if today is Monday, the daily close price of specified stocks or foreign currencies in Tuesday, Wednesday, Thursday, Friday and next Monday will be provided
The “monthly” price is defined to provide the average close price of the stock or currencies in every week within one month ahead. For example, given today is Monday, the average close price from Tuesday to next Monday is defined to be the weekly averaged close price by next Monday. In general there are four weeks per month, thus, the predicted weekly averaged close price for the next four weeks will be provided in our web application.

Apart from the “point” prediction on close prices, in order to provide more reliable information regarding the uncertainty of the predictions, the 70% confidence interval of price predictions would also be provided in the application, the definition of the 70% confidence interval were calculated based on the z-score and the properties of normal distribution and will be explained in the Implementation section of this report.

The following figures illustrates the definition of our prediction goals in short-term price prediction tasks:

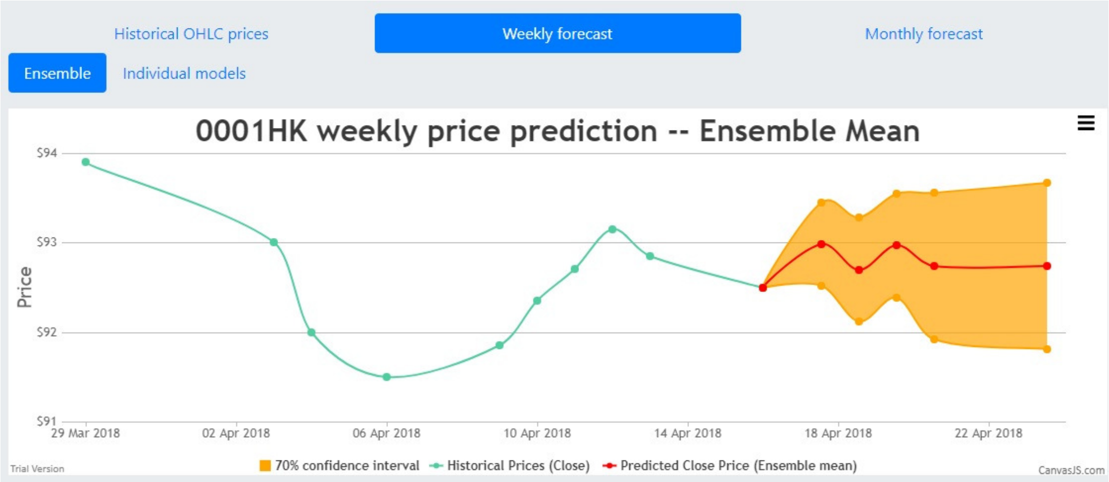


Figure 9: The illustration of short-term weekly price prediction generated by our web application, the predicted close prices within one week ahead and corresponding confidence intervals were provided.

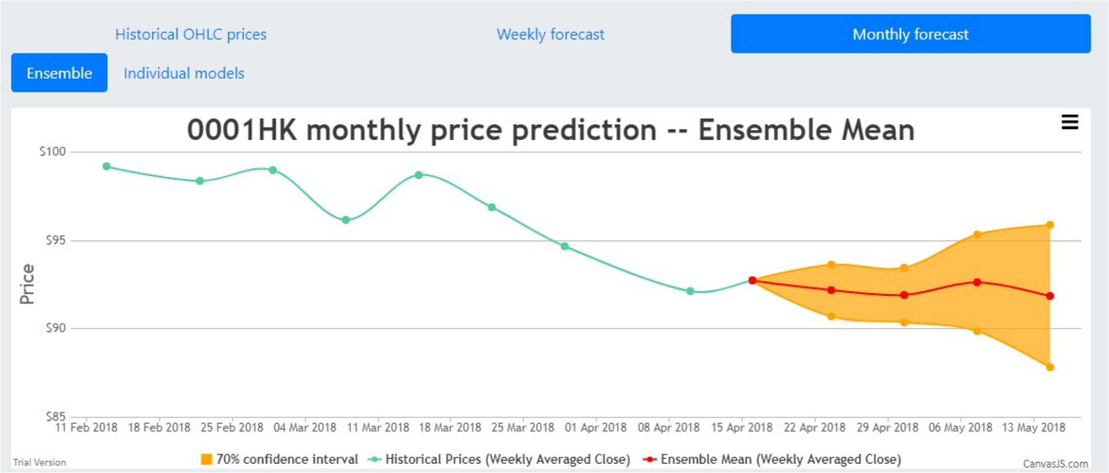


Figure 10: The illustration of short-term monthly price prediction generated by our web application, the predicted close prices within one week ahead and corresponding confidence intervals were provided.

For price trend prediction, it can be defined by comparing the difference between the average price between the specified time frame. For example, for monthly trend prediction, the monthly trend within the past month will be determined by the difference between the 20-days simple moving average of today and the 20-days simple moving average of 20 business days before. If the difference in average prices is greater than 0, it is determined as a rising monthly trend in the past month. Similarly, if the difference in monthly average prices is lower than 0, then it is interpreted as a falling monthly trend.

The definitions can be summarized in the following table and figures:

	Short term		Long term	
	Weekly	Monthly	Monthly	Quarterly
Timespan	Every day within next week	Every weeks within next month	20 days	60 days
Prediction target	$\mathrm{SMA}(\mathrm{t})$		$\Delta\mathrm{P} = \mathrm{SMA}(\mathrm{t}) - SMA(0)$ If $\Delta\mathrm{P} > 0$: Rise trend If $\Delta\mathrm{P} < 0$: fall trend	

Table 1: The definition of various key terms in our prediction system, the variable t refers to the timespan ahead

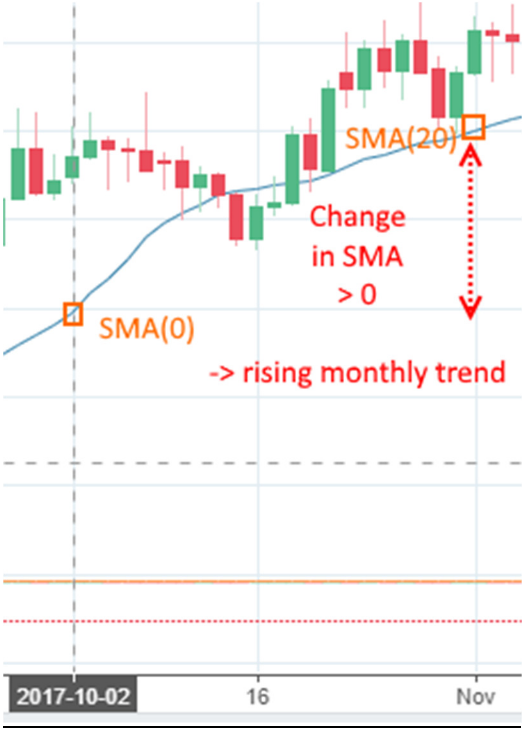


Figure 11: The graphical illustration of monthly trend determination,

The blue curve is the 20-days SMA of USD/JPY forex currencies pairs,

Which shows a rising monthly trend from October 2017 to November 2017.

In our online prediction system, the following forex currency pairs are chosen as the targets for forex prices predictions:

Forex currency pairs	**Market Share % in 2016**
USD/EUR	23.1%
USD/JPY	17.8%
USD/GBP	9.3%
USD/AUD	5.2%
USD/CAD	4.3%
USD/CNY	3.8%
USD/CHF	3.6%
USD/MXN	1.8%
USD/SGD	1.6%
USD/KRW	1.5%

Table 2: The forex currency pairs with prediction available in our prediction system and their corresponding market share in 2016.

The reason we choose these forex currency pairs as our prediction targets in our system is that they are the top 10 most traded forex currency pairs in 2016.

For the Hong Kong stocks, the following ten stocks as constituents of the Hang Seng Index(HSI) are chosen to have prediction provided in our system:

Stock Code	Name	Corresponding aspect as the constituent of HSI
5	HSBC Holdings	Finance
2388	BOC Hong Kong	Finance
388	HKEX	Finance
2	CLP Holdings	Utilities
836	China Res Power	Utilities
4	Wharf Holdings	Properties
2007	Country Garden	Properties
1	CKH Holdings	Commerce & Industry
700	Tencent	Commerce & Industry
992	Lenovo Group	Commerce & Industry

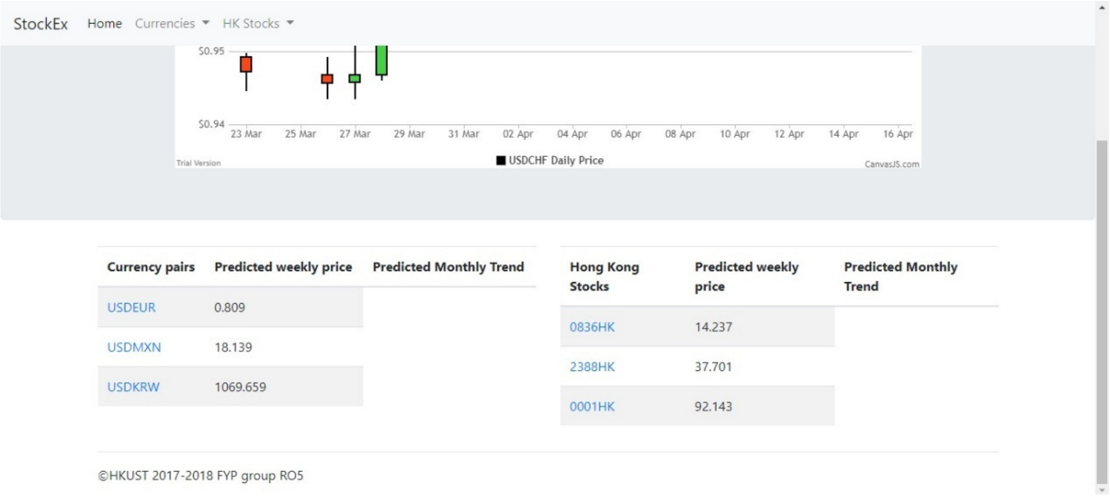
Table 3: The stocks selected to have predictions provided in the prediction system

2.1.3Design of web user interface

In order to provide a user-friendly, informative and easily-interpretable interfaces for the users to obtain accurate historical price information and the predicted price of the specified stocks and forexes, we decided to implement the user interface of our web application in a responsive and graphic-dominated approaches.

The responsive web design refers to the web pages which can adjust the layout of the interfaces and the content based on the screen size or resolutions of the Devices, with the drastic increase in popularity and usages of mobile devices in recent years, the design of web pages to be responsive become necessary.

The following figures shows an example of responsive web design in our web application:

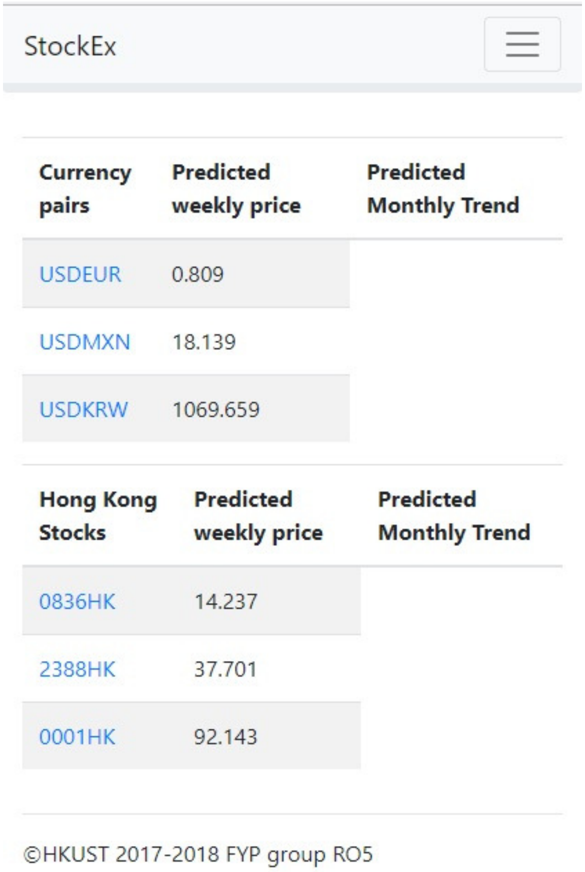


Currency pairs	Predicted weekly price	Predicted Monthly Trend
USDEUR	0.809	
USDMXN	18.139	
USDKRW	1069.659	

Hong Kong Stocks	Predicted weekly price	Predicted Monthly Trend
0836HK	14.237	
2388HK	37.701	
0001HK	92.143	

©HKUST 2017-2018 FYP group RO5

Figure 12: Screenshot of our web application in full screen of laptop/PC



Currency pairs	Predicted weekly price	Predicted Monthly Trend
USDEUR	0.809	
USDMXN	18.139	
USDKRW	1069.659	

Hong Kong Stocks	Predicted weekly price	Predicted Monthly Trend
0836HK	14.237	
2388HK	37.701	
0001HK	92.143	

©HKUST 2017-2018 FYP group RO5

Figure 13: Screenshot for screen with much smaller size (smartphones, tablets, etc.), Notice the alignments of tables and contents inside were adjusted automatically as screen size changes

For the illustration of the historical price information, the candlestick chart representation is adopted in our web application.

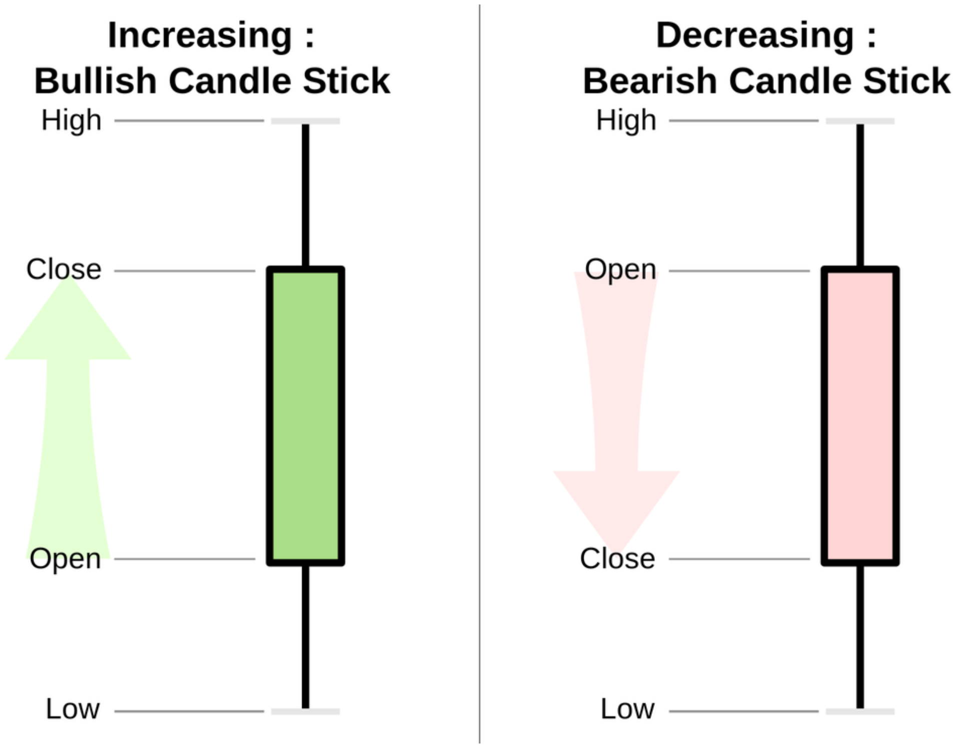
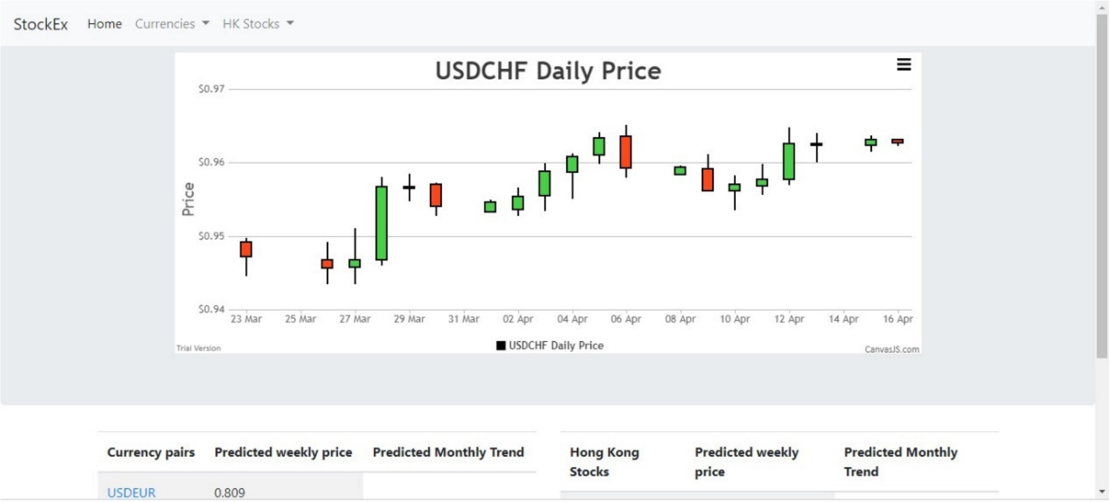


Figure 14: Definition of different type of prices represented candlestick chart

With the candlestick representation, the user can interpret the price information and the potential trend of the specified stock or forex easily in an intuitive way.

The following figures summarize the basic design of our web application:



Currency pairs	Predicted weekly price	Predicted Monthly Trend
USDEUR	0.809	

Hong Kong Stocks	Predicted weekly price	Predicted Monthly Trend

Figure 15: The screenshot of the index page of web application,

A candlestick chart for a random selected stock/forex and the tables summarize several stocks/ forexes are included. Users can select their desired stock/forex through the dropdown menu in the header.

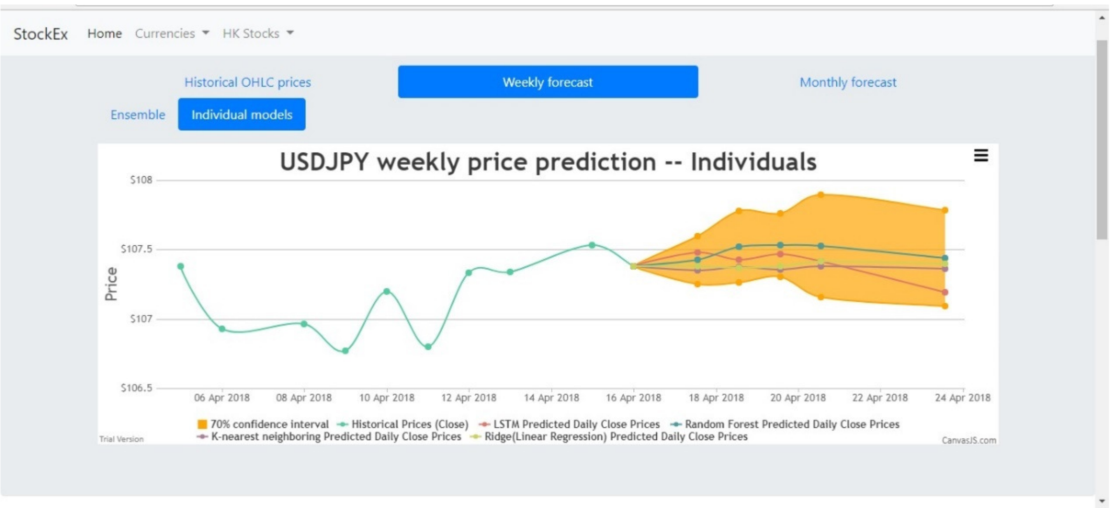
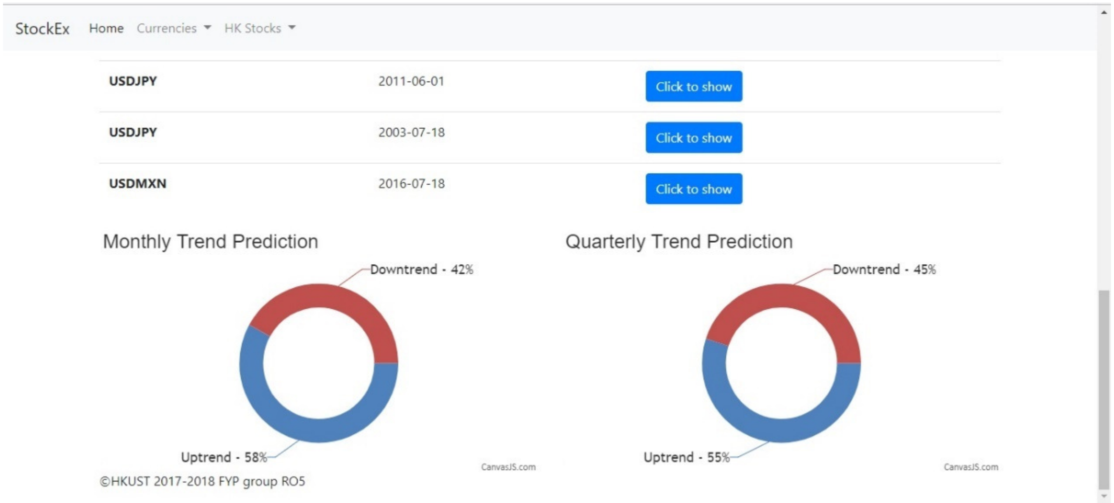


Figure 16: The screenshot of the prediction page of the web application, the user can select to view historical prices for the stock / forex or the forecasts by clicking the blue button lists to switch between different charts.



USDJPY	2011-06-01	Click to show
USDJPY	2003-07-18	Click to show
USDMXN	2016-07-18	Click to show

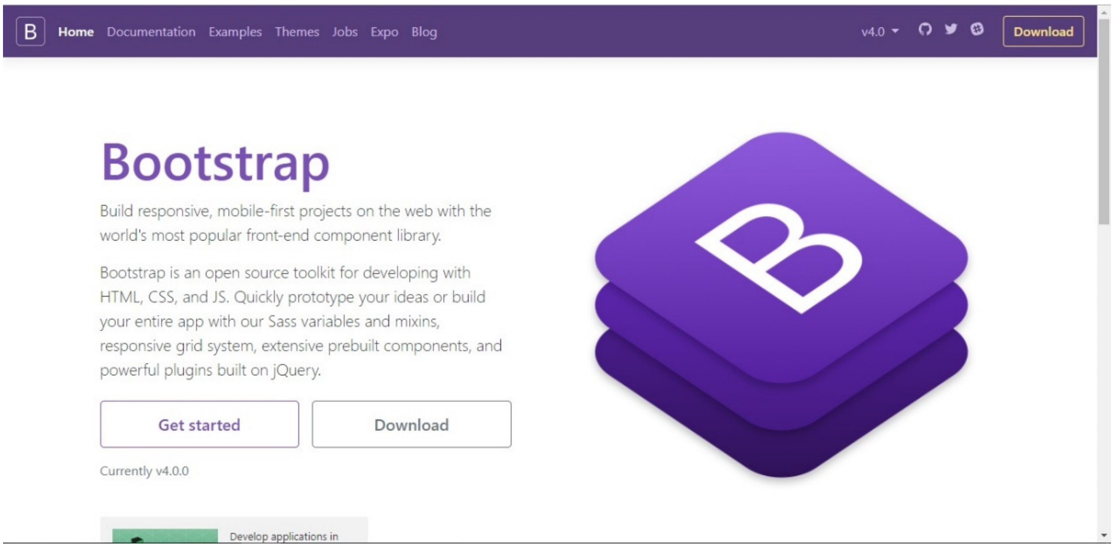
Figure 17: The bottom part of the prediction page, donut charts showing the predicted trend and the prediction confidences are provided.

2.2 Implementation

In order to construct the three main components of the online prediction system, we have implemented them by adopted various programming frameworks and relevant algorithms to achieve the goals in efficient ways

2.2.1Construction of web user interface

For the construction of the web user interface, it can be mainly divided into two major parts, the front-end layout and the back-end programming operations.
For the front-end layout and design, we mainly adopted the bootstrap layout framework to construct the basic appearance of the webpages.



Bootstrap

Build responsive, mobile-first projects on the web with the world's most popular front-end component library.

Bootstrap is an open source toolkit for developing with HTML, CSS, and JS. Quickly prototype your ideas or build your entire app with our Sass variables and mixins, responsive grid system, extensive prebuilt components, and powerful plugins built on jQuery.

Currently v4.0.0

Develop applications in

Figure 18: The main page and descriptions of Bootstrap framework

Regarding the display of daily price information, CanvasJs, a javascript-based library for data visualization was used to display the historical daily prices of selected forex currency pairs or stocks in the candlestick-chart format.

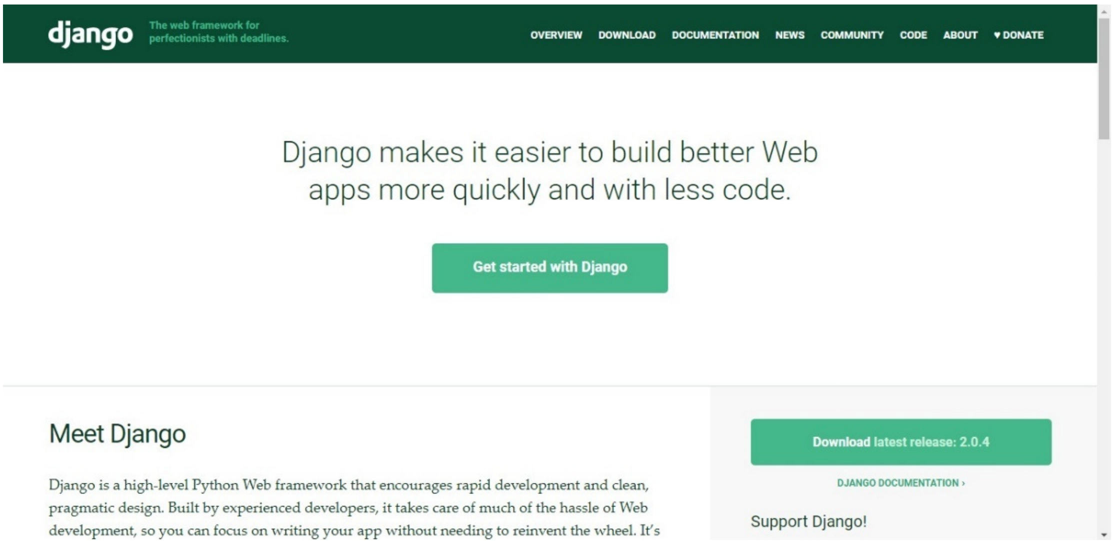
Beautiful HTML5 JavaScript Charts

Responsive HTML5 Charting Library with a Simple API and 10x Performance – makes your dashboards fly!

- Simple API
- 10x Performance
- 30 Chart Types
- Well Documented
- Supports Chrome, Firefox, Safari, IE8+
- Support Directly From Developers

Figure 19: The main page of CanvasJS chart library

For the back-end programming and logics, Django, a python-based web framework was used to construct the background programming logics and rendering of pages based on the front-end templates.



Django makes it easier to build better Web apps more quickly and with less code.

Meet Django

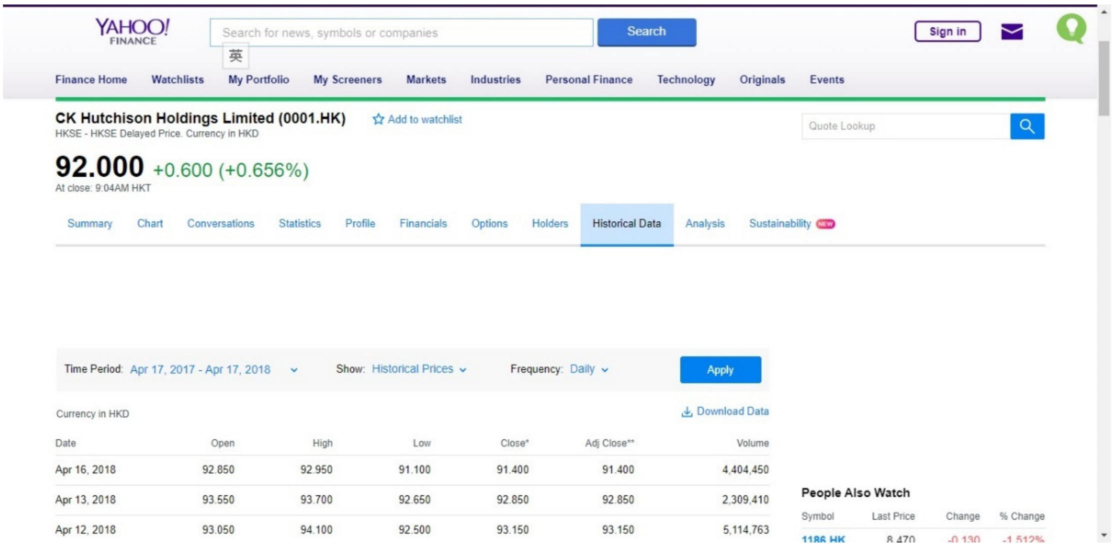
Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It's

Support Django!

Figure 20: The main page of Django web framework

2.2.2Data scraping and storage

In order to obtain the daily price data of the specified stocks and forexes, a data scraper has been constructed to download and parse the data from the targeted webpages and data sources. For stock prices, Yahoo Finance, a worldwide financial news and data provider, is selected as the data sources of daily prices in our project. For forex, Investing.com is selected as the forex prices data source.



CK Hutchison Holdings Limited (0001.HK)
HKSE - HKSE Delayed Price. Currency in HKD

92.000 +0.600 (+0.656%)
At close: 9:04AM HKT

Time Period: Apr 17, 2017 - Apr 17, 2018 Show: Historical Prices Frequency: Daily

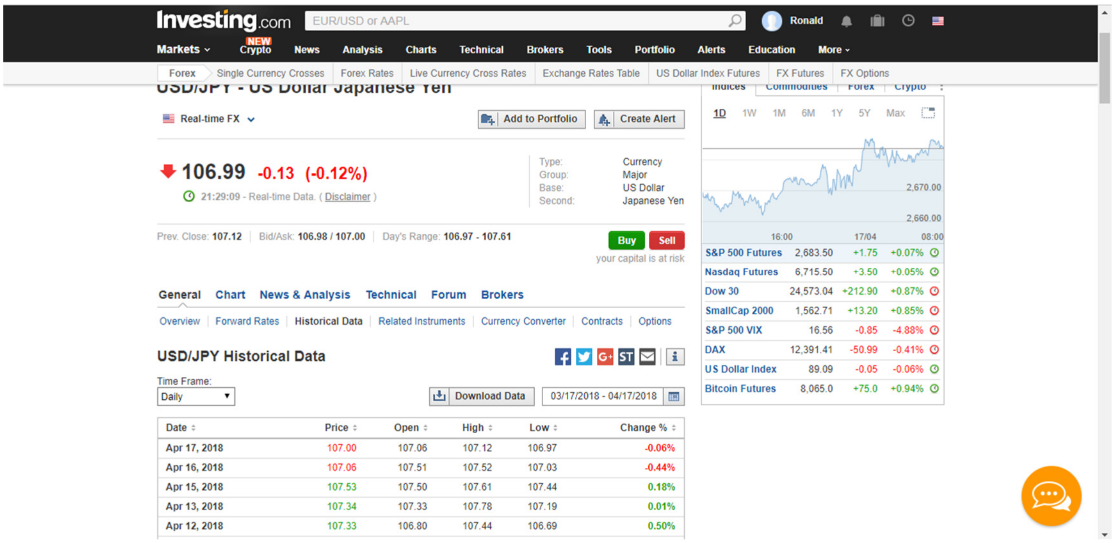
Currency in HKD

Date	Open	High	Low	Close*	Adj Close**	Volume
Apr 16, 2018	92.850	92.950	91.100	91.400	91.400	4,404,450
Apr 13, 2018	93.550	93.700	92.650	92.850	92.850	2,309,410
Apr 12, 2018	93.050	94.100	92.500	93.150	93.150	5,114,763

People Also Watch

Symbol	Last Price	Change	% Change
1186.HK	8.470	-0.130	-1.512%

Figure 21: The page of Yahoo Finance, notice that the table for historical prices exists.



USD/JPY - US Dollar Japanese Yen

106.99 -0.13 (-0.12%)

21:29:09 - Real-time Data. (Disclaimer)

Type:	Currency
Group:	Major
Base:	US Dollar
Second:	Japanese Yen

Prev. Close: 107.12 | Bid/Ask: 106.98 / 107.00 | Day's Range: 106.97 - 107.61

your capital is at risk

USD/JPY Historical Data

Time Frame: Daily

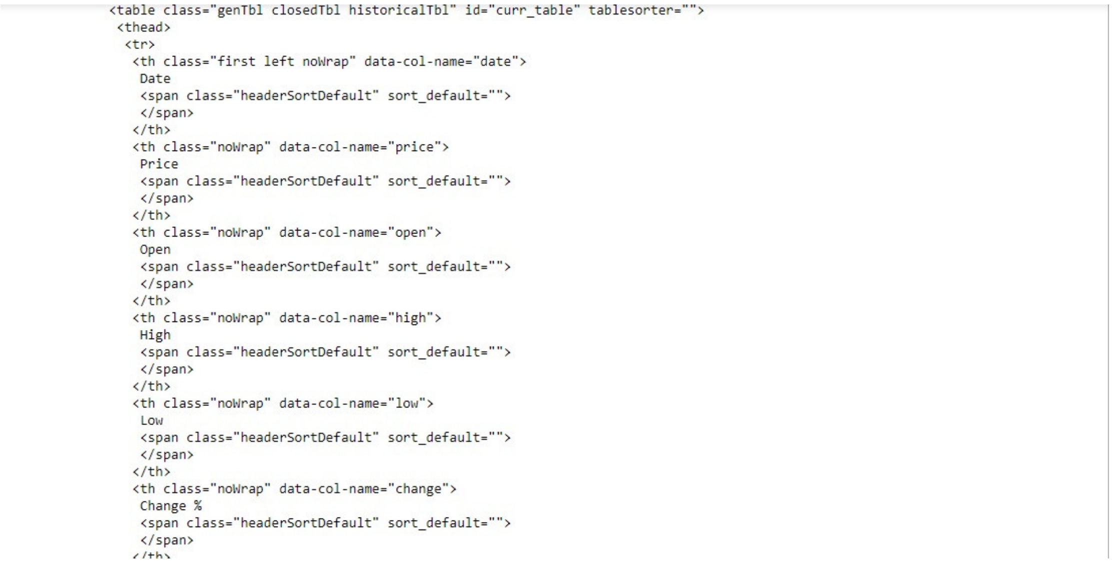
03/17/2018 - 04/17/2018

Date	Price	Open	High	Low	Change %
Apr 17, 2018	107.00	107.06	107.12	106.97	-0.06%
Apr 16, 2018	107.06	107.51	107.52	107.03	-0.44%
Apr 15, 2018	107.53	107.50	107.61	107.44	0.18%
Apr 13, 2018	107.34	107.33	107.78	107.19	0.01%
Apr 12, 2018	107.33	106.80	107.44	106.69	0.50%

S&P 500 Futures	2,683.50	+1.75	+0.07%
Nasdaq Futures	6,715.50	+3.50	+0.05%
Dow 30	24,573.04	+212.90	+0.87%
SmallCap 2000	1,562.71	+13.20	+0.85%
S&P 500 VIX	16.56	-0.85	-4.88%
DAX	12,391.41	-50.99	-0.41%
US Dollar Index	89.09	-0.05	-0.06%
Bitcoin Futures	8,065.0	+75.0	+0.94%

Figure 22: The page of Investing.com, notice that the table for historical prices exists.

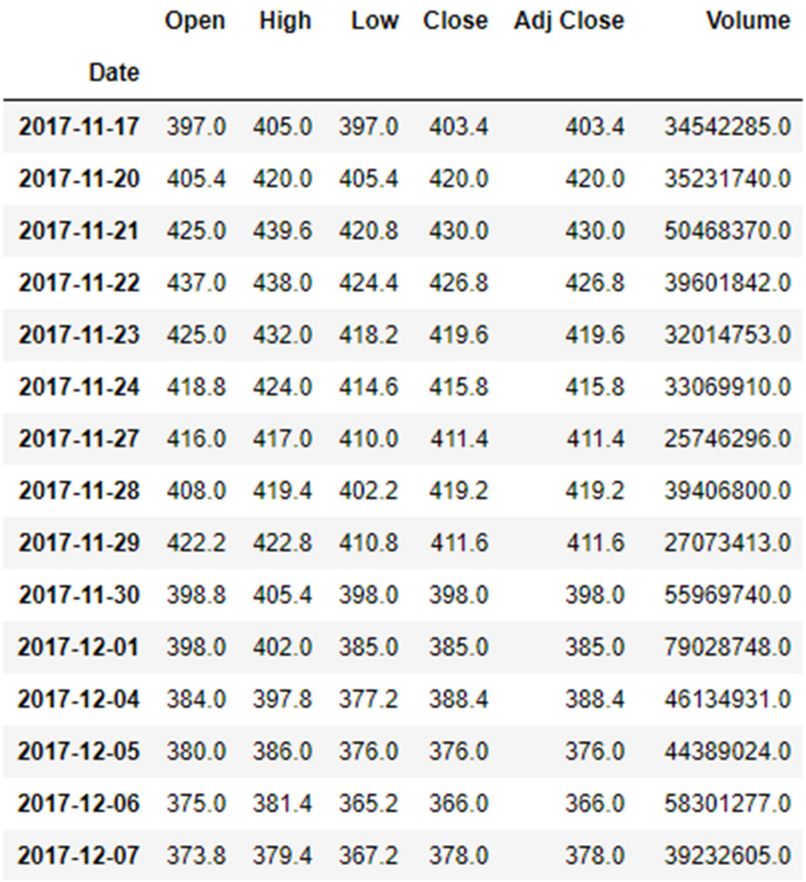
The price data are downloaded automatically everyday from the corresponding data sources in the format of html files through sending the “GET” requests from the server of the web application by using the “requests” library of Python.



```
<table class="genTbl closedTbl historicalTbl" id="curr_table" tablesorter="">
 <thead>
  <tr>
   <th class="first left noWrap" data-col-name="date">
    Date
    <span class="headerSortDefault" sort_default="">
    </span>
   </th>
   <th class="noWrap" data-col-name="price">
    Price
    <span class="headerSortDefault" sort_default="">
    </span>
   </th>
   <th class="noWrap" data-col-name="open">
    Open
    <span class="headerSortDefault" sort_default="">
    </span>
   </th>
   <th class="noWrap" data-col-name="high">
    High
    <span class="headerSortDefault" sort_default="">
    </span>
   </th>
   <th class="noWrap" data-col-name="low">
    Low
    <span class="headerSortDefault" sort_default="">
    </span>
   </th>
   <th class="noWrap" data-col-name="change">
    Change %
    <span class="headerSortDefault" sort_default="">
    </span>
   </th>
```

Figure 23: Example of contents inside the web page from the data sources, notice that the table element exists for further parsing and fetching of price data

After getting the content of the web pages, another python library, BeautifulSoup, is applied to parse the elements inside the downloaded html files. With the historical prices were stored as the “table” elements inside the html files, the historical prices can be obtained by parsing the table elements and their contents inside.



	Open	High	Low	Close	Adj Close	Volume
Date						
2017-11-17	397.0	405.0	397.0	403.4	403.4	34542285.0
2017-11-20	405.4	420.0	405.4	420.0	420.0	35231740.0
2017-11-21	425.0	439.6	420.8	430.0	430.0	50468370.0
2017-11-22	437.0	438.0	424.4	426.8	426.8	39601842.0
2017-11-23	425.0	432.0	418.2	419.6	419.6	32014753.0
2017-11-24	418.8	424.0	414.6	415.8	415.8	33069910.0
2017-11-27	416.0	417.0	410.0	411.4	411.4	25746296.0
2017-11-28	408.0	419.4	402.2	419.2	419.2	39406800.0
2017-11-29	422.2	422.8	410.8	411.6	411.6	27073413.0
2017-11-30	398.8	405.4	398.0	398.0	398.0	55969740.0
2017-12-01	398.0	402.0	385.0	385.0	385.0	79028748.0
2017-12-04	384.0	397.8	377.2	388.4	388.4	46134931.0
2017-12-05	380.0	386.0	376.0	376.0	376.0	44389024.0
2017-12-06	375.0	381.4	365.2	366.0	366.0	58301277.0
2017-12-07	373.8	379.4	367.2	378.0	378.0	39232605.0

Figure 24: The price data obtained from the web page after parsing and further preprocessing

The storage of price data is initialized by the previously available prices stored initially in .csv files, and to be updated by merging the latest prices obtained through the data scraper with these initialized data for further review and prediction purposes.

2.2 Construction of Prediction Models

For the short-term average price predictions, the models were implemented by using several supervised machine learning regression models and the recurrent neural network with the prediction goals defined as multi-steps time series predictions.

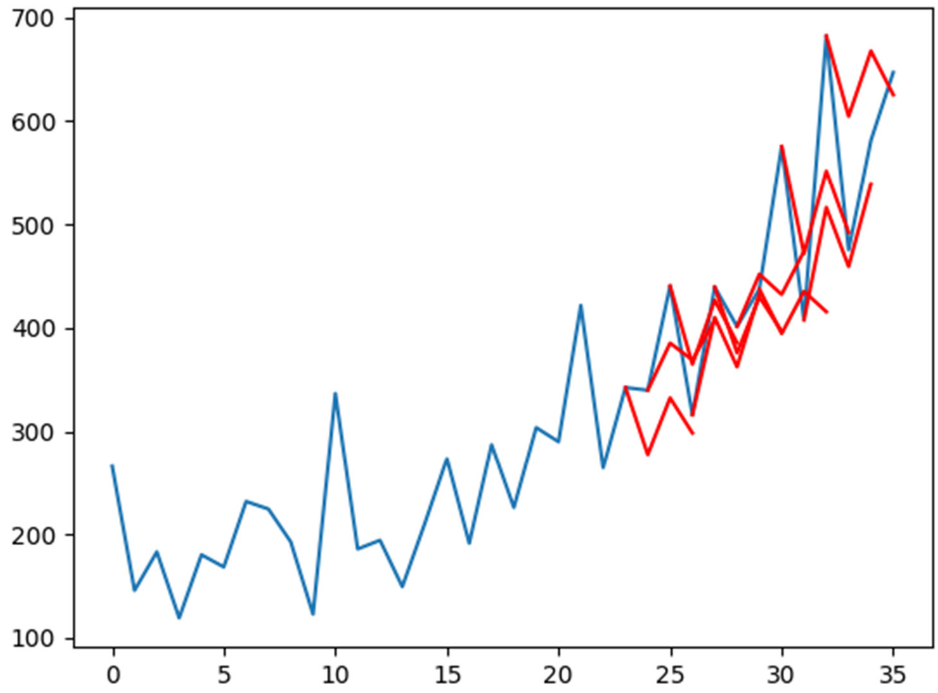


Figure 25: The illustration of a multi-steps time series prediction (x-axis refers to time, y-axis refers to the values), notices that the blue line refers to the previous data (prices in our project) and the red lines refer to the predictions provided in different instance of time, which included several time steps ahead.

In order to prepare the data for further construction of prediction models, several data preprocessing techniques have been employed to transform the data.

As the historical price data of stocks and forexes often involve obvious trend and fluctuations, transformation of the price series into stationary time series is necessary. In our project, the first-order differencing techniques has been applied to perform the transformation by subtracting every price in the series with the price with one time step behind. For example, the daily close price of 17 April is subtracted by the daily close price of 16 April, and the price of 16 April is subtracted with 15 April and so on.

Formula 3: The general equation of first-order differencing:

$$\mathrm{First\ \ order\ \ differencing} = \mathrm{P(t)} - P(t-1)$$

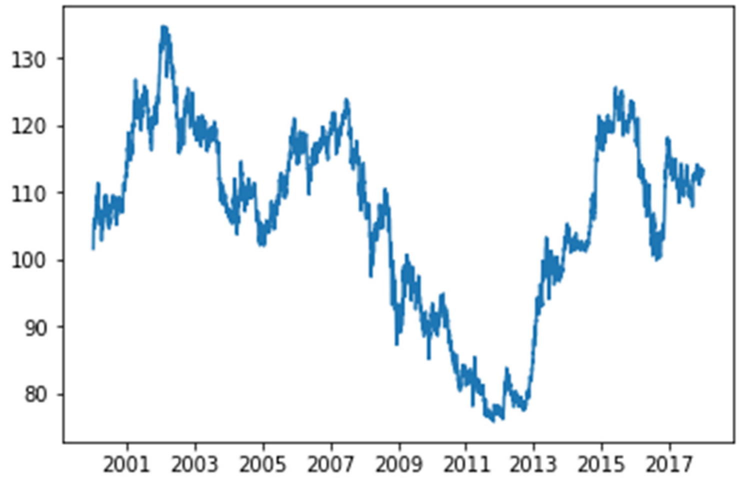


Figure 26: The original price series of USDJPY within the prediction time interval (x-axis refers to years and y-axis refers to the price) ,notice there are obvious trend and fluctuation with time.

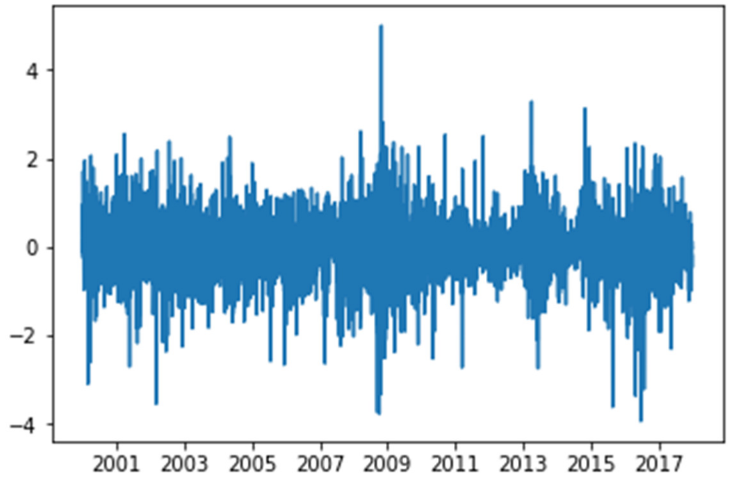


Figure 26: The transformed price series of USDJPY within the same prediction time interval (x-axis refers to years and y-axis refers to the price), notice the series tend to have steady mean and fluctuation.

Regarding the input features of the dataset for price predictions, the stationarized price series with 20 days time span has been selected for the weekly price predictions and 50 days time span has been selected for monthly prediction. That is, given today is 17 April, for weekly price predictions, the stationarized prices for everyday from one month before (around 17 March) will be included as the input feature of weekly price prediction of today, and the stationarized prices from 50 days before will be included as the input feature of monthly price prediction of today.

Apart from transformation towards stationary time series, standardization of price data is also employed to remove the mean and scaling to unit variance so as to ensure different features can have similar scale.

Formula 4: The general equation of standardization:

$$\text{Standardized price} = \frac{\text{price}_{\text{stationarized}} - \ mean}{(standard \ \ deviation)^2}$$

In order to perform reliable performance analysis of the trained models, The train-test ratio of splitting the dataset in the project is set as 66% :33%, which 66% of price data in previous time period were selected as training set for construction of models, and 33% of the price data in more recent time period were included in the test set. In the case of the price data starting from 1 January 2000 to 31 December 2017, the cut off of the train and test set is approximately at around the year of 2012, where the prices from 2000 to 2012 were included as the training set and the prices from 2012 to 2017 were considered as the test set.

In particular, instead of training the models individually for every stocks or forex, we have decided to combine the training set of every stocks and forex together based on their nature. For example, the training set of the stocks (0001HK, 0002HK,…,2388HK) were combined together as the combined training set for stocks, and the training set of the forexes (USDAUD, USDCAD…, USDSGD) were combined together as the combined training set for forexes, the prediction models for combined stocks and combined forexes then are trained separately. The reason we adopted the combination training approach is that the approach enables extension of training and test set size, with combination of all stocks and all forexes together based on their features, the size of training sets extends from thousands towards twenty to thirty thousands (20,000 to 30,000), which reduced the risk of overfitting and enable more training data available for the models to increase the prediction accuracies. At the same time, as the prices were stationarized and standardized, the training set of individual stocks and forexes thus can be combined with similar scales.

For the training of model, the following machine learning algorithms have been utilized:

1. Random Forests (With 100 decision trees for ensemble forecasts)
2. K-nearest neighboring (With k = 5 nearest neighbors)
3. Linear regression
4. Recurrent neural networks (with Long-short term memory(LSTM) neurons)
5. Ensemble mean of models

The models 1 to 3 had been trained by using the algorithms provided by scikit-learn, a Python machine learning library. Model 4 (Recurrent neural networks) had been trained by Keras and Tensorflow deep learning libraries.

For ensemble mean of models, it is basically calculated by obtain the mean of predictions by model 1 to 4.

Formula 5: Ensemble mean of multi-models:

$$\mathrm{Ensemble\ \ mean} = \frac{(\mathrm{Price}_{\mathrm{rf}} + Price_{knn} + Price_{LR} + Price_{rnn})}{4}$$

The reason of including ensemble means in the predictions is to average out the errors produced by different models and obtain their consensus to reduce the errors generated by individual models.

Apart from point prediction of prices, the 70% confidence intervals of prediction were also implemented in the project based on the prediction results of individual decision tree of the random forests model.
The individual decision tree of random forests generates diverse predictions. With the assumption of normal distribution for these diverse predictions, the envelop of 70% data can be obtained within 1.04 standard deviation of the diverse predictions mean.

Formula 6: Calculation of 70% prediction envelop:

$$70\%\mathrm{envelop}_{upper\ bound} = \ mean_{individuals} + 1.04\ \ SD_{individuals}$$

$$70\%\mathrm{envelop}_{lower\ bound} = \ mean_{individuals} - 1.04\ \ SD_{individuals}$$

$$\begin{aligned} &70\%\ \ \mathrm{prediction\ \ confidence\ \ intervals} \\ &\qquad = [\mathrm{Price}_{\mathrm{rf}} - 70\%\mathrm{envelop}_{lower\ bound}, \mathrm{Price}_{\mathrm{rf}} \\ &\qquad + \ \ 70\%\mathrm{envelop}_{upper\ bound}] \end{aligned}$$

The illustrations of the price predictions can be represented in the following four figures:

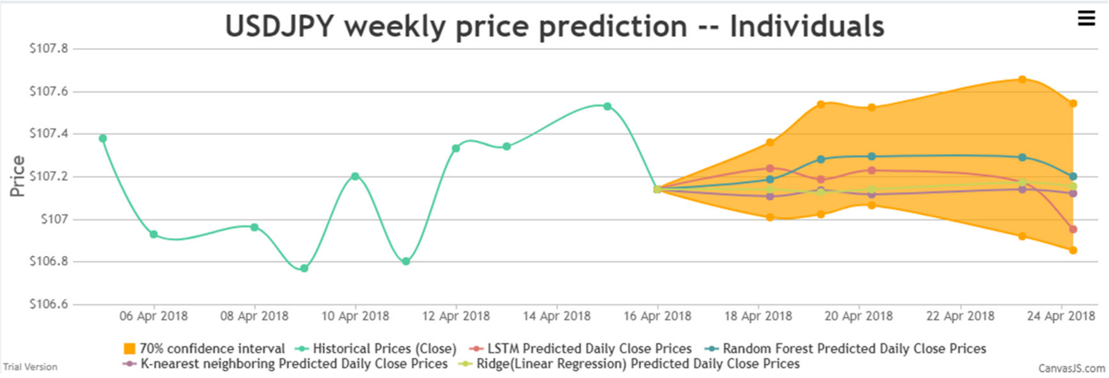


Figure 27: The weekly price prediction generated by different models and corresponding 70% envelop/confidence interval of prediction.

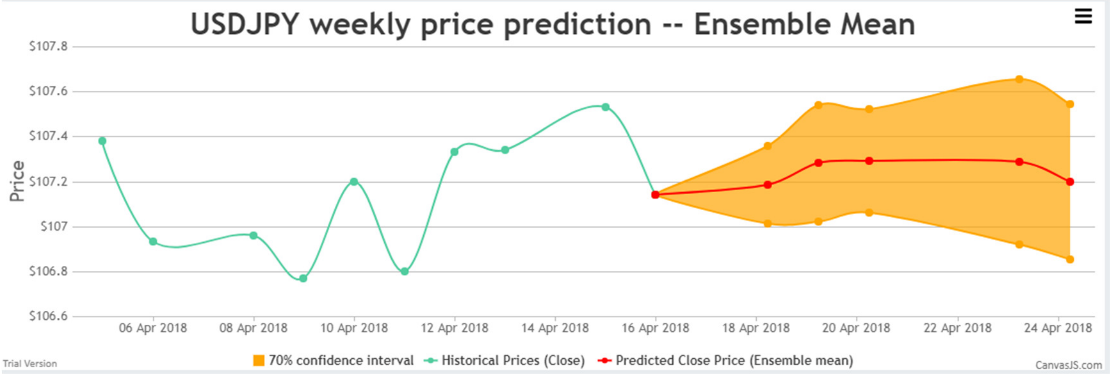


Figure 28: The ensemble weekly price prediction generated based on random forest model and corresponding 70% envelop/confidence interval of prediction.

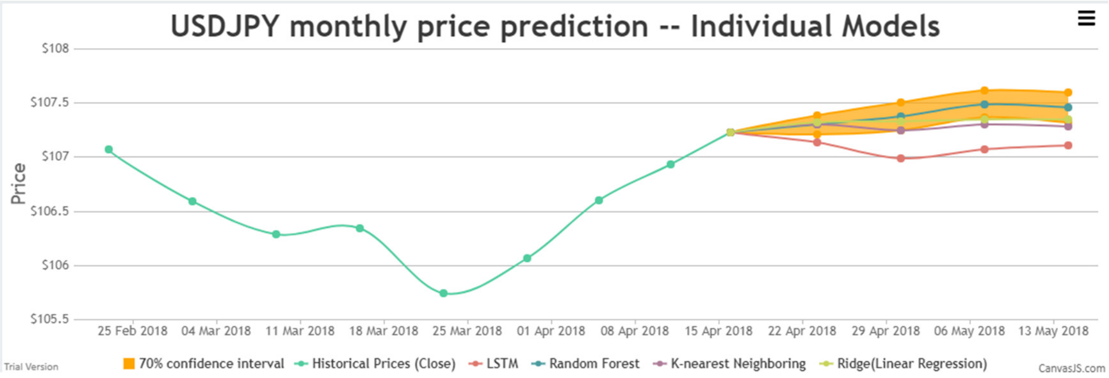


Figure 29: The monthly price prediction generated by different models and corresponding 70% envelop/confidence interval of prediction.

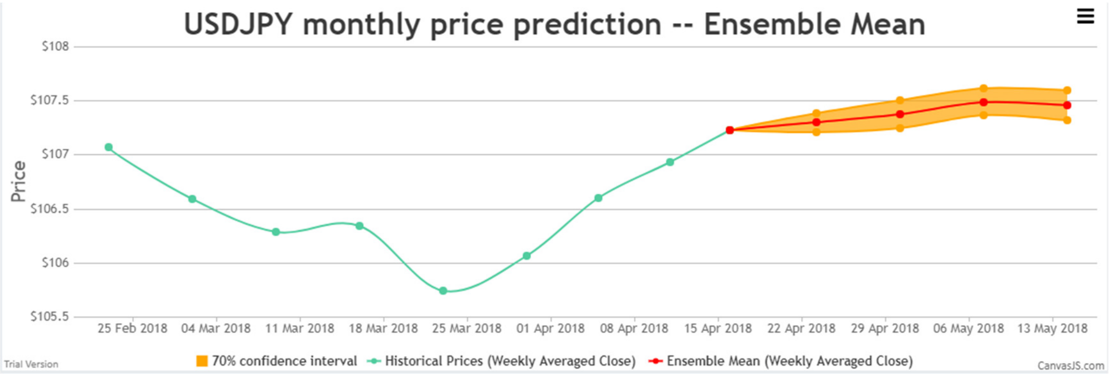


Figure 30: The ensemble monthly price prediction generated based on random forest model and corresponding 70% envelop/confidence interval of prediction.

For long-term price trend predictions, the prediction model was implemented as binary classification task based on the random forest model, with the following features included:

1. Relative Strength Index (RSI)
2. Moving Average Convergence/Divergences (MACD)
3. Differences between the daily price and the simple moving average in specified time frame (20-days SMA for monthly prediction, 60-days for quarterly prediction)
4. Differences between the simple moving average with specified time frame and the simple moving average with halve time frame (SMA(10) – SMA(20) for monthly prediction, SMA(30) – SMA(60) for quarterly prediction)

Similar to the standardization of short term price prediction tasks, standardization of features in trend prediction tasks are also performed in order to ensure the features have similar scales.

The trend prediction results were represented by the prediction trend and their corresponding probability, and to be represented by donut charts, as shown in the figures below:

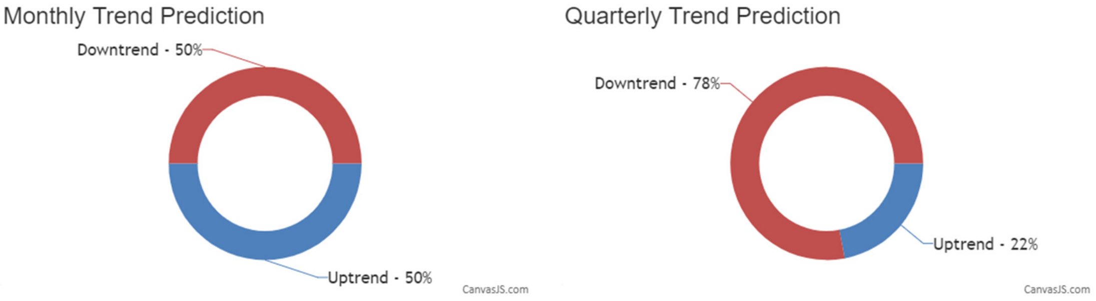


Figure 31: The monthly and quarterly trend predictions provided in our web application

The reason that we provided the probability of trend predictions is to provide a more informative prediction to the web application users, which they will be able to know how confident the possible future trends in the specified time period.

For the construction of prediction models and evaluation of model performance, we mainly based on the Python 3.6 platform with the relevant packages or libraries.

3. Testing and Evaluation

In order to analyze the prediction performance for the prediction models in both short term price predictions and long term price trend predictions, several metrics for performance and accuracy determinations have been chosen.

For short term price predictions, the mean absolute percentage error (MAPE) has been selected for measuring the error between the predicted prices and actual prices in terms of percentage, MAPE can be defined by the following formula:

Formula 7: The definition of MAPE:

$$\mathrm{MAPE} = \frac{100}{\mathrm{n}} \sum_{t=1}^{n} \left| \frac{Pt_{actual} - Pt_{predicted}}{Pt_{actual}} \right|, \mathrm{Pt} = \mathrm{price\ \ at\ \ time\ \ t}$$

For long term price predictions, the accuracy of prediction has been selected for measuring the degree of accuracy for the trend predictions, which can be defined as the following formula:

Formula 8: The definition of accuracy:

$$\mathrm{Accuracy} = \frac{\mathrm{number\ \ of\ \ correct\ \ predictions}}{\mathrm{total}\ \ number\ \ of\ \ predictions} * 100\ \ \%$$

Intuitively, with accuracy closer to 100%, the models are more able to produce prediction nearer towards perfect accurate.

With the train-test split ratio as 66%:33% mentioned before, the test sets used for performance evaluation were the more recent 33% of the whole dataset.

Regarding the performance evaluation of short term price predictions, the results can be summarized in the following table:

Table 4: The MAPE of weekly price predictions for stocks:

Model\Time	+1 business day	+2 business day	+3 business day	+4 business day	+5 business day
Random Forest	**1.185%**	**1.741%**	**2.150%**	**2.508%**	**2.812%**
Recurrent neural networks	**1.366%**	**2.017%**	**2.488%**	**2.861%**	**3.162%**
K nearest neighboring	**1.234%**	**1.806%**	**2.231%**	**2.602%**	**2.932%**
Linear regression	**1.162%**	**1.698%**	**2.104%**	**2.464%**	**2.776%**
Mean MAPE	1.237%	1.816%	2.243%	2.609%	2.921%

Table 5: The MAPE of weekly price predictions for forex:

Model\Time	+1 business day	+2 business day	+3 business day	+4 business day	+5 business day
Random Forest	**0.377%**	**0.534%**	**0.663%**	**0.768%**	**0.862%**
Recurrent neural networks	**2.014 %**	**1.951%**	**2.345%**	**6.921%**	**2.374%**
K nearest neighboring	**0.397%**	**0.556%**	**0.687%**	**0.793%**	**0.888%**
Linear regression	**0.378 %**	**0.539%**	**0.672%**	**0.785%**	**0.887%**
Mean MAPE	0.792%	0.895%	1.092%	2.317%	1.253%

Table 6: The MAPE of monthly price predictions for stocks:

Model\Time	+1 week	+2 week	+3 week	+4 week
Random Forest	**2.345%**	**3.818%**	**4.931%**	**5.835%**
Recurrent neural networks	**2.892%**	**4.432%**	**5.653%**	**6.915%**
K nearest neighboring	**2.649%**	**4.310%**	**5.506%**	**6.479%**
Linear regression	**2.308%**	**3.740%**	**4.816%**	**5.712%**
Mean MAPE	2.549%	4.075%	5.227%	6.235%

Table 7: The MAPE of monthly price predictions for forex:

Model\Time	+1 week	+2 week	+3 week	+4 week
Random Forest	**0.728%**	**1.164 %**	**1.475%**	**1.714%**
Recurrent neural networks	**6.012%**	**7.824%**	**5.246%**	**9.670%**
K nearest neighboring	**0.818%**	**1.320%**	**1.674%**	**1.959%**
Linear regression	**0.733%**	**1.162%**	**1.463%**	**1.802%**
Mean MAPE	2.073%	2.868%	2.465%	3.786%

The statistics marked as red refer to the entries with below average MAPE, which indicates a smaller error between predicted price and actual prices, and better accuracy in predicting prices in the corresponding time period (weekly or monthly).

According to the statistics shown in the tables above, it can be observed that the random forest and linear regression models in general perform better than average in both weekly and monthly price predictions.

Regarding the prediction accuracy, the combined test set of stocks consists of 13061 samples for monthly prediction and 12840 samples for quarterly prediction. The combined test set of forex consists 15243 samples of for monthly prediction and 15848 samples for quarterly prediction.
Both combined test sets in general start from around the year of 2012 to 31 December 2017.

The accuracy of the trend prediction in various time periods can be summarized with the table below:

Table 8: The MAPE of monthly price predictions for forex:

Type of asset\ Time period of prediction	Monthly	Quarterly
Stock	73.9%	76.1%
Forex	73.2%	77.2%

According to the accuracy percentages shown in the table above, it shows that the trend prediction models constructed by random forest perform much better than the prediction made by random guessing (which is expected to have around 50% accuracy in the case of random guessing of future trends).

4. Discussion:

It is originally expected that the recurrent neural networks model should perform relatively better price predictions than the other models with simpler structure, for example, the k-nearest neighboring and linear regression models, and complex models without consideration of the sequential ordering of features, like random forests. However, the accuracy metrics shown in the test and evaluation section reveal the completely opposite image. The recurrent neural network model tends to overfit with the historical price data and tried hard to completely replicate the historical price motion and unable to generalize well for the price data in the test set and the new coming price data. The evaluation results shown that the simpler models such as k-nearest neighboring and linear regression performed particularly well, especially in weekly price predictions, which may indicate that these simpler models are able to generalize well for new data and be able to strike a good balance between fitting the general price motion of the past data and fetching the correlations between the historical price motions and possible future change in prices. For monthly price predictions, random forest and linear regression performed better than the k-nearest neighboring and the recurrent neural networks models, which also revealed that the consideration of sequential ordering may not be as important as expected in the tasks for stock and forex price predictions.

For the monthly and quartering trend predictions, the trend prediction models constructed by random forest for both time periods are able to perform satisfactory results in predicting the possible trend with the accuracy much greater than 50%, which is a benchmark for distinguishing the accuracy between the prediction models and random guessing.

With the short term price prediction models and long term trend prediction models, we believe the users of the web application will be able to have a much more reliable information regarding the possible future price motions and trends and made better investment decisions.

5. Conclusion

To conclude, in this project we have implemented a web application for illustrating the prediction results on short term price prediction and long term price trend predictions by using various python libraries. For the construction of web user interface, we adopted the bootstrap framework to build a user-friendly and responsive web layout. For the backend web server logic, Django, a Python-based web framework has been used to construct the backend logics and execution of related python scripts for making predictions. Regarding the models that are responsible for making price and trend predictions, we have implemented several models based on different supervised machine learning algorithms like Linear regression, random forests and recurrent neural networks. It was discovered that the models with simpler algorithms tend to perform better than those with complex algorithms.

Due to the consideration of limited time, capitals and human resources, several proposed features and ideas did not implement in our web application. However, some of these ideas will be worth-considering to be the possible further improvement of this web application or other financial market related projects:

1. Deployment of the web application towards real production environment
2. Extension towards other types of financial assets, including encrypted currencies (for example, Bitcoin), Gold, Oil and etc.
3. Integration of prediction models with sentiment/textual analysis of social media (for example, twitter, facebook) and financial news articles.

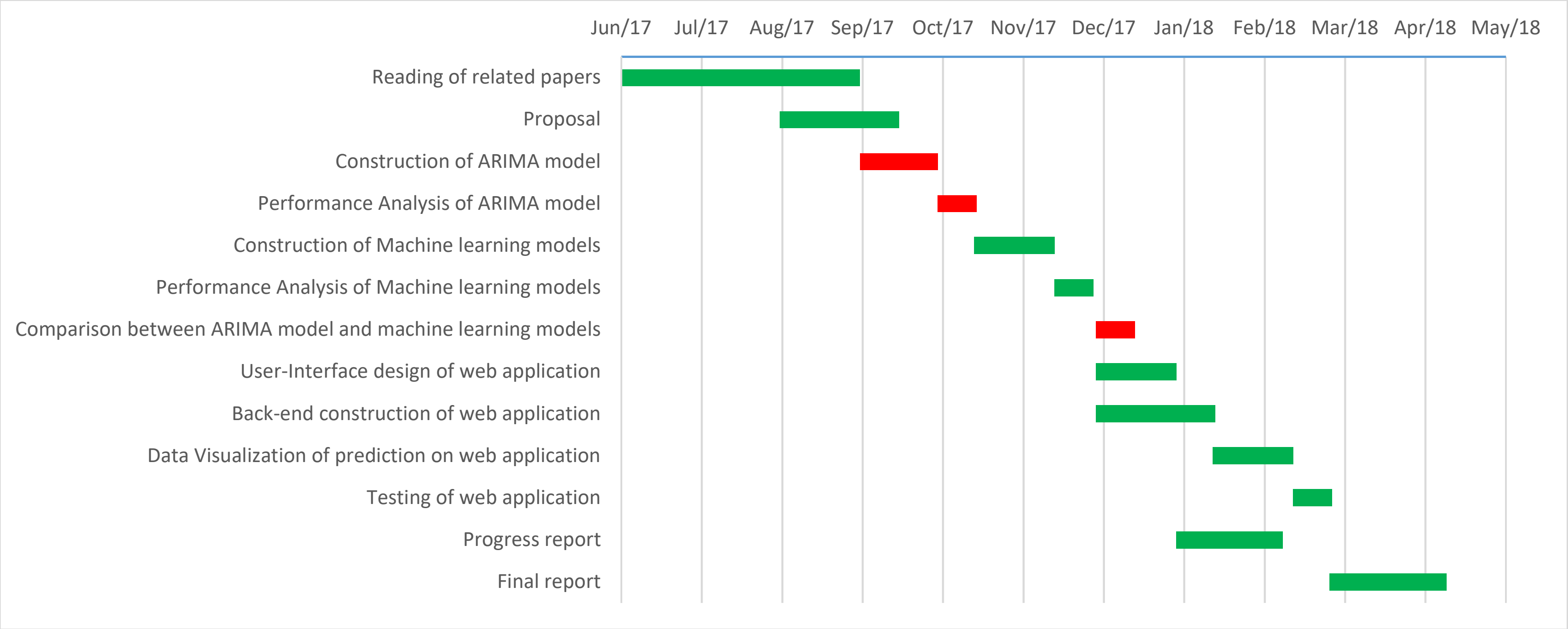
6. Project Planning

6.1 Division of work

Task	Ho Cheuk Yin (Ronald)	Wong Hoi Ming (Henry)
Reading of related papers	L	-
Proposal	L	-
Construction of Machine learning models	L	-
Performance Analysis of Machine learning models	L	-
Comparison between different price and trend prediction models	L	-
User-Interface design of web application	L	-
Back-end construction of web application (integration of database and front-end web application, Python script integration with web application)	L	-
Data Visualization of prediction on web application	L	-
Testing of web application	L	-
Progress report	L	-
Final report	L	-

L = Leader and person in charge

6.2 GANTT Chart



Green = completed tasks, Red = cancelled tasks, Orange = tasks in progress

7. Hardware and Software Requirements

7.1 Hardware Requirements:

- Development PC: PC with Windows 10 or MacOS

7.2 Software Requirements:

Software Packages	**Description**
Anaconda Python package (version 4.4.0)	Integration of libraries for machine learning and data science (numpy, pandas, scikit-learn and etc.)
Python 3.6.0	Main programming language for constructing and testing of models for predictions.
Django	Python-based web framework for integrating front-end web component to back-end programming operations and handling communication with database
Bootstrap	Front-end web framework for construction of user interfaces
Scikit-learn	Python-based machine learning library for construction of predictive models
Tensorflow	Deep learning framework for construction of deep learning models
Keras	High-level API for deep learning frameworks(e.g tensorflow)
Microsoft Visual Studio Code	Code editor for website construction.

8. References:

[1] Investopedia, “Technical Indicator”, 2018;
https://www.investopedia.com/terms/t/technicalindicator.asp

[2] Investopedia, “Simple Moving Average – SMA” , 2018;
https://www.investopedia.com/terms/s/sma.asp

[3] Investopedia, “Relative Strength Index – RSI” , 2018;
https://www.investopedia.com/terms/r/rsi.asp

[4] Investopedia, “Moving Average Convergence Divergence - MACD” , 2018;
https://www.investopedia.com/terms/m/macd.asp

[5] Investopedia, “ Autoregressive Integrated Moving Average – ARIMA” , 2018;
https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

Appendix A - Meeting Minutes

9.1 Minutes of the 1st Project Meeting

Date: September 12, 2017

Time: 4:00 PM

Place: Room 3554, Prof. Rossiter's office

Present: Professor Rossiter, HO Cheuk Yin (Ronald), WONG Hoi Ming (Henry)

Absent: None

Recorder: Ronald

1. Approval of minutes

This was the first formal group meeting, so there were no minutes to approve.

2. Report on progress

2.1. Started to review literature and research papers that we received in June related to stock price and exchange rate predictions from June.

2.2. Finished the draft of our proposal in August.

2.3. Started implementing the ARIMA model for foreign exchange rate prediction in September (the model for the USD/JPY currencies pair)

2.4. Started investigating features/indicators in technical analysis and macroeconomic aspects for future implementations of machine learning models for exchange rate predictions.

3. Discussion items

3.1. Professor Rossiter suggested we widen the scope of the investigation to developing models for financial assets instead of over-focusing on implementing only one model for exchange rate prediction.

3.2. Professor Rossiter pointed out the importance of linking the relationship and our project works to the project title

- We are considering a change to the project title in order to describe the project focus more accurately.

4. Goals for the coming month(October)

4.1 Start testing the ARIMA models for different currency pairs by measuring the

root-mean-squared-error of predictions and the accuracy of future trend predictions with type 1 / 2 errors parameters

4.2 Start implementing the supervised machine learning models (SVM, Random Forest and etc.) by using features of technical analysis (RSI, MACD, Momentum, Bollinger band and etc.) and macroeconomic indicators(GDP, CPI, unemployment rate and etc.) indicators

5. Meeting adjournment and next meeting

The meeting was adjourned at 4:30 PM.

The next meeting is expected to be held in the middle of October. The time and place will be set later by e-mail.

9.2 Minutes of the 2nd Project Meeting

Date: October 19, 2017

Time: 2:45 PM

Place: Room 3554, Prof. Rossiter's office

Present: Professor Rossiter, HO Cheuk Yin (Ronald), WONG Hoi Ming (Henry)

Absent: None

Recorder: Ronald

1. Approval of minutes

The minutes of the last meeting were approved without amendment.

2. Report on progress

2.1 Perform basic analysis of the prediction performances of several machine learning models on various forex currency pairs

2.2 Started develop the regression machine learning models for daily price forecasting.

3. Discussion items

3.1 Professor Rossiter asked if there are correlation between the features selected in the machine learning models have correlations with the predicted long-term trend.

3.2 Professor Rossiter recommended us to use Django web framework for effective construction of the web user interface

3.3 Professor Rossiter suggested us to insert some informative figures in our presentation report and slides for easier understanding and interpretation.

4. Goals for the coming week

4.1 Refining the price prediction models by adding more features, including major stock market indexes.

4.2 Start exploring the possible approaches of web app construction

5. Meeting adjournment and next meeting

The meeting was adjourned at 03:15 PM.

The next meeting is expected to be held in the middle of November. The time and place will be set later by e-mail.

9.3 Minutes of the 3rd Project Meeting

Date: November 23, 2017
Time: 2:45 PM
Place: Room 3554, Prof. Rossiter's office
Present: Professor Rossiter, HO Cheuk Yin (Ronald), WONG Hoi Ming (Henry)
Absent: None
Recorder: Ronald

1. Approval of minutes

The minutes of the last meeting were approved without amendment.

2. Report on progress

2.1 Perform basic analysis of the regression prediction performances of several machine learning models on various forex currency pairs and Hong Kong stocks

2.2 Started exploring the approaches for data visualization of daily price information.

3. Discussion items

3.1 Professor Rossiter suggested us to implement the prediction system as online so as to keep training the models with new daily price data

3.2 Professor Rossiter advised us to be aware of the labeling of figures / graphs for effective communication to the audiences.

3.3 Professor Rossiter advised us to choose representative forex currency pairs and Hong Kong stocks for our prediction system.

4. Goals for the coming week

4.1 Test the prediction models by trying more test data from various representative forex currency pairs and Hong Kong stocks..

4.2 Start construction of the web application

5. Meeting adjournment and next meeting

The meeting was adjourned at 03:30 PM.

The next meeting is expected to be held in the middle of December. The time and place will be set later by e-mail.

9.4 Minutes of the 4th Project Meeting

Date: December 19, 2017

Time: 11:00 AM

Place: Room 3523

Present: Professor Rossiter, HO Cheuk Yin (Ronald), WONG Hoi Ming (Henry)

Absent: None

Recorder: Ronald

1. Approval of minutes

The minutes of the last meeting were approved without amendment.

2. Report on progress

2.1 Make presentation of our progress from September to December

2.2 Started exploring the methodologies of making regression forecasts using recurrent neural network..

3. Discussion items

3.1 Professor Rossiter advised us to state our goal clearly in the presentation

3.2 Professor Rossiter suggested us to always perform high-level brainstorming before start coding tasks.

4. Goals for the coming week

4.1 Start construction of web user interface

4.2 Start preparation of the progress report.

5. Meeting adjournment and next meeting

The meeting was adjourned at 11:30 AM.

The next meeting is expected to be held in the middle of December. The time and place will be set later by e-mail.

9.5 Minutes of the 5th Project Meeting

Date: February 23, 2018

Time: 2:30 PM

Place: Room 3554, Prof. Rossiter's office

Present: Professor Rossiter, HO Cheuk Yin (Ronald), WONG Hoi Ming (Henry)

Absent: None

Recorder: Ronald

1. Approval of minutes

The minutes of the last meeting were approved without amendment.

2. Report on progress

2.1 Make presentation of our progress from December to February

2.2 Started construction of web user interface for the web application

3. Discussion items

3.1 Professor Rossiter suggested us to make use of the existing libraries to scrape price data from the web

3.2 Professor Rossiter suggested us to finish all coding task by the middle of April, one week before the submission of final report.

4. Goals for the coming week

4.1 Start construction of back end server logics and integration of prediction models with the web user interface.

4.2 Start constructing the data scrapers of price data based on several python library.

5. Meeting adjournment and next meeting

The meeting was adjourned at 03:00 PM.

The next meeting is expected to be held in the middle of March. The time and place will be set later by e-mail.

9.5 Minutes of the 6th Project Meeting

Date: March 28, 2018

Time: 2:30 PM

Place: Room 3554, Prof. Rossiter's office

Present: Professor Rossiter, HO Cheuk Yin (Ronald), WONG Hoi Ming (Henry)

Absent: None

Recorder: Ronald

1. Approval of minutes

The minutes of the last meeting were approved without amendment.

2. Report on progress

2.1 Make presentation of our progress from February to March

2.2 Continue construction of web user interface for the web application

3. Discussion items

3.1 Professor Rossiter suggested us to implement one possible new feature of the web application to show similar historical prices and corresponding stocks or forexes to the users

3.2 Professor Rossiter suggested us to finish all coding task by the middle of April, one week before the submission of final report.

4. Goals for the coming week

4.1 Continue construction of back end server logics and integration of prediction models with the web user interface.

4.2 Preparation of complete the final report.

5. Meeting adjournment and next meeting

The meeting was adjourned at 03:00 PM.

The time and place of the possible next meeting will be set later.