Short-Term Stock Price Prediction Models Based on Economic Background

Ruien Luo
Sino-Mac International Academy,
Shenzhen 518000, China
ruienluo@foxmail.com

ABSTRACT

In recent years, many different types of prediction models have occurred, and many of them are utilized in the stock market, to predict the stock price for a relatively short period of time, to speculate. Nonetheless, they have both advantages and disadvantages, some models are only reliable to a certain extent, or some specific conditions. In this case, predicting the fluctuation of the stock prices based on a single model possibly lead to an inaccurate value, and this paper aims to analyze the weakness and benefits of several models in these three categories, including pure Statistics, time series, and deep learning. And this article purposes to collect historical data, and utilize the data to analyze different models from the three categories, study the performances of different models in the various economic background, refer to the realistic macro data in different periods. Deriving a new method, giving the consequences from the models' new meanings according to the economic background, is the innovation of this study.

KEYWORDS

Stock market prediction(short-term), Traditional statistics, Time series, Deep learning, Economic background

1 Introduction

The stock market is one of the most significant financial markets, it provides many opportunities to investors and speculators, indicating the conditions of the enterprises. Nowadays, owing to the particular circumstances, the spread of the COVID-19 pandemic, leading to the recession, different stock markets are experiencing a more uncertain period. Especially, the stock market of the United States triggered the circuit breaker four times previously, which the stock index, the S&P 500 tumbled by 30% approximately in March. Moreover, the protection regulations of the stock market in China will be changed owing to the announcement, "t+1" the trading time will alter to "t+0", which means that the stock prices can fluctuate to a larger extent. Therefore, a grasp of the short-term stock prices forecasting skill is more significant.

Besides, the outcomes of the models based on the economic background, many macro indexes, the speculators and investors can avoid the trap of the investors that have a large amount of money, especially refrain this phenomenon in the A-share market. Many short-term predictions by the investors that only own a bit of stock are influenced by the investors who have an amount of stocks, owing to their attention to some micro indexes, such as MACD, moving average convergence divergence. Nevertheless, the utilization of models according to the macroeconomic background enables those investors to earn a profit.

Last but not least, the predictions of these models that combine with the economic background may avoid the huge economic bubbles that created by the over speculate in some conditions, like the A-share market's bubble in a very short period. The short-term prediction based on the economic background may provide fluctuations of a longer period, avoid the risk of trading within few days, and the disadvantages after the bubbles burst, protecting the stock market to a larger degree. The stock market's slump usually has a conductive impact on the economy, which leads to results like unemployment. Hence, the predictions based on the economic background is pretty significant.

2 Related Work

Contemporary, many algorithms that predict short-term stock price movements appeared and many works target machine learning methods, especially long short-term memory (LSTM). However, most of the works only use two or three models, which is not quite comprehensive; on the other hand, some works include different models, but it doesn't compare and contrast with the economic background of various periods. Whereas, this article concludes three categories of methods of short-term stock price movement prediction and compares them with the economic background to make the results more accurate.

Kai Chen, Yi Zhou, and Fangyan Dai utilize the long short-term memory(LSTM) model to forecast the stock prices of the short-term in China, and they found out that normalization and the SSE index increase the accuracy.

Xiaodong Li, Haoran Xie, Ran Wang, Yi Cai, Jingjing Cao, Feng Wang, Huaqing Min, and Xiaotie Deng utilize extreme learning machine(ELM), comparing with back- propagation neural

ICITEE2020, December 03–05, 2020, Changde City, Hunan, China

ACM ISBN 978-1-4503-8866-5/20/12…$15.00.
https://doi.org/10.1145/3452940.3452947

networks(BP-NN) and support vector machine(SVM), building trading signal mining platform.

Zou Cunzhu, Luo Jiping, Bai Shengyuan, Wang Yuanze, Zhong Changfa, and Cai Yi predict the stock prices based on long short-term memory(LSTM), and auto-regressive integrated moving average model(ARIMA), utilize the data of the S&P 500.

Naman Arora and Parimila M, utilize three models to predict and found out that long short-term memory(LSTM) performs better than both back-propagation(BP) and support vector machine(SVM)

Sanjiv R. Das, Karthik Mokashi, and Robbie Culkin offer a small experiment to assess the potential of a deep learning network, using the data from 1963 to 2016, finding the market efficiency, comparing several models.

Hongju Yan and Hongbing Ouyang explore a method that combines the wavelet analysis with long short-term memory (LSTM), and LSTM performs the best predicting the actual financial time series that can alert the risks.

Generally, most of the previous researches is based on limited models, such as LSTM and SVM. Most of these researches are limit to machine learning methods, especially deep learning, traditional statistic methods, and time series analysis; nevertheless, these works didn't comprehensively compare and contrast the outcomes with the economic background.

3 Algorithms

In order to obtain more accurate results, this paper utilizes 8 models from three different categories, pure statistic, time-series, and deep learning, that already exist, and combine with the economic background, involving many factors; then compare and contrast the outcome of different models in the different economic index and monetary policy.

- Random Forest algorithm
- Adapting boost algorithm(Adaboost)
- Naive Bayes model(NBM)
- Auto-regressive Integrated Moving Average model(ARIMA)
- Prophet model(fbprophet)
- Long Short-Term Memory model(LSTM)
- Gated Recurrent Unit model(GRU)
- Temporal Convolutional Network(TCN)
- Economic Background

3.1 Random Forest Algorithm

Random forest is a stable and strong explanatory model that construct by multiple decision trees. In this case, understand the decision tree is necessary. The decision tree is a method that represents the relationship between an object and the factors that influence the object. The principle of the decision tree is to utilize the decision points to replace different questions, utilize the probability branches to replace different outcomes possible, and eventually calculate different outcomes by algorithms. On the other hand, the more trees under the random forest algorithm, the better the generalization results.

Advantages:

◆ can handle both classification and regression features

◆ reduce the risk of over-fitting by averaging decision trees

◆ wrong predictions will only be made when over half of the base classifiers have errors, which means that it's more stable

Disadvantages:

◆ over-fitting is still appeared, especially when there's noise in the training data

◆ it's more complicated than a decision tree and has a higher computation cost, as well as the time of predicting and training

◆ not good at predicting when features are missing or when there are variables out of the range

3.2 Adapting boost Algorithm (Adaboost)

Adaboost algorithm is a famous and representative algorithm based on the boost theory, so understand boosting is essential. The basic principle of boosting is when there's a complicated problem, the determination by many experts is better than the determination by only one expert.

Adaboost is an adaptive iterative algorithm based on a boosting algorithm. Initially, all training samples are weighted the same, and the wrong classified sample will be used to train the next classifier. Plus, Adaboost is sensitive to both noise and abnormal data, and the accuracy of AdaBoost only needs to be better than random guesses, greater than 0.5. Since when the resulting multiple weak classifiers are linearly superimposed, they can be assigned coefficients, which can also improve the the outcome of classification.

Advantages:

◆ low generalization

◆ improve the accuracy of the classifier

◆ cooperate with other algorithms with weighting

◆ avoid over-fitting to a certain extent

Disadvantages:

◆ very sensitive to outliers, easy to get wrong, which gradually influences the classifiers

◆ difficult to determine the number of times of iteration, may need to utilize cross-validation

◆ training time is very long

3.3 Naive Bayes Model(NBM)

The Naive Bayes model is a classifier based on the Bayes' theorem, and it's necessary to understand the theorem first. Bayes' theorem is a theorem about the conditional probability of random events A and B, where $P(A|B)$ is the probability that A will happen if B happens. Based on this theory, the Bayesian classifier is designed, and the naive Bayes' model is derived from the Bayesian classifier.

The reason why Naive Bayes is naive is the assumption that attributes are independent, so if there is a strong relationship between the attributes, the accuracy will decrease. Contrasting to the Bayesian classifier, Naive Bayes' model is much simpler. In addition, Laplacian correction can increase the accuracy of this model. If one attribute doesn't appear in the training set with a category, then directly calculating will lead to an inaccurate consequence.

Advantages:

◆ originated in classical mathematical theory, it has a mathematical foundation and stable classification efficiency

◆ required few estimated parameters, is not sensitive to missing data

◆ the algorithm is simple and fast

Disadvantages:

◆ although it has the smallest error rate compared to other models theoretically, for some conditions, attributes are not independent of each other, there are relationships between them

◆ need to know the prior probability

◆ an error occurred in the classification decision

3.4 Auto-Regressive Integrated Moving Average Model (ARIMA)

ARIMA, the auto-regressive integrated moving average model, is one of the time series analysis models. ARIMA(p, d, q), is the combination of AR the auto-regressive(p the past value used to predict future value), MA the moving average(q the past predicted error used to predict future value), and I the integration(d the order difference), as well as extend of ARMA(p, q).

There are several steps to building the model. Firstly, get the time-series sequence of observations. Secondly, test it through the stability test. If it passes, then keep on, and if it doesn't pass, then do the difference operation, the order depends on the model. Keep repeat this step until it's stable. Next, do the white noise test, if it passes, the model is built, but if it is a "no", then do the fitting of the ARMA, auto-regressive moving average, until it passes.

Advantages:

◆ the model is very simple, only endogenous data are required

Disadvantages:

◆ have to be stationary

◆ can only capture linear relations

◆ relative not good at very-short term prediction

3.5 Prophet Model(fbprophet)

Prophet model is also a time series model, it is designed and developed by Facebook, and it can handle the situation where there are some outliers in the time series. Plus, it can deal with the deformation of some missing values, and it can almost predict the future trend. And the things that the prophet model do is the following steps.

a. input the time and corresponding value of a known time series

b. input the length of the prediction of the time series

c. output the future trend

d. output the trend that can provide statistical indicators, including indicators like fitting curve (YHAT), upper bound (YHAT_UPPER), and lower bound (YHAT_LOWER).

Advantages:

◆ flexibility, adjust the period and trend freely

◆ fast-fitting speed

◆ the parameters are explainable

Disadvantages:

◆ the model is relative easy, trends and seasonality could be both added or times, causing the under-fitting during training, couldn't learn certain complex ones

3.6 Long Short-Term Memory Model(LSTM)

Long short-term memory model, LSTM, is one of the recurrent neural network models(RNN). Moreover, LSTM is a special RNN, which can tackle the gradient problem in the long sequence.

There are 3 gates in the model LSTM. Among these 3 gates, any of them can be called a conventional artificial neuron, as in a multi-layer feed-forward network (Arora and M, 2019)

a. Forget gate, this gate is to forget the unimportant part and remember the remember part

b. Input gate, selecting the important parts in the input gate and remember them

c. output gate, this gate will determine whether it can be the output of the current stage.

Advantages:

◆ high accuracy

◆ strong parallel distributed processing capabilities

◆ store information, memorize

Disadvantages:

◆ require a large number of parameters

◆ learning time is long

3.7 Gated Recurrent Unit model(GRU)

There are many variants of LSTM, among which the best one is the Gated Recurrent Unit (GRU), which was proposed by Cho, et al. (2014). It combines the forget gate and input gate into a single update gate. The effect is similar to LSTM, but the parameters are one-third less. There are only two gates in GRU, update gate and reset gate.

a. reset gate: determine how influential is h(t-1) to h(t) the new memory

b. update gate: determine how much do h(t-1) transfer to h(t) the new memory, influencing the hidden state, which is the combination of h(t-1) and h(t)

Advantages(compare to LSTM):

◆ less parameters that avoid over-fitting

Disadvantages(compare to LSTM):

◆ lower accuracy to a certain extent

3.8 Temporal Convolutional Network(TCN)

The temporal convolutional network, TCN, is one of the variants of the convolutional neural network, CNN. And TCN includes causal and dilated convolution, and get rid of the gates and utilization of residual connections is the change of the TCN. TCN includes two layers of convolution and nonlinear mapping, and WeightNorm and Dropout are added to each layer to regularize the network.

Advantages:

◆ Flexible receptive field.

◆ Stable gradient.

◆ Lower memory.

Disadvantages:

◆ TCN may not have such strong adaptability in transfer learning.

◆ TCN is after all a variant of a convolutional neural network. Although the use of extended convolution can expand the receptive field, it is still limited.

3.9 Economic Background

First and foremost, with the structural change in the world by both technology and politics, many micro statistical indexes become less important, such as MACD, KDJ index, RSI, relative strength index, and so on. The periods of these indexes are relatively short, and the period of RSI is fourteen days, and usually nine days for KDJ. In this case, these indexes only apply to the micro bases, to a short-term base, so it possibly misses the trend of the big change in the world, owing to its inability of macro bases. However, the macro indexes such as M2 growth rate, entrepreneur confidence index, and FDI, foreign direct investment. These macro indexes could reflect the macro trend of the direction of development, which the investors can grasp the opportunities.

What's more, the eight models that the paper introduced are very unique and pragmatic in the research on the basis of economic background. The first model, random forest, is based on decisions, which is one of the most stable models related to logic. The second model is AdaBoost, which is special in that it can increase accuracy by combining different algorithms. The third model is the naive Bayes model since most of the macro indexes don't related to each other and the smallest risk theologically. The fourth model is ARIMA, which can reveal the macro trend well, though it lacks accuracy. The fifth model is Prophet, which possibly enables investors to react to some accidental events, such as holidays. The sixth model is LSTM, which is one of the best models that usually has the least RMSE, root mean square error. The seventh model is the GRU, which has fewer parameters that can avoid over-fitting, since there are many macro indexes, and this model is excelled in a certain area. The last model is TCN, the new model that could predict with stable gradient and less memory, might works better in relative short stock price movement. Therefore, these models are tested owing to their uniqueness and how the models cooperate with the macro indexes, including different categories, traditional pure statistical methods, time series, and deep learning.

4 Experiment

Many macro factors possibly influence the SSEC, Shanghai Securities Composite Index, which eventually affect the individual stocks. And all the experiment is under the prudent monetary policy.

Data of Factors collected:
1. M2 supply
2. M2 year-to-year growth rate
3. total GDP, Gross Domestic Product
4. GDP year-to-year growth
5. total FDI, Foreign Direct Investment
6. FDI month-to-month growth
7. number of new investors
8. month-to-month growth of the number of new investors
9. CPI, Consumer Price Index
10. year to year growth of CPI
11. PPI, Producer Price Index
12. PMI, Purchasing Managers' Index(manufacturing industry)
13. year-to-year growth PMI(manufacturing industry)
14. the total market value of the Shanghai Stock Exchange
15. the difference between the total market value of the Shanghai Stock Exchange
16. the total turnover of the Shanghai Stock Exchange
17. the difference between the total turnover of the Shanghai Stock Exchange
18. the total share price of the Shanghai Stock Exchange
19. the difference of the total share price of Shanghai Stock Exchange
20. CCI, the consumer confidence index
21. year-to-year growth of consumer confidence index
22. month-to-month growth of consumer confidence index
23. entrepreneur confidence exponent
24. year-to-year growth of entrepreneur confidence exponent
25. month-to-month growth of entrepreneur confidence exponent
26. prosperity index of enterprises
27. year-to-year growth of prosperity index of enterprises
28. season-to-season growth of prosperity index of enterprises
29. customs export value
30. year-to-year growth of customs export value
31. customs import value
32. year-to-year growth of customs export value
33. total retail sales of consumer goods
34. year-to-year growth of the total retail sales of consumer goods
35. month-to-month growth of the total retail sales of consumer goods
36. new credit
37. year-to-year growth of the new credit
38. month-to-month growth of the new credit

Data of Factors selected:

1. M2 year-to-year growth rate: change point of the trend(change maintains at least 1.5 months)= trough(0.5 of change from both previous and later or over 1 of change for one)+4.5 months; acceleration(affect 8 weeks that is nearest to the point)= peak(0.5 of change from both previous and later or over 1% of change for one)+5.5 months; weight: 1, 20.87%
2. FDI: change point of the trend(change maintains at least 1.5 months)(bear market)= peak(above 13.7 billion dollars)+2.25 months; change point of the trend(change maintains at least 1.5 months)(bull market)= peak(above 13.7 billion dollars)+4.5 months; weight: 0.376, 7.85%
3. month-to-month growth of number of new investors: change point of the trend (change maintains at least 1.5 months)= peak(month-to-month growth above 50%)+1.75 month; weight: 0.506, 10.56%
4. change of PMI: change point of the trend(change maintains at least 0.5 month, later ones combines with the previous ones)= [the absolute value of the sum of the three continuous changes that are in the range of (-0.1 to 0.5 or -0.5 to 0.1) is equal or less than 0.6, and the sign of the first change is different than the third sign] +4 months; weight: 0.65, 13.56%
5. season-to-season growth of the entrepreneur confidence exponent(season): change point of the trend(change maintains at least 1.5 months)= the season that the season-to-season growth of the entrepreneur confidence exponent(season) above 5%+2.25 months; weight: 0.875, 18.26%
6. month-to-month growth of consumer confidence index, CCI: change point of the trend(change maintains at least 1.75 months) = the month that the month-to-month growth of consumer confidence index(below -2.3%)+2.25 months; weight: 0.8, 16.69%
7. month-to-month growth of the new credit: change point of the trend(change maintains at least 1.75 months)= the month that the month-to-month growth of the new credit(above 100%) +4.5; weight: 0.585, 12.21%

Figure 1&2: 38 data collected and the algorithm of the 7 data selected(strong related data)

This study utilizes the close price of SSEC.

Figure 2: Close of SSEC

Figure 3: M2 year-to-year growth rate

Figure 4: FDI

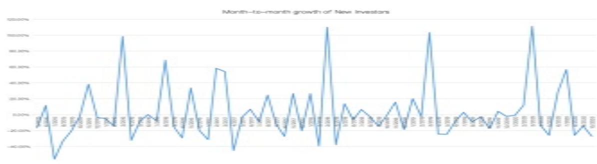

Figure 5: New investors month-to-month growth rate

Figure 6: CCI month-to-month growth

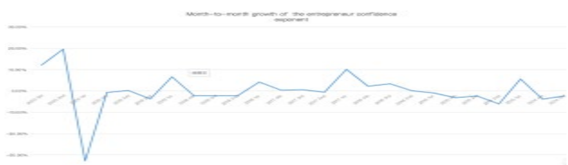

Figure 7: Entrepreneur confidence exponent month-to-month growth

Figure 8: Change of PMI

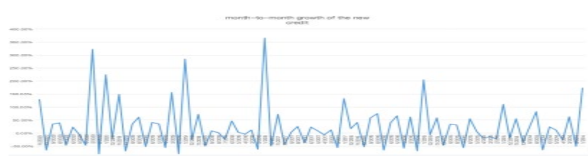

Figure 9: New credit month-to-month growth

These seven factors are strongly related to the stock price and it has been weighted by accuracy. A predicted rise is numbered as 1 and a predicted drop is numbered as 0. If the weighted and combined outcome is greater than 0.5, then it's a predicted rise, vice versa. The more the outcome closer than 1, the more possibility that it will rise, vice versa.

	M2	FDI	new investors	change of PMI	entrepreneur confidence exponent	CCI	new credit	outcome	real trend
7/20	-	-	-	-	-	1	-	1 rose	3025.98 rose
7.5/20	1	-	-	-	-	1	1	1 rise	3361.3 rise
8/20	1	1	-	-	1	0.5	1	0.89 rise	3367.97 rise
8.5/20	1	1	-	-	1	0.5	1	0.89 rise	3360.1 rise
9/20	1	1	-	-	1	0	1	0.78 rise	3410.61 rise
9.5/20	1	1	0	0.5	1	0	-	0.61 rise	3295.68 rise
10/20	0.5	-	0	0	-	-	-	0.23 drop	3218.05 drop
10.5/20	0.5	-	0	0.25	-	-	-	0.35 drop	3332.18 rise
11/20	-	1	-	1	1	-	-	1 rise	3225.12 drop
11.5/20	-	1	-	0.5	1	-	-	0.83 rise	3346.97 rise
12/20	-	1	-	0	1	-	-	0.65 rise	3451.94 rise
12.5/20	-	1	-	0.5	1	-	-	0.83 rise	3367.23 drop

Figure 10: Experiment

7/20 means the start of July, 2020, and 7.5/20 means the middle of July, 2020. This experiment is based on the data from 2014 to 2020, and eventually utilizes the newest data(from 7/2020 to 12/2020) to calculate the accuracy of the trend, which is 0.75(the actual trend is based on the stock price per day). The argmax is the fluctuations of the M2 year-to-year growth rate that reach the requirement. There are 3 errors and 2 of them that could be improved, the holiday factor(National day 10.1) and the presidential election of the United States, and these are both micro factors. The lack of the economic background model is that it can only provide the trend but not the accurate price, so the accuracy cannot be calculated regularly using RMSE. On the other hand, this model is limited since there's restrictions of the data, so the model is limited in some conditions, especially in some periods of the bear market.

Ultimately, many models that are not sensitive to the changes of the significant economic background, and that the reason why some model like ARIMA isn't that pragmatic. Nevertheless, the model that refers to the economic background could solve this problem to a certain extent, although some micro factors might affect the accuracy. Therefore, if this economic background could be an adjustment for these eight models, then it could be extraordinary.

5 Conclusion

In conclusion, this paper analyzes the influences of economic background and introduces and analyzes eight algorithms, including the Random Forest, Adaboost, Naive Bayes model, ARIMA model, Prophet model, LSTM algorithm, GRU algorithm, and TCN algorithm. And this paper introduces the principle of each model, explaining why it's extraordinary referring to the economic background, and analyzes the advantages and disadvantages, which some of the models can be complements to a certain extent, such as GRU and LSTM, accuracy and fitting. In addition, Adaboost can combine the models as a weak learner, which can fit the prediction based on the macro data better. The reason why combining the prediction is owing to the specific condition this year and the new direction of the world, which means that based on the macro data possibly is more reliable. Moreover, this study collects thirty-eight macro data, and some of it has a strong direct relationship with SSEZ. And the weighted combined data reached the accuracy of 0.75. Furthermore, comparing to many short-term trades that trade in a couple of days, the short-term trade that is based on the economic background is relatively longer, which is approximately from two weeks to a month according to the indexes changes. However, during this period, investors can avoid some of the risks and possibly notice the peak better. What's more, this can also avoid over-speculate, which could cause a series of severe issues. Hence, the short term prediction refers to economic background could be more advantageous possibly.

6 Future Work

Since the model based on the economic background can only provide the trend in some situations that fit the model, it's limited. If the model can combine with the eight models above as a modifier, choosing the best model that suits it, and consider some micro events, then it would be extraordinary.

REFERENCES

[1] Arora, Naman, and Parimala M. Financial Analysis: Stock Market Prediction Using Deep Learning Algorithms, 24 Feb. 2019, ssrn.com/abstract=3358252.
[2] Jin, Jerry. The Principle and Difference Between RNN&LSTM&GRU. 27 Sept. 2018, www.cnblogs.com/jins-note/p/9715610.html.
[3] Sim, Hyun Sik, et al. Is Deep Learning for Image Recognition Applicable to Stock Market Prediction?, 19 Feb. 2019, doi.org/10.1155/2019/4324878.
[4] Tipirisetty, Abhinav. "Master's Theses and Graduate Research." Stock Price Prediction Using Deep Learning, May 2018, doi.org/10.31979/etd.bzmm-36m7.
[5] Yan, Hongju, and Hongbing Ouyang. Financial Time Series Prediction Based on Deep Learning. 6 Dec. 2017, doi.org/10.1007/s11277-017-5086-2.