A LSTM-based method for ETF close prediction : A case study of Taiwan stock market

***Abstract*—**

Predictions on stock market prices are a great challenge due to the fact that it is an immensely complex, chaotic and dynamic environment. There are many studies from various areas aiming to take on that challenge and Machine Learning approaches have been the focus of many of them. Short-term price prediction on general stock using purely time series data of stock price is the domain that currently has the worst prediction accuracy. This article studies the usage of LSTM networks, XGBoost, and SVR to predict future trends of stock prices based on the price history, alongside with technical and chip analysis indicators. We applied complicated models to pre-process the time series data before running ML models. We didn't find any works that combined these financial technical and chip indicators with machine learning algorithms like we did. And we got better results than other papers we found in this particular problem domain. The results that were obtained are promising, getting up to an average of 70% of accuracy when predicting if the price of a particular stock is going to go up or not by percent in the near future.

Keywords—stock prediction, classification, LSTM

I. Introduction

The stock market has long been characterized by its dynamic, complicated, and non-stationary nature [1]. Researches about stock price prediction have flourished in recent years [2], [4], [5]. Market movements are dependent upon various factors ranging from political events, firms policies, economic background, commodity prices, exchange rates, movements of other stock markets to a psychology of investors [3], [6], [7], [8]. In this paper, we will focus on short-term price prediction on general stock using time series data of stock price. From our preliminary findings, we know that SVM with a Gaussian kernel performs well, so we decided to implement support vector regression on the data to see whether we can get a better result.

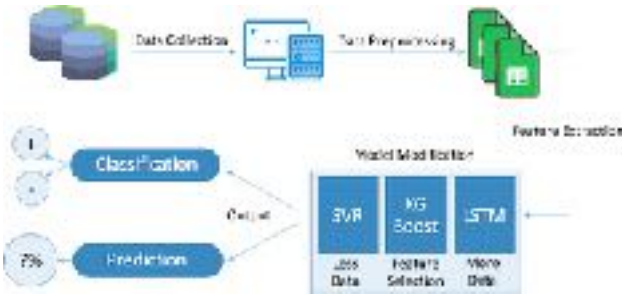
Long Short-Term Memory (LSTM) is one of the most successful RNNs architectures. LSTM introduces the memory cell, a unit of computation that replaces traditional artificial neurons in the hidden layer of a network.

A support vector machine (SVM) is a supervised learning technique from the field of machine learning applicable to both classification and regression. Rooted in the Statistical Learning Theory developed by Vladimir Vapnik and co-

workers at AT&T Bell Laboratories in 1995, SVMs are based on the principle of Structural Risk Minimization. And SVR (support vector regression) is on SVM （support vector machine） for regression, we use SVR to predict ETF which data is less than five years.

XGBoost (eXtreme Gradient Boosting) is one of the most loved machine learning algorithms at Kaggle [9]. Teams with this algorithm keep winning the competitions. It can be used for supervised learning tasks such as Regression, Classification, and Ranking, we use XGBoost to find out the most relevant parameters

II. Method



A. Data Collection

We acquired the historical stock data for Taiwan stock market from Taiwan Stock Exchange (daily record of high/low/ open/close/volume), Taiwan Futures Exchange (futures/options/institutional investors), Bloomberg (news). It has 47500 daily records of 18 ETFs from 2008/1/1 to 2018/06/22. To initialize the training process, we apply 15 consecutive daily data for predicting next five days and also train the time gap as a variable. One sequence contained 15 consecutive daily stock data. We used 90% of sequences for training purpose and 10% sequences for validation.

B. Data Preprocessing

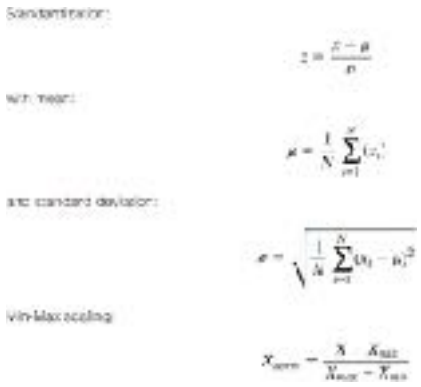
- Cleaning

The objective is to structure the data to facilitate the data analysis we set out for prediction. Variable manipulations such as aggregation, filtering, reordering, transforming and sorting are used in our case. On top of the historic price data, in order to reduce random variation and noise on the pricing series, data inserting was performed through sequential data (ex. the value of

yesterday) to cause improvements on the prediction capability. Also on top of the price data, A binary class y is assigned to each entry of the dataset, ”1” will indicate that the price will go up on the following time step, and ”0” that it won't. Therefore, given that i is the current moment and j is the following, then j = i + timestep, and for this project timestep is equivalent to 5 days.

- Scaling

Historic price data for particular stocks are gathered in the format of a time series of candles (open, close, high, low and volume). With the data in hand, a log-return transformation is performed as means of normalization as well as to stabilize the mean and variance along the time series. In order to normalize the data, we apply the result of standardization. Features will be rescaled so that they'll have the properties of a standard normal distribution with μ=0 and σ=1
where μ is the mean (average) andσis the standard deviation from the mean; standard scores (also called z scores) of the samples are calculated as follows:



C. *Feature Selection*

A set of technical indicators is generated using domain knowledge by our partner traders. Such indicators are mathematical calculations intended to determine or predict characteristics from stocks based on their historical data. A total of 40 features are selected from technical and chip analysis, and they are intended to represent a very diverse set of characteristics of the stock, like the future price, volume to be traded, the intensity of the current movement tendency among others. Main categories of features are daily record of high/low/open/close/volume, futures/options/institutional investors and news.

D. *Model Modification*

We applied LSTM for more data to enhance the accuracy of predictions. However, some ETFs data are not available online, so they are not enough to fit the data amount of LSTM. The accuracy of LSTM model for less data is extremely low to 10%. Therefore, we chose SVR to solve the problems of less data. To reduce noises of insignificant features, features are needed to filter for SVR mode. Applying XGBoost could solve this problem and select features which are most related to classification and prediction for future values of ETF.

III. Experiment and Result

This model predicts the daily closing price and the ups and downs in the coming week through the data collected over the past few years. The way to calculate the ups and downs is to compare the closing price of the day with the previous day. In order to implement this system, we use Keras to construct LSTM model. The code is as the following:

Fig. code

```
model.add(LSTM(
    neurons[1],
    input_shape=(days, num_of_features),
    return_sequences=True))
model.add(Dropout(0.2))
model.add(GRU(
    neurons[2],
    return_sequences=False))
model.add(Dropout(0.2))
model.add(Dense(neurons[3],kernel_initializer="uniform",activation='relu'))
model.compile(loss='mse', optimizer="adam", metrics=['mse'])
```

The results produced by each model are shown below.

A. Long Short Term Memory (LSTM)

Fig. Loss function displays the Mean Squared Error (MSE) during the process of the training model.

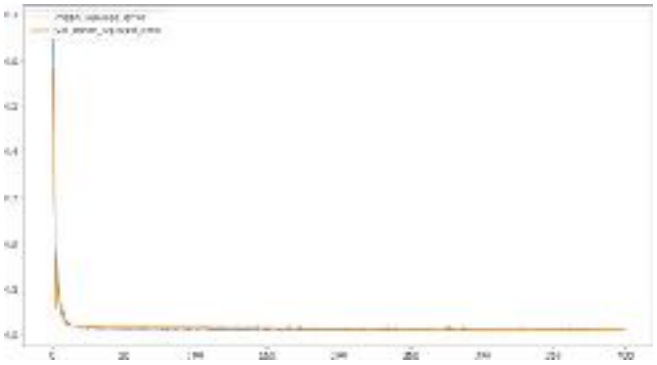


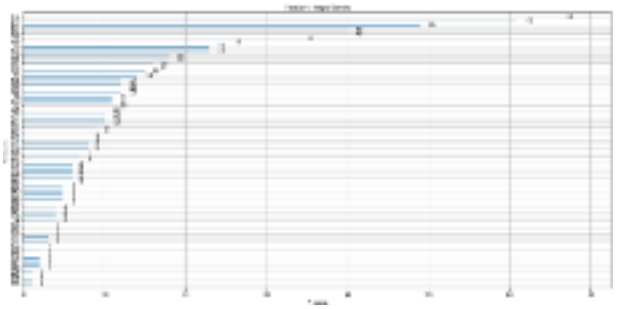
Fig. LSTM result shows the trend of stock closing prices.

Table. Results of Mean Square Error for stock predictions via compared methods

Methods	Features	MSE
Random	NA	0.6476
M1	Close, Volume	0.9780
M2	Close, Volume	0.6322
M3	High, Low, Open, Close, Volume	0.6223
M4	M3 plus features	0.6539
M5	M3 plus put call	0.6552

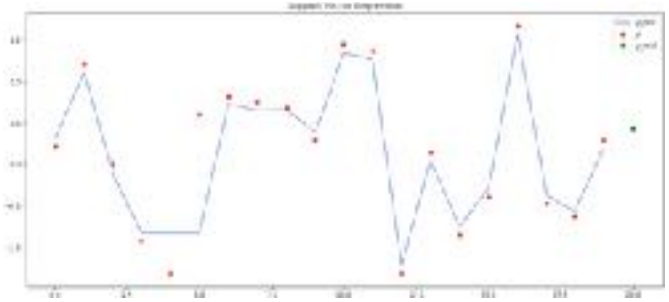
B. EXtreme Gradient Boosting (XGboost)

Besides, in order to decide the importance of our feature, we apply Extreme Gradient Boosting(XGboost) to our model. We obtain much information from Taiwan Futures Exchange and use those as the input to select the feature. From the graph below, we can see that the top 6 results are "Number of transactions"(f30), "Put/Call Ratio"(f0), "XIF Futures Top 10"(f12), "GTF Futures Top 5"(f1), "TX Futures Top 5"(f9), "Opening Price"(f25). The importance of Futures data is significant, so we decide to use Futures data as our feature.



C. Support Vector Regression (SVR)

We use the SVR model to predict the amplitude of the ups and downs. The interval is 20 days, as shown below. The red dots are actual values, and the blue polyline is the result of the model prediction. Generally, we see that correct rate can reach 60% to 70%.



IV. Conclusion and Future Work

Our value is predicting accurately for the first consecutive value by applying enough data for LSTM. Therefore, we could provide results of classification and values for traders to narrow their possible strategies. During this work, we encounter some difficulties, for example, the accuracy of our model fluctuates. The ups and downs are very difficult to predict due to the real- time information. Adding instant news to our model is needed for future improvements.

References

1. E. F. Fama, "The behavior of stock-market prices," *The journal of Business*, vol. 38, no. 1, pp. 34–105, 1965.
2. K. Kim and I. Han," Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index," *Expert Systems with Applications*, vol.19,no.2,pp.125- 132,August 2000.
3. G. Gidofalvi and C. Elkan, "Using news articles to predict stock price movements," *Department of Computer Science and Engineering, University of California, San Diego*, 2001.
4. P. C. Chang and C. H. Liu," A TSK fuzzy rule based system for stock price prediction," *Expert Systems with Applications*, vol.34,no. 1,pp.135-144,January 2008.
5. X. Lin, Z. Yang, and Y. Song," Short-term stock price prediction based on echo state networks," *Expert Systems with Applications*, vol. 36,no.3,pp.7313-7317,April 2009.
6. J. Bollen, H. Mao, and X. J. Zeng," Twitter mood predicts the stock market," *Journal of Computational Science*, vol.2,no.1,pp.1- 8,March 2011.
7. N. P. Committee, "Understanding asset prices," Nobel Prize Committee, Nobel Prize in Economics documents 2013-1, 2013.
8. X. Ding, Y. Zhang, T. Liu, and J. Duan," Deep Learning for Event-Driven Stock Prediction,"*24th IJCAI 2015*.
9. T. Chen and C. Guestrin," Xgboost: A scalable tree boosting system," *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,pp.785–794.