금융공학 석사학위 논문

수리적 모델과 딥러닝을 이용한 코스피200 선물 변화 예측

Prediction of KOSPI200 Future's Change Using Mathematical Model and Deep learning

아주대학교대학원

금융공학과

김기범

Prediction of KOSPI200 Future's Change Using Mathematical Model and Deep learning

지도 교수 : 배 형 옥

이 논문을 금융공학 석사학위 논문으로 제출함.

2018년 2월

아주대학교 대학원

금융공학과

김기범

김기범의 금융공학 석사학위 논문을 인준함.

심사위원장 배형옥 인

심사위원 구형건 인

심사위원 김하영 인

아주대학교 대학원

2017년 12월 21일

Abstract

The most talked-about thing in any fields ever since 2016 is machine learning. Attempts to apply machine learning along with the emergence of *AlphaGo* have been taking place all over the industry, and have also received much attention in the financial sector. In the financial sector, machine learning has begun to be studied in a variety of financial areas such as investment banking, trading, and credit evaluation, especially, in the prediction of future stock price trends, researches were actively carried out by methods other than existing methods. The past studies of predicting stock price were the making mathematical modeling or complex statistical modeling, but in this case, there were critical problem to apply in a real situation because the prediction rate is low or the model is too generalized not to explain a spot.

This paper aims to predict the price change of KOSPI200 futures using LSTM method which is a one of the machine learning techniques. A total of 33 variables are used, and the VPIN model and the Rama Cont model, which were modeled for predicting and statistically superior, are included in the LSTM model. Prior to this, the data is used to generate transaction unit data, which is used in the HFT (High Frequency Trading) field, we designed a market microstructure model to meet the price change in a short period of time.

Keyword : VPIN, CST model, HFT, machine learning, LSTM

Contents

Chapter Ⅰ. Introduction

Section ⅰ. Background and Purpose of the Paper

In this paper, we try to predict the situation of price changes in the stock market. If we can predict the flow in the stock market, we can pursue high returns and design stable portfolios. Many researchers have done a lot of mathematical studies and statistical studies to find models with good prediction rates. Recently, studies using machine learning algorithms have become active.

In this study, I have also studied the algorithm using the machine learning, especially LSTM(Long-Short Term Memory) which is one of the recurrent neural networks. LSTM is an algorithm that is widely used in various fields by solving the problems of existing recurrent neural networks. Therefore, this study is differentiated from other papers that predicted the price change (flow) of the stock market using LSTM at two aspects.

First, I used the data which is the HTF(High Frequency Trading) raw stock data.

Second, the raw data is originated in mathematical model value, which makes better output.

The data of this study were collected from the raw data of KOSPI 200 futures, and the data were processed to derive numerical values as technical variables and mathematical models. The step is training, testing and final simulation investment by using the LSTM model. The data was collected through Xing api of *eBest Investment Securities* and then used the simulated trading system.

The evaluation of the LSTM model set in this study is divided into two parts. First, we compare the accuracy of the model with the existing model. Secondly, we compare the yield through the simulated trading with the holding yield without any investment activity using the set model.

The main contributions of this study are as follows.

1) New prediction model about tick data based on deep learning technology

2) Verification of actual forecast using actual data (KOSPI 200 futures)

3) Comparison and analysis of forecasting rates with existing models

Section ⅱ. Previous Research and Industry Trends

Research to predict future stock price's changes has been one of the key research for the stock market at the past. In the past, there has been an active field of modeling stock prices through a mathematical approach. This approach was the same through micromarket. The typical approaches are Time series models which is represented of economic and statistical model. There are a lot of statistical model which are included Johansen cointegration test, AR model, VECM model and so on. In addition, there are other famous model like the Madhavan model which interpreted the main components such as volatility and spread, the Kyle model which classifies three market participants (informed trader, liquidity trader and market maker) and the participants' strategy make volatility, and the volatility forms stock price, and PIN model whose modeler David Easley and three others make Kyle model's value easy by using MLE[1]. These models attempted to explain the stock price change through the mathematical modeling, assuming a few specific situations, so there were some cases where the explanatory power in the actual situation was poor.

The existing studies explain the movements of the stock price through mathematical approach, but recently, the rapid development of hardware reevaluated the big data and machine

[1] Maximum Likelihood Estimation

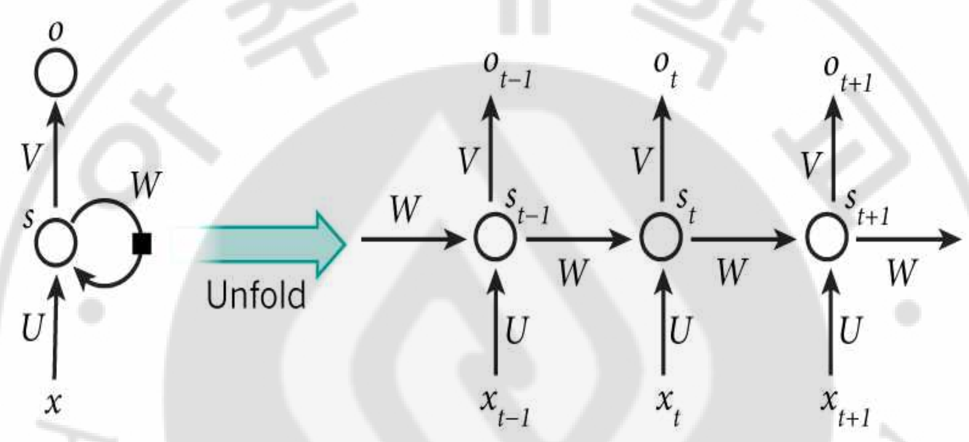
learning field, and research and development using two fields for stock prediction began active. Koscom has built a *Big Data Center* for stock forecasting models based on big data. Overseas, companies such as Kensho, AlphaSense, and Deep Value have developed and commercialized stock forecasting models based on Big Data. Various studies have been carried out in the field of prediction and prediction using machine learning. In general, the stock price prediction was based on Support Vector Machine. Anthony Lai et al. created a DBN (Deep Belief Network) model for predicting the direction and size of futures contracts. Son and two others published a paper that predicts the tendency of KOSPI 200 Index using learning classification.

Section ⅲ. Configuration of the Paper

Chapter 1 introduces the background and purpose of the paper, the existing research and industry trends, and the configuration of the paper. In Chapter 2, I explain the concept of artificial neural networks and the properties of LSTM. In Chapter 3, I introduce the mathematical models which used to process the data in LSTM. Chapter 4 explains the data and LSTM model required for learning, and Chapter 5 analyzes the results using the actual data processed in Chapter 4 using the LSTM model. Also, I simulate investment to seek the return using the most optimal LSTM model. Finally, Chapter 6 will conclude the conclusion and future work.

Chapter Ⅱ. Artificial Neural Network

In chapter 2, I introduce LSTM among the artificial neural network algorithms applied to study the stock price prediction model using big data.



2)

Figure 1. General structure of RNN

Section ⅰ. The Concept of RNN

RNN (Recurrent Neural Network) literally means that networks have a cyclical structure in learning. The advantage of this recursive structure is that it is affected by previously generated information, in other words, it memorizes the past information so that it affects the current output value. While other artificial neural models determine the output value through the relationship

[2] Figure Reference : http://aikorea.org/blog/rnn-tutorial-1

of the input values at a specific time, the RNN can be said to determine the output value considering the relation with the input values that is, the past input value. Due to the nature of these RNN, there are many applications where information of the past is important. like the field of natural information processing and time series analysis.

x_t: Input value at time t

s_t : Hidden state at time t. It is calculated by the hidden state value of the previous time point (t-1) and the input value of the current time point (t) to the memory part of the network.

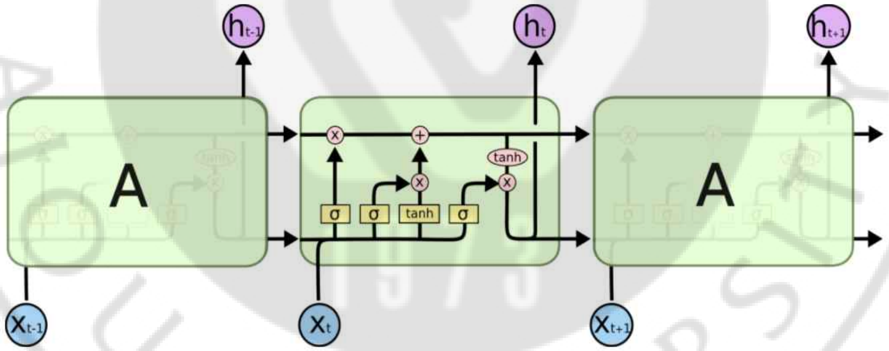
o_t : Output value at time t

The o_t of RNN is solely determined by the value s_t which is mirrored by the information of the past time points. However, the data at too far away points are insignificant. Also, in the general artificial neural network model, the variables are different from each other in each layer, RNN has a recurrent feature, and all the variables are shared at the specific point of time. U, V, and W in Figure 1 mean this. However, the biggest problem with RNN is that long - term dependency problems arise. To solve this problem, LSTM emerged.

Long-term dependency problem

The long-term dependency problem means that the RNN does not properly reflect the data at the distant point mentioned above. In Hochreiter and Benjio's paper, "Learning Long-Term Dependencies with Gradient Descent is Difficult" (1994) explained it, and sovled the problem by using LSTM

Section ⅱ. Concepts of LSTM



[3]

LSTM is a kind of RNN introduced by Hochreiter and Schmidhuber in "Long Short-Term Memory". LSTM has solved the long-term dependency problem in that RNN vulnerability, which

[3] Figure Reference : Understangding LSTM Networks, (2015, August 27) from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

gets solved by LSTM easily to memorize the distant past information.

The LSTM has a chain structure like a general recurrent structure, but has different point compared with structure of the RNN at single module. The LSTM determines the final output value by adding the operation of the elements in each chain gate to a connection line called a cell state. The gate is a device for selectively determining information, and is used to judge the influence of the data on the future through a sigmoid layer σ which has output value 0 and 1. Let's look at how LSTM learn the information through cell states and gates.

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

σ : Sigmoid layer

W_f : h_{t-1}, x_t weight

b_f : bias

The first step in learning LSTM is through the Forget gate layer which is made with a sigmoid function.

The Forget gate determines whether to keep or forget the input variables h_{t-1}, x_t through the sigmoid layer. At this time, if the output value is 1, the value is held. If the output value is 0, the value is discarded.

2. Input gate : $i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i)$ (2)

3. Cell state : $\widehat{C}_t = \tanh(W_c \bullet [h_{t-1}, x_t] + b_c)$ (3)

The second step is to determine whether to store the new incoming information in the cell state. Like the Forget gate, the information passes through the input gate layer which is made with the sigmoid layer. In this section, it is determined which value is to be updated, and in the cell state, new information is generated through the tanh function.

4. Update Cell state : $C_t = f_t \bullet C_{t-1} + i_t \bullet \hat{C}_t$ (4)

The fourth step is to update to the new state using the f_t, C_{t-1}, i_t, $\widehat{C}_t$ created through the above processes.

5. Output : $O_t = \sigma(W_\sigma [h_{t-1}, x_t] + b_o)$ (5)

6. Final value : $h_t = O_t \bullet \tanh(C_t)$ (6)

The final step is to produce the final output value by multiplying the value of the updated cell state and the output value through the sigmoid gate and the updated cell state. In this case, the cell state generates a value between -1 and 1 via the tanh function, and multiply the cell state and the sigmoid output to complete the final output of the LSTM.

Chapter Ⅲ. Market Microstructure

Market microstructure is an area of economics that is derived from economics such as microeconomics and econometrics. A key aspect of microeconomics is how each participants distribute resources to maximize their profit in economic activities. From a general economics viewpoint, they are naturally distributed by market activity in a single balanced price and optimal distribution, but supposing that the trading process in the market is static and collective, there is an additional discussion was needed. It is the market microstructure that studies this process in more detail. It is also the purpose of the market microstructure to see how the outcome of price formation and distribution changes when information outside the market affects price and when the market or transaction structure are changed. In Section 3, I will discuss the theory of the market microstructure used in data processing to learn through LSTM.

Section ⅰ. Trading participants

The most basic principle in microscopically observing the securities market is to classify the market participants into three major categories.

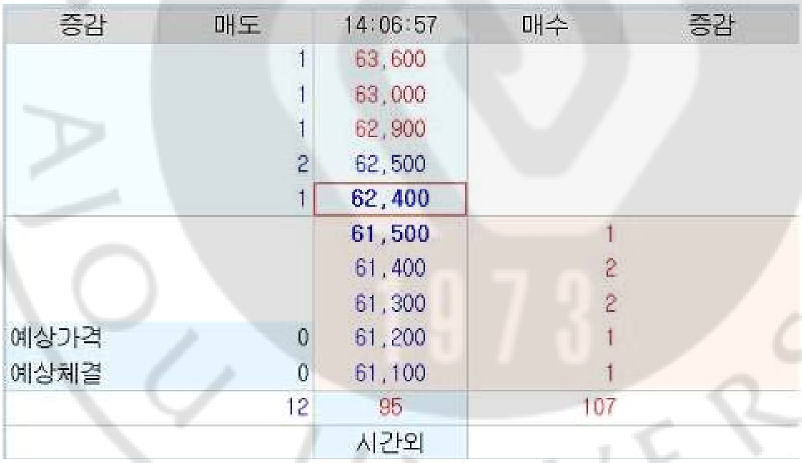
- **Informed Trader** : It is a trader who can predict the direction of the stock price through his own market information. In general, market information refers to information which is open to the public, that is, information which is not publicly prevalent in the market, but is not disclosed to the other participants or to the market. Therefore, informed traders are generally highly likely to be inside traders. Since informed traders know the direction of stock prices, I assume that they bought stocks through market orders, which are mostly liquidity-consuming forms. The fact that informed traders buy or sell stocks is an act that exposes their information to the market. Therefore, they participate in the market through strategies that minimize the leakage of their information.

- **Market Maker** : As a market participant who does not have information, it refers to a trader who observes transaction patterns of informed traders, estimates the information held by informed traders, and calculates the value of the information to supply liquidity to the prices. Therefore, it is generally assumed that they will participate in the market through limit orders.

- **Liquidity Trader or Noise Trader** : Like market makers, they do not have the information, but they trade for their own purposes. As the name *Noise Trader* suggests, the transaction pattern is random because they are doing their own business.

Section ⅱ. Order Book Data

In order to distinguish the above mentioned market participants, we need to observe the records in the market. In general, these records are recorded and transmitted at the central stock exchange of the stock market. These records are called tick data. There are many ways to get price data, but to get the data that is the closest to raw data, you have to use API provided by securities firm.



증감	매도	14:06:57	매수	증감
	1	63,600		
	1	63,000		
	1	62,900		
	2	62,500		
	1	62,400		
		61,500	1	
		61,400	2	
		61,300	2	
예상가격	0	61,200	1	
예상체결	0	61,100	1	
	12	95	107	
		시간외		

Figure 2. Order Book

The tick data is divided into quotation information and contract information. The quotation information refers to the information such as the price, the remaining amount, and the number of orders in each order book. The remaining amount means the number of purchases or the quantity to be sold at each price point. In other words, it is the quantity of waiting for entering the limit order and waiting for the contract. On the other hand, the market order will exhaust the waiting order quantity (remaining amount). Ultimately, from a microscopic point of view, the flow of stock prices will move depending on how much the market price order can absorb the remaining amount. Therefore, in order to observe the flow of stock prices, it is necessary to measure the intensity of market price orders that flow in each time.

The contract information includes information such as the contract price, quantity and cumulative quantity of the contract when the contract according to the market order is generated. By analyzing market data, we can read the overall stock flow, even if it is not an accurate forecast for the stock market.

Section ⅲ. Probability of Informed Trading Model (PIN)

The PIN model is a theory published in David Easley et al., "Liquidity, Information, and Infrequently Traded Stocks" (1996). Since it is assumed the informed trader buy or sell at the market price, the information that they have is exposed in the transaction itself. The PIN model quantifies informed traders by estimating the inflow of information with the number of buy or sell generated per unit time.

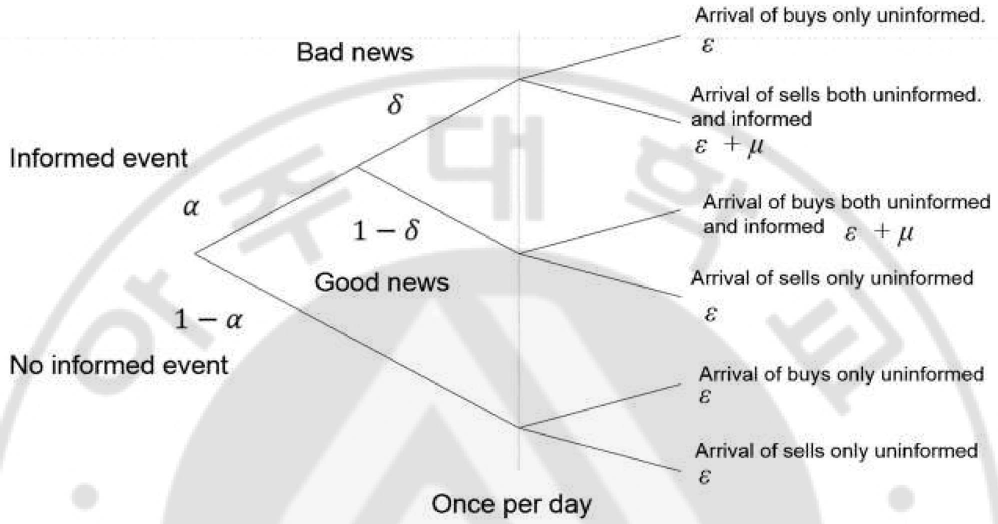

Bad news
δ
Informed event
α
$1-\delta$
Good news
$1-\alpha$
No informed event
Once per day
Arrival of buys only uninformed.
ε
Arrival of sells both uninformed. and informed
$\varepsilon + \mu$
Arrival of buys both uninformed and informed $\varepsilon + \mu$
Arrival of sells only uninformed
ε
Arrival of buys only uninformed
ε
Arrival of sells only uninformed
ε

Figure 3. Illustrated PIN model

- α : The probability of information being present
- δ : Probability that information is bad (1 - δ : Probability that information is good)
- μ : Transaction intensity of information-based traders
- ϵ : Trading Strength of Noise Traders

First, it largely depends on the presence or absence of information. In the absence of information, informed traders do not trade, so market makers also can not trade. Therefore, the transactions observed in the market consist of transactions by noise traders. On the other hand, if the information is present, it can be classified as good or bad. In the case of a good thing, informed traders will buy and sell if it is bad. Noise traders, of course, will continue to trade, regardless of whether they are informed or not.

Figure 3 illustrates the PIN model using the above information. If these assumptions are given, the buying strength and selling strength of the stocks can be obtained, and the Trading Intensity of informed traders can be estimated..

- Buying Strength =

$$\begin{aligned}&\alpha(1-\delta)(\mu+\epsilon_B)+\alpha\delta\epsilon_B+(1-\alpha)\epsilon_B\\&=\alpha(1-\delta)\mu+\epsilon_B\end{aligned} \tag{7}$$

- Selling Strength =

$$\begin{aligned}&\alpha(1-\delta)\epsilon_S+\alpha\delta(\mu+\epsilon_B)+(1-\alpha)\epsilon_S\\&=\alpha\delta\mu+\epsilon_S\end{aligned} \tag{8}$$

- Total Trading Intensity = Buying Strength + Selling Strength =

$$\alpha\mu+\epsilon_B+\epsilon_S \tag{9}$$

- Trading Intensity of informed traders =

$$\alpha\mu \tag{10}$$

- PIN =

$$\frac{(10)}{(9)}=\frac{\alpha\mu}{\alpha\mu+\epsilon_B+\epsilon_S} \tag{11}$$ [4)]

[4)] For detailed proofs, see the author's paper. Easley, D., Kiefer, N. M., O'hara, M., & Paperman, J. B. (1996). Liquidity, information, and infrequently traded stocks. The Journal of Finance, 51(4), 1405-1436.

In order to apply the PIN model, five variables $\alpha, \mu, \delta, \epsilon_B, \epsilon_S$ must be obtained. However, since the only values that can be seen in the market are the buying strength and the selling strength, accurate values should be estimated. In this case, assuming the Poisson distribution of each variables, we construct the likelihood function and then numerically estimate the parameters of the model using Maximum Likelihood Estimation (MLE). The problem is that it takes a very long time to estimate if it goes through this process. Therefore, if you have to trade for a short period of time, such as high frequency trading(HFT), there are many aspects that are not suitable for learning of short period of data in short time, as in this paper.

Section ⅳ. Volume-Synchronized Probability of Informed Trading(VPIN) model

The above-mentioned PIN value indicates the proportion of the informed traders in total transaction intensity. The VPIN model also takes these PIN model definitions similarly. The only difference is that the VPIN model is designed to obtain the PIN value without having to estimate the variables in the PIN model. The core of VPIN is to keep the volume constant, unlike the PIN model, which keeps the time interval constant. This is called Volume Bucket. In VPIN, volume buckets are used instead of time units, so volume can be estimated through the quantity distribution of buy and sell. Suppose that the trading volume of the noise traders has the same weight in the buy and sell.($\epsilon_S = \epsilon_B$) If a single volume bucket has the same number of buy and sell, it is unlikely that informed traders will be involved in the transaction, as bucket is likely to be traded only by noise traders. On the other hand, if the buying and selling ratios are different, it is likely that informed traders are involved in transactions. The inclusion of the volume buckets is considered to be a high probability that information-based transactions (informed traders) are high because the market-to-market transactions, that is, only actual transactions are counted, Thus, the buying and selling ratios within the bucket allow us to estimate $\alpha\mu$ and δ in the PIN model.

- Trading Intensity of Noise traders =

$$\epsilon_S = \epsilon_B \tag{12}$$

- Trading Intensity of informed traders =

$$\alpha\mu \approx E[|S - B|] = \frac{1}{n}\sum_{i=1}^{n}|S_i - B_i| \tag{13}$$

- Total Trading Intensity =

$$\alpha\mu + \epsilon_S + \epsilon_B = E[V^S + V^B] = V \tag{14}$$

- PIN =

$$\frac{\alpha\mu}{\alpha\mu + \epsilon_S + \epsilon_B} = \frac{\alpha\mu}{V} \approx \frac{\sum_{i=1}^{n}|V_i^S - V_i^B|}{nV} = \quad \textbf{VPIN} \tag{15}$$

We can see that the difference between the buying and selling ratios can be regarded as the ratio of the informed traders, so $\alpha\mu$ is the average of the difference between the buying amount and the selling amount in the whole transaction.

In addition, since buckets are the same size of transactions, the trading intensity corresponding to the denominator in PIN can be viewed as V. The condition for the above equations is that $E[|S-B|]$ should approximate the intensity of the informed trader. Therefore, the more tradable securities, ie, the more buckets, the more likely they are to meet the conditions. Therefore, VPIN is possible to quickly obtain a numerical value that conforms to the PIN definition, without having to obtain complicated variables that would have to be sketched in the PIN. Easley's paper stated that the VPIN thus obtained can be used as a substitute for variability.

Section ⅴ. Rama Cont model (CST model)

Rama Cont, et al. published a paper, "Order Book Dynamics in Liquid Markets: Limit Theorems and Diffusion Approximations" (2011). In this paper, the writers assumed Poisson probabilities for Limit orders, Market orders, and Cancellations, which are spell-influencing spells, and Bid, Ask, The motions of the best Bid

& Ask are explained by the mathematical movement using the heavy traffic limit theory.

- h = Minimum order quantity

- λ = The probability of a limit order for the best bid, ask

- μ = The probability of a cancellation occurring at best bid (ask) or The probability of a market order

- η = The probability that a cancellation order occurs on one side of the top priority order and that the other side's top priority order is placed on the other side

Suppose that λ, μ, η follow the Poisson distribution. Also, suppose that the function of the best bid and the ask queue size follow the Markov process[5], and that the order probability at the best bid(ask) is the same.($\lambda_a = \lambda_b = \lambda$, $\mu_a = \mu_b = \mu$)

The equations for the transition rate for function $f = (X_t^b, Y_t^a)$ are :

$$E[X_{t+\Delta t} - X_t | X_t, Y_t] = E[Y_{t+\Delta t} - Y_t | X_t, Y_t] = h(\lambda - \mu)\Delta t + o\Delta t \tag{16}$$

$$E[(X_{t+\Delta t} - X_t)^2 | X_t, Y_t] = E[(Y_{t+\Delta t} - Y_t)^2 | X_t, Y_t] = h^2(\lambda + \mu + 2\eta)\Delta t + o\Delta t \tag{17}$$

[5] Rama Cont, Adrian de Larrard, "Price dynamics in a markovian limit order market", 2010

$$E[(X_{t+\Delta t}-X_t)(Y_{t+\Delta t}-Y_t)|X_t, Y_t]=h^2(2\eta)\Delta t+o\Delta t \tag{18}$$

$$\rho=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}=\frac{-2\eta}{\lambda+\mu+2\eta} \tag{19}$$

Assuming that $\lambda=\mu$, that is, the probability of a market order and a limit order are the same, the equations become simpler.

$$E(X)=E(Y)=0 \tag{20}$$

$$\sigma_X^2=\sigma_Y^2=2h^2(\lambda+\eta) \tag{21}$$

$$\rho=-\frac{\eta}{\lambda+\eta} \tag{22}$$

The simplest form of continuous time diffusion model is Brownian motion with correlation. Therefore, in order to express the order of the best bid(ask) as a simplified model (Reduced form model), the amount of the best bid and ask order are assumed to be Brownian motion.[6]

Let the median values of X_t and Y_t be <X> and <Y>. In order to see the movement of the best bid from a macro perspective, we define a new variable,

$$x=\frac{X}{<X>}, y=\frac{Y}{<Y>} \tag{23}$$

[6] Avellaneda et al,(2010), Forecasting Prices in the Presence of Hidden Liquidity

(x_t, y_t) follows the Bownian motion.

$$dx_t = \sigma d W_t^{(1)} \tag{24}$$

$$dy_t = \sigma d W_t^{(2)}, \quad \sigma^2 = \frac{2h^2(\lambda + \mu)}{< X >^2}, \tag{25}$$

$$E(d W^{(1)} d W^{(2)}) = \rho dt, \quad \rho = \frac{-\eta}{\lambda + \mu} \tag{26}$$

Let (x_t, y_t) be a probability function that increases the price of a change. According to the diffusion theory, $u(x,y)$ must satisfy the following differential equation.

$$\sigma^2(u_{xx} + 2\rho u_{xy} + u_{yy}) = 0, \qquad x > 0, y > 0 \text{[7]} \tag{27}$$

Let's look at the boundary conditions. The price must be raised when the best ask quantity becomes zero, that is, when the incoming market buying order is no longer the best ask quantity, so must be contracted at the next ask price order. That is, it will be $\forall x > 0, \ \ u(x,0) = 1$.

x, y are greater than 0, since it means the quantity of the order book. On the contrary, the price does not rise when the best bid quantity becomes zero, so that a new incoming market selling order must be contracted at the buying order of the next price. In other words, $\forall y > 0, \ \ u(0,y) = 1$

[7] Rama Cont "Order book dynamics in liquid markets : limit theorems and diffusion approximation"

The solution of the differential equation with boundary conditions is as follows.

$$u(x,y) = \frac{1}{2}\left(\frac{1 - Arctan(\sqrt{\frac{1+p}{1-p}}\frac{y-x}{y+x})}{Arctan(\sqrt{\frac{1+p}{1-p}})} \right)\text{[8]} \tag{28}$$

$\rho = 0$, That is, if the movements of the best bid and ask are independent of each other, the equation becomes simpler.

$$u(x,y) = \frac{2}{\pi} Arctan(\frac{y}{x}) \tag{29}$$

[8] Rama Cont "Order book dynamics in liquid markets : limit theorems and diffusion approximation"

Chapter Ⅳ. Data and Model Design

Chapter 4 explains the data preprocessing of the LSTM model used for stock price forecasting and explains the predictive model constructed.

Section ⅰ. Data Collection

The data used for stock price forecast are tick data of KOSPI 200 futures, and the period is from 6 February , 2017 to 31 July , 2017. In the case of the future, I collected data for three month future, and there were two maturity days (9 March and 8 June) during the collection period. In this regard, the six - month future is replaced with three - month future after the expiration of it in the market, so the D+3 six - month future can be defined as three - month future for the following three months after the end of the three - month future. This assumption can be valid because the two data are actually different data and should be defined economically, but the collection data is defined as a three-month record of the period. In fact, the data for the two futures were considered to be continuous because the closing price of the three - month future was equal to the price of the D + 3 six - month future.

In addition, due to the nature of futures, the trading volume increases as the maturity day approaches.

In this paper, the change in trading volume(the trading volume increases as the maturity expires) can be regarded as the trading volume on the day when the trading volume is high irrespective of the maturity date, because the objective is to predict the price change by using only micro data(Limit Order Book).

The data collected and calculated during the collection period is a total of 30 data.

- Initial Price : The initial price during the completion of a trading volume bucket

- High Price : The highest price during the completion of a trading volume bucket

- Low Price: The lowest price during the completion of a trading volume bucket.

- Closing Price: The final price during the completion of a trading volume bucket

- Buying Volume: The number of buying during the completion of a trading volume bucket

- Selling Volume: The number of selling during the completion of a trading volume bucket

- Tick: The number of tick changes during the completion of a trading volume bucket

- VPIN Values: Volume Synchronized Probability of Informed Trading for the next trading volume bucket.

- Selling Strength: Percentage of selling volume of total trading volume during the completion of a trading volume bucket

- CST probability: Rama Cont paper's Probability of rising prices at the next trading volume bucket

- 1~5 bid and ask quote : The quantity specified in the quotes for up to 5 during the completion of a trading volume bucket

- 1~5 bid and ask quote net change : The variation for each quotes during the completion of a trading volume bucket

In the case of the net change in the bid and ask, qutoes are changed if market price is changed until the bucket is completed. In order to solve this problem, I set the base prices of the bucket at the created point, and then defined the net change as the difference in the order quantity of the bid price during completion of the bucket ignoring the market price change.

Data was reworked on the basis of the trading volume bucket mentioned above in 50 trading volume. The collection period of the data for a day was set from 9:00 am, which is the beginning time of the future market, to 3:35 pm, 10 minutes before closing time of the future market. The reason for excluding the 10-minute closing time was that simultaneous quotation transactions were made during this period and no actual transactions were made.

The most important point in dealing with trading volume buckets is the issue of how many buckets contains and how many contracts contains. Because large contracts (more than 100 in the case of KOSPI 200 futures) are frequent, if bucket size is small, the buckets are treated as a large number of the same trading buckets in succession. Therefore, it is important to get the proper bucket size. For the KOSPI 200 futures used in the paper, the minimum trading volume during the collection period is 47,579, the maximum trading volume is 514,040 on May 17, and the average trading volume is 151,280. If one bucket is generated every 10 seconds except for the simultaneous call transaction at the market start, the average volume of bucket is about 50. Therefore this transaction amount is set as a standard of bucket volume.

	Max Volume	Min Volume	Average Volume
KOSPI 200 Future	514,040	47,579	151,280

Table 1. Characteristics of trading volume of KOSPI 200 futures during the collection period (06.02.2017 ~ 04.07.2017)

The price data is changed to the tick unit value by the following equation.

$$Tick\ \mathrm{degree}\ of\ change = \frac{Current\ \mathrm{Price} - The\ last\ bucket\ closing\ price}{0.05}$$

The market price, the high price, the low price, and the closing price are transformed by the above equation, that is, the difference of the value of the previous bucket is divided by the unit of the future value(0.05). In other words, moving prices in the future will be data about how many ticks have moved, and expressed in integers such as +1 and -1. Therefore, the predicted results used in the machine learning are from the three results shown in Table 3.

Forecast Case	Output
Price increase	1
Price not change	0
Price decrease	-1

Table 3. Results by conditions

The reason for this simple prediction is that it is a prediction of a rise in ticks. As I mention through the nature of the empirical data on the back, the main degree of tick change is 0, 1, -1, and the change of 2 ticks or more is few. Therefore, it is reasonable to assume that the actual tick changes are 0, -1, 1 except for the price spread at the end of the simultaneous call and the closing and the next day.

Section ⅱ. Building Learning Data

The prediction rate may vary depending on how many time steps are used in the LSTM. Especially, in the case of stock price, it is important to set the time step to learn the past data because the flow changes instantaneously. In this paper, we constructed various time steps in total of 6 steps. This is called variable '*Window_size*', and the total of 6 steps is 1, 2, 3, 5, 10, 20.

In order to increase the prediction rate using machine learning, it is necessary to learn through classification of data. In Section 4.1, the outputs of the data are limited to -1, 0, 1. The proportions of the outputs are about 0.169 for 1, 0.665 for 0, and 0.164 for -1. When learning through this data, the proportion of 0 is very high, so even if you do not learn, you can get a 66% prediction rate. Of course, if the model matches more than 66%, it can be called a model with a high prediction rate. But before that, it should be a model that can pursue higher returns and a model with high returns is a condition called good model. Therefore, there is a premise that classification should be done well before showing a high prediction rate for the purpose of the paper. Therefore, in order to increase the learning rate of LSTM, the data whose output is 0 is artificially adjusted to the number of -1 and 1 according to the time step. In order to optimize the model performance, we added the net changes in the quantity of buying(selling), ticks, the sell strength to the 30 variables collected in the last chapter. This is because the data at the n-1 bucket point is lost in the process of reworking the output, so it is applied to maximize the data.

Before the time step, the data were generated with a total of 211,845 × 34 sizes. During the time step working, the size of the data increases exponentially. In order to prevent the possibility of data being lost in this case, the data was saved to be used by Hadoop using DB format.

Tick change	Data ratio
Increase	1
No change	1
Decrease	0.98

Table 4. Final data rate, where the number of ratios depends on how the time step is done.

Section ⅲ. The Experimental Design and The Evaluation ways

9)

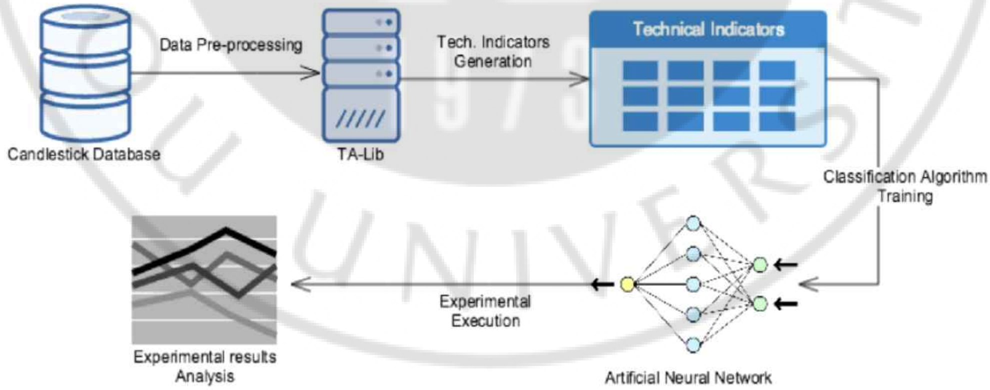

Data Pre-processing
Candlestick Database
TA-Lib
Tech. Indicators Generation
Technical Indicators
Classification Algorithm Training
Artificial Neural Network
Experimental Execution
Experimental results Analysis

Figure 5. A schematized experimental procedure

[9] Figure Reference: David M.Q. Nelson, Adriano C.M. Pereira, Renato A. de Oliveira, Stock Market's Price movement prediction with LSTM

The biggest problem in designing the LSTM model is how to solve the over-fitting problem. When designing too many LSTM layers in designing a model, even with a small number of tests, over-fitting problems occur and fail to find optimized models . Also, the larger the number of output units in the LSTM model, the more the over-fitting, the longer the experiment time, and the lower the predictability. To overcome this over-fitting problem, I added a dropout, which is added to each LSTM layer to suppress the over-fitting problem as much as possible.

As a result of the experiment, the activation functions are Softsign and Softmax, which were showed good performance. Generally, LSTM uses tanh a lot. In this experiment data condition, not only the over-fitting is shown but the performance is lower than the above two functions in the classification condition. In the case of the experiment, the number of experiments was 200 because it showed convergence as it got closer to 200, and more times when it did not converge even at 200.

The ultimate goal of this experiment is classification by variables. Therefore, we have adopted loss functions and conditions that have shown excellent performance in classification. The model that was finally designed is as follows.

Experimental conditions	Model settings
The number of LSTM	2 ~ 5 (depending on time step)
Dropout	Combination of 0.8, 0.5 and 0.2
The number of Dropout	Add to all LSTMs (2 to 5)
Activation function	LSTM = softsign , Classification = softmax
Optimizers	lr = 0.0001, beta = 0.9, beta2 = 0.999, ϵ=1e-0.8, decay = 0.000001
Loss function	Categorial crossentropy
Epochs	200, 400

Table 5. Summary of LSTM model design

The data set used in the experiment is divided into a training set and a test set, and the division ratio is set to about 0.7 to 0.3. Depending on the time step, data may be lost in the middle of the process, so some correction work has been done.

Two methods are used to evaluate the performance of the model. First, the prediction rate is evaluated. The prediction rate is as follows.

$$Accuracy = \frac{The\ number\ of\ predicted\ output}{The\ number\ of\ total\ data}$$

However, as mentioned above, since the number of data has the largest number of 0, there is a case in which all the states are predicted as 0(no change) at the time of testing even if the learning is good.

In other words, returning this prediction rate as much as 0 (66%) proportion may be the best learning rate model. However, this model can not be a model with good returns, so a model that better predicts -1(fall) and 1(rise) should be a better model even if the forecast for 0 falls. Naturally, it better predicts the -1 and 1 conditions, so it is a model with higher returns.

$$Accuracy = \frac{The\ number\ of\ predicted\ output\ 1,-1}{The\ number\ of\ total\ data}$$

Chapter Ⅴ. Experiments and Evaluation

Chapter 5 analyzes the results of applying the proposed model through KOSPI 200 futures data generated in Section 4.

Section ⅰ. Analysis of data properties

We will look at the nature of the data(KOSPI 200 futures) during the collection period. First, KOSPI 200 futures are futures trading for Korea's KOSPI 200 stock index. Since the underlying asset is the KOSPI 200 index, the trading unit is the KOSPI 200 index × 125,000. The quotation unit of the index is 0.05, and the value of one tick is 12,500 won. There are three exchanges where KOSPI futures can be traded, including the Korea Exchange, Exture and the Chicago Mercantile Exchange. Based on the Korea Exchange, regular trading hour is from 9:00 am to 3:35 pm, and simultaneous call deals are made from 3:35 to 3:45. From 6:00 pm to 5:00 am the following day, night market trading is possible. On the expiration date of the future, the regular trading time ends at 3:20 pm. The expiration date is the second Thursday of the month in March, June, September, and December and there were two expiration dates (9 March and 8 June) while data were collected.

When trading in the market, you can use limit price order, market price order, conditional limit price order, and the best advantageous price order.

In this paper, I try to predict the price change of KOSPI 200 futures for three months. Since futures, not 3-months future, are relatively lower volume than 3-months future in the market, so the futures data are selected based on 3-months future, and the future at the point of time when they are traded in the market for 3-months is regarded as 3-months Therefore, there are three pieces of futures data during this data collection period, and data are 101M3000, 101M6000, and 101M9000.

Figures 5 and 6 are graphs of the price and rate of change of the futures during the collection period. It shows a steady upward trend, and there are many days that showed the positive rate of change.

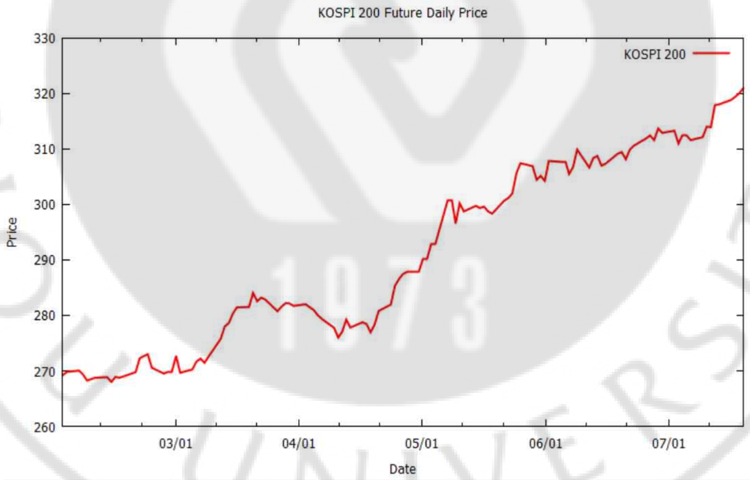


Figure 6. KOSPI 200 futures daily stock price change

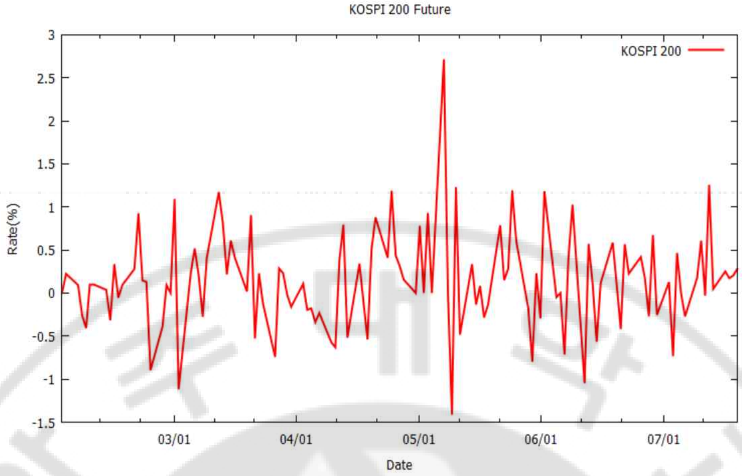


Figure 7. KOSPI 200 futures daily yield

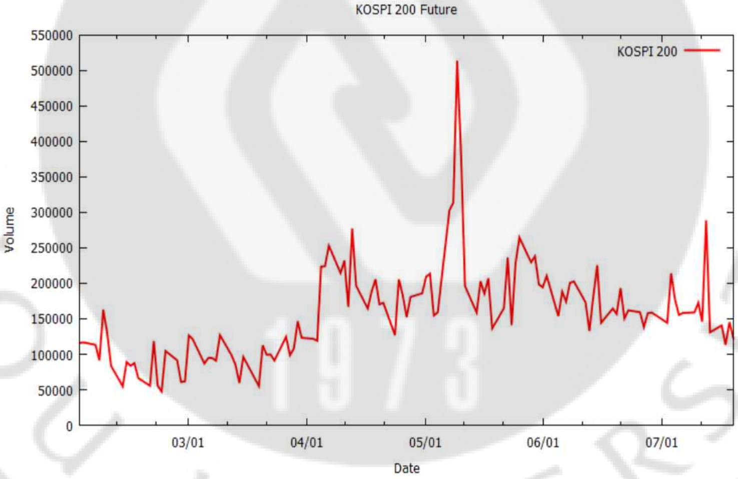


Figure 8. KOSPI 200 daily trading volume change

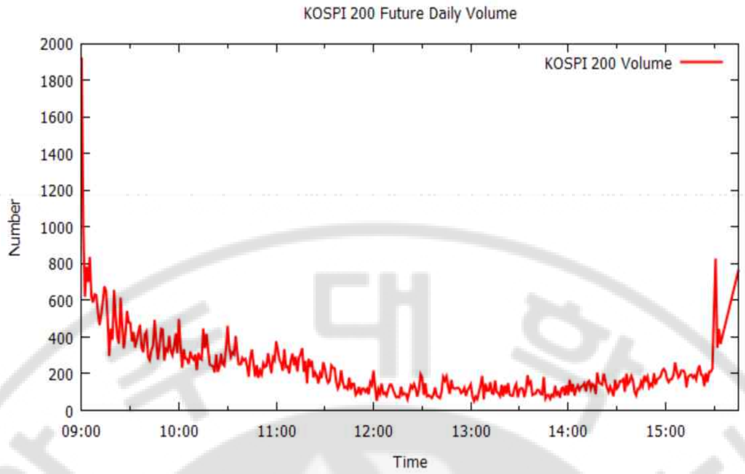


Figure 9. Progress of average daily trading volume of KOSPI 200 futures

Let's look at the average change of the day. Here, a day means average data for each time point (unit of seconds) during a collection period, not a specific period.

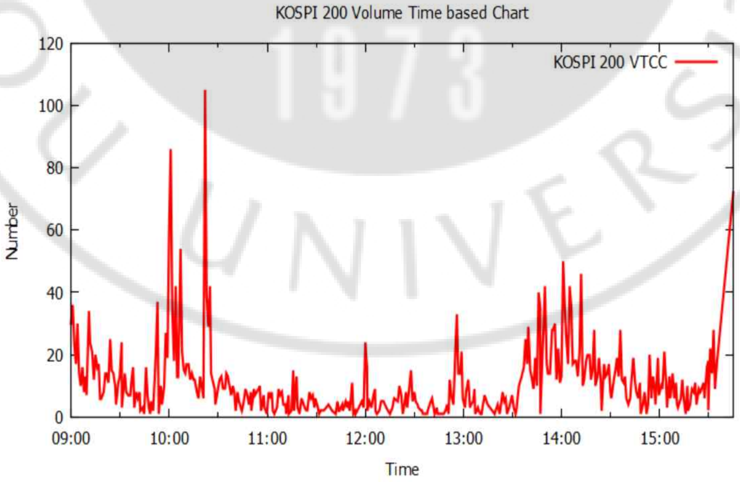


Figure 10. Change in bucket number generated during KOSPI 200 futures trading

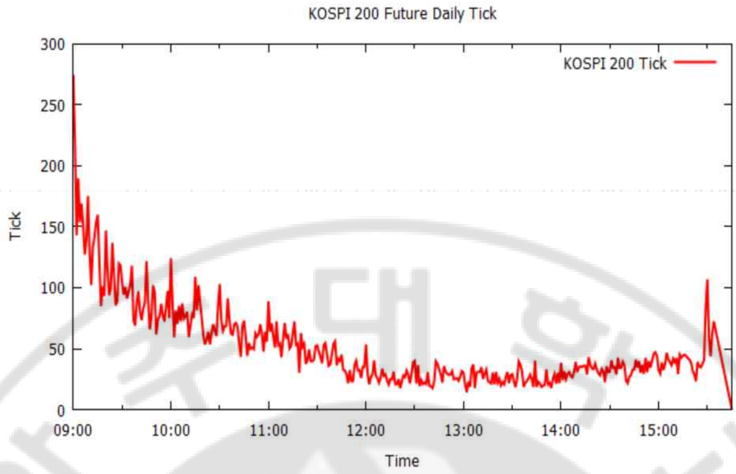


Figure 11. Tick changes during KOSPI 200 futures trading

Figure 9 shows the change in trading volume in 30-second increments. A lot of transactions are done with the simultaneous call price at 9:00 am, and we confirm that about half of daily trading volume is in the morning. It can also be seen that the trading volume increases sharply even after 3 o'clock when the market is closed. In general, the volatility of the price increases naturally when the trading volume is high. Therefore, if the machine for price prediction is applied to investment, it can be expected that the prediction possibility will be actively performed in the morning market.

Let's look at the data processed based on trading volume. In order to apply the VPIN value mentioned above, the data is processed in the volume bucket unit rather than the time unit basis. In this case, the standard transaction volume is 50. The amount of bucket produced is shown in the graph of Figure 10 in one-minute increments. In Figure 8, we can see that the trading volume is largely focused on the beginning and closing of the market and the buckets are also large at the beginning and end of the chapter. Finally, Figure 11 is a graph of the change in ticks of the data reprocessed in units of trading volume. It took about 6 to 7 seconds for one bucket to be generated based on the volume of transactions during the collection period. The change in data during this time was often less than one tick. In other words, there were more cases where the stock price did not change than the stock price changed over a short period of time.

Tick Change	Amount
Excess 1	1,096
1	34,578
0	140,903
-1	34,241
Under -1	1,027

Table 6. Learning Data

Tick Change	Amount
Rise	35,268
No Change	140,903
Fall	35,674

Table 7. Output Used for Learning

Tick Change	Amount
Excess 1	106
1	5,812
0	22,899
-1	5,630
Under -1	114

Table 8. Imitation Investment Data

Tick Change	Amount
Rise	5,918
No Change	22,899
Fall	5,744

Table 9. Output Used for Imitation

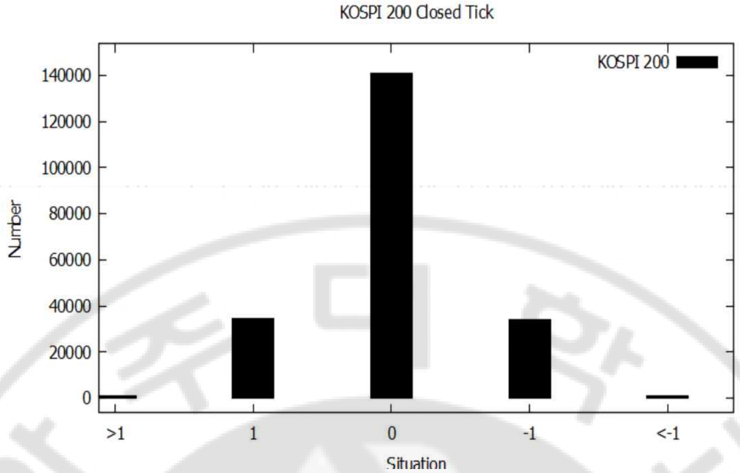


Figure 12. Graph of KOSPI200 Future tick change

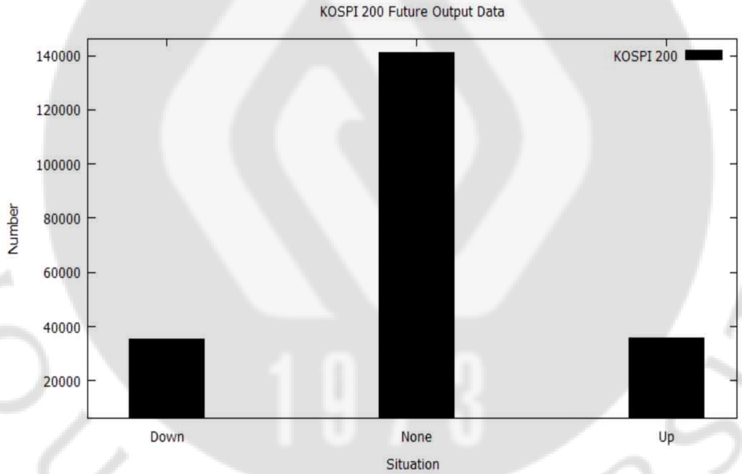


Figure 13. KOSPI 200 Future Learning Output

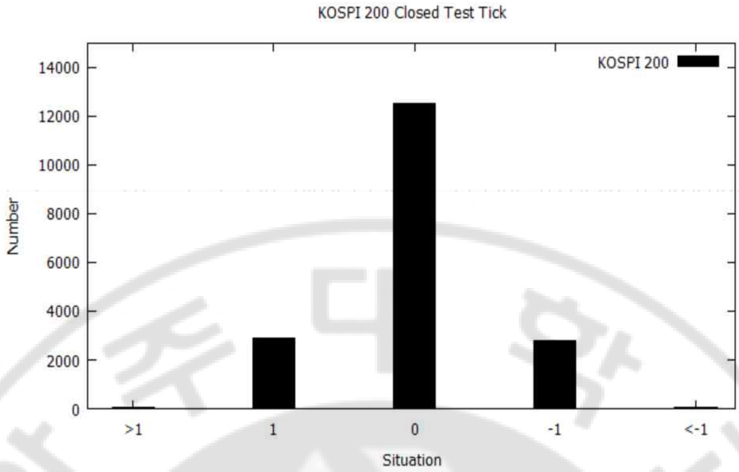


Figure 14. KOSPI200 Future Imitation Investment Data

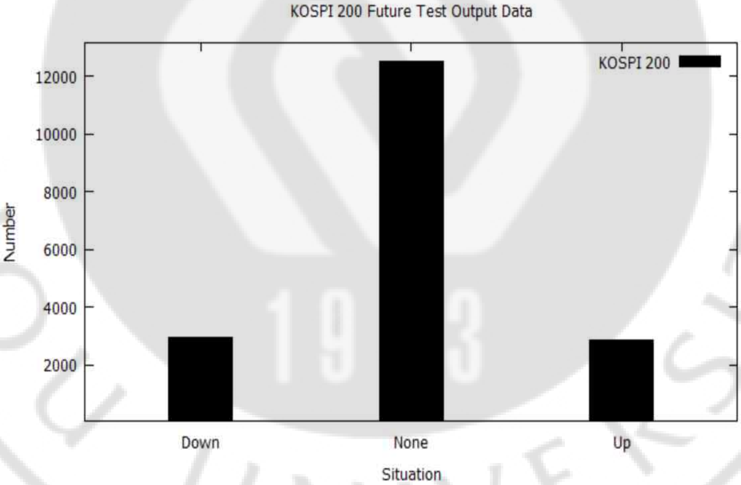


Figure 15. KOSPI 200 Future Imitation Investment Output

Section ⅱ. Experimental Results and Analysis

I compared the performance of the LSTM with the baseline model and the various time steps to see how much the LSTM model is performed and how much it could predict for stock changes. Unfortunately, there is not yet a model for estimating probability using similar data. Therefore, I use the K-mean algorithm model, which is widely used among the classification models, for how to classify the next situations, which is the definition of model accuracy.

Table 10 summarizes the probability of each situation for this situation. The LSTM model based deep-learning shows better prediction than the K-means algorithm. The K-means algorithm showed a huge hit ratio for the 'no change' situation, which can be interpreted as the over-fitting of the model. This is because the LSTM has a higher prediction rate than the K-average model for the ‘rise and fall’ of the forecasting situation. On the other hand, in the case of LSTM-based models, the probability of correctly recognizing "no change" may be somewhat lower than the K-means model, but the prediction of the ‘rise and fall’ was even better.

Also, The LSTM model shows that the prediction rate varies depending on how the time step is set. As the time step increases, the prediction rate increases, and the best prediction rate is obtained when the time step is 10. Not only the overall forecast, but also the forecasts for 'rise' and 'fall' outperformed other time steps. Therefore, when we look at how the actual return on investment is, we will use the 10 time step model.

	Probability by Situation(%), n = Time Step					
	K-Mean	LSTM (n=2)	LSTM (n=3)	LSTM (n=5)	LSTM (n=10)	LSTM (n=20)
Rise	40.2	48.5	49.2	45.7	51.9	50.1
No Chagne	93.3	83.2	89.6	90.1	91.1	91.7
Fall	41.6	50.1	47.7	49.6	51.2	51.4
Total prob	51.6	56.1	59.3	60.3	61.5	60.7

Table 10. Probability of rise, fall and no change by each time steps

Figures 16 to 19 are graphs of changes in accuracy with respect to the number of learning times per time step.

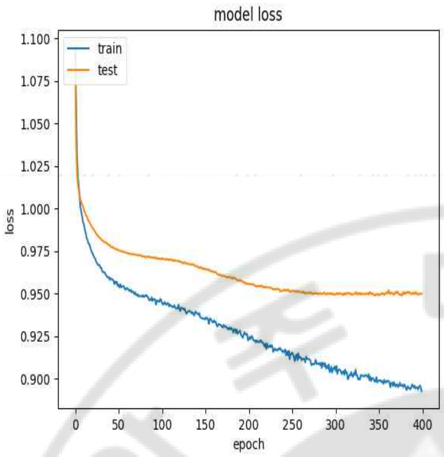


Figure 16. Graph of error variation according to learning times (time step:2)

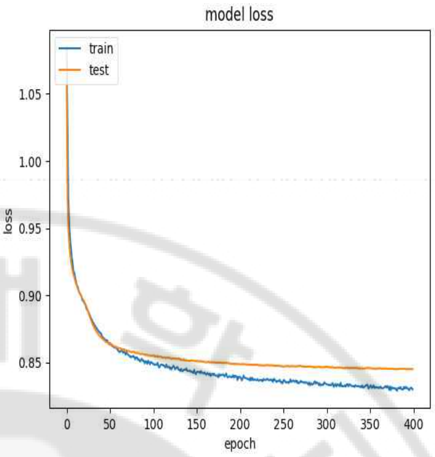


Figure 17. Graph of error variation according to learning times (time step:3)

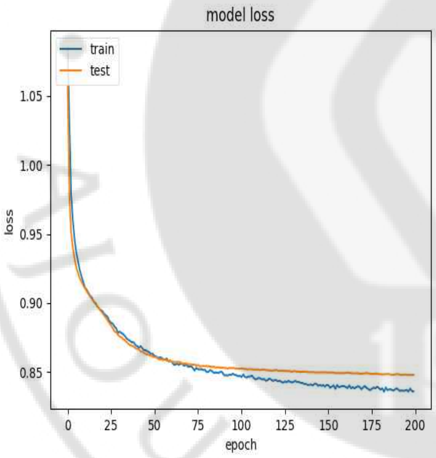


Figure 18. Graph of error variation according to learning times (time step:5)

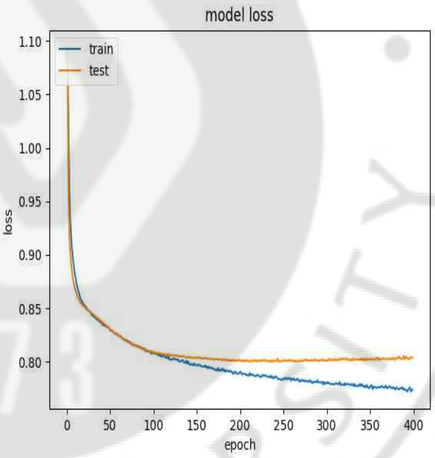


Figure 19. Graph of error variation according to learning times (time step:10)

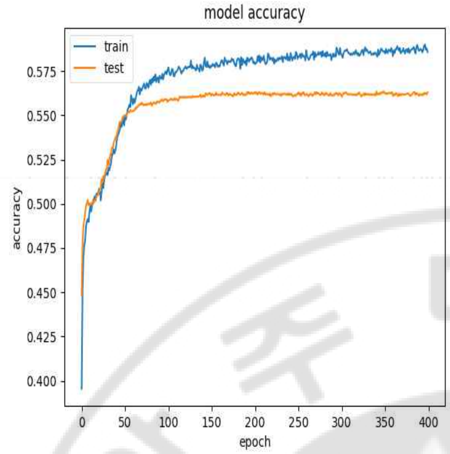


Figure 20.Graph of accuracy change according to learning times (time step:2)

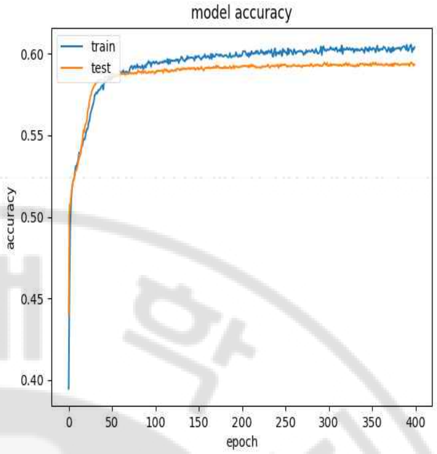


Figure 21. Graph of accuracy change according to learning times (time step:3)

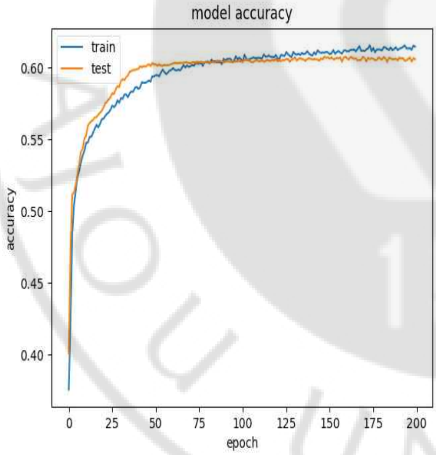


Figure. 22 Graph of accuracy change according to learning times (time step:5)

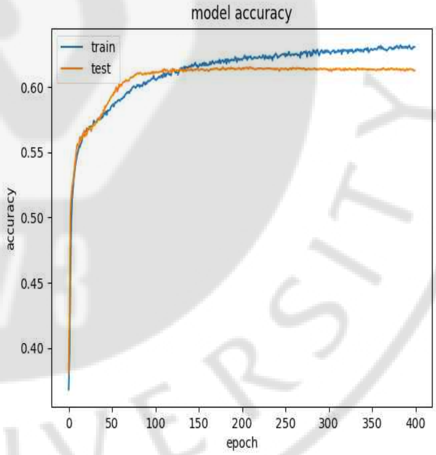


Figure 23. Graph of accuracy change according to learning times (time step:10)

Next, we will examine how the parameters of the mathematical model, CST and VPIN, influence the prediction rat. Table 11 summarizes the prediction rates of learning models excluding VPIN, excluding CST, and excluding both variables using the time step 10 model.

	Probability by Situation(%)			
	Regular	Except VPIN	Except CST	Except Both
Rise	51.9	51.4	51.7	51.2
No Chagne	91.1	90.8	91	90.7
Fall	51.2	51	50.9	51.1
Total prob	61.5	61.2	61.4	61.1

Table 11. Probability of rise, fall and no change by each situations

As a result, VPIN, and CST values did not have a significant influence on the prediction rate. However, we can see that the VPIN value has a little more influence than the CST probability.

First, I consider two reasons why CST probability is less influential than expected. First, the CST probability itself is the data created by using only the bid and ask, so it is highly likely that these two values alone will not predict the next probability. In fact, the prediction rate is very low when the following situation is predicted only by the CST probability.[10]

Second, a large time interval per bucket can also have an impact. Second, a large time interval per bucket can also have an impact. In Rama Cont 's paper, they apply the CST probability to

[10] It simply means that the CST probability compares well with the next situation.

very small time, that is, stocks with high volume trading volume. In the case of KOSPI200 used in this paper, it takes more time for buckets to be generated because trading volume is smaller than those used for demonstration in Rama Cont's paper. Therefore, there is a high possibility that information during that time is distorted.

On the other hand, VPIN is more influential than CST probability because of the basic nature of VPIN. In the paper[11] that introduces VPIN, VPIN can be used as a substitute for volatility. In other words, the cumulative VPIN is highly correlated with the stock price trend because it shows the stock price change during the cumulative period.

Finally, we look at how much profit can be made using LSTM (n = 10) model which showed the most accurate prediction rate. In this case, we assumed the simulated investment situation when it can be possible to ignore fees,

We set equally the trading volume of 50 units to minimize the effect of transactions by machine algorithm. The important thing about machine algorithm investment is that if you anticipate the following situation in advance, then you will carry out the investment behavior according to the situation. Likewise, you should not do the opposite of the situation. In next rising situation, if you sell stocks as a declining signal, you will see the damage to the rise. On the other hand, if you take the downside signal as an upside, you will lose money by making purchases.

A brief description of the algorithm follows.

[11] Easley, D., López de Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. The Review of Financial Studies, 25(5), 1457-1493.

- Rising forecast: Maintain existing assets and buy assets.
- No change Predict: Hold investment.
- Prediction of decline: Dispose all existing assets.

Actual Data (Amount)	Fall	3,148	1,270	1,494
	No change	1,170	20,723	1,006
	Rise	1,562	1,275	2,907
		Fall	No change	Rise
		Prediction Data (Amount)		

Table 12. The nature of the simulated investment data

Table 12 shows the results of the LSTM machine for simulated investment data. The forecast for the drop was 53.2%, and the forecast for the rise was 50.7%. The rate of return was -0.17% on the LSTM machine. but -1.61% if the asset was held without any investment,

ROI Comparison (19.July.2017 ~ 31.July.2017)	Return (%)
Long	- 1.61
LSTM Machine	- 0.17

Table 13. Comparison of returns during the simulation period

Chapter Ⅵ. Conclusion and Future Analysis

In this paper, we researched a new investment forecasting machine with a mathematical model through transaction data used in high frequency trading using LSTM machine which was used by existing papers on stock market prediction.

In order to carry out the study, we first collected the raw data of the KOSPI 200 futures data from 6 February, 2017 to 31 July, 2017, and combine market data with numerical value VPIN and CST derived mathematical models. In addition, the data used for the prediction is processed again as the trading volume data, not the time data. To make the LSTM model with high prediction rate, we divide the data into various time steps, drop-out and initial model weights were adjusted to reduce over-fitting problem in the model.

As a result, the best model for market prediction was LSTM model of time step 10. In the case of a 10 time step, although the prediction of 'no change' compared with the K-means model was a slight decrease, it was the best model in that the 'rise' and 'fall' were somewhat better than the other models.

LSTM 10 model was used to simulate the return on investment. As a result of the experiment, it showed a slightly higher return than holding the futures without doing anything.

But there were a lot of things that were disappointing. First, prediction rate was lower than expected. I expected that as the number of time steps increases, the prediction rate of rise and fall will increase as more time steps are taken, but 20 time step showed a lower probability of rise and fall than 10. Also, as the time step increases, the over-fitting phenomenon occurs more rapidly. These problems created a decline in yields during trading. Since the final return have to exclude taxes, the actual rate of return will be lower than this.

I also concluded that the rate of return could be higher depending on how the nature of the data is viewed. One bucket contains four situations during the completion of the bucket, the beginning, the highest, the lowest, and the closing price. If the highest and lowest prices are different, for example, let's say the price has risen while the bucket was being built, but it has fallen again. In this case, if the LSTM model predicted 1, the profit would have occurred. Likewise, It is possible that we missed this opportunity simply because we saw the final price as the result.

Looking the data one by one, there were a lot of gaps between the closing price and the highest price. Of course, the opposite situation when the middle price is 0, but closing price is 1 also happens. However, considering a profitable model, I think it would be more advantageous to include these situations. Therefore, we can expect a model with better prediction rate learning through data considering all of these situations.

It is expected that the model will have a higher prediction rate than the existing model when learning through longer period data. Of course, I think we can apply the model to various other stocks as well as KOSPI 200 futures data. In particular, mathematical models (CST model, VPIN model) are models with good predictions when using stocks with higher trading volume(The stocks used for verifying the papers were Citibank, GE which have averages daily trading volume of over 10,000,000). So, for more data volume, more buckets will of course be created, and in this case, more data and more accurate mathematical model values will be created and a model with a higher prediction rate.

Also, the prediction output of this research model were only used as a simple signal in trading. However, studies on stock market trading using reinforcement learning have been actively conducted in recent years. Therefore, the model created in this study can be used as an additional indicator in the model using reinforcement learning. In the future, it would be a good research topic to develop a reinforcement learning model combined with this research model.

Reference

Korean Reference

[1]Sohyun Gil, Rooda Lee, Bowon Yang and Kwang-soon Choi(2016), *A Study on the Method of Energy Demand Prediction Using Deep Learning*, KETEP, No.20132010101850, 1014-1015

[2]Sungtak Kim(1998), *A evaluation of strategic trader models in the securities market microstructure theory,* 271-281

[3]Il seon Seo, Sangsu Yeo and Heaujo Kang(2014), *A Study on the Suggestion of Domestic Stock Market Analysis Scheme using Big Data*, Proceedings of KIIT Summer Conference, 550-554.

[4]Yongjae Lee, Woo Chang Kim(2013), *A stochastic Model for Order Book Dynamics: An Application to Korean Stock Index Futures,* Management Science and Financial Engineering, 19(1), 37-41.

[5]Jaewon Lee(2013), A Stock Trading System based on Supervised Learning of Highly Volatile Stock Price Patterns, Journal of KIISE : Computing Practices and Letters 19(1), 167-172

[6]Jungkwon Lee(2017), Single image super-resolution using spatial LSTM(masters dissertation), Seoul National University, Seoul

[7]Jihoon Lee(2016), Stock Price Prediction Model using Deep Learning(masters dissertation), Soongsil University, Seoul

[8]Sanghyun Choo and Hyunsoo Lee(2016), Stock Price Prediction and Investment Strategy based on Deep Learning, Proceedings of

KIIS Spring Conference 26(1), 11-12

Foreign References

[9]Chen, K., Zhou, Y., & Dai, F. (2015). A LSTM-based method for stock returns prediction: A case study of China stock market. In Big Data (Big Data), 2015 IEEE International Conference, 2823-2824

[10]Cont, R., & De Larrard, A. (2012). Order book dynamics in liquid markets: limit theorems and diffusion approximations.

[11]De Jong, F., & Rindi, B. (2009). The microstructure of financial markets. Cambridge University Press.

[12]Fischer, T., & Krauß, C. (2017). Deep learning with long short-term memory networks for financial market predictions (No. 11/2017). FAU Discussion Papers in Economics

[13]Gunduz, H., Cataltepe, Z., & Yaslan, Y. (2017, May). Stock market direction prediction using deep neural networks. In Signal Processing and Communications Applications Conference (SIU), 2017 25th, 1-4

[14]Jiang, J. (2015). Volume-Synchronized Probability of Informed Trading (VPIN), Market Volatility, and High-Frequency Liquidity.

[15]Marco Avellaneda, Josh Reed, Sasha Stoikov(2011), Forecasting prices from level-1 quotes in the presence of hidden liquidity, Algorithmic Finance, vol. 1, no. 1, 35-43

[16]Nelson, D. M., Pereira, A. C., & de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. In Neural Networks (IJCNN), 2017 International Joint Conference,

1419-1426

[17]O'hara, M. (1995). Market microstructure theory (Vol. 108). Cambridge, MA: Blackwell.

[18]Raza, K. (2017). Prediction of Stock Market performance by using machine learning techniques. In Innovations in Electrical Engineering and Computational Technologies (ICIEECT), 2017 International Conference, 1-1

[19]Sagitov, S. (2013). Weak Convergence of Probability Measures. Chalmers University of technology and Gothenburg University.

[20]Silva, E., Castilho, D., Pereira, A., & Brandao, H. (2014). A neural network based approach to support the market making strategies in high-frequency trading. In Neural Networks (IJCNN), 2014 International Joint Conference, 845-852

[21]Türkmen, A. C., & Cemgil, A. T. (2015). An application of deep learning for trade signal prediction in financial markets. In Signal Processing and Communications Applications Conference (SIU), 2521-2524

초 록

2016년 이후 분야를 막론하고 가장 화두로 떠오른 학문은 기계학습이다. 알파고의 등장과 함께 기계학습을 응용하려는 시도는 산업 전 분야에 걸쳐서 일어났으며, 금융권에서도 마찬가지로 주목받기 시작하였다. 금융권에서 기계학습은 투자 은행, 트레이딩, 신용평가와 같은 다양한 금융 영역에서 연구되기 시작했으며, 특히 향후 주가의 흐름에 대한 예측하는데 있어서도 기존 방법들과 다른 방법으로써 활발히 연구가 이뤄졌다. 기존의 주가 예측은 수리적인 모델링을 통하거나 복잡한 통계적인 모델링을 통하여 연구하였는데, 이러한 경우 일반화된 모델을 설정하여 예측률이 떨어지거나, 응용하기가 어려워 실제 상황에서 투자하기 어려운 경우가 많았다.

본 논문은 기계 학습 기술 중 LSTM 방법을 이용하여 코스피200 선물의 가격 변화에 대한 예측을 목표로 한다. 총 33개의 변수가 이용되며, 사용되는 변수들 중에서 예측을 목표로 만들어지고, 수리, 통계적으로 우수한 모델링이었던 VPIN모델과 Rama Cont 모델을 LSTM의 변수를 포함한다. 이에 앞서, 데이터는 HFT(High Frequency Trading)분야에서 많이 사용하는 거래량단위 데이터를 만들어서, 단기간 하에서의 가격 변화를 맞추려는 시장미시구조 모델을 설계하였다. 본 논문은 다양한 타임 스텝과 변수들을 통하여 최적화된 예측률을 보이는 LSTM 모델을 설계하는 것을 목표로 한다.

주제어 : VPIN, CST model, HFT, 기계학습, LSTM