International Journal of Advanced Smart Convergence Vol.11 No.1 19-27 (2022)
http://dx.doi.org/10.7236/IJASC.2022.11.1.19

IJASC 22-1-3

Prediction Model of Real Estate ROI with the LSTM Model based on AI and Bigdata

Jeong-hyun Lee*, Hoo-bin Kim**, Gyo-eon Shim ***

**PhD Candidate, Department of Real Estate Studies, Konkuk University, Korea*
***Master Candidate, Department of Artificial Intelligence, Yonsei University, Korea*
****Professor, Department of Real Estate Studies, Konkuk University, Korea*
E-mail leejohn.news@gmail.com, source234@naver.com, x1000@konkuk.ac.kr

Abstract

Across the world, 'housing' comprises a significant portion of wealth and assets. For this reason, fluctuations in real estate prices are highly sensitive issues to individual households. In Korea, housing prices have steadily increased over the years, and thus many Koreans view the real estate market as an effective channel for their investments. However, if one purchases a real estate property for the purpose of investing, then there are several risks involved when prices begin to fluctuate. The purpose of this study is to design a real estate price 'return rate' prediction model to help mitigate the risks involved with real estate investments and promote reasonable real estate purchases. Various approaches are explored to develop a model capable of predicting real estate prices based on an understanding of the immovability of the real estate market. This study employs the LSTM method, which is based on artificial intelligence and deep learning, to predict real estate prices and validate the model. LSTM networks are based on recurrent neural networks (RNN) but add cell states (which act as a type of conveyer belt) to the hidden states. LSTM networks are able to obtain cell states and hidden states in a recursive manner. Data on the actual trading prices of apartments in autonomous districts between January 2006 and December 2019 are collected from the Actual Trading Price Disclosure System of the Ministry of Land, Infrastructure and Transport (MOLIT). Additionally, basic data on apartments and commercial buildings are collected from the Public Data Portal and Seoul Metropolitan Government's data portal. The collected actual trading price data are scaled to monthly average trading amounts, and each data entry is pre-processed according to address to produce 168 data entries. An LSTM model for return rate prediction is prepared based on a time series dataset where the training period is set as April 2015~August 2017 (29 months), the validation period is set as September 2017~September 2018 (13 months), and the test period is set as December 2018~December 2019 (13 months). The results of the return rate prediction study are as follows. First, the model achieved a prediction similarity level of almost 76%. After collecting time series data and preparing the final prediction model, it was confirmed that 76% of models could be achieved. All in all, the results demonstrate the reliability of the LSTM-based model for return rate prediction.

Keywords: *Real Estate, AI, Bigdata, Prediction, LSTM, Machine learning, Deep learning, Time series forecasting*

1. Introduction

In Korean society, 'real estate' tends to account for a majority of a household's wealth and thus have long

Manuscript Received: January. 10, 2022 / Revised: January. 17, 2022 / Accepted: January. 20, 2022
Corresponding Author: x1000@konkuk.ac.kr
Tel: +82-02-450-3364
Professor, Department of Real Estate Studies, Konkuk University, Korea

been considered as an important form of investment. When investing into real estate, it is necessary to measure ‘return rates.’ For several years, there have been efforts to use AI to increase return rates in the stock market. On the other hand, this trend is only just now being explored in the real estate market with research on measuring return rates and predicting the future with AI.

For many years, real estate has been perceived as a ‘stable asset.’ However, from a macro perspective, there have been momentary pauses in the rise of real estate prices, and it has been proven that prices are affected by a wide range of variables. Hence, many Koreans are sensitive to changes in real estate prices, given that it could potentially place a significant portion of their assets at risk. Such uncertainty could lead to excessive speculation in the real estate market or make investors reluctant to buy or sell real estate property for fear of a market crash in the future.

This study focuses on the immovability of real estate. Given that real estate properties are built on land, price fluctuations according to location are inevitable. Various efforts have been made to predict such price fluctuations using a variety of approaches, including those involving AI and deep learning. This research aims to predict ‘return rates’ of real estate prices by utilizing deep learning technology.

2. Research Method

This study uses the long short-term memory (LSTM) technique. As there are several existing LSTM models for return rate predictions in the stock market, this study aims to explore the potential of LSTM models for return rate predictions in the real estate market with time series data. As previously mentioned, the LSTM approach has been widely applied in the stock market for return rate prediction, which is in contrast to the lack of implementation in the real estate market. A key purpose of this study is to investigate the feasibility of LSTM models for predicting return rates in the real estate market. By analyzing data, this study explores the possibility of predicting return rates for highly sought-after residential areas. RNN is known to gradually decrease the gradient during backpropagation if the distance between the relevant information and the point where the information is used is far, resulting in a significant decrease in learning ability. This is called the varnish gradient problem. The LSTM is designed to overcome this problem.

For this purpose, this study uses data on actual trading prices of apartments in Korea provided by MOLIT. As real estate prices form time series data that follow seasonal trends, it is important to pre-process the data to separate and isolate trends and periodic components.

To extract trends from the time series data, this study uses the HP filter (Hodrick-Prescott filter) method to improve the data. Additionally, the SOM (self-organizing map) method, a cluster analysis method based on artificial neural networks, is used for data mining to specify appropriate target market regions by clustering regions with similar price trajectories. SOM is an unsupervised neural network that arranges high-dimensional data into low-dimensional neurons that are easy to understand and shapes it in the form of a map. In other words, if the input variable of the actual space is close, it is also close to the map. This shaping is characterized by preserving the positional relationship of the input variable as it is, so we referred to this model. The explanatory powers of input variables are presented based on the similar cluster regions. To construct a prediction model for future price trends related to the analysis targets, several approaches were considered, including the SVR (support vector regression) method and deep learning methods. In the end, the LSTM (long short-term memory) method was selected as it is suitable for time series data prediction.

As LSTM models are often used for predictions involving time series data, this study aims to validate the prediction performance of the LSTM method for return rates in the real estate market.

3. Theoretical Considerations

Neural networks have been touted as powerful tools for single time series data prediction, and various techniques have been developed to compensate for internal issues of such networks. In addition to MLP (multilayer perceptron), notable examples include RNN (recurrent neural networks), ESN (echo state networks), GRNN (generalized regression neural networks), and LSTM (long short-term memory). In recent years, RNNs (a deep learning technique) have been widely applied to time series data prediction.

Hidden layers of current artificial neural networks consist of simple neuron configurations that do not consider context. While this structure is capable of time series data prediction by implementing hidden layers from the past, a recurrent neural network (RNN) has a directed cycle in which past events are able to influence future results. RNNs that process sequential information are mainly used for temporally correlated data. This results in a neural network model that considers past data as well as the correlation between previous data (t-1) and current data (t) to predict t+1 data. However, RNNs have the disadvantage of vanishing gradients when there are great distances between the information and the point at which the information is used. The LSTM technique provides a solution to this issue by adding cell states (which act as a type of conveyer belt) to the hidden states of an RNN and then recursively obtaining the cell states and hidden states.

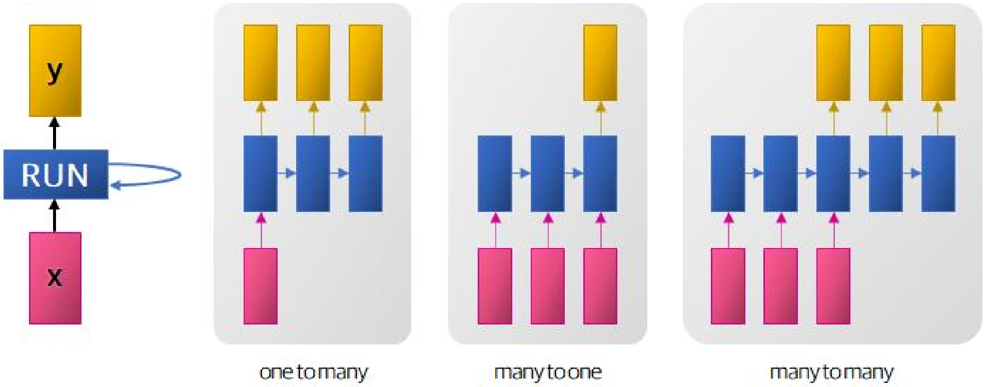


Figure 1. Basic structure

The LSTM network method was first introduced by Hochreiter and Schmidhuber (1997) as a means of improving upon the long-term memory of vanilla RNNs and as an extension of the concept of memory cells and gating mechanisms that regulate information flow in RNNs. LSTM networks resolve the shortcomings of RNNs by storing information and reducing errors in the network over longer periods of time. The basic structure of LSTM centers around the network learning what to store, discard, and read over long periods. This study aims to apply an LSTM model made for time series data to predict return rates in the real estate market.

4. Data Composition

4.1 Data collection

This study utilized data on the actual trading prices of multi-unit housing and domestic apartments from MOLIT and the Public Data Portal. The actual trading prices of apartments in various autonomous districts between January 2006 and December 2019 were gathered from the Actual Trading Price Disclosure System of MOLIT. Additionally, basic apartment and commercial building information were gathered from the Public Data Portal and Seoul Metropolitan Government’s data portal. The collected actual trading price data were scaled to monthly average trading amounts, and each data entry was pre-processed according to address to produce 168 data entries. The final data presented the average monthly price of each autonomous district, and the data were divided into training data and model validation data according a 4:1 ratio (135 data entries for training, 33 data entries for validation). <Table 1> provides an overview of the collected data. Using the data, we attempted to predict the 'return rate' by region.

Table 1. Data collection

Table	DF1	DF2	DF3	DF4
Description	Domestic apartment information	Apartment actual trading prices	Commercial area analysis information	Population information
Columns	Name, address, parcel sale format, no. of households, construction company, etc.	Name, address, contract year/month, dedicated area, trade price	Baseline date, commercial area address, no. of workers (male/female/total)	Autonomous district, time series population information by month
Source	Public Data Portal	MOLIT	Seoul Metropolitan City Data	-

- DF: DataFrame

4.2 Data configuration

The address information in DF2 was set as the baseline for the datasets. The datasets were constructed by deriving explanatory variables based on address and time series. In addition, since future trade prices are the ultimate goal of the model, the trade price in DF2 was set as the target variable. <Table 2> shows the basic configuration of the datasets, and <Table 3> lists the datasets that were produced. The parcel sale period and construction period etc. of DF1 can also be applied as explanatory variables in the datasets. This below regression model is very simple. However, the prediction model of real estate ROI proposed in this study is meaningful because it is the first step as an attempt in the real estate market, so first, I will try to study with a simple prediction model. The datasets such as actual trading price, population data, final data set, commercial area data, and applied elements, were collected from the data portal for public opening and sharing (www.data.go.kr is operated by the Korean government). And all the datasets are composed for actual training.

Table 2. Basic dataset configuration

* Configured in the format of Y(t) = a*X1(t) + b*X2(t) + c
* X1 = dedicated area / X2 = no. of residents / Y = trade price

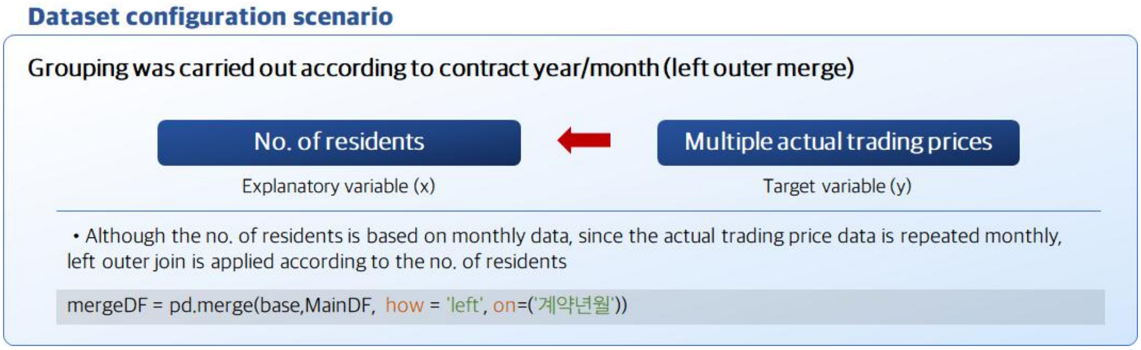
Table 3. Datasets

	Baseline	Combined element	Time series	Explanatory variable (x)	Target variable (y)
Actual trading price	DF2	**Address**	**Contract year/month**	**Dedicated area**	**Trade price**
Population data	DF4	**-**	**Baseline year/month**	**No. of residents**	**-**
Final data set	**-**	**Address**	**Baseline**	**No. of residents,**	**Trade price**

			year/month	**dedicated area**	
Commercial area data	DF3	-	Baseline quarter	Floating population (workplaces)	-
Applied elements	DF1	Address	-	Parcel sale period, etc.	-

4.3 Scenario and dataset configuration

Through various simulations, the data were configured to predict the 'return rate' of the Dongjak-gu district over three months. An LSTM model capable of predicting prices in Dongjak-gu three months into the future was configured, and the prepared training and test data are shown in <Figure 2>. Firstly, address was limited to Dongjak-gu to specify scale for primary training. And secondly, left outer merge based on month and year of contract. Number of residents is monthly data standard but real sale price data overlaps since they are monthly, left outer was applied based on number of residents. Thirdly, pivoted based on address and month and year of contract to remove overlapped data resulted from real transaction price → Standardization job.


Dataset configuration scenario
Grouping was carried out according to contract year/month (left outer merge)
No. of residents
Multiple actual trading prices
Explanatory variable (x)
Target variable (y)

• Although the no. of residents is based on monthly data, since the actual trading price data is repeated monthly, left outer join is applied according to the no. of residents

```
mergeDF = pd.merge(base,MainDF, how = 'left', on=('계약년월'))
```

Figure 2. Dataset configuration scenario

4.4 Data pre-processing for LSTM implementation (predictions three months into the future)

To implement an LSTM model for return rate predictions, a three-dimensional data input structure of batch size, time steps, and input lengths was used. <Table 4> describes the data pre-processing process for 'return rate' prediction. We operate various dataset such as Data set, y value, Statistics, Difference (preprocessing), Time shift, Date data, X data, Y data, train point, min, max, Normalization, Normalization x RAW, LSTM input value, LSTM Train, LSTM valid, and LSTM test as follow. The Table 3 shows all the dataset with the source and interpretation each.

Table 4. Data pre-processing process

	Data set	Source	Description
Dataset	**RawDF**	**Data set**	**Configure initial dataset**
y-value	yDF	RawDF	Define y-values of target variable
Statistics	StatDF	RawDF	Statistical analysis of current state
Difference (pre-processing)	DiffDF	RawDF	Configure time series data in RNN format

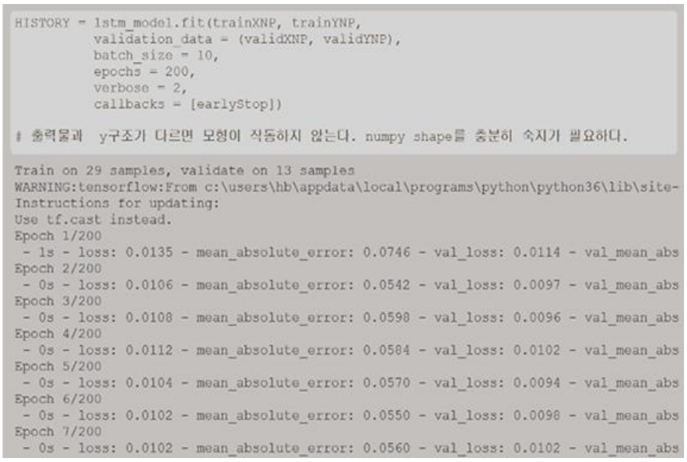
Period shift	**ShiftDF**	diffDF+yDF	Configure time series data in RNN format
Date data	dateDF	-	-
X data	**shiftXDF**	-	-
Y data	shiftYDF	-	-
Training period	trainDF	shiftDF	Generate training data
	~~trainYDF~~	trainDF	
	trainXDF	trainDF	
min,max	minTrainXSR	trainXDF	Preparations for normalization
	maxTrainXSR	trainXDF	
Normalization	**normXDF**	**shiftXDF, maxminTrainXSR**	**Normalization**
Normalization x RAW	normXrawYDF	normXDF, shiftYDF	-
LSTM input value	xNP	normXrawYDF	Form conversion suitable for LSTM inputs
	yNP	normXrawYDF	
	subDateDF	normXrawYDF	
LSTM Train	train XNP	xNP	For training
	train YNP	yNP	
LSTM valid	valid XNP	xNP	For validation
	valid YNP	yNP	
LSTM test	test XNP	xNP	For testing
	test YNP	yNP	

To properly apply the raw DF extracted via the above dataset configuration scenario to the LSTM model, a data pre-processing process involving data conversion and classification into training/validation/testing data is performed prior to the implementation of the LSTM model. In the aforementioned hypothesized equation, the y-value is defined as yDF, and DFs are configured for statistics required for each LSTM, differentials, and each shift. As shown in the LSTM modeling, each data structure is configured as shown above.

Finally, before defining each input value in the LSTM network, each data entry is normalized to be scaled to a suitable size for the LSTM model. After normalizing the DFs, the input, output, and shift values of the aforementioned equation are defined and each data entry is categorized into training, validation, and testing data to conclude the pre-processing process.

5. Experiments and Results

To implement the LSTM model for the prediction of return rates, the training period was set as April 2015~August 2017 (29 months), the validation period was set as September 2017~September 2018 (13 months), and the test period was set as December 2018~December 2019 (13 months) according to the aforementioned time series datasets. After designating separate periods for training, a total of 69 time series data entries were used for the model validation: training (40 data entries), validation (15 data entries), data (15 data entries), as shown in <Figure 3>.



```
HISTORY = lstm_model.fit(trainXNP, trainYNP,
          validation_data = (validXNP, validYNP),
          batch_size = 10,
          epochs = 200,
          verbose = 2,
          callbacks = [earlyStop])

# 출력물과  y구조가 다르면 모형이 작동하지 않는다. numpy shape를 충분히 숙지가 필요하다.

Train on 29 samples, validate on 13 samples
WARNING:tensorflow:From c:\users\hb\appdata\local\programs\python\python36\lib\site-
Instructions for updating:
Use tf.cast instead.
Epoch 1/200
 - 1s - loss: 0.0135 - mean_absolute_error: 0.0746 - val_loss: 0.0114 - val_mean_abs
Epoch 2/200
 - 0s - loss: 0.0106 - mean_absolute_error: 0.0542 - val_loss: 0.0097 - val_mean_abs
Epoch 3/200
 - 0s - loss: 0.0108 - mean_absolute_error: 0.0598 - val_loss: 0.0096 - val_mean_abs
Epoch 4/200
 - 0s - loss: 0.0112 - mean_absolute_error: 0.0584 - val_loss: 0.0102 - val_mean_abs
Epoch 5/200
 - 0s - loss: 0.0104 - mean_absolute_error: 0.0570 - val_loss: 0.0094 - val_mean_abs
Epoch 6/200
 - 0s - loss: 0.0102 - mean_absolute_error: 0.0550 - val_loss: 0.0098 - val_mean_abs
Epoch 7/200
 - 0s - loss: 0.0102 - mean_absolute_error: 0.0560 - val_loss: 0.0102 - val_mean_abs
```

Figure 3. Data training process

The LSTM model was trained using the return rate training data. In this process, optimal parameters (obtained via trial-and-error) resulting in result values with the lowest error were applied.

The LSTM model consists of one input layer, two hidden layers, and a final output layer. The optimal LSTM model (obtained via trial-and-error), resulting in result values with the lowest error was obtained as shown in <Figure 4>.

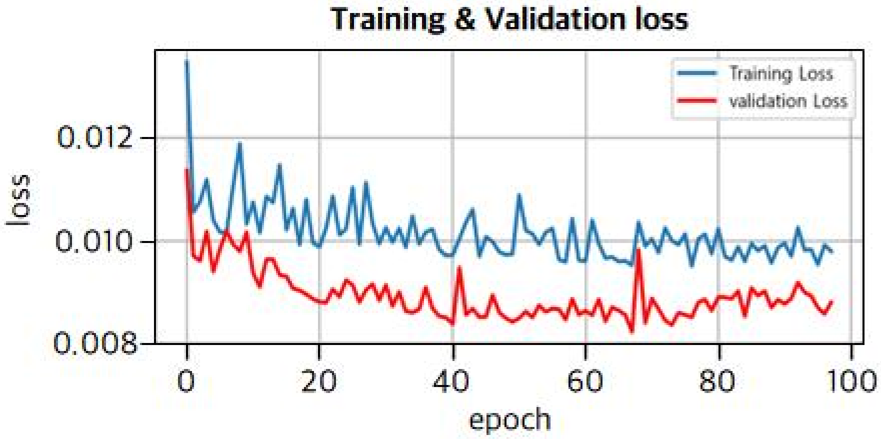


Figure 4. Comparative validation of the LSTM model

The return rate ‘epoch’ refers to the number of iterations for training. It is clear that the loss error rate decreases as the number of iterations increases. When the model is trained repeatedly with various training and validation combinations, the error rate is reduced due to the nature of the variables. A decreased error rate indicates that the property of a variable can be expressed as a general property, resulting in universal results like ’A is B’.

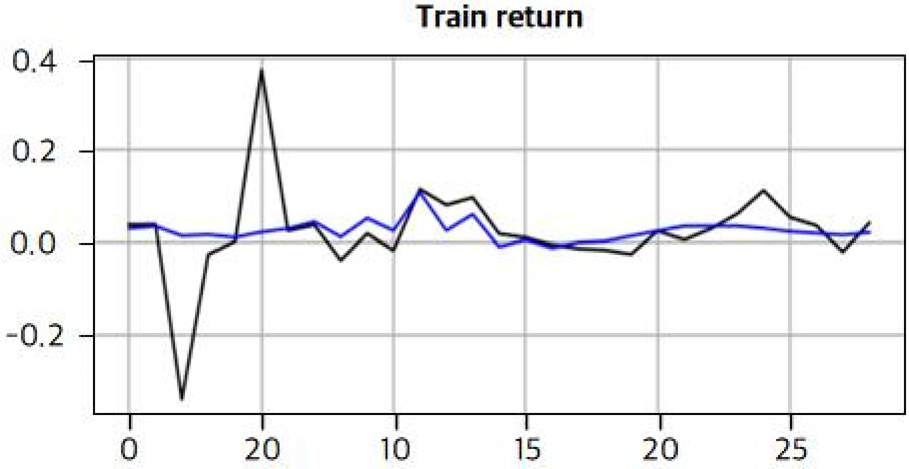


Figure 5. Comparison of 'return rate' training (blue: predicted values, black: actual return rate)

The training model was visualized by constructing a model using the training data. The y-values (trading prices) that were predicted when inputting the x-values (demographic data) of the training data to the model were marked with a blue line, and the actual y-values (trading prices) of the training data were marked with a black line to compare with the predicted values. As a result, the two lines were consistent except for certain overfitted instances. Using the tested model, the following results were obtained.

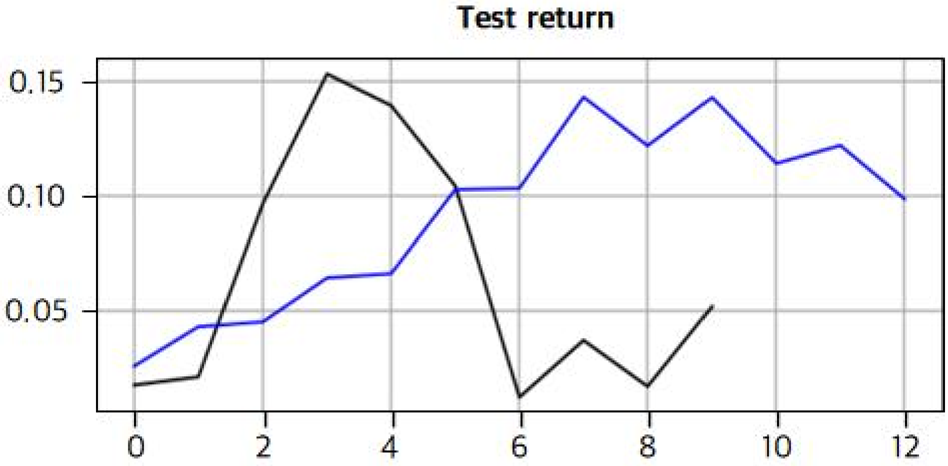


Figure 6. Return rate validation

Similar to the testing process, the y-values (trading prices) that were predicted when inputting the x-values (demographic data) of the data to the model were marked with a yellow line, and the actual y-values (trading prices) of the data were marked with a white line to compare with the predicted values. <Figure 6> illustrates the visualized results for 'return rate.' Here, it is shown that the return rate prediction values deviate somewhat from the actual trend over the past three years. Here, the trading price prediction results exhibited an upwards trend that was consistent with the trend of the actual values over the past three years.

The experimental results demonstrate that the predicted results showed similarities to the actual values in terms of trading prices and price index but deviated from the actual values in terms of return rate.

The results produced a correlation value of –0.206, which indicate a weak inverse correlation between the y-values (trading prices, rate_pred) produced by the model and the actual trading prices (rate_real). The final model exhibited an accuracy of 76% when comparing the trading price data (three months into the future) obtained using the model with the actual trading prices (y).

6. Conclusions and Limitations

This study, which aimed to predict 'return rates' in the real estate market, produced the following conclusions. First, the model that was produced resulted in a prediction similarity of nearly 76%.

In other words, it was confirmed that 76% of models could be configured by collecting time series data. Therefore, the study confirmed that the LSTM method could produce return rate predictions with a certain degree of reliability. In the process of developing the LSTM time series model, this study also explored various other modeling techniques required to select input data (explanatory variables) and obtain results. Most notably, this study is meaningful in that it confirmed the potential of such modeling techniques with real estate time series data. This resulted in the successful application of a technique widely used in the stock market to the real estate market, highlighting the diverse applicability of the LSTM model.

However, there are certain limitations in the research that future studies should address to achieve more accurate return rate predictions. First, it is necessary to secure a greater amount of data to raise the return rate prediction accuracy beyond 76%. Future studies should train AI with more data to improve the model. However, it is worth noting that there is a lack of real estate-related data that can be used for analysis. For this reason, it will be necessary to create prediction models specifically designed for the real estate market by collecting abundant data from public data or MyData sources. Additionally, there were limitations in the process of deriving appropriate parameters. Future studies should attempt to extract parameters that can ensure higher reliability by performing numerous simulations, even if via a trial-and-error approach.

References

[1] Antipov, E. A., and Pokryshevskaya, E. B.,"Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and A CART-based Approach for Model Diagnostics," Expert Systems with Applications, Vol. 39, No. 2, 2012, pp.1772-1778.

[2] Azadeh, A., Ziaei, B., and Moghaddam, M., "A Hybrid Fuzzy Regression-fuzzy Cognitive Map Algorithm for Forecasting and Optimization of Housing Market Fluctuations," Expert Systems with Applications, Vol. 39, No.1, 2012, pp. 298–315.

[3] Baek, Y., and Kim, H. Y., "ModAugNet: A New Forecasting Framework for Stock Market Index Value with An Overfitting Prevention LSTM Module and A Prediction LSTM Module," Expert Systems with Applications, Vol.113, 2018, pp. 457-480.

[4] Bin, O., "A Prediction Comparison of Housing Sales Prices by Parametric versus Semi-parametric Regressions," Journal of Housing Economics, Vol. 13, No. 1, 2004, pp. 68-84.

[5] Karevan, Z., and Suykens, J. A., "Transductive LSTM for Time-series Prediction: An Application to Weather Forecasting," Neural Networks, Vol. 125, 2020, pp.1-9.

[6] Patrick, J., Okunev, J., Ellis, C., and David, M., "Comparing Univariate Forecasting Techniques in Property Markets," Journal of Real Estate Portfolio Management, Vol. 6, No. 3, 2000, pp. 283-306

[7] Plakandaras, V., Gupta, R., Gogas, P., and Papadimitriou, T., "Forecasting the US Real House Price Index," Economic Modelling, Vol. 45, 2015, pp. 259-267.

[8] Wang, X., Wen, J., Zhang, Y., and Wang, Y., "Real Estate Price Forecasting Based on SVM Optimized by PSO," Optik, Vol. 125, No. 3, 2014, pp. 1439-1443