International Journal of Advanced Culture Technology Vol.10 No.1 274-283 (2022)
DOI https://doi.org/10.17703/IJACT.2022.10.1.274

IJACT 22-3-34

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

[1]Jeong-hyun Lee, [2]Hoo-bin Kim, [3]Gyo-eon Shim*

[1]PhD Candidate, Dept. of Real Estate Studies, Konkuk Univ., Korea
[2]Master Candidate, Dept. of Artificial Intelligence, Yonsei Univ., Korea
[3]Prof., Dept. of Real Estate Studies, Konkuk Univ., Korea
*leejohn.news@gmail.com, source234@naver.com, x1000@konkuk.ac.kr**

Abstract

Korea is facing a number difficulties arising from rising housing prices. As 'housing' takes the lion's share in personal assets, many difficulties are expected to arise from fluctuating housing prices. The purpose of this study is creating housing price prediction model to prevent such risks and induce reasonable real estate purchases. This study made many attempts for understanding real estate instability and creating appropriate housing price prediction model. This study predicted and validated housing prices by using the LSTM technique - a type of Artificial Intelligence deep learning technology. LSTM is a network in which cell state and hidden state are recursively calculated in a structure which added cell state, which is conveyor belt role, to the existing RNN's hidden state. The real sale prices of apartments in autonomous districts ranging from January 2006 to December 2019 were collected through the Ministry of Land, Infrastructure, and Transport's real sale price open system and basic apartment and commercial district information were collected through the Public Data Portal and the Seoul Metropolitan City Data. The collected real sale price data were scaled based on monthly average sale price and a total of 168 data were organized by preprocessing respective data based on address. In order to predict prices, the LSTM implementation process was conducted by setting training period as 29 months (April 2015 to August 2017), validation period as 13 months (September 2017 to September 2018), and test period as 13 months (December 2018 to December 2019) according to time series data set. As a result of this study for predicting 'prices', there have been the following results. Firstly, this study obtained 76 percent of prediction similarity. We tried to design a prediction model of real estate transaction price with the LSTM Model based on AI and Bigdata. The final prediction model was created by collecting time series data, which identified the fact that 76 percent model can be made. This validated that predicting rate of return through the LSTM method can gain reliability.

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

1. INTRODUCTION

'Housing' dominates a household's assets in the Korean society. Housing price fluctuations exert major influence on households. For a long time, people perceived real estate as a stable asset. However, in macroeconomic terms, housing prices were not always increasing but they have been affected by many

Manuscript received: February 15, 2022 / revised: March 1, 2022 / accepted: March 8, 2022
Corresponding Author: x1000@konkuk.ac.kr
Tel: +82-02-450-3364
Professor, Dept. of Real Estate Studies, Konkuk Univ., Korea

variables. In that sense, many people are sensitive toward housing price fluctuations since their entire asset can be endangered. Such an instability can bring about extreme speculative investment in real estate or abstinence in making real estate purchase due to the concern toward housing price fall. This study pays attention to immovability of real estate. Price fluctuation resulting from geographical conditions is inevitable since real estate is settled on land. There have been a number of attempts to predict price fluctuations; one of them is deep learning using Artificial Intelligence. This study aimed to predict real estate prices by using deep learning technology.

2. RESEARCH METHOD

As the method of this study, Long Short-Term Memory (in brief, LSTM) [1] models were used. There were many LSTM models regarding existing stock prices [2] and this study aimed to look into applicability of predictability by applying the LSTM model regarding times series data to housing prices. Such LSTM techniques have been widely used in stock market price predictions but the cases when it was applied to real estate are few. The main purpose of this study is identifying whether housing prices [3, 4, 5] can be predicted by the LSTM model in the real estate market [6, 7, 8] This study's data analysis can explore price predictability of housing districts with high preference. For this aim, this study used real sale prices of apartments in Korea provided by the Ministry of Land, Infrastructure, and Transport. Real estate's price data - time series data - have seasonal features and preprocessing procedure is important since trends and periodic components have been combined. They have been enhanced to high quality data by applying the HP filter (Hodrick-Prescott filter) technique in order to extract time series data trends. Also, in a bid to finalize appropriate target market districts, districts with similar prices have been clustered by applying the SOM (Self-organizing map) technique, one of data mining techniques, based on AI network. In order to provide explanation power toward input variables based on similar group districts and organize a prediction model regarding future price trends for analysis targets, out of the SVR (Support Vector Regression) technique and the deep learning technique, the LSTM technique appropriate for time series data prediction was used. LSTM is suitable for long-term time series prediction and extensively used for predicting time series data. This study tried to validate whether it has excellent prediction performance even toward real estate price prediction.

3. THEORETICAL CONSIDERATION

Neural networks has been supported as a strong tool for predicting single time series data and diverse techniques have been developed to supplement its internal problems. Together with MLP, some leading techniques are RNN (Recurrent Neural Networks), ESN (Echo State Networks), GRNN (Generalized Regression Neural Networks), and LSTM; recently, RNN is being widely applied in time series data prediction.

The hidden layer of existing neural networks has merely neurons which did not take into account context. While such a structure can predict time series data by adopting past hidden layer, RNN has directed cycle in which past events can influence future outcomes. RNN, which processes consecutive information, is mostly used for data which have correlations in terms of time. The next t+1 data are predicted by taking into account the correlation between immediate past data (t-1) and current data (t) and the neural network reflecting past data is created. In the case when the distance between information and the point using information is far, RNN has the disadvantage of having the vanishing gradient issue and LSTM is the model which supplemented the issue.

LSTM, which added cell state, conveyor belt role, to the existing RNN's hidden state, is the network which recursively calculates cell state and hidden state <Figure 1>.

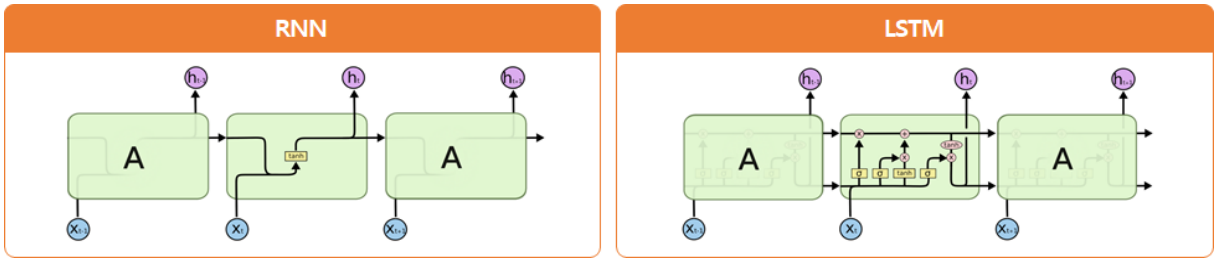


Figure 1. RNN and LSTM

The LSTM network was introduced by Hochreiter and Schmidhuber (1997) and they expanded the concept of memory cell and gating mechanism which control information flow at RNN by supplementing vanilla RNN's long-term memory. Through saving information and reducing errors for longer time on network, LSTM comes to supplement the disadvantages of RNN. It is learning saving, removing, and reading on long-term state with the basic structure of LSTM.

The purpose of this study is applying the LSTM model regarding the aforementioned time series data to real estate prices.

4. DATA MANIPULATION

4.1 Data Collection

This study used public housing real sale prices and domestic apartment information provided by the Ministry of Land, Infrastructure, and Transport and the Public Data Portal. Autonomous districts' apartment real sale prices from January 2006 to December 2019 were collected through the Ministry of Land, Infrastructure, and Transport's real sale price open system and basic apartment and commercial district information were collected through the Public Data Portal and the Seoul Metropolitan City Data. The collected real sale price data were scaled based on monthly average sale price and a total of 168 data were organized by preprocessing based on address. The ultimately organized data were shown in respective autonomous districts' average price. When dividing overall data into data for learning and data for validation in the rate of 4 to 1, there were a total of 135 data for learning and 33 data for validation. <Table 1> is data collection. This instructed the method of predicting respective districts' 'real estate prices'.

Table 1. Data collection

Table	Table information	Property information	Source
DF1	Domestic apartment information	Name, address, sale form, number of households, construction company, etc.	Public Data Portal
DF2	Apartment real sale price	Name, address, year and date of contract, area for exclusive use, and transaction cost	Ministry of Land, Infrastructure, and Transport
DF3	Commercial district analysis information	Standard date, commercial district address, number of working population (males/females/aggregate)	Seoul Metropolitan City Data
DF4	Demographic information	Autonomous district, time series population monthly information	-

DF: DataFrame

4.2 Data Set Composition

Data set was based on address information of DF2, which was the essential baseline. Data set was organized by deriving explanatory variables based on address and time series and since the ultimate target is future sale price, sale price of target variable DF2 was selected. <Figure 2> is the basic data set composition and <Table 2> is the data set. DF1's sale period and construction period can be added as explanatory variables in the data set.

$$Y(t) = aX_1(t) + bX_2(t) + c$$

$$s.t.$$

$$X_1 : dedicated\ area \quad X_2 : no.of\ residents \quad Y : trade\ price$$

Figure 2. Basic data set composition

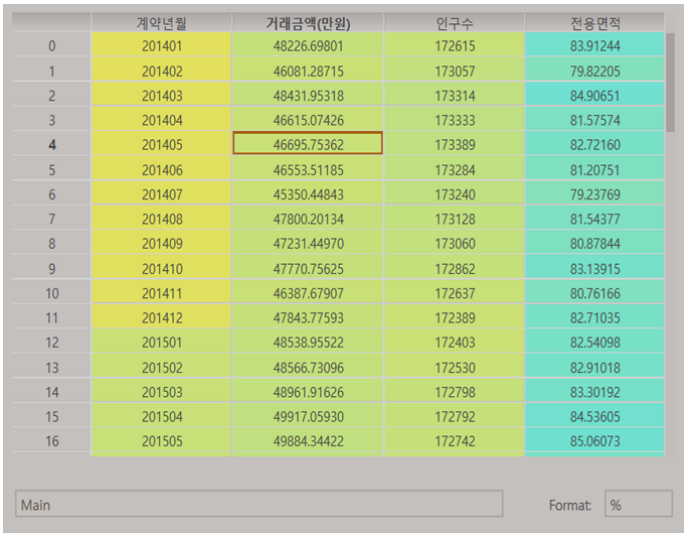
Table 2. Data set

	Real sale price	Demographic information	**Final data set**	Commercial district information	Applied factor
Baseline	DF2	DF4	**-**	DF3	DF1
Combined element	**Address**	**-**	**Address**	-	Address
Time series	**Month and year of contract**	**Standard month and year**	**Standard month and year**	Standard quarter	-
Explanatory variable (x)	**Area of exclusive use**	**Number of residents**	**Number of residents, area of exclusive use**	Floating population (workplace)	Sale period, etc.
Target variable (y)	**Transaction cost**	**-**	**Transaction cost**	-	-

4.3 Scenario and Data Set Composition

After undergoing several simulations, data were organized to predict 'price' of Dongjak-gu's 3 months. The LSTM model which can predict market prices of Dongjak-gu 3 months from now and training data and test data were organized as in <Figure 3>.

Figure 3. Data set composition scenario



	계약년월	거래금액(만원)	인구수	전용면적
0	201401	48226.69801	172615	83.91244
1	201402	46081.28715	173057	79.82205
2	201403	48431.95318	173314	84.90651
3	201404	46615.07426	173333	81.57574
4	201405	46695.75362	173389	82.72160
5	201406	46553.51185	173284	81.20751
6	201407	45350.44843	173240	79.23769
7	201408	47800.20134	173128	81.54377
8	201409	47231.44970	173060	80.87844
9	201410	47770.75625	172862	83.13915
10	201411	46387.67907	172637	80.76166
11	201412	47843.77593	172389	82.71035
12	201501	48538.95522	172403	82.54098
13	201502	48566.73096	172530	82.91018
14	201503	48961.91626	172798	83.30192
15	201504	49917.05930	172792	84.53605
16	201505	49884.34422	172742	85.06073

Main Format: %

Figure 4. Final composition data set example

4.4 Data Preprocessing to Apply LSTM (Predict 3 Months from Now)

Input data structure for implementing the LSTM model for predicting prices is three-dimensional – Batch Size, Time Steps, and input lenghts – and divided into three parts: number of arrangement, sequence length, and input dimension. <Table 3> shows data preprocessing procedure and contents for predicting 'price'. We use 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 3. Data preprocessing procedure

	Data set	Source	Interpretation
Data set	**RawDF**	**Data set**	**Initial data set composition**
y value	yDF	RawDF	Definition of target value y
Statistics	StatDF	RawDF	Current statistics analysis
Difference (preprocessing)	DiffDF	RawDF	Time series data set in RNN form
Time shift	**ShiftDF**	diffDF+yDF	Time series data set in RNN form
Date data	dateDF	-	-
X data	**shiftXDF**	-	-
Y data	shiftYDF	-	-
train point	trainDF	shiftDF	Creation of training data
	~~trainYDF~~	trainDF	
	trainXDF	trainDF	
min,max	minTrainXSR	trainXDF	Preliminary work prior to normalization
	maxTrainXSR	trainXDF	
Normalization	**normXDF**	**shiftXDF, maxminTrainXSR**	**Normalization**

Normalization x RAW	normXrawYDF	normXDF, shiftYDF	-
LSTM input value	xNP	normXrawYDF	Form conversion suitable for LSTM input
	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 test
	test YNP	yNP	

In order to adequately apply Raw DF, which was ultimately extracted through the aforementioned data set composition scenario, complete data preprocessing procedure for operating LSTM after data conversion work and training/validation/test data classification.

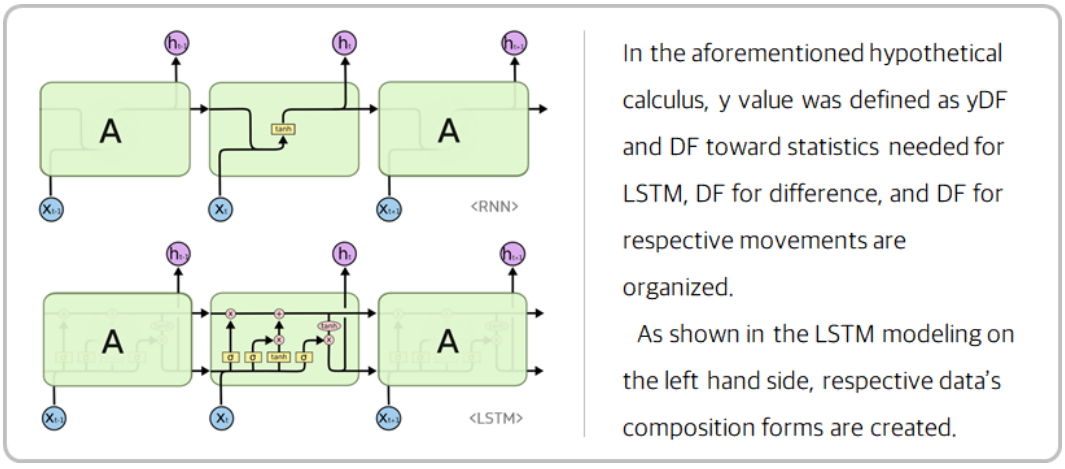


Figure 5. RNN, LSTM modeling

Prior to defining respective input values in the LSTM in the final stage, organize appropriate data through respective data's normalization work and scaling data into scale and size fit for the LSTM model.

The preprocessing procedure is complete once the ultimately normalized DF are defined as input value, output value, and shift value and classified into for training, for validation, and for test.

5. EXPERIMENT AND RESULTS

The LSTM implementation process for predicting real estate prices was conducted as the following: 29 months of training (April 2015 to August 2017), 13 months of validation (September 2017 to September 2018), and 13 months of test period (December 2018 to December 2019) (refer to <Figure 6>).

```
training 기간 : 201504 ~ 201708
valid 기간    : 201709 ~ 201809
test 기간     : 201812 ~ 201912
```

Figure 6. Period setting

Period was divided to conduct time series training and test, a total of 69 validations – 40 trainings, 15 validations, and 15 data – were conducted.

The LSTM model was learned with the training data for real estate price prediction and when it came to the parameter needed for the LSTM model regarding the LSTM learning process, optimized parameter with the lowest error in result value after undergoing trials and errors was applied. <Figure 7> is the details of data parameter.

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 5, 3)	0
lstm_0 (LSTM)	(None, 5, 64)	17408
lstm_1 (LSTM)	(None, 64)	33024
dense (Dense)	(None, 1)	65

Total params: 50,497
Trainable params: 50,497
Non-trainable params: 0

Figure 7. Details of data parameter

The LSTM model is comprised of one input layer, two hidden layers, and the final output layer and the model was comprised as in <Figure 8> based on optimized values with the lowest error result value.

Figure 8. Comparison validation of LSTM model

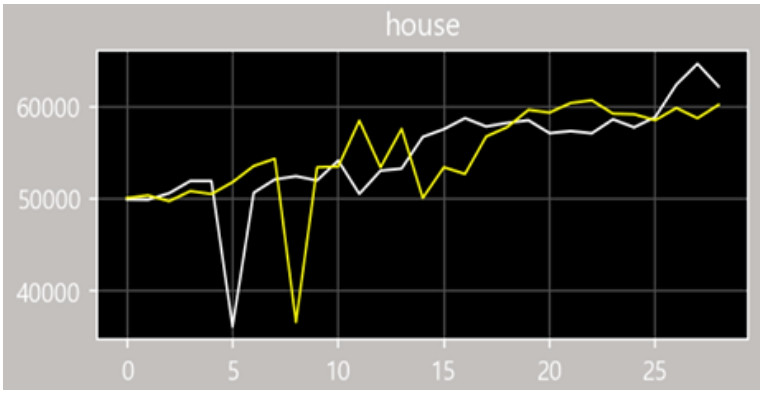


Figure 9. 'Average transaction price' training comparison (yellow is prediction value and white is actual price)

Real estate price 'epoch' refers to number of repetitions in learning. As the repetition number increases, loss error rate decreases. This can be interpreted as error rate decreasing owing to variable property when model repeatedly learns several training pairs and validation pairs. Error rate decrease, expressing variable property as a certain ordinary property, can be said to create universal results such as 'A is B'.

Visualization of training model was conducted through training data. The comparison was made by predicting y value (transaction price) after entering training data x value (demographic information), marked as 'yellow' line, and marking actual training data y value (transaction price) in 'white' line. As a result of the comparison, excluding certain overfitting, they coincided. Based on previously conducted test model, the following results have been derived.

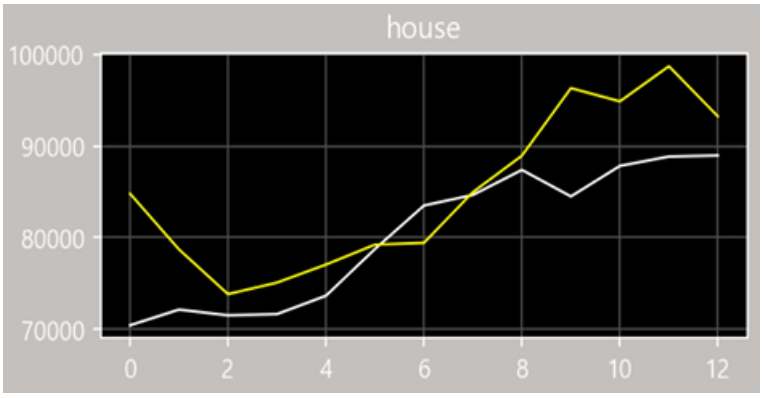


Figure 10. Average price validation

Similar to the aforementioned test, prediction value was compared by marking y value (transaction price) prediction as 'yellow' line and data's y value (transaction price) as 'white' line. As a result of the comparison. <Figure 10> is price index prediction and as a result of predicting transaction prices, it somewhat coincided with past 3 years of progress. As a result of the experiment, there was slight difference between transaction price coincidence and rate of return but index coincided.

```
realYKDF = pd.DataFrame(testYNP, columns = ['rate_real'])
predYKDF = pd.DataFrame(rate_pred, columns = ['rate_pred'])
resultDF = pd.concat([realYKDF, predYKDF], axis = 1)
correlation = resultDF.corr()
print('[ Correlation ] ', correlation.iloc[0,1])
correlation
```

```
[ Correlation ]  -0.206110653060629
```

	rate_real	rate_pred
rate_real	1.000000	-0.206111
rate_pred	-0.206111	1.000000

```
directionAccL = np.sign(resultDF['rate_real']) == np.sign(resultDF['rate_pred'])
directionAcc = directionAccL.mean()
print('[ Direction Accuracy ] ', directionAcc)
```

```
[ Direction Accuracy ]  0.7692307692307693
```

Figure 11. Correlation analysis results

<Figure 11> is correlation analysis results. As a result of the analysis, -0.206 index emerged and a slight reverse correlation between this model's y value (transaction price, rate_pred) and original transaction price (rate_real) was found. There was approximately 76 percent of accuracy between the final model's transaction price (after 3 months) data and original transaction price (y).

6. CONCLUSION AND LIMITATION

We tried to design a prediction model of real estate transaction price with the LSTM Model based on AI and Bigdata. As a result of this study designed for predicting real estate prices, the following results have been identified. First, there was approximately 76 percent of prediction similarity rate. The final prediction model was created by collecting time series data and it has been found that 76 percent model can be made. This implies that predicting rate of return through the LSTM method can be somewhat reliable. This study identified diverse modeling techniques necessary for acquiring setting input data and obtaining results through the LSTM time series model. In particular, it is significant that such a possibility has been identified through real estate time series data. This study validated diverse application ability of LSTM by succeeding in expanding the technique, which had been applied in the stock market, to the real estate market.

However, for more accurate real estate price prediction, there were some limitations and some aspects which follow-up studies should supplement. It has been found that more data need to be secured to increase rate of return prediction ability to higher than 76 percent. In the future, it is necessary to conduct AI learning after securing more data. Particularly, in the case of data related to real estate, amount of data for analysis was scarce. In that sense, it is necessary to secure enough data through public Bigdata [9, 10] and Mydata in the future and create prediction models inherent for real estate. In addition, there was limitation in deriving appropriate parameter. Even though there may be some trials and errors in follow-up studies, extracting parameters which have higher reliability is needed through diverse simulations.

REFERENCES

[1] 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. DOI: https://doi.org/10.1016/j.neunet.2019.12.030

[2] 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. DOI: https://doi.org/10.1016/j.eswa.2018.07.019

[3] 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. DOI: https://doi.org/10.1016/j.eswa.2011.07.020

[4] Fan, G. Z., Ong, S. E., and Koh, H. C.,"Determinants of House Price: A Decision Tree Approach," Urban Studies, Vol. 43, No. 12, 2006, pp. 2301-2315. DOI: https://doi.org/10.1080/00420980600990928

[5] 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. DOI: https://doi.org/10.1016/j.jhe.2004.01.001

[6] Plakandaras, V., Gupta, R., Gogas, P., and Papadimitriou, T., "Forecasting the US Real House Price Index," Economic Modelling, Vol. 45, 2015, pp. 259-267. DOI: https://doi.org/10.1016/j.econmod.2014.10.050

[7] 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. DOI: https://doi.org/10.1016/j.ijleo.2013.09.017

[8] 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. DOI: https://doi.org/10.1080/10835547.2000.12089608

[9] Y.I. Kim, S.S. Yang, S.S. Lee, S.C. Park, "Design and Implementation of Mobile CRM Utilizing Bigdata Analysis Techniques", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 14, No. 6, 2014, pp. 289-294. DOI: https://doi.org/10.7236/JIIBC.2014.14.6.289

[10] S.J. Oh, "Design of a Smart Application using Bigdata", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 15, No. 6, 2015, pp. 17-24. DOI: https://www.earticle.net/Article/A259710