UNIVERSITY OF INFORMATION TECHNOLOGY

FACULTY OF INFORMATION SYSTEMS

Using Statistical Models and Machine Learning To Predict Housing Prices

- PHAM QUOC ANH - 20521077
- NGUYEN TRAN CHI DUC - 20521199
- NGUYEN CHI KIEN - 18520953

Using Statistical Models and Machine Learning To Predict Housing Prices

Pham Quoc Anh
Faculty of Information Systems
University of Information Technology
Thu Duc City, Vietnam
20521077@gm.uit.edu.vn

Nguyen Tran Chi Duc
Faculty of Information Systems
University of Information Technology
Thu Duc City, Vietnam
20521199@gm.uit.edu.vn

Nguyen Chi Kien
Faculty of Information Systems
University of Information Technology
Thu Duc City, Vietnam
18520953@gm.uit.edu.vn

***Abstract*— Real estate is always one of the topics that are always interesting in Vietnam, especially in big cities like Ho Chi Minh City, Hanoi, and Da Nang... The development of a housing prices prediction model can assist a house seller or a real estate agent to make better-informed decisions based on house price valuation.**

Keywords—Real estate, housing prices

I. Introduction

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry's operational mechanism and driving factors.

There are several approaches that can be used to determine the price of a house, one of them is the prediction analysis. The first approach is a quantitative prediction. A quantitative approach is an approach that utilizes time-series data. The time-series approach is to look for the relationship between current prices and prevailing prices. The second approach is to use linear regression based on hedonic pricing. Previous research conducted by Gharehchopogh, using a linear regression approach get 0,929 errors with the actual price. In linear regression, determining coefficients generally uses the least square method, but it takes a long time to get the best formula.

In this field, since the data used for prediction and analysis are nonlinear and have time-dependent issues, we will use linear regression, non-linear regression, TBATS model, ARIMA model, and machine learning regression as LSTM to develop a house price prediction model. We use house price data in Ho Chi Minh city.

II. Related work

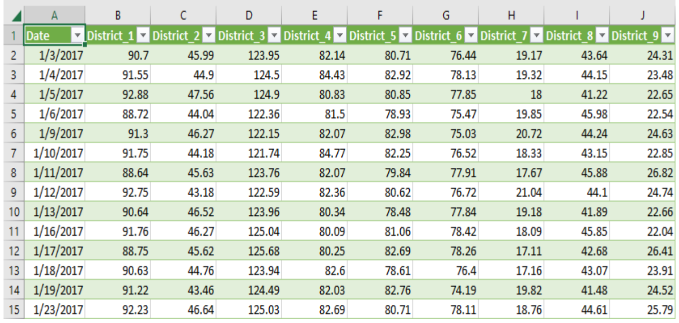
1) Predict Housing Prices using TBATS and ARIMA model; Research Arpad Gellert[1], Ugo Fiore[2], Adrian Florea[3], Radu Chis[4], Francesco Palmieri [5], [1,3,4] Computer Science and Electrical Engineering Department, Lucian Blaga University of Sibiu, Romania, [2,5]Department of Computer Science, University of Salerno, Italy; In this paper the author uses TBATS and ARIMA Model to evaluate and compared them with other models on datasets from a household equipped with photovoltaics and an energy management system.
2) Predict Housing Prices using multiple non-linear regression and multiple linear regression; Research Sheung-Chi Chow[1], Juncal Cunado[2], Rangan Gupta[3], Wing-Keung Wong[4], [1]Hang Seng Management College, Sha Tin, Hong Kong, [2] Faculty of Economic and Business Sciences, University of Navarra, Pamplona, Spain, [3]Department of Economics, University of Pretoria, Pretoria, South Africa, [4] Department of Finance, Asia University, Taichung, Taiwan; Department of Economics, Lingnan University, New Territories, Hong Kong; In this paper the author using non-linear regression and linear regression to found that there is only a linear causal relationship from India's EPU growth to China's housing returns.
3) Predict Housing Prices using LSTM model; Research Rui Liu, Lu Liu, Center for Assessment and Development Research of Real Estate, Shenzhen, 518000, China; In this paper the author uses LSTM model to overcome the problems of traditional models, a long short-term memory approach is proposed to predict the housing price of a city by using historical data.

III. Dataset

As mentioned above, our article will provide a house price prediction. For this purpose, we use a dataset that provides house price prediction data, the dataset we choose is selected from Kaggle. It contains prices from district 1 to district 9 with 10 columns and 1364 rows, and the data collected from 2017 to 2022.

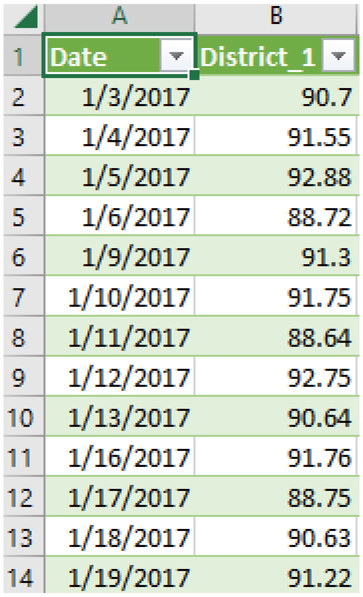
We chose District 1 because this is a crowded place for tourists from home and abroad, transactions in goods and entertainment also often take place here, and this is also the center of the city, house prices tend to be higher than in neighboring areas.

DATA AFTER PROCESSING



	A	B	C	D	E	F	G	H	I	J
1	Date	District_1	District_2	District_3	District_4	District_5	District_6	District_7	District_8	District_9
2	1/3/2017	90.7	45.99	123.95	82.14	80.71	76.44	19.17	43.64	24.31
3	1/4/2017	91.55	44.9	124.5	84.43	82.92	78.13	19.32	44.15	23.48
4	1/5/2017	92.88	47.56	124.9	80.83	80.85	77.85	18	41.22	22.65
5	1/6/2017	88.72	44.04	122.36	81.5	78.93	75.47	19.85	45.98	22.54
6	1/9/2017	91.3	46.27	122.15	82.07	82.98	75.03	20.72	44.24	24.63
7	1/10/2017	91.75	44.18	121.74	84.77	82.25	76.52	18.33	43.15	22.85
8	1/11/2017	88.64	45.63	123.76	82.07	79.84	77.91	17.67	45.88	26.82
9	1/12/2017	92.75	43.18	122.59	82.36	80.62	76.72	21.04	44.1	24.74
10	1/13/2017	90.64	46.52	123.96	80.34	78.48	77.84	19.18	41.89	22.66
11	1/16/2017	91.76	46.27	125.04	80.09	81.06	78.42	18.09	45.85	22.04
12	1/17/2017	88.75	45.62	125.68	80.25	82.69	78.26	17.11	42.68	26.41
13	1/18/2017	90.63	44.76	123.94	82.6	78.61	76.4	17.16	43.07	23.91
14	1/19/2017	91.22	43.46	124.49	82.03	82.76	74.19	19.82	41.48	24.52
15	1/23/2017	92.23	46.64	125.03	82.69	80.71	78.11	18.76	44.61	25.79

Figure 1: Dataset



	A	B
1	Date	District_1
2	1/3/2017	90.7
3	1/4/2017	91.55
4	1/5/2017	92.88
5	1/6/2017	88.72
6	1/9/2017	91.3
7	1/10/2017	91.75
8	1/11/2017	88.64
9	1/12/2017	92.75
10	1/13/2017	90.64
11	1/16/2017	91.76
12	1/17/2017	88.75
13	1/18/2017	90.63
14	1/19/2017	91.22

Figure 2: Data of District 1

IV. Methodology

In this paper, we use 4 statistic models (linear regression, non-linear regression, TBATS, and ARIMA) and 1 machine learning (LSTM).

A. *Linear Regression*

Linear regression is a data analysis technique that predicts the value of unknown data using another known and related data value. It mathematically models unknown or dependent variables and known or independent variables as a linear equation. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

The formula for the linear regression equation is given:

y = a + bX

Where:

b = Slope of the line.

a = Y-intercept of the line.

X = explanatory variable.

Y = dependent variable.

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas of business and academic study. You will find that linear regression is used in everything from biological, behavioral, environmental, and social sciences to business. Linear regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear regression models are well understood and can be trained very quickly.

Result of Linear Regression:

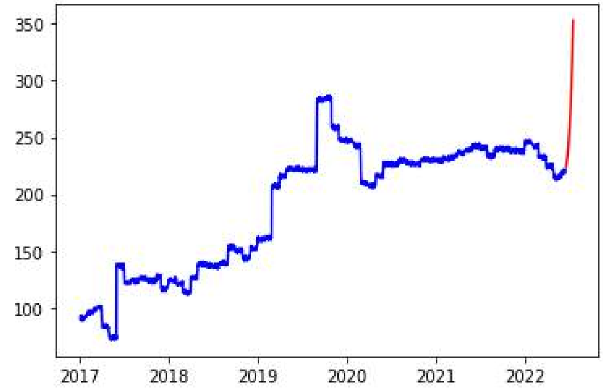


Figure 3: Linear Regression

We use our data to perform multi Linear Regression.
Summary output:

Regression Statistics	
Multiple R	0.977369358
R square	0.955250863
Adjusted R Square	0.954986466
Standard Error	11.76349248
Observations	1363

	Coefficients	***Standard Error***	***t Stat***	***P-value***
District 1	-23.592	3.004	-7.852	8.2165E15
District 2	0.210	0.075	2.768	0.005713517
District 3	0.544	0.034	15.756	1.74234E-51

District 4	1.062	0.045	23.593	2.2491E103
District 5	-0.179	0.027	-6.511	1.0434E10
District 6	-0.592	0.063	-9.406	2.11841E-20
District 7	0.880	0.069	12.675	7.08794E-35
District 8	-0.233	0.112	-2.080	0.03770665
District 9	0.973	0.120	8.095	1.25809E-15

The independent value is District 1.
The dependent value is from District 2 to District 9.
After analysis by using Linear Regression, we have the result:
Y = -23.592 + 0.210x_1 + 0.544x_2 + 1.062x_3 - 0.179x_4 - 0.592x_5 + 0.880x_6 - 0.233x_7 + 0.973x_8

B. *Non-Linear Regression*

Non-linear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function. Simple linear regression relates two variables (X and Y) with a straight line (y = mx + b), while nonlinear regression relates the two variables in a nonlinear (curved) relationship.
Nonlinear regression uses nonlinear regression equations, which take the form:
$Y = f(X, \beta) + \varepsilon$
Where:
X = a vector of p predictors,
β = a vector of k parameters,
f (-) = a known regression function,
ε = an error term.
It's a method for performing more flexible nonlinear analysis to obtain proper outputs such as choices, categorization, or inferences when similar future states or inputs are present. This method can provide impressive results and frequently beats people in performance, stability, and precision.
The goal of the model is to make the sum of the squares as small as possible. The sum of squares is a measure that tracks how far the Y observations vary from the nonlinear (curved) function that is used to predict Y.
It is computed by first finding the difference between the fitted nonlinear function and every Y point of data in the set. Then, each of those differences is squared. Lastly, all of the squared figures are added together. The smaller the sum of these squared figures, the better the function fits the data points in the set. Nonlinear regression uses logarithmic functions, trigonometric functions, exponential functions, power functions, Lorenz curves, Gaussian functions, and other fitting methods.

Result of Non-Linear Regression:

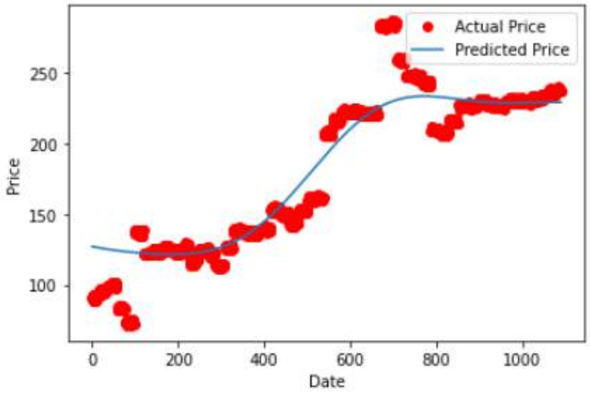


Figure 4: Non-linear Regression

C. *ARIMA*

An Autoregressive Integrated Moving Average, or ARIMA, ARIMA is a statistical model used for forecasting time series data. The ARIMA equation is a regression-type equation in which the independent variables are lags of the dependent variable and/or lags of the forecast errors.
The equation of the ARIMA model is given as :
$y'(t) = c + \phi1* y'(t-1) + \cdots + \phi p*y'(t-p) + \theta1*\varepsilon(t-1) + \cdots + \theta q*\varepsilon(t-q) + \varepsilon t$
There are three terms in the equation:
AR: Auto Regression: The time series is regressed with its previous values i.e., y(t-1), y(t-2), etc. The order of the lag is denoted as p.
I: Integration: The time series uses differencing to make it stationary. The order of the difference is denoted as d.
MA: Moving Average: The time series is regressed with residuals of the past observations i.e., error ε(t-1), error ε(t-2), etc. The order of the error lag is denoted as q.
The ARIMA model uses differenced data to make the data stationary, which means there's a consistency of the data over time. This function removes the effect of trends or seasonality, such as market or economic data.
Seasonality occurs when data exhibits predictable, repeating patterns. It is critical to control for seasonality because it could impact the accuracy of the results.
ARIMA models can be built using seasonal and nonseasonal formats. A seasonal model must take into account the number of events in each season in addition to the autoregressive, differencing and average terms for each season.
ARIMA models can be built in an array of software tools, including Python. Before deciding on an ARIMA model, the data scientist must confirm that the process in question fits the model. If the data is an appropriate fit for the ARIMA model, the data scientist builds the model and trains it on a dataset before inputting live data to develop and plot a forecast.

The result of the ARIMA model:

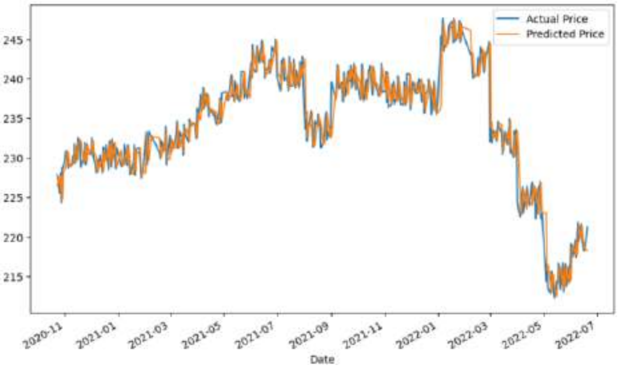


Figure 5: ARIMA model

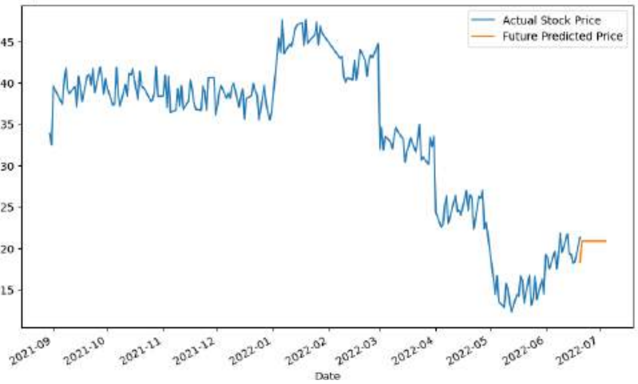


Figure 6: The prediction 30 days later

D. TBATS

TBATS is a time series model that is useful for handling data with multiple seasonal patterns, i.e., the data that changes over time. The TBATS is preferred over BATS as the Trigonometric seasonality (TBATS) can deal with complex and high frequency. TBATS is an acronym for key features of the model: T: Trigonometric seasonality B: Box-Cox transformation A: ARIMA errors T: Trend S: Seasonal components This recipe demonstrates an example of TBATS modeling of time series.

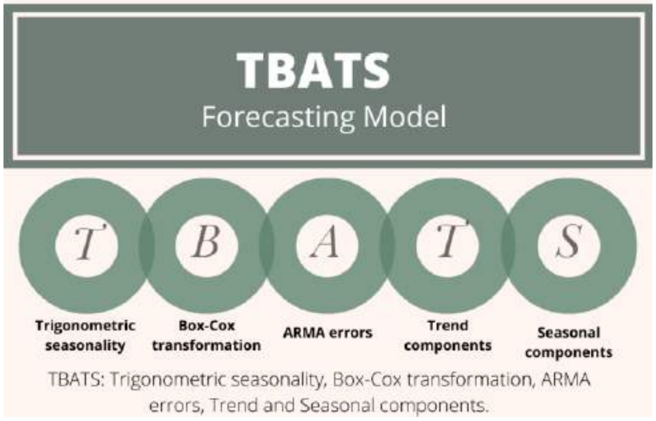

TBATS
Forecasting Model
T
B
A
T
S
Trigonometric seasonality
Box-Cox transformation
ARMA errors
Trend components
Seasonal components
TBATS: Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components.

Figure 7: TBATS

TBATS uses the trigonometric representation of seasonal components based on the Fourier series.

Model

$$y_t^{\lambda} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} S_{t-m_i}^{(i)} + d_t$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t$$

$$b_t = \phi b_{t-1} + \beta d_t$$

$$d_t = \sum_{i=1}^{p} \varphi_i \, d_{t-i} + \sum_{i=1}^{q} \theta_i \, e_{t-i} + e_t$$

Where:

y_t^{λ} – Time series at moment t (Box-Cox transformed)

$s_t^{(i)}$ – ith seasonal component

l_t – local level

b_t – trend with damping

d_t – ARMA(p,q) process for residuals

e_t – Gaussian white noise

Seasonal Part

$$s_t^{(i)} = \sum_{j=1}^{(k_i)} s_{j,t}^{(i)}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos(\omega_i) + s_{j,t-1}^{*(i)} \sin(\omega_i) + \gamma_1^{(i)} d_t$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \sin(\omega_i) + s_{j,t-1}^{*(i)} \cos(\omega_i) + \gamma_2^{(i)} d_t$$

$$\omega_i = 2\pi j / m_i$$

Model paraments:

T **– Amount of seasonalities**

m_i – Length of ith seasonal period

k_i – Amount of harmonics for ith seasonal period

λ – Box-Cox transformation

β ,α - Smoothing

ϕ – Trend damping

φ_i θ_i – ARMA(p,q) coefficienst

$\gamma_2^{(i)}$ $\gamma_1^{(i)}$- Seasonal smoothing (two for each period)

The result of the TBATS model:

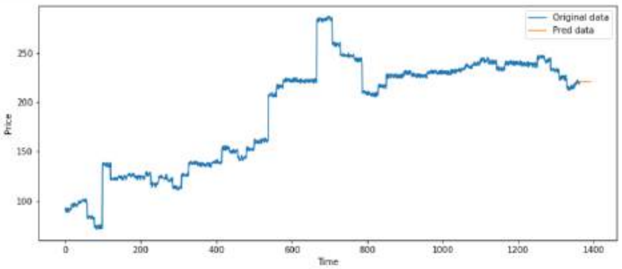


Figure 8: The prediction 30 days later

E. LSTM

Long short-term memory (LSTM)[is an artificial neural network used in the fields of artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Such a recurrent neural network (RNN) can process not only single data points (such as images) but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition, machine translation, robot control, video games, and healthcare.

The name LSTM refers to the analogy that a standard RNN has both "long-term memory" and "short-term memory". The connection weights and biases in the network change once per episode of training, analogous to how physiological changes in synaptic strengths store long-term memories; the activation patterns in the network change once per time-step, analogous to how the moment-to-moment change in electric firing patterns in the brain store short-term memories. The LSTM architecture aims to provide a short-term memory for RNN that can last thousands of timesteps, thus "long short-term memory".

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

LSTM networks are well-suited to classifying, processing, and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

The structure of LSTM:

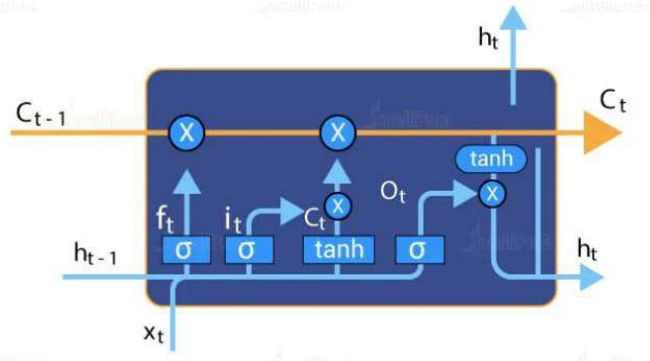


Figure 9: LSTM

The result of the LSTM model:

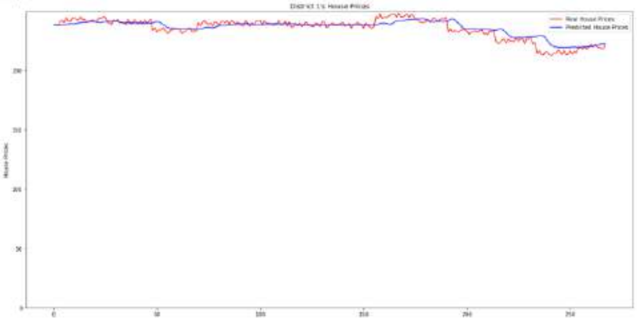


Figure 10: LSTM model with train/test 8/2

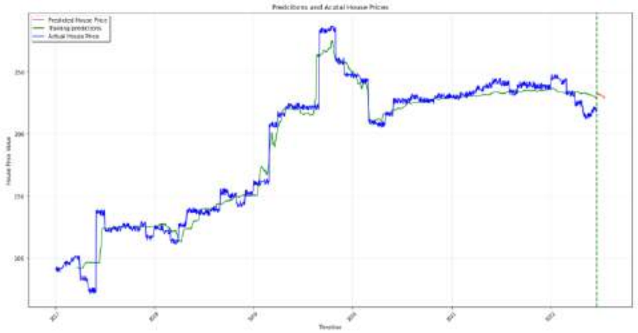


Figure 11: The prediction 30 days later with train/test 8/2

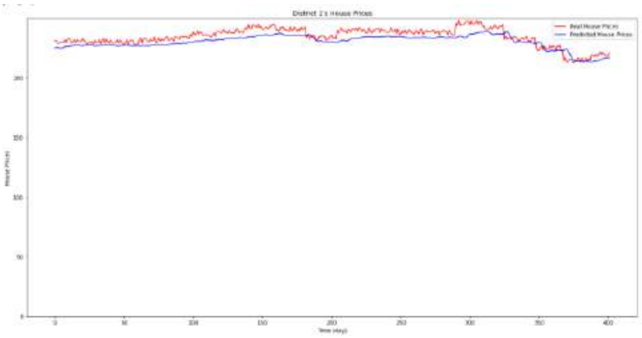


Figure 12: LSTM model with train/test 7/3

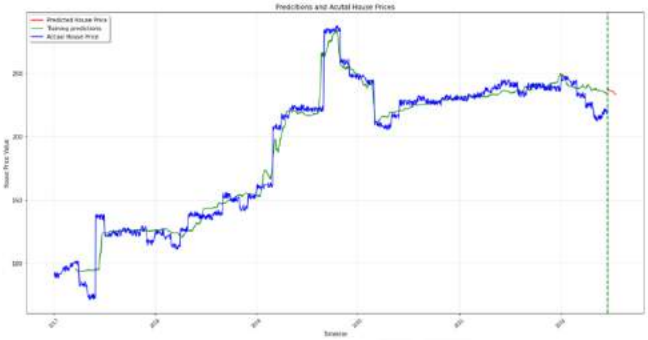


Figure 13: The prediction 30 days later with train/test 7/3

V. CONCLUSION

Data table on train test, MAPE, RMSE of Linear Regression, NonLinear Regression, ARIMA, TBATS, LSTM models.

	Train Test	**RMSE**	**MAPE**
LR	10,0	26.16	23.09%
NLR	10,0	19.02	6.42%
LSTM	8,2	4.11	1.28%
	7,3	5.39	2.02%
ARIMA	8,2	16.8232881 3581944	0.06472455 276376615
	7,3	15.3203058 31709565	0.05928648 754035152

TBATS	7,3	8.25256082831204	2.56509913070307
	8,2	9.05383881082447	2.6680975403299
	9,1	12.3227860797379	4.12957157538224

Looking at the data table, we see that LSTM is the model with the best predictive data, with train test (8.2) (7.3), RMSE with 4.17 and 3.56, MAPE with 4.12%, and 3.04%. The models utilized apartment transaction data over 66 months. The results produced a set of price-prediction models for District 1 from Ho Chi Minh City. Despite the limited interpretation of the predictions over the pre-existing price data, the merit of the approach taken in this study lies in its ability to predict housing prices, both in the long and short term, with improved accuracy. This is meaningful, as it can be used to evaluate and simulate changes in the regional and local housing markets.

The modeling process shows that distance measures have a significant impact on housing prices. The price trend in the area fluctuated strongly from the beginning of 2019 to the end of 2019, and tended to decrease from the middle of 2020 due to the impact of the Covid-19 epidemic, after this period, house prices tended to stabilize and slightly increase. Future studies should also take a combined approach using complementary methods such as regression-based modeling to determine the individual effects of the relevant variables. This approach can also determine whether the proximity effect is detrimental to nearby housing prices.

The application of machine learning techniques for constructing a housing price model was outstanding. In this study, the LSTM managed to produce a set of price-prediction models efficiently, with less margin of error than a traditional time series model. However, there are also some models such as SARIMAX, and LSTM correctly predicts data 30 days later, when home values show signs of decreasing compared to reality. The reason is the Vietnamese government has intervened in real estate to prevent price inflation from capitalists.

With respect to the implication of the research, the modeling approach employed in this study can assist in the decision-making process in the planning, development. Researchers can use these models to make their own plans, they can plan, and use their own capital to build and develop. In addition, city planners can add assessment processes to ensure economic stability and viability in neighborhoods. With further expansion of the modeling approach, different socioeconomic dimensions can be tested to address urban problems, including infrastructure/resource shortages.

Lastly, the superiority of the machine learning technique in analyzing large time-series data can be valuable in assessing the impact of various built environment factors, market factors, and the impact of housing policies over time. This can be useful in understanding the effect of certain housing policy and market dynamics over time based on big-data, and simulating the response of the housing market

VI. REFERENCE

[1] Master's in Data Science, ARIMA Modeling.

[2] Nadeem, Time Series Forecasting using TBATS Model, Nov 21, 2021.

[3] Chen Xing, Why TBATS?, 2022-05-07.

[4] Shweta Tyagi, Introduction to Time Series Forecasting — Part 2 (ARIMA Models), Bengaluru, Karnataka, India, Jul 31, 2021.

[5] Rui Liu, Lu Liu, Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm, 19 Feb, 2019.

[6] Arpad Gellert, Ugo Fiore, Adrian Florea, Radu Chis, Franceso Palmieri, Forecasting Electricity Consumption and Production in Smart Homes through Statistical Methods, Jan 2022.

[7] Bruce Nguyen, Aalto University, End-to-End Time Series Analysis and Forecasting: a Trio of SARIMAX, LSTM and Prophet (Part 1), Finland, Feb 7, 2021.

[8] Ken Riippa, Bruce Nguyen, Taeyoung Kee, Atreya Ray, Son Nguyen and Duong Tran, Finnish Housing Market Prediction using Time Series Analysis.

[9] Sheung-Chi Chow, Juncal Cunado , Rangan Gupta und Wing-Keung Wong, Causal relationships between economic policy uncertainty and housing market returns in China and India: evidence from linear and nonlinear panel and time series models, 4. September 2017.

[10] Will Kenton, What Is Nonlinear Regression? Comparison to Linear Regression, May 29, 2022.

[11] What is LSTM? Introduction to Long Short Term Memory, 13th Dec, 2022.

[12] What is Linear Regression? - AWS.

[13] From AR to SARIMAX: Mathematical Definitions of Time Series Models, https://phosgene89.github.io/sarima.html.

[14] https://blog.tenthplanet.in/time-series-forecasting-tbats/, Time-Series Forecasting using TBATS model.

[15] https://blog.tenthplanet.in/time-series-forecasting-tbats/, Time-Series Forecasting using TBATS model.

[16] https://en.wikipedia.org/wiki/Long_short-term_memory, Long short-term memory.

[17] https://en.wikipedia.org/wiki/Long_short-term_memory, Long short-term memory.