A Machine Learning and Deep Learning Based Solution for Predicting Land Values

Bimali Y.M.Y.
Department of Software Engineering
Faculty of Computing, Sri Lanka Institute of Information Technology (SLIIT)
Malabe, Sri Lanka.
yashinkabimali@gmail. com

Rodrigo U.S.D.
Department of Software Engineering
Faculty of Computing, Sri Lanka Institute of Information Technology (SLIIT)
Malabe, Sri Lanka.
supundileepar@gmail.com

Gamage M.P.A.W.
Department of Information Technology
Faculty of Computing, Sri Lanka Institute of Information Technology (SLIIT)
Malabe, Sri Lanka.
anjalie.g@sliit.lk

Rathnayaka P.B.
Department of Information Technology,
Faculty of Computing, Sri Lanka Institute of Information Technology (SLIIT)
Malabe, Sri Lanka.
pasangi.r@sliit.lk

***Abstract*— Real Property are the most valuable possession of most of the common people. Getting the proper valuation for these real properties is very much important. This document analyses an innovative solution proposed to facilitate land valuation based on recent sales, prediction of future price and the effect of proposed development work on the land, so that real-estate customers and owners of real estate companies can be benefitted and make smarter property related decisions. This intelligent tool can help people to identify the land they are going to buy in terms of current value and future value. Machine learning, and optimization are the main research components of this system. The system utilizes LSTM model as well as KNN and MLR model in making predictions. LSTM model can make predictions with an accuracy of over 0.75 in current value prediction and also future value predictions with reasonable accuracy. This paper discusses about the research methodology we have used in identifying the most suitable algorithms which can serve our intended purpose.**

Keywords—Valuation, AI- Artificial Intelligence, ML- Machine learning, ANN- Artificial Neural Network, LSTM- Long Short Term Memory, RNN- Recurrent Neural network, MLR- Multivariate Regression, ARIMA- Auto Regressive Integrated Moving Average, MAE- Mean Absolute Error, MSE- Mean Squared Error, RMSE- Root Mean Squared Error

I. Introduction

Real Property are the most valuable possession of most of the common people. In Sri Lankan culture, most of the people tend to think that owning a real estate is a better investment than having that money saved in a bank. Therefore, getting the proper valuation for these real properties is very much important.

Land valuation is the process of assessing the characteristics of a given piece of land based on experience and judgment.[1] The determination of a land parcel value depends on several physical and economic characteristics which must be taken into consideration very carefully in a land valuation procedure.[1] These values can be affected by various social factors too. For example, if there is a crime happened in that land, it can cause a negative effect on the value.

Hence, real estate appraisal it is a challenging multidimensional problem that involves estimating many facets of a property, its neighborhood, and its city.[2]

Since, Sri Lanka is lacking a good data platform to gather all these data, considering all these factors can take ages to do proper valuation considering all these factors. The manual process is a time-consuming slow task which needs to be done by an experienced professional valuer. The valuation approaches used by those professionals are limited due to the lack of digital data in Sri Lanka. Also, it is a known fact that the valuation process can be so subjective to the person. Ideally, the systematic process of valuation consists of four different stages as physical and legal identification, identification of property rights to be valued, gathering and analysis of market data, applying convenient valuation approach. The major convenient valuation approaches are, Sales Comparison Approach, Income Approach, Cost Approach [3]. Analyzing the previous land sale details and trends in those fluctuations and considering those data to predict the valuation is called the sales comparison approach.[3]

The task of automatically estimate the market value of houses can be seen as a regression problem, where the price (or the price per square meter) is the dependent variable, while the independent one is the available information that could help to determine the price correctly. [2] When the neighborhood economical value is combined with effect of neighborhood factors such as walkability etc. we believe it is possible to give a accurate, fair prediction of the value of the land.

The influence of technology on daily life of the Sri Lankans has increased immensely. People tend to use traffic data, online shopping more than ever.

Since the manual process is too slow and dependent to make a quick better decision of the worthiness of the land and suitability of it for the purpose of the customer, our attempt is to digitally assist the people in property related decision making by providing them accurate predictions of the values and future studies of the land. The main research problem is to develop an automated system to evaluate the land based on its neighborhood economical value and identify the possible effects of development work on the value of the land in the future. This requirement of a solution to predict the current value and future value came from an expertise. While reviewing the literature, by means of supervisor meetings, we identified another aspect as an improvement, which is to predict the effect of future development work on a particular land, since Sri Lanka is a developing country, although the rate of development may vary, infrastructure development projects are carried out frequently.

We can never underestimate the duty of a valuation officer as the estimations are affected by numerous factors of particular to the area. But these factors are subjected to perception of each other's experience, according to Vaz J.[4], the discretionary and the appraisers' subjectivity that characterize traditional real estate valuation are still allowed to take part in the formation of the asset price even when respecting international standards (EVS, IVS) or Appraisal Institution´s regulations (TEGOVA, RICS, etc.). For example, an experienced valuer who is familiar with the area maybe biased towards the effect of regional factors, social factors, than the physical factors compared to a fairly new valuer who still sticks to the land valuation theories and follow the proven procedure. Therefore, manual valuation can be considered as a more sensitive approach.

Our intention is to provide people with fair accurate prediction of the land they are going to buy, so that they can decide the investment is fruitful for them. We believe this is an area improvement is needed because we can assist people in making decisions related to property, which would be the largest investment most probably in many people's lives.

During the AI Asia Summit 2018, the summit panelists Dr. Yasantha Rajakarunanayake, Dr Rukshan Baduwita , Dr. James Shanahan and Dr. Chrisantha Fernando agreed that Sri Lanka is behind in terms of AI startups[5], despite the fact software industry is vastly growing area. According to the survey conducted under research done by Karunanda et al[6], carried out in 2014, this is due to the lack of popularity, knowledge, experts, requirements and sponsorship for the AI related software projects[6].

But when analyzing local news, we can see that AI based applications has become a trend. For Example, Dialog has its own AI powered voice service to support its product service framework.

There are researches that have been conducted to predict the Stock prices of Sri Lanka with the usage of Artificial Intelligence and Machine Learning approaches, tilted A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market[7], and Comparison of Support Vector Regression and Artificial Neural Network Models to Forecast daily Colombo Stock Exchange[8]. According Li et al, [1]to the real estate valuation researches evaluating the use of GIS technology have been conducted. But there is no information regarding application of AI technology in real estate value prediction in Sri Lankan context.

The use of AI for residential value forecasting has been suggested in the literature from 1990s. [9]. Although Sri Lanka is lacking an automated land valuation system, many up and running, reliable solutions have been implemented in developed countries like New Zealand, England and Wales, USA etc. It is obvious with the well-structured digital data infrastructure of those countries, they can implement very accurate systems.

Zillow is an online real estate database company that was founded in 2006, and was created by Rich Barton and Lloyd Frink, former Microsoft executives and founders of Microsoft spin-off Expedia. [10] Zillow.com supports United States of America (USA) and Canadian property listing. Zestimate determines an estimation for 12 months for a house based on neighbourhood comparable houses. Accuracy of zestimate depends on the amount of data used as the underlying approach is Hedonic regression analysis based proprietary algorithm [11] which analyses of several features of the house. The forecasted value is interpolated using cubic spline to connect to current value. [11]

Trulia is also a product offered in USA, which offers a range of services for real estate sector. The price estimates are based on publicly available information the home's physical characteristics (e.g. location, number of bedrooms, etc.), Property tax information, Recent sales of similar nearby homes.

It involves more community interaction, for example, Trulia Neighbourhoods provide photographs, drone footage, etc. so that who are interested about the neighbourhood can refer. Trulia provides price using public data which shows the price fluctuation of a house, comparative to the other homes with same ZIP code.

Quotable Value (QV) provides independent and authoritative information on any home in New Zealand on or off the market [12] QV.co.nz and their mobile App QV homeguide is known to be providing more accurate values of real estate property and key details to assist people to make instant decisions regarding property. QV with CoreLogic, a company which analyzes information assets and data to provide clients with analytics and customized data services provide a range of reports valuable to the user.

Creating a methodology that would bring more sophisticated information, greater accuracy and analytical rigor to the United Kingdom (UK) residential property market is the motivation behind HousePrice.ai. Their proprietary model provides a combination of multi-disciplinary experiences of AI and Big Data to provide most accurate estimations. HousePrice.ai has Horizon app, which calculates capital, rental and gross development values for a single property or an entire portfolio. [13] it produces accurate property valuations both in the present time and can offer future predictions. Valuations are based on objective measurable values, creating a fact-based result as opposed to a subjective one [14]. This tool allows the user to adjust, add and remove factors within the surrounding areas to determine how external changes will affect property prices.

Our intention is to identify the ways to use their underlying methodology in a suitable manner in Sri Lankan context.

II. Methodology

A. Data Collection

The study focuses on Homagama diviosional secretariat division, Colombo which experiences relatively high infrastructure development and higher sales rate of lands.

Primary data have been collected through questionnaires, interviews and personal visits to land area to know the present situation of the market and the secondary data are collected mainly through various survey department, land estate agents, newspaper advertisements, and land sale website contents. The data are useful for assessing the performance of property as a key to predict land price.

The cross-sectional data collected for current price prediction to be used with non-time-series algorithm were collected through a questionnaire where residents in Homagama, Colombo district responded and by means of including publicly available data in newspaper and website advertisements. The questionnaire mainly asked for price of the land, location of the land, nearest bus route, and distance to the nearest bus route, along with the buying price and details of valuation history with above 200 samples. These properties were selected based on findings of [18] which suggested of a better valuation model suitable for Sri Lanka rather than the previous annual value index model.

The time series data collected to predict the current value from a land sale company which had monthly land values from the same area over a period of 10 years, containing above 200 samples.

B. System Design

When a customer goes to a land he is willing to buy, they can input the current location through the application. Based on that location, the suitable recent sales data are selected. Then those data will be analyzed by the AI model to predict the current value. That predicted value is optimized to produce the most accurate current value. The application of machine learning, and deep learning algorithms have been tested in each of the components with suitable data.

C. Tested models

1) Multivariate Linear Regression

MLR is an algorithm used in both the components of current value prediction and future value prediction. Simply, it is assuming that there is linear relationship between price predictions and other contributing factors.

MLR has several advantages than other algorithms. The ability to determine the relative influence of one or more predictor variables to the criterion value. multivariate techniques provide a powerful test of significance compared to univariate techniques.[15] multivariate techniques to give meaningful results, they need a large sample of data; otherwise, the results are meaningless due to high standard errors. [15] Standard errors determine how confident you can be in the results, and you can be more confident in the results from a large sample than a small one.

MLR model implementation finds the best fitting line using model coefficients. Process of optimizing the model is to minimize the error of the predicted value.

The MLR algorithm used for current value prediction component analyzed the factors location, distance to the main bus route, accessibility index, size of the land during testing.

2) Random forest regressor

Random forest regressor operates by constructing a multitude of decision trees to fit the observations into groups based on their attribute values and outputs the mean prediction of the individual trees. As the name suggests, “decision tree” model builds a reversed tree-like structure, where the “root” is at the top, followed by multiple branches, nodes and leaves. The end of each branch is a decision leaf, which is the model’s predicted value, given the values of the attributes represented by the path from the root node to the said decision leaf.

This model was tested for current value prediction component with the same features tested with MLR model.

3) Artificial Neural Networks

ANN design concept is based on human brain. The purpose of ANN is to imitate human learning process. This model consists of mainly three types of layers namely, input layer, hidden layer and output layer, each layer having artificial neurons contribute in adjusting weights for the input features and attempt making conclusions just like the human brain is doing.

The ANN was also trained for the current value prediction with same dataset used for MLR. Through a trivial trial and error process suitable model was identified and compared with the others.

4) LSTM – Recurrent Neural Network

Considering the fact that time has a direct influence on land prices time-series algorithms were also tested for selecting best prediction model for current price. What makes LSTM different from typical neural network is that it has feedback connections. To test this model , timeseries dataset having monthly land values from the area over a period of 10 years was used. The dataset was having lags of unknown duration hence out of available RNN types, LSTM was the best option.

5) ARIMA model

ARIMA standing for Auto Regressive Integrated Moving Average is the most popular and commonly used statistical method for time series prediction. It is a combination of the two models based on Auto Regressive process (AR) and Moving Average (MA) process, where both the processes are stochastic processes.

AR process is defined as in (1) below where p is the order of the process found out using Partial Auto Correlation function (PACF).

$$x_t = \mu + \varepsilon_t + \sum_{i=1}^{p} \theta_i \in_{t-i} \qquad (1)$$

While MA process is defined as in (2) below, where q is the order of the process found out using Auto Correlation Function (ACF).

$$x_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \in_{t-i} \qquad (2)$$

Procedure to follow with this model is split the training dataset into train and test sets, use the train set to fit the model, and generate a prediction for each element on the test set. A rolling forecast is required given the dependence on observations in prior time steps for differencing and the AR model. A crude way to perform this rolling forecast is to re-create the ARIMA model after each new observation is received.

III. Results and discussion

The results obtained by testing the above models in the two different domains of current value prediction is discussed here.

This was carried out as two phases testing time-series algorithms and vice versa. As mentioned above in II, machine learning models LSTM and ARIMA were tested with time series data while MLR, Random forest regressor, and ANN was tested with cross sectional data. These models were evaluated in terms of mean absolute error (MAE), mean standard error(MSE) and root mean squared error (RMSE).Test results for these models can be summarized as follows.

Table 1 : Summary of findings

	MAE	**MSE**	**RMSE**
MLR	12578.2076	37057375442	192502.923
Random Forest Regressor	69388.61903	17241415729	131306.572
ANN	495306.848	351944183254.44	593248.838
LSTM	12150.774	1834424960	42830.187
ARIMA	26549.4523	4559474.12	2135.29251

According to above results, time series algorithms predicted values with comparatively less error than the others. It can be concluded that ARIMA model has outperformed all the other machine learning models in land price prediction.

The data used for ARIMA model have been resampled with monthly mean as depicted in Fig. 1 below. The correlogram in Fig. 2 depicts that the number of significant correlations at the first or second lag followed by correlations that are not significant.

For the term of AR, using the PACF in Fig. 3 we will be using three. Based on the pattern of ACF depicted in , we cannot infer the terms for MA, zero will be the best option. As per the standardized residual plot in Fig. 4, we can observe that most of the data are distributed around zero. The density graph Fig. 5, also displays a normal distribution.

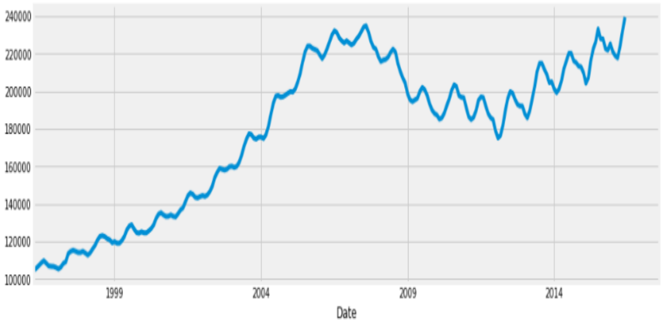

Fig 1. Monthly average of land prices

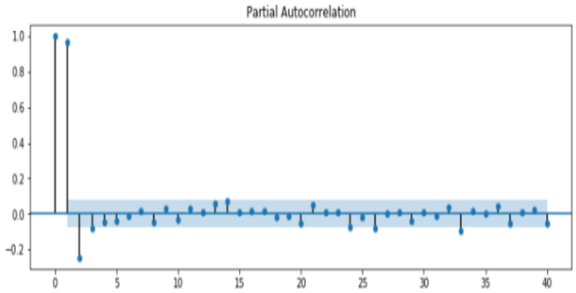

Fig 2. Partial Auto Correlation Function

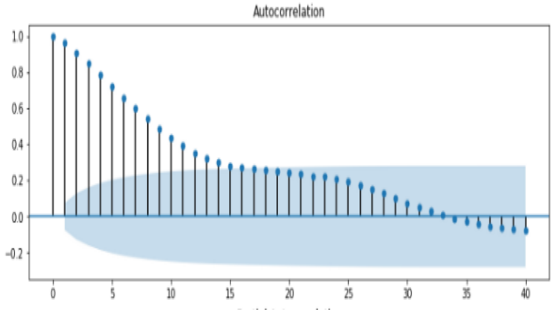

Fig 3. Auto Correlation Function

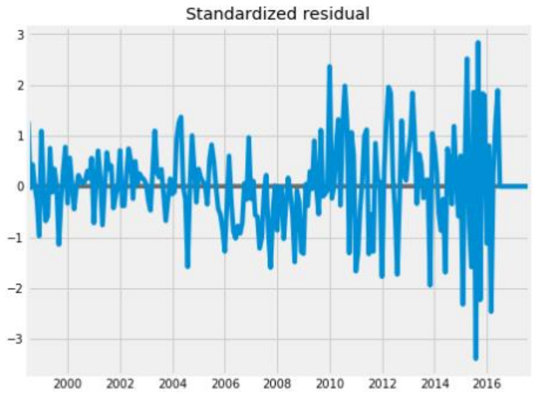

Fig 4. Standardized residual function

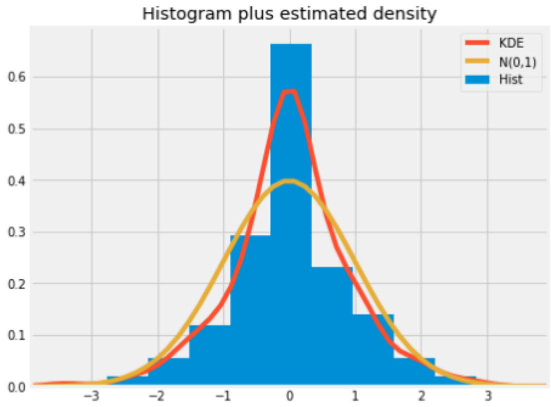

Fig 5. Density distribution

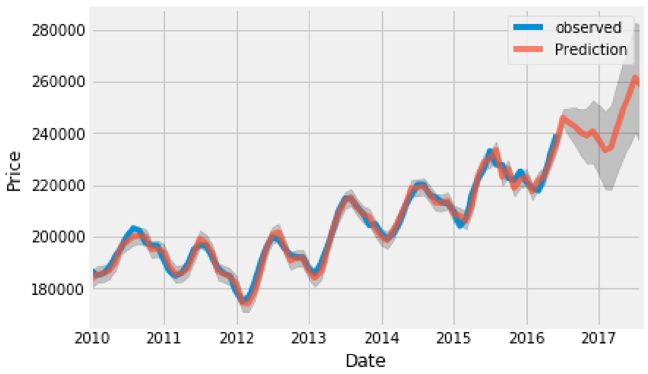

Fig 6. Predictions compared with observed dataset

The graph Fig. 6 depicts the predictions got through rolling forecast, which depicts well fitted predictions for the testing data. Hence, the predictions made with ARIMA model can be taken as the most suitable for our purpose.

IV. Conclusion

Based on the observations above, we can conclude that ARIMA model has the least error among the other tested models in predicting current value. But there can be tradeoffs, depending on the dataset being used and its sample size. Hence, further work on these models are recommended with different features considered based on different valuation models and with greater sample size.

Acknowledgment

We would like to express our gratitude to the Faculty of Computing, Sri Lanka Institute of Information Technology (SLIIT) for providing us with a good academic environment and facilities to complete this project. We also want to express our thankful words to our families, friends and lectures for their understanding and support on us in completing this project.

References

[1] Li, L., Prussella, P.G.R.N.I., Gunathilake, M.D.E.K., Munasinghe, D.S. and Karadana, C.A., 2015. Land Valuation Systems using GIS Technology Case of Matara Urban Council Area, Sri Lanka. Bhumi, The Planning Research Journal, 4(2), pp.7–16.

[2] Nadai, M. D., & Lepri, B. (2018). The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). doi:10.1109/dsaa.2018.00043

[3] Schulz, R. (2003). Valuation of properties and economic models of real estate markets. Erscheinungsort nicht ermittelbar: Verlag nicht ermittelbar.

[4] Vaz, J. (2015). REAL ESTATE APPRAISAL AND SUBJECTIVITY. European Scientific Journal March 2015, ISSN: 1857 – 7881(e - ISSN 1857- 7431), pp.55, 63.

[5] De Andrado, M. (2018). Aiming for a Smarter Future With the AI Asia Summit 2018 – README. [online] README. Available at: https://www.readme.lk/slasscom-ai-asia-summit-2018-post-event/ [Accessed 20 Feb. 2019].

[6] Karunananda, A., Asanka, P., Fernando, H., Adhikari, T. and Pathirage, I. (2014). State of Artificial Intelligence in Sri Lankan Software Industry. [online] Available at: https://www.researchgate.net/publication/281224224_State_of_Artificial_Intelligence_in_Sri_Lankan_Software_Industry [Accessed 17 Feb. 2019].

[7] A. J. P. Samarawickrama and T. G. I. Fernando, "A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market," 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), Peradeniya, 2017, pp. 1-6.

[8] Chandrasekara, Vasana & Tilakaratne, Chandima. (2011). Comparison of Support Vector Regression and Artificial Neural Network Models to Forecast daily Colombo Stock Exchange.

[9] Chaphalkar, N.B, & Sayali Sandbhor. (n.d.). Use of Artificial Intelligence in Real Property Valuation. Retrieved from http://www.enggjournals.com/ijet/docs/IJET13-05-03-087.pdf

[10] Zillow. (2019, February 21). Retrieved from https://en.wikipedia.org/wiki/Zillow#Zestimate [Accessed 23 Feb. 2019].

[11] Hagerty, James R. "How Good Are Zillow's Estimates?", The Wall Street Journal, 2007-02-14. Retrieved on 2009-02-25.[Accessed 23 Feb. 2019].

[12] “QV Homeguide App Now Available.” New Zealand Property Investors Federation, 3 Mar. 2015, www.nzpif.org.nz/news/view/56971. [Accessed 24 Feb. 2019].

[13] “ABOUT US.” Houseprice.AI-What's the Fair Price ?, www.houseprice.ai/about. [Accessed 24 Feb. 2019].

[14] “Introducing Houseprice.AI: The Must Have Tool for Every Developer.” Bridging Loans | Development Loans | AvamoreCapital, 29 May 2018, avamorecapital.com/introducing-houseprice-ai-the-must-have-tool-for-every-developer/. [Accessed 24 Feb. 2019].

[15] J. Jackson, “Multivariate Techniques: Advantages and Disadvantages,” The Classroom | Empowering Students in Their College Journey, 10-Jan-2019. [Online]. Available: https://www.theclassroom.com/multivariate-techniques-advantages-disadvantages-8247893.html. [Accessed: 04-Aug-2019].

[16] A. Singh, “A Practical Introduction to K-Nearest Neighbor for Regression,” Analytics Vidhya, 07-May-2019. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/. [Accessed: 04-Aug-2019].

[17] “How to Create an ARIMA Model for Time Series Forecasting in Python,” Machine Learning Mastery, 26-Apr-2019. [Online]. Available: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python. [Accessed: 04-Aug-2019]

[18] A. Wasantha, K. Weerakoon and N. Wickramaarachchi, "Rating Valuation Model for Residential Properties in Sri Lanka: Case Study in Homagama", Sri Lankan Journal of Real Estate Department of Estate Management and Valuation, no. 06, pp. 61 - 76, 2010.