Housing Price Prediction using Machine Learning

LKTG Liyanaarachchi#, IA Wijethunga and MKP Madushanka

Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka

#senethd@icloud.com

Abstract - As housing price increases annually, offering unusual prices for houses that are not worth so much is a current problem faced by those who plan to buy a house. Moreover, most property investors also mislead by using fake facts without knowing the trend of houses for a certain location. So, the proposed system allows to evaluate the performance and the predictiveness of a model that supervises collected data from a certain area. The system is stricter on providing accurate values for the houses than the existing systems. This project expects to build a good mutual understanding between buyer and seller. It will endeavor to give the best rates among different calculations when utilizing the public dataset in preparing real-world. The project shows the factors that are affecting Housing Price Prediction on real-world. Furthermore, the observational outcomes show that crime rates, store rates, and public spots impact the house costs contrarily, whereas expansion, year, and joblessness rate sway the house costs emphatically. Overall, in the modern world, with the rapid development of technology and digitalization, a software like this is really required to defend from sellers, who deceive customers verbally and physically by showing inaccurate prices for properties which are not worth that much. It will help investors to achieve their economic ideas without any doubt too

Keywords: prediction, machine learning, housing price, regression

I. INTRODUCTION

Housing price prediction or estimation is a major trending project related to the field of Machine learning. Now days, most of the people are used to buy a house rather than building a one. So, eventually there are lot of housing projects going around the world. Investors are happy to invest on those projects as it is more profitable than other business. There are lot of telecasting ads, website notices, radio announcements about those housing projects. But when buying a house, the buyer has to know whether that price is fair enough or not according to the features of that house. If then, there will be a good understanding between the buyer and seller. Otherwise, people get caught for the houses that are not worth for that much. By introducing a prediction system, people can get use to know about the house and rates before buying a house. They can compare the prices of the seller and the system to buy the most suitable house according to the money which they have.

In Sri lanka currently housing price is predicted using a manual process of Valuation. But mainly it takes around 2-3 weeks for the entire process and result will be subjected to various fluctuations due to many reasons. The main reason is that the government rates are not congruent with the original rates of market. With the rapid growth of housing schemas and housing projects, availability of Housing price prediction system is essential for Sri Lanka.

II. A STUDY ON PRICE PREDICTION AND RELATED WORKS

As machine learning is an advanced topic crucial review about the related works will done in this chapter. The problems in the past systems, the solutions they used, technologies they used and the difficulties they faced when creating a housing price will be clarified. The Case of Melbourne City, Australia by Danh Phan is a major related work for the Housing price prediction. It mainly focused on several features related to the city Melbourne is Australia. Land size, property count, longitude, latitude, year are some of them. On this study, they have used histograms as a descriptive model to show how the prices are changed with each feature. In this

work they have prepared data before using by data reduction and exploration. It is said that it will increase the performance and accuracy. Machine learning algorithms which analyzed is that the Regression, Polynomial Regression, Regression Tree, stepwise, Neural Networks and SVM. By all of them the combination of stepwise and SVM which produces the lowest error on this dataset, is the most competitive models. Before using data, they have arranged according to the importance of them.

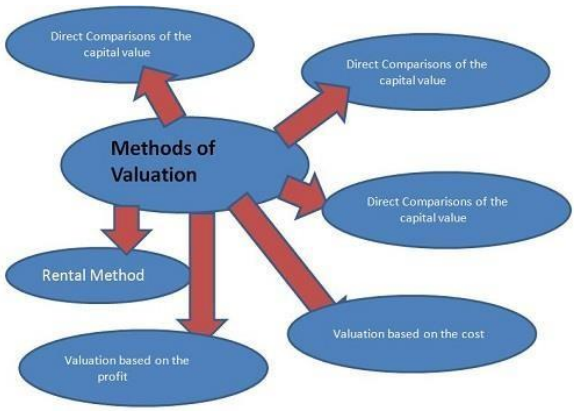

Figure 1: Valuation

For a country like Sri Lanka, it easier to go through with valuation theme because the things like housing price index will not updated regularly in here. Though the valuation process seems like very easy, there is a challenging part also. When once a valuer valued the property with a price with time various new development projects like highway, water projects. Sanitary projects etc will exist and then the price should be change.

Mainly it is consisted with neural networks. Neural network is an advanced technique used in machine learning. This research work is more focused on model training. In here, artificial neural network based on memristor is created to work on several variable regression with back propagation method. Memristor based means including any kind of non- volatile memory. By using artificial neural network, it can learn about the models and predict the values which are closer to original ones. As an example, in this study they have used ANN to learn about linear regression model of some houses in US, and they have predict the housing prices to be close to the real data. In here they have used, Mean Square error which tells about the closeness of linear regression line to a set of data points. Linear regression line is a linear answer to a set of scaler identities with one or more variable.

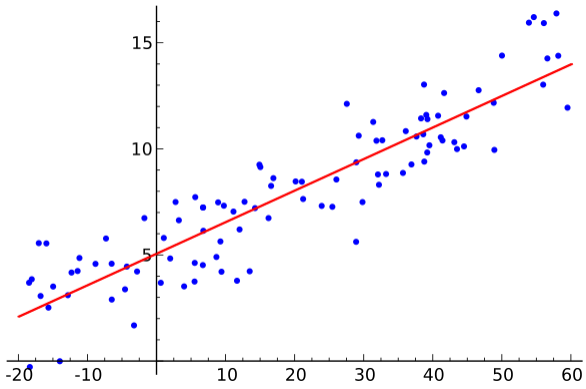

Figure 2: Linear Regression

Mean square error is calculated by taking the distance of each point to the regression line and squaring them. It is calculated as below according to an equation.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Yi\text{-}Y'i) \wedge 2$$

Table 1: Variables for equation

MSE	=	Mean squared error
n	=	Number of data points
Yi	=	Observed values
Y'i	=	Predicted values

To calculate mean squared error;

1. Find the regression line.
2. Insert your X values into the linear regression equation to find the new Y values (Y').
3. Subtract the new Y value from the original to get the error.
4. Square the errors.

By taking mean square error they had used to check the performance of the network. They had said that to decrease the MSE you have to increase the training data. When predicted prices are reached from training and testing data they must close to the target values.
Otherwise, there is a huge problem in the accuracy.

An important study which has done with minimum errors and same as the machine learning project which is suitable to Sri lanka. Mainly, in this one features are according to the importance of those green buildings. Unlikely in other projects green buildings focus on the features like energy, indoor environment status, site planning, water, and resources etc. In here, they have used five machine learning techniques namely linear regression, decision trees, random forest, ridge, and lasso.

Features	Description
Transaction Price	Dispose price/sqf (RM)
Date of Transaction	Building Transaction/Months
Lot Size	Lot Size
MFA	Main Floor Area
Tenure	Freehold/Leasehold
Type of Property	Residential/Commercial
No of bedroom	Number of bedrooms
Level Property	Level Property Unit
Floor	Building Floor
Building Facade	City/Park/Lake/Klcc
Age of Building	Age of Building
Distance	Distance to Central Business District
Accessibility	Ease of accessibility
Mukim	Mukim
Certificate	Green Certificate/Non-Green Building
Density	Population Density
Security	Security of Building
Infrastructure	Infrastructure Development

Figure 3: Features

Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyze current and historical facts to make predictions about future or otherwise unknown events.

Machine learning techniques have the potential to unearth patterns and insights we did not see before, and these can be used to make unerringly accurate predictions.

Another important tool that was used in system is that the co-relation. Co-relation is the strength and the direction of linear relationship between two variables. It is very useful to get the relationship of string variables in machine learning. If the association is (+) value then the relationship is positive and if the association is (-) , then the relationship is negative.

```
correlations = training.corr()
correlations = correlations["TransactionPrice"].sort_values(ascending=False)
features = correlations.index[1:6]
correlations
```

```
TransactionPrice              1.000000
Main Floor Area               0.715018
Lot Area                      0.714555
Population Density            0.452463
Mukim                         0.365324
No of Bedroom                 0.339879
Tenure                        0.071255
Building Floor                0.067840
Green Certificate             0.023308
Level Property Unit          -0.059356
Building Façade              -0.070940
Date of Transaction          -0.136980
Distance                     -0.182399
Age of Building              -0.209287
Type of Property             -0.273990
Security of Building         -0.328196
Acessibility                       NaN
Infrastructure Development         NaN
Name: TransactionPrice, dtype: float64
```

Figure 4: Co-Relation

It is hardly pushed because to solve some problems in SVM. Selection of decision boundaries of parameters, when matrix scale subjected by the training data, and also when finding solutions for quadratic programming issues LSSVM is very useful. Though this is not a very popular algorithm which use for machine learning it is highly recommended for the prediction functions like in this work. LSSVM regression model can be described as follows.

$$Y = z \times \phi(x) + 1$$

ϕ(x) shows the non-linear mapping function and z is the weight factor. *l* is coefficient of invariable.

The last algorithm that they have used is that the Partial least square regression algorithm. It is a special algorithm used for statistical data analysis. In a prediction system it is widely used to find the best function which maps the original data which reduce the sum of squares error. When number of parameters are greater than number of data points PLS is used to construct the model. The mathematical equation of the PLS algorithm can be defined as below.

Where, *x* is n*m of predictors, y is n*p responses T and U are n*l matrices. P & Q are projections of x and y, respectively.

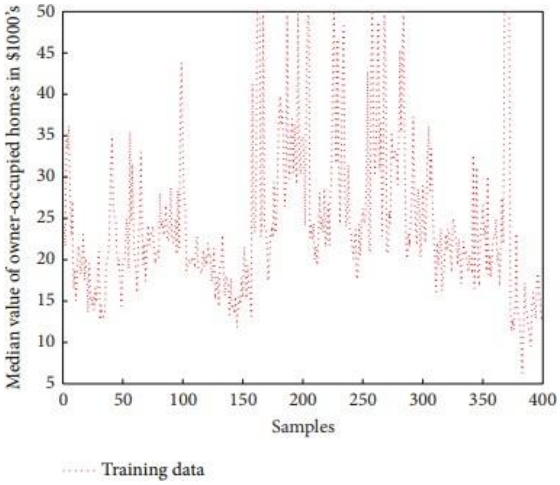

Figure 5: Training Data

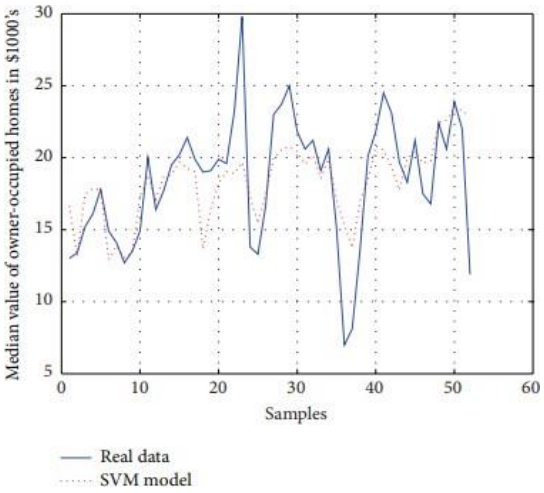

Figure 6: SVM

Study related to the housing price prediction. In this work, it is said that future works can be enabled by the past data. Prediction was reached by a sample data set which consisted of the attributes related to cost prediction. A dataset of real estate agents in US was used to evaluate the cost of houses. The main features that they have used are size of plot, number of bedrooms, number of bathrooms, number of floors, included gas and water facility etc. Data analysis has been done using exploratory approach to minimize data anomalies and to discover patterns.

Mainly their system based on the machine learning algorithm which is called regression. Under the regression their focus had been to linear regression which is based on the equation y=mx+c. Training of the model has been done by understanding the slopes and positions of all the data points with the line.

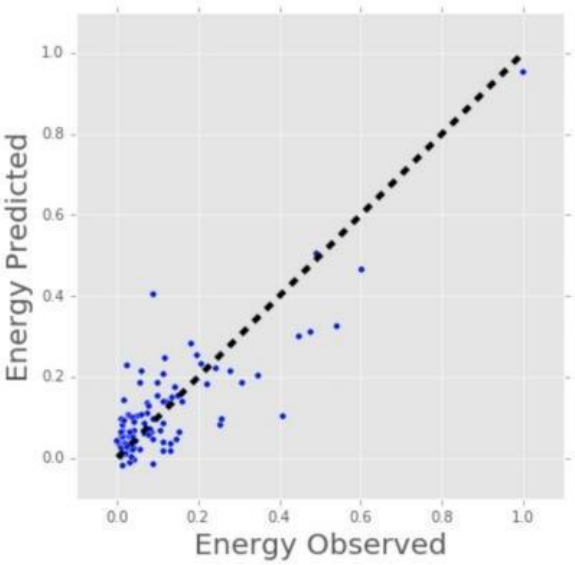

Figure 7: Clustering

The specialty in this work is that it predicts whether a customer will purchase the house or not status also. For that they have used logistic regression as their prediction algorithm and based on single and multiple inputs it predicts the occurrence of the event. Another very important tool that was used was Clustering. It defined that whether the output uses the algorithm instead of using output data for training. Visualization is used to monitor the clustering effects.

By using a user-friendly interface, the work has been able achieve 95% accuracy with the defined variables. Following shows the clear user interface that will display to the client.

III. DESIGN AND IMPLEMENTATION OF THE METHODOLOGY

Data flow diagram and sequence diagrams will be used to describe the design of the project graphically. The design architecture of the system with its individual units and their relationship with relevant to each other is described here. Requirement analysis, need of the project, the domain, and physical design of the project with its challenges will addressed by the design diagrams.

A.Requirement Analysis

Requirement analysis is a major part of a software project. It can change the path of the project towards success or towards the fail. In the

research papers most of the systems fail due to incorrect method of requirement gathering or in the 1st phase of the software development life cycle. As this is a housing price prediction project data related to housing prices with correct feature details should be acquired. In order to get the data related to the Housing prices should contact housing sales agencies because government rates are subjected to various negotiations, and they are diverted from the real-world values. By collecting data accurately, it provides 50% of accuracy because the model, website or price prediction app, everything depends on those requirements.

Projecting on requirement analysis, the 1st step of the analysis was Data collection. Data collection is a process of gathering data on depending on variables related to a project. So, in the case of housing price prediction project, the data collection is process of collecting information on number of bedrooms, number of washrooms, number of perches, number of floors etc in related to the price of that house.

There are nine variables and price of the house in the data collection process related to the housing price project. When selecting those variables, focusing whether they are dependable on housing price is a major challenge in this project. But it was solved by contacting several housing sale companies. Not only solved, but they also helped to give the data needed on those variables related to the housing price. As this is a machine learning project which relies accuracy on the amount of data, 1000 data targets were achieved with the help of two companies.

B. Sequence Diagram

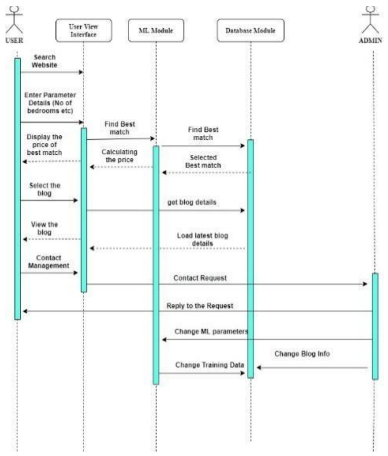

Figure 8:Sequence Diagram

The arrangements of the process when dealing with housing price prediction system is defined by sequence diagram.

C. Data-flow Diagram

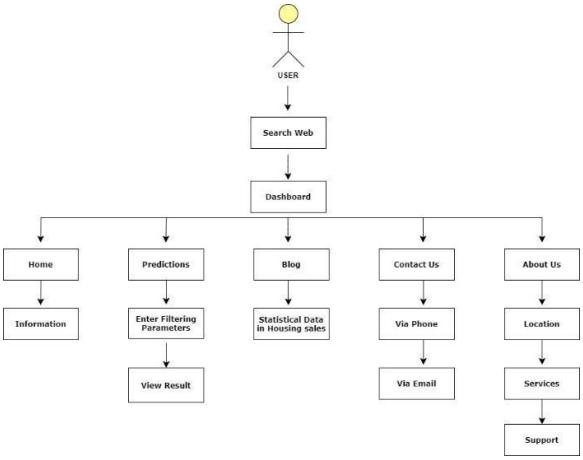

Figure 9: Data-Flow Diagram

Following figure shows the data flow design of the user's website for the housing price prediction system.

D. Need of the project

Offering unusual prices for the houses that are not worth so much, fake details which are provided to the investors to lose their business path, too much bargaining the prices of the sellers related to the housing sector leads to promote a system of predicting values of the houses before buying or selling them. But in Sri lanka, as most of the systems are manually ongoing, the prediction also done in manual ways. Though a predicting process is available, the results of those systems are subjected to lot of errors. Those systems are not stricter to provide actual market results which outcomes bad relationships between buyers and sellers.

With the rapid development of technology in the world, other countries are emerged with automated systems and results. But in Sri lanka, there are not any system to predict the prices of the houses with the automated technology. This housing price prediction system is targeting to produce accurate results of Sri Lankan market by approaching the Digitalization system and to defend sellers, buyers, and investors to achieve their goals related to the housing sector.

E. Domain

The domain or the subjected area of a software project is the area which the project will hit on. So, in this case the domain will be the Housing market. This domain is special because the domain decreasing per year is 0 % because the amount of houses per year is just only increasing but not decreasing. With the rapid growth of population in Sri lanka, the domain of housing market will not decrease in the future too. So, building an automated system to predict the prices of houses in Colombo area will be an emerging software idea in Sri lanka.

F. Physical Design

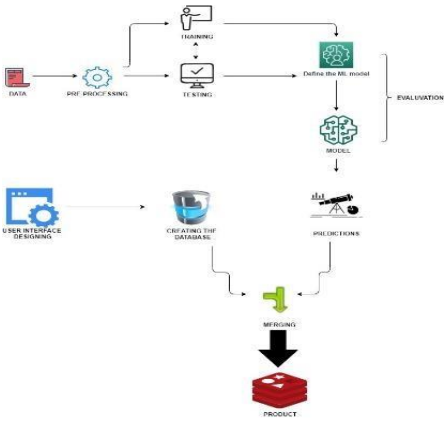

Figure 10: Physical Design

IV. ORIGINALITY AND INNOVATIVENESS`

In Sri Lanka, there are not any software to measure the price of a house before buying it. Over the last few years, the theme of housing price prediction has been highly discussed area in Sri Lankan economy. Not only in economy, but also in academical view it has a major trend because it carries huge employment opportunities. A simple mislead of housing market affects hugely to many social and economic stats of Sri lanka. In Sri lanka, the characteristic features which leads to housing price prediction has changed over the time.

But, in manual process it takes long time to insert those changes for the prediction. Due to those factors the accuracy of those results is subjected to discussion. So, a software system for the housing price prediction will make an interest all over the country and if government could involve in this it will be a massive effort in all aspects as it carries major role in Sri Lankan economy.

Though in Sri lanka we can find house sales agents or websites, there are not any system to predict the prices of those houses by using an automation method. There are only investment analysis agents, and they are also providing only the information related to investments. So, a system like this will focus hugely on Lankan people and also it will help for the investors to build up new housing projects to rise economy of Sri Lanka.

Also, by using a Machine learning approach it is easy to identify the patterns in the housing field and how those patterns change with price. Support vector machines (SVM), linear regression, logistic regression, K-Nearest neighbour, Random Forest are some of the examples for Machine learning techniques. To get the maximum accuracy, have to feed more data into the algorithms as training and test data. Machine learning algorithms use those data to train the model to predict the values of housing price.

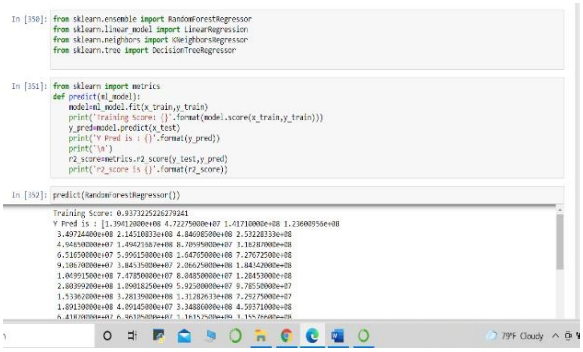

Figure 11: ML Code

V. CONCLUSION AND FURTHER WORKS

This system is mainly based on machine learning using python language. Primarily, we clean the all the Data (Variables) and will group them into necessary fields. And then we can classify the parameters according to the importance of them in our system according to pricing value of them.

Nearby local amenities such as railway station, supermarket, school, hospital, temple, parks etc will be our unique approach in addition to our variables. By using Google maps, we can analyze those local amenities near 1km of radius - circle around the main town. And we can increase the price of those Houses in related to each other when we find such amenities in our circle. In the

system, it will use the Linear Regression, Forest regression and Boosted Regression as the algorithms and those results will inputted to a neural network to compare the predictions and to get the accurate results. And finally, the most accurate result will be displayed on the system.

REFERENCES

Belov, A. (2019) 'Tallinn House Price Prediction'.

Burgt, E. van der (2017) 'Data Engineering for house price prediction Master ' s Thesis', p. 75.

Ghosalkar, N. N. and Dhage, S. N. (2018) 'Real Estate Value Prediction Using Linear Regression', *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018,* pp. 1–5. doi: 10.1109/ICCUBEA.2018.8697639.

Jamil, S. et al. (2020) 'Machine Learning Price Prediction on Green Building Prices', *2020 IEEE Symposium on Industrial Electronics and Applications, ISIEA 2020,* pp. 0–5. doi: 10.1109/ISIEA49364.2020.9188114.

Malik, N. (no date) 'Employing Machine Learning for House Price Prediction'.

Mu, J., Wu, F. and Zhang, A. (2014) 'Housing Value Forecasting Based on Machine Learning Methods', *Abstract and Applied Analysis*, 2014. doi: 10.1155/2014/648047.

Phan, T. D. (2019) 'Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia', Proceedings - *International Conference on Machine Learning and Data Engineering*, iCMLDE 2018, pp. 8–13. doi: 10.1109/iCMLDE.2018.00017.

Satish, G. N. et al. (2019) 'House Price Prediction Using Machine Learning', *International Journal of Innovative Technology and Exploring Engineering*, 8(9), pp. 717–722. doi: 10.35940/ijitee.i7849.078919.

Varma, A. et al. (2018) 'House Price Prediction Using Machine Learning and Neural Networks', *Proceedings of the International Conference on Inventive Communication and Computational Technologies*, ICICCT 2018, pp. 1936–1939. doi: 10.1109/ICICCT.2018.8473231.

Wang, J. J. et al. (2018) 'Predicting House Price with a Memristor-Based Artificial Neural Network', *IEEE Access*, 6(c), pp. 16523–16528. doi: 10.1109/ACCESS.2018.2814065.

AUTHOR BIOGRAPHIES

T. G. Liyanarachchi is a 4th year undergraduate student of the Department of Computer Science at Genaral Sir John Kotelawala Defence University. His main goal of this research is to predict the housing prices to help the people who suffered difficulties in whenever they are going to buy or sell their houses.

Mrs. I.A. Wijethunga received her Bachelor of Science (Special) in Computer Science form University of Kelaniya and Master of Computer Science from University of Colombo School of Computing. Currently, she is working as Lecturer (Probationary) at University of Moratuwa and formerly at General Sir John Kotelawala Defence University. Her research interests include Computer Vision, Cloud Computing and Machine Learning.

Mrs.MKP. Madushanka is a Lecturer Probationary) at Department of Computer Science and Reading for the MPhil in Computing at General Sir John Kotelawala Defence University Sri Lanka. Her obtained Bsc (Special) in Computer Science from South Eastern University of Sri Lanka and Master of Computer Science from University of Peradeniya. Her research intrest include the Machine Learning, Big Data analysis and Image Processing.