E-Valuer

(An Intelligent tool to assist in making smarter property related decisions)

Mrs. M.P.A.W. Gamage

Mrs. Pasangi Rathnayake

Bimali Y.M.Y. IT 16 42 3534

Project ID: 19-010

Sri Lanka Institute of Information Technology

Bachelor of Science Special (Honors) in Information Technology Specializing in Software Engineering

September 2019

E-Valuer

AN INTELLIGENT TOOL TO ASSIST IN MAKING SMARTER PROPERTY DECISIONS

Project ID: 19-010

Final Report Submission (Individual)

(Dissertation submitted in partial fulfillment of the requirements for the degree Bachelor of Science Special (Honors) in Information Technology Specializing in Information Technology)

Bachelor of Science (Honors) in Information Technology Specialized in Software Engineering

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2019

DECLARATION

I declare that this is my own work and dissertation does not incorporate without acknowledgment of any material previously submitted for a Degree or Diploma in SLIIT or any other university or institute of higher learning. To the best of my knowledge and belief the document does not contain any material previously published or written by another person except where the acknowledgment is made in the text

Project ID : 19-010

Project Team Members:

Student Name	Registration No	Signature
Bimali Y.M.Y.	IT16423534	

The above candidates are carrying out research for the undergraduate dissertation under my supervision.

Signature of the supervisor : .....

Date : .....

Signature of the co-supervisor : .....

Date : .....

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 willing to buy. Our first priority would be given to current valuation of the land since it is difficult to find out the reliable average prices of the land around manually, especially if the area is unfamiliar. We can get the service of a professional valuer, but that process is known to be subjective to the person and time. The system utilizes ensemble model of MLR and ARIMA model as well as KNN and MLR model in making predictions. The ensemble model can make predictions with an accuracy of over 0.75 in current value prediction and also future value predictions with reasonable accuracy. Since Sri Lanka is a rapidly developing country, there are many ongoing and proposed road and infrastructural development projects. If someone needs to find out the effect of such development work on the selected land, it is very difficult to gather and find out such data based on the current weak digital infrastructure available in Sri Lanka. With E-valuer, users just need to input the location of the land then the predictions are given at the speed of lightning. Machine learning , deep learning and optimization are the main research components of this system. This system would be of great assistance to make better property decisions, which adds value to user's money spent on buying land plots, which is a massive investment as well as a very important decision in one's life.

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

ACKNOWLEDGEMENT

The work described in this research paper was carried out as our 4th year research project for the subject Comprehensive Design Analysis Project. The completed final project is the result of combining all the hard work of the group members and the encouragement, support and guidance given by many others. Therefore, it is our duty to express our gratitude to all who gave us the support to complete this major task. We are deeply indebted to our supervisor Ms. M.P.A.W. Gamage and our co-supervisor Ms. Pasangi Rathnayake, lecturers of Sri Lanka Institute of Information Technology whose suggestions, constant encouragement and support in the development of this research, particularly for the many stimulating and instructive discussions. We are also extremely grateful to Mr. Jayantha Amararachchi, Senior Lecturer/ Head-SLIIT Centre for Research who gave and confirmed the permission to carry out this research and for all the encouragement and guidance given. We also wish to thank all our colleagues and friends for all their help, support, interest and valuable advices. Finally, we would like to thank all others whose names are not listed particularly but have given their support in many ways and encouraged us to make this a success.

Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 Background Literature

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 this real property 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 a number of 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 follows.

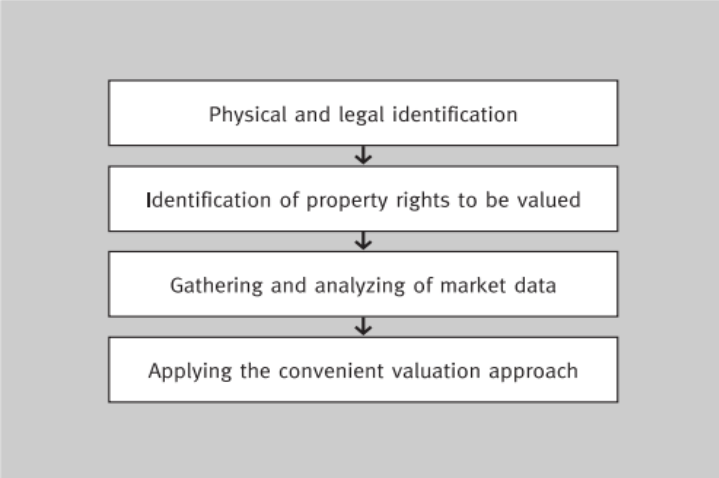


Figure 1.1: Different stages of the appraisal process for estimating the market value

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

The major convenient valuation approaches are,

1. Sales Comparison Approach
2. Income Approach
3. 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 neighbourhood economical value is combined with effect of neighbourhood factors such as walkability etc. we believe it is possible to give an 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.

Our training environment determines the best out of conventional Multivariate Regression Analysis (MRA), Random Forest Regressor and non-conventional Artificial Neural Networks (ANN) methods to do the most accurate prediction. According to Sampathkumar, et al. [4], both the models are found to be well fit with the data set of the land price in all locations, the model using NN (correlation 98%) shows better accuracy than the regression model (correlation 96%), while Zurada, Levitan and Guan, 2011 [5] concludes no single obvious non-conventional method that can be expected to consistently outperform traditional multivariate linear regression in predicting residential real estate sales prices. In the least, the non-conventional methods may be used as a complement to the traditional, multiple regression-based methods [5].

Chaphalkar et al [6], have compared use of ANN, use of fuzzy logic, use of expert system and genetic algorithm and other techniques such as decision tree by previous literature aggregation and suggest that ANN performs better than MRA but its black box nature has led to reach different conclusions in the observations.

Existing solutions

The use of AI for residential value forecasting has been suggested in the literature from 1990s [6]. 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. Our intention is to identify the ways to use their underlying methodology in a suitable manner in Sri Lankan context.

1.Zillow Zestimate

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. [7] Zillow.com supports United States of America (USA) and Canadian property listing. Zillow compliments that Zestimate provides forecast for 12 months with below accuracy rates.

Model	Average Absolute % Error	Improvement over Naïve
Naïve Forecast	7.35%	0%
County Forecast	6.47%	11.9%
Zestimate Forecast	5.84%	20.5%

Table 1.1 Zillow Zestimate prediction evaluations

Features :

- Estimates for 12 months

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 [8] which analyses of several features of the house. The forecasted value is interpolated using cubic spline to connect to current value. [8]

2.Trulia

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.

Below is the accuracy report of Trulia estimates.

National	Within 5% of Sale Price	Within 10% of Sale Price	Within 20% of Sale Price	Median Error
United States	48.2%	67.7%	82.3%	5.3%

Table 1.2 Trulia Accuracy report

Features -

- Crime map - Crime map data is sourced from CrimeReports.com and SpotCrime.com, which aggregate crime data from law enforcement agencies and news reports.
- Local schools with schools rating - Data of the schools around the premises with details such as Grades taught, GreatSchool Score.
- Commute times at a glance - Using data from OpenStreetMaps and General Transit Feed Specification (GTFS) feeds, the user can get an idea of commute times at a glance.[9]

3.QV.co.nz - QV homeguide

Quotable Value (QV) provides independent and authoritative information on any home in New Zealand on or off the market [10] 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.

Features - QV homeguide app

- Online Value Estimation - Provides the likely selling price of a property during that particular time
- Sales activity - Sales activity specific property found on the app
- Suburb Demographics - Median price data, Demographic data, Current listings, and latest auction results [11]
- E-Valuer Report - Subjected to a fee complete valuation report of the property can be downloaded.

4.HousePrice.ai

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. [10]

Features-

- Current and Future value prediction - 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 [12]. This tool allows the user to adjust, add and remove factors within the surrounding areas to determine how external changes will affect property prices

- Distance to Schools, commutes etc.

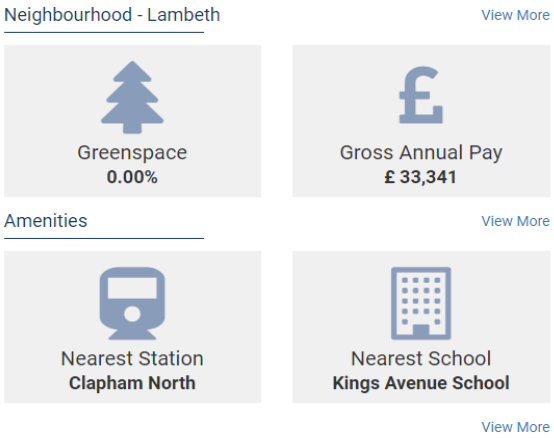


Figure 1.2: Brief Neighbourhood analysis

Source : *Sample Valuation Report - HousePrice.ai , Horizon*

https://myhorizon.io/valueReport?id=59ddcdc7a699d278745b81e1

- Historical data relevant to location

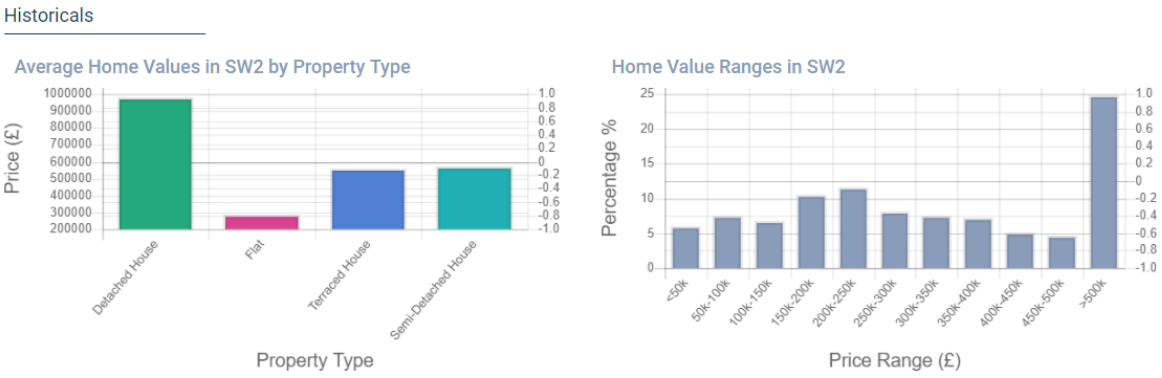


Figure 1.3: Historical Sales analysis

Source : *Sample Valuation Report - HousePrice.ai , Horizon*

https://myhorizon.io/valueReport?id=59ddcdc7a699d278745b81e1

Comparison of Existing Systems

	Zillow	Trulia	QV-CoreLogic	HousePrice.ai	Our Product
Current Value Prediction	Yes	Yes	Yes	Yes	Yes
Provision of possible natural hazards of the area	No	Yes	No	Yes	Yes
Use of machine learning algorithm	Yes	Unknown	No	Yes	Yes
Data used for predictions	Statistical data	Publicaly available data	Statistical data	Publicaly available data	Data provided by Valuers, and relevant departments
Mobile / Desktop / Web Application	Web	Web	Web	Web	Web
Available for Sri Lanka	No	No	No	No	Yes

Table 1.3: Comparison of existing systems

1.2 Research Gap

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[14], despite the fact software industry is vastly growing area. According to the survey conducted under research done by Karunanda *et al*[15], carried out in 2014, this is due to the lack of popularity, knowledge, experts, requirements and sponsorship for the AI related software projects[15].

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 Machine Learning approaches as well as Deep Learning approaches, tilted *A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market*[16], and *Comparison of Support Vector Regression and Artificial Neural Network Models to Forecast daily Colombo Stock Exchange*[17].According Li *et al,* [18]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 or machine learning in real estate value prediction in Sri Lankan context.

There are numerous factors which affect the value of a land such as physical factors, economic factors, and social factors. The details of the physical factors and economic factors along with recent valuation details will be considered in this application when deciding the terminal output, prediction of the current value. In this application, the effect of social factors will not be taken into consideration as much as above mentioned because of the unavailability of proper information infrastructure to analyze the same factors. But that will not affect the accuracy of the output since recent valuations are taken into consideration.

Since machine learning and deep learning algorithms which were proven to be suitable for real estate valuation will be used accuracy of the predictions can be guaranteed.

1.3 Research Problem

The main research problem is to develop an automated system to evaluate the land based on its neighbourhood 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.[19], 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.

1.4 Objectives

The goal is to assist people by providing them with accurate valuation, facts about how the land is going to be affected by various means of development projects, ultimately to decide whether it would be useful for their expected purpose.

Main objective

The main objective of our research is to develop a portable application which can provide instant report of a selected land parcel which can provide the users with an insight of the land with current value and future value.

Specific objectives

- Identifying the most accurate cross-sectional algorithm from conventional Multiple Regression Analysis (MRA) and non-conventional Artificial Neural Networks (ANN) in the domain of providing values in the domain of current value prediction following the Sales Comparison Approach.

- Identifying the most accurate time-series algorithm from Long Short Term Memory (LSTM), Recurrent Neural Network (RNN) and Auto Regressive Integrated Moving Average (ARIMA) model in the domain of providing values in the domain of current value prediction following the Sales Comparison Approach

- Identifying method to predict future value based on the fluctuation rates and records of weather conditions.

- Identifying the effect of proposed development plans on the future price of the selected land plot

- Creating a concise yet complete report based on the selected land plot which can be used to assist in making smarter property related decisions.

2. RESEARCH METHODOLOGY

2.1 System overview

E-valuer is focused on providing lighting fast predictions of valuation focused on location of the land helping the user to decide the suitability of the land for their purpose. 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, ensemble model predicts the current value. Then the future value will be predicted by collaboration of two units, one which considers the fluctuation rates of past pricing values and weather effects, while the other calculates the effect of proposed development projects in the area. All these units generate a report which depicts these two types of data with relevant other information about the land in a simpler way anyone can understand.

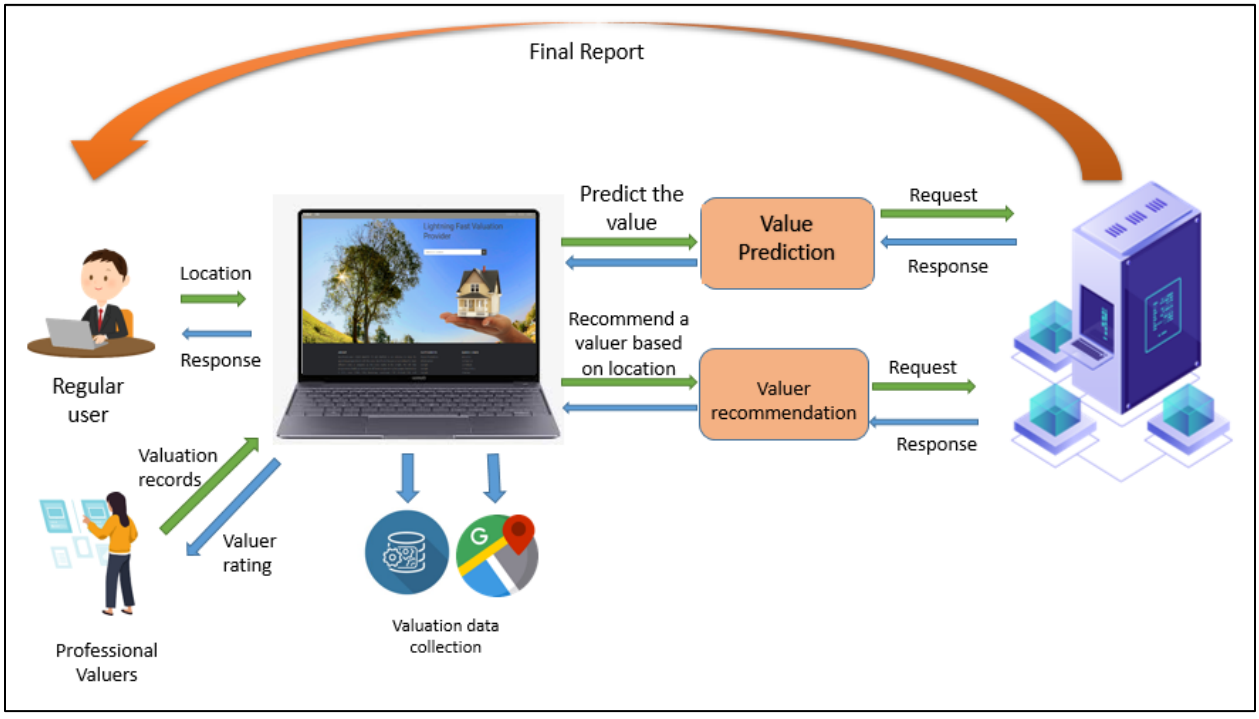


Figure 2.1. System Diagram

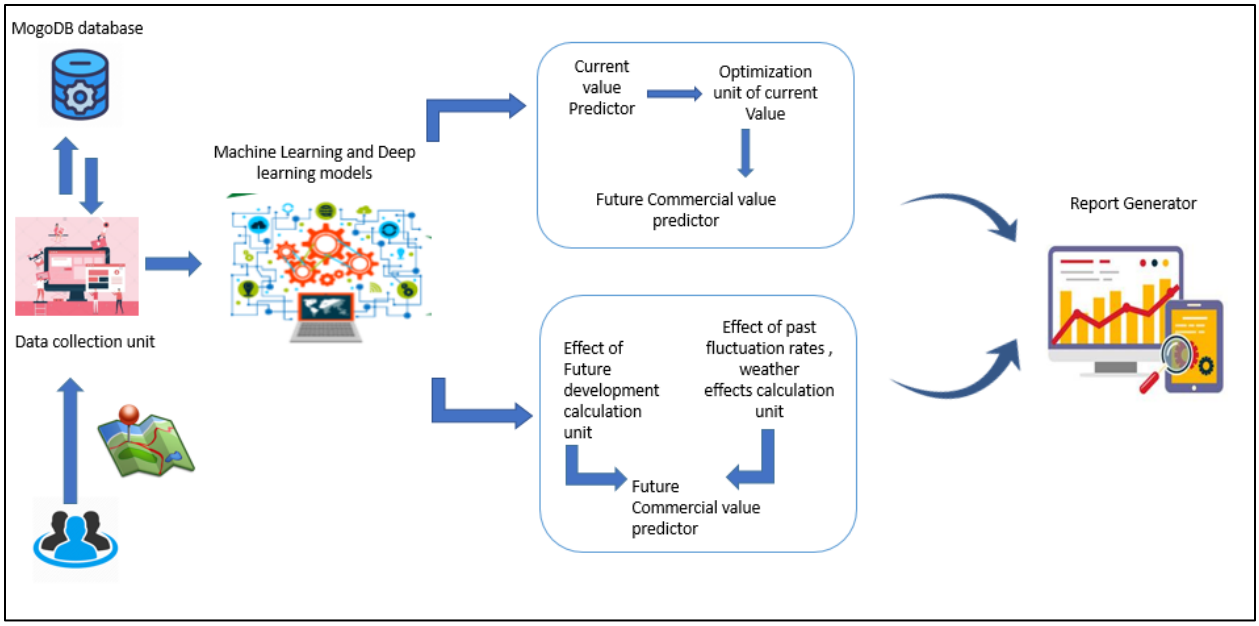


Figure 2.2 . High level system architecture

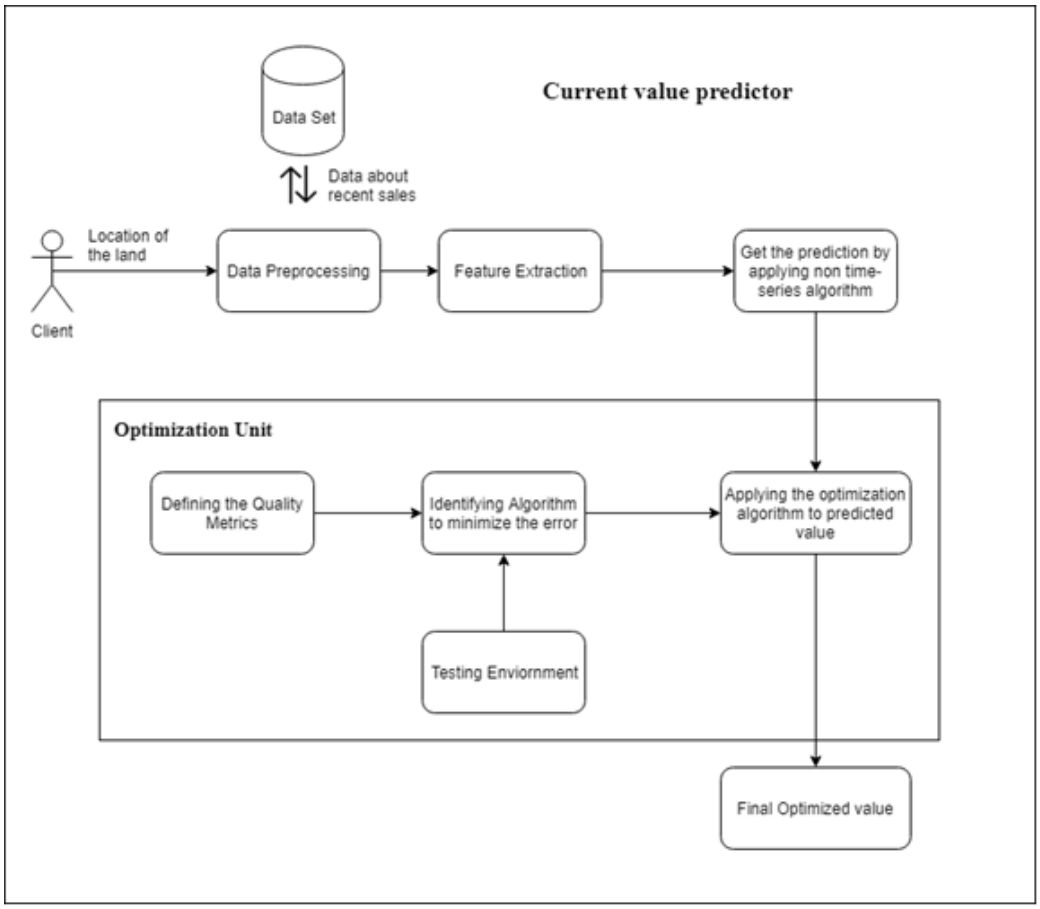


Figure 2.3 Current value predictor component design

2.2 Features

2.2.1 Finding the geo-location of the land

This feature is responsible for identifying the geo-location of the device or the location that is input by the user to get the necessary values. This feature has been implemented utilizing our Google location API. It can intelligently manage underlying location technology while meeting various development needs when implementing location-based features where it is able to pin down the user's location. Since the application is based on Colombo, geo coordinate restrictions have been made to the results displayed .

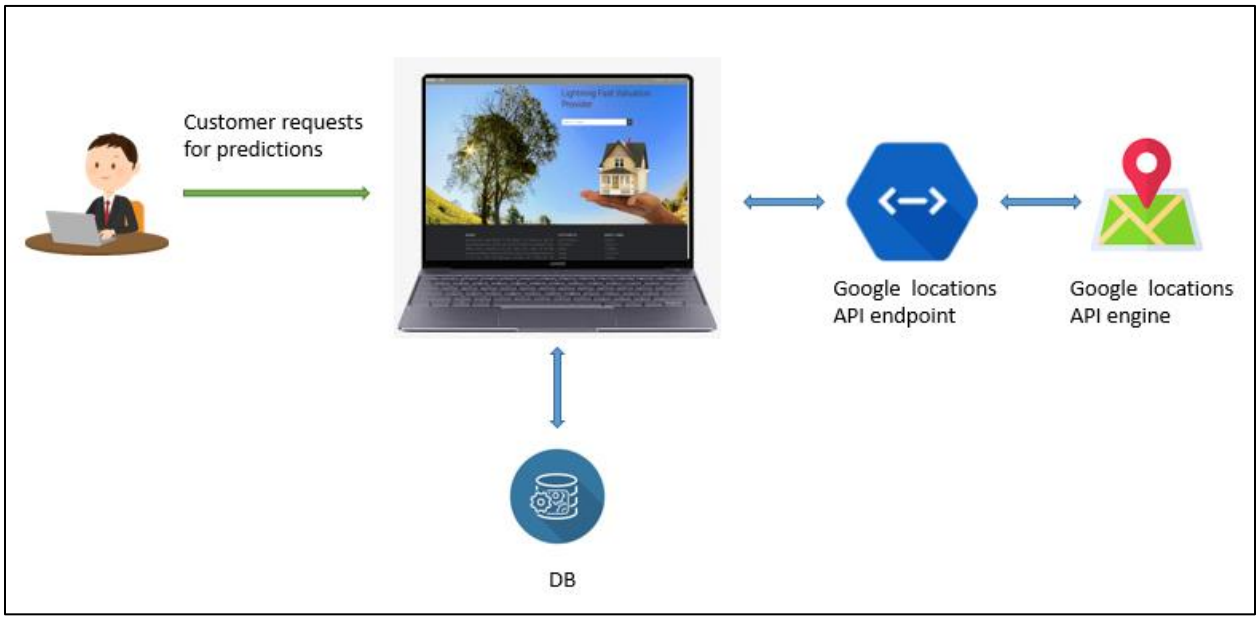


Figure 2.4 Retrieving geo-location process

Figure 2.5 User Interface

2.2.2 Price prediction for a particular location

When the user inputs the location, geo-coordinates are taken by the above feature mentioned at 2.1.1, is sent to the ensemble model of multivariate linear regression and ARIMA model make the prediction for the given location.

One of the major objectives of this project is to test the non-time series generic machine learning algorithms for their performance on land value prediction. For that, Multivariate Linear Regression (MLR) , Random Forest Regression, and Artificial Neural Network (ANN) models have been tested on a cross sectional dataset collected manually through a questionnaire and evaluated in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Hence, this component had each of the above-mentioned models trained and tested.

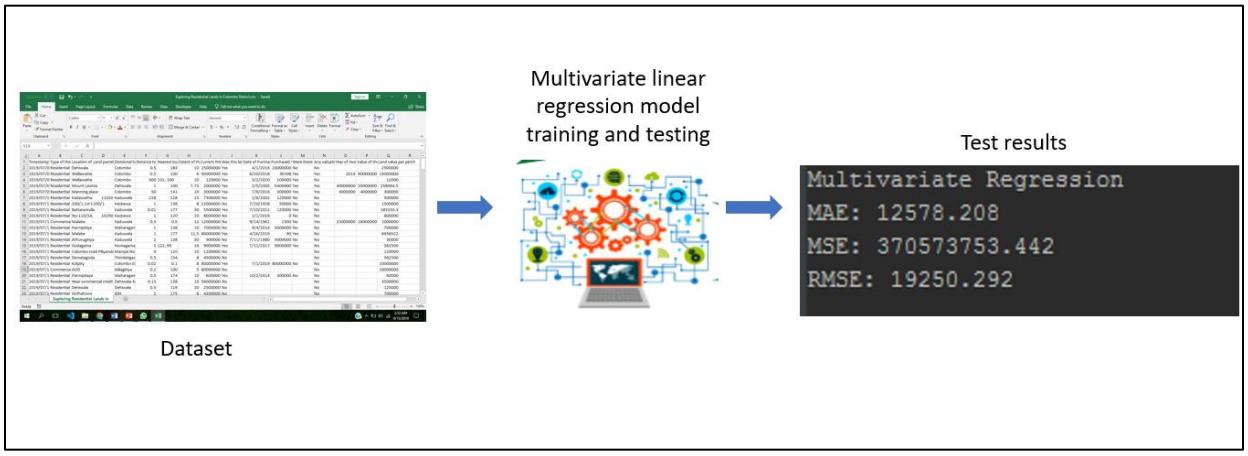


Figure 2.6 MLR training and testing

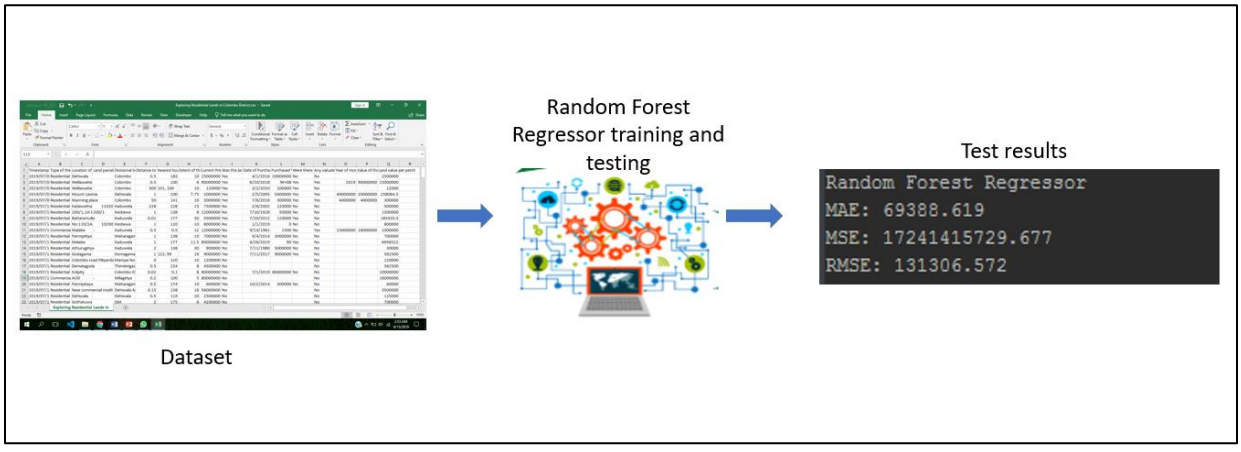


Figure 2.7 Random Forest Regressor training and testing

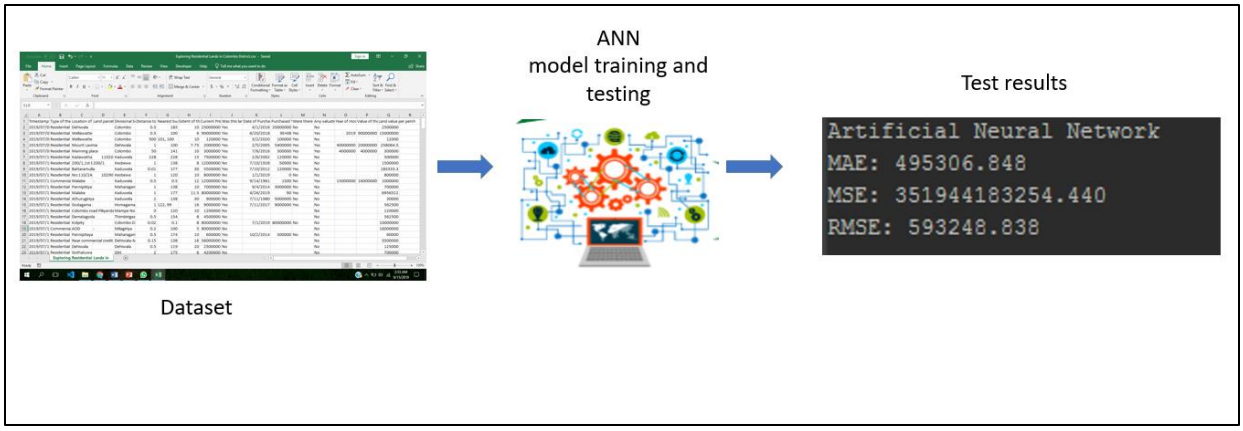


Figure 2.8 ANN model training and testing

The test results proved that the ARIMA model (time-series algorithm) performed best in predicting land prices. But since the new predictions are solely based on the location of the land, ensemble model having the results of both MLR model and ARIMA model was used in the system.

Along with the current price prediction per perch, list of valuers who added recent records of valuation jobs carried out in that area are displayed as contributions. If the users need to find a valuer who is experienced in particular area, they can easily find a suitable person. This way, the valuers get promoted too, since not only the buyer, but banks , courts , land sales companies etc who need the service of a valuer can make use of this feature to find out actively engaged valuer.

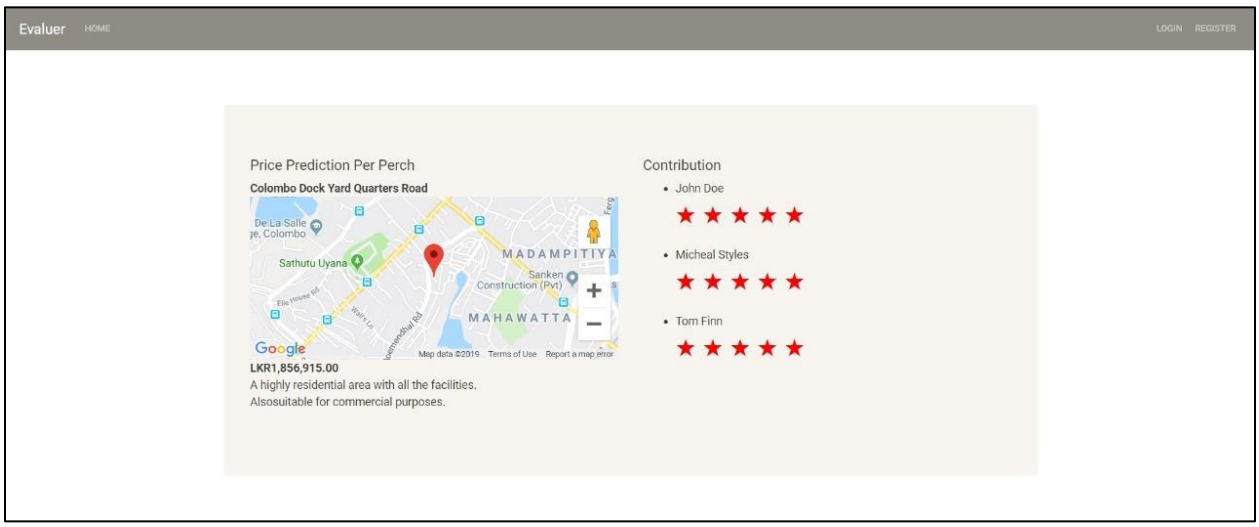


Figure 2.9 User Interface for displaying current value predictions

2.2.3 Recommendation of valuers

When the user requests for a price prediction in a certain area, along with the prediction per perch, list of valuers who added recent records of valuation jobs carried out in that area are displayed as contributors. If the users need to find a valuer who is experienced in particular area, they can easily find a suitable person. This way, the valuers get promoted too, since not only the buyer, but banks, courts , land sales companies etc who need the service of a valuer can make use of this feature to find out actively engaged valuer.

The users can pick a valuer who is experienced in carrying out valuation jobs in the given area. When recommending the valuers, those who added the valuation records most recently would be the top picks.

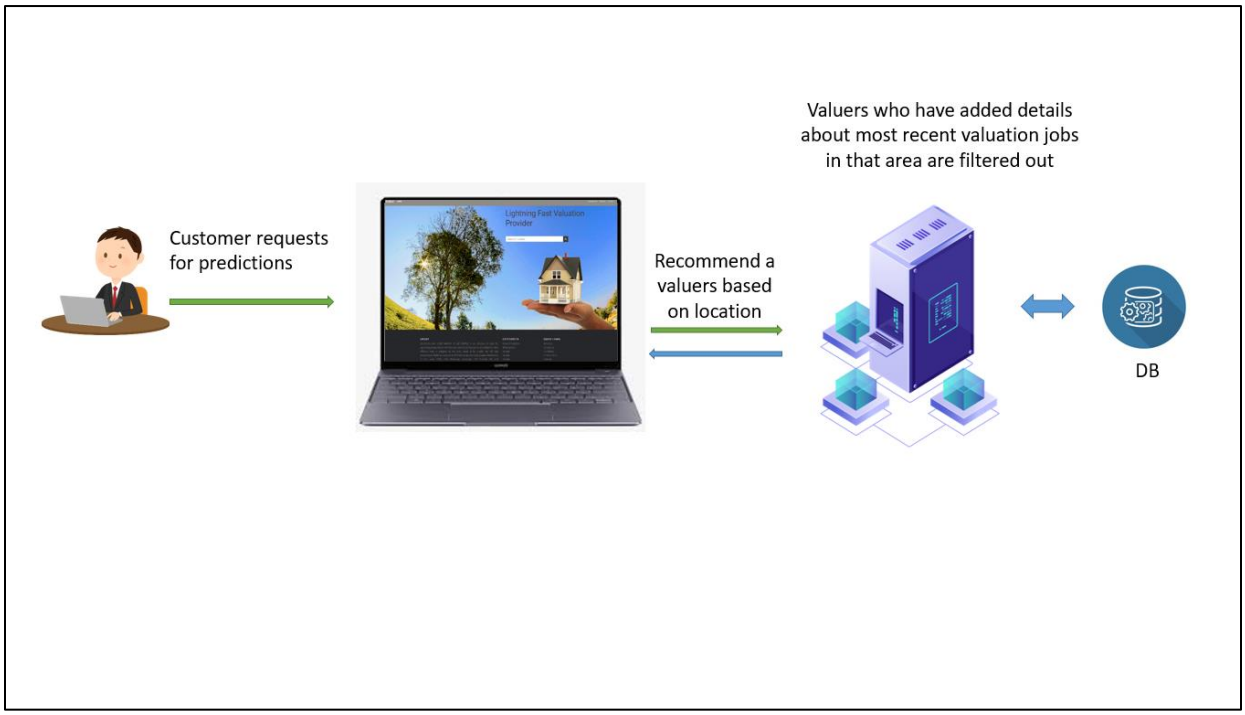


Figure 2.10 Valuer recommendation process

2.3 Methodology

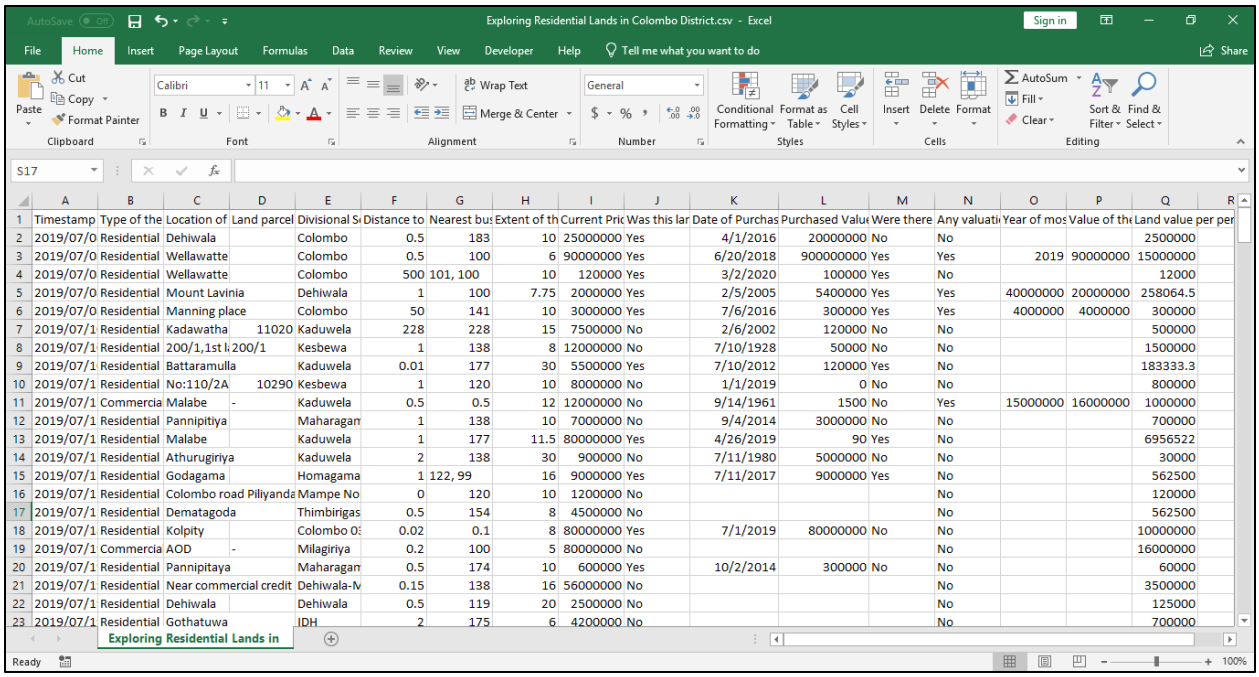
This component is responsible for developing a machine learning model which predicts the current price of a land parcel upon submission of the location based on data gathered by analyzing those submitted by valuers and other key factors identified as significant to the area, following a non time-series based algorithm. One of our specific objectives is to test the accuracy of time-series algorithms and vice versa in the domain of land value prediction. Here several non-time series algorithms were tested and evaluated in terms of Mean squared error(MSE), Mean Absolute error (MAE), and root mean squared error (RMSE) to select the best performing algorithm in the domain of current value prediction based on cross sectional data.

First phase of developing the module included selecting best algorithm to predict the current value of a land. For that, multivariate linear regression analysis, random forest regressor, and Artificial neural network model were tested. These models were selected based on the previous literature.
Next phase of the current value prediction unit is developing the API and the imputation module to estimate the missing factors necessary for the flask module in addition to location, to perform the prediction.
Finally, the most important component in terms of commercialization is developed. That is the data collection unit by the valuers.

2.3.1 Data collection

The study focuses on Colombo which experienced relatively high infrastructure development.
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 collected through an online questionnaire where residents in Colombo district responded. This sample space was expanded 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.



	Timestamp	Type of the	Location of	Land parcel	Divisional S	Distance to	Nearest bus	Extent of th	Current Pric	Was this lar	Date of Purchas	Purchased Valu	Were there	Any valuati	Year of mos	Value of the	Land value per per
2	2019/07/0	Residential	Dehiwala		Colombo	0.5	183	10	25000000	Yes	4/1/2016	20000000	No	No			2500000
3	2019/07/0	Residential	Wellawatte		Colombo	0.5	100	6	90000000	Yes	6/20/2018	900000000	Yes	Yes	2019	90000000	15000000
4	2019/07/0	Residential	Wellawatte		Colombo	500	101, 100	10	120000	Yes	3/2/2020	100000	Yes	No			12000
5	2019/07/0	Residential	Mount Lavinia		Dehiwala	1	100	7.75	2000000	Yes	2/5/2005	5400000	Yes	Yes	40000000	20000000	258064.5
6	2019/07/0	Residential	Manning place		Colombo	50	141	10	3000000	Yes	7/6/2016	300000	Yes	Yes	4000000	4000000	300000
7	2019/07/1	Residential	Kadawatha	11020	Kaduwela	228	228	15	7500000	No	2/6/2002	120000	No	No			500000
8	2019/07/1	Residential	200/1,1st l	200/1	Kesbewa	1	138	8	12000000	No	7/10/1928	50000	No	No			1500000
9	2019/07/1	Residential	Battaramulla		Kaduwela	0.01	177	30	5500000	Yes	7/10/2012	120000	Yes	No			183333.3
10	2019/07/1	Residential	No:110/2A	10290	Kesbewa	1	120	10	8000000	No	1/1/2019	0	No	No			800000
11	2019/07/1	Commercia	Malabe	-	Kaduwela	0.5	0.5	12	12000000	No	9/14/1961	1500	No	Yes	15000000	16000000	1000000
12	2019/07/1	Residential	Pannipitiya		Maharagam	1	138	10	7000000	No	9/4/2014	3000000	No	No			700000
13	2019/07/1	Residential	Malabe		Kaduwela	1	177	11.5	80000000	Yes	4/26/2019	90	Yes	No			6956522
14	2019/07/1	Residential	Athurugiriya		Kaduwela	2	138	30	900000	No	7/11/1980	5000000	No	No			30000
15	2019/07/1	Residential	Godagama		Homagama	1	122, 99	16	9000000	Yes	7/11/2017	9000000	Yes	No			562500
16	2019/07/1	Residential	Colombo road Piliyanda		Mampe No	0	120	10	1200000	No				No			120000
17	2019/07/1	Residential	Dematagoda		Thimbirigas	0.5	154	8	4500000	No				No			562500
18	2019/07/1	Residential	Kolpity		Colombo 03	0.02	0.1	8	80000000	Yes	7/1/2019	80000000	No	No			10000000
19	2019/07/1	Commercia	AOD	-	Milagiriya	0.2	100	5	80000000	No				No			16000000
20	2019/07/1	Residential	Pannipitaya		Maharagam	0.5	174	10	600000	Yes	10/2/2014	300000	No	No			60000
21	2019/07/1	Residential	Near commercial credit		Dehiwala-M	0.15	138	16	56000000	No				No			3500000
22	2019/07/1	Residential	Dehiwala		Dehiwala	0.5	119	20	2500000	No				No			125000
23	2019/07/1	Residential	Gothatuwa		IDH	2	175	6	4200000	No				No			700000

Figure 2.11 Sample from dataset

The features were selected following a research [20] conducted in Sri Lanka, to find out the best valuation model suitable for Sri Lanka. The research suggests a model of order as below.

$$Y = b0 + b1\,(AOP) + b2\,(LOP) + b3(EOL) + b4(DTMR) \quad (1)$$

Y is the value of the land, b1 to b4 are the regression coefficients of independent variables. The independent variables are. Extent of the land (EOL), Accessibility of the property (AOP), Location of the Property (LOP), Distance to the Main road (DTMR).

The coefficient values elaboration is as below.

Model	Description	Un standardized Coefficients		Standardized Coefficients	t Sig.	
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	299027.88	4626.709		64.631	.000
	Distance to the Main Road	-134.92	16.255	-.768	-8.300	.000
2	(Constant)	280447.35	9555.246		29.350	.000
	Distance to the Main Road	-120.78	16.916	-.687	-7.140	.000
	Extent of Land	861.83	392.148	.212	2.198	.033
3	(Constant)	235524.95	24131.614		9.760	.000
	Distance to the Main Road	-117.54	16.468	-.669	-7.138	.000
	Extent of Land	828.88	380.312	.203	2.179	.034
	Accessibility to the property	462.69	229.531	.175	2.016	.050
4	(Constant)	163090.29	34768.385		4.691	.000
	Distance to the Main Road	-116.50	15.419	-.663	-7.556	.000
	Extent of Land	894.42	356.763	.220	2.507	.016
	Accessibility to the property	643.99	224.790	.244	2.865	.006
	Location of the Property	581.69	212.272	.232	2.740	.009

Table 2.1 Model Elaboration

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.768(a)	.589	.581	16046.37882
2	.792(b)	.628	.612	15442.13078
3	.811(c)	.658	.636	14962.20019
4	.841(d)	.707	.681	14004.12193

Table 2.2 Model Summary

Distance to the main Road (DTMR) represent 58.1% of the land value in the model. It is significant at 1%. The adjusted R square in the model (b) is 0.628 and it indicates the two variables account for 62.8% of land value variations. Entry of Extent of the land into the model increases the R square value by 3.9%. Accessibility to the Property is the third variable to enter the equation (c) and its explanatory power of the regression equation increases to 65.8 percent that means the entry of ACTP in the model increases R^2 by 3%. The final model, the explanatory power of the regression equation increases 70.7 percent and the last variable to enter the equation is LOP. Entry of LOP in the model increases R^2 by 4.9 percent [20].

Though satisfying results have been acquired, the values of the lands may be vastly different when evaluated by a professional valuer since we considered price value to train the models.

For that, to train the models in order to enhance the predictions, the data collection unit was implemented to gather data from the valuers, which include a valuer recommendation module based on the number of feedbacks given by him or her providing a platform for the valuers to market themselves by actively interacting with the system.

2.3.2 Implementation and testing

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.

Regression analysis is the conventional technique used in manual valuation process. The best linear equation is found, as measured by the least squares method, even if the relationship is actually nonlinear. But the coefficient of determination, R^2 , will measure how well the line fits the data points after finding the one which minimizes the sum of the squares of the vertical distances – the residuals – between the line and the points. An R^2 of 100% indicates that the equation explains 100% of the variation in the dependent variable around its mean within the relevant range of the sample. An R^2 of zero indicates that regression can find no relationship between the dependent variable and the independent one(s), or no line that fits any better than any other one [5].

MLR has several advantages than other algorithms. The ability to determine the relative influence of one or more predictor variables to the criterion value is one. Multivariate techniques provide a powerful test of significance compared to univariate techniques.[15] Multivariate techniques to give meaningful results, 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.

Random forest regressor

Random Forest is a learning algorithm first created by Tin Kam Ho [21], a computer scientist at IBM, and later extended by Leo Breiman and Adele Cutler [22] [23] . It has been in the literature of stock price prediction for quite a long time. But there are very limited approaches have been made to utilize this model in real property value prediction. Hence, we tested the random forest regressor too.

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. Decision trees are said to have very low bias and high variance. This results in overfitting the training sets. Random Forest overcomes this problem by training multiple decision trees on different subspace of the feature space at the cost of slightly increased bias. This means none of the trees in the forest sees the entire training data. The data is recursively split into partitions. At a particular node, the split is done by asking a question on an attribute. The choice for the splitting criterion is based on mean squared error, which is equal to variance reduction as feature selection criterion, or the mean absolute error in regression using random forest.

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

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 complexity of the ANN depends on the number of neurons and the inter-neural connections. The input layer is formed by neurons which contain the exogenous information, translated in terms of the pulse for the neurons of the upper level [24]. Output layer is where the result is returned. In between these layers , the hidden layers are present where weights of the connections are decided in order to convert the input to output. In addition to that dropout layers have been introduced in between each of hidden layers to regularize and prevent overfitting.

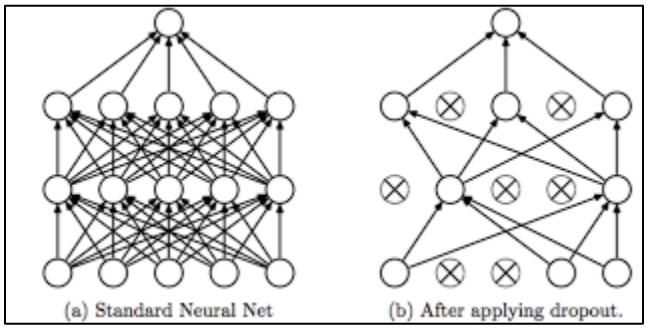


Figure 2.11 Standard Neural net (left) and Neural net with dropout layers

The process happening within ANN is of black box nature, where only the input and output are visible. The transference of information between neurons happen by sigmoid activation function, which is modeled continuously between 0 and 1 [24]. The connections are present only between neurons of two successive levels, and the pulses of the neurons are direct (one way) from the input layer to the output layer [24]. The diagram of a neuron base is described in figure 2.12, in which the neuron is divided into two parts: in the first part, on the left, the pulses received are added; in the second part, the output is determined through the activation function.

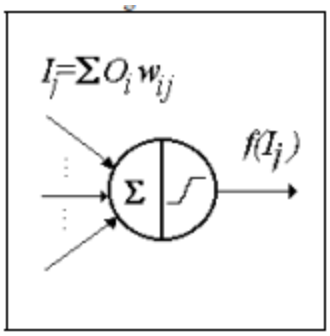


Figure 2.12 Training scheme of a neuron base

Since the ANN is capable of learning a particular task , learning technique should be defined, to appropriately update the weights of the connection of the network. In this research supervised learning technique was used.

In accordance to standard analytical practice, the estimative sample has been divided in a random basis into two sets, the "training set" and the "test set". The training set includes 80% of the sample, corresponding to over 160 transactions, leaving the remaining 20% of the total cases as the test set 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.

The Ensemble model

The current price prediction system was implemented using an ensemble model of MLR and time series ARIMA model which had the accuracy of above 0.75. This is an averaged ensemble model where significance of both ARIMA and MLR models are considered to be equal.

Then the h5 model was generated for each best algorithm to serve the API calls to make the predictions for new user instances.

2.3.3 Testing

Software testing can be mentioned as,

- Unit Testing
- Component Testing
- Integration Testing
- System Testing

In this scenario, since we deal with real data, the best testing strategy is real user monitoring and comparison for the accuracy of the model.

Other than API testing, above mentioned testing strategies can be carried out to ensure the consistency of the system.

Unit Testing

Each unit is tested individually to find whether it's fit for use. This is used to identify smallest part of problems earlier stages of testing, and most important thing in unit testing is identify the bug than correcting it.

Component Testing

Each component testing done in the application separately also its known as program testing here it found the bugs or defect and take the actions to correct it.

Integration Testing

Each module of the software combined and tested as a group. It must be test after unit testing.
System Testing

This is the level of testing where complete software and integrated software is tested. It verified as system whether it meets the requirements. This will ensure the quality level of the system.

2.4 Tools and Technologies

2.4.1 Hardware interfaces

For the developer end, a computer with

- CPU: Quadcore Processor
- RAM: At least 8 GB
- Storage: 1 TB

2.4.2 Software interfaces

- Database

  MongoDB is the database used in this system
- Miniconda

  We use Miniconda as the application launcher. It allows us to launch applications and easily manage conda packages, environments and channels without the need to use command line commands.
- Jupyter Notebook

  Jupyter notebook is an IDE we have used to develop our machine learning models and it is powerful interactive development environment for the Python language with advanced editing, interactive testing, debugging and introspection features.
- PyCharm

  Python development IDE by Jetbrains

2.5 Commercialization aspects

We believe our product is going to be a better option for entrepreneurship since this is the first of its kind in Sri Lanka. This application would be useful for regular customers as well as land developers, and land sale owners. Also, the facts mentioned under section 1.3 above makes the application unique and useful for the users. We hope to offer a free trial of the product for a month and then have an option to subscribe with the system for a reasonable fee than offering few services for free and then providing the report document subjected to a fee. The latter option might be suitable if the reports are recognized by local authorities, banks like organizations as a substitute to a valuation report provided by a valuation officer in the future. Therefore, we believe, there are numerous other features which can be added to the application which makes it more valuable.

Accuracy of the system is determined through the training/testing environment of AI model development. Our intention is to provide the best predictions by finding out the most accurate algorithms to be used with AI model.

We expect our application to provide the outputs within an optimal minimum time so the users identify the product as instant, reliable and effective one of its kind which makes their effort on identifying the lands they are going to buy easier. Since we will be developing a web application UI responsiveness is also important until an Android/IOS apps are developed.

We consider about the scalability of our product to be of the same importance as accuracy because there can be number of users accessing the resources at a time when the product is published.

The system complexity can affect the cost of the services provided. But the service would be much cheaper than the manual process since it can tradeoff the indirect costs and effort of travelling, gathering data etc.

This application should be hosted to be accessible by public. We can add new features like

- Giving a suggestion of the type of suitable building to be built whether it is of some business value, suitable for residence etc.
- Prediction of possible schools a child can enroll when living in that area
- Check for neighborhood suitability, crime rate in the area etc.

to replace the entire valuation process.

3. RESULTS AND DISCUSSION

3.1 Results

As mentioned above, the models tested have been evaluated in terms of MAE, MSE, and RMSE. The results can be summarized as follows for the generic machine learning models tested in this component.

	MAE	**MSE**	**RMSE**
MLR	12578.2076	370573753.442	60874.769
Random Forest Regressor	69388.61903	17241415729.677	131306.572
ANN	495306.848	351944183254.44	593248.838

Table 3.1 Model evaluation summary

Out of the three models MLR found out to be having the best performance.

As mentioned above, in 2.2.2 , an ensemble model has been used in the final system. First reason is because the end user input only the location, and the system has to provide location specific predictions. Second reason was the unavailability of enough data to build an ARIMA model of higher dimensions.

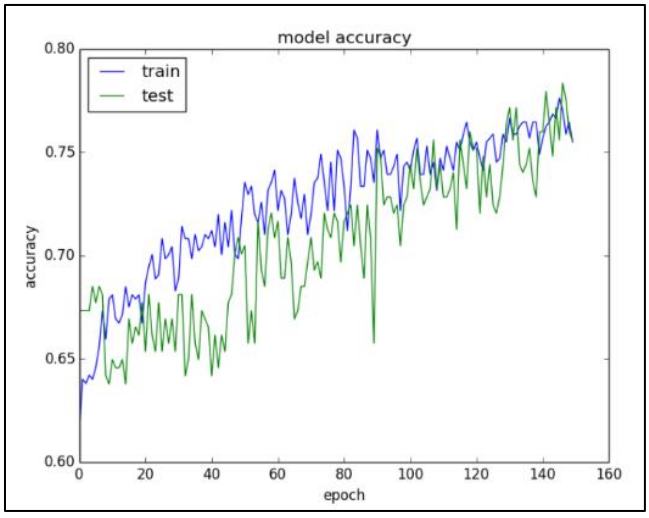


Figure 3.1 : Accuracy curve for ensemble model

Below is the best fitting curve obtained for our MLR model depicting the correlation between extent of the land, distance to the main road and price. Since there were some outliers the scattered data looks somewhat distorted. Hence, the extents greater than 500000 square feet were discerned. Accuracy score for the MLR model was 0.7033086763381378.

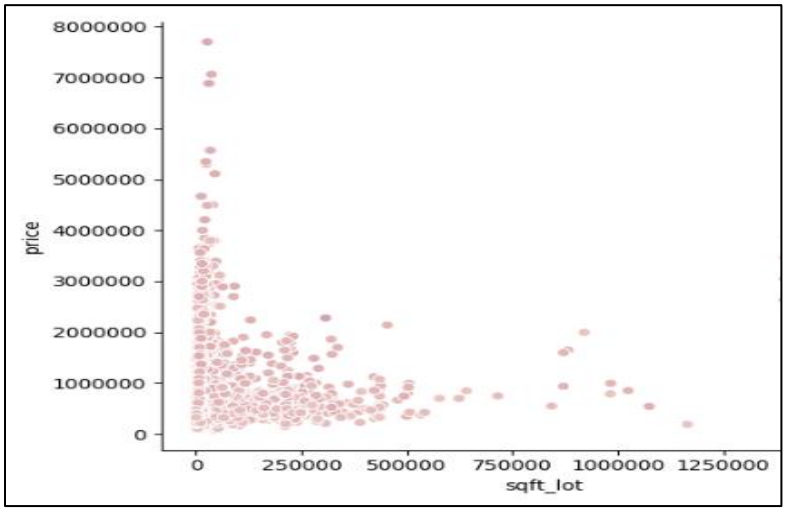


Figure 3.2 Plot of extent of the land, distance to main road features

Then the Random Forest regressor also did not perform very well as expected like it usually does in stock price prediction. Though the accuracy score was 0.5619, it had a higher MSE than MLR. We expected the ANN to make the best predictions out of these, but it did not. The graphs plotted in 12 consecutive epochs of ANN are displayed below. The red curves show training data accuracy while the blue curves display training data. That depicts model is kind of overfitting to the training data that it performed well with training data but when it comes to test predictions, the accuracy is really low.

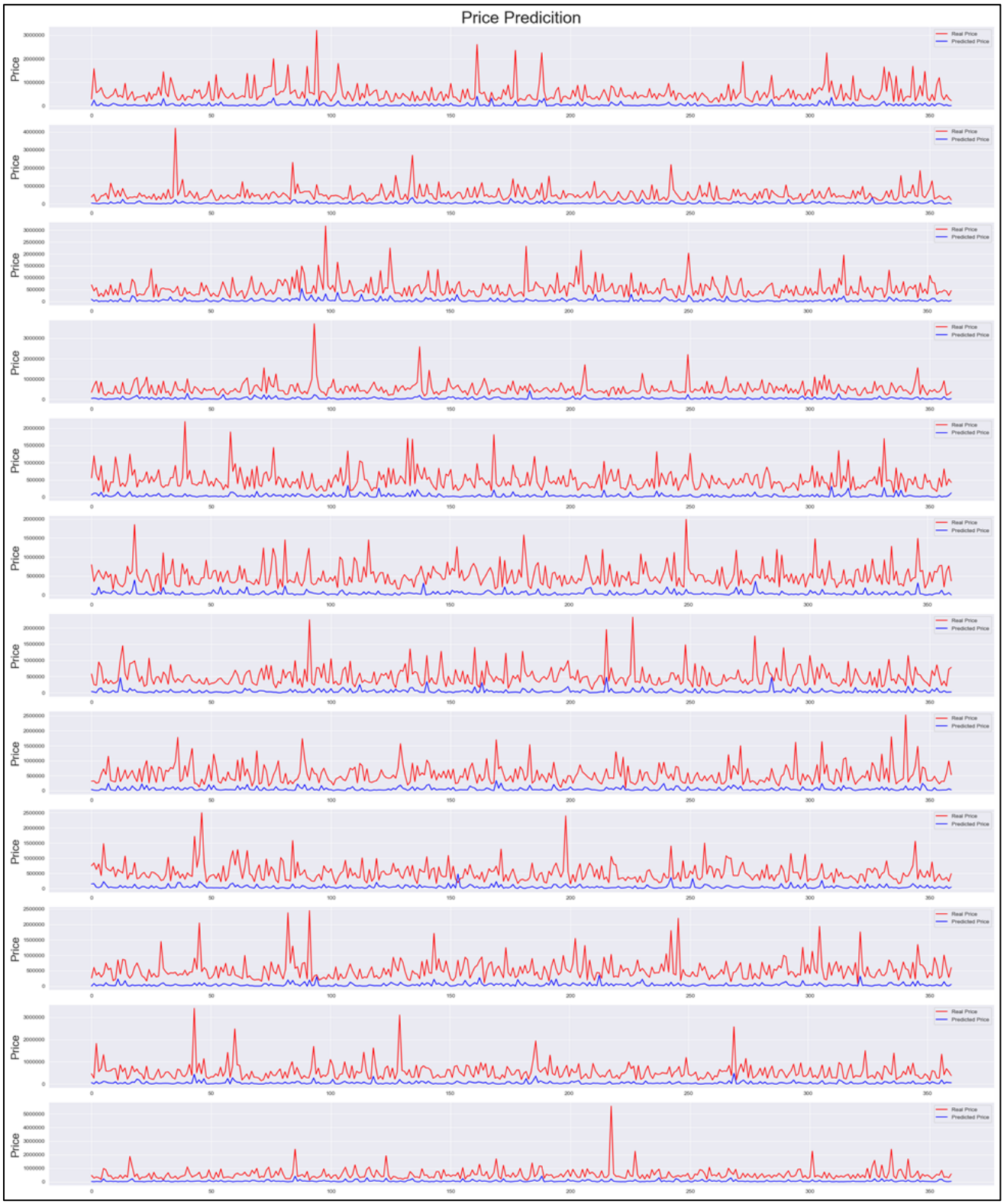


Figure 3.3 Train accuracy (in red) and test accuracy (in blue) curves of ANN model

3.2 User Interfaces

Figure 3.4 is the home page of our application where the user can input the location

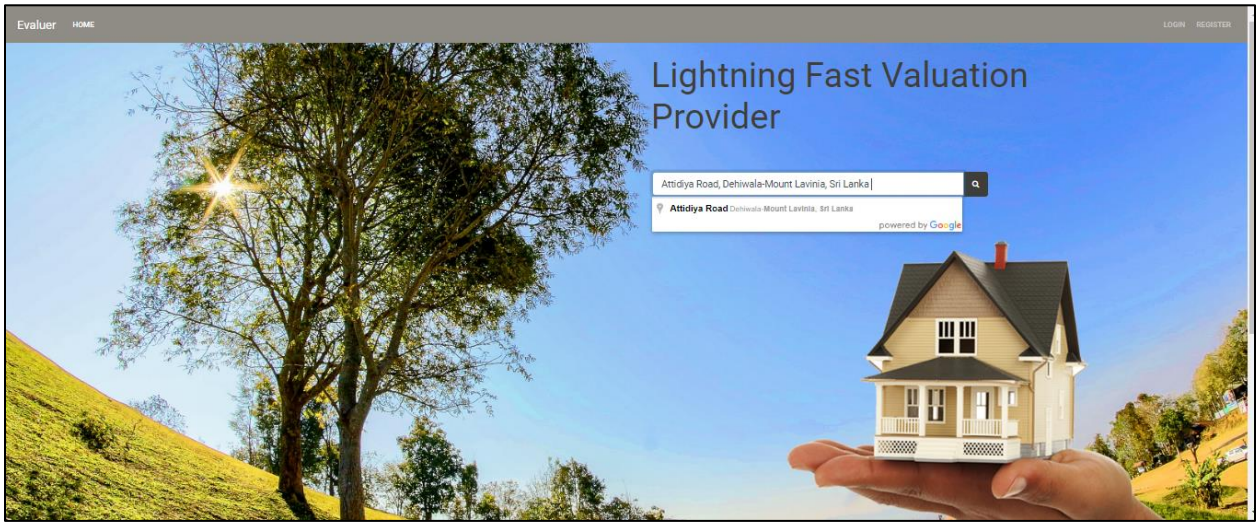


Figure 3.4 Home page

Below figure 3.5 is the current price prediction view.

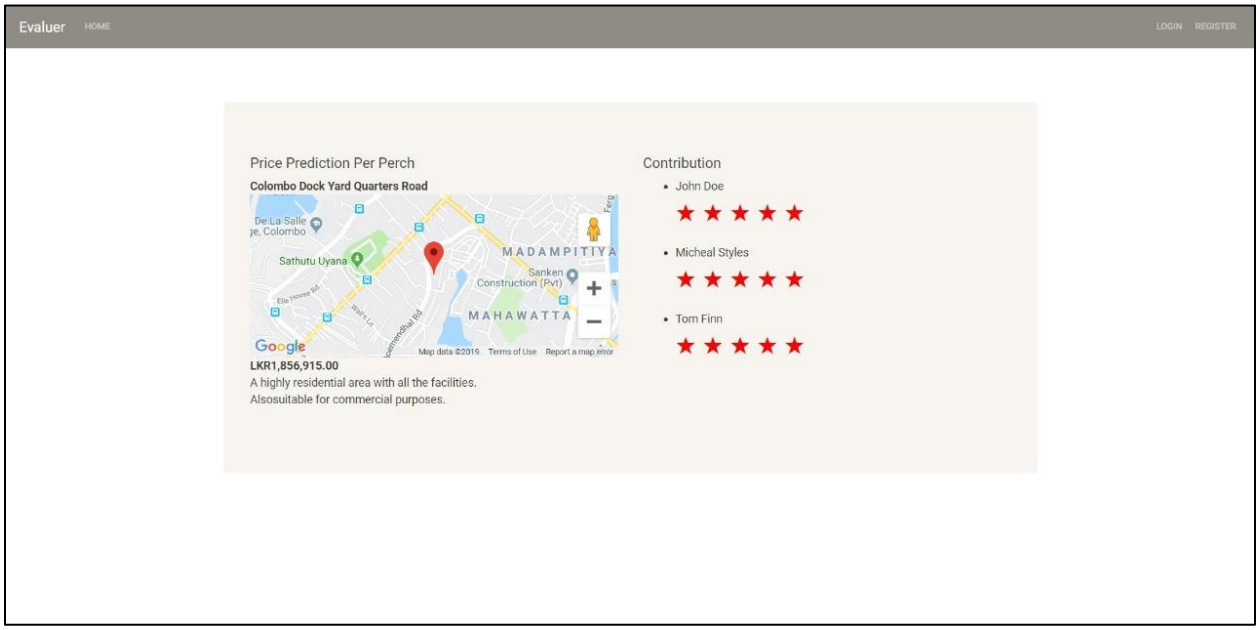


Figure 3.5 Current value prediction along with valuer recommendation

3.3 Discussion

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). It can be concluded that MLR gives out better predictions when the dataset is small. According to the results, time series algorithms predicted values with comparatively less error than the others. Based on the above findings, we can conclude that the ARIMA model predicts the current value with higher accuracy than any other model. Finally, it can be concluded that ARIMA model has outperformed all the other machine learning models in price prediction. But to consider the location when predicting, we needed to use an ensemble model of MLR model and ARIMA model.

4. CONCLUSION

In this research, we discovered efficient generic architectures to make better predictions based on relatively small dataset that provides a better performance within a limited duration. The findings of this research are helpful in automating the land valuation process. Also, it lays foundation for conducting future researches considering different features of the land and measuring effectiveness of different combinations of features and machine learning and deep learning algorithms.

Based on the observations above, we can conclude that we have developed a satisfactory system to predict the land values.

Our research findings prove that ARIMA model has the least error among the other tested models and it can achieve an accuracy of around 0.75 in predicting current value when an ensemble model of ARIMA and MLR models. 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.

To enhance the benefits of the system we can add new features like

- Giving a suggestion of the type of suitable building to be built whether it is of some business value, suitable for residence etc.
- Prediction of possible schools a child can enroll when living in that area
- Check for neighborhood suitability, crime rate in the area etc.

to replace the entire valuation process.

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] Sampathkumar, V., Santhi, M. and Vanjinathan, J. (2015). Forecasting the Land Price Using Statistical and Neural Network Software. Procedia Computer Science, 57, pp.112-121.

[5] Zurada, J., Levitan, A. and Guan, J. (2011). Non-Conventional Approaches To Property Value Assessment. Journal of Applied Business Research (JABR), 22(3).

[6] 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

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

[8] 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].

[9] Trulia. (n.d.). Retrieved February 22, 2019, from https://en.wikipedia.org/wiki/Trulia [Accessed 23 Feb. 2019].

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

[11] “Part 4: Quotable Value Limited's QV Homeguide Application.” Office of the Auditor-General New Zealand, www.oag.govt.nz/2018/digital-access/part4.htm. [Accessed 24 Feb. 2019].

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

[13] “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].

[14] 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].

[15] 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].

[16]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.

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

[18] Li, L., Prussella, P., Gunathilake, M., Munasinghe, D. and Karadana, C. (2015). Land Valuation Systems using GIS Technology Case of Matara Urban Council Area, Sri Lanka. Bhumi, The Planning Research Journal, 4(2), p.7.

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

[20] 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

[21] Tin Kam Ho. “Random decision forests”. In: Document analysis and recognition, 1995., proceedings of the third international conference on. Vol. 1. IEEE. 1995, pp. 278–282

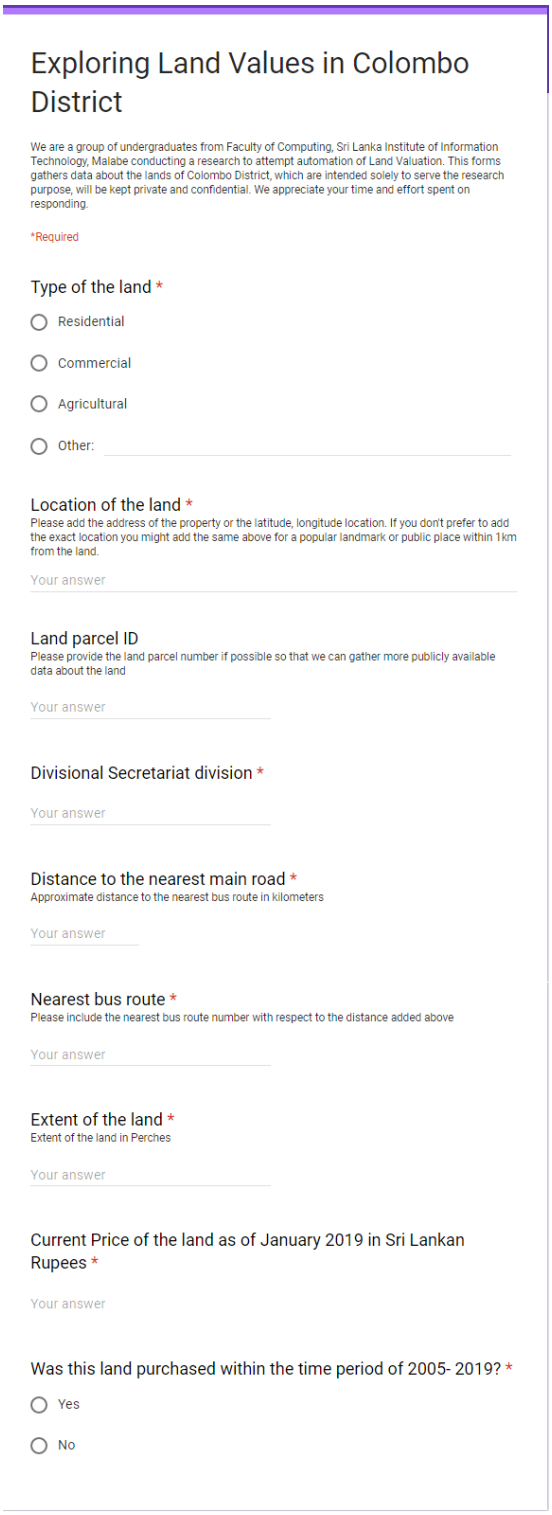
[22] Leo Breiman. “Random forests”. In: Machine learning 45.1 (2001), pp. 5–32

[23] Random Forests by Leo Breiman and Adele Cutler. URL: https://www.stat.berkeley. edu/˜breiman/RandomForests/.

[24] P. Morano, F. Tajani and C. Torre, "Artificial intelligence in property valuations An application of artificial neural networks to housing appraisal", *Semanticscholar.org*, 2015. [Online]. Available: https://www.semanticscholar.org/paper/Artificial-intelligence-in-property-valuations-An-Morano-Tajani/57de46ad59391c6b120d076863e5387111a60a1d. [Accessed: 15- Apr- 2019].

APPENDIX 1

Appendix 1.1 Google form to collect data – Part 1



Exploring Land Values in Colombo District

We are a group of undergraduates from Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe conducting a research to attempt automation of Land Valuation. This forms gathers data about the lands of Colombo District, which are intended solely to serve the research purpose, will be kept private and confidential. We appreciate your time and effort spent on responding.

*Required

Type of the land *

- Residential
- Commercial
- Agricultural
- Other:

Location of the land *

Please add the address of the property or the latitude, longitude location. If you don't prefer to add the exact location you might add the same above for a popular landmark or public place within 1km from the land.

Your answer

Land parcel ID

Please provide the land parcel number if possible so that we can gather more publicly available data about the land

Your answer

Divisional Secretariat division *

Your answer

Distance to the nearest main road *

Approximate distance to the nearest bus route in kilometers

Your answer

Nearest bus route *

Please include the nearest bus route number with respect to the distance added above

Your answer

Extent of the land *

Extent of the land in Perches

Your answer

Current Price of the land as of January 2019 in Sri Lankan Rupees *

Your answer

Was this land purchased within the time period of 2005- 2019? *

- Yes
- No

Appendix 1.2 Google form – part 2

Recently Purchased Lands

Please complete this section if your land was purchased between 2005 - 2019

Date of Purchase *

Please provide the approximate date of purchase

Date

mm/dd/yyyy

Purchased Value in Sri Lankan Rupees *

Your answer

Were there any buildings in the land during the time of purchase? *

- ○ Yes
- ○ No

Appendix 1.3 Google form – Part 3

Valuation of land

Any valuation carried on the property? *

- Yes
- No

Details of the past Valuation done on the land

Year of most recent valuation *

Your answer

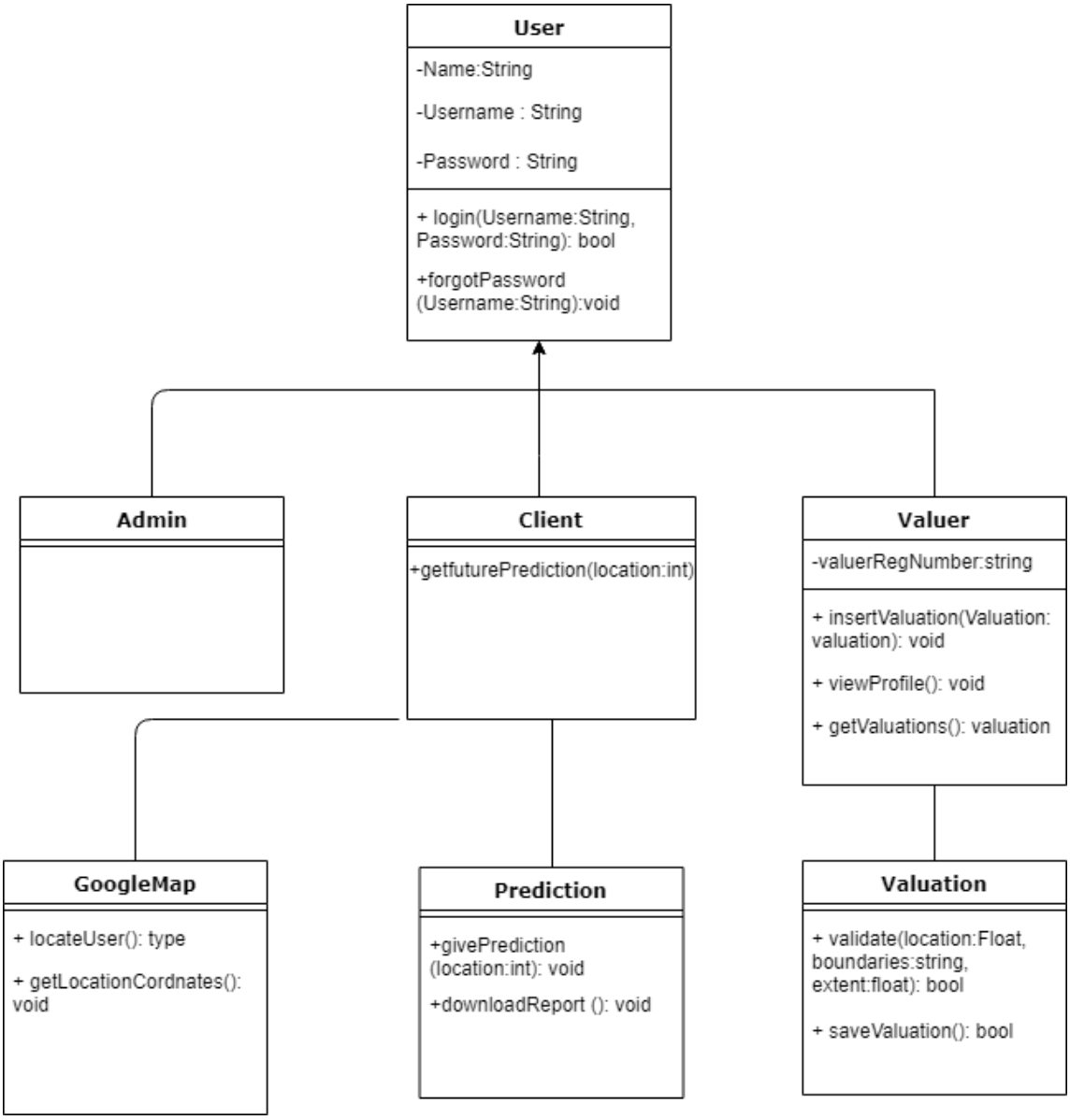
Value of the land as per valuer's report in Sri Lankan Rupees *

Your answer

GET LINK

Never submit passwords through Google Forms.

Appendix 1.4 Class diagram for current value prediction



Appendix 1.5 Use case diagram for current value prediction