E-Valuer

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

Mrs. M.P.A.W. Gamage

Mrs. Pasangi Rathnayake

Rodrigo U.S.D.	IT 16 15 4490

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
Rodrigo U.S.D.	IT16154490	

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

In order to study the impact of various factors on the land price, different prediction models based on deep learning and machine learning are built to determine the existing data of the land prices in order to more accurately predict the land prices or its changing trend in the future.

Professional valuers estimate land values based on current bid prices (open market values). However, there is no reliable forecasting service for land values with past bid prices being taken as the best indicator of current price movement.

In this paper, Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA), trained using national land transaction time-series data, which forecasts future trends within the land market.

Keywords**—**Deep learning, ARIMA, LSTM, land valuation.

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.

TABLE OF CONTENTS

LIST OF FIGURES

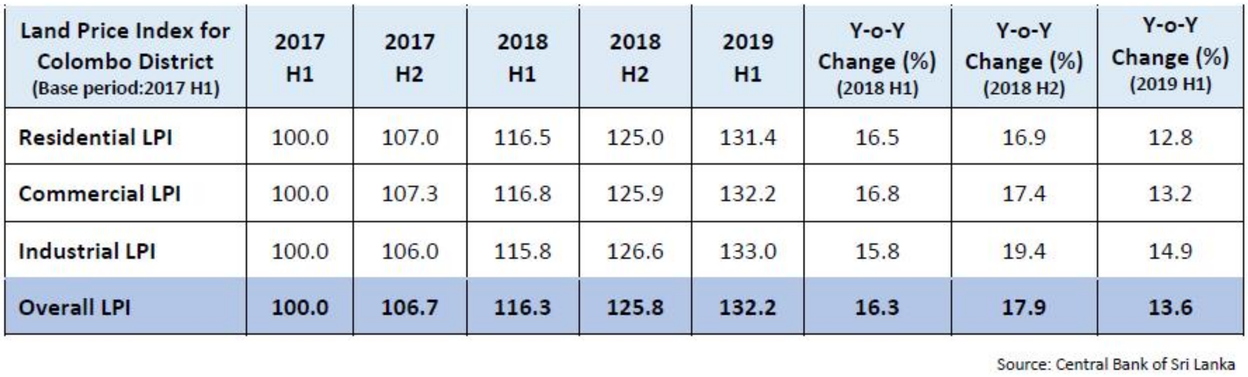
LIST OF TABLES

1. INTRODUCTION

1.1 Background literature

Major trends in real property changes in Sri Lanka can make a huge impact on economy in general.

Land Price Index for Colombo District, compiled by the Central Bank has reached 132.2 during the 1st half of 2019, recording an increase of 13.6 percent compared to the 1st half of 2018[1]



Land Price Index for Colombo District (Base period:2017 H1)	2017 H1	2017 H2	2018 H1	2018 H2	2019 H1	Y-o-Y Change (%) (2018 H1)	Y-o-Y Change (%) (2018 H2)	Y-o-Y Change (%) (2019 H1)
Residential LPI	100.0	107.0	116.5	125.0	131.4	16.5	16.9	12.8
Commercial LPI	100.0	107.3	116.8	125.9	132.2	16.8	17.4	13.2
Industrial LPI	100.0	106.0	115.8	126.6	133.0	15.8	19.4	14.9
Overall LPI	**100.0**	**106.7**	**116.3**	**125.8**	**132.2**	**16.3**	**17.9**	**13.6**

Source: Central Bank of Sri Lanka

Figure 1: Land price indexes and growth rates

Source: https://www.lankabusinessonline.com/wpcontent/uploads/2019/09/LPI.jpg

This proves the fact real property is a good investment with better returns. the property tax is the main source of income in local government authorities in Sri Lanka [2].

Valuation is the way toward evaluating the attributes of a given land parcel and the gauge of the value of landed property dependent on experience and judgment of the person carrying out the process. Sometimes this dependency on subjectivity can be the main disadvantage of manual process of valuation because it might make deviations from the predefine valuation process. Hence there is a higher probability of not getting the proper valuation for the property. During valuation process, several features such as physical features, social features, heritage value etc. should be taken into consideration. However, it is not seen any such national level standard on real estate as a product except some selected standards on some materials [3] when detecting the market value too.

Also, the pace of re-advancement and extreme demand has directly influenced the increment value of the property than the present capital estimation of the property. In addition to that the property owner always concerned about getting the highest and the best value of the property. If the valuation is done correctly following the standards fair value can be defined for each of the models. However, in Sri Lanka there is no reliable forecasting service for real property values with current prices, physical features being taken as the best indicator of price movement because of the afore-mentioned subjectivity. Therefore, the prices are being fluctuated as per a few property business owner's choice. This research is attempting to identify the relationships between such indicators using modern mathematical approaches and technological trends such as deep learning in order to build a reliable portal which can provide its users with an accurate information to avoid the drawbacks of present manual valuation process.

When examining the previous literature, Zurada et al, 2011[4] 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 [4]. Assuming that land value prediction as a time-series problem since it tends to change with time, Wilson I. et al [5] propose that non-linear time series models such as neural networks are more reliable in predicting values for real property. Time series forecasting is a challenging problem that has been there since a while, that attracted the consideration of investors and scholastics. The procedure is equivalent with modeling, where the result of an obscure variable is created from known or controllable factors. Most time series consist of members that are serially dependent in the sense that one can estimate a coefficient or a set of coefficients that describe continuous members of the series from specific, time-lagged (previous) members [5]. This relation is called autoregression. But it has been identified that implementation of nonlinear systems is quite troublesome due to the black-box nature and also the high sensitivity of such algorithms where they try to fit in every instance in the training environment.

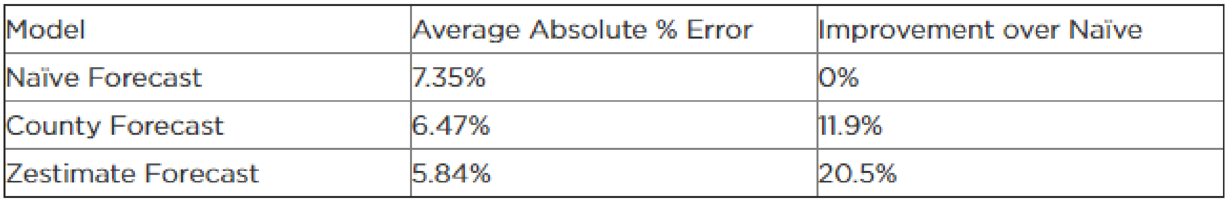
This particular component of our training environment tests the Long Short Term Memory (LSTM) – Recurrent Neural Network (RNN) versus Auto Regressive Integrated Moving Average model with a dataset along with monthly time-series values from 1996 up until 2017.

Existing solutions

The use of AI for residential value forecasting has been suggested in the literature from 1990s. [6]. There are as many as similar applications in developed countries. The main barrier in implementing such a system in Sri Lanka is the unavailability of rich digital information platform to take initiative. However, we managed to gather publicly available suitable data for our models. Below is a brief analysis of existing systems in developed countries.

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. [5] 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%

Features:

- Estimates for 12 months

Zestimate determines an estimation for 12 months for a house based on neighborhood comparable houses. Accuracy of estimate depends on the amount of data used as the underlying approach is Hedonic regression analysis based proprietary algorithm [6] which analyses of several features of the house. The forecasted value is interpolated using cubic spline to connect to current value. [6]

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 Neighborhoods provide photographs, drone footage, etc. so that who are interested about the neighborhood 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%

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, Great Schools Score.
- Commute times at a glance - Using data from OpenStreetMap and General Transit Feed Specification (GTFS) feeds, the user can get an idea of commute times at a glance.[7]

3.QV.co.nz - QV home guide

Quotable Value (QV) provides independent and authoritative information on any home in New Zealand on or off the market [8]. QV.co.nz and their mobile App QV home guide 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 [9]

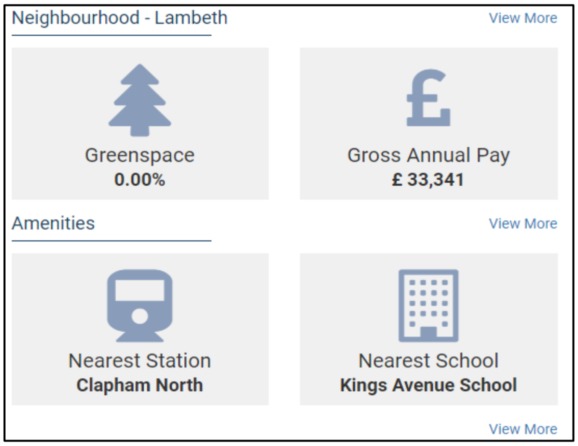
- E-Valuer Report - Subjected to a fee complete valuation report of the property can be downloaded.

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 [11]. 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.



Neighbourhood - Lambeth

Greenspace
0.00%

Gross Annual Pay
£ 33,341

Amenities

Nearest Station
Clapham North

Nearest School
Kings Avenue School

Figure 1.2: Brief Neighborhood analysis

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

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

- Historical data relevant to location

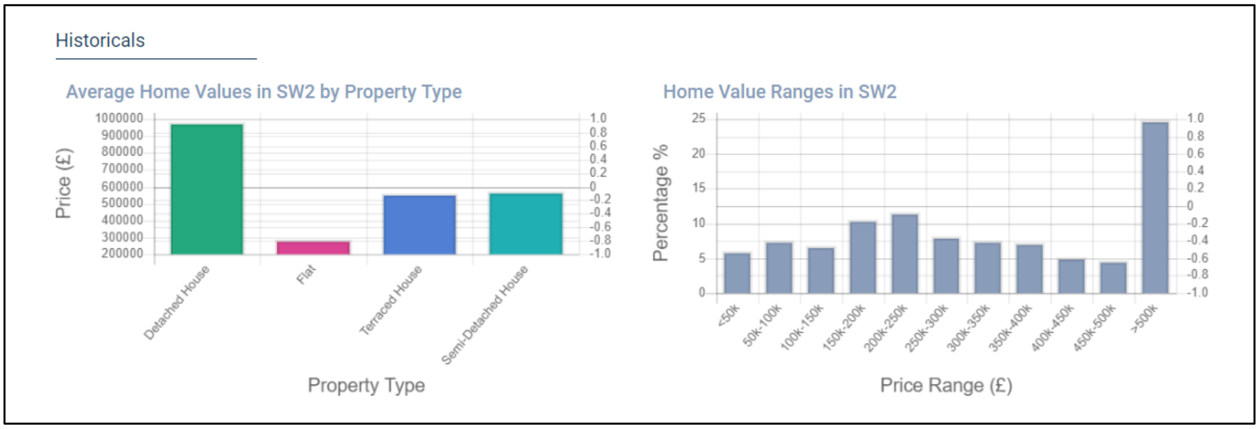


Figure 1.3: Historical Sales analysis

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

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

1.2 Research Gap

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	Publicly available data	Statistical data	Publicly 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

Several industries in Sri Lanka have started using Artificial intelligence (AI), Machine Learning (ML). Usage of augmented reality, virtual reality has become trends in the gaming industry of Sri Lanka. There are several tech companies who utilize these technologies effectively in creating international quality games. Dialog the main communication service provider in Sri Lanka uses AI based service in their support service. However, there are no notably successful machine learning or deep learning-based applications implemented within the country. Several researches have been conducted in application of those concepts in areas like stock price prediction. The problem with applying these algorithms in land value prediction is that there are certain features which data cannot be gathered about like social factors contributing to the value of the land. Hence, in this research those factors are considered to be constant. The time series analysis component is considering just the past value fluctuations of the selected area.

1.3 Research Problem

The main research problem is to develop a system to provide the lands with fair predictions so as to avoid the errors happening due to the subjectivity of the person. Hence, this research is an improvement to the existing manual valuation process. Here the effectiveness of application of LSTM and ARIMA in land value prediction is tested. It is a domain where these algorithms have not been tested as frequently as linear regression and Artificial Neural Networks (ANN). Therefore, the findings of this research will contribute immensely for the future development of this area. 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) more accurate 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 centered around giving quick predictions of valuation concentrated on area of the land helping the client to choose the appropriateness of the land for their purpose. When a client goes to a land, he wants to purchase, they can enter the location of land through the application. Then the ensemble model predicts the present value. At that point the future worth will be anticipated by coordinated effort of two units, one which considers the fluctuation rates of past values and climate impacts, while the other computes the impact of proposed advancement extends in the territory. All these units together create a report which delineates these two kinds of information with applicable other data about the land in an easier manner anybody can understand it.

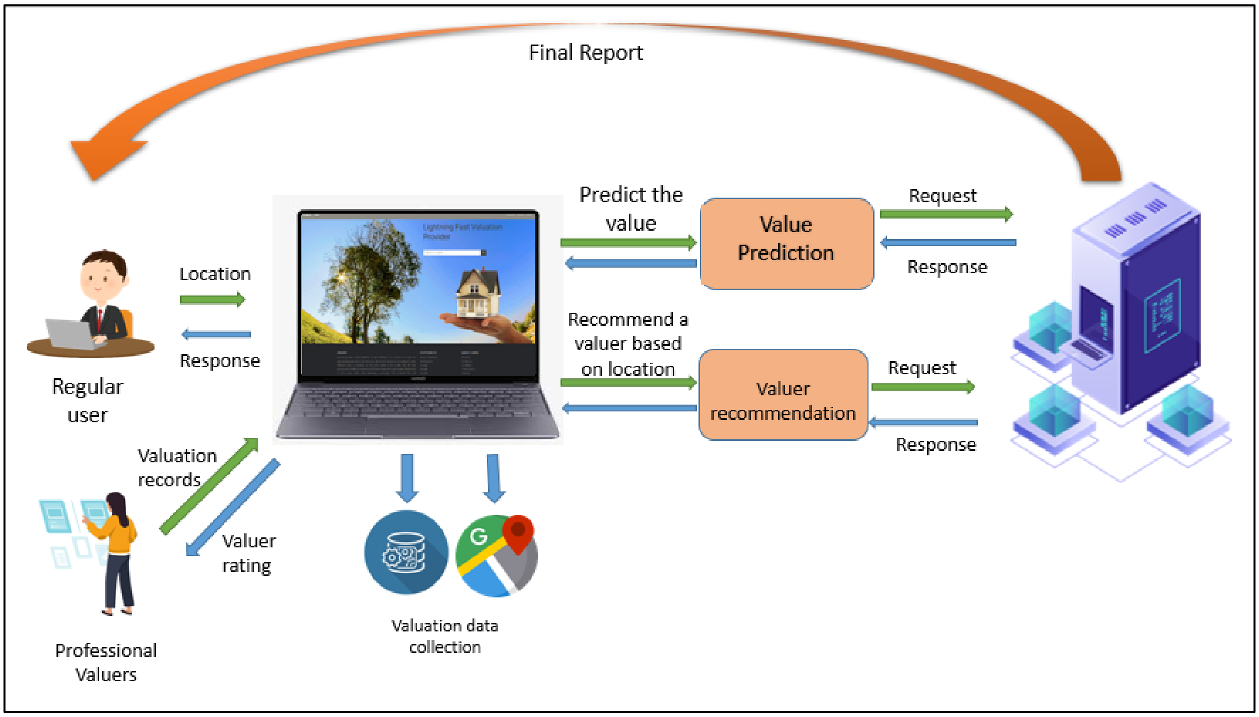


Figure 2.1. System Diagram

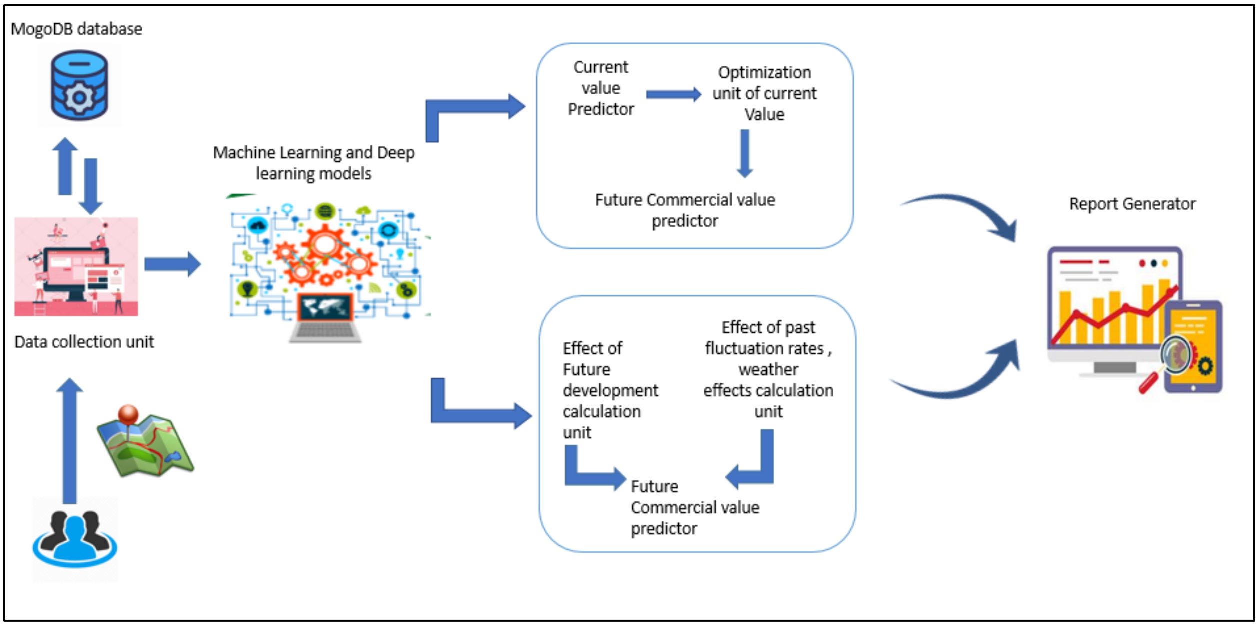

MogoDB database
Data collection unit
Machine Learning and Deep learning models
Current value Predictor
Optimization unit of current Value
Future Commercial value predictor
Effect of Future development calculation unit
Effect of past fluctuation rates , weather effects calculation unit
Future Commercial value predictor
Report Generator

Figure 2.2. High level system architecture

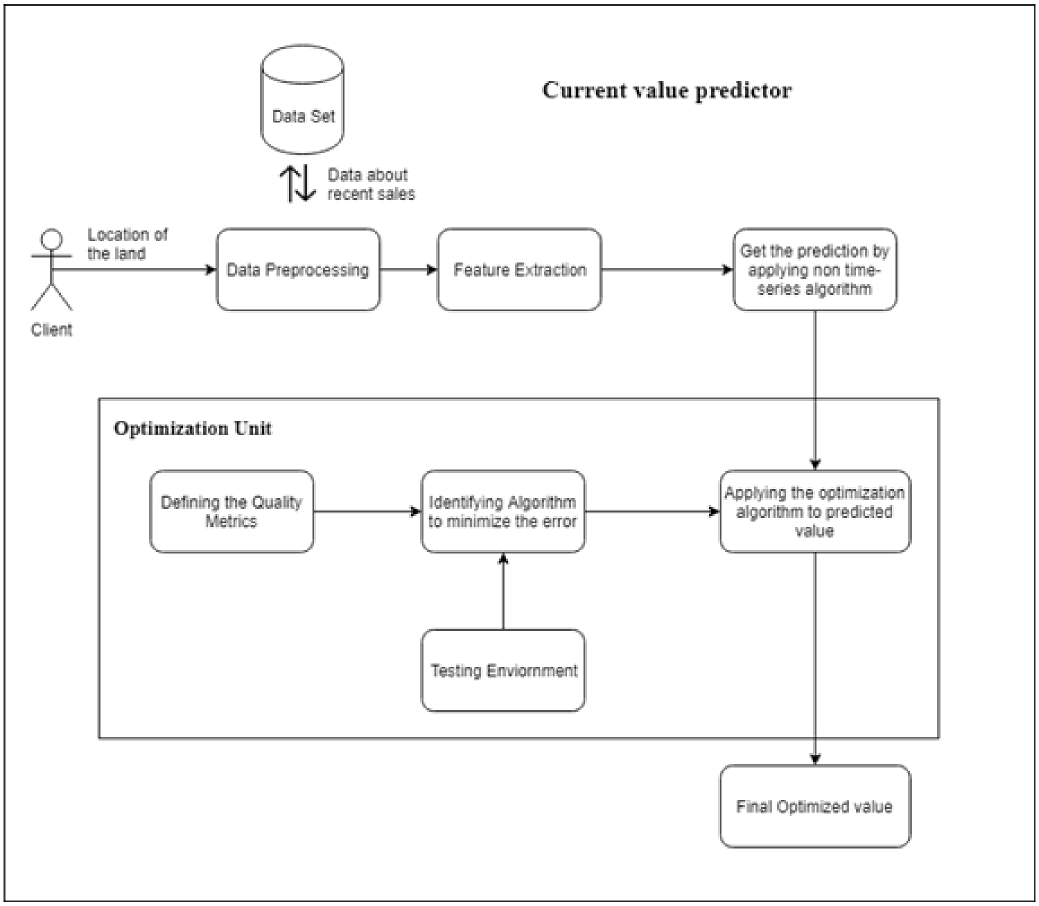

Data Set
Data about recent sales
Current value predictor
Location of the land
Client
Data Preprocessing
Feature Extraction
Get the prediction by applying non time-series algorithm
Optimization Unit
Defining the Quality Metrics
Identifying Algorithm to minimize the error
Applying the optimization algorithm to predicted value
Testing Enviornment
Final Optimized value

Figure 2.3 Current value predictor component design

2.2 Features

2.2.1 Finding the geo-location of the land

This component is in charge of distinguishing the geo-location of the device or the area that is input by the client to get the results. This feature has been implemented using our Google location API. It can intelligently oversee fundamental location technology while meeting different improvement needs when implementing area-based features where it can bind the results to custom location. Since the application centered around Colombo, geo coordinate restrictions have been imposed to the outputs showed.

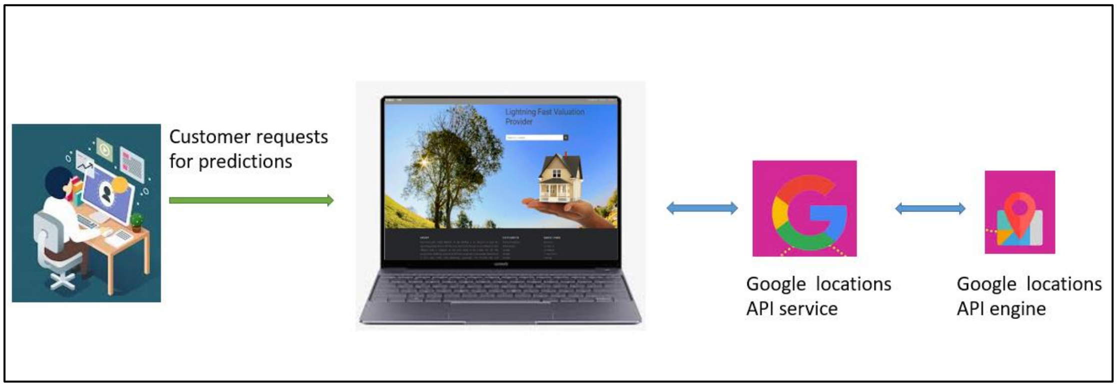
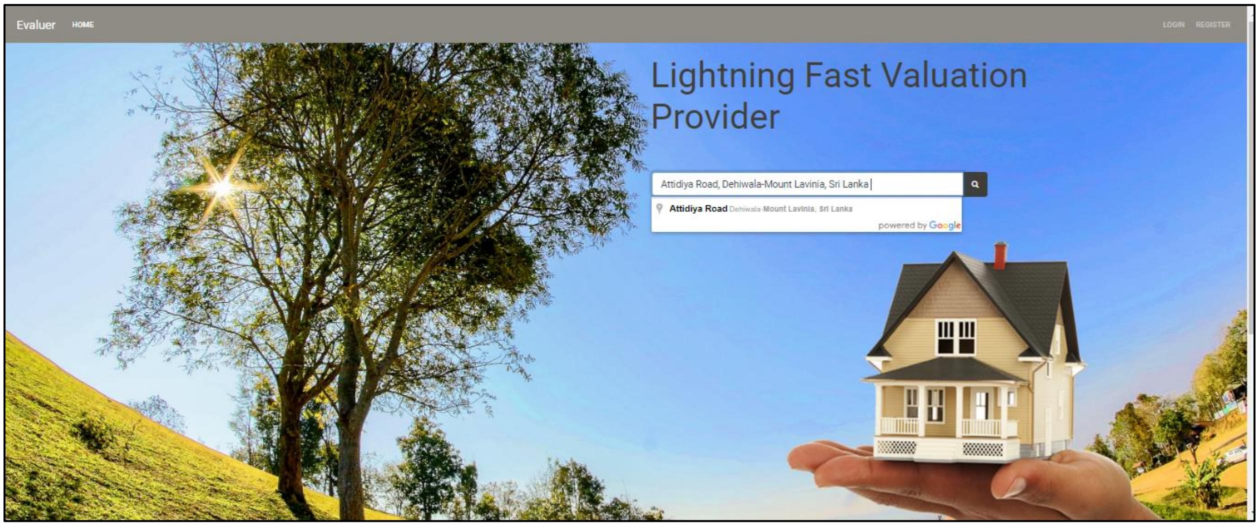

Customer requests for predictions
Lightning Fast Valuation Provider
Google locations API service
Google locations API engine

Figure 2.4 Retrieving geo-location process



Attidiya Road, Dehiwala-Mount Lavinia, Sri Lanka

Attidiya Road Dehiwala-Mount Lavinia, Sri Lanka

Figure 2.5 User Interface

2.2.2 User Authentication

We expect our system to be used by mainly three types of users, namely, Super Administrator which is the product owner, the valuers and the other clients. JWT token-based authentication has been implemented for valuers to login and register themselves.

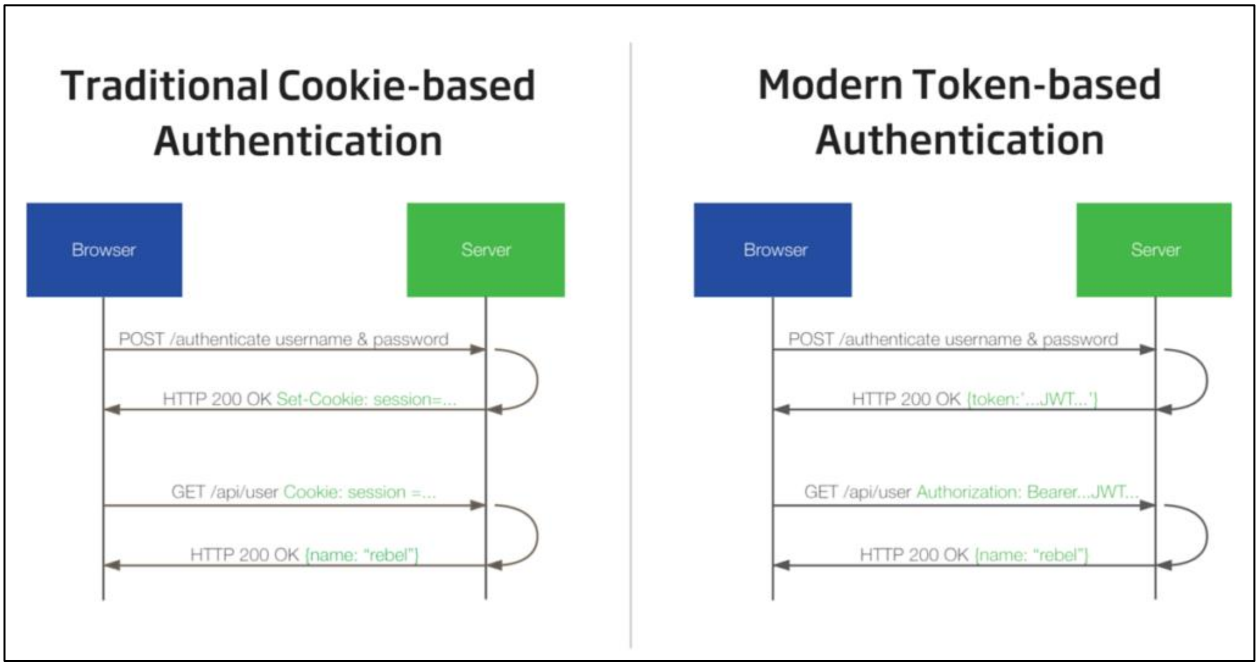

Traditional Cookie-based Authentication
Browser
Server
POST /authenticate username & password
HTTP 200 OK Set-Cookie: session=...
GET /api/user Cookie: session =...
HTTP 200 OK {name: "rebel"}
Modern Token-based Authentication
Browser
Server
POST /authenticate username & password
HTTP 200 OK {token:'...JWT...'}
GET /api/user Authorization: Bearer...JWT...
HTTP 200 OK {name: "rebel"}

Figure 2.6 Cookie based authentication vs Token-based authentication

Source: https://stormpath.com/wp-content/uploads/2016/05/Cookie-v-Token-Diagram-v1-3-1024x536.png

The data sent by the JWT are encoded and signed, but not encrypted. [20] Encoding is used to transform structure of the data while the signing allows the receiver to verify the authenticity of the sender. One problem with JWT being not encrypted is that it does not guarantee the sensitive data security. Here, we are using a JWT that is signed by the HS256 algorithm where only the authentication server and the application server know the secret key. The application server receives the secret key from the authentication server when the application sets up its authentication process. Since the application knows the secret key, when the user makes a JWT-attached API call to the application, the application can perform the same signature algorithm as used in the creation step of JWT. The application can then verify that the signature obtained from its own hashing operation matches the signature on the JWT itself (i.e. it matches the JWT signature created by the authentication server). If the signatures match, then that means the JWT is valid which indicates that the API call is coming from an authentic source. Otherwise, if the signatures don't match, then it means that the received JWT is invalid, which may be an indicator of a potential attack on the application. Therefore, by verifying the JWT, the application adds a layer of trust between itself and the user.[20]

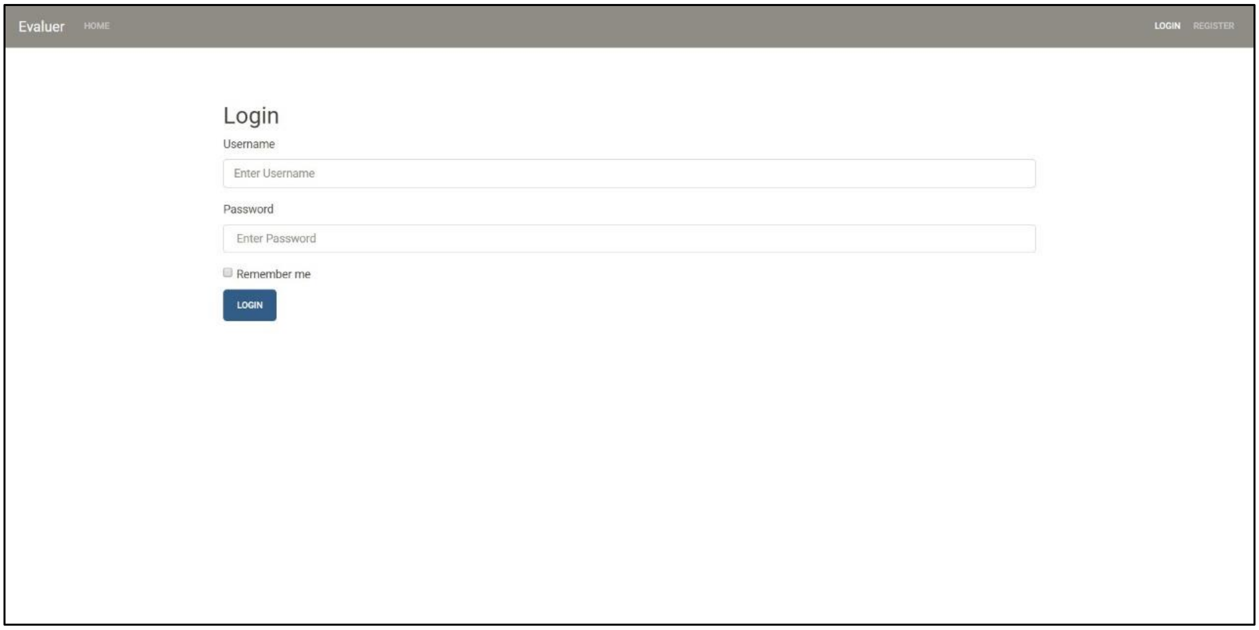
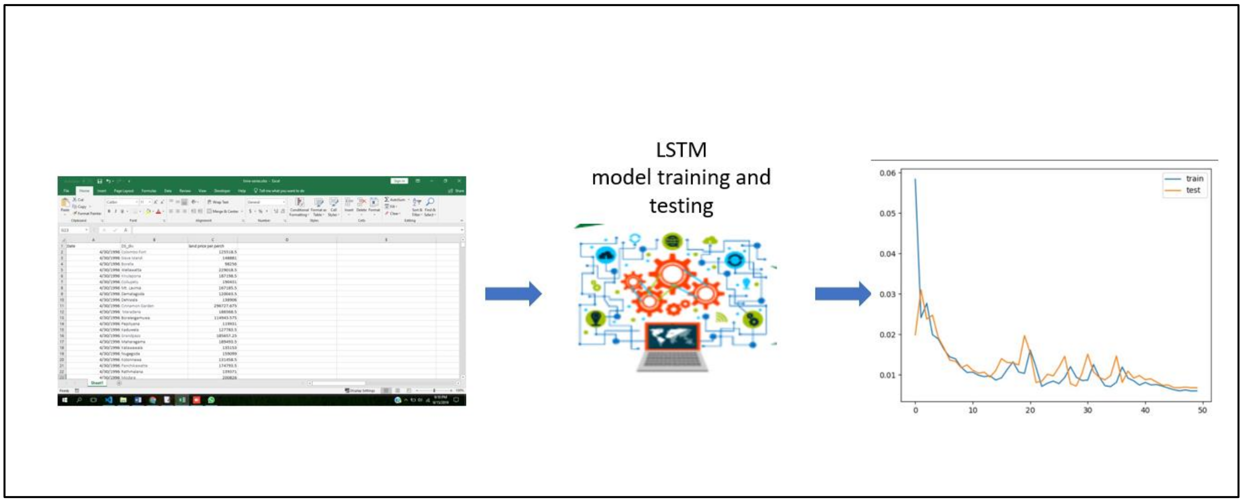


Figure 2.6 Login screen

2.2.3 Price prediction for a particular location

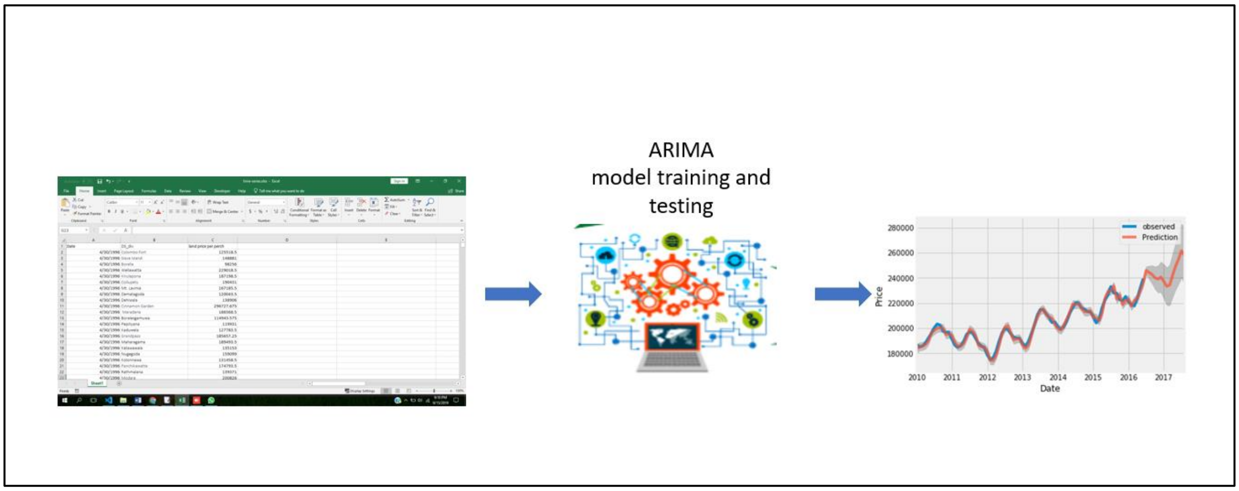
When the user inputs the location, geo-coordinates are taken by the above 2.2.1 feature and sent to the ensemble model of multivariate linear regression and ARIMA model make the prediction for the given location.

One of the major objectives of this project is to test time series deep learning algorithms for their performance on land value prediction. For that, Long Short-Term Memory (LSTM) recurrent neural network and Autoregressive Integrated Moving Average models have been tested on a time series dataset collected manually and evaluated in those models 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 ultimately compared with the results of non-time series algorithms. 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.



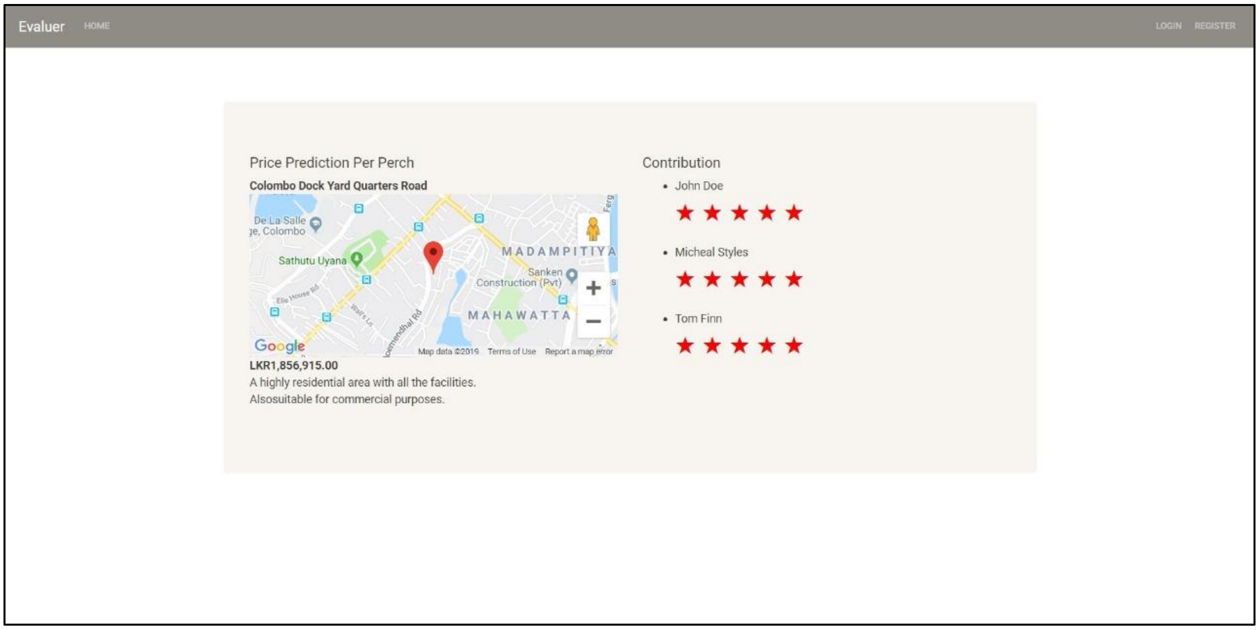
	A	B	C
1	Date	DS_div	land price per perch
2	4/30/1996	Colombo Fort	125518.5
3	4/30/1996	Slave Island	148881
4	4/30/1996	Borella	98256
5	4/30/1996	Wellawatta	229018.5
6	4/30/1996	Kirulapona	187198.5
7	4/30/1996	Collupity	190431
8	4/30/1996	Mt. Lavinia	167185.5
9	4/30/1996	Dematagoda	120043.5
10	4/30/1996	Dehiwala	138906
11	4/30/1996	Cinnamon Garden	296727.675
12	4/30/1996	Maradana	188568.5
13	4/30/1996	Boralesgamuwa	114943.575
14	4/30/1996	Papiliyana	119931
15	4/30/1996	Kaduwela	127783.5
16	4/30/1996	Grandpass	185657.25
17	4/30/1996	Maharagama	189493.5
18	4/30/1996	Kalubowila	135153
19	4/30/1996	Nugegoda	159099
20	4/30/1996	Kolonnawa	131458.5
21	4/30/1996	Panchikawatte	174793.5
22	4/30/1996	Rathmalana	139371
23	4/30/1996	Modara	200826

Figure 2.7 LSTM model training and testing



	A	B	C
1	Date	DS_div	land price per perch
2	4/30/1996	Colombo Fort	125518.5
3	4/30/1996	Slave Island	148881
4	4/30/1996	Borella	98256
5	4/30/1996	Wellawatta	229018.5
6	4/30/1996	Kirulapona	187198.5
7	4/30/1996	Collupity	190431
8	4/30/1996	Mt. Lavinia	167185.5
9	4/30/1996	Dematagoda	120043.5
10	4/30/1996	Dehiwala	138906
11	4/30/1996	Cinnamon Garden	296727.675
12	4/30/1996	Maradana	188568.5
13	4/30/1996	Boralesgamuwa	114943.575
14	4/30/1996	Papiliyana	119931
15	4/30/1996	Kaduwela	127783.5
16	4/30/1996	Grandpass	185657.25
17	4/30/1996	Maharagama	189493.5
18	4/30/1996	Kalubowila	135153
19	4/30/1996	Nugegoda	159099
20	4/30/1996	Kolonnawa	131458.5
21	4/30/1996	Panchikawatte	174793.5
22	4/30/1996	Rathmalana	139371
23	4/30/1996	Modara	200826

Figure 2.8 ARIMA model training and testing



Price Prediction Per Perch

Colombo Dock Yard Quarters Road

LKR1,856,915.00

A highly residential area with all the facilities.
Alsosuitable for commercial purposes.

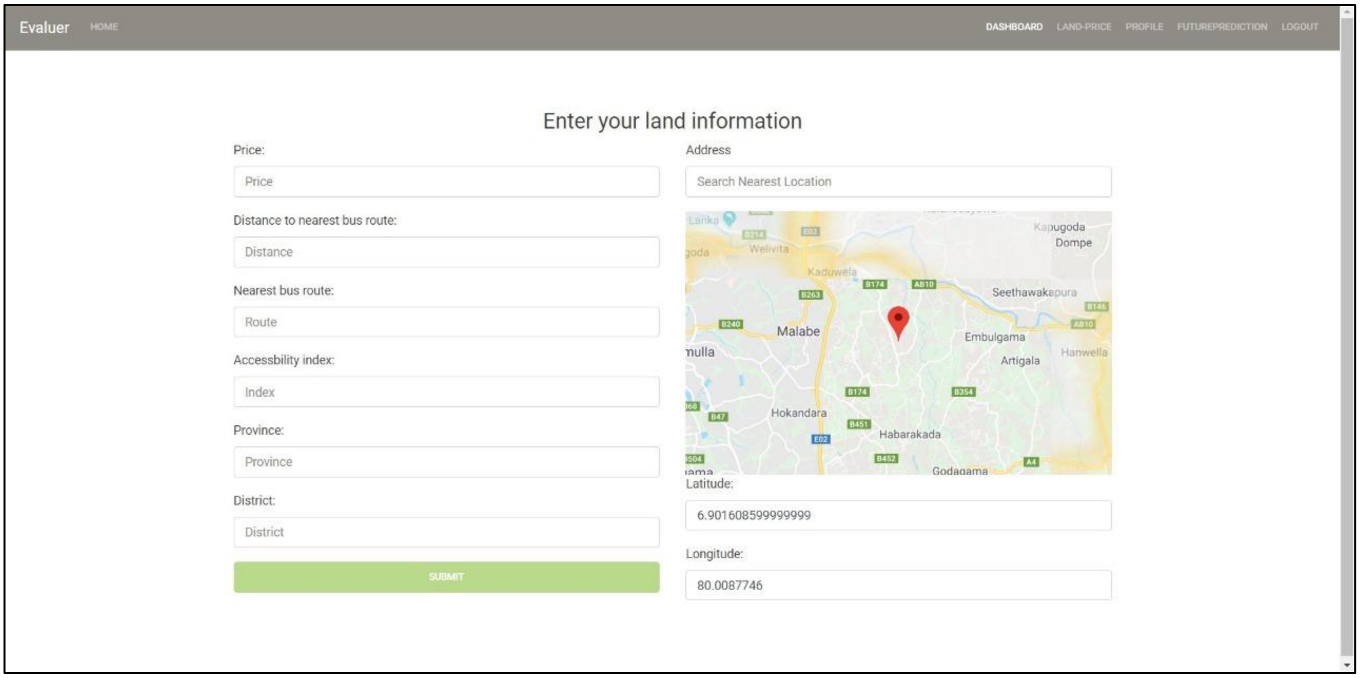
Contribution

- John Doe
  ★ ★ ★ ★ ★
- Micheal Styles
  ★ ★ ★ ★ ★
- Tom Finn
  ★ ★ ★ ★ ★

Figure 2.10 User Interface for displaying current value predictions

2.2.4 Data input for the Data collection unit

This feature of data collection unit allows the authorized valuers to input data about the valuation tasks carried out by them into the system hence, they get promoted along when users seek for price values. The input data will be stored in the mongo DB database. The feature values collected from valuers are selected based on the valuation model used to collect training data for generic algorithm testing which is described in [2].



Enter your land information

Price:

Price

Distance to nearest bus route:

Distance

Nearest bus route:

Route

Accessbility index:

Index

Province:

Province

District:

District

Address

Search Nearest Location

Latitude:

6.901608599999999

Longitude:

80.0087746

Figure 2.11 Data input screen for data collection unit

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 time-series 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 two time series algorithms, namely LSTM and ARIMA models 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 time series data.

First phase of developing the module included selecting best algorithm to predict the current value of a land. For that, Long Short-Term Memory (LSTM) and ARIMA model were tested.

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 time series data collected to predict the current value from a land sale company which had divisional secretariat division wise monthly fluctuation rates of land values from the same area over a period of 20 years from April 1996 up until April 2017, containing above 200 samples.

Though the true price values have been acquired, the values of the lands may be vastly different when evaluated by a professional valuer since market value is not always the value provided by the valuer.

For that, we have implemented a data collection unit to gather data from the valuers, which include a valuer recommendation module based on the number of feedbacks given by him or her, and those ratings will be utilized by the banks, real estate buyers to select an experienced valuer.

2.3.2 Implementation and testing

Long Short-Term Memory (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.

A LSTM unit is a recurrent network unit that excellent at remembering values for either long or short durations of time. The key to this ability is that it uses no activation function within its recurrent components. Thus, the stored value is not iteratively squashed over time, and the gradient or blame term does not tend to vanish when backpropagation through time is applied to train it [21]. This system is solely based on time series since the data set contains only the divisional secretariat wise prices from April 1996.

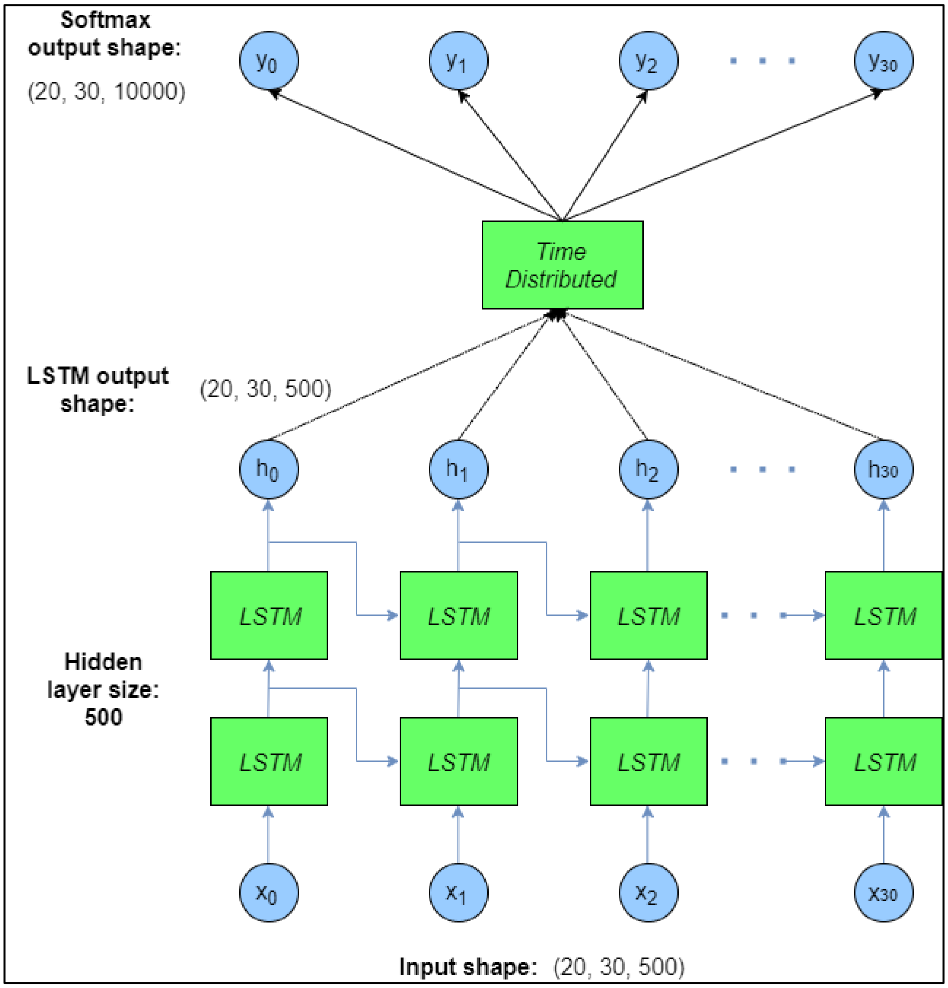


Figure 2.12 Example for a LSTM

To test this model, timeseries dataset having monthly land values from the area over a period of 20 years was used. The dataset was having lags of unknown duration, hence, out of available RNN types, LSTM was the best option.

ARIMA model

ARIMA is a popular model applied in financial time series forecasting. The applications of an ARIMA model are well documented in Barras (1983), Box and Jenkins (1976), Chow and Choy (1993), Cleary and Levenbach (1982), Hanke and Reitsch (1986), Herbst (1992), and Nazem (1988) for example. Box-Jenkins methodology requires that time series values must be stationary and invertible before one recognizes any pattern in the data and attempt to fit any of the ARIMA model. An ARIMA model uses an iterative approach of identifying a possible useful model from a general class of models. Another tool for identification of stationarity in an ARIMA model is the ordinary and partial autocorrelation. Non-stationarity may be present if the values plotted in the correlogram do not diminishes at large lags. When the original series or correlogram exhibits non-stationarity, successive differencing is carried out [22].

ARIMA standing for Auto Regressive Integrated Moving Average is the most popular and commonly used statistical method for time series prediction. This model was utilized in both current value prediction and future value prediction units.

Procedure to follow with this model is splitting 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.

The ARIMA forecasting for stationary time series is nothing but linear equation (like linear regression). The predictor depends on (p, d, q) of Arima model.

The ARIMA model dependent on following components.

1. Number of AR (Auto regressor) term **(p)**: AR term is lag of dependent variable. If p is 3 then predictor for x(t) will be x(t-1), x(t-3).
2. Number of MA (Moving Average) term **(q)**: MA term is lag of forecast error of predictor equation. If q is 3 then error for x(t) will be e(t-1), e(t-3)

3. Number of Differences **(d)**: The number of times that the raw observations are differenced, also called the degree of differencing.

- To determine p and q we will use two plots

1. Auto Correlation Function **ACF**: It is a measure of correlation between TS and lagged of TS (q) which can be depicted as follows. Here x_t and ε_t are actual values and random errors (or random shock) at time 'i'. θ_i is the model parameter. Mean of the sample data is depicted by the symbol μ. q is known as the order of the model.

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

2. Partial Auto Correlation Function **PACF**: This measures the correlation between the TS with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons. (p) . The model can be expressed in the form of (2) , where x_t and ε_t are actual values and random errors (or random shock) at time 'i'. θ_i is the model parameter and p is known as the order of the model.

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

The reason for selecting this model in this component is that it is known to be the most successful time series model when it comes to price prediction. Here we attempt to successfully generate a model to be used in local context using the same dataset as with the above LSTM model.

The Ensemble models

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 Commercialization aspects

We believe our product will be a superior alternative for business enterprise since this is the first of its sort in Sri Lanka. This application would be valuable for real estate clients just as real state dealers. We would like to offer a free trial of the application for a month and afterward have a choice to subscribe with the service for a reasonable charge. We could have let the users to use the application freely and let them get the documents subjected to a fee but it would be suitable if the reports provided through the system are recognized as legitimate by the relevant authorities.There are various different features which can be added to the application which makes it more useful.

Precision of the system is resolved through the training/testing environment of AI model development. We will likely give the best forecasts by discovering the most precise calculations to be utilized with AI model.

We anticipate that our application should give the results within least possible time, so the clients recognize the system as instant, reliable and effective one of its sort which helps them to identify the lands. They are going to purchase. Since we are implementing a web application UI responsiveness is equally significant until an Android/IOS applications are created.

We consider about the scalability of our system to be of a similar significance as accuracy. There can be number of clients utilizing the services when the product is published.

The system complexity can influence the fee of the services provided. Be that as it may, the service would be a lot less expensive than the manual procedure since it can tradeoff the indirect expenses and effort, gathering information 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.

2.5 Tools and Technologies

2.5.1 Hardware interfaces

For the developer end, a computer with

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

2.5.2 Software interfaces

- Database

  For creating our application database, we use MongoDB.
- 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

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.

	MAE	**MSE**	**RMSE**
LSTM	12150.774	1834424960	42830.187
ARIMA	26549.4523	4559474.12	2135.29251

Table 3.1 – Model evaluation

Out of the two models ARIMA model found out to be having the best performance.

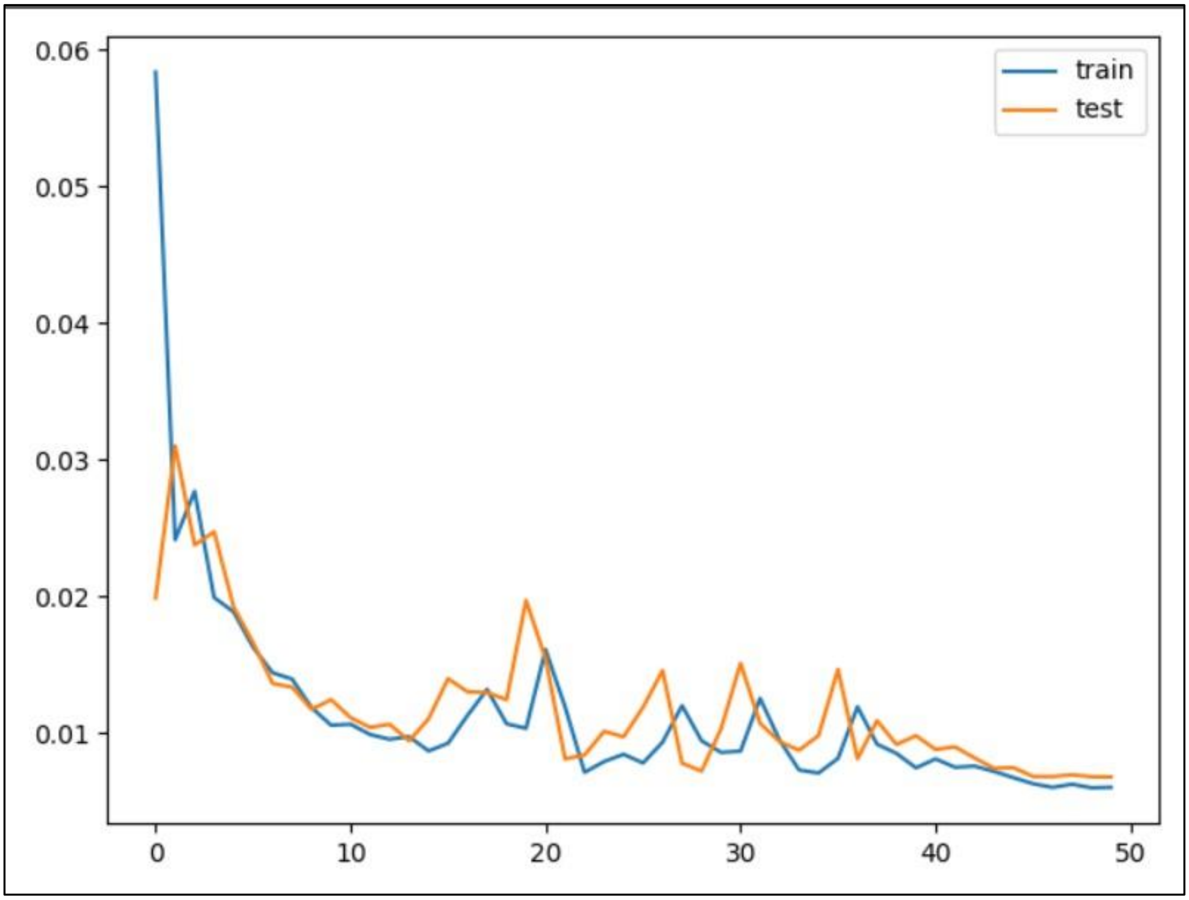


Figure 3.1 LSTM loss function

Though the error has been reduced with training process the LSTM model had a comparatively higher error. Hence, ARIMA model is suitable to predict time-series values of land prices.

Below are the PACF and ACF functions of the ARIMA model.

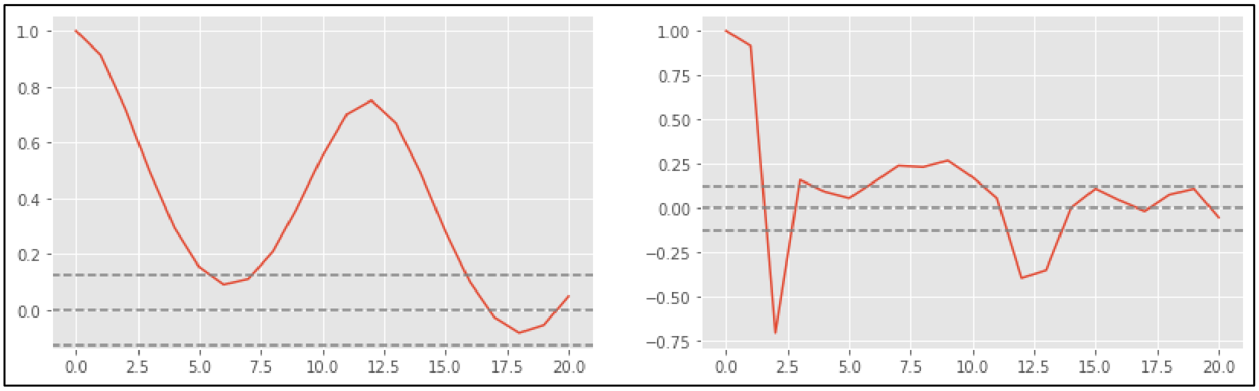


Figure 3.2 PACF and ACF functions of the ARIMA model

The dotted lines in confidence interval, this can be used to determine **p** and **q**.

- p: The lag value where the **PACF** chart crosses upper chart for first time.
- q: The lag value where **ACF** chart crosses upper chart for first time. Here p = 5, q = 1, order = (5,1,1)

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.

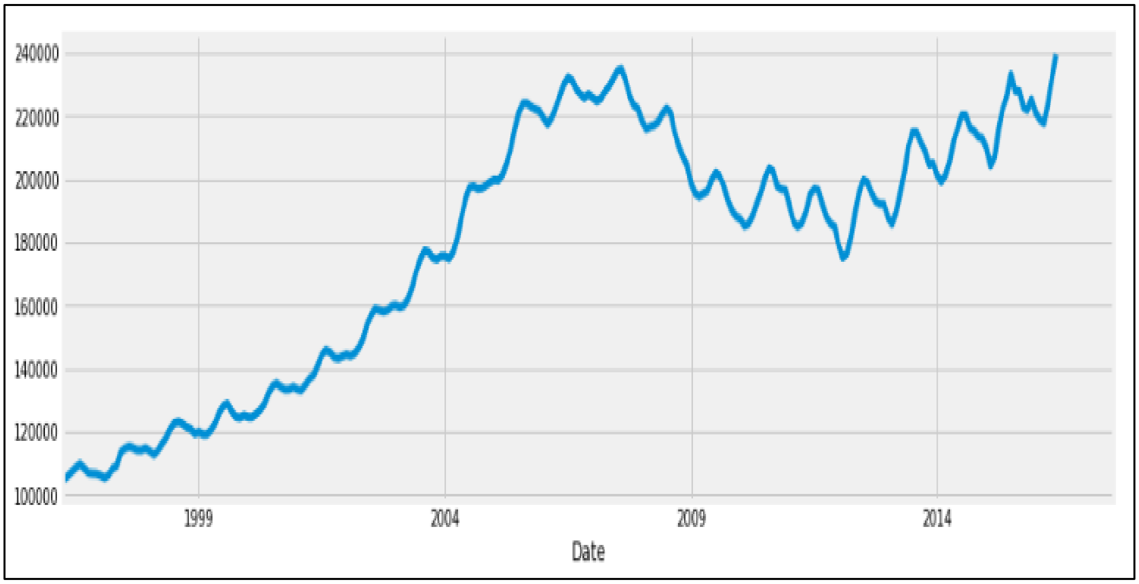


Figure 3.3 Monthly average of land prices

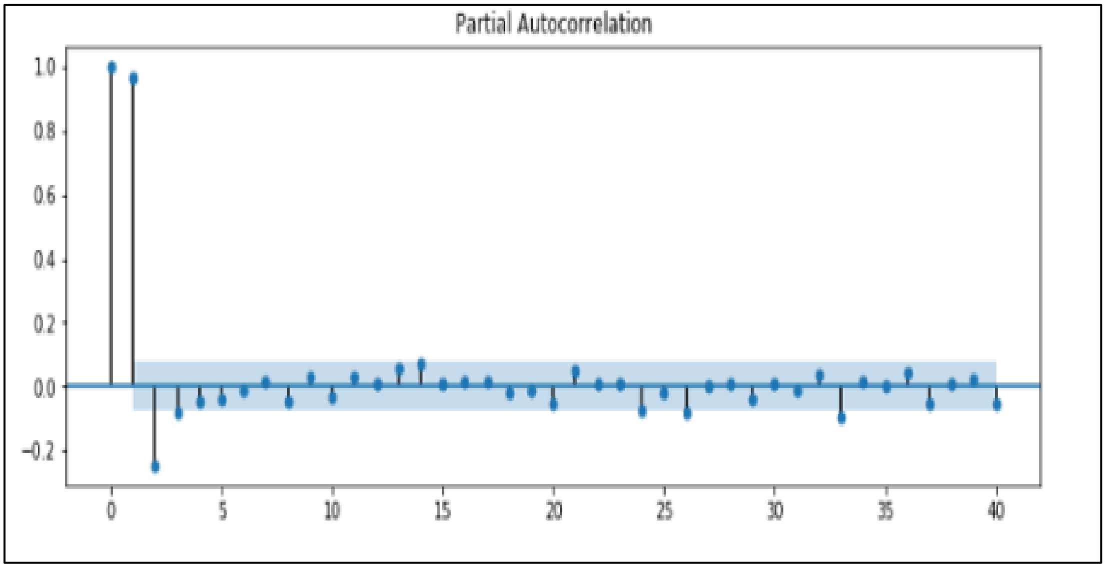


Figure 3.4 Partial Auto Correlation Function

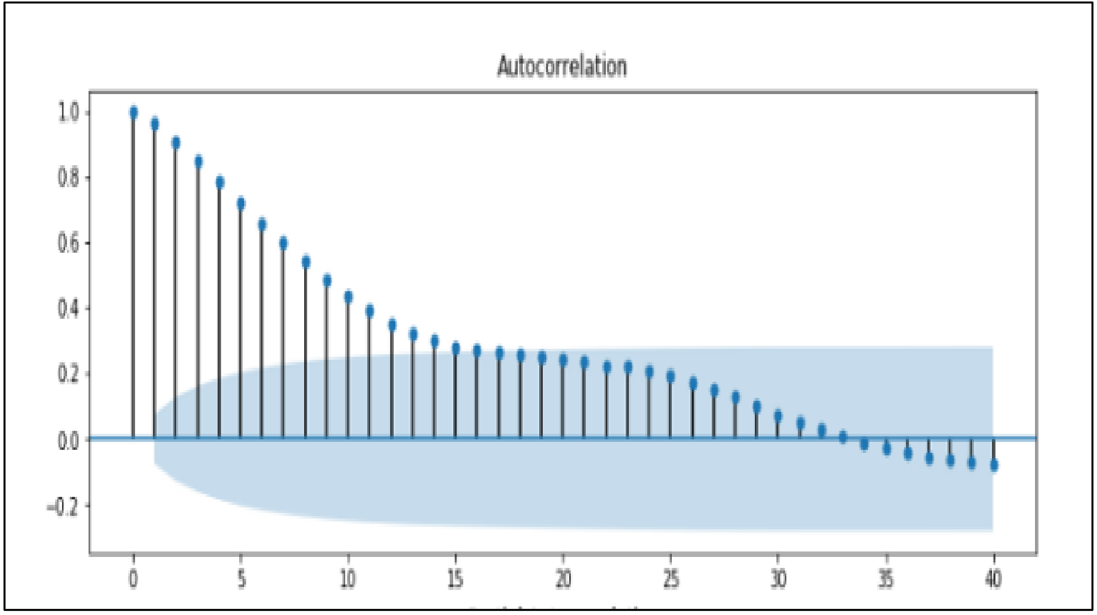


Figure 3.5 Auto Correlation Function

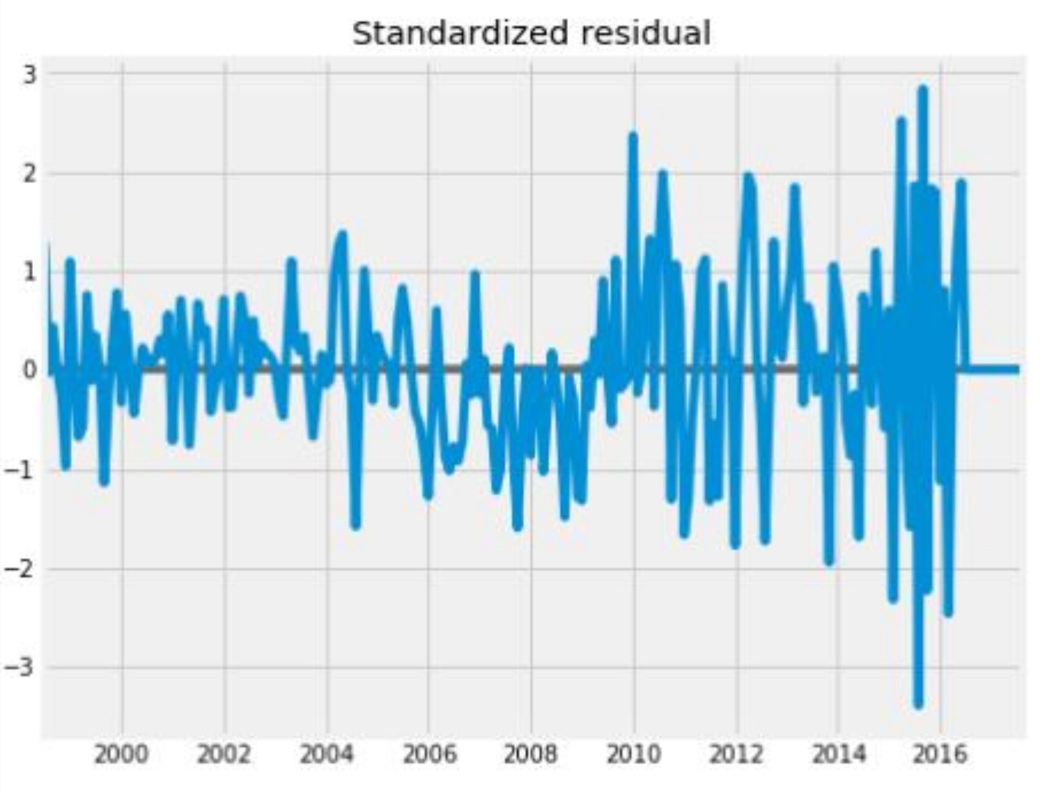


Figure 3.6 Standardized residual function

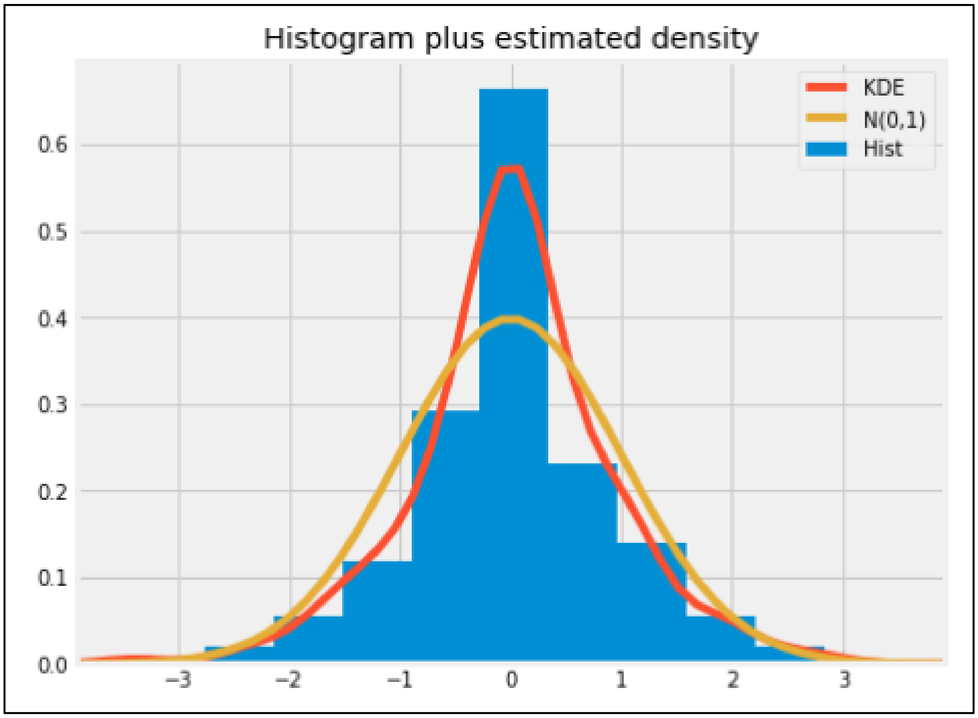


Figure 3.7 Density distribution

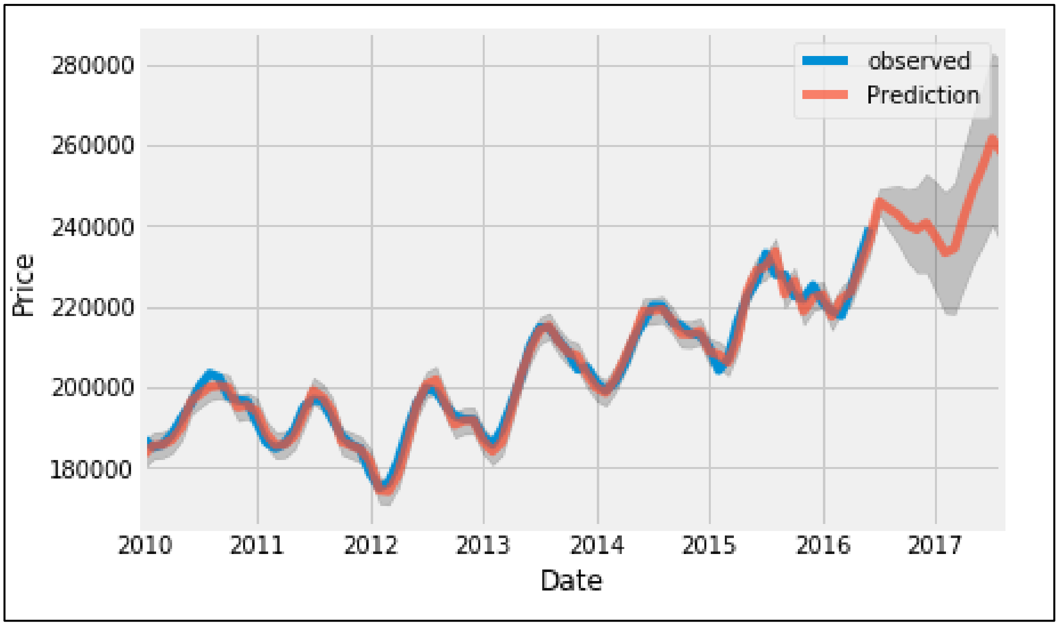


Figure 3.8 Predictions compared with observed dataset

The graph Figure 3.8 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.

As mentioned above, in 2.2.3, 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.

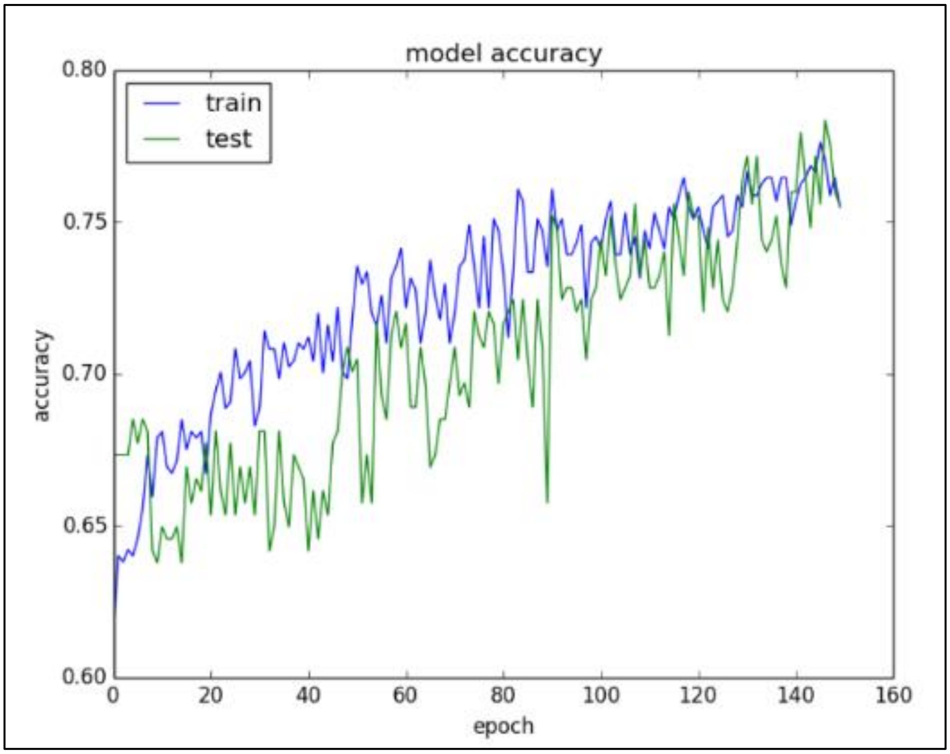
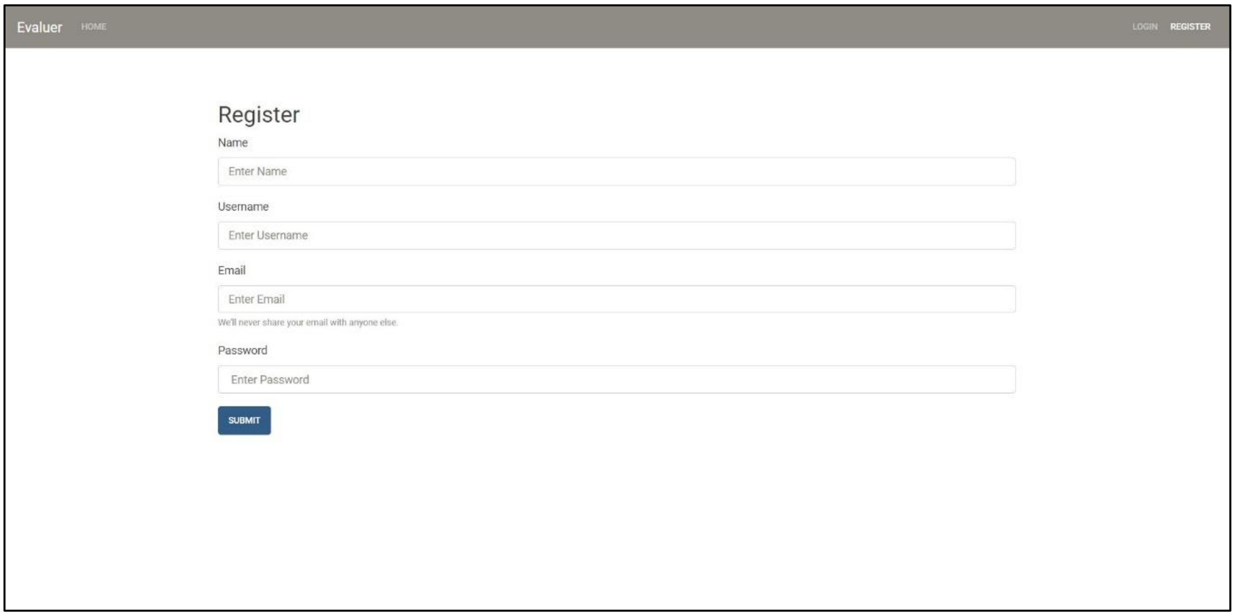


Figure 3.9: Accuracy curve for ensemble model

3.2 User Interfaces



We'll never share your email with anyone else.

Figure 3.10 Registration form

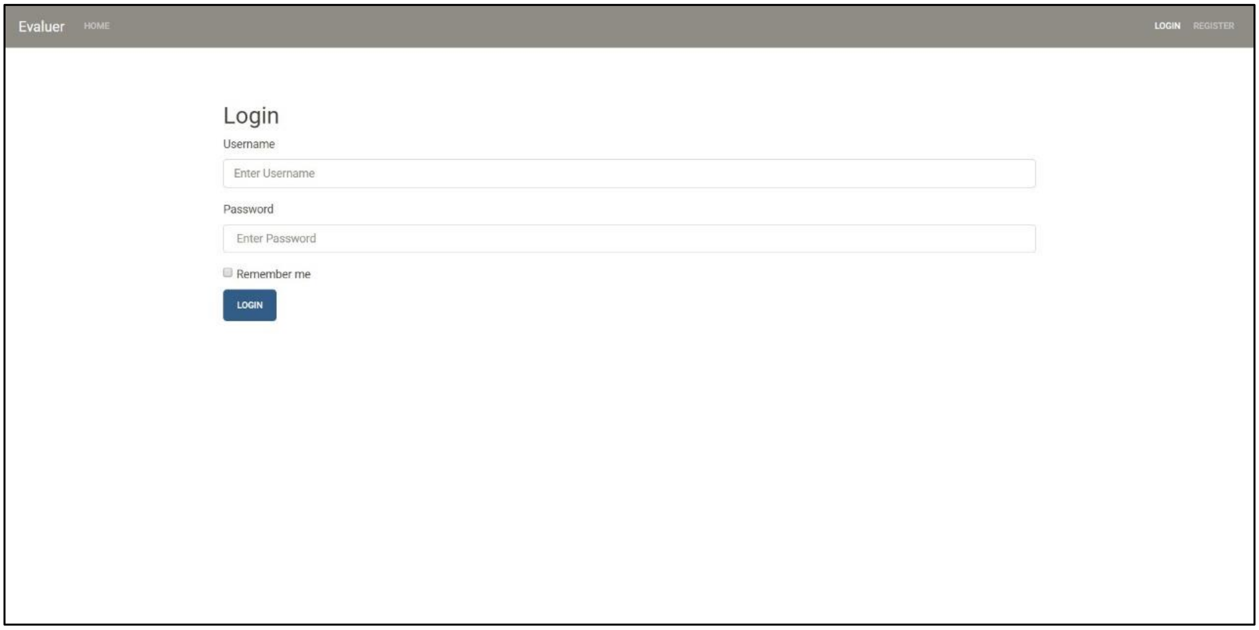
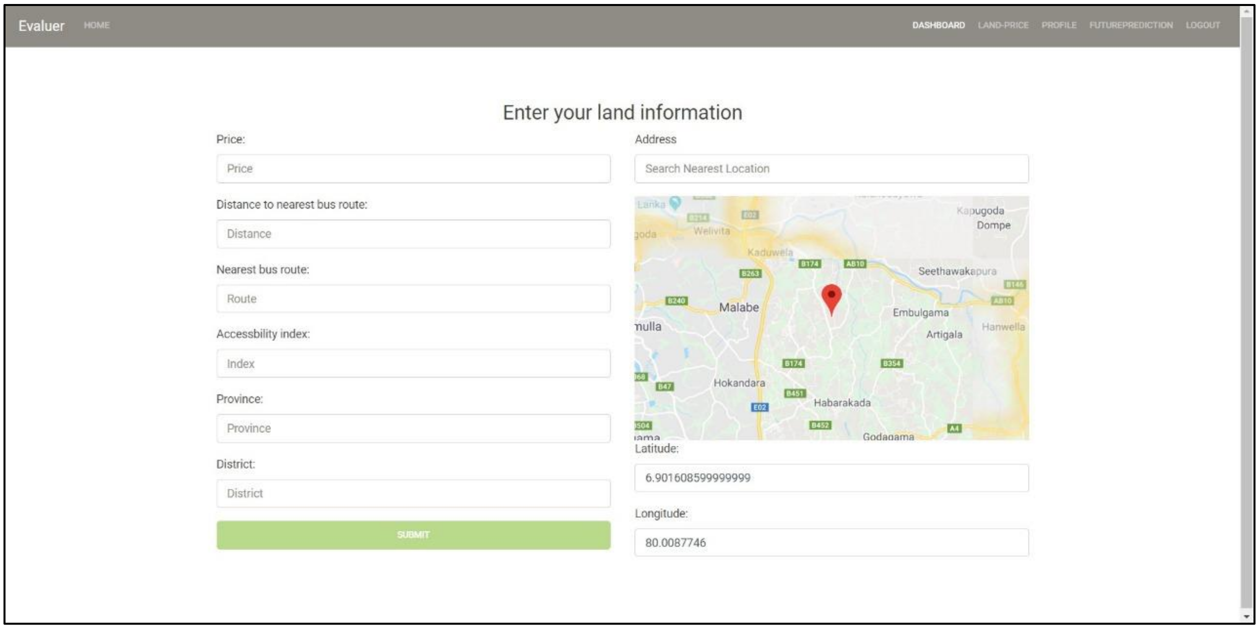


Figure 3.11 Registration form



Enter your land information

Price: Price

Distance to nearest bus route: Distance

Nearest bus route: Route

Accessbility index: Index

Province: Province

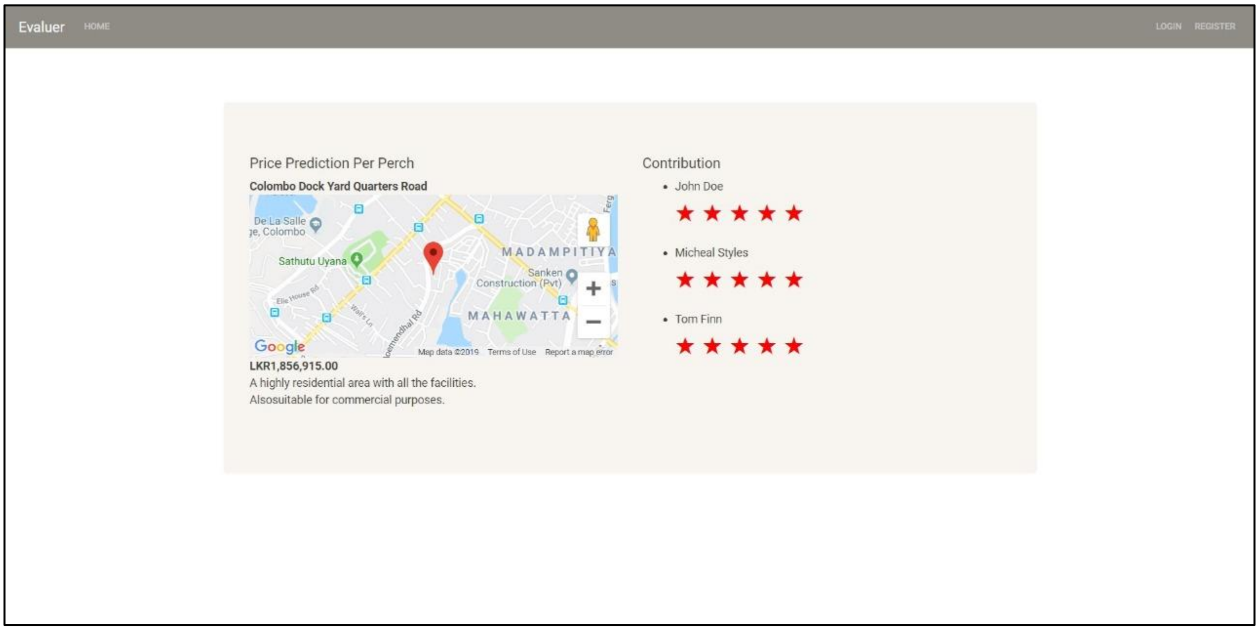
District: District

Address: Search Nearest Location

Latitude: 6.901608599999999

Longitude: 80.0087746

Figure 3.11 Data input form



Price Prediction Per Perch

Colombo Dock Yard Quarters Road

LKR1,856,915.00

A highly residential area with all the facilities.
Alsosuitable for commercial purposes.

Contribution

- John Doe ★★★★★
- Micheal Styles ★★★★★
- Tom Finn ★★★★★

Figure 3.12 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 2.3, 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] "Colombo District Land Price Index records 13.6-pct increase in 1H 2019", *Lanka Business Online*, 2019. [Online]. Available: https://www.lankabusinessonline.com/colombo-district-land-price-index-records-13-6-pct-increase-in-1h-2019/. [Accessed: 21- Sep- 2019].

[2] 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

[3] R. Ariyawansa and T. Gunawardhana, *PROPERTY STANDARDS ON PROPERTY MANAGEMENT AND PROPERTY VALUES: ANALYSIS OF COLOMBO PROPERTY MARKET, SRI LANKA*. 2019.

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

[5] I. Wilson, S. Paris, J. Ware and D. Jenkins, "Residential property price time series forecasting with neural networks", *Knowledge-Based Systems*, vol. 15, no. 5-6, pp. 335-341, 2002. Available: 10.1016/s0950-7051(01)00169-1.

[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.

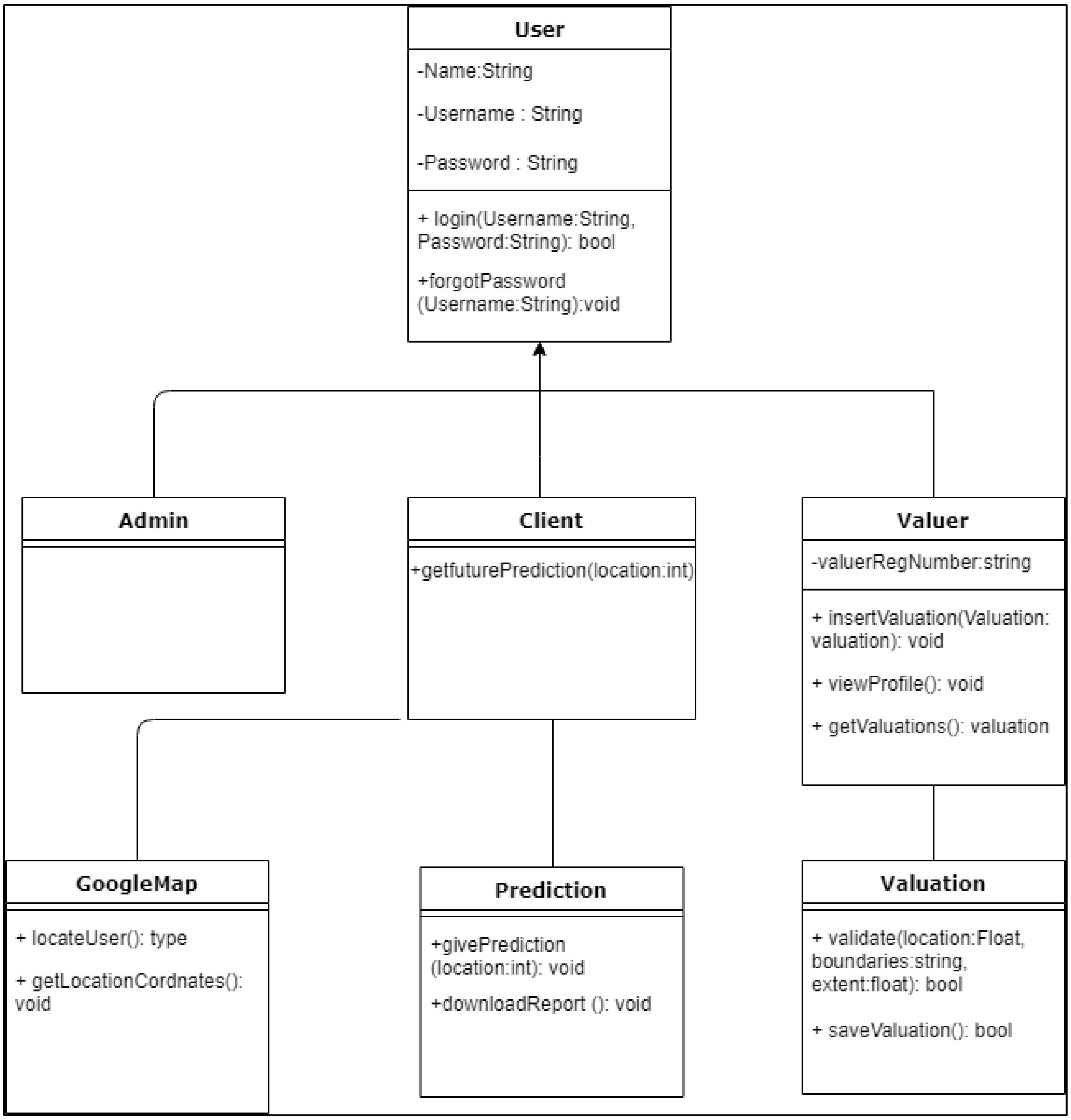
[20] "5 Easy Steps to Understanding JSON Web Tokens (JWT)", *Medium*, 2018. [Online]. Available:https://medium.com/vandium-software/5-easy-steps-to-understanding-json-web-tokens-jwt-1164c0adfcec. [Accessed: 16- Jul- 2019].

[21] Chen, X., Wei, L. and Xu, J. (n.d.). *House Price Prediction Using LSTM*. [online] arXiv.org. Available at: https://arxiv.org/abs/1709.08432 [Accessed 15 May 2019].

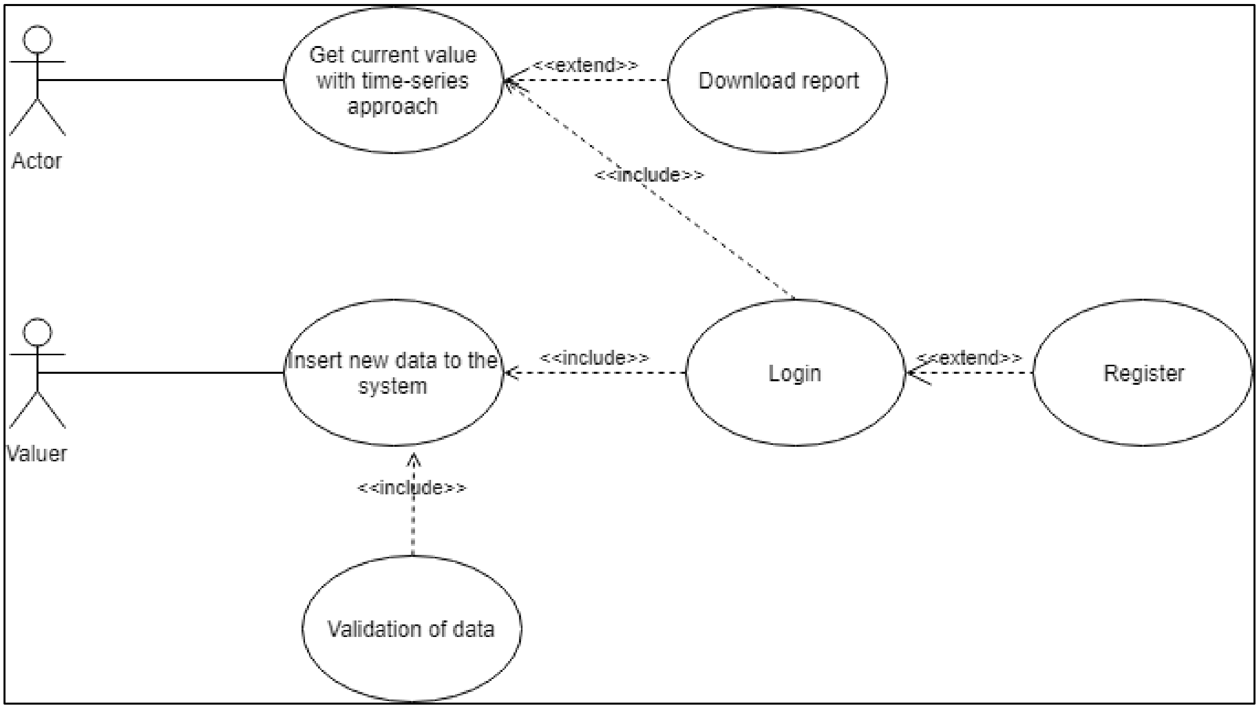
[22] Tse, R. (1997). An application of the ARIMA model to real-estate prices in Hong Kong. *Journal of Property Finance*, 8(2), pp.152-163.

APPENDIX

Appendix 1.1 Class diagram for current value prediction


User
-Name:String
-Username : String
-Password : String
+ login(Username:String, Password:String): bool
+forgotPassword (Username:String):void
Admin
Client
+getfuturePrediction(location:int)
Valuer
-valuerRegNumber:string
+ insertValuation(Valuation: valuation): void
+ viewProfile(): void
+ getValuations(): valuation
GoogleMap
+ locateUser(): type
+ getLocationCordnates(): void
Prediction
+givePrediction (location:int): void
+downloadReport (): void
Valuation
+ validate(location:Float, boundaries:string, extent:float): bool
+ saveValuation(): bool

Appendix 1.2 Use case diagram for current value prediction


Actor
Get current value with time-series approach
<<extend>>
Download report
<<include>>
Valuer
Insert new data to the system
<<include>>
Login
<<extend>>
Register
<<include>>
Validation of data