



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. (IT16423534), Rodrigo U.S.D.(IT16154990),

Thenuka Dharmaseelan (IT16116566), K.Thayalini (IT15156884)

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 (Group)

(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

We 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	
Rodrigo U.S.D.	IT16154490	
T. Dharmaseelan	IT16116566	
K. Thayalini	IT15156884	

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.

TABLE OF CONTENTS

1. INTRODUCTION	1
1.1 Background literature	1
1.2 Research Gap.....	9
1.3 Research Problem.....	10
1.4 Objectives.....	11
Main objective	11
Specific objectives	11
2. RESEARCH MEHODOLOGY	12
2.1 System Overview	12
2.2 Features	14
2.2.1 Finding the user’s Geo-location	14
2.2.2 Methods of delivering the services	15
2.2.3 Choose the service providers	16
2.2.4 User Authentication	17
2.2.5 Price prediction for a particular location	19
2.2.6 Valuer recommendation module	22
2.3 Predictions	23
2.3.1 Type of predictions	23
2.4 Methodology	23
2.4.1 Data Collection	23
2.4.2 Implementation and testing	25
2.5 Commercialization aspects of the system	33
2.6 Tools and Technologies	34
3 RESULTS AND DISCUSSIONS.....	35
3.1 Results	35
3.1.1 Current Value Dataset	35
3.1.2 Past Value Dataset	35
3.1.3 Future Infrastructure data	36
3.1.4 Comparison of models with different datasets	36

3.2 User Interfaces.....	50
3.3 Discussion	53
3.3.1 Current Value Prediction	53
3.3.2 Future Value Prediction.....	53
4 CONCLUSION.....	54
REFERENCES	55
APPENDIX.....	58

LIST OF FIGURES

Figure 1.1: Different stages of the appraisal process for estimating the market value	2
Figure 1.2: Brief Neighbourhood analysis.....	6
Figure 1.3: Historical Sales analysis	7
Figure 2.1. System Diagram	13
Figure 2.2 . High level system architecture	13
Figure 2.3 Retrieving geo-location process	14
Figure 2.4 User Interface	14
Figure 2.8 LSTM model training and testing.....	20
Figure 2.9 ARIMA model training and testing	20
Figure 2.10 User Interface for displaying current value predictions	21
Figure 2.11 Valuer recommendation process	22
Figure 2.12 Standard Neural net (left) and Neural net with dropout layers	27
Figure 2.13 Training scheme of a neuron base	27
Figure 2.14 Example for a LSTM.....	29
Figure 3.1 LSTM loss function.....	37
Figure 3.2 PACF and ACF functions of the ARIMA model	38
Figure 3.3 Monthly average of land prices	39
Figure 3.4 Partial Auto Correlation Function	39
Figure 3.5 Auto Correlation Function.....	40
Figure 3.6 Standardized residual function	40
Figure 3.7 Density distribution	41
Figure 3.8 Predictions compared with observed dataset.....	41
Figure 3.9 Plot of extent of the land, distance to main road features	42
Figure 3.10 Train accuracy (in red) and test accuracy (in blue) curves of ANN model.....	44
Figure 3.11 : Accuracy curve for ensemble model	45
Figure 3.13 ARIMA model dicky fuller test results	46
Figure 3.14 RMSE versus K values.....	46
Figure 3.17 Loss curve for neural network.....	49
Figure 3.18 Registration.....	50
Figure 3.19 Login.....	50
Figure 3.20 Data input screen for valuers	51
Figure 3.22 Future value prediction.....	52

LIST OF TABLES

Table 1.1: Average absolute percentage error of the 12-month forecast.....	3
Table 1.2: Trulia accuracy report.....	4
Table 1.3: Comparison of existing systems	8
Table 3.1 Price prediction model evaluation	36
Table 3.2 Comparison of MLR and ARIMA performance in future value prediction	47

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 since the land prices are going up day-by-day. 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.

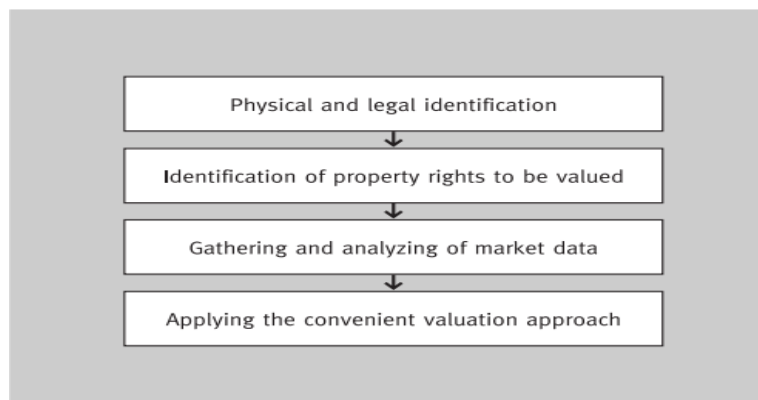


Figure 1.1: Different stages of the appraisal process for estimating the market value

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

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 a accurate, fair prediction of the value of the land.

Our training environment determines the best out of time series algorithms and vice versa to do the most accurate current value 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 Artificial Neural Network (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 Multiple Regression Analysis (MRA) but its black box nature has led to reach different conclusions in the observations.

Initial step is to implement an algorithm to create a suitable model preceding by the development of model using algorithms to forecast the future commercial value of the land selected, by comparing the past values and feature values of the land market. Then the percentage fluctuation of past value and feature commercial value is found out and graph to those comparisons using statics will be generated. The K-NN algorithm is one of the methods used for classification analysis, but the last few decades the KNN method has also been used for prediction of continuous values[7].

Existing solutions

The use of Artificial Intelligence (AI) for residential value forecasting has been suggested in the literature from 1990s. [8]. 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. [9] 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: Average absolute percentage error of the 12-month forecast.

Source: <https://www.zillow.com/research/zestimate-forecast-methodology/>

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 [9] which analyses of several features of the house. The forecasted value is interpolated using cubic spline to connect to current value. [9]

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

Source:/www.trulia.com/info/trulia-estimates/

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.[10]

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 [11] 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. [12]

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 [13]. 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.

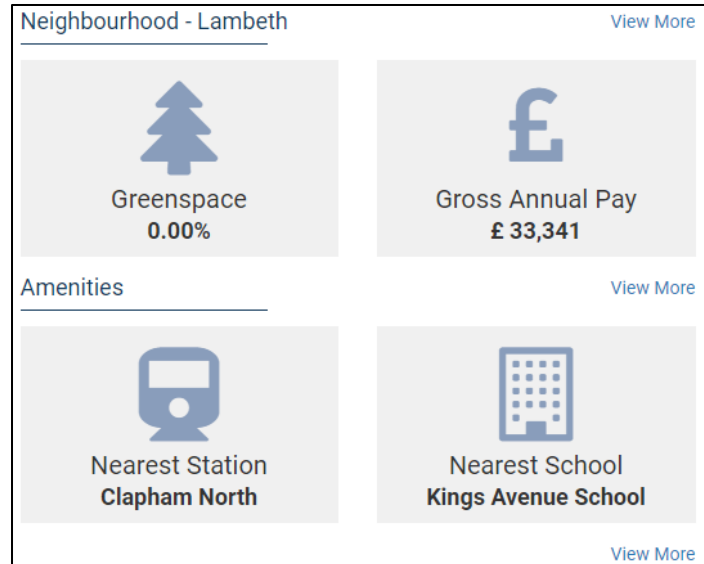


Figure 1.2: Brief Neighbourhood analysis

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

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

- Historical data relevant to location

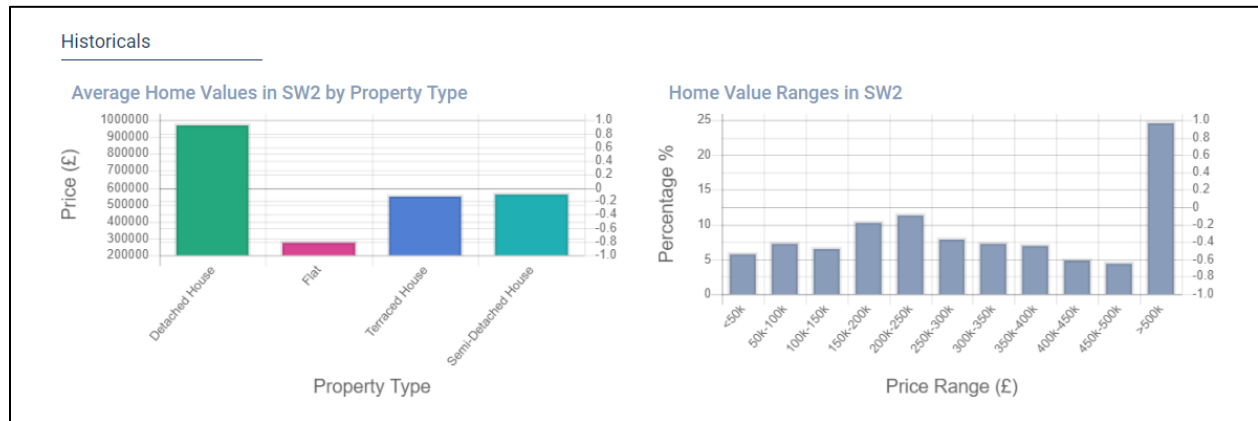


Figure 1.3: Historical Sales analysis

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
Use of AI/ Machine Learning	Yes	Unknown	No	Yes	Yes
Future Value forecasting	For 12 months	No	Yes, with E-Valuer report	For 3 years	For 5 years
Future Development effects prediction	No	No	Yes	No	Yes
Available for Sri Lanka	No	No	No	No	Yes

Table 1.3: Comparison of existing systems

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.

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*[16].According Li *et al*, [17]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.[18], 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) as 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 methods to optimize the values provided as current value prediction so that the procedure could be used as a reliable option.
- 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 MEHODOLOGY

2.1 System Overview

E-Valuer a land value prediction system which will be assist the users in making smart property related decisions. Mention will be made throughout this document of selected probable features of future releases. The purpose of this is to guide developers in selecting a design that will be able to accommodate the full-scale application. It provides the clear overall design of the system which will provide the foundation for the development of the final end product.

The purpose of this application is to provide its users with accurate prediction of the value of the land which is similar to an estimation given by a professional valuer. Users can get the detailed report of the land with minimal effort needed which includes current value prediction based on the correct analyzing of the significance of the properties that can be computationally analyzed to help users for deciding the return of investment (ROI) and suitability of the land for their intended purposes.

Our system uses model to do the estimations and forecasts. We design the application as much as inexpensive and cost effective to the user, simplified user interfaces, with accurate results which provides a concise but complete report of the analysis of the land.

We provide all the estimations and predictions for a fair price which is beneficial in terms of time and cost over the manual method of valuation.

User Interface should be appealing to the users despite of their educational or social level for an application to be useful. We believe our application will provide services to all categories of users, mostly the buyers, then surveyors, valuation officers etc. We can provide a simple user interface to input the location of the land and ultimately produce the report which includes current estimation, future value prediction, and details of possible development projects of the area which can be referred and understood by almost anyone.

Below figure 2.1 depicts the System Diagram while figure 2.2 depicts the high

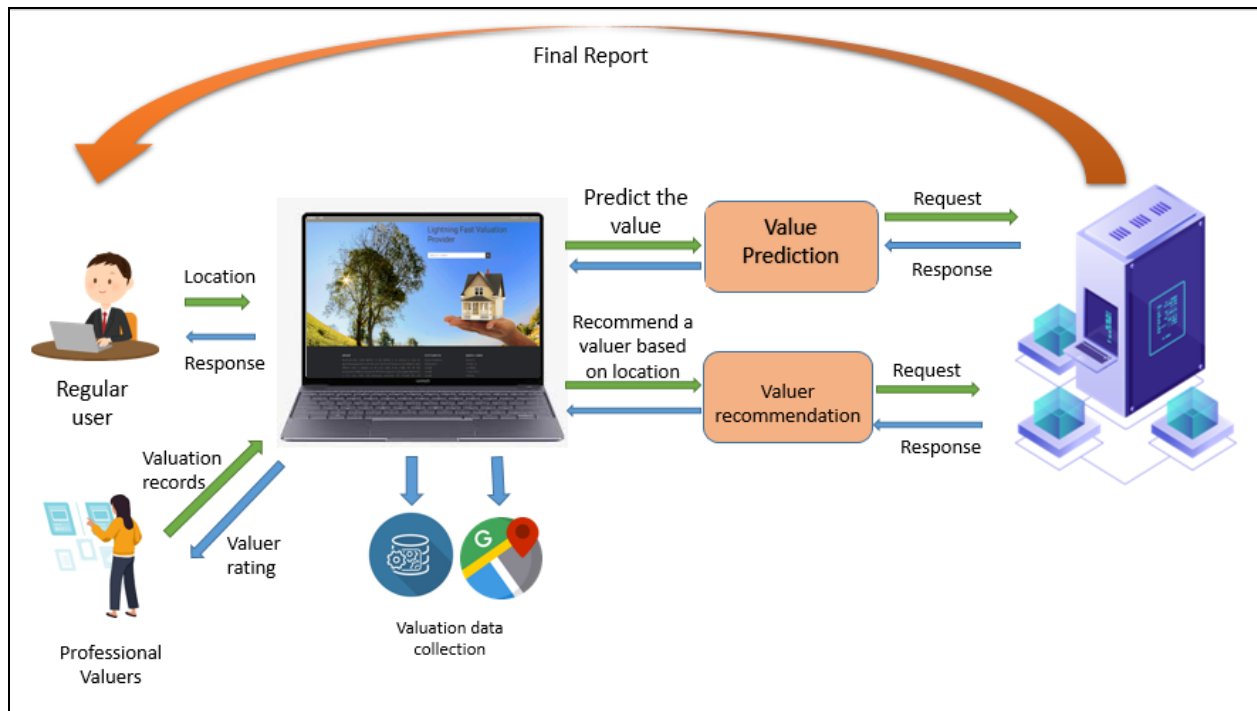


Figure 2.1. System Diagram

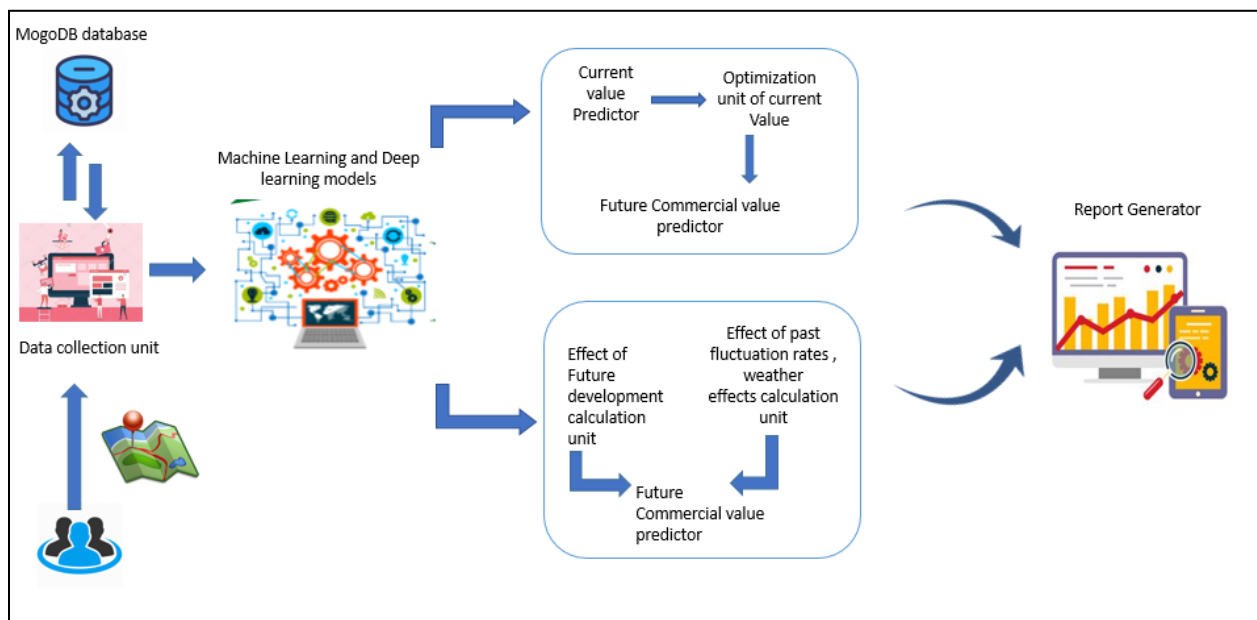


Figure 2.2 . High level system architecture

2.2 Features

2.2.1 Finding the user's Geo-location

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 .

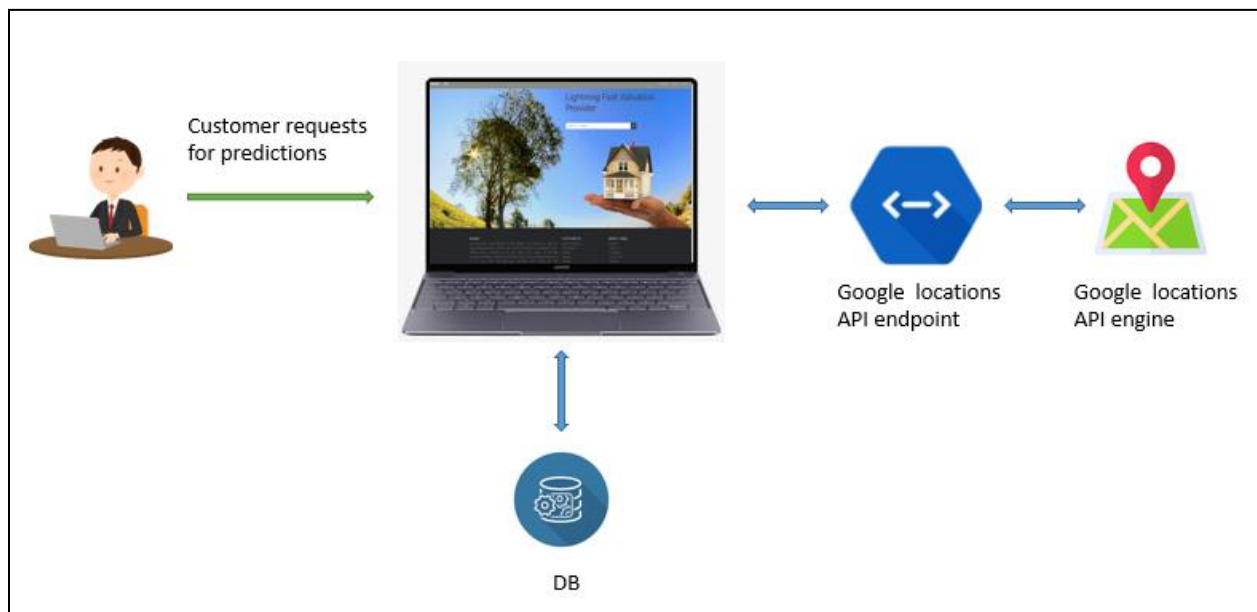


Figure 2.3 Retrieving geo-location process

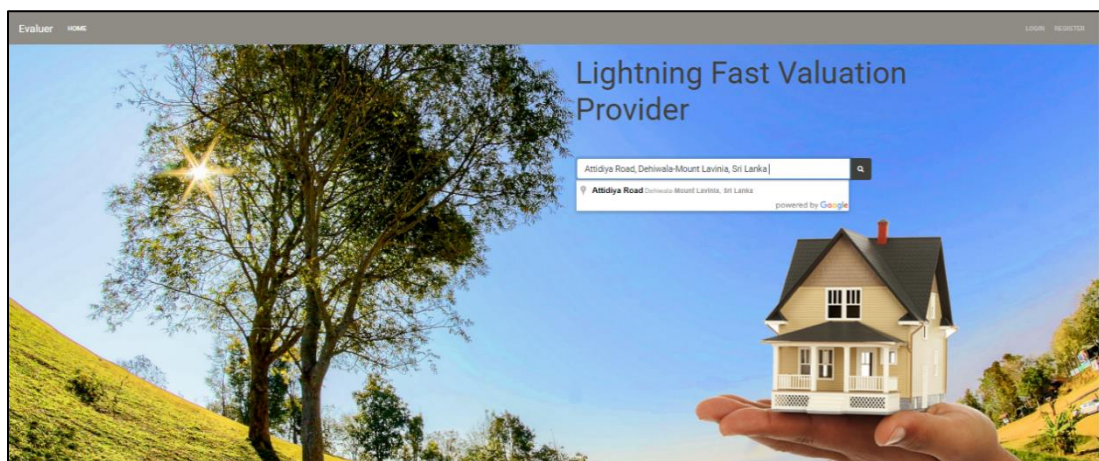


Figure 2.4 User Interface

2.2.2 Methods of delivering the services

Application delivery refers to the pool of services that combine to provide application functionality, usually web-based software applications, from the databases where the required data processing and computing is executed to be utilized by clients or end-users with the use of an API. The services for delivering applications on a network infrastructure aim to provide a reliable user experience by providing, load balancing, security, latency and TCP optimizations which combine to provide application content seamlessly.

When we get the services from Evaluer, it can deliver the services using 2 methods.

1. Current price prediction for particular area
2. Future value prediction

The two types of services delivered through this application are Current price prediction for particular area and Future value prediction. The usage of geo location of the user useful because it is flexible and allows the customer to interact in a setting that is often most comfortable for them. This is a portable application which can provide the users with an insight of the land with future value, just giving the location. This website can help people to get an idea about the land they are hoping to buy, to decide the suitability of their purpose.

2.2.3 Choose the service providers

Our application is mainly targeted for real-estate customers and owners of real estate companies. But any person can use this application. This intelligent tool can help people to identify the land they are willing to buy.

The forecast of the current value is the first service offered to the customer to decide if this plot of land is suitable for their needs. The terrain, because it is difficult to know manually the reliable average prices of the land that surrounds it, especially if the region is unknown.

Our second customer service provides forecasts of the future value of the land, as well as a report that indicates the effect of the development projects proposed in the area.

This system would be of great help in making better property decisions, which is a huge investment and a very important decision in everyone's life.

2.2.4 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.

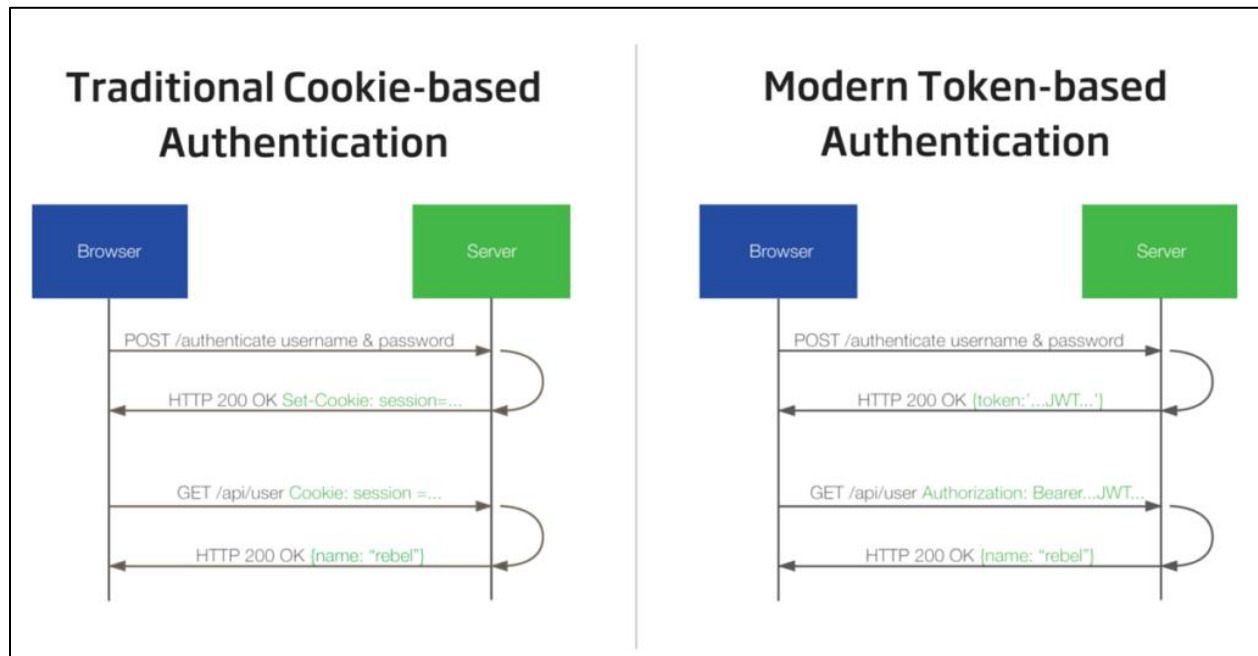


Figure 2.5 Cookie based authentication vs Token-based authentication

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. [19] 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.[19]



The screenshot shows a web application interface. At the top, there is a dark grey header bar. On the left side of the header, the text 'Evaluator' is followed by a link 'HOME'. On the right side, there are links 'LOGIN' and 'REGISTER'. The main body of the page is white. In the center, there is a 'Login' section. It starts with the word 'Login' in a bold font. Below it is the label 'Username' followed by a text input field containing the placeholder text 'Enter Username'. Underneath that is the label 'Password' followed by a text input field containing the placeholder text 'Enter Password'. Below the password field is a checkbox with the label 'Remember me'. At the bottom of the login section is a blue button with the text 'LOGIN' in white capital letters.

Figure 2.6 Login screen

2.2.5 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.

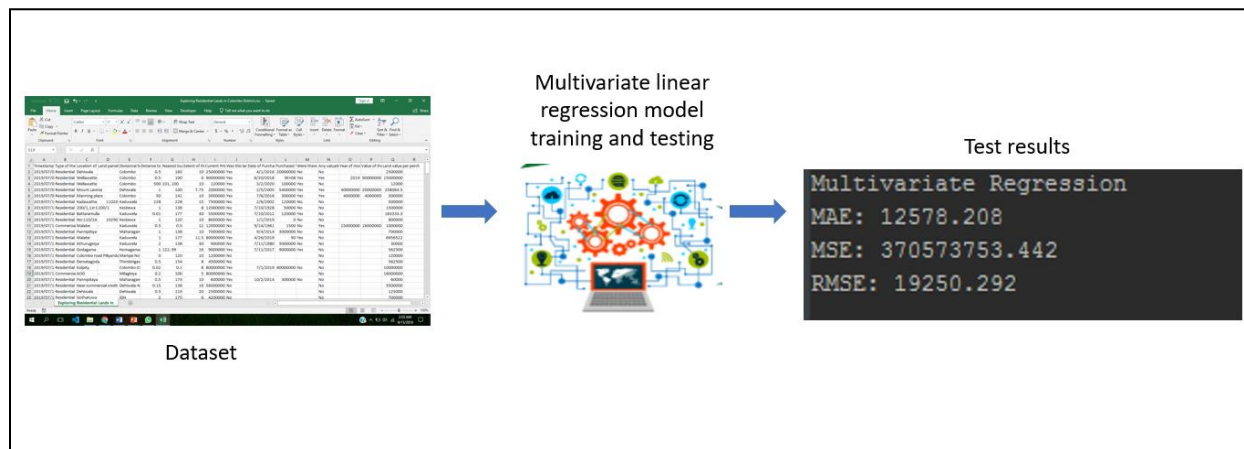


Figure 2.6 MLR training and testing

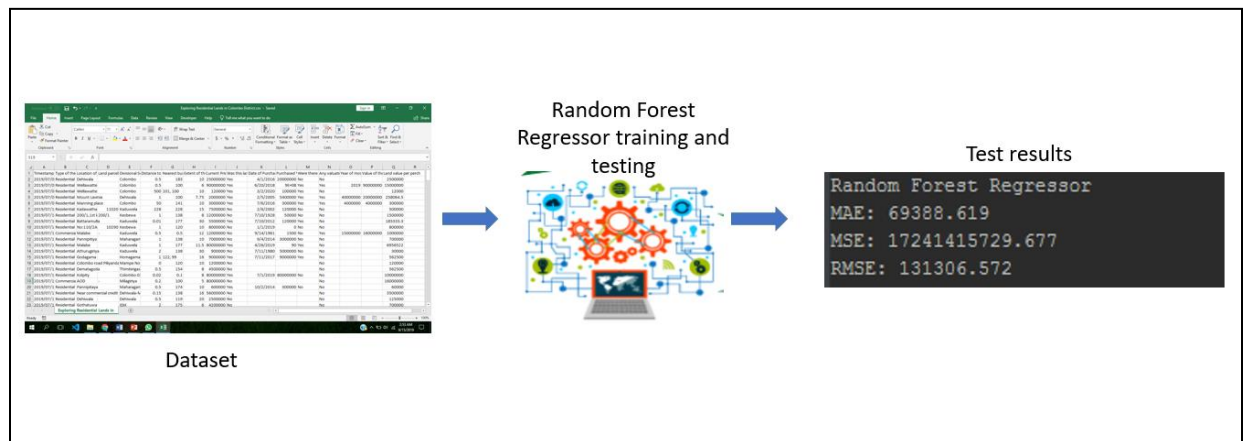


Figure 2.7 Random Forest Regressor training and testing

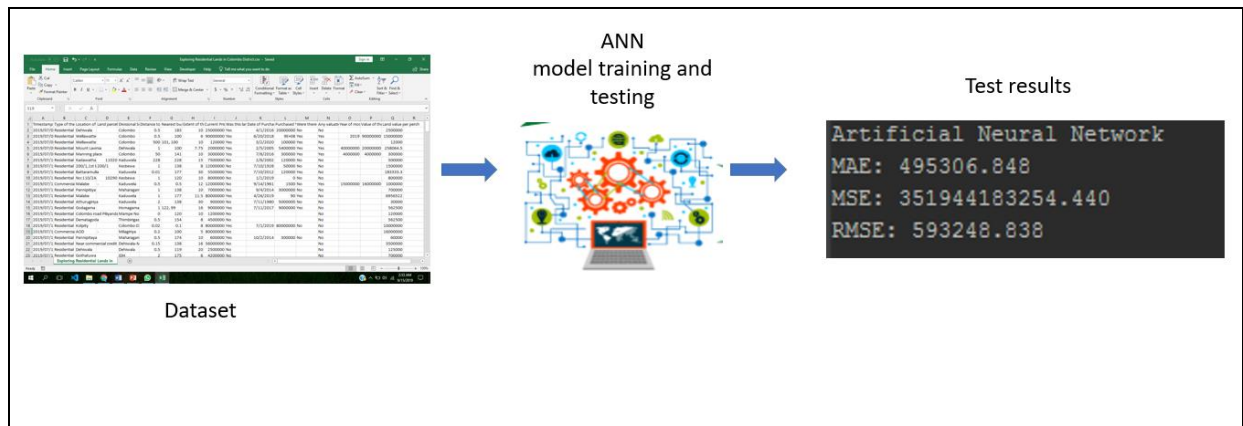


Figure 2.8 ANN model training and testing

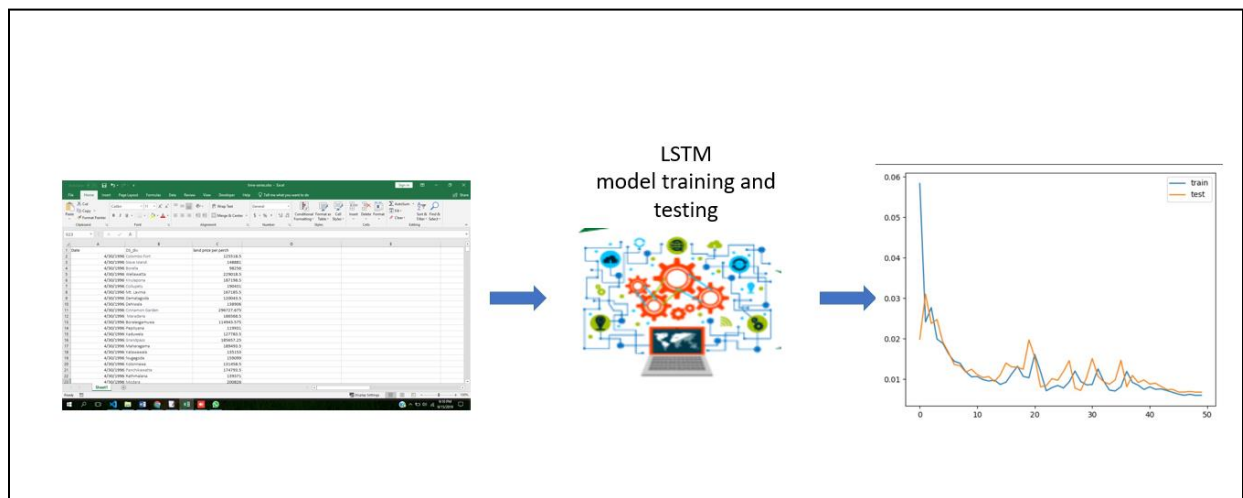


Figure 2.8 LSTM model training and testing

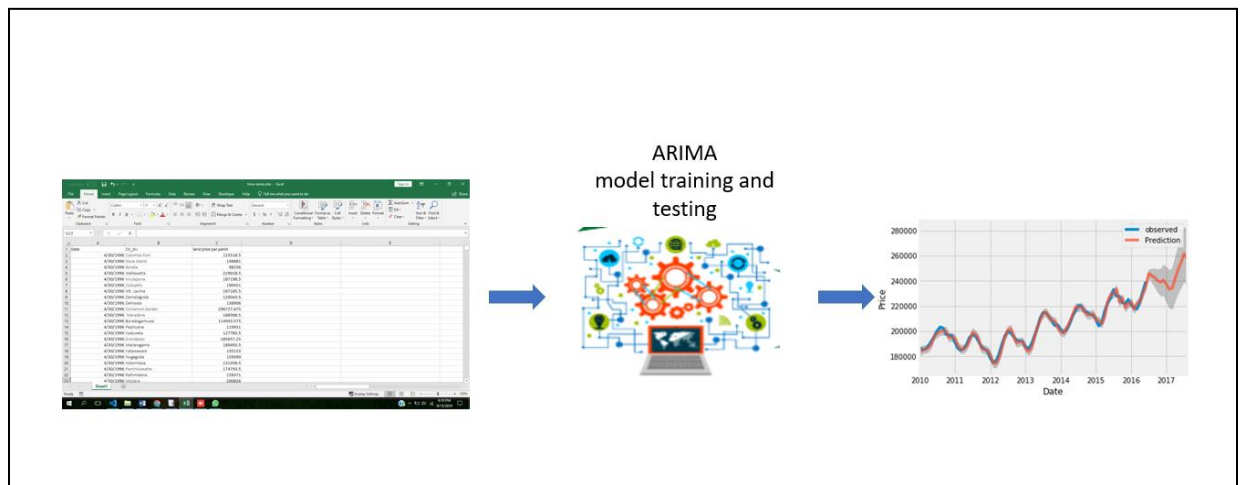


Figure 2.9 ARIMA model training and testing

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.

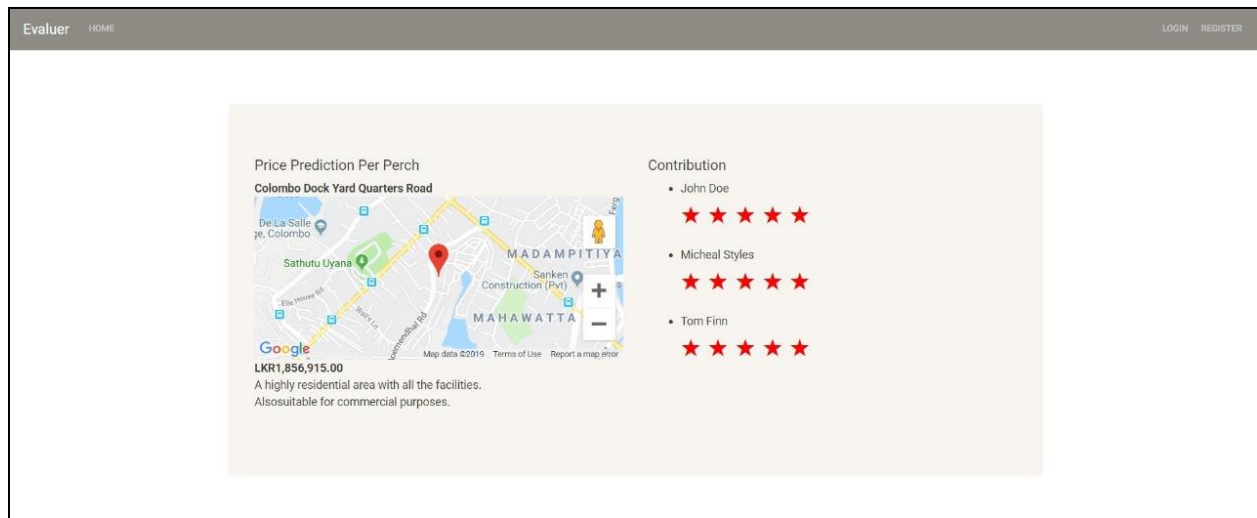


Figure 2.10 User Interface for displaying current value predictions

2.2.6 Valuer recommendation module

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.

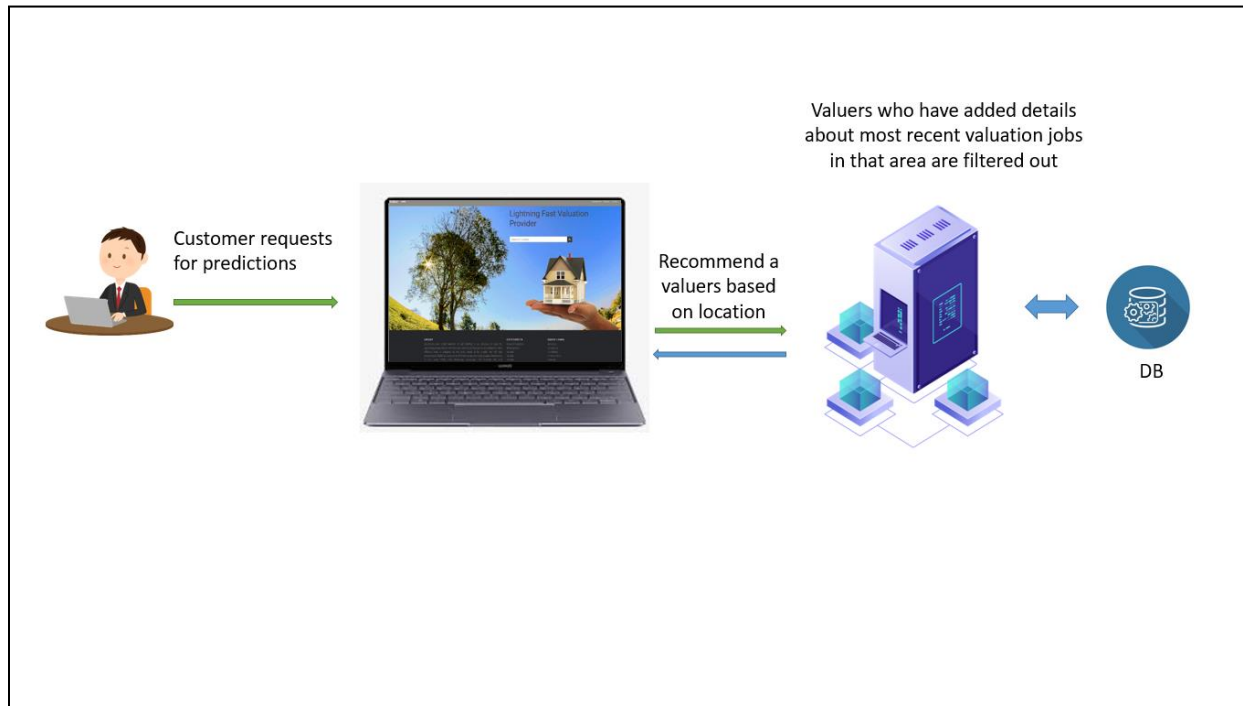


Figure 2.11 Valuer recommendation process

2.3 Predictions

2.3.1 Type of predictions

Current Value prediction

This component is expected to predict current price, which is a continuous value according to the given location. Since the location is the only feature input by user in the web application, the rest of the features contributing to the prediction should be imputed based on the location and other publicly available data with relevant to the location using the third-party API used. The flask model of the most accurate algorithm identified during the testing phase will be used here to make the prediction using location and imputed feature values.

Future value prediction

This component is expected to predict the future value, which is also a continuous value. The future value is predicted analyzing the past trend using several machine learning algorithms as well as in terms of future infrastructure development i.e., the percentage increase of the land value with the addition of new infrastructure such as schools, shopping complexes etc.

2.4 Methodology

2.4.1 Data Collection

The study focuses on Colombo district which experiences relatively high infrastructure development. For training and testing purposes, primary data has been collected through questionnaires, interviews and personal visits to land area to know the present situation of the market and the secondary data is collected mainly through various survey department, real estate agents, newspaper advertisements, and land sale website contents.

But the accuracy and suitability of this data seem to be more deviated from a valuer's point of view. Therefore, a data collection unit will be implemented to get data from the professional valuers across the country.

The cross-sectional data collected for current price prediction to be used with non-time-series algorithm were 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.

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 suggested a model of order as below.

$$Y = b_0 + b_1 (AOP) + b_2 (LOP) + b_3(EOL) + b_4(DTMR) \quad (1)$$

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 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.

2.4.2 Implementation and testing

The main components of the system are the machine learning and deep learning models used to make predictions. For the purpose of selecting the most appropriate model, below algorithms were tested on the data collected.

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.

MLR used for future value prediction unit analyzed relationship between land price and latest other factors such as state, city, zip code, price, pollution index, hospital distance, tourist score, bank/ATM, school distance, distance to town, population index and weather conditions.

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.

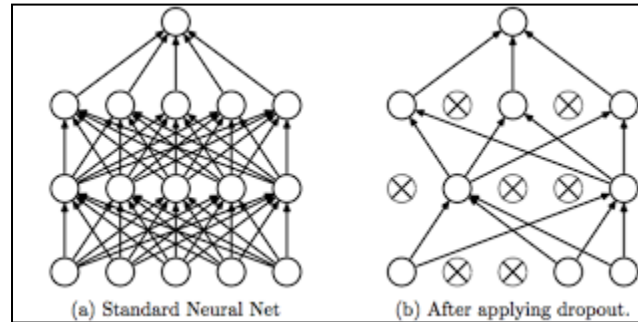


Figure 2.12 Standard Neural net (left) and Neural net with dropout layers

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.

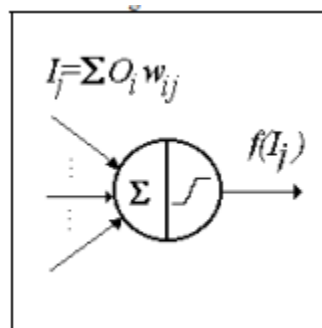


Figure 2.13 Training scheme of a neuron base

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.

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 [25]. This system is solely based on time series since the data set contains only the divisional secretariat wise prices from April 1996.

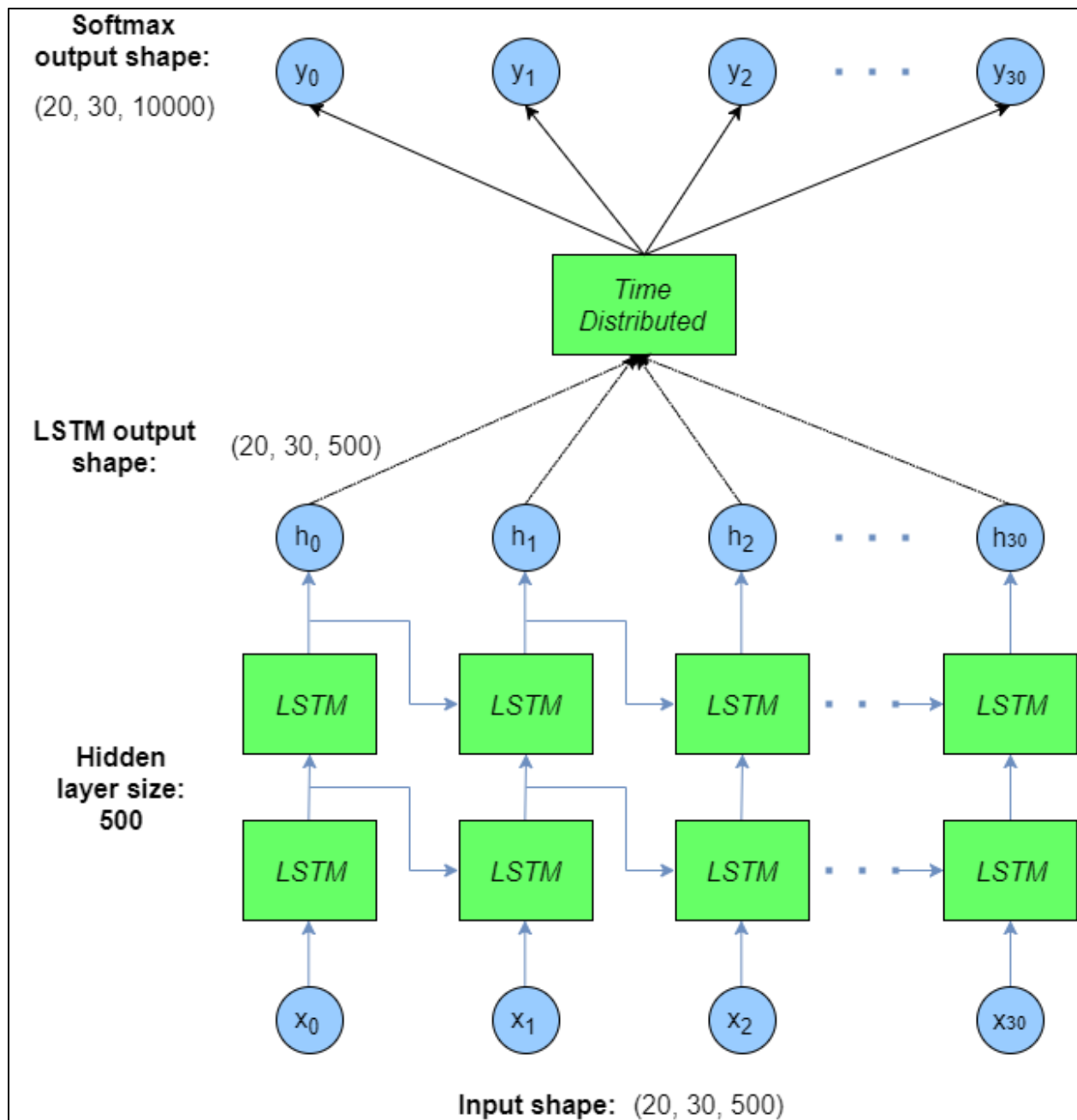


Figure 2.14 Example for a LSTM

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 [26].

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.

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

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' respectively. θ_i is the model parameter and p is known as the order of the model.

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

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.

KNN algorithm

KNN can be used for both classification and regression problems. The algorithm uses 'feature similarity' to predict values of any new data points.

This model was used to predict future values with categorized data being used as training sets and test sets (40% allocate for test data). First, initialize the model. After that fit the train data to store RMSE values for different k values, fit the model. Then make the prediction on test data set and calculate the RMSE value and store the RMSE value.[27]

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

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 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 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.

2.5 Commercialization aspects of the system

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.

2.6 Tools and Technologies

For the developer end, we used a computer with

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

Developer End:

- Python (Numpy, Pandas, Matplotlib, scikit-learn)
- Anaconda distribution,
- Angular 6
- Google API

3 RESULTS AND DISCUSSIONS

3.1 Results

Comparison of RMSE, MSE and loss MAE of the ANN, LSTM for current value prediction and MLR, ARIMA, Random forest regressor pre-trained architectural models are presented in this section. To do the training and testing processes, we consider current value, past value, future infrastructure-based value datasets. All the models were trained and validated on the same training and test dataset and using the same parameters to compare results. Each of these architectures takes various time periods per epoch to train and test particular data sets. All these models mentioned above were processed with a 12GB RAM and Intel Core i7 2.50 GHz processor in Microsoft Windows 10 operating system environment.

3.1.1 Current Value Dataset

The cross-sectional data collected for current price prediction to be used with non-time-series algorithm collected through a questionnaire where residents in Colombo district responded and by means of including publicly available data in newspaper and website advertisements. The questionnaire mainly asked for price of the land, location of the land, nearest bus route, and distance to the nearest bus route, along with the buying price and details of valuation history with above 200 samples (Appendix 1). The time series data collected to predict the current value from a land sale company which had monthly land values from the same area over a period of 20 years, containing above 200 samples.

3.1.2 Past Value Dataset

Dataset used for future value prediction has Land price in places in Colombo district from 20012 to 2018 which the algorithm is going to predict the future price for 10 years with features state, city, zip code, price, pollution index, hospital distance, tourist score, bank/ATM, distance to school, distance to town, population index, bid date having above 500 samples. Each places has 8 or more than 8 samples.

3.1.3 Future Infrastructure data

Primary data has been collected through interviews and personal visits to land area to know the present situation of the market and the secondary data is collected mainly through various survey department, land estate agents. The data is usefully for assessing the performance of property as a key to future land price.

Here, Score Model segment that attempts to anticipate the test information. The train model gives a prescient calculation that a score model employments. The Scored Labels segment characterizes the expectation of sales amount. For the straight relapse, we will take a gander at "Coefficient of Determination". This worth educates us about the exactness regarding the model and can change somewhere in the range of 0 and 1. In the event that the worth is close 0.8 or 1, the straight relapse model is dependable.

3.1.4 Comparison of models with different datasets

Current value prediction

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

Test results for these models can be summarized as follows in table 3.1.

	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
LSTM	12150.774	1834424960	42830.187
ARIMA	26549.4523	4559474.12	2135.29251

Table 3.1 Price prediction model evaluation

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

As mentioned above, the models tested have been evaluated in terms of MAE, MSE, and RMSE. The results can be summarized as in tables .

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

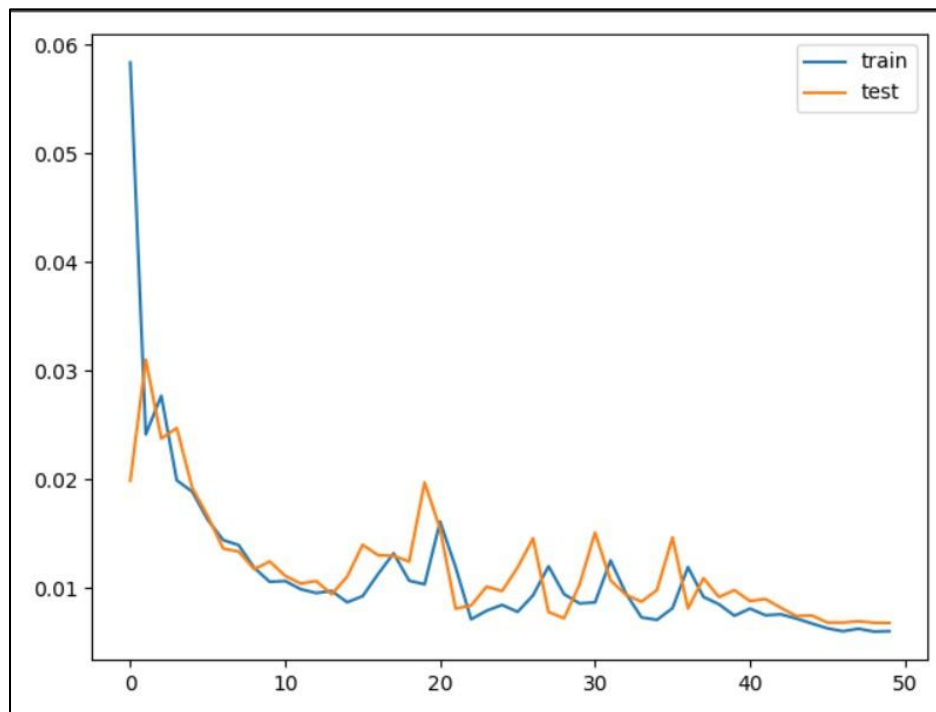


Figure 3.1 LSTM loss function

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.

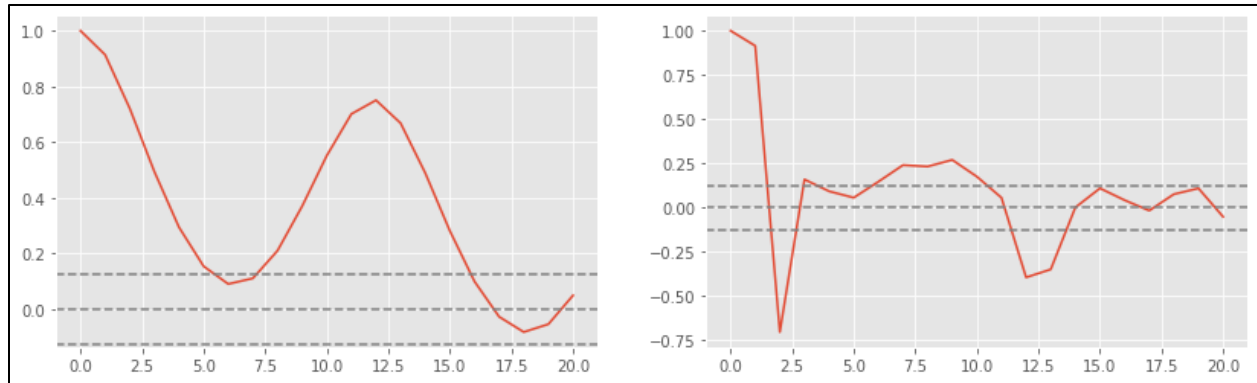


Figure 3.2 PACF and ACF functions of the ARIMA model

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. 3.3 below. The correlogram in Fig. 3.4 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. 3.5 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. 3.6, we can observe that most of the data are distributed around zero. The density graph Fig. 3.7, also displays a normal distribution.

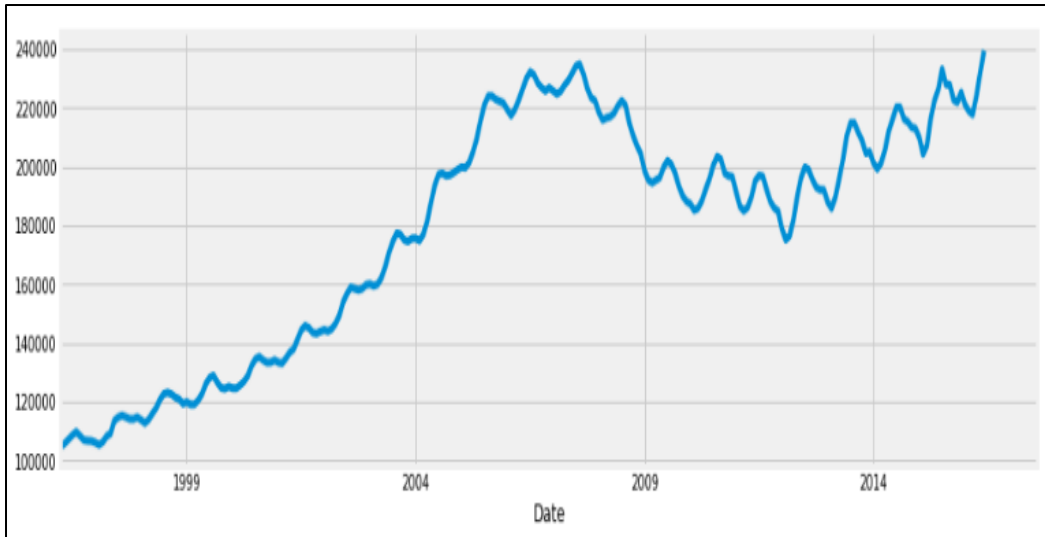


Figure 3.3 Monthly average of land prices

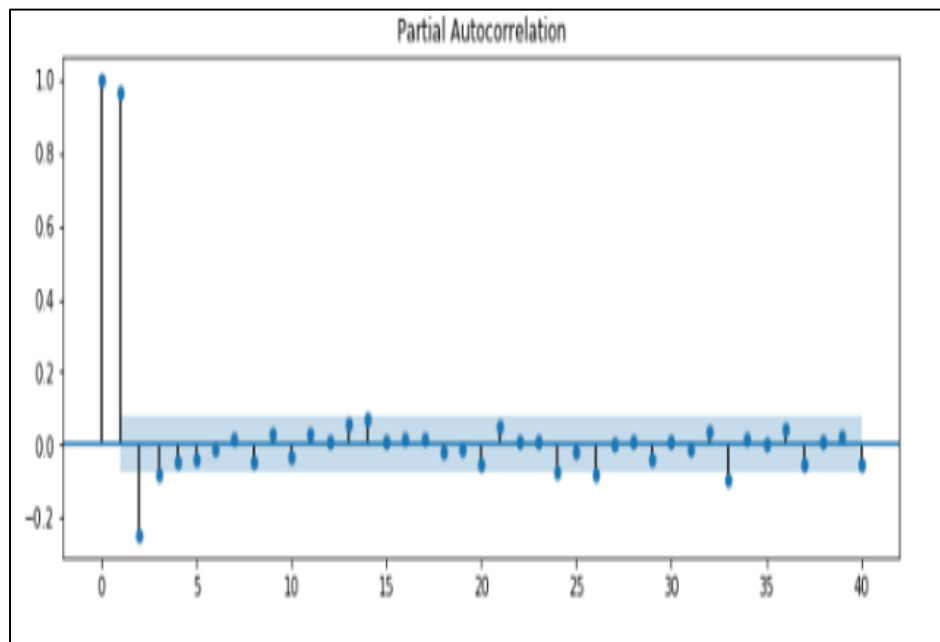


Figure 3.4 Partial Auto Correlation Function

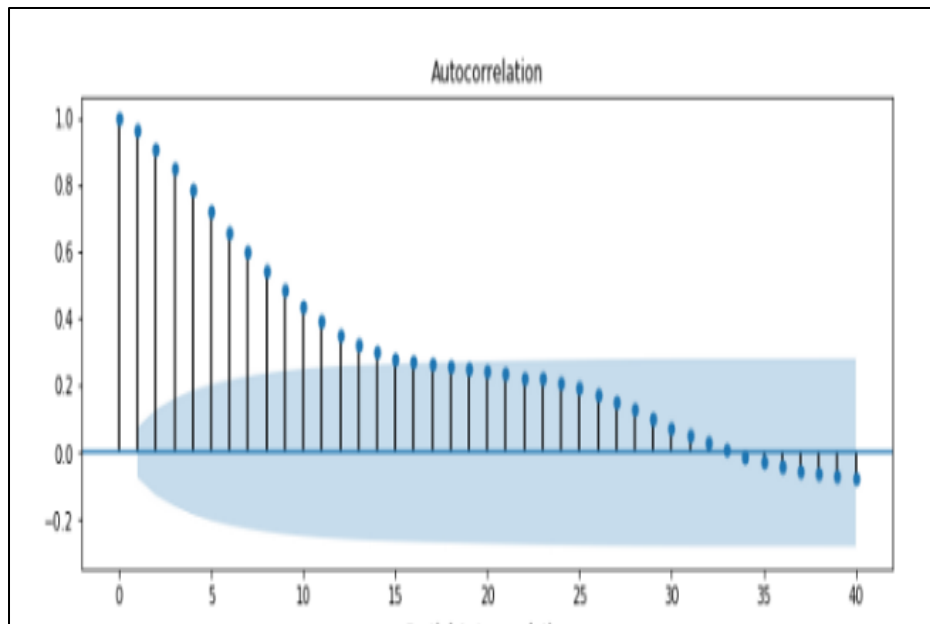


Figure 3.5 Auto Correlation Function

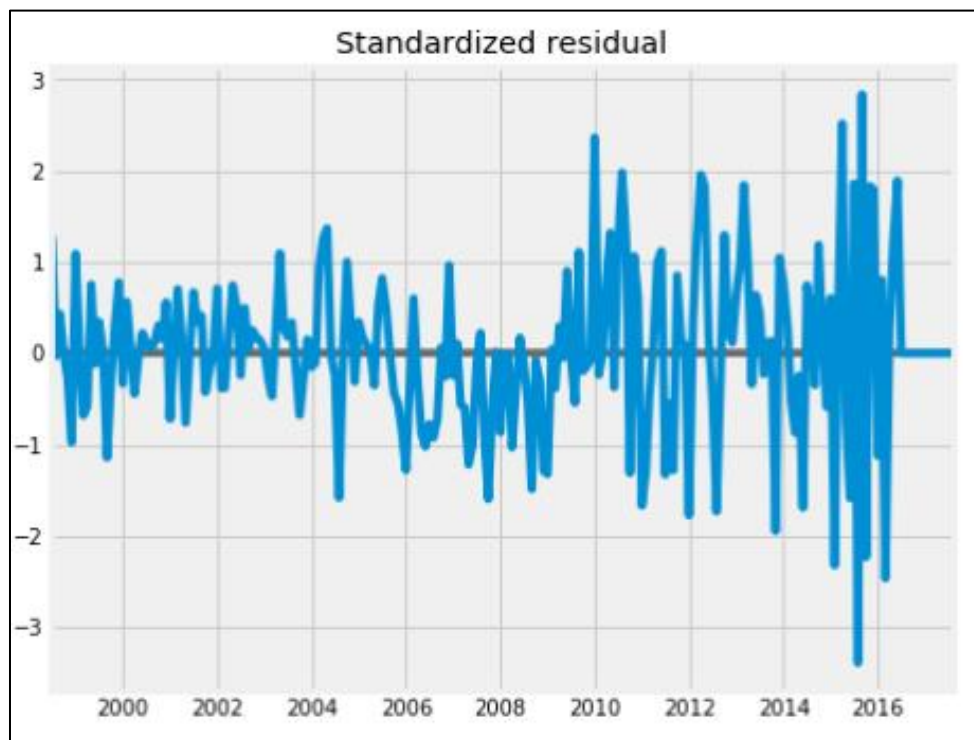


Figure 3.6 Standardized residual function

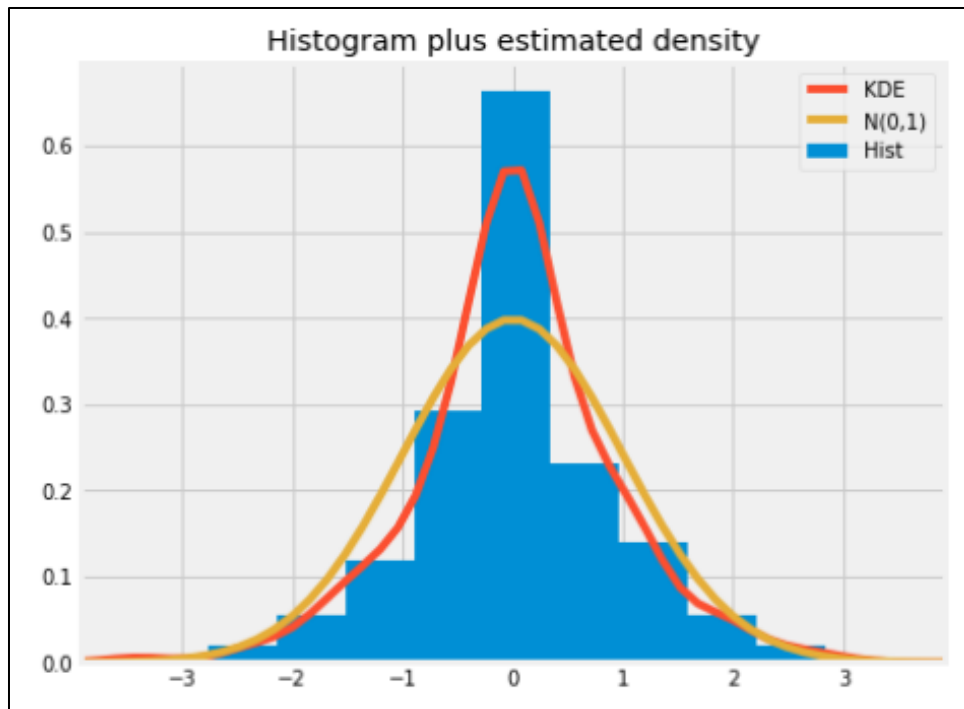


Figure 3.7 Density distribution

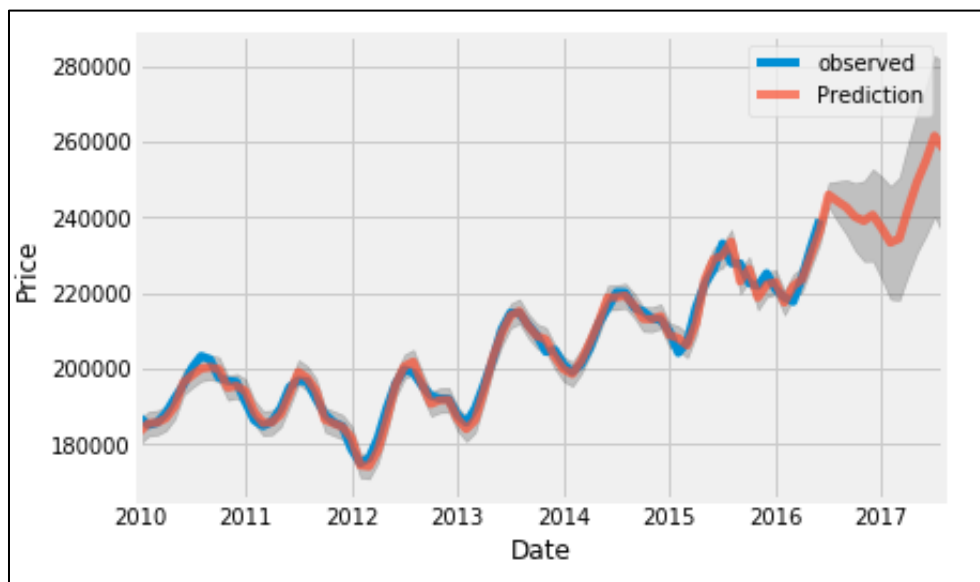


Figure 3.8 Predictions compared with observed dataset

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.

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.

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.

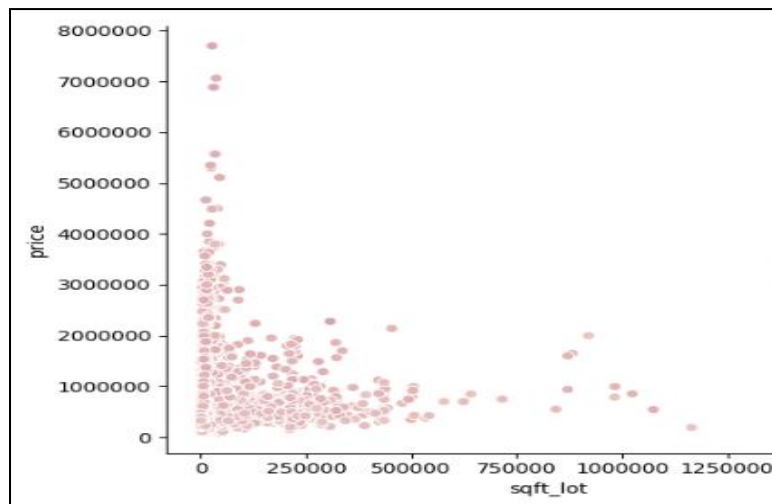


Figure 3.9 Plot of extent of the land, distance to main road features

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 0.5619605842677327, 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 testing 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.

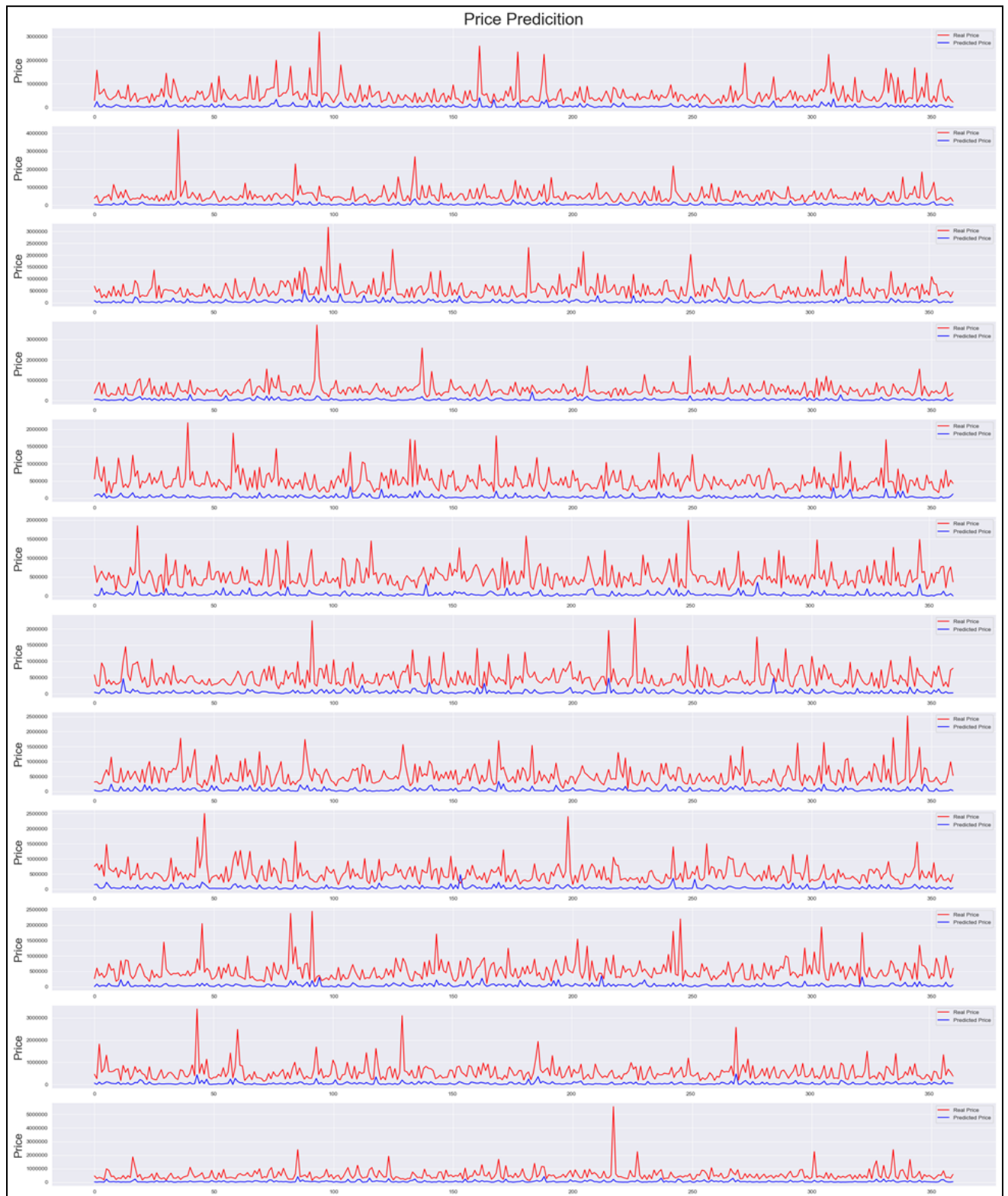


Figure 3.10 Train accuracy (in red) and test accuracy (in blue) curves of ANN model

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.

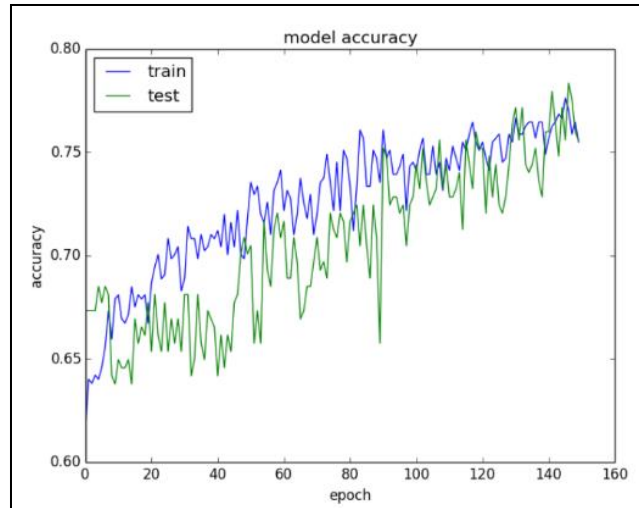
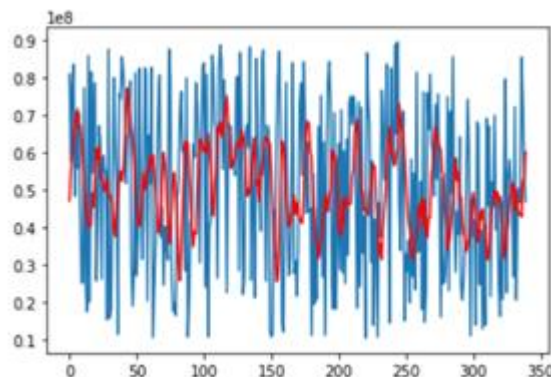


Figure 3.11 : Accuracy curve for ensemble model

Future Value Prediction

First approach was based on features of the land and historical data of price fluctuation rates. There has been used a combination KNN and MLR versus ARIMA model. Results can be concluded as follows. These two models when evaluated, displays MSE value for multivariant linear regression is 590293123907492.1, MSE value for ARIMA model is 605023251112851.9. Therefore, MLR and KNN combination can be seen as a more accurate option for future price prediction.



3.12 ARIMA model forecasts

ARIMA Model Results						
=====						
Dep. Variable:	D.y	No. Observations:	998			
Model:	ARDMA(5, 1, 0)	Log Likelihood	-18426.140			
Method:	css-mle	S.D. of innovations	25231482.665			
Date:	Sun, 04 Aug 2019	AIC	36866.279			
Time:	11:00:00	BIC	36900.619			
Sample:	1	HQIC	36879.332			
=====						
	coef	std err	z	P> z	[0.025	0.975]

const	-2253.6727	2.27e+05	-0.010	0.992	-4.47e+05	4.43e+05
ar.L1.D.y	-0.8700	0.031	-27.770	0.000	-0.931	-0.809
ar.L2.D.y	-0.7050	0.041	-17.356	0.000	-0.785	-0.625
ar.L3.D.y	-0.5085	0.043	-11.691	0.000	-0.594	-0.423
ar.L4.D.y	-0.3001	0.041	-7.372	0.000	-0.380	-0.220
ar.L5.D.y	-0.1418	0.031	-4.519	0.000	-0.203	-0.080
Roots						
=====						
	Real	Imaginary	Modulus	Frequency		

AR.1	0.5117	-1.3305j	1.4255	-0.1916		
AR.2	0.5117	+1.3305j	1.4255	0.1916		
AR.3	-1.5044	-0.0000j	1.5044	-0.5000		
AR.4	-0.8177	-1.2800j	1.5189	-0.3405		
AR.5	-0.8177	+1.2800j	1.5189	0.3405		

Figure 3.13 ARIMA model dicky fuller test results

Below is the variation of RMSE with K value as concluded by the KNN model used for price prediction.

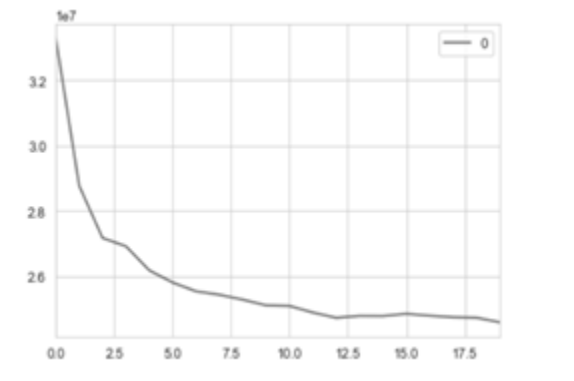


Figure 3.14 RMSE versus K values

The results can be summarized as follows in table 3.2.

	MAE	MSE	RMSE
MLR	21125295.265937325	590293123907492.1	24295948.7138801
ARIMA	21397292.126597658	648244532873425.1	25460646.748922642

Table 3.2 Comparison of MLR and ARIMA performance in future value prediction

KNN Algorithm training results are as follows

RMSE value for k= 1 is: 33236335.5633499
RMSE value for k= 2 is: 29350206.021491293
RMSE value for k= 3 is: 27730285.611861072
RMSE value for k= 4 is: 26450157.466238208
RMSE value for k= 5 is: 25962758.90996831
RMSE value for k= 6 is: 25563133.749448232
RMSE value for k= 7 is: 25271278.39128187
RMSE value for k= 8 is: 24852237.123635385
RMSE value for k= 9 is: 24849072.09022946
RMSE value for k= 10 is: 24659752.571712613
RMSE value for k= 11 is: 24568955.481309887
RMSE value for k= 12 is: 24557293.04777497
RMSE value for k= 13 is: 24439868.888457965
RMSE value for k= 14 is: 24297442.464820538
RMSE value for k= 15 is: 24216999.15279268
RMSE value for k= 16 is: 24103323.314502314
RMSE value for k= 17 is: 24175568.901250575
RMSE value for k= 18 is: 24133684.631889954
RMSE value for k= 19 is: 24139068.93470548
RMSE value for k= 20 is: 24042353.731431223

Second approach of predicting future values is based on infrastructural facilities. The prediction is made with relative to the future development projects in the area. This includes the infrastructure facilities such as school, hospitals, highway, and apartment. The commercial value of a land in future is calculated based on the percentage ratio. Each of these infrastructure facilities are given a specific percentage value based on studies made on the effect on land value with emerging facilities. The dataset collection plays a major role here which involves the percentage calculation. The machine learning field is to be used in order to predict the land value. In order to achieve these objectives, the machine is trained and tested with the dataset to predict the future commercial value of the land with effect of different infrastructural additions. For example, being the current land price Rs.2000, assuming after five years to have school, a hospital so land value percentage increased by school is 20% and by hospital 40%, hence future land value will be $2000 + 20\% + 40\%$ kind of prediction is done. In the prediction system there mainly use one algorithm to predict the data that is MLR the purpose of the using two algorithms is to provide a most accreted point. In doing that thing, first of all, need to train the algorithm using future infrastructure dataset using the past information collected.

Here, Score Model segment that attempts to anticipate the test information. The train model gives a prescient calculation that a score model employs. The Scored Labels segment characterizes the expectation of Sales Amount.

For the straight relapse, we will take a gander at "Coefficient of Determination". This worth educates us about the exactness regarding the model and can change somewhere in the range of 0 and 1. In the event that the worth is close 0.8 or 1, the straight relapse model is dependable.

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	76767	11	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.800,0.900]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.700,0.800]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.600,0.700]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.500,0.600]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.400,0.500]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.300,0.400]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.200,0.300]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.100,0.200]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713
(0.000,0.100]	0	0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.713

Figure 3.15 Prediction and prediction probability calculation.

Experiment created on 6/25/2019 > Evaluate Model > Evaluation results

Metrics

Mean Absolute Error	3156.034607
Root Mean Squared Error	3842.607571
Relative Absolute Error	1.075045
Relative Squared Error	1.279317
Coefficient of Determination	-0.279317

Figure 3.16 MAE,MSE for the given model

When we run our model, we will see a coefficient of assurance about 0.9. It implies that our model is practically more accurate. Predicting future values is based on infrastructural facilities. The prediction is made with relative to the future development projects in the area. This includes the infrastructure facilities such as school, hospitals, highway, and apartment. The commercial value of a land in future is calculated based on the percentage ratio.

In the prediction system there mainly use one algorithm to predict the data that is MSE the purpose of the using algorithms is to provide a most accreted point. In doing that thing, first of all, need to train the algorithm using future infrastructure dataset using the past information collected. Here, Score Model segment that attempts to anticipate the test information.

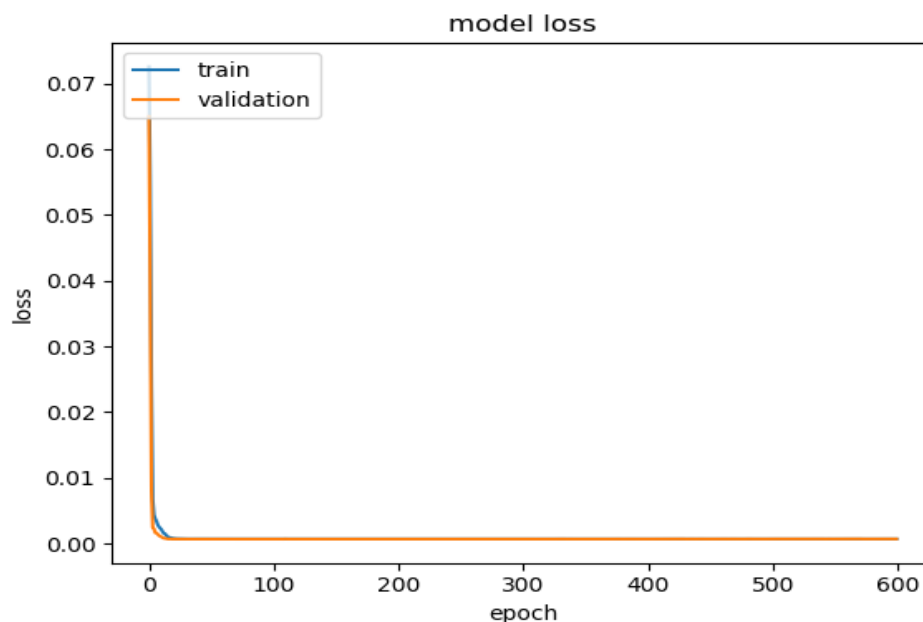
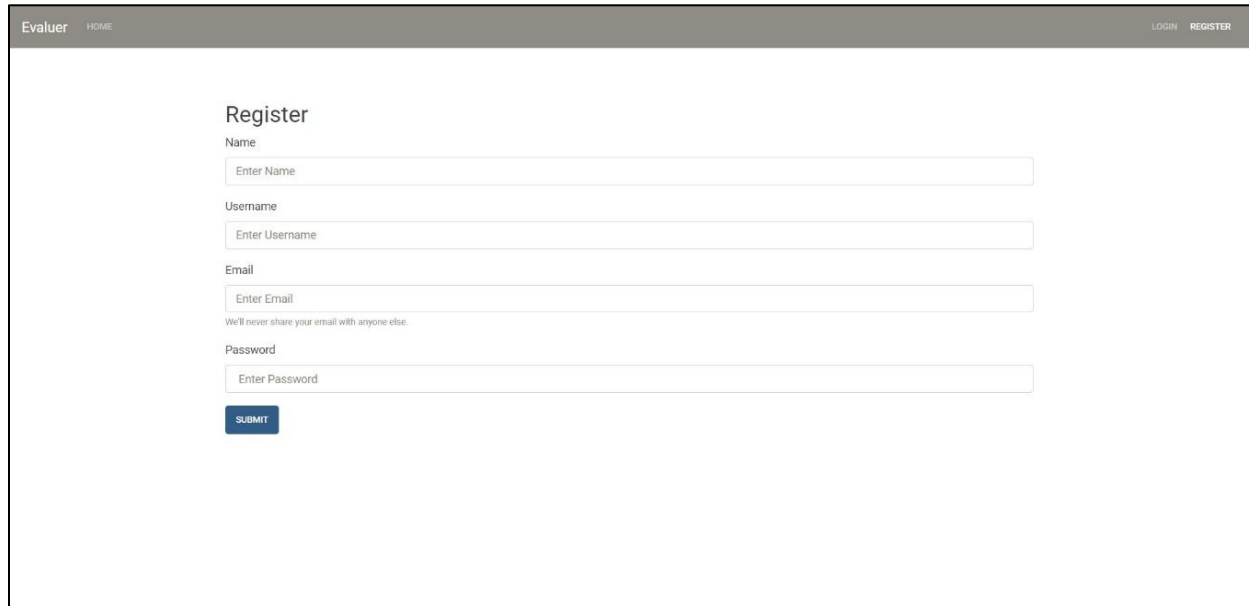


Figure 3.17 Loss curve for neural network

3.2 User Interfaces

Mobile responsive front end was developed for fulfilling the system user interface requirements of the system




The registration form is titled "Register" and is located on a page with a header containing "Evaluator" and "HOME" on the left, and "LOGIN" and "REGISTER" on the right. The form fields are: "Name" with a placeholder "Enter Name", "Username" with a placeholder "Enter Username", "Email" with a placeholder "Enter Email" and a note "We'll never share your email with anyone else.", and "Password" with a placeholder "Enter Password". A blue "SUBMIT" button is at the bottom.

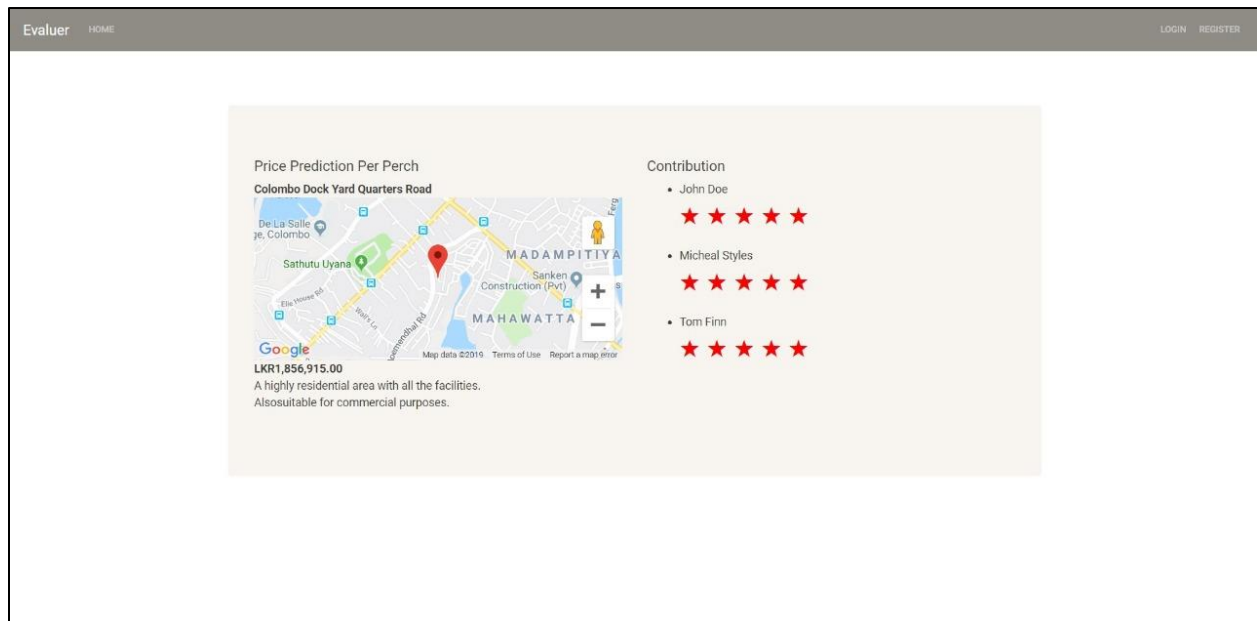
Figure 3.18 Registration



The login form is titled "Login" and is located on a page with a header containing "Evaluator" and "HOME" on the left, and "LOGIN" and "REGISTER" on the right. The form fields are: "Username" with a placeholder "Enter Username" and "Password" with a placeholder "Enter Password". There is a checkbox labeled "Remember me" and a blue "LOGIN" button at the bottom.

Figure 3.19 Login

Evaluator	HOME	DASHBOARD	LAND PRICE	PROFILE	FUTURE PREDICTION	LOGOUT
<h2 style="text-align: center;">Enter your land information</h2>						
Price:	Address					
<input type="text"/>	<input type="text"/>					
Distance to nearest bus route:	Search Nearest Location					
<input type="text"/>						
Nearest bus route:						
<input type="text"/>						
Accessibility index:						
<input type="text"/>						
Province:						
<input type="text"/>						
District:						
<input type="text"/>						
<input type="button" value="SUBMIT"/>						
<p>Latitude: <input type="text"/> 6.901608599999999</p> <p>Longitude: <input type="text"/> 80.0087746</p>						



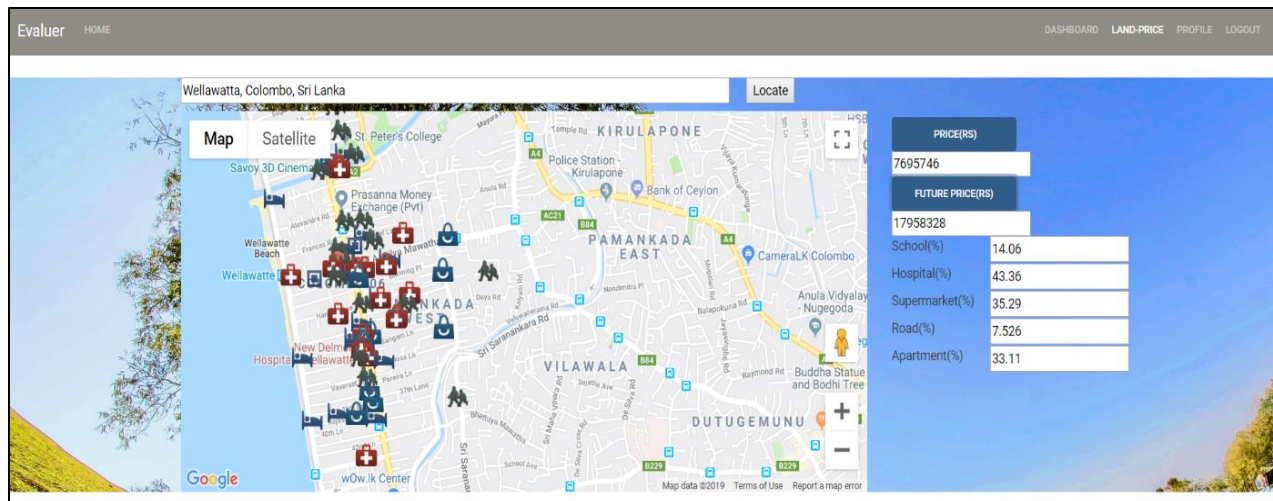


Figure 3.22 Future value prediction

3.3 Discussion

3.3.1 Current Value Prediction

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).According to above 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.

3.3.2 Future Value Prediction

Future value prediction was carried out based on two approaches. Then the specific values for each sector are coming as the output. Then the output will be compared with other machine learning algorithms. Extracted structured data are gained by the fellow members and then the data are fed into the algorithm. After that a comparison runs between these data and price with aid of the algorithm. Hereafter, the output which comes from this algorithm and other algorithm is compared. Finally, the decision is made which the most trustworthy prediction. First approach was based on features of the land and historical data of price fluctuation rates. MLR getting high accuracy and it predicts the non-time series algorithm and ARIMA used for time series. Therefore, when combined the two, accuracy of the predicted values can be increased.

4 CONCLUSION

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. In predicting future values, the combination of KNN and MLR was identified to be outperforming the ARIMA model with lower MSE value. Again, same as in LSTM, tradeoffs are possible.

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] Mukhlisin, M. (2017). Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor - IEEE Conference Publication. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/document/8276357> [Accessed 19 Feb. 2019].
- [8] Zillow. (2019, February 21). Retrieved from <https://en.wikipedia.org/wiki/Zillow#Zestimate> [Accessed 23 Feb. 2019].
- [9] 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].
- [10] Trulia. (n.d.). Retrieved February 22, 2019, from <https://en.wikipedia.org/wiki/Trulia> [Accessed 23 Feb. 2019].
- [11] "QV Homeguide App Now Available." New Zealand Property Investors Federation, 3 Mar. 2015, www.nzpif.org.nz/news/view/56971. [Accessed 24 Feb. 2019].
- [12] "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].

- [13] “ABOUT US.” Houseprice.AI-What's the Fair Price ?, www.houseprice.ai/about. [Accessed 24 Feb. 2019].
- [14] “Introducing Houseprice.AI: The Must Have Tool for Every Developer.” Bridging Loans | Development Loans | AvamoreCapital, 29 May 2018, avamorecapital.com/introducing-houseprice-ai-the-must-have-tool-for-every-developer/. [Accessed 24 Feb. 2019].
- [15] 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].
- [16] 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].
- [17] 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.
- [18] Chandrasekara, Vasana & Tilakaratne, Chandima. (2011). Comparison of Support Vector Regression and Artificial Neural Network Models to Forecast daily Colombo Stock Exchange.
- [19] 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.
- [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] Vaz, J. (2015). REAL ESTATE APPRAISAL AND SUBJECTIVITY. *European Scientific Journal March 2015*, ISSN: 1857 – 7881(e - ISSN 1857- 7431), pp.55, 63.
- [22] 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
- [23] 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
- [24] Leo Breiman. “Random forests”. In: Machine learning 45.1 (2001), pp. 5–32

- [25] Random Forests by Leo Breiman and Adele Cutler. URL: <https://www.stat.berkeley.edu/~breiman/RandomForests/>.
- [26] 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.
- [27] 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

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 1 km 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

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

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

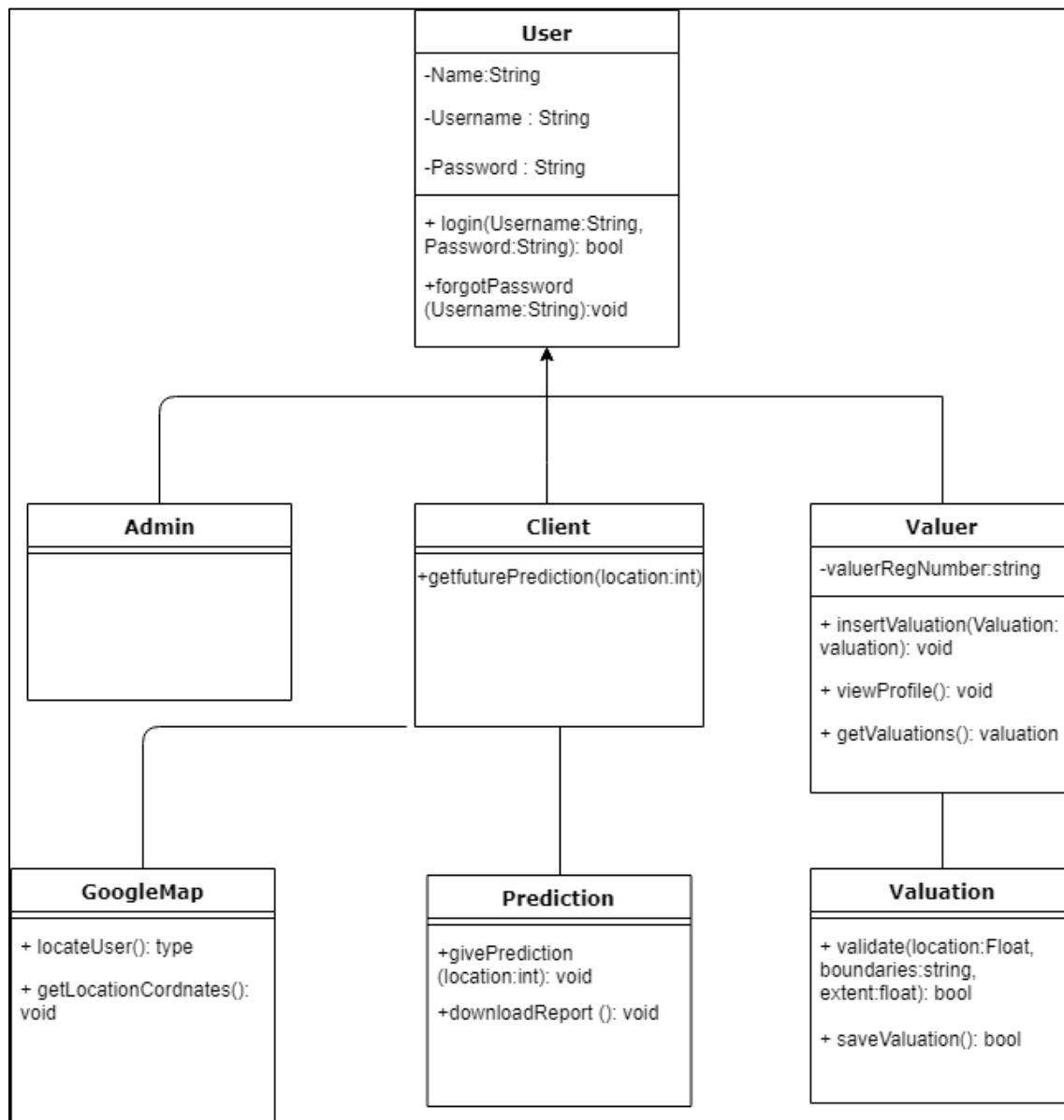
Value of the land as per valuer's report in Sri Lankan Rupees *

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

