A Model for the Estimation of Land Prices in Colombo District using Web Scraped Data

R.A.G.Naotunna

2022

A Model for the Estimation of Land Prices in Colombo District using Web Scraped Data

A dissertation submitted for the Degree of Master of Business Analytics

R.A.G.Naotunna
University of Colombo School of Computing

2022

Abstract

Sri Lankan people have been showing keen interest in real estate investments, especially in the Colombo district, as these assets do not depreciate with time like most tangible assets and as these investments cause a significant outflow of money from their overall wealth. However, at present lands in Sri Lanka are valued based on the experience and judgment of the individual valuation officers which could be highly subjective and questionable as the way of analyzing the features and providing a value could vary from person to person. In an attempt to address the above-mentioned issue, this research focuses on developing a machine learning model to estimate the land prices in the Colombo district by utilizing web scraped data.

To achieve the above objective, web advertisements posted in the ikman.lk on lands for sale in the Colombo district for a 3 months period were scraped and obtained the land related data. These data were amalgamated with land price determinants data obtained from other web sources and formed the dataset which contained 3725 records distributed over 43 land price determinants. Further, when developing the required dataset, steps have been taken to collect data about different sub-categorical levels of each price determinant as it could add more value and make the dataset being built more meaningful.

This dataset is utilized to fit five machine learning algorithms, namely; Multiple linear regression, Random Forests Regression, Support Vector Regression, Extra Trees Regression and Extreme Gradient Boosting. The performance of each machine learning model is gradually increased through feature reduction and hyper-parameter optimization. In feature reduction, two different approaches; a wrapper method (Recursive Feature Elimination) and a filter method (SelectKBest) were utilized, and selected the approach which provided the optimum results. Out of the five machine learning algorithms utilized, the hyper-parameter optimized Random Forests regression model outperformed the other linear, nonlinear, tree-based and ensemble machine learning models. The model performed exceptionally well for unseen data with R^2 value of 90.24% and MAPE, MAE and RMSE values of 17.88%, 0.098065 and 0.313154 respectively

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: R.A.G. Naotunna

Registration Number:2019/BA/020

Index Number:19880202

Signature: Date:19/11/2022

This is to certify that this thesis is based on the work of

Mr./Ms.

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: SPW

Signature: Date: 19th Nove 2022

Acknowledgement

I would like to express profound gratitude to my supervisor, Prof. Prasad Wimalaratne, for his invaluable support in guiding the selection, scoping, and completion of this research study. Further, I would like to thank all my colleagues for their help in finding relevant research material, for sharing knowledge and experience, and for their encouragement. I am as ever, especially indebted to my family for their love and support throughout my life. Finally, I wish to express my gratitude to all my colleagues at work, for the support given to manage this MSc research work.

Table of contents

List of Figures

List of Tables

1. Introduction

The real estate segment in Sri Lanka has witnessed a boom during recent times as evidenced by the increasing prices and increasing demand("Lanka Property Web - Find the average Sri Lanka House & Land Prices | Sri Lanka House Price Index," n.d.). In the backdrop of high levels of inflation("Measures of Consumer Price Inflation | Central Bank of Sri Lanka," n.d.), investors as well as the public have shown interest towards new investment avenues. Among such avenues, due to the appreciating nature of the asset("Land Valuation Indicator - First Half of 2021 | Central Bank of Sri Lanka," n.d.), investing in property in the form of land has become a popular choice. While the importance of lands in Colombo district is widely discussed, lands in Sri Lanka are currently valued based on the experience and judgement of the individual valuation officers. This manual and individual oriented method is considered highly subjective and questionable as the way of analyzing the features and providing a value could vary from person to person(Ariyawansa, 2016) and could lead to the determination of different values to the same land by two different valuation officers(Li et al., 2015). Further, machine learning models are considered superior to human expert estimations as they produce more accurate predictions than the estimations made by real estate professionals(Kim et al., 2020). Even though importance of implementing a land price estimation model in Sri Lanka is highlighted in the past in (Ariyawansa, 2016),(Li et al., 2015), only few attempts have been made to address this issue, specifically by considering the importance of location related variables and utilizing machine learning models.

Thus, the proposed study would contribute to the price predictions by filling these research gaps by attempting to develop a machine learning model to estimate land prices in Colombo district by considering the impact of price determinants at more granular sub-categorical levels utilizing Machine Learning techniques. One of the major challenges in developing a land price prediction model in the Sri Lankan context is the non-availability of publicly available data with respect to land prices in a structured manner. Therefore, the required dataset in this study is developed by using land sales advertised on the web, which is considered (Pai and Wang, 2020) as a more reliable mode of collecting data when developing price prediction models. Research on this area will enable the people who are willing to buy or sell a property in Colombo district to identify a reasonable price for their potential transactions. Further, it will provide a data driven platform to the land valuers in Sri Lanka to benchmark and compare their professional estimations.

1.1. Aims and Objectives

Identification of the right price determinants impacting the land prices of a particular geographical area is identified as the key success factor in modelling land prices(Zhang et al., 2021)(Derdouri and Murayama, 2020)(Córdoba et al., 2021). Consideration of the impact of price determinants at granular sub-categorical levels(Zhang et al., 2021) rather than considering them at a very high level could add high validity to the model being built. Further, usage of price related data available on the web(Pai and Wang, 2020) is considered a more convenient mode of communication for the general public when developing price prediction models.

This study focused on contributing to the field of land price predictions by proposing a machine learning model for estimation of land prices in Colombo district by utilizing web scraped data. Further, price determinants are considered at more granular sub categorical levels in this study rather than just considering the basic form of the price determinants. As the structured form of publicly available data on land prices and related variables is not available in the Sri Lankan context, a data set with required variables was built in this study through scraping a land price listing website in Sri Lanka. Moreover, this study aimed at identifying the most important determinant that drives the land prices in Colombo district. Overall objectives of the study can be summarized as follows:

- To conduct a systematic review of literature to identify the land price determinants and machine learning techniques for modelling land prices in Colombo district
- To extract required land prices and other land related data from a land price listing website through web scraping
- To collect data on land price determinants through freely available sources based on the recommendations from previous similar studies mentioned in literature
- To analyze the applicability of those identified determinants in the Sri Lankan context
- To synthesizing all the collected and web scraped data to form a combined dataset
- To identify and apply appropriate machine learning algorithms
- To identify and apply appropriate accuracy evaluation metrices to evaluate the accuracy of the model predictions

1.2. Scope

This project focuses on building a model to estimate land prices in Colombo district by using land price data advertised in www.ikman.lk and consent from the ikman.lk authorities has been obtained to use the details available on the website to conduct this study (Attachment I). Geographical area of interest for this study is the lands available for sale in Colombo district which are listed in the above-mentioned real estate listing website.

Land price, extent of the land (land size), land usability purpose (land type) and the location of the land, price scale, posted date of the advertisements have been obtained by scraping the above real estate listing website. Data with respect to other proposed variables are obtained through publicly available websites mentioned under the data collection sub section of the Methodology chapter of this dissertation. Over 4900 records of land price listing data advertised in www.ikman.lk from 31.07.2021 to 21.09.2021 have been scraped and combined with other price determinants data to form the final dataset.

Python (Version 3.8.3) is primarily used for data extraction and data cleansing requirements of this project. Along with Python (Version 3.8.3), MS Excel and MS PowerPoint software were used for other analytical and data cleansing purposes required during the project life cycle.

1.3. Structure of the dissertation

This dissertation consists of six chapters inclusive of the Introduction chapter. Similar related work done in the past and Machine learning models previously used in the projects similar to this study and their inherited pros and cons are discussed in Chapter 2; Background and Related work. Data collection for the project, variable selection for data modelling, machine learning models and model performance evaluation of the project are discussed in Chapter 3; Methodology. Analysis and the results of the study along with discussion on the model performance are discussed in Chapter 4; Evaluation. To wrap up things, chapter 5 concludes the study by highlighting the general discussion, summarizing key findings, limitations, and future work along with the conclusion of the study. The references used in the thesis is stated in chapter 6.

2. Background and Related Work

This chapter will compare the approaches and machine learning algorithms utilized in the previous studies to estimate the land valuations and will highlight the significance of capturing the most important determinants of land prices for studies of this nature.

A proper mechanism to accurately estimate the land prices is an important consideration in any nation. Zheng(Zhang et al., 2021) highlighted the importance of accurate mapping of residential land prices and regular monitoring of spatiotemporal changes based on three main reasons. Firstly, as it helps to analyze the neighborhood and location characteristics which affect real estate values and explain the facility preferences of residents(Liu et al., 2018). Secondly as it helps to assess and monitor the local residential land market(Hu et al., 2016). Thirdly, as it indicates the direction and pattern of urban expansion to some extent(Mendonça et al., 2020), thus reflects the evolution of urban spatial structure. Furthermore, knowledge and continuous monitoring of land values in the market is considered highly important as it plays a major role in overall land planning, contributing towards control of land price speculations and identification of the zones with higher or lower valuation(Córdoba et al., 2021) which eventually guides government intervention to promote more equitable land development. Finding accurate methods to estimate and map land prices at macro scale based on publicly accessible and low cost spatial data is an essential step in producing a meaningful reference for regional planners(Derdouri and Murayama, 2020). This would assist them in making economically justified decisions which could be used by key investors for development projects and post disaster recovery efforts. Figure 2.1 shows the taxonomy of literature referred in this study which is categorized based on focus, data collection methodology, feature selection methodology and approach utilized in the respective study. In the subsequent sections in this chapter, challenges in modelling land prices, determinants that could be utilized to build land price prediction models and different approaches taken by previous scholars to build land price prediction models are discussed.

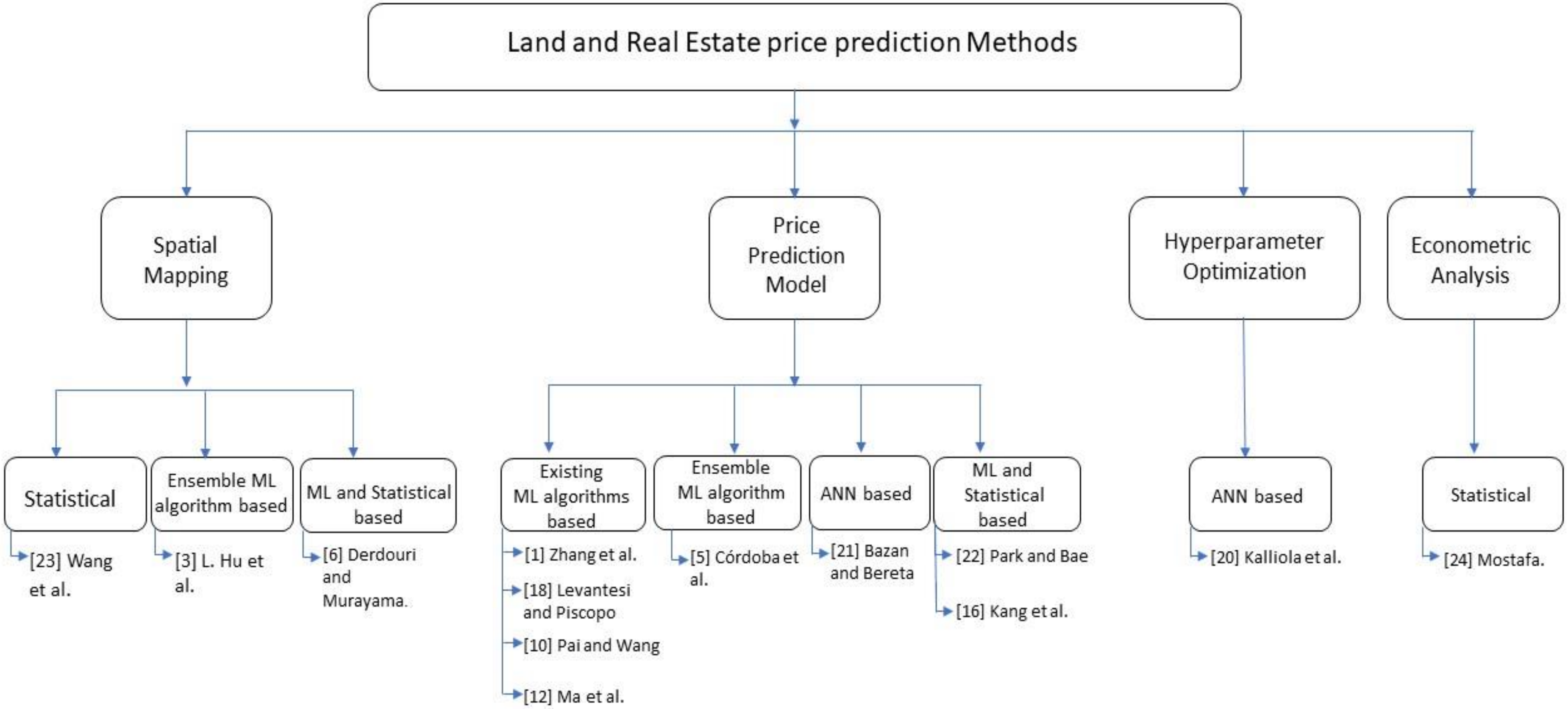


Figure 2.1 Taxonomy of land and real estate price prediction models

2.1. Challenges in modelling land prices

Even though accurate mapping of land prices with respective determinants is an important milestone to any country, modelling residential land prices in most developing countries still remains as a major challenge. In countries such as China and Iran(Mirkatouli et al., 2018), information with respect to residential land price use is not publicly available. Even though aggregated data for residential land prices can be obtained through manual surveys and official statistics(Nakamura, 2019), those sources do not adequately provide micro level attributes required for fine scale residential land price mapping. Further, such manual data collection processes consume significant amounts of time and labor and even the collected data have relatively coarse spatial and temporal resolution.

Outdated cadastral data, financial speculation and inflationary processes(Córdoba et al., 2021) are considered as major barriers for comprehensive land valuations as these hinder the price formation mechanism from the traditional markets. The process of examining the variation in land prices in a wide area is considered more challenging(Derdouri and Murayama, 2020) due to the significant budget and time consuming process behind extracting land price maps covering a whole region, which requires costly and lengthy field surveys. Further, the available samples do not usually cover the entire study area in question as the data collected from dispersed locations.

In the Sri Lankan context too, availability of data with respect to land prices remains a major barrier for studies focusing land price evaluations(Ariyawansa, 2016). Real transactional data related to lands remains undisclosed with the government regulatory bodies and only data available with respect to land price quotations are in newspapers and land price listing websites. Further, non-availability of geospatial distributions of land prices and difficulty in obtaining data with respect to determinants of land prices restricts local researchers from conducting their studies in this area.

2.2. Potential determinants of land prices

Commercial and educational facilities (Hu et al., 2016) are considered important determinants of the residential land price distribution in Wuhan, China. Further on the temporal perspective, there exists higher impact from natural amenities and educational facilities on the determination of land prices than the impact from commercial facilities and public transportation. However, this study does not consider the impact of various facilities on the residential land prices at different levels such as size, quality or grade of facilities.

Factors influencing real estate prices could be quantitative or qualitative. The quantitative factors could contain unemployment rates, share index, current account of a country, industrial production, and gross domestic product (Pai and Wang, 2020). Subject preferences of decision makers such as building styles and living environment are identified as qualitative factors which could influence land prices. However, most of the qualitative factors mentioned above suffer from lack of measurements and difficulties with respect to data collection.

According to Haizhen Wen(Wen et al., 2018), variables that could affect the land prices can be grouped into three main types; individual factors referring to the characteristics of land parcel (size and shape), neighborhood factors related to the characteristics of land parcel including socioeconomic variables, external environment, and amenities; and location determinants depicting traffic patterns and distance to the central business district. Further, there exists close association of the economic value of land with variables such as population density, proximity to railways, schools and other facilities(Derdouri and Murayama, 2020). Strong association of elevation and job density with land economic values is also highlighted in this study. Selection process of suitable factors depends mainly on elements such as the setting of the designated target area, type of land price (eg. Residential and commercial), and the availability of spatial data.

Derdouri and Murayama selected explanatory variables for their study based on land parcels within urban and rural areas, land parcels not only available for residential purposes and the availability of free spatial data(Derdouri and Murayama, 2020). Based on these three criteria, they selected distance to the nearest railway station, area of rice fields, area of other agricultural land, area of forest, area of cultivated land, area of roads, area of railways, area of other land uses, area of water bodies, area of seashore, area of the surface of the sea, area of golf courses, population density, urbanization promoting area, number of enterprises, number of employees and elevation as the explanatory variables for their study, which are publicly available and no cost data from different sources.

For the proposed study, land price determinants were decided after considering the variables mentioned in literature and by considering the availability of data in the Sri Lankan context. Data with respect to price of the land, area of the land and city are available in the Real estate listing websites and was extracted through careful web scraping procedure. Data with respect to price determinants such as Healthcare facilities, Educational facilities, commercial facilities, public transportation etc. are publicly and freely available in the internet which could be readily accessed.

2.3. Approaches to model land prices

Table 2.1 summarizes the statistical and machine learning approaches undertaken previously to model land prices, algorithms used, findings and the limitations of these studies.

Table 2.1 Comparison of Statistical and Machine learning approaches to land price estimations

Price estimation method	**Approach**	**Statistical or Machine Learning algorithms utilized and their pros and cons**	**Findings**	**Limitations**
Spatial Mapping	Statistical	Multiple Regression analysis with ordinary least squares (MRA with OLS), Geographically weighted regression (GWR) model, Geographically and temporally weighted regression (GTWR) model Pros- Geographically weighted regression allows to fit geographically varying coefficients in to regression model Cons- Geographically weighted regression models generally suffer from multicollinearity in local coefficients, multiple hypothesis testing, and the incapability of decomposing the global estimates into local estimates	Level of community can be effective for the mass appraisal modeling with annual average price and other meaningful attributes. GWTR model outperformed every other model in algorithmic performances	Wang(Wang et al., 2020) highlighted that community data considered for statistical modelling in his study is a mathematical processing of the original individual transaction data which may result in loss of some important information. Further he explained that for a larger dataset, which would result in multiple increases of the amount of calculations could be challenging to the stability of the regression model.
	Ensemble Machine Learning Algorithm based	Integrated Machine learning models with hedonic model(Hu et al., 2016) to map spatial patterns. Machine Learning algorithms used were random forest regression (RFR), extra-trees regression (ETR), gradient-boosting regression (GBR), support vector regression (SVR), multi-layer perceptron neural network (MLP-NN) and k nearest neighbor algorithm (k-NN) Pros- output of the tree-based algorithms are generally more interpretable compared to neural networks. For larger datasets, neural networks provide more accurate predictions compared to tree-based algorithms Cons- Tree based algorithms generally take more time to train the model and can suffer with over fitting very often. Neural networks suffer from interpretability related issues of the output.	tree-based bagging algorithms (RFR, ETR) outperformed the rest	S.Hu(Hu et al., 2016) showed that Insufficient consideration of data with respect to temporal aspect restrict the study ensembled machine learning techniques to further extend for time series analysis and long-term monitoring(Hu et al., 2016) .Further consideration of single type of data should be further improved by integrating more types of data from several sources to improve the prediction accuracy of the developed model.

	Machine Learning and statistical based	Geostatistical mathematical models of regression kriging (exponential, gaussian and spherical) Generalized linear model, generalized additive model using splines, support vector machines with linear kernel, multivariate adaptive regression spline, k nearest neighbors(KNN), support vector machines with radial basis function kernel, Cubist, stochastic gradient boosting and random forests(Derdouri and Murayama, 2020) Pros- support vector machines work relatively well when there is a clear margin of separation between classes and in high dimensional spaces. KNN is non parametric and doesn't require validation with assumptions. Cons-Support vector machines don't perform well when the dataset is larger and has more noise. Generalized linear models are parametric and data should comply with several assumptions in order to apply the algorithm.	Random Forests outperformed all the geostatistical and machine learning methods based on the calculation of Mean absolute error, Root mean squared error, and R2	In Geostatistical Mathematical models, the main limitation is the non-possibility of obtaining spatial data due to availability issues and high costs(Derdouri and Murayama, 2020). Derdouri and Murayama considered only one machine learning model for the entire prefecture which resulted in overestimated land prices in urban areas and underestimated land prices in suburban areas. Further, their study relied entirely on published literature to select potential land price determinants which could significantly depend on the settings of the target area and did not consider ensembled methods by combining multiple models which could result in further improvement to accuracy of the predictions.
Price Prediction Modelling	Existing Machine Learning algorithms based	Zhang(Zhang et al., 2021) utilized support vector regression (SVR) with radial basis function, SVR with a linear kernel, random forests regression, extra trees regression (ETR) and multiple linear regression to predict residential land prices. Pai and Wang (Pai and Wang, 2020) utilized least square support vector regression, classification and regression trees, general regression neural networks, backpropagation neural networks to predict the prices of real estate To investigate the influential factors on land values, Jun Ma(Ma et al., 2020) has utilized random forest (RF), Gradient Boosting Decision Tree (GBDT), Multi Linear Regression (MLR), Linear Support Vector Regression (SVR), Multilayer Perceptron (MLP) Regression, and K-Nearest Neighbor (KNN) Regression algorithms	Both Zhang(Zhang et al., 2021) and Jun Ma(Ma et al., 2020) showed that Tree based and Nonlinear machine learning algorithms perform better than traditional linear methods. However, Pai and Wang(Pai and Wang, 2020) showed that least square support vector regression performs better compared to other linear and nonlinear models that they have utilized in their study.	Zhang(Zhang et al., 2021) highlights that his study on building a price prediction model by utilizing existing machine learning algorithms did not consider the Impact of various facilities on the residential land prices at different levels. Further, integration of Machine learning algorithms has not been considered which could improve the interpretability of the final model. Pai and Wang (Pai and Wang, 2020) highlight that non consideration of diverse data types such as comments of real estate attributes, prices from social media, images from Google maps, and economic indicators as one of the main limitation in their land price prediction model
	Ensemble Machine learning algorithm based	In a study to model the relationship between a set of environmental variables and rural land values, Córdoba(Córdoba et al., 2021) has used spatial quantile regression forests, linear regression,	Spatial random forests outperformed the rest	Spatial quantile regression forests(Córdoba et al., 2021) lack automation and could result in interpretability related issues

		regression kriging, spatial random forests algorithms Extreme gradient boosting (XG Boost) (Alshboul et al., 2022) for cost prediction of green buildings	XG Boost outperformed Deep Neural networks and Random Forests (RF)	XG Boost is very sensitive to outliers since every classifier is forced to fix the errors in the predecessor learners
	ANN based	A study conducted to build a predictive model by analyzing the possible use of planning documents by Bazan and Michal (Bazan-Krzywoszanska and Bereta, 2018) have utilized deep neural networks with one, two and three hidden layers, and linear regression for comparison	Neural network with three hidden layers outperformed the rest with respect to prediction accuracy	A small number of the available historical transactions, missing values and different attributes' types(Bazan-Krzywoszanska and Bereta, 2018) (numerical, nominal and binary) caused difficulties with the preparation of appropriate input to the networks in ANN which require numeric attributes without missing values
	Machine Learning and statistical based	C4.5, RIPPER, Naïve Bayesian, and AdaBoost were utilized by Park and Bae (Park and Bae, 2015)to predict housing prices	RIPPER model outperformed all selected methods	Considered only specific regions and specific types of residential properties only and considered attributes might not match the other available regions and residential property types (Park and Bae, 2015). Further, Performance evaluation was only based on classifiers
Hyperparameter optimization	ANN based	Kalliola (Kalliola et al., 2021) optimized ANN model by fine-tuning hyper-parameters (such as activation functions, optimization algorithms, etc.) of the ANN architecture for higher accuracy using the Bayesian optimization algorithm	Optimization of model hyper-parameters improved the performance by a good margin (the R2 value improved by 0.05 and the RME value improved by 2.5%)	Difficulty in explaining relationships between inputs and outputs and inability of human intervention directly to these relationships. Further implementation of ANN requires very large datasets(Kalliola et al., 2021)
Econometric Analysis	statistical	To monitor the impact of selected factors on the residential land prices in Kuwait, Mostafa(Mostafa, 2018) has conducted a study by utilizing ordinary least square regression, spatial autoregressive regression and spatial error model	Spatial autoregressive regression model outperformed the rest	Considered only one type of land; residential lands Does not consider the impact of other hedonic factors such as proximity to parks, views of green spaces, the seaside, lakes and waterfalls, degree of education, income per capita and the presence of marginal or segregated neighborhoods on land prices

2.4. Chapter Summary

Many research works have been carried out globally to model estimation of land and real estate prices by utilizing statistical and machine learning methods. Applicability of these methods primarily depend on the nature and quantity of data available as well as on the location of application. These studies highlighted the importance of considering the impact of price determinants at granular sub-categorical levels(Zhang et al., 2021) and utilization of price and price related data available on the web(Pai and Wang, 2020); which is a more convenient mode of communication for the general public, when modelling price prediction models. In the Sri Lankan context too, non-availability of proper pricing mechanism to evaluate the real estate prices(Ariyawansa, 2016) is identified as a major issue in connection with price formation of lands and landed properties. Therefore, to address the above highlighted void in literature, proposed study attempted to model the impact of various land price determinants in Colombo district at different sublevels by utilizing web scraped publicly available data. Further, proposed study will enable the people who are willing to buy or sell a property to identify a reasonable price for their potential transactions and will provide a data driven platform to the land valuers to benchmark and compare their professional estimations.

3. Methodology

This chapter will discuss the methodology of the proposed study to develop the machine learning model to predict land prices in Colombo district. Figure 3.1 illustrates the summary of the methodology undertaken in this study. The approach and techniques used under each step are explained in the subsequent sections.

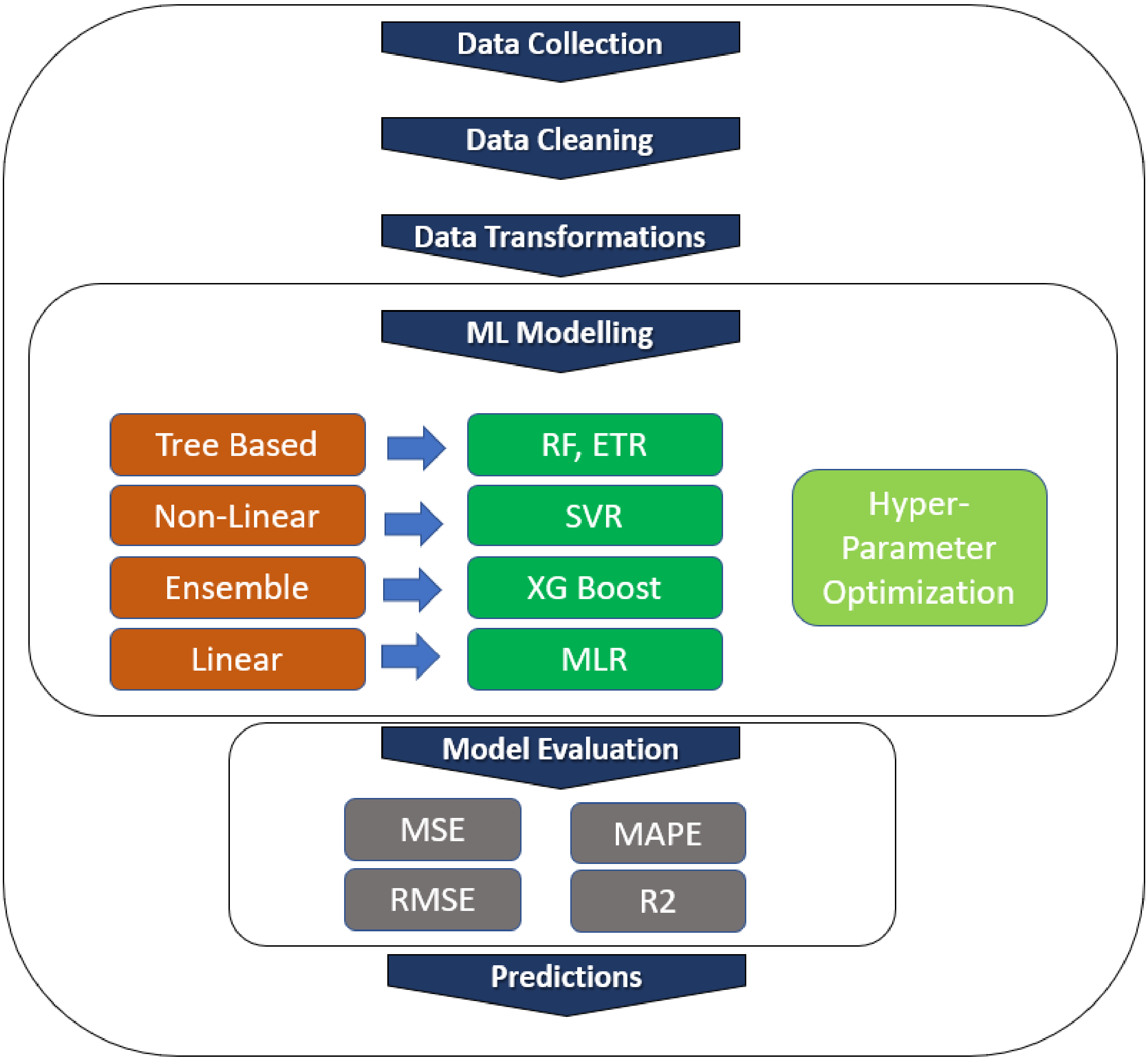


Figure 3.1 Methodological Framework

3.1.Data collection

For the proposed project, data has been gathered primarily from three sources. Summary of the data collection and final dataset preparation approach is shown in figure 3.2

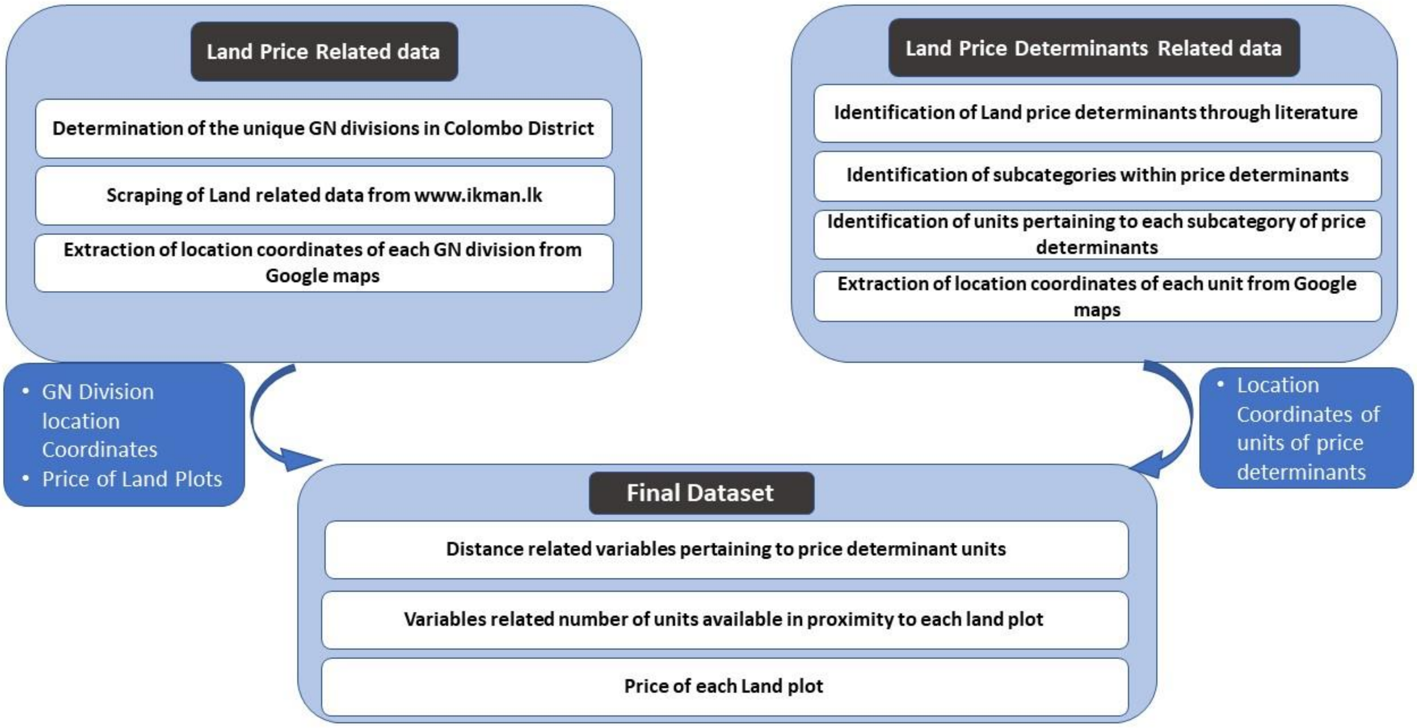
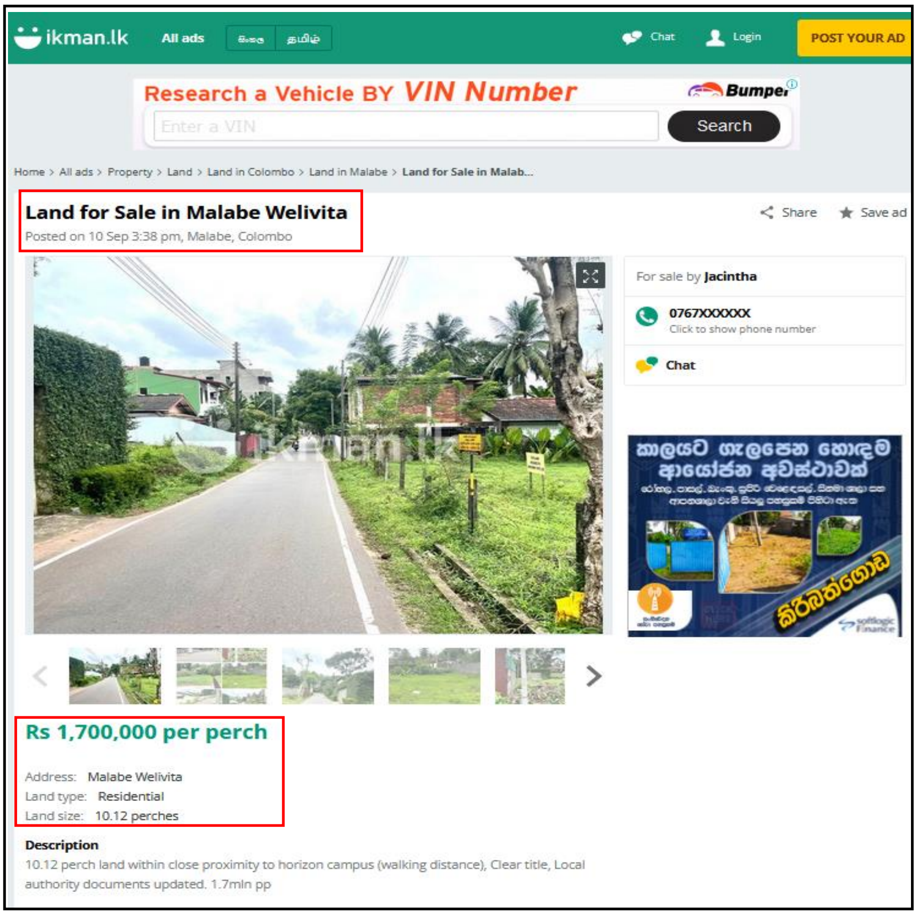

Land Price Related data
Determination of the unique GN divisions in Colombo District
Scraping of Land related data from www.ikman.lk
Extraction of location coordinates of each GN division from Google maps
Land Price Determinants Related data
Identification of Land price determinants through literature
Identification of subcategories within price determinants
Identification of units pertaining to each subcategory of price determinants
Extraction of location coordinates of each unit from Google maps
• GN Division location Coordinates
• Price of Land Plots
• Location Coordinates of units of price determinants
Final Dataset
Distance related variables pertaining to price determinant units
Variables related number of units available in proximity to each land plot
Price of each Land plot

Figure 3.2 Data Collection and Dataset Preparation process

Firstly, land related data is collected by scraping ikman.lk. Thereafter, price determinants and their respective locations are gathered by referring to the official website and sources of the respective price determinants. Finally, location coordinates of each unit of these price determinants are extracted through google maps. Explanation of the process involved in each of these data collection procedures are mentioned below.

3.1.1.Collection of data related to land prices

Data related to land plots is collected through scraping advertisements on lands for sale posted in www.ikman.lk website from 31.07.2021 to 21.09.2021. Even though prices of land plots for sale are mentioned in the websites of the corporate real estate sellers like Prime Lands and Homelands; those data have not been considered for this study due to four main reasons. Firstly, a price for the advertised land plot is not shown by them in their respective websites. Secondly, such sites do not facilitate the general public to sell lands through their websites which would restrict the opportunity for competitive prices. Thirdly, these sites contain restrictions on scraping and use of automated programs on their sites which could prevent the extraction of data for the proposed study. Finally, due to inconsistency in the type of data available from one advertisement to another.



Home > All ads > Property > Land > Land in Colombo > Land in Malabe > Land for Sale in Malab...

Land for Sale in Malabe Welivita

Posted on 10 Sep 3:38 pm, Malabe, Colombo

For sale by Jacintha

0767XXXXXX

Click to show phone number

Rs 1,700,000 per perch

Address: Malabe Welivita

Land type: Residential

Land size: 10.12 perches

Description

10.12 perch land within close proximity to horizon campus (walking distance), Clear title, Local authority documents updated. 1.7mln pp

Figure 3.3 Land Sale advertisement posted in ikman.lk

As shown in figure 3.3, price per perch, location of the land, land size, land type, advertisement posted date are extracted through web scraping each land sale advertisement posted in the website during the given period.

3.1.2.Collection of data related to price determinants

Individual factors related to the characteristics of the land parcel(Derdouri and Murayama, 2020) such as size and shape; neighborhood factors related to the characteristics of land parcel including socioeconomic variables, external environment, and amenities; and location determinants depicting central business district are three main grouping or categories of the variables affecting land prices. Further, variables derived based on type of land, contact with road, distance from railroad, distance from waste treatment facilities and big projects (Kim and Kim, 2016) are also important considerations when estimating the land prices.

Based on these findings and by considering the availability of data in local context from verified and reliable sources like government websites; five different categories of land price determinants were identified for the proposed study including education, transportation, financial Institutions, healthcare facilities and utilities. Even though previous land price prediction studies have utilized variables pertaining to above mentioned price determinant categories, none of the studies have considered the sub-categorical available within each main category of land price determinants. Therefore, in order to fill that void, proposed study focused on creating variables based on the subcategories within the main categories of land price determinants identified based on previous studies. Table 3.1 shows the different subcategories considered in the proposed study for variable creation.

Table 3.1 Categories and subcategories of price determinants

Category	Sub-Categories
Education	Government Schools-Class A
	Government Schools-Class B
	Semi-Government Schools
	International Schools
	Universities
Transportation	Expressway entrances
	Railway Stations
Financial Institutions	Banks
	Finance Companies
Healthcare Facilities	Government Hospitals
	Private Hospitals
	Private Medical Centers
Utilities	Supermarkets
	Fuel Stations

For each of above-mentioned subcategories, price determinant units were identified by referring to the previous studies and by considering the data availability in the local context. Table 3.2 summarizes the data sources of the above-mentioned subcategories, units of subcategories and rationale behind selecting those subcategories.

Table 3.2- Selection of units of price determinants

Sub-Categories	Sources of getting data	Units selected	Rationale behind selection
Government Schools-Class A	Web page of Colombo Zonal Education Office ("Colombo Zonal Education Office," n.d.) School Performance Indices -GCE A/L ("Statistics - Department of Examinations - Sri Lanka," n.d.)	159 Government Schools	Latest list of Government schools available in the web page of Colombo Zonal education office Government Schools were separated into Class A and Class B based on the performance index of GCE A/L. Class A- Top 20 performed Schools Bio, Physical sciences, Commerce and Arts stream Class B- rest of the schools
Government Schools-Class B			
Semi-Government Schools		30 Semi Government Schools	
International Schools	Official Website of 'The International Schools of Sri Lanka' Association ("TISSL - The International Schools of Sri Lanka," n.d.)	15 International Schools	International Schools which have the membership of 'The International Schools of Sri Lanka' (TISSL) which is the Association of the Premier International Schools in Sri Lanka
Universities	Official website of the University Grants Commission ("Universities," n.d.)	25 Universities	All the universities registered under the University Grants Commission
Expressway entrances	Official website of Expressway Operation Maintenance and Management Division-Road Development Authority ("EOM&M Division," n.d.)	14 Entrances	All the Expressway entrances located within Colombo district
Railway Stations	Official website of Sri Lanka Railways ("Station Details," n.d.)	126 railway stations	All the railway stations located in Colombo district

Banks	Official website of CBSL which mention on the authorized financial institutions in Sri Lanka (“Authorized Financial Institutions | Central Bank of Sri Lanka,” n.d.) Most valued local banks as per Brands Sri Lanka Ratings (“Sri Lanka 100 2022 | Brand Value Ranking League Table | Brandirectory,” n.d., p. 100)	65 BOC Branches, 98 People’s Bank branches, 63 HNB branches, 41 NSB branches, 62 Sampath bank branches, 33 NDB branches,48 Combank branches, 6 Cargills bank branches and 6 Amana bank branches	All the branches located in Colombo district of top 9 most valued banks in Sri Lanka
Finance Companies	Official website of CBSL which mention on the authorized financial companies in Sri Lanka (“Licensed Finance Companies | Central Bank of Sri Lanka,” n.d.)	112 branches of 6 Finance companies	All the branches located in Colombo district of the top 6 most valued Finance companies in Sri Lanka
Government Hospitals	Official website of the Ministry of Health Sri Lanka (“Ministry Of Health - HOSPITALS,” n.d.)	16 Hospitals	There exist 5 major Government Hospital Categories in the Colombo District- Base Hospital-Type A, Base Hospital-Type B, Teaching Hospital Divisional Hospital-Type A, Hospital for women, National Cancer Institute, National Eye Hospital, National Hospital, National Institute of Mental Health
Private Hospitals	Official website of the Private Health Services Regulatory Council (“Registered Institutes,” n.d.)	22 Private Hospital	Private Hospitals registered under Private Health Services Regulatory council of Sri Lanka
Private Medical Centers		30 Private Medical Centers	Private Medical Centers registered under Private Health Services Regulatory council of Sri Lanka
Supermarkets	Official Websites of respective supermarkets	240 outlets belonging to 6 supermarkets	Outlets of Keels, Cargills, Arpico, Laughfs and SPAR Supermarket Chains
Fuel Stations	Official websites of Lanka IOC, Ceypetco and Laugfs fuel stations	69 Fuel stations belonging to 3 fuel distributors	Considered Lanka IOC, CEPETCO and Laugfs Fuel Stations

3.1.3.Collection of data related to location coordinates

When deriving the variables from above-mentioned sub categories, previous studies(Kim and Kim, 2016),(Derdouri and Murayama, 2020) have shown that number of determinant units within a given radius and closest distance to nearest unit as two major considerations. Therefore, in order to extract these two metrices with respect to each identified price determinant unit, location coordinates of each price determinant unit were extracted by referring to google maps. Sample of the location coordinates obtained for some of the units of subcategories of Healthcare Facilities category are mentioned in table 3.3. This exercise is repeated for all the sub categories to derive the variables from each of these sub categories.

Table 3.3- Location coordinates of a sample of the units of the Healthcare category

Category	Sub-Category	Unit	Latitudes	Longitudes
Healthcare Facilities	Government Hospitals	National Cancer Institute Maharagama	6.8372496	79.9181309
		National Eye Hospital Colombo	6.9185823	79.8630093
		Castle Street Hospital for Women	6.9106465	79.8825794
	Private Hospitals	Kings Hospital (Pvt) Ltd.	6.8947568	79.8795433
		Asiri Hospitals Holdings PLC.	6.8949797	79.8867986
		Asiri Surgical Hospital PLC.	6.8946164	79.877354
	Private Medical Centers	Confidence medical centre	6.925883	79.86543
		Mediquick (Pvt) Ltd.	6.880489	79.86075
		Norris Clinic	6.921145	79.86397

Finally, by considering the location coordinates of these price determinant units and the location coordinates of each GN division; variables with respect to nearest determinant unit from each GN division and number of determinant units in the vicinity of each GN division were created. Variables created for each sub-category are mentioned in table 3.4

Table 3.4- Variables created under each sub-category of price determinants

Category	Sub-Categories	Variables Created
Education	Government Schools-Class A	1)No. of Class A Govt Schools within 5km radius 2)Distance to nearest Class A Govt School
	Government Schools-Class B	3)No. of Class B Govt Schools within 5km radius 4)Distance to nearest Class B Govt School
	Semi-Government Schools	5)No. of Semi Govt Schools within 5km radius 6)Distance to nearest Semi Govt School
	International Schools	7)No. of International Schools within 5km radius 8)Distance to nearest International School
	Universities	9)No. of universities within 5km radius 10)Distance to nearest university
Transportation	Expressway Entrances	11)Distance to the nearest Expressway entrance
	Railway Stations	12)Distance to the nearest railway station
Financial Institutions	Banks	13)No. of Bank branches located within 2km radius 14)Distance to the nearest bank branch
	Finance Companies	15)No. of Finance companies located within 2km radius 16)Distance to the nearest Finance company
Healthcare Facilities	Government Hospitals	17)No. of Government Hospitals located within 5km radius 18)Distance to the nearest Government Hospital
	Private Hospitals	19)No. of Private Hospitals located within 2km radius 20)Distance to the nearest Private Hospital
	Private Medical Centers	21)No. of Private Medical Centers located within 2km radius 22)Distance to the nearest Private Medical Center
Utilities	Supermarkets	23)No. of Supermarkets located within 2km radius 24)Distance to the nearest Supermarket
	Fuel Stations	25)No. of Fuel stations located within 2km radius 26)Distance to the nearest Fuel station

All the variables mentioned in table 3.4 were computed for each GN division in the Colombo district. Thereafter, these variables were merged with the variables extracted from web advertisements (section 3.1.1) through the location variable and formed the final dataset. Variables of the final dataset after merging data from two sources as mentioned in dataset preparation process are mentioned in table 3.5

Table 3.5 Variables for the final dataset

	Variable Name		**Variable Name**
1	Land Size (in perches)	16	No. of Bank branches located within 2km radius
2	Address (Location of the land)	17	Distance to the nearest bank branch
3	Type of the land	18	No. of Govt. Hospitals located within 5km radius
4	Price per perch	19	Distance to the nearest Govt. Hospitals
5	No. of Class A Govt Schools within 5km radius	20	No. of Finance companies located within 2km radius
6	Distance to nearest Class A Govt School	21	Distance to the nearest Finance Company
7	No. of Class B Govt Schools within 5km radius	22	No. of Pvt. Hospitals located within 2km radius
8	Distance to nearest Class B Govt School	23	Distance to the nearest Pvt. Hospitals
9	No. of Semi-Govt Schools within 5km radius	24	No. of Pvt. Medical Centers located within 2km radius
10	Distance to nearest Semi-Govt School	25	Distance to the nearest Pvt. Medical Center
11	No. of Intl. Schools within 5km radius	26	No. of Supermarkets located within 2km radius
12	Distance to nearest Intl. School	27	Distance to the nearest Supermarket
13	No. of universities within 5km radius	28	No. of Fuel Stations located within 2km radius
14	Distance to nearest university	29	Distance to the nearest Fuel Station
15	Distance to the nearest Expressway entrance	30	Distance to the nearest railway station

3.2. Data cleansing

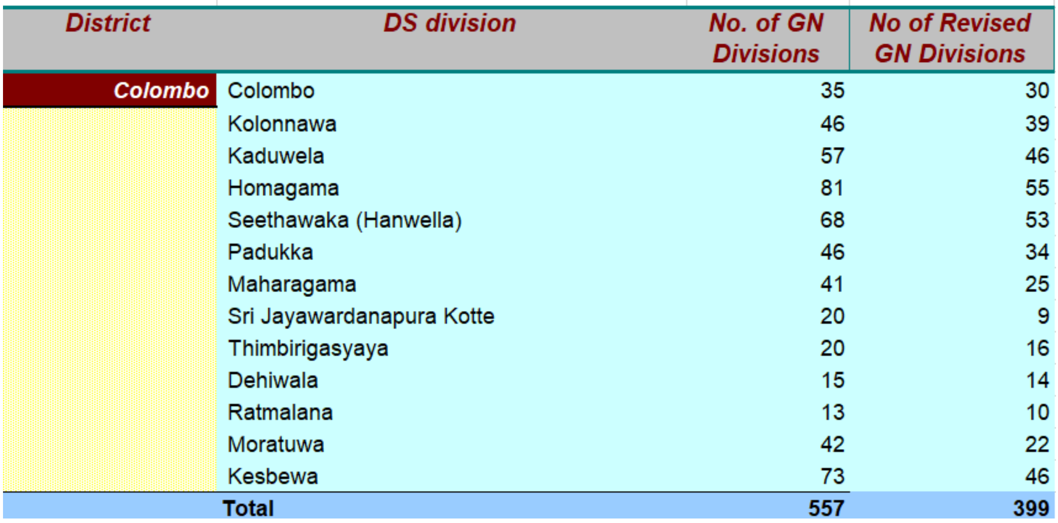
Data cleansing of this project primarily consisted of three main stages. Firstly, GN divisions were identified by removing the cardinality directions of the GN divisions. Thereafter improper location references available in the advertisements were converted to proper format and finally removed the outliers available in the dataset. Detailed explanations of the above three stages are mentioned below.

3.2.1. Data preprocessing related to location variable

Location of the land is highly important data as this relates to model generalization and variable derivation. Geographical area based for this study; Colombo district, consists of 13 Divisional Secretariat (DS) divisions and these DS divisions consist of 557 Grama Niladari (GN) divisions(“Grama Niladhari Division,” n.d.). In this study, GN division is used as the location of each land plot as it is the maximum granularity of location that can be extracted from each web advertisement posted in www.ikman.lk.

Out of the 557 GN divisions in Colombo district, some have repeated base names with cardinal directions; i.e. North, South, East and West. For example, Kotahena East and Kotahena West are considered as two separate GN divisions. However, such distinct identification of GN divisions cannot be extracted from the advertisement posted in the ikman.lk website (as per the given example, base name Kotahena is provided in the website advertisement as the location of the land plot which is located in either Kotahena East or Kotahena West). Therefore, GN divisions with same base name but different cardinal directions as suffix of the name were identified in each GS division and merged to form the distinct GN divisional names (in the example given earlier, both Kotahena East and Kotahena West were merged as Kotahena). This reduced the final distinct GN division count in Colombo District to 399 GN divisions. DS division-wise sorted summary of the above procedure is shown in the table 3.6

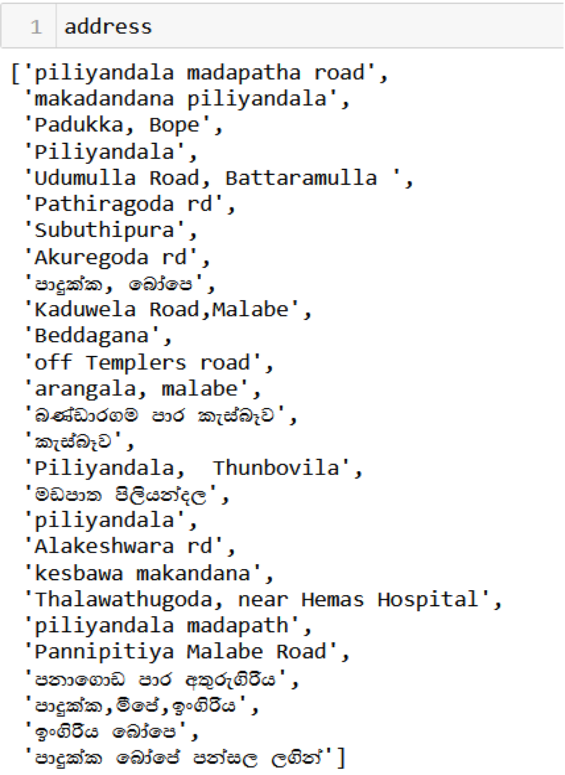
Table 3.6 Summary of the revised number of GN divisions by DS divisions



District	DS division	No. of GN Divisions	No of Revised GN Divisions
Colombo	Colombo	35	30
	Kolonnawa	46	39
	Kaduwela	57	46
	Homagama	81	55
	Seethawaka (Hanwella)	68	53
	Padukka	46	34
	Maharagama	41	25
	Sri Jayawardanapura Kotte	20	9
	Thimbirigasyaya	20	16
	Dehiwala	15	14
	Ratmalana	13	10
	Moratuwa	42	22
	Kesbewa	73	46
	Total	557	399

3.2.2.Cleansing of location references

After identification of the unique GN divisions, from each web advertisement, the GN division to which the advertised land belongs is identified through the location references. However, some of the advertisements did not contain these location references in proper format to be directly used in this study. As shown in figure 3.4, advertisements with location referenced in Sinhala language, advertisements with location referenced using the names of the nearby roads and advertisements with same GN division name in different spellings are some of the examples for improper location references available in the web advertisements.



```
address
```

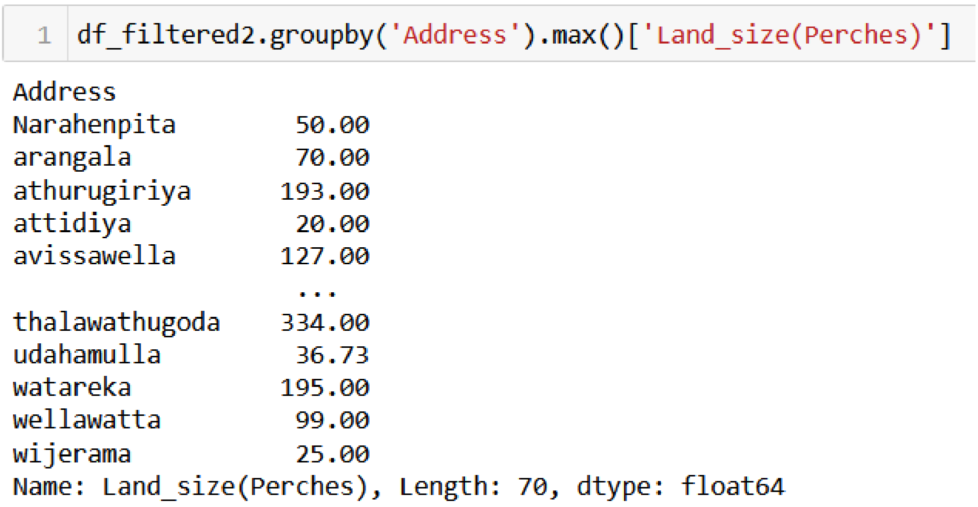
```
['piliyandala madapatha road',
 'makadandana piliyandala',
 'Padukka, Bope',
 'Piliyandala',
 'Udumulla Road, Battaramulla ',
 'Pathiragoda rd',
 'Subuthipura',
 'Akuregoda rd',
 'පාදුක්ක, බෝපෙ',
 'Kaduwela Road,Malabe',
 'Beddagana',
 'off Templers road',
 'arangala, malabe',
 'බණ්ඩාරගම පාර කැස්බෑව',
 'කැස්බෑව',
 'Piliyandala,  Thunbovila',
 'මඩපාත පිලියන්දල',
 'piliyandala',
 'Alakeshwara rd',
 'kesbawa makandana',
 'Thalawathugoda, near Hemas Hospital',
 'piliyandala madapath',
 'Pannipitiya Malabe Road',
 'පනාගොඩ පාර අතුරුගිරිය',
 'පාදුක්ක,මීපේ,ඉංගිරිය',
 'ඉංගිරිය බෝපෙ',
 'පාදුක්ක බෝපේ පන්සල ලගින්']
```

Figure 3.4 Location references in improper formats

Therefore, data cleansing was done to bring the location references into a uniform format through transformations including conversion of location names posted in Sinhala to respective English GN divisional names, extraction of GN divisional name by analyzing the road names and other location references mentioned in the advertisement, and merging of same location references with minor spelling mistakes to unified name.

3.2.3.Removal of the outliers

Outliers are the extreme values that reside outside the range of what is expected and show a major deviation from the rest of the values available for a particular variable. In the developed dataset, outliers were noted primarily with respect to 'price per perch' and 'land size' variables. However, all the records pertaining to 'price per perch' or 'land size' variable were not considered at once when identifying the outliers, as ranges of values vastly differ from one GN division to another. For an example, average price per perch in Kurunduwatta (Colombo 7) GN division is around LKR 18 Mn while average price per perch in Kadugoda GN division is around LKR 120,000. Similar behavior is observed with respect to the 'land size' variable where the maximum value of each GN division varies from one GN division to another as shown in figure 3.5.



```
df_filtered2.groupby('Address').max()['Land_size(Perches)']
```

```
Address
Narahenpita        50.00
arangala           70.00
athurugiriya      193.00
attidiya           20.00
avissawella       127.00
                   ...
thalawathugoda    334.00
udahamulla         36.73
watareka          195.00
wellawatta         99.00
wijerama           25.00
Name: Land_size(Perches), Length: 70, dtype: float64
```

Figure 3.5 Maximum value of the size of land in some of the GN divisions

Therefore, outliers were identified with respect to each GN division rather than identifying the outliers with respect to a variable. In each GN division, an observation with price per perch greater than or less than 2 standard deviations from mean were considered as an outlier(Zhang et al., 2021) and removed from the dataset. This procedure removed 136 observations from the total dataset. Similarly, 142 outliers from the dataset with respect to ‘land size’ variable were identified and removed. Altogether 278 observations were removed from the dataset as outliers.

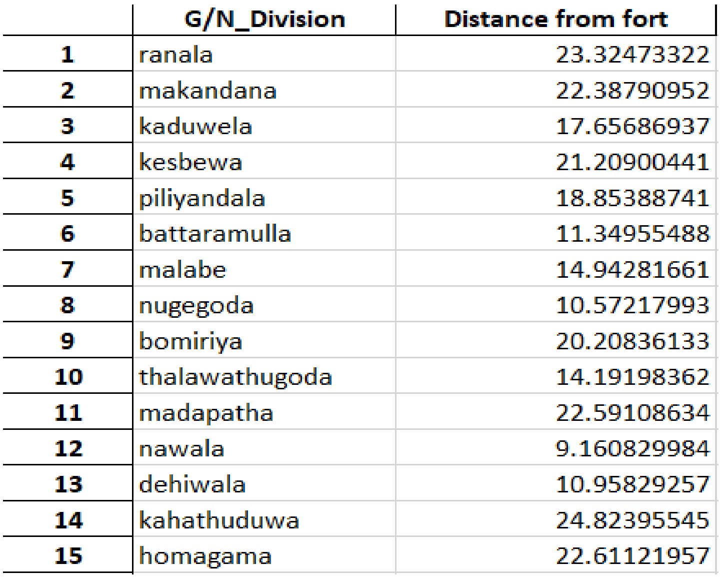
3.3. Data transformations

Data transformation is the phase which prepares the data available in the amalgamated dataset as the inputs with desired qualities for the machine learning modelling. In this study, transformation of the categorical GN divisional name into numerical representations, conversion of categorical ‘land type’ variable into numerical variables and scaling of variables were the main data transformations. Detailed explanations of these transformations are given in the following subsections.

3.3.1. Transformation of GN divisional names in to numerical form

Distinct GN divisions available in the data set were identified and obtained the location coordinates of each of these through google maps. As the model generalization is required for future predictions, the name of the GN divisions needed to be replaced with a quantitative variable generated based on some distance measure. Previous literature(Derdouri and Murayama, 2020) too showed that when computing these distances, location to the central business district can be considered as a base measure. Therefore, distance from each of the GN divisions to Fort-Colombo, which is the location generally referred to when calculating distances in Sri Lanka, was calculated considering the difference of location coordinates and stored as a separate variable for the purpose of model generalization. Some of the calculated distances from Fort to respective GN divisions are mentioned in table 3.7

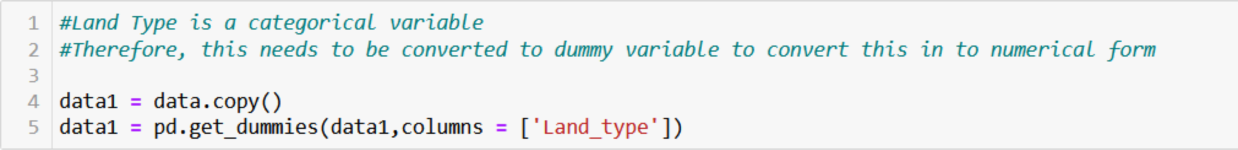
Table 3.7 GN Divisions and distance from Fort to each GN division



	G/N_Division	Distance from fort
1	ranala	23.32473322
2	makandana	22.38790952
3	kaduwela	17.65686937
4	kesbewa	21.20900441
5	piliyandala	18.85388741
6	battaramulla	11.34955488
7	malabe	14.94281661
8	nugegoda	10.57217993
9	bomiriya	20.20836133
10	thalawathugoda	14.19198362
11	madapatha	22.59108634
12	nawala	9.160829984
13	dehiwala	10.95829257
14	kahathuduwa	24.82395545
15	homagama	22.61121957

3.3.2.Conversion of categorical variable in to numerical variable

There exist two types of categorical variables; nominal and ordinal. Nominal variables have no intrinsic ordering to its categories while ordinal variables have a clear ranking for the values available. In machine learning modelling, as machines can only understand the numbers, categorical columns need to be converted into numerical representations in order for machine learning algorithms to properly understand. This process is known as categorical encoding. Two of the most widely used techniques for categorical encoding are label encoding and one hot encoding. As label encoding assigns integers in ordinal manner for each category within a variable, this type of encoding is considered more appropriate for categorical variables with ordinal classes. However, one hot encoding creates a separate variable for each distinct value available within the categorical variable and each record is represented by a binary notation. One hot encoding doesn't assign any ordering for the values available in a given categorical variable. In this study, 'Type of land' is the only categorical variable available in the dataset. This categorical variable consists of 4 main categories, namely; Agricultural, Commercial, Residential and other. However, as some of the advertised lands were categorized under multiple categories, there exist 14 combinations of categories (values) under the 'Type of land' variable. As the ‘Type of land’ variable is a nominal categorical variable, one hot encoding was utilized to convert this variable to 14 binary variables. Code snippet used to encode the ‘Land_type’ variable and the variables resulted due to encoding of ‘of Land Type’ variable are shown in figure 3.6 and table 3.8 respectively.



```
#Land Type is a categorical variable
#Therefore, this needs to be converted to dummy variable to convert this in to numerical form

data1 = data.copy()
data1 = pd.get_dummies(data1,columns = ['Land_type'])
```

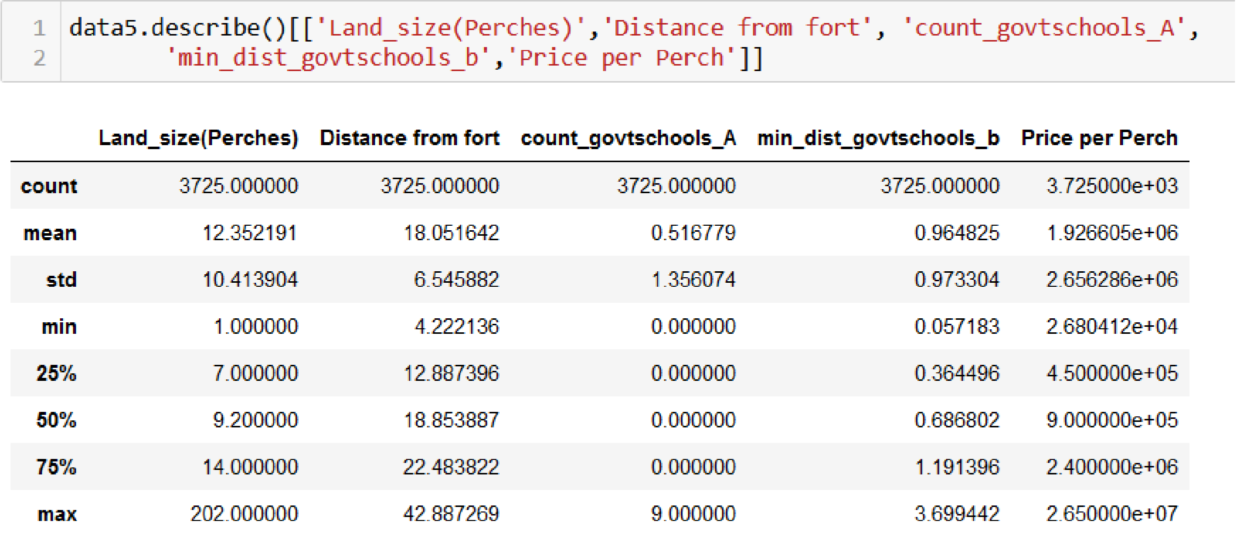
Figure 3.6 Code used to encode the ‘Land Type’ variable

Table 3.8 Binary variables created after encoding of ‘Land Type’ variable

1	Land_type_Agricultural	8	Land_type_Commercial
2	Land_type_Agricultural, Commercial	9	Land_type_Commercial, Other
3	Land_type_Agricultural, Commercial, Other	10	Land_type_Commercial, Residential
4	Land_type_Agricultural, Commercial, Residential	11	Land_type_Commercial, Residential, Other
5	Land_type_Agricultural, Commercial, Residential, Other	12	Land_type_Other
6	Land_type_Agricultural, Residential	13	Land_type_Residential
7	Land_type_Agricultural, Residential, Other',	14	Land_type_Residential, Other

3.3.3.Log transformation of variables

Variables that are measured at different scales do not contribute equally to the models being fit and could end up creating bias. Thus, to deal with this potential problem, feature scaling is used prior to model fitting.



```
data5.describe()[['Land_size(Perches)','Distance from fort', 'count_govtschools_A',
        'min_dist_govtschools_b','Price per Perch']]
```

	Land_size(Perches)	Distance from fort	count_govtschools_A	min_dist_govtschools_b	Price per Perch
count	3725.000000	3725.000000	3725.000000	3725.000000	3.725000e+03
mean	12.352191	18.051642	0.516779	0.964825	1.926605e+06
std	10.413904	6.545882	1.356074	0.973304	2.656286e+06
min	1.000000	4.222136	0.000000	0.057183	2.680412e+04
25%	7.000000	12.887396	0.000000	0.364496	4.500000e+05
50%	9.200000	18.853887	0.000000	0.686802	9.000000e+05
75%	14.000000	22.483822	0.000000	1.191396	2.400000e+06
max	202.000000	42.887269	9.000000	3.699442	2.650000e+07

Figure 3.7 Comparison of range of the ‘Price per perch’ variable with some of the other variables

Log transformation is used to reduce the variability in data, especially when there exist outlying observations(Feng et al., 2014). This makes the variable to which the log transformation applied easy to handle and make the model evaluations more interpretable. As shown in figure 3.7, ‘Price per perch’ variable has a wide range of values compared to other variables in the dataset which spans from Rs.26,804 per perch to Rs.26,500,000 per perch. If the model is evaluated by utilizing the ‘Price per perch’ variable without log transformation, MSE and RMSE values could get very high due to this variation in ranges and misinterpretation of model error could happen. Therefore, log transformation of price per perch variable was done in order to make the variations of ‘Price per perch’ variable more interpretable.

3.4. Machine learning modelling

Selection of suitable machine learning algorithms to learn the complex relationship between the land prices and potential determinants is one of the most significant steps in any land price prediction study. Prior studies (Zhang et al., 2021) , (Derdouri and Murayama, 2020) , (Levantesi and Piscopo, 2020) have shown that tree based algorithms outclassed most of the other linear and non-linear machine learning algorithms for land and real estate price predictions in terms of performance and accuracy. In a study conducted in Wuhan to predict the residential land prices utilizing point of interest and night time light datasets, Zhang (Zhang et al., 2021) has explored the ability of Machine learning algorithms by developing several land price prediction models based on five machine learning algorithms, namely; support vector regression(SVR) with radial basis function, SVR with a linear kernel, random forests regression, extra trees regression (ETR) and multiple linear regression. Experimental results showed that Support vector regression with radial basis function and ETR algorithms have the best prediction performance irrespective of the different temporal regions. Further, a study conducted in Fukushima prefecture (Derdouri and Murayama, 2020) to map geostatistical data to land prices by utilizing generalized linear model, generalized additive model using splines, support vector machines with linear kernel, multivariate adaptive regression spline, k nearest neighbors, support vector machines with radial basis function kernel, cubist, stochastic gradient boosting and random forests have showed that better performance of the random forests relative to other considered ML algorithms in terms of all errors and accuracy indicators. Levantesi and Piscopo (Levantesi and Piscopo, 2020) too showed that predictive models based on Random forest performed better compared to generalized linear model-based regression approach to evaluate the importance of economic variables on the London Real estate market. However, a more advanced version of regression forests was proposed by Córdoba(Córdoba et al., 2021) by the name spatial Quantile Regression Forests (sQRF) to model the relationship between a set of environmental variables and rural land values. He experimentally showed the superior performance of his proposed model over linear regression, regression kriging and spatial random forest algorithms through several model validation measures

Artificial neural networks are identified as one of the most widely used machine learning algorithms in land valuation. Accuracy of the neural networks is considered better(Demetriou, 2017) than linear regression models in some instances. Random Forests and Quantile random forests algorithms which are powered by resampling are considered more resistant to overfitting and more robust to noise in the data than regular regression tree models, hence these algorithms have been used for mass appraisal in residential real estate(Wang and Li, 2019). Concept of big data (Singh et al., 2020) has been used to predict housing sales by utilizing three models; linear regression, random forests, and gradient boosting. The numerical results indicated that the gradient boosting model outperforms other forecasting models in terms of forecasting accuracy.

Rather than comparing the results of different statistical or Machine Learning models, Kaliola (Kalliola et al., 2021) attempted to predict real estate prices through hyperparameter optimization of Neural networks. Different options of the hyper-parameter values were investigated and the analysis showed that improvement can be obtained in the model performance through hyperparameter tuning. A house price prediction model developed by Park and Bae(Park and Bae, 2015) by utilizing machine learning approaches indicated that repeated incremental pruning obtains more accurate forecasting results than the other forecasting methods.

As per above mentioned previous studies of similar nature; tree-based machine learning algorithms such as Extra trees regression, Random forests regression algorithms and non-linear algorithms like support vector regression have consistently shown excellent performance in numerical predictions. Furthermore, literature has also shown the importance of training and testing the dataset with a linear algorithm in order to compare and verify the performance advantages of nonlinear Machine Learning algorithms. Considering these facts, two tree-based machine learning algorithms, namely; Random forests (RF) and Extra trees regression (ETR), a non-linear machine learning algorithm; Support vector regression (SVR) and an ensemble-based Machine learning algorithm; Extreme gradient boosting (XGBoost) were utilized in this study to fit the developed land price dataset. Further, a linear machine learning algorithm, Multiple linear regression is used as the base model to compare the performance of other types of machine learning algorithms with a linear model. Detailed explanation of the machine learning algorithms that are used in this study is given below.

3.4.1.Support Vector Regression

SVR is a supervised machine learning algorithm for prediction and curve fitting for both linear and nonlinear regression types. SVR is based on the support vectors which are points closer to the generated hyperplane in an n-dimensional feature space that distinctly segregates the data points about the hyperplane(Parbat and Chakraborty, 2020). SVR has unique advantages for small datasets and can maintain a good generalization ability. SVR is composed of two main components; hyperparameters and the kernel functions(Zhong et al., 2019). Hyperparameters determine the support vectors while the kernel functions determine the properties of high-dimensional feature spaces. In order to improve the performance of SVR, hyperparameters should be optimized by utilizing methods such as grid search.

As SVR is a kernel-based algorithm, performance of SVR relies heavily on kernel functions. Commonly used kernel functions in prediction models are radial basis function, polynomial function and linear function which project input data into high-dimensional feature space.

3.4.2.Random Forests

RF is a tree-based ensemble method and was developed to address the shortcomings of traditional classification and regression tree (CART) method(Ahmad et al., 2018). RF consists of a large number of weak decision tree learners, which are grown in parallel to reduce the bias and variance of the model. Random forests algorithm is applied by getting bootstrapped sample sets from the original dataset and growing an unpruned regression trees from each bootstrapped sample. In this step, a fixed number of randomly sampled predictors are used as split candidates instead of using all available predictors. These steps are repeated until a sufficient number of such trees are grown, and new data is predicted by aggregating the prediction of those trees. Bagging is used in RF to increase the diversity of the trees by growing them from different training datasets which results in reducing the overall variance of the model. RF facilitates to assess the relative importance of input features, which is useful when dimensionality reduction is required.

Random Forests is considered as a powerful and useful model for prediction because it can handle high-dimensional databases and complex variables(Chowdhury, n.d.). For regression analysis, random forest is considered as a better algorithm mainly due to reduction in the classification error and the rate of overfitting.

3.4.3.Extra Trees Regression

ETR is developed as an extension of the random forest algorithm which is a relatively recent machine learning technique(Ahmad et al., 2018). ETR employs the same principle as random forests and uses a random subset of features to train each base estimator. In contrast to regression forests which use bootstrap replicas to train the model, ETR uses a whole training dataset to train each regression tree.

3.4.4.Extreme Gradient Boosting

XGBoost is a scalable tree optimization machine learning methodology that has been evolved recently in data analysis disciplines. The boosting concept is the root of this algorithm(Alshboul et al., 2022), which merges the forecasting of weak learners with additive training methods to develop a strong learner. XGBoost architecture is organized in such a way that simplified objective functions allow the prediction and regularization terms to be combined while preserving the fastest possible processing speed.

3.4.5.Multiple Linear Regression

MLR is utilized to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. As MLR is based on several assumptions like homogeneity of variance, independence of observations, normality and linearity; several transformations should be done with the data being modelled before applying the MLR algorithm. When there are many explanatory variables it is essential to log transform for linear conversion. The data cleaning is necessary for linear regression to remove the noise and the outliers in the numerical output variable(Chowdhury, n.d.). Pairwise correlation should be utilized to identify highly correlated variables before applying the algorithm in order to avoid the overfitting problem. Literature also states that standardization of numerical input variables too could lead to more accurate predictions.

3.5. Model performance optimization

Performance improvements in machine learning modelling is the process of improving the prediction accuracy of the respective algorithm. There exist many techniques to improve performance in a machine learning model and some of the most important techniques highlighted in literature are discussed below.

3.5.1.Train test split of the dataset

Accommodation of too much noise in the dataset to the underlying model causes a model to become overfit and non-accommodation of the general patterns in the dataset to the underlying model causes a model to become underfit. In order to address this issue and to strike a balance between overfitting and underfitting of the given dataset to underlying models, the dataset of this study was split into two parts; namely as a train set and a test set. 70% of the observations were used as the train set to train the model and the remaining 30% of the observations were utilized to test the model and evaluate the performance of different machine learning models. Dataset in this study was split by utilizing the in-built “train_test_split” module available in scikit-learn library of the python package.

3.5.2.Feature Selection

Process of selecting the most relevant, consistent and non-redundant features out of all the available features in a dataset to use in model construction is known as the feature selection. The main objective of feature selection is to improve the predictive accuracy of the model while reducing the computational cost of modelling. Feature selection algorithms can be broadly classified into two categories(Refaeilzadeh, n.d.); filter methods and wrapper methods. Filter methods are less computationally expensive, depend on some intrinsic characteristics of data and selection of features in this method is independent of any machine learning algorithms. Wrapper methods on the other hand, are more computationally expensive than filter methods and depend on a specific machine learning algorithm to find the optimum subset of features. Comparisons of filter and wrapper methods(Suto et al., 2016) have shown that filter methods are frequently used in many machine learning applications because of their applicability to any type of machine learning technique and faster execution. However, this also has highlighted the fact that wrapper methods are more efficient than the filter methods as they take into consideration the classifier hypothesis and handle feature dependencies.

Zohre (Ebrahimi-Khusfi et al., 2021) employed wrapper methods, namely; Boruta, Multivariate Adaptive Regression Splines (MARS) and recursive feature elimination (RFE) to select the most influencing variables to forecast number of dusty days in desert wetlands. His study revealed that RFE is a robust method for extracting all relevant features than the other mentioned feature selection methods and is capable of modelling nonlinear relationships and modifying the collinearity effect between the independent variables. Further, he showed that Boruta and MARS are sensitive to the existence of correlation between predictive variables. However, rather than depending on a single methodology to select the best feature subset, utilization of two different approaches and comparison of resulting model performances is highlighted in the literature (Refaeilzadeh, n.d.),(Ebrahimi-Khusfi et al., 2021),(Suto et al., 2016) as a more effective feature selection approach.

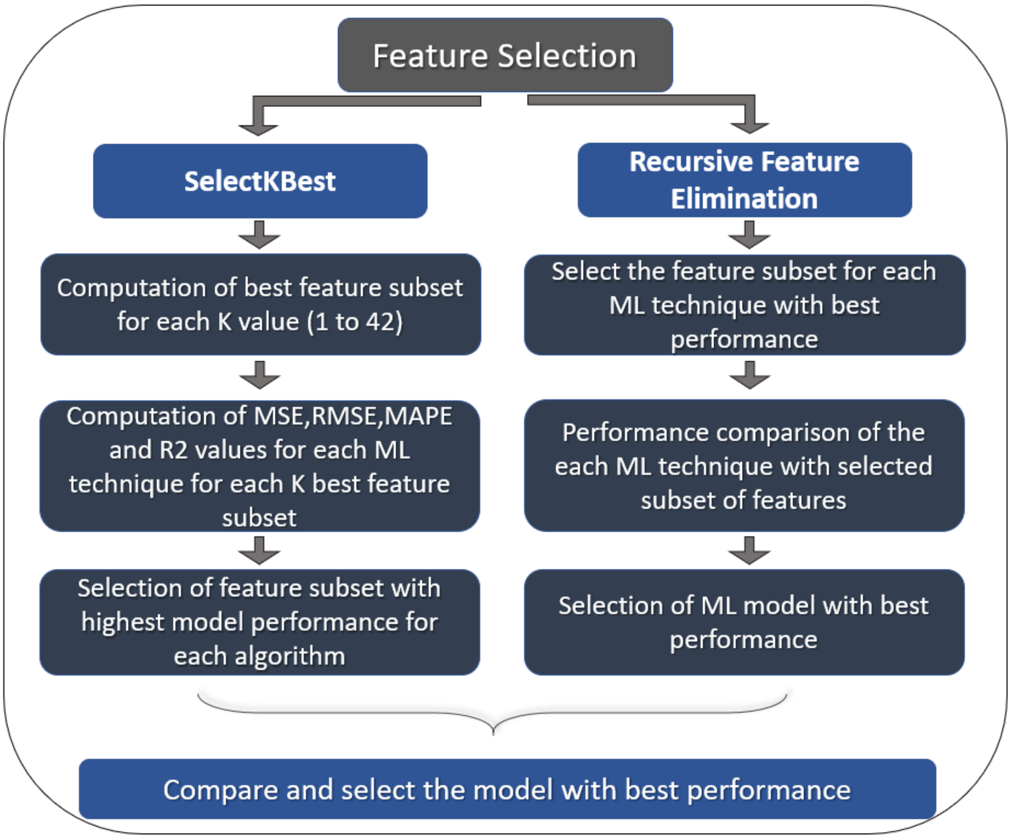

Feature Selection
SelectKBest
Recursive Feature Elimination
Computation of best feature subset for each K value (1 to 42)
Select the feature subset for each ML technique with best performance
Computation of MSE,RMSE,MAPE and R2 values for each ML technique for each K best feature subset
Performance comparison of the each ML technique with selected subset of features
Selection of feature subset with highest model performance for each algorithm
Selection of ML model with best performance
Compare and select the model with best performance

Figure 3.8 Feature selection process

Therefore, in this study, a wrapper method; Recursive feature elimination with cross validation (RFECV) and a Filter method; SelectKBest, were utilized to select the most influencing features for the models being built and the best method was selected based on the model performances. Summary of the approach followed in the current study is mentioned in figure 3.8

3.5.3.Hyperparameter Optimization

Hyperparameters are the configuration variables that control the model training process but do not change during the model training process. Hyperparameter optimization provides optimized values for hyperparameters which maximize the predictive accuracy of the model. K-fold cross validation has been used commonly (Ahmad et al., 2018) as a resampling procedure to find optimal values for a model's hyper parameters on a limited data sample before evaluating the performance of machine learning models. This is a scientific approach to define the train-test split in machine learning modelling and helps to find the best parameters for models being built. This procedure has a single parameter 'k' that defines the number of samples to which the dataset will be divided (i.e. if k=10, dataset will be divided into 10 samples). The value of k is chosen such that each train/test group of data samples is large enough to be statistically representative of the broader dataset. In k-fold cross validation, initially the dataset is shuffled randomly and thereafter is be divided into k samples and one sample is selected as the test set and remainder of samples will be considered train sets. The model is fitted to this train test split and the accuracy of the prediction is noted. Thereafter, a sample different to the initially selected sample is selected as the train set and the same procedure is followed. This process needs to be iterated until all the initial samples become test sets. Finally, the results of the iterations are summarized and used for model evaluations. Literature considers k-fold cross validation prone to less variation("Making Predictive Models Robust," n.d.) as it uses the entire training dataset for train-test split and subsequent evaluations. However, higher computational costs and need of the model to be trained k times in case of larger k are considered difficulties that a researcher could face when validating a developed model using k-fold cross validation.

In the scikit learn package of Python, two main hyperparameter optimization techniques are available; namely, GridSearchCV and RandomSearch CV. In GridSearchCV, every combination of the pre-defined list of hyper-parameter values are tested and choose the best combination based on the cross-validation score. Main advantage of this approach is that it provides the best combination of hyper parameters from the given set of hyper-parameters. However, as this method tests for all possible combinations of the given list of hyper-parameters, this method is considered computationally expensive and time consuming for datasets with larger samples. In contrast, RandomSearchCV tests only for random combinations of the pre-defined list of hyper-parameters and chooses the best out of the sample taken. Further, this method has a higher probability of providing the optimal set of hyperparameters from a randomly selected combination. As the dataset considered for this study has nearly 3800 observations with respect to 43 variables, Grid search could consume more time and computational resources. Therefore, by considering the complexity of the dataset and time factor, RandomSearchCV was utilized in this study for hyperparameter optimization purposes.

3.6. Model Evaluation

When creating machine learning models, evaluation of the model accuracy is an important component as this mechanism represents how accurate the model is performing in its predictions. In accuracy evaluation(Chowdhury, n.d.), predicted value through the machine learning model is compared with the original target by utilizing a set of metrices. A performance metric is defined as a logical and mathematical construct (Botchkarev, 2019) designed to measure how close are the actual results to the expected or predicted results. These metrices help us to understand how accurate the predictions are and what is the deviation it has from the actual values.

Regression predictive models involve predicting a numerical value and the performance of the regression model must be reported as an error score. This error score will convey how close our predictions are to the expected values. A high value of the error score generally means our model performed poorly and a low value generally means our model performed well.

To evaluate the accuracy of predictions of random forests, XG boost and support vector regression algorithms, Chowdury(Chowdhury, n.d.) utilized MAE, MSE and RMSE as evaluation metrices. In his study, random forests showed the least value for all three metrices. Among three metrices for random forests, MSE got the minimum value. However, the prediction accuracies of the model were enhanced after cross validation where the model accuracy further changed by 0.16% to 1.1%. Zhou (Zhou et al., 2022) has utilized R^2, MAE, RMSE and mean absolute percentage error (MAPE) to evaluate the performance of multiple linear regression (MLR) , artificial neural networks (ANN), K nearest neighbors (KNN), random forest (RF) and support vector regression (SVR) machine learning algorithms. His comparison revealed that machine learning methods have very high R^2 compared to other methods. Zhou further showed that the accuracy of regression-based modelling despite the addition of other predictor variables does not improve compared to MLR. However, in his study KNN showed the most accurate predictions compared to other algorithms which contradicts with Chowdury's (Chowdhury, n.d.) findings.

These scenarios in literature show the importance of the proper evaluation metrices to derive an error score which is critical for training an accurate model. Certain metrices have properties that help the model to learn in a specific manner. Some of these metrices put more weight on outliers and others will put more weight on the majority. Further, rather than relying on one single metric, it is important to evaluate a model on a set of metrices (Botchkarev, 2019) as there is no exact metric which suits best for all scenarios. Khaledian (Khaledian and Miller, 2020) has also shown the importance of using few metrices together to compare the performance of machine learning models as these metrices of model accuracy respond differently to different patterns of error within the dataset. Mean squared error (MSE), mean absolute error (MAE) ,root mean squared error (RMSE), mean absolute percentage error (MAPE) and coefficient of determination (R2) are identified in literature(Zhong et al., 2019)(Ahmad et al., 2018)(Botchkarev, 2019) as the error metrices that are being commonly used to quantify and evaluate the performance of the individual machine learning algorithms by comparing the predicted results with actual results.

Based on the above-mentioned facts and references, pros and cons of each method, nature of the predictions and nature of the algorithms utilized; MSE, RMSE, MAPE and R2 are used in this study to evaluate the performance of the machine learning algorithms. Brief descriptions of each evaluation metric used are mentioned below.

3.6.1.Coefficient of Determination (R^2)

Coefficient of determination can be interpreted as the percentage of variance in the response variable of a model that can be explained by the predictor variables. Therefore, a higher value of R^2 generally considers a better fit of the dataset to a given model. Coefficient of determination can be given through the following equation (1);

$$r^2 = 1 - \frac{\Sigma\,(y - y')^2}{\Sigma\,(y - \overline{y}')^2} \qquad (1)$$

However, there exist exceptions. If the R^2 value is low but the predictors are statistically significant, conclusions can still be made on how changes in the predictor values are associated with changes in response value. On the other hand, high R^2 does not necessarily indicate that the model has a good fit. This happens when a linear model is fitted to a dataset with nonlinear variations.

3.6.2.Mean Squared Error

The Mean Squared Error (MSE) is calculated by getting the difference between the model’s predictions and the actual values, squaring it and averaging it across the whole dataset. Lower MSE indicates a better fit of the dataset to a given model. MSE is formally defined by the following equation (2):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (2)$$

In the given equation, ‘N’ is the total number of observations and ‘y_i’ and ‘$y^\wedge_i$’ are the i^{th} value of experimental and predicted data. MSE helps to ensure that the trained model has no outlier predictions with significant errors as it puts larger weight on these errors through the squaring part of the function. However, if the model makes an extremely bad prediction, the squaring part of the function magnifies the error. Therefore, if the dataset contains many outliers, careful investigation should be carried out before implementing this evaluation metric.

3.6.3.Mean Absolute Percentage Error (MAPE)

Mean absolute percentage error is a relative measure that scales mean absolute deviation to be in percentage units instead of the variable’s units. MAPE is generally defined by the following equation (3), where ‘n’ is the number of fitted points, ‘A_t’ is the actual value and ‘F_t’ is the forecast value.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \qquad (3)$$

MAPE is often effective in analyzing large sets of data, but cannot be utilized if the dataset contains null values. This is because the calculation would require division by zero, which is impossible. MAPE is an easy to interpret evaluation metric that depends specifically on the data being evaluated, providing an accurate assessment on the reliability of the forecast

3.6.4.Root Mean Squared Error

RMSE is computed by taking the square root of MSE. RMSE measures the average magnitude of the errors and is concerned with the deviations from the actual value. Lower RMSE indicates a better model fit. RMSE can be given through the following equation (4), where ‘y_i’ and ‘$y^\wedge_i$’ are the i^{th} value of experimental and predicted data.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (4)$$

RMSE is computationally efficient and does not penalize the errors as MSE due to the consideration of the square root. Further, RMSE is measured in the same units as the response variable. Thus, the interpretation is much more straightforward than MSE. However, one of the major drawbacks of RMSE is its sensitivity to outliers.

When assessing how well a model fits a dataset, literature suggests calculating both the RMSE and the R^2 values as each metric interprets some different aspect. RMSE depicts how well a regression model can predict the value of the response variable in absolute terms while R^2 shows how well a model can predict the value of the response variable in percentage terms.

3.6.5.Model performance evaluation approach

In this study, overall performance evaluation is done based on the approach summarized in figure 3.9. by utilizing the available data, ML algorithms, feature selection methods and model evaluation techniques.

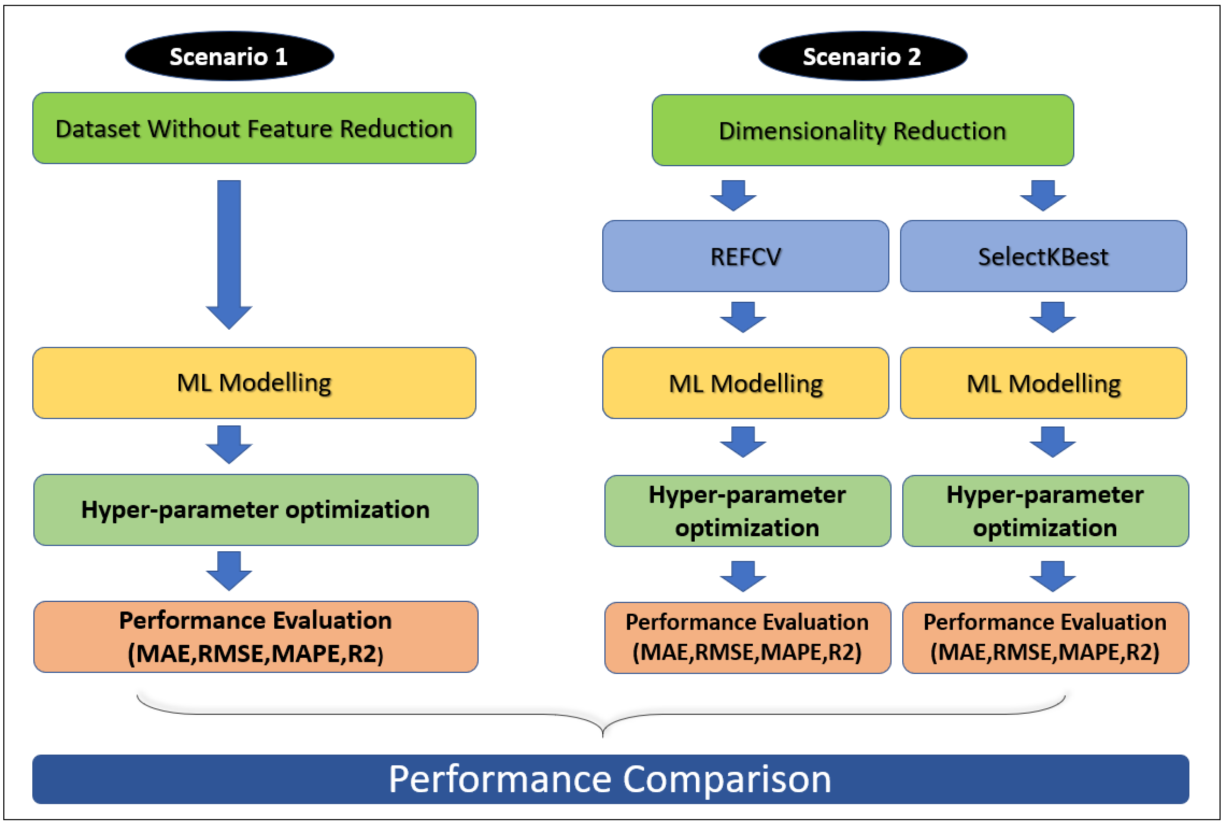

Scenario 1
Scenario 2
Dataset Without Feature Reduction
Dimensionality Reduction
REFCV
SelectKBest
ML Modelling
ML Modelling
ML Modelling
Hyper-parameter optimization
Hyper-parameter optimization
Hyper-parameter optimization
Performance Evaluation (MAE,RMSE,MAPE,R2)
Performance Evaluation (MAE,RMSE,MAPE,R2)
Performance Evaluation (MAE,RMSE,MAPE,R2)
Performance Comparison

Figure 3.9 Model performance evaluation

ML models of this study were built and evaluated based on two scenarios as shown in figure 3.9. As the scenario 1, raw dataset without any dimensionality reduction is utilized as the benchmark and the resulting performance is compared with the performance of the feature reduced datasets, which is referred to under scenario 2. Feature reduction is done based on RFECV and SelectKBest methods as discussed in the feature selection section of the methodology. Finally, the performance of each resulting model is compared against the benchmark model and selected the best performing model.

3.7. Chapter Summary

Theories and methodologies which have been used, starting from the data collection for the dataset preparation to machine learning model evaluation are explained in this chapter. Furthermore, data preparation and cleansing done prior to analysis were also explained. This chapter facilitates the reader to get a good understanding about the flow of the study and the logic behind selection of methodologies. Next chapter will present the analysis conducted on the developed dataset by following the methodology explained in this chapter.

4. Evaluation

This chapter will firstly describe the preliminary descriptive analysis done prior to the machine learning modelling on some of the important variables. Thereafter, the model building by utilizing machine learning algorithms, feature selection and hyperparameter optimization to achieve an optimized model will be discussed.

4.1. Preliminary analysis

Based on the methodology proposed in the previous chapter, the land price dataset was prepared with 3725 records which span over 43 variables (including dummy variables). Before the application of the machine learning algorithms, distribution of the price variable; which is the response variable in this study, is visualized to get an idea on the price distribution in different grama niladari divisions in Colombo district. Figures 4.1 and 4.2 show the per perch prices of highest priced and lowest priced lands in Colombo district.

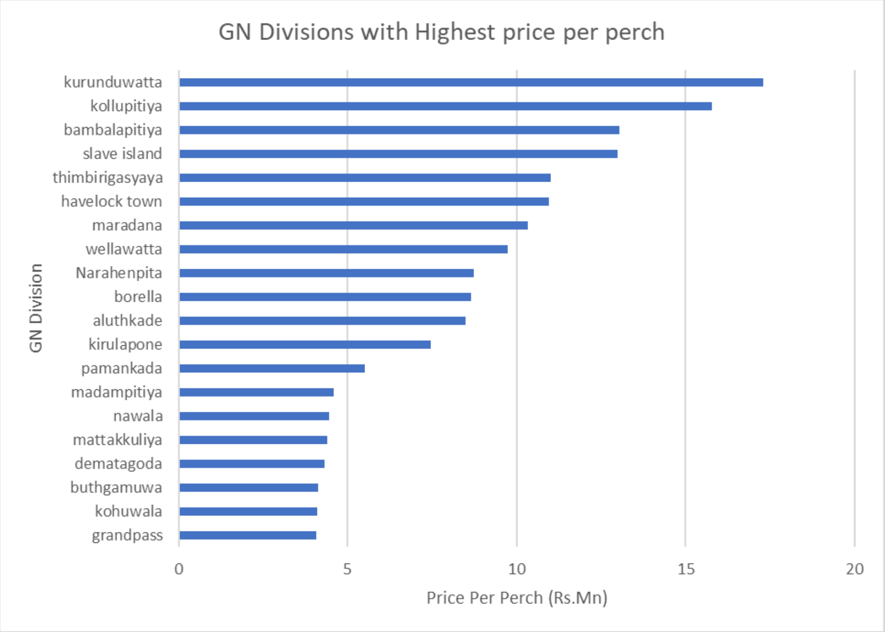


Figure 4.1 GN divisions with the highest price per

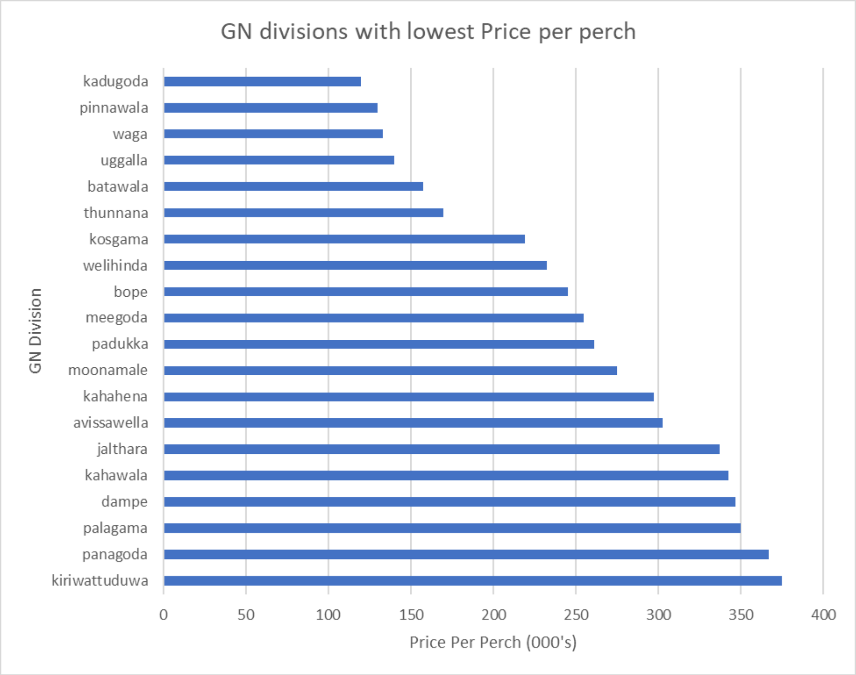


Figure 4.2 GN divisions with the lowest price per

Figure 4.1 shows that Kurunduwatta, Kollupitiya and Bambalapitiya as the grama niladari divisions with highest price per perch in Colombo district and figure 4.2 shows that Kadugoda, Pinnawala and Waga as the grama niladari divisions with lowest price per perch in Colombo district. Above two figures show that the land prices utilized to train this model spans over a wide range of prices from Rs. 125,000 per perch to Rs.1,800,000 per perch.

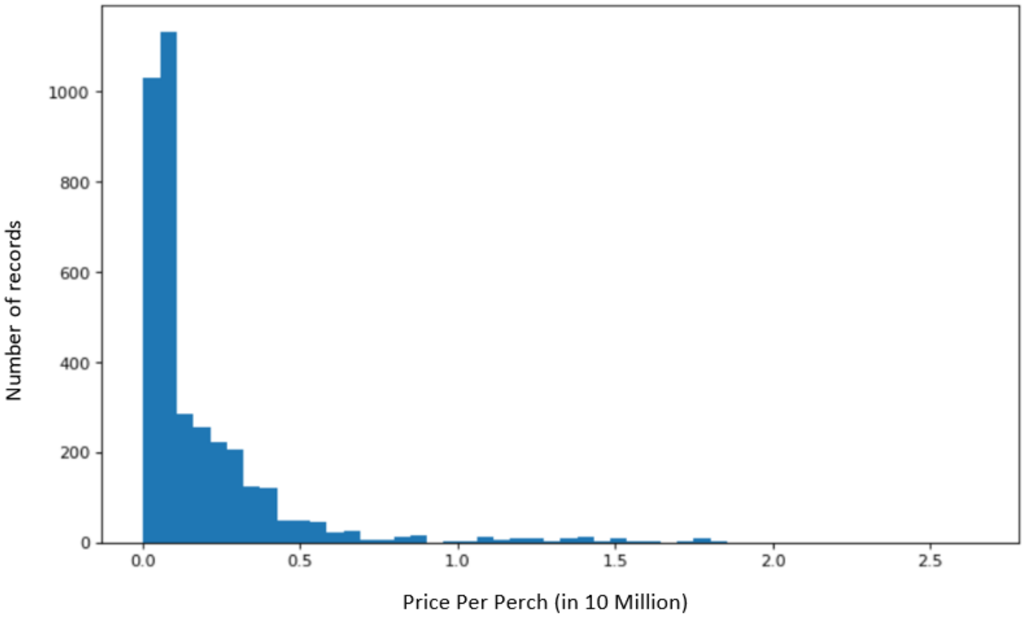


Figure 4.3 Distribution of land prices

Distribution of land prices among the 3725 records is shown in figure 4.3. This shows the response variable of this study; price per perch is positively skewed. This is because the majority of the advertised lands are priced below Rs. 2,500,000 per perch and there exist relatively fewer records available from 5,000,000 to Rs. 18,000,000 per perch.

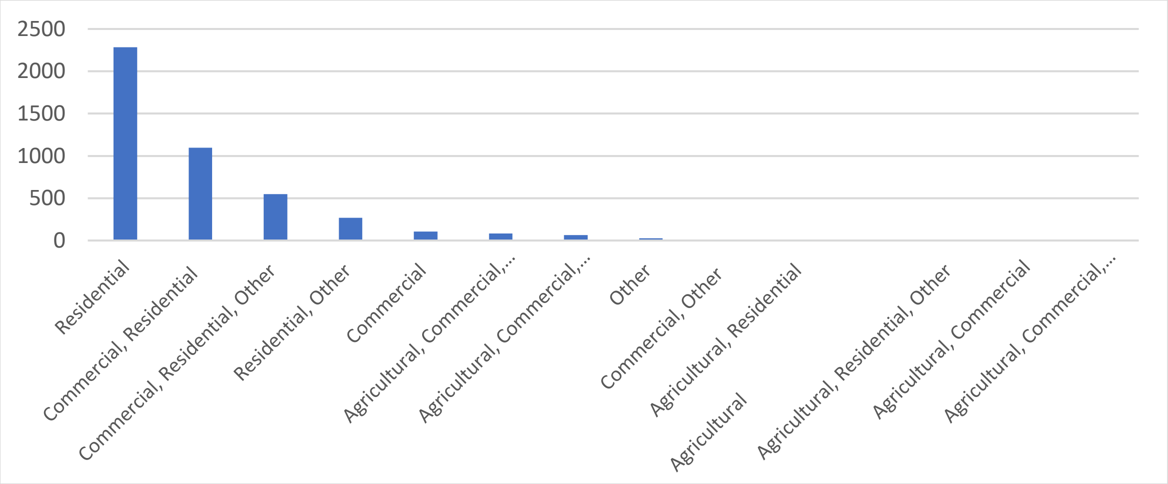


Figure 4.4 Counts of land types in each category

Figure 4.4 shows the counts of lands distributed across different land type categories available in the dataset before the application of variable encoding. As shown here, even though there exist 14 sub categories of land types, more than 90 % of the lands available in this dataset are distributed across 4 sub categories of land types, namely; 'Residential', 'Commercial, Residential', 'Commercial, Residential, Other' and 'Residential, Other'

4.2.Performance Evaluation of the models built

Machine learning algorithms discussed in the methodology section were applied on the dataset under two scenarios and compared the performance of each algorithm after tuning the hyperparameters. As the first scenario, machine learning algorithms were applied on the whole dataset without the application of feature selection mechanism. As the next scenario, feature selection was applied on the dataset with respect to SelectKBest and RFECV methods and performance of each machine learning model were evaluated. Thereafter, performance of machine learning models before and after application of dimensionality reduction were compared and selected the best performing model. As discussed in the previous section, five machine learning algorithms, namely; Random Forests (RF), Support Vector Regression (SVR), Extra Trees Regression (ETR), Extreme Gradient Boost (XGBoost) and Multiple linear regression (MLR) were utilized for this study. Performances of machine learning algorithms in each of these scenarios were assessed in terms of Mean Squared Error (MSE), Root Mean

Squared Error (RMSE), Mean Absolute Performance Error (MAPE) and Coefficient of determination (R2). Results obtained in each of these iterations are discussed below.

4.2.1.Performance Evaluation before feature selection (Scenario 1)

As the first scenario, each machine learning algorithm was applied on the whole dataset and compared the performance of each algorithm compared to the benchmark model. MLR algorithm is used as the benchmark model in this study to compare the performance of other tree based, non-linear and ensemble-based machine learning models. Results obtained after the application of MLR algorithm and other non-linear, tree based and ensemble-based machine learning models are given in table 4.1

Table 4.1 Performance Evaluation before feature elimination

Evaluation Metric	**Multiple Linear Regression (MLR)**	**Random Forests (RF)**	**Support Vector Regression (SVR)**	**Extra Trees Regression (ETR)**	**XG Boost**
MSE	0.122954822	**0.108154277**	0.126502467	0.137365304	0.112734342
RMSE	0.350649143	**0.328868176**	0.355671853	0.370628256	0.335759351
MAPE	23.92%	**18.08%**	23.51%	19.75%	18.88%
R2	87.76%	**89.23%**	87.41%	86.33%	88.78%

As shown in table 4.1, RF and XG Boost algorithms outperformed the MLR algorithm (Base case) in terms of MSE, RMSE and R2 evaluation metrices but the performance of SVR and ETR algorithms underperformed compared to MLR algorithm in terms of the mentioned evaluation metrices. These results show that RF and XG Boost algorithms have captured the underlying relationship among predictor and response variables relatively better compared to MLR, SVR and ETR models. To further evaluate these findings, feature selection has been applied on the dataset under each of these algorithms and compared the performances as the scenario 2.

4.2.2.Performance Evaluation after feature selection (Scenario 2)

Feature selection was applied on the dataset under two methods, namely; Recursive Feature Elimination with cross validation and SelectKBest methods. Recursive feature elimination (RFE) was applied on each of the above-mentioned machine algorithms and identified the most important feature subsets to each algorithm out of 42 features available. Summary of features removed or kept in each algorithm based on RFE and the number of features selected for each algorithm are mentioned in table 4.2

Table 4.2 Features selected for each algorithm under Recursive Feature Elimination (RFE)

	Multiple Linear Regression (MLR)	**Random Forests (RF)**	**Support Vector Regression (SVR)**	**Extra Trees Regression (ETR)**	**XG Boost**
No of Features selected	36	30	34	6	37
Features Removed/Kept	**Feature Removed:** Land_size(Perches) count_govtschools_A min_dist_nearest_Fuel_station Land_type_Agricultural Land_type_Agricultural, Commercial, Other min_dist_govtschools_b	**Features Removed:** count_govtschools_A count_intlschools count_uni count_Pvt_Hospital count_Fuel_Stations_within2km Land_type_Agricultural Land_type_Agricultural, Commercial, Other Land_type_Agricultural, Commercial, Residential, Other Land_type_Agricultural, Residential Land_type_Agricultural, Residential, Other Land_type_Commercial, Other Land_type_Other	**Feature Removed:** 'Land_size(Perches)' 'count_govtschools_A' 'count_govtschools_B' 'count_uni' 'Land_type_Agricultural' 'Land_type_Agricultural, Commercial, Other' 'Land_type_Agricultural, Residential' min_dist_govtschools_a	**Features Kept:** Land_size(Perches) 'Distance from fort' 'count_banks_within_2km' min_dist_govtschools_a min_dist_intlschools count_Supermarkets_within2km	**Features Removed:** count_Govt_Hospitals Land_type_Agricultural Land_type_Agricultural, Commercial, Other 'Land_type_Agricultural, Residential' count_Pvt_Med_Centers'

As shown in the table 4.2, dummy variables “Land Type-Agricultural” and “Land Type-Agricultural, Commercial and other” are identified under each modeling approach as the non-important features and were not considered for the respective optimum feature subset. RFE with ETR removed 36 features out of the 42 available features as non-important features. In general, most of the machine learning models identified dummy variables as non- important variables and removed them from subsequent considerations.

After removal of the non-important features through RFECV, dimensionality reduced training dataset was again fit to the machine learning models and performances were evaluated. Results obtained for the dimensionality reduced versions of each model under RFECV are mentioned in table 4.3

Table 4.3 Performance Evaluation after feature elimination through RFECV

Evaluation Metric	**Multiple Linear Regression (MLR)**	**Random Forests (RF)**	**Support Vector Regression (SVR)**	**Extra Trees Regression (ETR)**	**XG Boost**
MSE	0.122912192	**0.105665927**	0.126134956	0.124987853	0.112229973
RMSE	0.350588351	**0.325062958**	0.355154834	0.353536212	0.335007423
MAPE	23.92%	**18.04%**	23.47%	18.79%	18.54%
R2	87.76%	**89.48%**	87.44%	87.56%	88.83%

Minor performance improvement was observed in each machine learning algorithm after the application of RFECV as shown in table 4.3. Similar to the performances obtained under scenario 1, RF and XG Boost models outperformed the MLR model but SVR and ETR models underperformed compared to MLR model in terms of MSE, RMSE and R2. However, in terms of MAPE, all the models outperformed the benchmark MLR model. Feature reduced RF model showed superior performance among all the utilized models in terms of all the evaluation metrices.

To compare the performance of dimensionality reduced models obtained after RFECV, feature selection was done again on the complete dataset by utilizing a filter method; SelectKBest. Under this method, Machine learning models were fit for the range features from 1 to 42 and the iteration which showed the best performance for each machine learning model was selected. Number of features obtained for the best performing model under each algorithm is given in table 4.4.

Table 4.4 Number of features selected for each machine learning model under SelectKBest

	Multiple Linear Regression (MLR)	**Random Forests (RF)**	**Support Vector Regression (SVR)**	**Extra Trees Regression (ETR)**	**XG Boost**
No of Features selected	34	32	42	13	38

Machine learning algorithms were applied on the dimensionality reduced dataset obtained after the application of SelectKBest method and noted the performances. Summary of the performance obtained for each machine learning model under SelectKBest method is shown in table 4.5

Table 4.5 Performance Evaluation after feature elimination through SelectKBest

Evaluation Metric	**Multiple Linear Regression (MLR)**	**Random Forests (RF)**	**Support Vector Regression (SVR)**	**Extra Trees Regression (ETR)**	**XG Boost**
MSE	0.122957	**0.105813**	0.126502	0.126093	0.118753
RMSE	0.350652	**0.325289**	0.355672	0.355096	0.344606
MAPE	23.96%	**18.09%**	23.51%	22.33%	18.68%
R2	87.72%	**89.21%**	87.41%	87.47%	88.63%

As shown in table 4.5, model performance obtained through the application of SelectKBest method is similar in nature to the model performance obtained for the ML algorithms after feature reduction through RFECV. RF model outperformed all the other ML models. Both RF and XG Boost models outperformed the benchmark MLR model. However, SVR and ETR models underperformed relative to the benchmark MLR model.

As the final step of the feature selection, results of two feature selection methods were compared and identified the feature selection method which caused the highest performance improvement in each model. Summary of the comparison of RFECV feature reduced models and SelectKBest feature reduced models is mentioned in table 4.6

Table 4.6 Comparison of the results of RFECV and SelectKBest Feature selection methods

	MLR		**RF**		**SVR**		**ET**		**XGB**	
	RFECV	**SelectK Best**	**RFECV**	**SelectK Best**	**RFECV**	**SelectK Best**	**RFECV**	**SelectK Best**	**RFECV**	**SelectK Best**
MSE	0.1229	0.1230	0.1057	**0.1058**	0.1261	0.1265	0.1250	0.1261	0.1122	0.1188
RMSE	0.3506	0.3507	0.3251	**0.3253**	0.3552	0.3557	0.3535	0.3551	0.3350	0.3446
MAPE	23.92%	23.96%	18.04%	**18.09%**	23.47%	23.51%	18.79%	22.33%	18.54%	18.68%
R2	87.76%	87.72%	89.48%	**89.21%**	87.44%	87.41%	87.56%	87.47%	88.83%	88.63%

As shown in table 4.6, comparatively better performances in each machine learning algorithm is observed when these were applied on the dimensionality reduced models obtained through RFECV compared to that obtained through K best feature selection method. These results are in line with results shown in previous studies (Suto et al., 2016) which is due to the prominence given by wrapper methods to the underlying machine learning algorithm and ability of the wrapper methods to handle the feature dependencies.

4.2.3.Hyperparameter Optimization

Machine learning algorithms have unique model configurations called model hyper parameters. Results obtained so far in this study were based on the default model hyper parameters. Therefore, in order to further enhance the performance of these machine learning algorithms hyperparameter optimization was conducted to find the best set of hyperparameters for each of these algorithms and to compare the subsequent performance with the above results. Hyperparameters available in each algorithm, hyperparameters tuned in this study and the related performances obtained are discussed in the below sub sections.

Linear regression

Linear regression module of scikit-learn library contains 4 hyperparameters that can be optimized further to yield improved performance for the model. Out of these 4 hyperparameters "normalize' and "n-jobs" hyperparameters were considered in this study for further optimization by fitting various parameter combinations through RandomSearchCV. Hyperparameter "normalize" can normalize the regressors before regression by subtracting the mean and dividing by l2-norm. Hyperparameter "n-jobs" can decide the processing power to be used for the given computation. However, RandomSearchCV too provided equivalent model performances to that obtained through default configurations of the model. Hence, we couldn't achieve any improvement to the performance of the linear regression model that we obtained after the feature selection method.

Random Forests

Random forest module of the scikit-learn library contains 18 hyperparameters that can be optimized further to yield improved performance for the model. Out of these 18 parameters, “n_estimators”, “criterion”, “max_depth”,” max_features”, “bootstrap”, “min_samples_split” and “min_samples_leaf” hyperparameters were considered in this study for further optimization by fitting various parameter combinations through RandomSearchCV. Hyperparameter “n_estimators” defines the number of decision trees being built in the forest. This hyperparameter is mostly correlated to the size of the data and more decision trees are needed to encapsulate the trends in data. “criterion” is the function that is used to measure the quality of splits in a decision tree. In case of regression problems, mean absolute error and mean squared error can be used to measure the quality of splits. “max_depth” defines the maximum levels allowed in a decision tree and if this is set to “None”, the decision tree will keep on splitting until purity is reached. “max_features” defines the maximum number of features used for a node split process. “bootstrap” defines whether to use bootstrap samples or the whole dataset when building every decision tree. “min_samples_split” defines the minimum number of samples required to split an internal node. This hyperparameter decides the further subdivision of each internal node based on the given threshold value. “min_samples_leaf” sets the minimum number of data point requirements in a node of the decision tree. It affects the terminal node and basically helps in controlling the depth of the tree. Table 4.7 shows the range of hyperparameter values tested with Random search method and the optimum hyperparameter values obtained after cross validation.

Table 4.7 Hyperparameter values tested for RF algorithm and the optimum set of hyperparameter values obtained

Hyperparameter	Default value	Range of parameter values tested	Parameter value which provided the optimum result
n_estimators	100	75,100,150	**150**
criterion	"mse"	'mse','mae'	**"mae"**
max_depth	None	None,2,4,6	**2**
max_features	"auto"	auto, "sqrt", "log2"	**"log2"**
bootstrap	TRUE	TRUE, FALSE	**TRUE**
min_samples_split	2	2,4,6	**2**
min_samples_leaf	1	1,2,3	**2**

With the application of above optimized hyperparameter values, significant improvement was observed in the performance of the dimensionality reduced RF model obtained in the previous stage. Comparison of the model performance before and after the application of hyperparameter tuning is listed in table 4.8

Table 4.8 Performance of the RF model before and after hyperparameter optimization

Evaluation Metrices	**Performance of the Model before hyperparameter optimization**	**Performance of the Model after hyperparameter optimization**
MSE	0.105665927	**0.098065**
RMSE	0.325062958	**0.313154**
MAPE	18.04%	**17.88%**
R2	89.48%	**90.24%**

Support Vector Regression

Support vector regression module of the scikit-learn library contains 11 hyperparameters that can be optimized further to improve the model performance. Out of the 11 hyperparameters, “C”, “kernal” and “gamma” hyperparameters were considered in this study for further optimization by fitting various hyperparameter combinations through RandomSearchCV. “C” is the regularization parameter which represents how much misclassification is allowed in the model. By changing the regularization parameter, error in classifying data can be increased or decreased by changing the width of the margin. “kernal” is the function used by the algorithm to transform the one-dimensional data points into higher dimensions to make it linearly separable. “gamma” decides the influence made by data points which are located at a certain distance from the hyperplane. Points near to the hyperplane will have a higher impact if the gamma is high and vice versa. Table 4.9 shows the range of hyperparameter values tested with Random search method and the optimum set of hyperparameter values obtained after cross validation.

Table 4.9 Hyperparameter values tested for SVR algorithm and the optimum set of hyperparameter values obtained

Hyperparameter	**Default value**	**Range of parameter values tested**	**Parameter value which provided the optimum result**
C	1	1.0,10.0,20.0,50.0,100.0	**50**
gamma	'scale'	'scale','auto'	**scale**
kernel	'linear'	'linear', 'poly', 'rbf', 'sigmoid'	**rbf**

With the application of above optimized hyperparameter values, significant improvement was observed in the performance of the dimensionality reduced SVR model obtained in the previous stage. Comparison of the model performance before and after the application of hyperparameter tuning are listed in table 4.10

Table 4.10 Performance of the SVR model before and after hyperparameter optimization

Evaluation Metrices	**Performance of the Model before hyperparameter optimization**	**Performance of the Model after hyperparameter optimization**
MSE	0.126135	**0.107927**
RMSE	0.355155	**0.328523**
MAPE	23.47%	**20.78%**
R2	87.44%	**89.26%**

Extra trees regression

The Extra trees regression module of the scikit-learn library contains 18 hyperparameters that can be optimized further to improve the model performance. Out of the 18 parameters, “n_estimators”, “criterion”, “min_samples_split”, “min_samples_leaf” and “max_features” hyperparameters were considered in this study for further optimization by fitting various parameter combinations through RandomSearchCV. “n_estimators” is the number of decision trees used in the model. Number of decision trees used in the model can be increased until the model performance stabilizes. “criterion” is the function to measure the quality of a split. “min_samples_split” defines the number of samples required to split an internal node. “min_samples_leaf” is the minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least “min_samples_leaf” training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression. “max_features” defines the maximum features to consider when looking for the best split. Table 4.11 shows the range of hyperparameter values tested with Random search method and the optimum set of hyperparameter values obtained after cross validation.

Table 4.11 Hyperparameter values tested for ETR algorithm and the optimum set of hyperparameter values obtained

Hyperparameter	**Default value**	**Range of parameter values tested**	**Parameter value which provided the optimum result**
n_estimators	100	50,75,100,150,200,300	**75**
min_samples_split	2	2,3,4,5,8,10	**3**
min_samples_leaf	1	1,2,3,4,6,8	**2**
max_features	'auto'	'auto','sqrt','log2'	**'sqrt'**
criterion	'mse'	'mse','mae'	**'mse'**

With the application of above optimized hyperparameter values, significant improvement was observed in the performance of the dimensionality reduced ETR model obtained in the previous stage. Comparison of the model performance before and after the application of hyperparameter tuning are listed in table 4.12

Table 4.12 Performance of the Extra Trees Regression model before and after hyperparameter optimization

Evaluation Metrices	**Performance of the Model before hyperparameter optimization**	**Performance of the Model after hyperparameter optimization**
MSE	0.124988	**0.101804**
RMSE	0.353536	**0.319067**
MAPE	18.79%	**18.79%**
R2	87.56%	**89.87%**

Extreme Gradient Boosting (XGBoost)

The Extreme Gradient Boosting (XGBoost) module of the scikit-learn library contains 37 hyperparameters that can be optimized further to improve the performance of the model. These 37 hyperparameters are divided into 4 main categories, namely; general parameters, booster parameters, learning task parameters and command line parameters. General parameters guide the overall functioning of the XGBoost model. Booster parameters decide the functioning of the selected booster. Learning task parameters defines the optimization objective metric to be calculated at each step. Command line parameters are a set of parameters used only in the console version of XGBoost. Because of the time limitations and related computational expense, a selected set of hyperparameters from each of the four sub categories were considered in this study for further optimization. Explanations on the hyperparameters that were considered for optimization in this study are given below.

'booster' hyperparameter helps to choose which booster to be utilized and run in each iteration of the model. It has 3 options, namely; gbtree, dart and gblinear. Gbtree and dart use tree-based models while gblinear uses linear models. 'max_depth' defines the maximum depth of the decision tree and this controls the overfitting as higher depth will allow the model to learn relations very specific to a particular sample. 'min_child_weight' defines the minimum sum of weights of all observations required in a child which is used to control over fitting. Higher values prevent the model from learning relations which might be highly specific to the particular sample selected for a tree and too high values could lead to under-fitting. Therefore, cross validation is necessary for this hyperparameter to select the optimum value. 'tree_method' defines the tree construction algorithm used in XGBoost. Values for this parameter should be selected based on the size of the given dataset. 'scale_pos_weight' controls the balance of positive and negative weights and is useful for imbalanced classes. 'objective' defines the loss function to be minimized in the model. 'eval_metric' is the metric to be used for validation data. 'n_estimators' defines the number of gradient boosted trees to be utilized in the model. Table 4.13 shows the range of hyperparameter values tested with Random search method and the optimum set of hyperparameter values obtained after cross validation.

Table 4.13 Hyperparameter values tested and the optimum set of hyperparameter values obtained

Hyperparameter	**Default value**	**Range of parameter values tested**	**Parameter value which provided the optimum result**
tree_method	'exact'	'auto', 'exact', 'approx', 'hist', 'gpu_hist'	**'auto'**
scale_pos_weight	1	1,2,3,5,10	**1**
'objective'	'reg:squarederror'	'reg:squarederror','reg:squaredlogerror' ,'reg:logistic'	**'reg:squarederror'**
n_estimators	100	50,75,100,150,200	**75**
min_child_weight	1	1,2,3,4,5,10	**10**
max_depth	6	4,5,6,8,10,12,20	**4**
eval_metric	None	'rmse','mae','logloss'	**'logloss'**
booster	'gbtree'	'gbtree', 'gblinear','dart'	**'dart'**

With the application of above optimized hyperparameter values, significant improvement was observed in the performance of the dimensionality reduced XGBoost model obtained in the previous stage. Comparison of the model performance before and after the application of hyperparameter tuning is listed in table 4.14

Table 4.14 Performance of the XGBoost model before and after hyperparameter optimization

Evaluation Metrices	**Performance of the Model before hyperparameter optimization**	**Performance of the Model after hyperparameter optimization**
MSE	0.11223	**0.099244**
RMSE	0.335007	**0.31503**
MAPE	18.54%	**19.09%**
R2	88.83%	**90.12%**

Table 4.15 Performance Evaluation after hyperparameter optimization

Evaluation Metric	**Multiple Linear Regression (MLR)**	**Random Forests (RF)**	**Support Vector Regression (SVR)**	**Extra Trees Regression (ETR)**	**XG Boost**
MSE	0.122912192	**0.098065**	0.107927	0.101804	0.099244
RMSE	0.350588351	**0.313154**	0.328523	0.319067	0.31503
MAPE	23.92%	**17.88%**	20.78%	18.79%	19.09%
R2	87.76%	**90.24%**	89.26%	89.87%	90.12%

Table 4.15 summarizes the performance of the machine learning models after hyperparameter optimization. All the machine learning models outperformed the benchmark model (MLR) in terms of all the evaluation metrices considered in this study. MLR and SVR models performed poorly compared to the RF, ETR and XGBoost models. This could be due to the over dependency of these algorithms on the linear assumption as shown by Zhang (Zhang et al., 2021). In contrast, two tree-based models; RF and ETR, and ensemble model; XG Boost performed well with respect to all the evaluation metrices. The above results highlight the better performance of non-linear and ensemble-based machine learning algorithms in modelling complex real-world problems. Among these models , RF is the most robust method in terms of all performance indicators in agreement with the conclusion of Derdouri and Murayamas' study (Derdouri and Murayama, 2020), which is similar in nature and context to current study. RF showed the lowest MAPE of 17.88% and its MAE and RMSE are 0.098065 and 0.313154 respectively. Further, it showed a superior R^2 value of 90.24% compared to the rest of the algorithms considered in this study. Features on which the final RF model was trained are mentioned in table 4.16

Table 4.16 Features of the best performed RF model

	Feature Name		Feature Name
1	Land size (Perches)	16	Distance to the nearest Govt Hospital
2	Distance from fort	17	Number of Govt hospitals within 5km radius
3	Distance to the nearest Govt -type A School	18	Distance to the nearest Pvt Hospital
4	Number of Govt-type B schools in 2km radius	19	Distance to the nearest Pvt Medical Center
5	Distance to the nearest Govt -type B School	20	Number of Private Medical Centers within 2km radius
6	Number of Semi-Govt Schools in 2km radius	21	Distance to the nearest Supermarket
7	Distance to the nearest Semi-Govt School	22	Number of Supermarkets within 2km radius
8	Distance to the nearest International School	23	Distance to the nearest fuel station
9	Distance to the nearest University	24	Type of the land-Agricultural, Commercial
10	Distance to the nearest Expressway entrance	25	Type of the land-Agricultural, Commercial, Residential
11	Distance to the nearest Railway station	26	Type of the land- Commercial
12	Distance to the nearest Bank	27	Type of the land- Commercial, Residential
13	Number of Banks within 2km radius	28	Type of the land- Commercial, Residential, other
14	Distance to the nearest Finance company	29	Type of the land- Residential
15	Number of Finance companies within 2km radius	30	Type of the land- Residential, other

Moreover, it can be found that ensemble machine learning algorithms like Extreme Gradient Boosting (XGBoost) scored better results than linear and nonlinear methods like Multiple Linear Regression and Support Vector Regression. Even though its performances are slightly below the performances of RF, it showed MSE, RMSE, MAPE and R^2 values of 0.099244, 0.31503, 19.09% and 90.12% respectively which are highly satisfactory.

4.3. Chapter Summary

This chapter focused on obtaining a machine learning model to predict the land prices in Colombo district by fitting 5 machine learning models to the developed dataset. Evaluation of the models has been done under two scenarios; with and without the application of the feature selection mechanisms. Two feature selection methods were applied on the complete dataset to identify the method which provided a set of features that optimize the model performances. Wrapper method utilized in this study for feature selection; RandomizedSearch, provided the best set of features and the subsequent hyperparameter tuning showed Random Forest as the best machine learning model to predict the land prices in Colombo district with R^2 value of 90.24 % by outperforming Multiple Linear regression, Support Vector Regression, Extra Trees Regression and Extreme Gradient Bosting models. Best performing model consists of 30 features out of the 42 predictor variables available in the original dataset. Next chapter will provide a general discussion about the study with conclusions, limitations and suggestions for improvements.

5. Conclusion

This chapter summarizes the results obtained through the analysis in the previous chapter and highlights the important findings and the conclusions derived as a result. This chapter also discusses the limitations of the study and concludes by proposing further areas of research.

Land valuation in Sri Lanka is currently done based on the experience and judgement of the individual valuation officers which is highly subjective and questionable as the way of analyzing the features and providing a value could vary from person to person. This problem is highly impactful to the districts like Colombo, where the lands are relatively high priced compared to other districts. In this study, land prices of Colombo district are attempted to be estimated through machine learning models by analyzing the web scraped data.

Machine learning models built in this study are evaluated based on two scenarios. In the first scenario, dimensionality reduction is not applied on the dataset and model evaluations are done based on the complete dataset with 42 features and 3725 observations. In the second scenario, dimensionality reduction is applied on the dataset under two feature selection mechanisms, namely; Recursive Feature Elimination (RFE) and K best. Model evaluation results obtained in the second scenario showed that feature reduction through RFE could yield superior performance in all the considered models compared to other scenarios. Out of the 42 features considered, 30 features are selected through RFE as the most important features for the RF model. These features showed that size of the land, distance from Fort, distances to the nearest government, semi government and international schools, number of government and semi-government schools nearby, distance to the nearest university, distance to the nearest expressway entrance, distance to the nearest railway station, distance to the nearest bank and financial institution, number of banks and finance companies nearby, distance to the nearest government and private hospitals, number of government hospitals nearby, distance to the nearest private medical center, distance to the nearest supermarket, number of supermarkets nearby, distance to the nearest fuel station and type of land as the factors which most influenced the land prices in Colombo district. Five machine learning models, namely Multiple Linear Regression (MLR), Random Forests (RF), Support Vector Regression (SVR), Extra Trees Regression (ETR) and Extreme Gradient Boosting (XGBoost) are fitted on the dimensionally reduced dataset and subsequently optimized the hyperparameters of each model. Prior to hyperparameter optimization, the RF model showed superior performance compared to the rest of models with MAPE, MAE, RMSE and R^2 values of 18.04%, 0.1057, 0.3251 and 89.48%

respectively. After hyperparameter optimization, performance of the model further improved with MAPE, MAE, RMSE and R^2 values of 17.88%, 0.098065, 0.313154 and 90.24% respectively. This model with 90.24% accuracy means, estimation of land prices in Colombo district by hyper parameter tuned random forest model is a success and gives reliable results.

Thus, this thesis provides a framework to understand the factors which primarily affect the land prices in Colombo district. More importantly, this dissertation proposes an application of using these factors through a machine learning model to bring out meaningful estimates that the decision makers can use. Models of this nature would facilitate the general public to identify a reasonable price for their potential land related transactions as well as would provide a data driven platform to the land valuers in Sri Lanka to benchmark and compare their professional estimations.

5.1.Limitations

The main limitation of this study was the non-availability of data with respect to land prices and land price determinants in a structured and organized form. This non availability made the dataset preparation task of this study extremely difficult. When calculating the distances to nearest places and number of specific places within a given radius, location coordinates of the GN division are utilized in this study due to limited availability of data in the web advertisement to get the exact location coordinates of the advertised land.

This study is conducted based on the data on land prices and land price determinants in Colombo district due to time and resource limitations. Further, data advertised in ikman.lk for a period of only two months was considered to construct the dataset due to time limitations. Further, due to time limitation, only a selected set of hyperparameters of the machine learning algorithms were optimized.

5.2.Future Work

Meaningfulness of the model could be further improved if the factors which determine the quality of land location like crime rate in the GN division, propensity to seasonal flooding, literacy level and average income level of the people who are living in the area, proportion of professionals living in the vicinity etc. could be incorporated to the model in the future research work. Accuracy and reliability of the results could be further improved if the exact location coordinates of the land plot are utilized to calculate the nearest distances and number of specific places within a given radius.

Applicability of this model could be further improved if the scope of the project further extended to provincial or island wide level. This would provide the model with more samples to train hence the reliability of results could be further enhanced. Hyperparameter tuning of this study helped to improve the model performances in significant terms. Therefore, hyperparameter optimization by utilization of all the available hyperparameters of the RF algorithm could further improve the performance of the model.

6. References

Ahmad, M.W., Reynolds, J., Rezgui, Y., 2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. Journal of Cleaner Production 203, 810–821. https://doi.org/10.1016/j.jclepro.2018.08.207

Alshboul, O., Shehadeh, A., Almasabha, G., Almuflih, A.S., 2022. Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction. Sustainability 14, 6651. https://doi.org/10.3390/su14116651

Ariyawansa, R., 2016. Review of Price Formation of Land and Landed Properties in Colombo: Is it a Myth or Reality?

Authorized Financial Institutions | Central Bank of Sri Lanka [WWW Document], n.d. URL https://www.cbsl.gov.lk/en/authorized-financial-institutions (accessed 4.30.22).

Bazan-Krzywoszanska, A., Bereta, M., 2018. The use of urban indicators in forecasting a real estate value with the use of deep neural network. Reports on Geodesy and Geoinformatics 106, 25–34. https://doi.org/10.2478/rgg-2018-0011

Botchkarev, A., 2019. A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. IJIKM 14, 045–076. https://doi.org/10.28945/4184

Chowdhury, A.Z., n.d. Exploring the Impact of Covid-19 on Taxi Demand in New York City Using Machine Learning 70.

Colombo Zonal Education Office [WWW Document], n.d. URL http://www.cmbzone.sch.lk/english/schoolcs.php (accessed 4.30.22).

Córdoba, M., Carranza, J.P., Piumetto, M., Monzani, F., Balzarini, M., 2021. A spatially based quantile regression forest model for mapping rural land values. Journal of Environmental Management 289, 112509. https://doi.org/10.1016/j.jenvman.2021.112509

Demetriou, D., 2017. A spatially based artificial neural network mass valuation model for land consolidation. Environment and Planning B: Urban Analytics and City Science 44, 864–883. https://doi.org/10.1177/0265813516652115

Derdouri, A., Murayama, Y., 2020. A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan. J. Geogr. Sci. 30, 794–822. https://doi.org/10.1007/s11442-020-1756-1

Ebrahimi-Khusfi, Z., Nafarzadegan, A.R., Dargahian, F., 2021. Predicting the number of dusty days around the desert wetlands in southeastern Iran using feature selection and machine learning techniques. Ecological Indicators 125, 107499. https://doi.org/10.1016/j.ecolind.2021.107499

EOM&M Division [WWW Document], n.d. URL http://www.exway.rda.gov.lk/index.php?page=expressway_network/exps (accessed 4.30.22).

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., Tu, X.M., 2014. Log-transformation and its implications for data analysis 26, 6.

Grama Niladhari Division [WWW Document], n.d. URL http://www.colombo.dist.gov.lk/index.php/en/administrative-structure/grama-niladhari-division.html (accessed 4.19.22).

Hu, S., Yang, S., Li, W., Zhang, C., Xu, F., 2016. Spatially non-stationary relationships between urban residential land price and impact factors in Wuhan city, China. Applied Geography 68, 48–56. https://doi.org/10.1016/j.apgeog.2016.01.006

Kalliola, J., Kapočiūtė-Dzikienė, J., Damaševičius, R., 2021. Neural network hyperparameter optimization for prediction of real estate prices in Helsinki. PeerJ Computer Science 7, e444. https://doi.org/10.7717/peerj-cs.444

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling 81, 401–418. https://doi.org/10.1016/j.apm.2019.12.016

Kim, B., Kim, T., 2016. A Study on Estimation of Land Value Using Spatial Statistics: Focusing on Real Transaction Land Prices in Korea. Sustainability 8, 203. https://doi.org/10.3390/su8030203

Kim, Y., Choi, S., Yi, M.Y., 2020. Applying Comparable Sales Method to the Automated Estimation of Real Estate Prices. Sustainability 12, 5679. https://doi.org/10.3390/su12145679

Land Valuation Indicator - First Half of 2021 | Central Bank of Sri Lanka [WWW Document], n.d. URL https://www.cbsl.gov.lk/en/news/land-valuation-indicator-first-half-2021 (accessed 4.20.22).

Lanka Property Web - Find the average Sri Lanka House & Land Prices | Sri Lanka House Price Index [WWW Document], n.d. . LankaPropertyWeb.com. URL https://www.lankapropertyweb.com/house_prices.php (accessed 3.22.22).

Levantesi, S., Piscopo, G., 2020. The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach. Risks 8, 112. https://doi.org/10.3390/risks8040112

Li, L., Prussella, P.G.R.N.I., Gunathilake, M.D.E.K., Munasinghe, D.S., Karadana, C.A., 2015. Land Valuation Systems using GIS Technology Case of Matara Urban Council Area, Sri Lanka. Bhumi, Planning Res. J. 4, 7. https://doi.org/10.4038/bhumi.v4i2.6

Licensed Finance Companies | Central Bank of Sri Lanka [WWW Document], n.d. URL https://www.cbsl.gov.lk/authorized-financial-institutions/licensed-finance-companies (accessed 4.30.22).

Liu, Y., Fan, P., Yue, W., Song, Y., 2018. Impacts of land finance on urban sprawl in China: The case of Chongqing. Land Use Policy 72, 420–432. https://doi.org/10.1016/j.landusepol.2018.01.004

Ma, J., Cheng, J.C.P., Jiang, F., Chen, W., Zhang, J., 2020. Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. Land Use Policy 94, 104537. https://doi.org/10.1016/j.landusepol.2020.104537

Making Predictive Models Robust: Holdout vs Cross-Validation, n.d. . KDnuggets. URL https://www.kdnuggets.com/making-predictive-model-robust-holdout-vs-cross-validation.html/ (accessed 5.1.22).

Measures of Consumer Price Inflation | Central Bank of Sri Lanka [WWW Document], n.d. URL https://www.cbsl.gov.lk/en/measures-of-consumer-price-inflation (accessed 4.20.22).

Mendonça, R., Roebeling, P., Martins, F., Fidélis, T., Teotónio, C., Alves, H., Rocha, J., 2020. Assessing economic instruments to steer urban residential sprawl, using a hedonic pricing simulation modelling approach. Land Use Policy 92, 104458. https://doi.org/10.1016/j.landusepol.2019.104458

Ministry Of Health - HOSPITALS [WWW Document], n.d. URL http://www.health.gov.lk/moh_final/english/hospital_government.php (accessed 4.30.22).

Mirkatouli, J., Samadi, R., Hosseini, A., 2018. Evaluating and analysis of socio-economic variables on land and housing prices in Mashhad, Iran. Sustainable Cities and Society 41, 695–705. https://doi.org/10.1016/j.scs.2018.06.022

Mostafa, M.M., 2018. A spatial econometric analysis of residential land prices in Kuwait. Regional Studies, Regional Science 5, 290–311. https://doi.org/10.1080/21681376.2018.1518154

Nakamura, H., 2019. Relationship among land price, entrepreneurship, the environment, economics, and social factors in the value assessment of Japanese cities. Journal of Cleaner Production 217, 144–152. https://doi.org/10.1016/j.jclepro.2019.01.201

Pai, P.-F., Wang, W.-C., 2020. Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. Applied Sciences 10, 5832. https://doi.org/10.3390/app10175832

Parbat, D., Chakraborty, M., 2020. A python based support vector regression model for prediction of COVID19 cases in India. Chaos, Solitons & Fractals 138, 109942. https://doi.org/10.1016/j.chaos.2020.109942

Park, B., Bae, J.K., 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications 42, 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

Refaeilzadeh, P., n.d. On Comparison of Feature Selection Algorithms 6.

Registered Institutes [WWW Document], n.d. URL http://www.phsrc.lk/pages_e.php?id=12 (accessed 4.30.22).

Singh, A., Sharma, A., Dubey, G., 2020. Big data analytics predicting real estate prices. Int J Syst Assur Eng Manag 11, 208–219. https://doi.org/10.1007/s13198-020-00946-3

Sri Lanka 100 2022 | Brand Value Ranking League Table | Brandirectory [WWW Document], n.d. URL https://brandirectory.com/rankings/sri-lanka/table (accessed 4.30.22).

Station Details [WWW Document], n.d. URL https://www.railway.gov.lk/web/index.php?option=com_content&view=article&id=165&Itemid=191&lang=en (accessed 4.30.22).

Statistics - Department of Examinations - Sri Lanka [WWW Document], n.d. URL https://www.doenets.lk/statistics (accessed 4.30.22).

Suto, J., Oniga, S., Sitar, P.P., 2016. Comparison of wrapper and filter feature selection algorithms on human activity recognition, in: 2016 6th International Conference on Computers Communications and Control (ICCCC). Presented at the 2016 6th International Conference on Computers Communications and Control (ICCCC), IEEE, Oradea, Romania, pp. 124–129. https://doi.org/10.1109/ICCCC.2016.7496749

TISSL - The International Schools of Sri Lanka [WWW Document], n.d. URL https://tissl.lk/members.php (accessed 4.30.22).

Universities [WWW Document], n.d. URL https://www.ugc.ac.lk/index.php?option=com_university&view=list&Itemid=25&lang=en (accessed 4.30.22).

Wang, D., Li, V.J., 2019. Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review. Sustainability 11, 7006. https://doi.org/10.3390/su11247006

Wang, D., Li, V.J., Yu, H., 2020. Mass Appraisal Modeling of Real Estate in Urban Centers by Geographically and Temporally Weighted Regression: A Case Study of Beijing's Core Area. Land 9, 143. https://doi.org/10.3390/land9050143

Wen, H., Chu, L., Zhang, J., Xiao, Y., 2018. Competitive Intensity, Developer Expectation, and Land Price: Evidence from Hangzhou, China. Journal of Urban Planning and Development 144. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000490

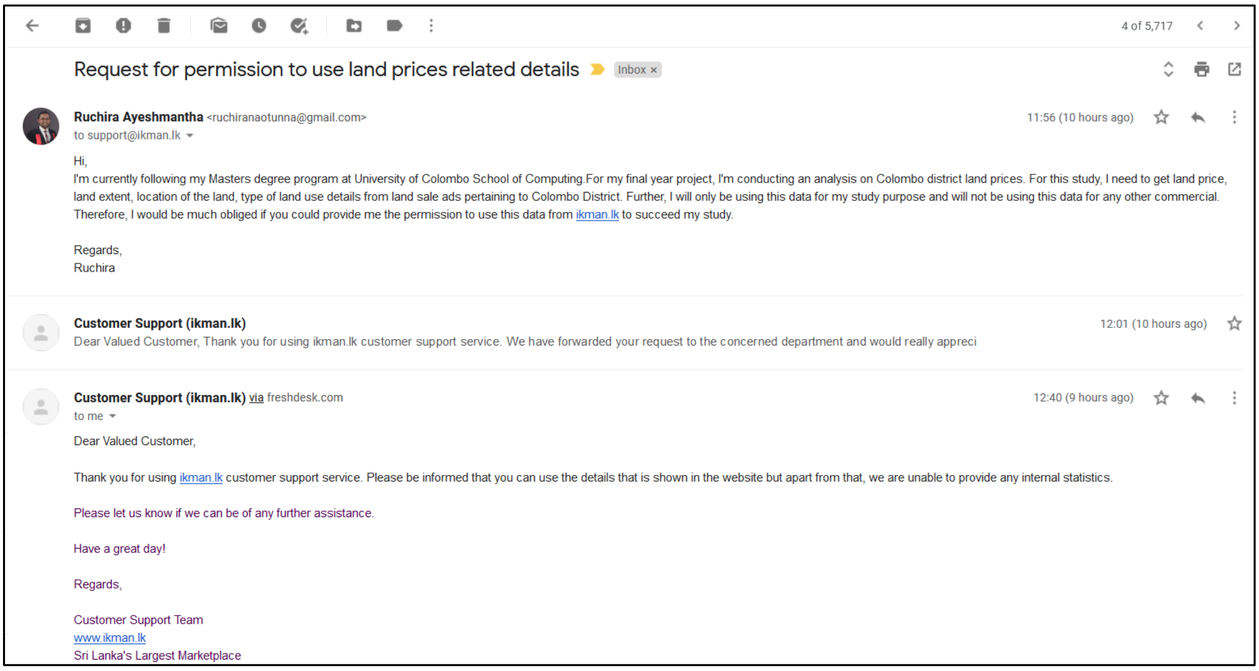
Zhang, P., Hu, S., Li, W., Zhang, C., Yang, S., Qu, S., 2021. Modeling fine-scale residential land price distribution: An experimental study using open data and machine learning. Applied Geography 129, 102442. https://doi.org/10.1016/j.apgeog.2021.102442

Zhong, H., Wang, J., Jia, H., Mu, Y., Lv, S., 2019. Vector field-based support vector regression for building energy consumption prediction. Applied Energy 242, 403–414. https://doi.org/10.1016/j.apenergy.2019.03.078

Zhou, G., Etemadi, A., Mardon, A., 2022. Machine learning-based cost predictive model for better operating expenditure estimations of U.S. light rail transit projects. Journal of Public Transportation 24, 100031. https://doi.org/10.1016/j.jpubtr.2022.100031

Attachments

Attachment I- Permission to use ikman.lk website details



Request for permission to use land prices related details

Ruchira Ayeshmantha <ruchiranaotunna@gmail.com>
to support@ikman.lk

11:56 (10 hours ago)

Hi,
I'm currently following my Masters degree program at University of Colombo School of Computing.For my final year project, I'm conducting an analysis on Colombo district land prices. For this study, I need to get land price, land extent, location of the land, type of land use details from land sale ads pertaining to Colombo District. Further, I will only be using this data for my study purpose and will not be using this data for any other commercial. Therefore, I would be much obliged if you could provide me the permission to use this data from ikman.lk to succeed my study.

Regards,
Ruchira

Customer Support (ikman.lk)
Dear Valued Customer, Thank you for using ikman.lk customer support service. We have forwarded your request to the concerned department and would really appreci

12:01 (10 hours ago)

Customer Support (ikman.lk) via freshdesk.com
to me

12:40 (9 hours ago)

Dear Valued Customer,

Thank you for using ikman.lk customer support service. Please be informed that you can use the details that is shown in the website but apart from that, we are unable to provide any internal statistics.

Please let us know if we can be of any further assistance.

Have a great day!

Regards,

Customer Support Team
www.ikman.lk
Sri Lanka's Largest Marketplace