House Value Prediction Using Machine Learning In Sri Lanka

- Semester 7/8 Report -

Anojan Satheesnathan
Sankeerthan Kasilingam

Department of Computer Engineering
University of Peradeniya

Final Year Project (courses CO421 & CO425) report submitted as a requirement of the degree of
B.Sc.Eng. in Computer Engineering

November 2019

Supervisor: Mr. Sampath Degalla (University of Peradeniya)

I would like to dedicate this thesis to my loving parents and “teachers” . . . who supported us to succeed our project.

Declaration

We hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is our own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgments.

Anojan Satheesnathan
Sankeerthan Kasilingam
November 2019

Acknowledgements

And we would like to acknowledge ... Mr. Sampath Degalla... Our supervisor, thank you for your all unwavering support, help and guidance.

Abstract

In Real estate market, houses (residential properties) took the biggest part and it has been considered as one of the largest asset classes from ancient times. Most people deal with houses directly in selling or buying. The house appraisal is usually done by experts known as valuation officers. Nowadays, in exploring unknown patterns from data, machine learning techniques become more traditional. This study aims to figure out how different machine learning models can be used to predict housing prices with good accuracy and reasonable predictive performance. In short, this study experiment in building various machine Learning models to fit house appraisal.

This project uses data set collected from ikman.lk by web scraping, which includes 11 explanatory features and 2185 data entries of housing sale advertisements in Colombo, Sri Lanka. 535 instances with 22 explanatory feature data set is created after preprocessing from collected data. Outliers detection method used in this research includes Tukey algorithm. Number of data points is reduced to 487 after outliers handling. Features have high correlation with price are identified by feature correlation matrix before training. Land size is identified as a factor that highly affecting house price with a correlation value 0.4. Regression models using Linear Regression, Support Vector Regression, Random Forest, Multi Linear Perceptron and K-Nearest Neighbour were built and results were compared using evaluation metrics R-Square and Root Mean Square Error. Random Forest performs well compared to other algorithms with R-Square score 0.632 and RMSE with 0.097.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

ANN	Artificial Neural Networks
CO	Carbon Monoxide
CV	Cross Validation
DT	Decision Tree
GDP	Gross Domestic Product
KNN	K-Nearest Neighbours
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi Linear Perceptron
MSE	Mean Squared Error
N/A	Non Available
NLP	Natural Language Processing
NO	Nitrogen Oxide
NO2	Nitrogen Dioxide
R2	R-Square
R2	R-Square

RBF	Radial Basis Function
RMSE	Root Mean Square Error
SO2	Sulfur Dioxide
SVM	Support Vector Machine
SVR	Support Vector Regression
UK	United Kindom
USA	United State of America

Chapter 1

Introduction

1.1 Background

1.1.1 Introduction

The relationship between house prices and the economy is a motivating factor for predicting house prices and there is no accurate measure of house prices [1]. A property's value is important in real estate transaction. House prices trends are not only the concerns for buyers and sellers, but they also indicate the current economic situations. Therefore, it is important to predict the house prices without bias to help both buyers and sellers make their decisions.

There are different machine learning algorithm can be used to predict the house prices. This research will use various regression algorithms and find out the suitable model to predict house prices in Colombo, Sri Lanka.

There are many factors affect house prices, such as land size, number of bed rooms and number of bath rooms. In addition choosing different combination of parameters in algorithms will also affect the prediction greatly. This project is guided by these questions: Which features are important for predicting price of a house?, Which parameters in various algorithms have better performance in predicting house price? How to find suitable regression model for house price prediction?.

1.1.2 Sri Lanka Housing Market

In the last decade, the housing market in Sri Lanka has been rapidly growing, with average housing prices increasing by 17% nationwide according to Lanka Property Web. Nearly 50% increase in Colombo district year over year. 10.3% of the GDP increased

in construction industry back in 2013 [2]. Together with stronger economic growth and increasing price expectations, this research presents good news to current home owners and potential home buyers looking for a safe long-term investment.

Fig. 1.1 Average Property Prices of Sri Lanka

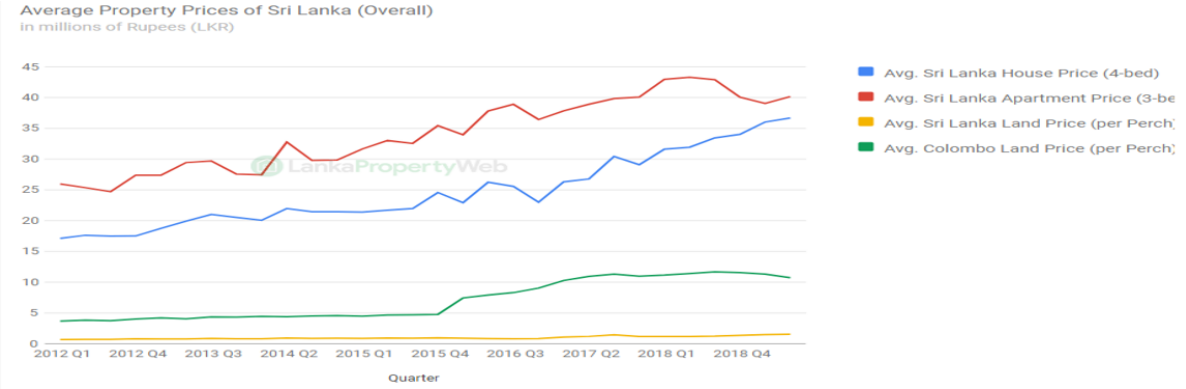


Figure 1.1 shows how 4-bed average house price, 3-bed average apartment price, average land price per perch and average Colombo land price per perch varying with the time with some fluctuations. Average house price increased up to 35 million in end of 2018.

1.2 The Problem

There is always uncertainty in the property market in Sri Lanka. There are lot of people involved in investment opportunities related to housing market. But most of the cases they use traditional data analysis techniques. As shown in figure 1.1 there was a suddenly reduction in house price in 2016 and there are some more in between start of 2012 to end of 2018. these uncertainties needs to be predicted beforehand to avoid risks related to investments.

Housing prices are currently estimated manually in Sri Lanka. This traditional approach needs a professional expert with the domain knowledge to predict the property value. In some cases appraiser can be paid by one party and he may conduct the appraisal in favour of them. There are difficulties in ensuring that the appraiser conducts a neutral appraisal in these cases.

1.3 The Proposed Solution

The proposed solution is to come up with a suitable house price prediction model that predicts all the house prices in Colombo, Sri Lanka.

Aims of this research are:

- Predict the house prices in Colombo
- Identify the factors affecting housing price

The objective of this research is to find out a suitable machine learning model by evaluating prediction accuracy and performance in order to predict the house prices in Colombo.

There are lot of researches were carried out in various countries in house price prediction using machine learning but it is new in Sri Lanka. Our plan is to build a prediction model by experimenting various algorithms and build ensemble model by combining multiple of those models in-order to get better accuracy and performance by comparing with previous researches.

The scope of this project is only limited to Colombo district since non-availability of data source for other districts in Sri Lanka. external factors such as economic affecting housing prices is not considered in this research. Data from ikman.lk is only used even though it is non-reliable because there are no other data sources available in Sri Lanka. Anomaly detecting techniques are used to remove anomalies by doing statistical analysis.

1.4 Deliverable and Milestones

These are the milestones completed in semester 7

- Related work inspection and Literature review
- Data collection
- Data pre-processing and create data set for experiments
- Initial experiment with selected algorithms and analyse the result
- Project Report for semester 7

These are the milestones to be completed in semester 8

- Increase data points and automate feature identification

- Experiment with more algorithm and build ensemble model
- Analyse feature interactions
- Find out suitable model by analysing performance and accuracy
- Final Project Report

1.5 Outline of the Report

This report gives the background information about house price prediction and why it is important in Sri Lanka in the Introduction section. It describes the details of previous work which have been done regarding house price prediction using machine learning in various countries in the Related Work section. It describes about the methodology followed and the implementation details of building the prediction model in Methodology and Experimental Setup and Implementation sections respectively. The document contains the experiment results and analysis in Results and Analysis section. The document also contains the future works required for further improvements in Conclusion and Future Works sections.

Chapter 2

Related work

2.1 Introduction

The study analyses 20 researches that used machine learning to predict house prices and aims to focus on various machine learning models used in predicting house prices along with different set of features, prediction accuracy and predictive performance in previous researches. An extensive study should required to analyse different algorithms selection and features selection and their results in-terms of accuracy and error metrics. Other than individual buyers and sellers, there are many investment firms directly linked with housing market. A report from MSCI (known as Morgan Stanley Capital Investment) states that real estate investment had increased to $8.5 trillion in 2017, while comparing it with previous year it was $1.1 trillion increment. Even Though the investment in housing/residential markets seem to be profitable ,but the prices are directly connected with global economy, GDP and political stability and demand is the key factor in determining house prices [3].

Multiple discipline people such as house owners, developers, investors, appraisers, tax assessors , mortgage lenders and insurers rely on house appraisal to make decisions [4]. The house price appraisal is usually done using hedonic models. In hedonic price theory, the house is viewed as a set of characteristics such as number of bedrooms, number of bathrooms, geolocation, house size and many more. The Coefficients of some characteristics exhibited unstable nature [4]. The main problem with hedonic models in house appraisal is the relationships between each characteristic and price should be known in advance [3].

Another alternative approach is in house appraisal is machine learning models. Various researches carried out in multiple countries including Spain, UK, New Zealand, USA, Turkey and Italy. House price prediction using machine learning performs better than

traditional models with reasonable error. By the nature of machine learning, there is no standard way to define a model to predict house price. The feature sets were considered in researches show difficulty in concluding one best model for this problem. The features sets took account of common known features and different features only belong to a particular region, such as environmental pollution and etc [5]. It is clear, the good data set with right feature selection and algorithm selection can make even better house appraisal model.

Machine learning is an internal part of artificial intelligence. In ML, computer learns automatically from data and information using different computer algorithms. Computer don't need to explicitly programmed. These can be improved and changed algorithms by themselves. There is a growing need for machine learning among companies for professionals and it is used all over the world [5].

2.2 Dataset

Out of 20 researches, 7researches used dataset collected from online realestate selling platforms [3] [5] [6] [7]. They did not use exact house price transactions instead of they assumed the price advertised in websites as real price. Only 5 researches used open Taiwan house data set, kaggle house data set, CoreLogic Dataset, and ValueguardAB dataset. The minimum number of data instances used in researches is 193 [5] and the maximum number of instances used in researche is more than 2.4 million [8]. The table 2.1 shows a detailed summary of data-sets used in various researches.

Table 2.1 Data-Sets Used in Researches

Research	**Location**	**Number of Instances**	**Source**
[3]	Salamanca district of Madrid, Spain	2226	Online website
[5]	Taranto(Italy)	193	Online Geographic Information System (GIS)
[9]	Christchurch,New Zealand	200	Online website
[10]	N/A	3000	Kaggle data set
[11]	Taiwan	74568	Open source data (data.gov.tw)
[6]	Ankara,Turkey	N/A	Online website
[1]	KingCounty, USA	21613	Kaggle data set
[12]	N/A	1500	Kaggle data set
[13]	Boston, USA	N/A	N/A
[14]	King County, Seattle	21000	N/A
[7]	Beijing, China	9600	Online website
[15]	Suburb, Boston	452	N/A
[16]	Turkey	5741	House hold budget survey data
[17]	S~aoPaulo, Brazil	N/A	N/A
[18]	Sweden	N/A	Valueguard AB's housing data set
[19]	Suburb, Boston	506	N/A
[8]	London	2.4 Million	Landon data store + Land registry (with their online searching tool)
[20]	Montr´eal, Canada	25000	online website

2.3 Conclusion

There are lots of researches related to house price prediction using machine learning. But there is no research exist for predicting housing prices in Sri Lanka. Researchers did not try multiple algorithms most of the cases instead of they built one model using particular algorithm. This is where our research focusing on getting better accuracy an performance by combining various model and building ensemble model.

Chapter 3

Methodology

3.1 Overview

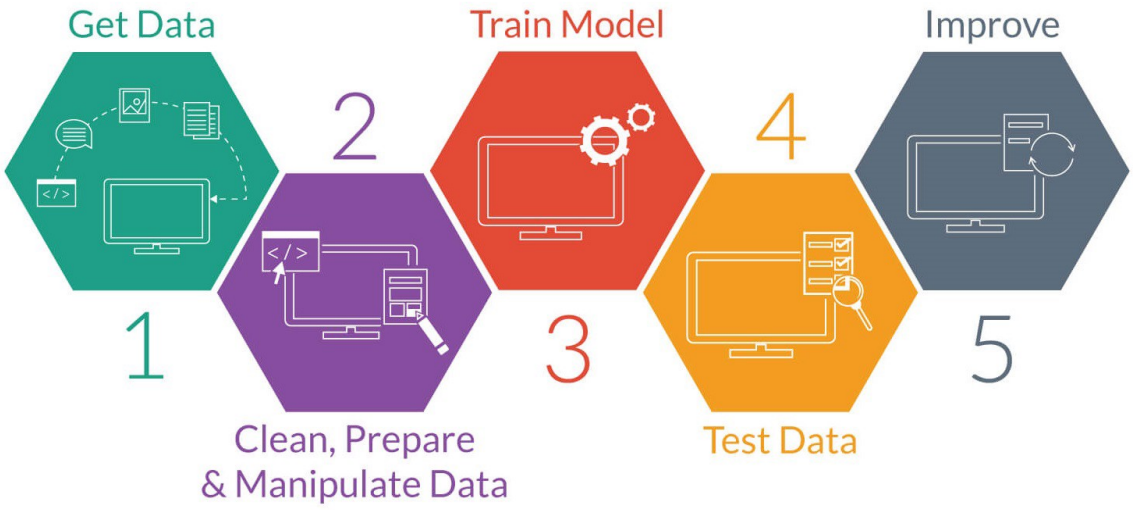
This chapter covers how we aligned multiple divisible steps into a sequence manner, selection criteria we used to select each methods (from data collection to evaluation of models), abstract idea of each steps and brief theory of algorithms .

After we reviewed literature, we found similar researches to house price prediction using machine learning were used in different ways and different methods. Especially algorithms selection to build model in literature made us confused, because some researches states justification for selected methods (data-source, algorithms, etc) and some are not. Finally we realized, the project is kind of research, that hypothesis and experimental is the way to form methodology.

3.2 Conceptual Design

Our project is machine learning project, so we followed and structured the project into multiple steps. The figure 3.1 shows the steps in sequence manner.

Fig. 3.1 Project Work Flow



3.3 Methodological Approach

The scope of project is Sri Lanka (house price prediction using machine learning in Sri Lanka), so we had to collect house data set of Sri Lanka. We observed five researches used data collected from online house selling platform (they assumed advertised price as sold price), seven researches used open source data set and others used collected data from government offices. Initially our aim was to find reliable house data set from government offices, we know having good data set is more than fine tuning models. Finally, we had to collect data from online house selling platform, because we contacted Valuation Department in Colombo, they responded that giving the historical data to outsiders (third-party) is not their standard practice and we had to take special permission. We attempted to take take special permission for two weeks, but we could not get. Then we decided to use collect data from online house selling platform.

The collected Data set contains 11 attributes and they cannot be used in building model directly and some attributes (posted address and date) are not useful. The considerable attributes were house size, land size, location, number of beds, number of baths, number of beds and price. We had to find out other attributes from description text of each house instance. To identify hidden features from description text, we analysed frequency of words and finalized features set.

Next, we decided and automated to find out hidden features from description text. We found process data using automation deviate from actual description (more explanation

on next chapter). By end of automated feature detection, the time had passed, to stick with time, we manually processed data to find out hidden features from description text.

Then outliers were identified and removed from the data set using Tukey algorithm. One of the simplest methods for detecting outliers is the use of box plots. A box plot is a graphical display for describing the distribution of the data. Box plots use the median and the lower and upper quartiles.

The Tukey's method defines an outlier as those values of the data set that fall far from the central point, the median. The maximum distance to the center of the data that is going to be allowed is called the cleaning parameter.

After processed and removed outlier, we had to build machine learning models. Before start training of model, We picked five algorithms (initially we planned to build with three). The algorithm selection was very difficult to us because, none of us have experience in working real machine learning project before.

Our first choice was Linear Regression, because if features exhibit linear relationship between price, there is no requirement of building complex models. A linear regression model attempts to explain the relationship between two or more variables using a straight line. Another algorithm we picked was SVR (Support Vector Regression). SVR finds a function that has most epsilon deviation from the instances and at the same time is as flat as possible. It's useful where problem cannot be solved by linear Regression.

ANN is mostly used algorithm in literature and it constitute human brain, for those reasons, we picked ANN to build model. The Random Forest was chosen because, it is an ensemble model and less sensitive to noisy data. It's good to mention here our main focus was find out best model for house price prediction in Sri Lanka along with highly impacted features.

The evaluation of models is another trickier part. There are various metrics to evaluate model (MSE, RMSE, MAE, R2 score). By comparing one evaluation metric between models is not good idea. We mainly used R2 score and RMSE to compare models. The simplest metrics MSE, we did not consider because incorrect prediction of few points (they may be outliers) with high deviation will result big MSE value. In our evaluation to compare models, we looked at R2 score, RMSE, and plot of graph (Predicted vs True prices).

Chapter 4

Experimental Setup and Implementation

The experiment compares the results with five algorithms which are LR,SVR, RF, MLP, K-NN. The goal of the experiment is to find out which algorithm performs well and what are the features highly affecting house price. The chosen evaluation metrics are R2 and RMSE.

These are the process flow in our experiment:

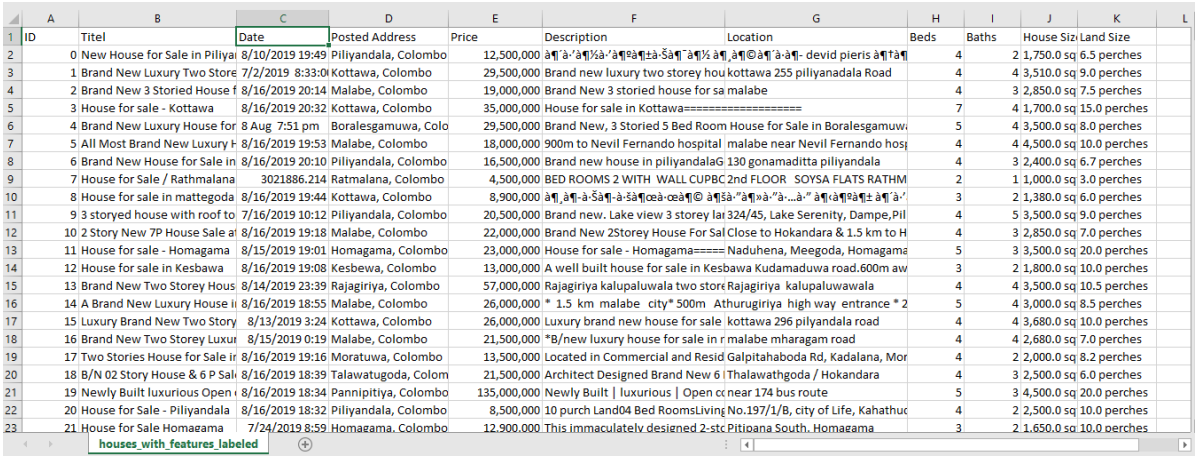
- Data Collection
- Data Preparation and Exploration
- Training and Testing

4.1 Data Collection

Data set was collected from ikman.lk which is a premier classified advertisement website operating in Sri Lanka using web scraping techniques. The site hosts user-generated classified advertisement, sorted by various categories.

The figure 4.1 shows some samples of collected data set. Each row represents one observation. 2185 data instances were collected with 11 attributes: ID, Title, Date, Posted Address, Price, Description, Location, Beds, Baths, House Size, Land Size.

Fig. 4.1 Collected Data Set



	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	Titel	Date	Posted Address	Price	Description	Location	Beds	Baths	House Siz	Land Size	
2	0	New House for Sale in Piliya	8/10/2019 19:49	Piliyandala, Colombo	12,500,000	à¶'à·'à¶¼à·'à¶ªà¶±à·Šà¶¯à¶½ à¶	à¶©à¶´à·à¶- devid pieris à¶†à¶	4	2	1,750.0 sq	6.5 perches	
3	1	Brand New Luxury Two Store	7/2/2019 8:33:0	Kottawa, Colombo	29,500,000	Brand new luxury two storey hou	kottawa 255 piliyanadala Road	4	4	3,510.0 sq	9.0 perches	
4	2	Brand New 3 Storied House f	8/16/2019 20:14	Malabe, Colombo	19,000,000	Brand New 3 storied house for sa	malabe	4	3	2,850.0 sq	7.5 perches	
5	3	House for sale - Kottawa	8/16/2019 20:32	Kottawa, Colombo	35,000,000	House for sale in Kottawa==========		7	4	1,700.0 sq	15.0 perches	
6	4	Brand New Luxury House for	8 Aug 7:51 pm	Boralesgamuwa, Colo	29,500,000	Brand New, 3 Storied 5 Bed Room	House for Sale in Boralesgamuw	5	4	3,500.0 sq	8.0 perches	
7	5	All Most Brand New Luxury H	8/16/2019 19:53	Malabe, Colombo	18,000,000	900m to Nevil Fernando hospital	malabe near Nevil Fernando hos	4	4	4,500.0 sq	10.0 perches	
8	6	Brand New House for Sale in	8/16/2019 20:10	Piliyandala, Colombo	16,500,000	Brand new house in piliyandalaG	130 gonamaditta piliyandala	4	3	2,400.0 sq	6.7 perches	
9	7	House for Sale / Rathmalana	3021886.214	Ratmalana, Colombo	4,500,000	BED ROOMS 2 WITH WALL CUPBC	2nd FLOOR SOYSA FLATS RATHM	2	1	1,000.0 sq	3.0 perches	
10	8	House for sale in mattegoda	8/16/2019 19:44	Kottawa, Colombo	8,900,000	à¶¸à¶·à·Šà¶œà·œà¶© à¶šà·"à¶»à·"à·...à·" à¶«à¶¤à¶± à¶'à·'		3	2	1,380.0 sq	6.0 perches	
11	9	3 storyed house with roof to	7/16/2019 10:12	Piliyandala, Colombo	20,500,000	Brand new. Lake view 3 storey ha	324/45, Lake Serenity, Dampe,Pil	4	5	3,500.0 sq	9.0 perches	
12	10	2 Story New 7P House Sale at	8/16/2019 19:18	Malabe, Colombo	22,000,000	Brand New 2Storey House For Sal	Close to Hokandara & 1.5 km to H	4	3	2,850.0 sq	7.0 perches	
13	11	House for sale - Homagama	8/15/2019 19:01	Homagama, Colombo	23,000,000	House for sale - Homagama=====	Naduhena, Meegoda, Homagama	5	3	3,500.0 sq	20.0 perches	
14	12	House for sale in Kesbawa	8/16/2019 19:08	Kesbewa, Colombo	13,000,000	A well built house for sale in Kesbawa Kudamaduwa road.600m aw		3	2	1,800.0 sq	10.0 perches	
15	13	Brand New Two Storey Hous	8/14/2019 23:39	Rajagiriya, Colombo	57,000,000	Rajagiriya kalupaluwala two store	Rajagiriya kalupaluwawala	4	4	3,500.0 sq	10.5 perches	
16	14	A Brand New Luxury House in	8/16/2019 18:55	Malabe, Colombo	26,000,000	* 1.5 km malabe city* 500m Athurugiriya high way entrance * 2		5	4	3,000.0 sq	8.5 perches	
17	15	Luxury Brand New Two Story	8/13/2019 3:24	Kottawa, Colombo	26,000,000	Luxury brand new house for sale	kottawa 296 pilyandala road	4	4	3,680.0 sq	10.0 perches	
18	16	Brand New Two Storey Luxur	8/15/2019 0:19	Malabe, Colombo	21,500,000	*B/new luxury house for sale in r	malabe mharagam road	4	4	2,680.0 sq	7.0 perches	
19	17	Two Stories House for Sale ir	8/16/2019 19:16	Moratuwa, Colombo	13,500,000	Located in Commercial and Resid	Galpitahaboda Rd, Kadalana, Mor	4	2	2,000.0 sq	8.2 perches	
20	18	B/N 02 Story House & 6 P Sal	8/16/2019 18:39	Talawatugoda, Colom	21,500,000	Architect Designed Brand New 6	Thalawathgoda / Hokandara	4	3	2,500.0 sq	6.0 perches	
21	19	Newly Built luxurious Open	8/16/2019 18:34	Pannipitiya, Colombo	135,000,000	Newly Built | luxurious | Open co	near 174 bus route	5	3	4,500.0 sq	20.0 perches	
22	20	House for Sale - Piliyandala	8/16/2019 18:32	Piliyandala, Colombo	8,500,000	10 purch Land04 Bed RoomsLiving	No.197/1/B, city of Life, Kahathuc	4	2	2,500.0 sq	10.0 perches	
23	21	House for Sale Homagama	7/24/2019 8:59	Homagama, Colombo	12,900,000	This immaculately designed 2-stc	Pitipana South, Homagama	3	2	1,650.0 sq	10.0 perches	

houses_with_features_labeled

Table 4.1 shows a sample observation from collected data with the values.

Table 4.1 Sample Instance from Collected Data

Feature	**Value**
ID	1
Title	“Brand New Luxury Two Storey House Piliyanadala”
Date	7/2/2019 8:33:00 AM
Posted Address	“Kottawa, Colombo”
Price	29500000
Description	“Brand new luxury two storey house for sale in kottawa 255 piliyanadala Road near to the Kottawa Junction. 4 bedrooms and 4 bathrooms. Comply with hot water, solar system,AC,CCTV camera, TV lobby, balcony and roof. Servant room and a bathroom. Complete pantry with a hob. Double carpoch. 100m to 255 Road.”
Location	“kottawa 255 piliyanadala Road”
Beds	4
Baths	4
House Size	“3,510.0 sqft”
Land Size	“9.0 perches”

The table 4.2 shows description about all the feature in collected data.

Table 4.2 Feature Description for Collected Data

Feature	Description
ID	Unique value for each instance to identify instances
Title	Add by the seller when posting advertisement
Date	Posted date
Posted Address	The location where he add the post
Price	Price of the house
Description	Features and description about the house
Location	House location
Beds	Number of bed rooms
Baths	Number of bath rooms
House Size	Size of the house in square feet
Land Size	Size of the land in perches

4.2 Data Preprocessing

The collected data has duplicate, inconsistent data instances and outliers. it was not processable. To prepare the data for the prediction system, some changes were made.

- More features are identified
- Feature types are identified
- Null values are handled
- Outliers are identified and removed

More features are identified with title and description field in the collected data since they are large text and more features needed to analyse the factors affecting house price.

4.2.1 Types of Variables

feature types are identified and all the variables are converted to numerical values to process the data.

36 locations were identified in the data set and all the locations are converted to numerical values. Table 4.3 shows the Id for each location.

Table 4.3 Location ID

Category ID	**Location**
1	Kottawa
2	Moratuwa
3	Colombo 10
4	Colombo 6
5	Colombo 8
6	Hanwella
7	Boralesgamuwa
8	Kesbewa
9	Athurugiriya
10	Colombo 15
11	Thalawatugoda
12	Colombo 5
13	Colombo 7
14	Colombo 9
15	Avissawella
16	Maharagama
17	Malabe
18	Battaramulla
19	Wellampitiya
20	Colombo 2
21	Kaduwela
22	Pilliyandala
23	Mount Lavina
24	Dehiwala
25	Homagama
26	Kohuwala
27	Colombo 14
28	Pannipitiya
29	Padukka
30	Rajagiriya
31	Nugegoda
32	Kotte
33	Ratmalana
34	Angoda
35	Nawala
36	Kolonnawa

There are 15 categorical variables identified. Table 4.4 shows the description of each categorical variable with values.

Table 4.4 Categorical Variables

Variables /Features	**Category**	
	0	**1**
New	Old House	New House
Garden	No Garden	Has Garden
Security	No Security	Has Security
AC	No AC	Has AC
Near Service	No Public Service Available Near	Public Service Available Near
Parking	No Parking	Has Parking
Luxury	Not a Luxury House	Luxury House
Pantry	No Pantry/Kitchen	Has Pantry/Kitchen
Hot Water	No Hot Water Available	Hot Water Available
Roller Gate	No Roller Gate	Has Roller Gate
CCTV	No CCTV	Has CCTV
Tiled	Not Tiled	Tiled
Designed	Not Designed Well	Designed Well
TV Lobby	No TV Lobby	Has TV Lobby
Location	(1-36) Different Locations	

There are 7 numerical variables are identified. Table 4.5 shows numerical variables and its types.

Table 4.5 Numerical Variables

Variable	**Type**
ID	Int
Price	Int
Beds	Int
Baths	Int
House Size	Float
Land Size	Float
Stories	Int

4.2.2 Null Values Handling

Figure 4.2 shows the created data with 535 instances set after feature identification.

Fig. 4.2 Created Data Set

ID	Location	Price	Beds	Baths	House Siz	Land Size	New	Garden	Security	AC	Near Serv	Stories	Parking	Luxury	Pantry	Hot Wate	Roller Gat	CCTV	Tiled	Designed	TV Lobby
2	17	19000000	4	3	2850	7.5	1	0	0	1	0	3	1	1	1	1	1	0	1	0	1
4	7	29500000	5	4	3500	8	1	0	1	0	1	3	1	1	1	0	0	0	0	0	0
5	17	18000000	4	4	4500	10	1	1	0	0	1	3	1	1	1	0	0	0	0	0	1
6	22	16500000	4	3	2400	6.7	1	0	1	0	1	1	1	0	1	0	0	0	0	0	0
7	33	4500000	2	1	1000	3	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
9	22	20500000	4	5	3500	9	1	0	0	0	0	3	1	1	1	0	1	0	1	1	0
10	17	22000000	4	3	2850	7	1	0	1	0	1	2	1	1	1	1	0	0	1	0	0
12	8	13000000	3	2	1800	10	1	0	1	0	1	1	1	0	1	0	0	0	1	0	0
13	30	57000000	4	4	3500	10.5	1	0	0	1	1	2	1	1	1	0	0	0	0	0	0
14	17	26000000	5	4	3000	8.5	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1
15	1	26000000	4	4	3680	10	1	0	0	0	1	1	1	1	1	0	0	0	0	0	1
16	17	21500000	4	4	2680	7	1	1	1	0	0	2	1	1	1	1	1	1	0	0	1
17	2	13500000	4	2	2000	8.2	0	0	0	0	0	2	1	0	1	0	0	0	0	0	0
18	11	21500000	4	3	2500	6	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0
19	28	1.35E+08	5	3	4500	20	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0
20	22	8500000	4	2	2500	10	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
21	25	12900000	3	2	1650	10	0	0	0	0	0	2	1	0	1	0	0	0	0	0	0
22	31	1.15E+08	5	5	6500	14.8	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0
23	18	1.95E+08	7	7	11500	20	1	0	0	0	0	3	1	1	1	0	0	0	0	0	0
24	22	24000000	5	4	3400	11.5	0	0	0	0	0	2	1	1	1	0	0	0	0	0	0
25	9	24000000	4	3	2200	19	0	1	0	0	1	1	1	0	1	0	0	0	0	1	0
26	26	1.05E+08	4	4	4600	37	0	0	0	0	0	2	1	1	1	0	0	0	0	0	0

Table 4.6 shows a sample observation from created data set.

Table 4.6 Sample Observation from Created Data Set

Feature	**Valuse**
ID	2
Location	17
Price	19000000
Beds	4
Baths	3
House Size	2850
Land Size	7.5
New	1
Garden	0
Security	0
AC	1
Near Service	0
Stories	3
Parking	1
Luxury	1
Pantry	1
Hot Water	1
Roller Gate	1
CCTV	0
Tiled	1
Designed	0
TV Lobby	1

After data set was created, data set was checked for null values. Figure 4.3 shows, attributes have no null values except "Luxury" which has one null value.

Fig. 4.3 Data Set With Null Values

```
In [5]: df.isnull().sum().sort_values(ascending = False)

Out[5]: Luxury          1
        TV Lobby        0
        Security        0
        Location        0
        Price           0
        Beds            0
        Baths           0
        House Size      0
        Land Size       0
        New             0
        Garden          0
        AC              0
        Designed        0
        Near Service    0
        Stories         0
        Parking         0
        Pantry          0
        Hot Water       0
        Roller Gate     0
        CCTV            0
        Tiled           0
        ID              0
        dtype: int64
```

Null Value was identified and corrected. Figure 4.4 shows, All the attributes have non-null values, meaning there are no null values in the data set.

Fig. 4.4 Data Set Without Null Values

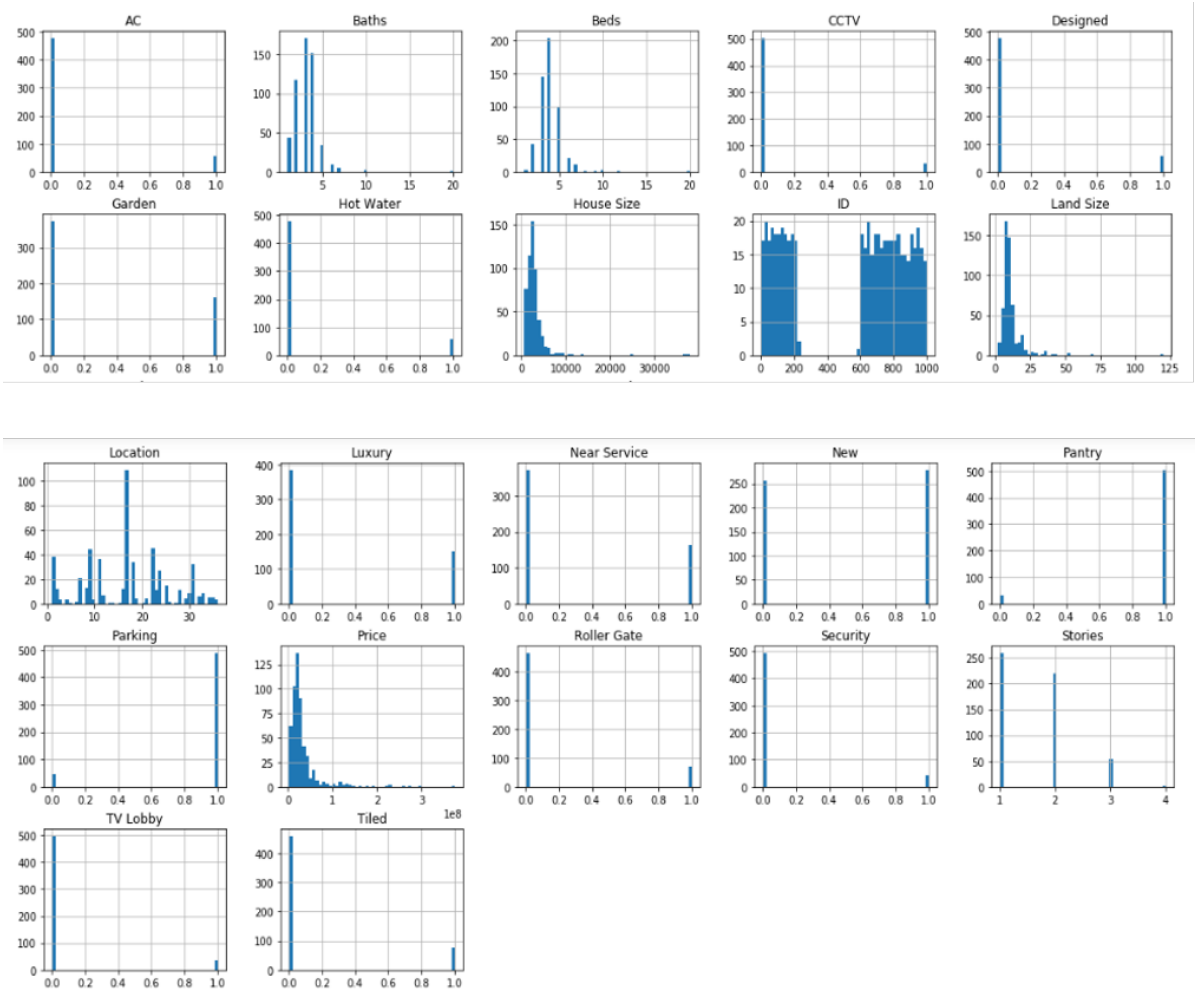
```
In [2]: import pandas as pd

        HOUSING_PATH = 'data_set_535_ano_sankee.csv'
        housing = pd.read_csv(HOUSING_PATH)
        housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 535 entries, 0 to 534
Data columns (total 22 columns):
ID              535 non-null int64
Location        535 non-null int64
Price           535 non-null int64
Beds            535 non-null int64
Baths           535 non-null int64
House Size      535 non-null float64
Land Size       535 non-null float64
New             535 non-null int64
Garden          535 non-null int64
Security        535 non-null int64
AC              535 non-null int64
Near Service    535 non-null int64
Stories         535 non-null int64
Parking         535 non-null int64
Luxury          534 non-null float64
Pantry          535 non-null int64
Hot Water       535 non-null int64
Roller Gate     535 non-null int64
CCTV            535 non-null int64
Tiled           535 non-null int64
Designed        535 non-null int64
TV Lobby        535 non-null int64
dtypes: float64(3), int64(19)
memory usage: 92.1 KB
```

4.2.3 Data Visualization

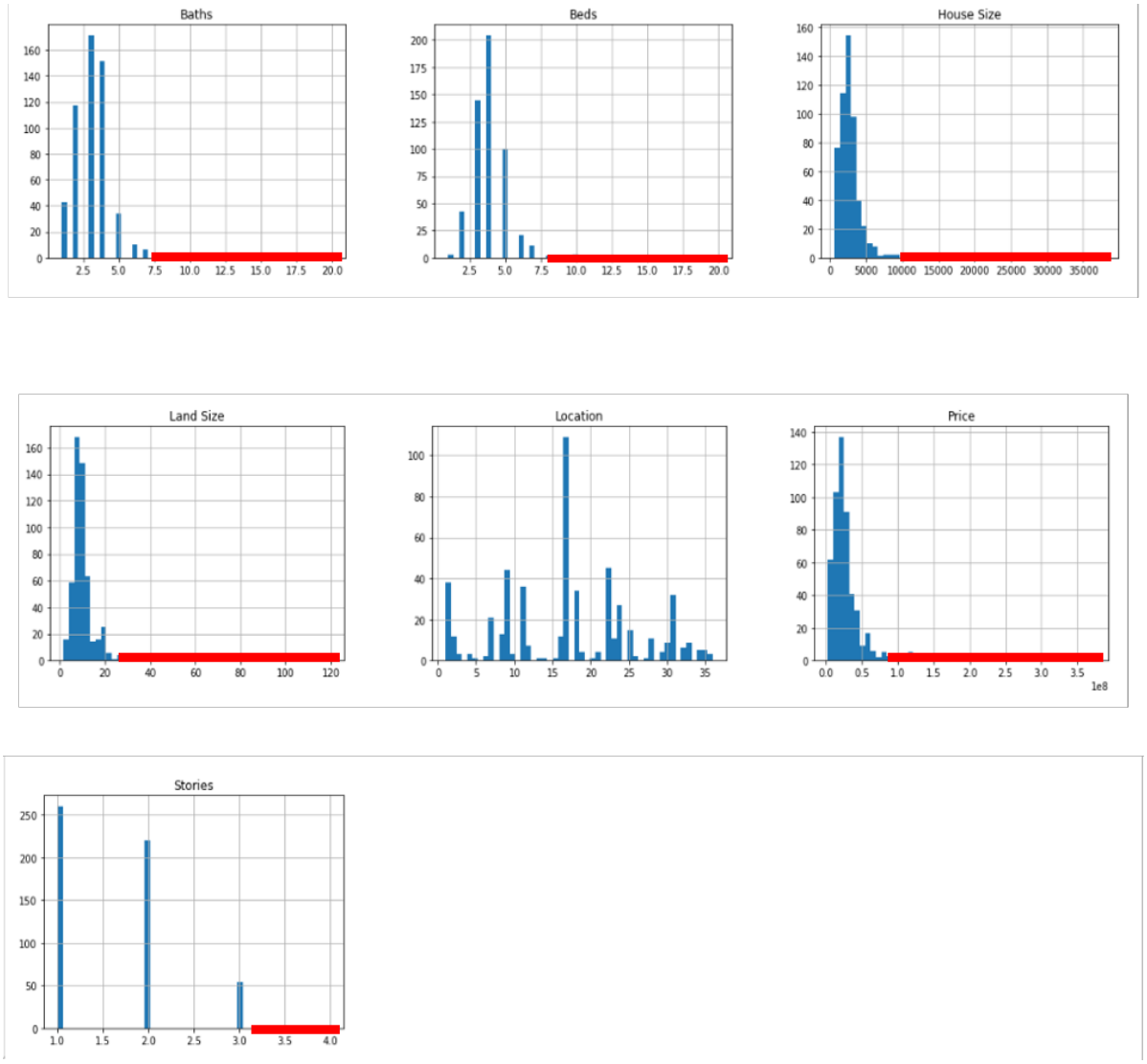
Feature distribution graphs was plotted for each attributes. Figure 4.5 shows, other than categorical variables (AC, CCTV, Designed, Garden, Hot Water, Location, Luxury, Near Service, New, Pantry, Parking, Roller Gate, Security, TV Lobby, Tiled), Numerical variables (Baths, Beds, House Size, Land Size, Price, Stories) have left skewed data. It is not good in statistics. it also indicates the data set has outliers.

Fig. 4.5 Feature Distribution



Histograms are tail heavily for some numerical attributes. Its harder to machine learning algorithms to detects patterns. These attributes need to be transformed more bell shaped distribution. As shown in figure 4.6, red line indicates the values may be outliers, these outliers will be calculated in the next section mathamatically.

Fig. 4.6 Feature Distribution For Numerical Features



4.2.4 Outliers Handling

Outliers were detected by doing statistical analysis. As shown in figure 4.7 Price, Beds, Baths, House Size and Land Size have high standard deviation and maximum values of those features are relatively high.

Fig. 4.7 Data Description

```
In [20]: df.describe()
```

Out[20]:

	ID	Location	Price	Beds	Baths	House Size	Land Size	New	Garden	Security	AC	Near Service
count	535.000000	535.000000	5.350000e+02	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000
mean	539.706542	17.026168	3.233493e+07	3.994393	3.188785	2888.479065	10.864579	0.521495	0.299065	0.076636	0.106542	0.302804
std	345.491451	8.997257	3.801996e+07	1.456603	1.454640	2761.046478	8.071565	0.500005	0.458277	0.266261	0.308819	0.459901
min	2.000000	1.000000	2.500000e+06	1.000000	1.000000	650.000000	1.700000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	148.500000	9.000000	1.470000e+07	3.000000	2.000000	1800.000000	7.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	676.000000	17.000000	2.250000e+07	4.000000	3.000000	2500.000000	9.800000	1.000000	0.000000	0.000000	0.000000	0.000000
75%	830.500000	23.000000	3.300000e+07	5.000000	4.000000	3300.000000	11.550000	1.000000	1.000000	0.000000	0.000000	1.000000
max	999.000000	36.000000	3.750000e+08	20.000000	20.000000	38000.000000	120.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Near Service	Stories	Parking	Luxury	Pantry	Hot Water	Roller Gate	CCTV	Tiled	Designed	TV Lobby
535.000000	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000	535.000000
0.302804	1.618692	0.914019	0.280374	0.943925	0.104673	0.132710	0.057944	0.140187	0.106542	0.067290
0.459901	0.670699	0.280599	0.449602	0.230281	0.306418	0.339579	0.233856	0.347506	0.308819	0.250758
0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	2.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	2.000000	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	4.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

With these information and visualisation graph in the previous section, we decided data has outliers. Outliers were detected using Tukey algorithm and removed from the data set. Figure 4.8 shows the implementation of tukey algorithm used in outlier detection.

Fig. 4.8 Tukey Algorithm

```
import numpy as np
def find_outliers_tukey(x):
    q1 = np.percentile(x, 25)
    q3 = np.percentile(x, 75)
    iqr = q3 - q1
    floor = q1 - 1.5 * iqr
    ceil = q3 + 1.5 * iqr
    outlier_indices = list(x.index[(x < floor) | (x > ceil)])
    outlier_values = list(x[outlier_indices])

    return outlier_indices, outlier_values
```

Price values above 60 millions calculated as output and removed from the data set. Figure 4.9 shows the outliers detected for Price.

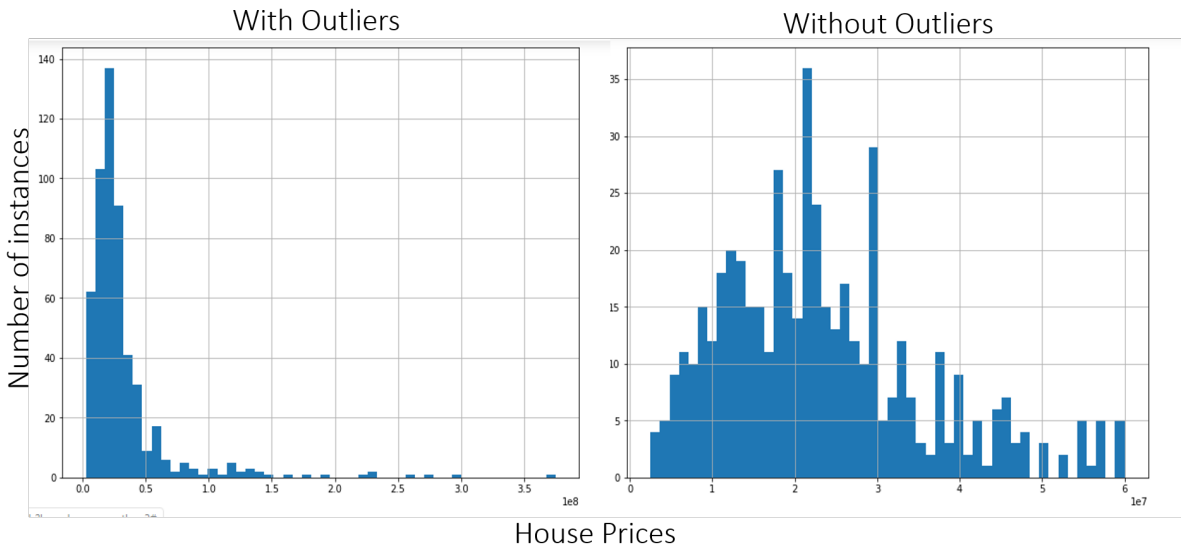
Fig. 4.9 Price Outliers

```
import pandas as pd
df = pd.read_csv('Machine-Learning-in-real-estate-industry/data_preprocessing/resource/data_set_535_ano_sankee.csv')
tukey_indices, tukey_values = find_outliers_tukey(df['Price'])
#print((tukey_indices))
tukey_values = tukey_values
print((tukey_values))
```

```
[135000000, 115000000, 195000000, 105000000, 375000000, 82000000, 62000000, 120000000, 110000000, 100000000, 140000000, 6750000
0, 80000000, 135000000, 65000000, 65000000, 180000000, 65000000, 80000000, 80000000, 127500000, 230000000, 85000000, 260000000,
75000000, 105000000, 127500000, 80000000, 68000000, 75000000, 85000000, 220000000, 160000000, 120000000, 95000000, 147500000, 1
15000000, 120000000, 65000000, 135000000, 275000000, 230000000, 85000000, 140000000, 300000000]
```

Figure 4.10 shows the Price distribution plot before and after outliers handling. The graph after outliers handling is more bell shaped than previous one.

Fig. 4.10 Price Distribution before and after Outliers Handling



This procedure is applied for all the numerical variables has outliers. Outliers were identified and removed from the data set. There were 487 data instances remains after outlier hadling for the training.

4.3 Training and Testing

Before building model, we implemented training and testing setup. One of the standard practice in machine learning project is split data set into two parts, one for testing another for training (1/3 test, 2/3 train). But in our case we only had few number of instances (487) after prepossessing and removal of outliers. If we split the data set like mentioned above, we may break some patterns exist in house data set.To ensure, all the data points which represents patterns are covered in training phase, we implemented both test-train split and cross validation for each and every model we built.

We split data set using library scikit-learn class train test split. Using this we can split easily the data set into training and test data set in various proportions. The parameters train size, test size, and random state will specify size of train data set size, test data set size, and seed of random number generator. The benefit of using this library class instead of splitting data set manually is it splits in random manner. The figure 4.11 shows usage of this class;

Fig. 4.11 Code Train Test Split

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=31)
```

Another approach we followed was we split the data set into three equal size bins and then pick one for testing and other two for training, likewise we changed the test set and training set. We did implemented three experiments for each and every models, then we averaged R2 score. We used library scikit-learn class KFold to perform above mentioned implementation(cross validation). The parameters are n splits, shuffle, and random state shown in figure 4.12. The n splits will specify number of splits(or numbers of bins) and shuffle will specify whether to shuffle the data before splitting into bins.

Fig. 4.12 Code KFold

```
shuffle = KFold(n_splits =3,
                shuffle=True,
                random_state =10)
```

The reason why we selected cross validation size as three (usually CV=10), if we selected CV = 10 then the probability of missing house instances represents some locations (The range of location attribute is 1, 36) in one bin is high (bin size will be approximately 48). To minimize of probability of missing some locations into one bin, the CV size = 3 was chosen.

Hyper parameters are the variables that govern the training process itself. For example in MLP the number of neutrons in hidden layers, number of hidden layers and activation function those are parameters known as hyper parameters. To find best hyper parameters , we have to train and see the results if results are poor then we have to adjust hyper parameters and repeat same procedure until we get reasonable results. This process is known as hyper parameter optimization or tuning,. To do this procedure manually will takes time. We used the library scikit-learn GridSearchCV to utilize time.

Fig. 4.13 Code GridSearchCV

```
grid_obj = GridSearchCV(estimator= pipeline,
                        param_grid = parameters,
                        cv =3,
                        verbose = 2,
                        n_jobs =1,
                        refit=True,
                        error_score='raise'
                        )
```

Linear Regression try to fit a straight line between house price and other features (beds, baths, ...) . We used library scikit-learn class LinearRegression. The parameters grids contains fit intercept , normalize values to tune using GridSearchCV. The Figure 4.13 shows the code.

Figure 4.14 shows the code for linear regression.

Fig. 4.14 Code Linear Regression

```
params_lr = {'fit_intercept':[True,False], 'normalize':[True,False]}
model(linear_model.LinearRegression(), params_lr, X_train, y_train, X, y, X_test, y_test)
```

Support Vector Regression finds a function that has most epsilon deviation from the instances and at the same time is as flat as possible. We used library scikit-learn class SVR. The parameters grid contains kernel functions and values for C show in figure 4.15.

Fig. 4.15 Code SVR

```
pipe_svr = Pipeline([('scl', StandardScaler()),
                     ('clf', SVR())])

param_svr = {'clf__kernel': ['linear', 'rbf', 'poly'],
             'clf__C': [1, 10, 100],
            }
```

K-NN is neighbour based algorithm. We used library scikit-learn KNeighborsRegressor to implement model. The grid params contains numbers of neighbors show in figure 4.16.

Fig. 4.16 Code K-NN

```
pipe_knn = Pipeline([('clf', KNeighborsRegressor())])
param_knn = {'clf__n_neighbors':[5, 10, 15, 25, 30]}
```

Random Forest is ensemble learning model. We used library scikit-learn RandomForestRegressor to implement model. The grid params contains max features, n estimators(trees), and max depth show in figure 4.17.

Fig. 4.17 Code Random Forest Regressor

```
pipe_forest = Pipeline([('clf', RandomForestRegressor())])
param_forest = {'clf__n_estimators': [10, 20, 50, 70, 100, 120, 130,140, 150, 200],
                'clf__max_features': [None, 1, 2,3],
                'clf__max_depth': [1, 2, 5, 10, 15, 20, 25]}
```

MLP is mimics of human brain and it's useful in finding unknown relationship between variables(features). We used library scikit-learn MLPRegressor. The grid params contains activation functions, hidden layers, learning rate, solver and alpha shown in figure 4.18.

Fig. 4.18 Code MLP Regressor

```
pipe_neural = Pipeline([('scl', StandardScaler()),
                        ('clf', MLPRegressor())])

param_neural = {'clf__alpha': [0.001, 0.01, 0.1, 1, 10, 100],
                'clf__hidden_layer_sizes': [(5),(10,10),(7,7,7),(20, 20,20, 20), (30,30, 30), (20, 20,20), (
                'clf__solver': ['lbfgs'],
                'clf__activation': ['relu', 'tanh'],
                'clf__learning_rate' : ['constant', 'invscaling']}
```

4.4 Research Instruments

We used python and scikit-learn library for our experiments.Usually prototype model will be built on python and actual model will implemented in other languages (C++).

So we also following same approach. The library scikit-learn is beginner friendly and have enough documentation.Vast community uses scikit-learn and most of algorithms are implemented in scikit-learn.

4.5 Pitfalls and workarounds

The website can be used by anyone, So initially panel suggested to use another data source if possible because of the non-reliability of the data. but we got to know that users required to provide a valid email address and need to pay in order to upload advertisements onto the site. All submitted ads are manually screened before being shown on the site as a security measure to ensure that the products and services being offered are authentic and legal.

We have contacted property valuation department in Colombo for additional data source. The attempt was given up because of getting delayed and there were some authorization related issues.

These house prices are not exactly the sold price of a house but we made an assumption since there is no much different in the advertisement price and sold price most of the cases. And outliers in the data-set can be also identified and removed by doing statistical analysis.

Our initial attempt was to identify the house features from ikman.lk data programmatically. But after some researches, we have found that it is a more complex problem and NLP has to be used. It was out of the scope of project at that time, It was given up. we will consider this in the next phase.

Chapter 5

Results and Analysis

5.1 Overview

This chapter presents the visualizations, feature relationship, model prediction result and analysis of results .

- Feature Relationship
- Model Prediction and Analysis

5.1.1 Features Relationship

Attributes of house are not independent, they have relationship between others. We looked at correlation matrix, it shows correlation between attributes. The figure 5.1 shows the correlation between attributes of all houses.

Fig. 5.1 Correlation of Features With Price (All Houses)

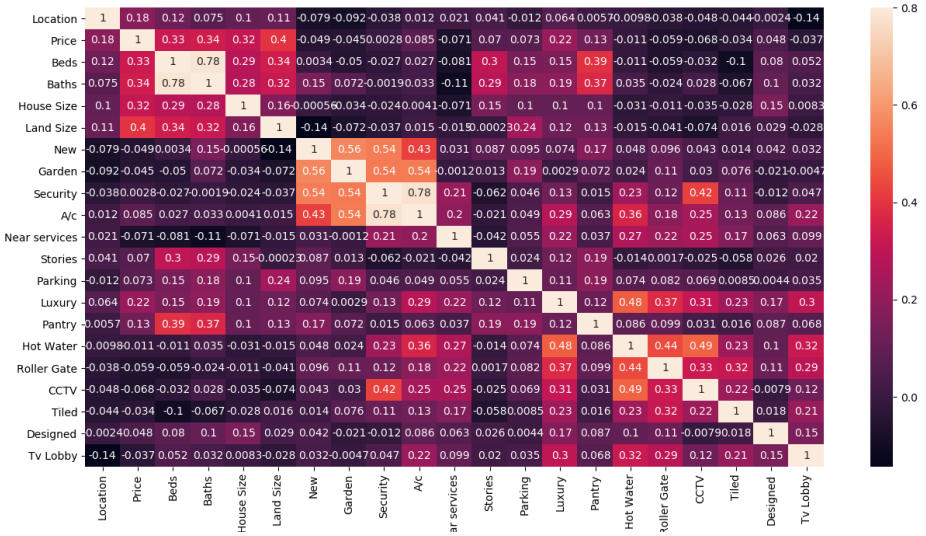


Table 5.1 Correlation of Features With Price (All Houses)

Features	**Correlation with Price**
Land Size	0.40
Baths	0.34
Beds	0.33
House Size	0.32
Luxury	0.22
Location	0.18
Pantry	0.13

By looking at table 5.1, the land size is highly correlated with house price. Correlation between house prices those house price is below 60 million shows some different to below case show in figure 5.2.

Fig. 5.2 Correlation of Features With Price (Price below 60 million)

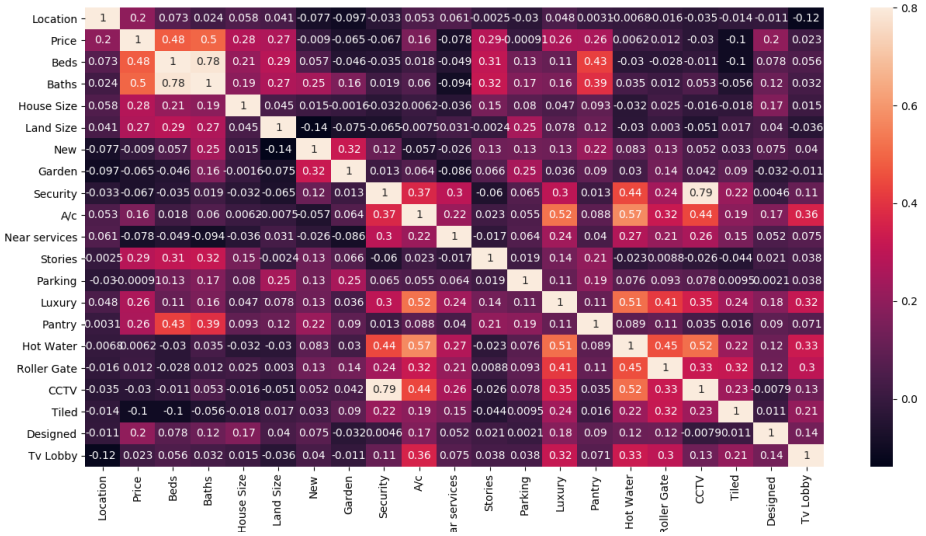


Table 5.2 Comparison of Correlation Change with Price

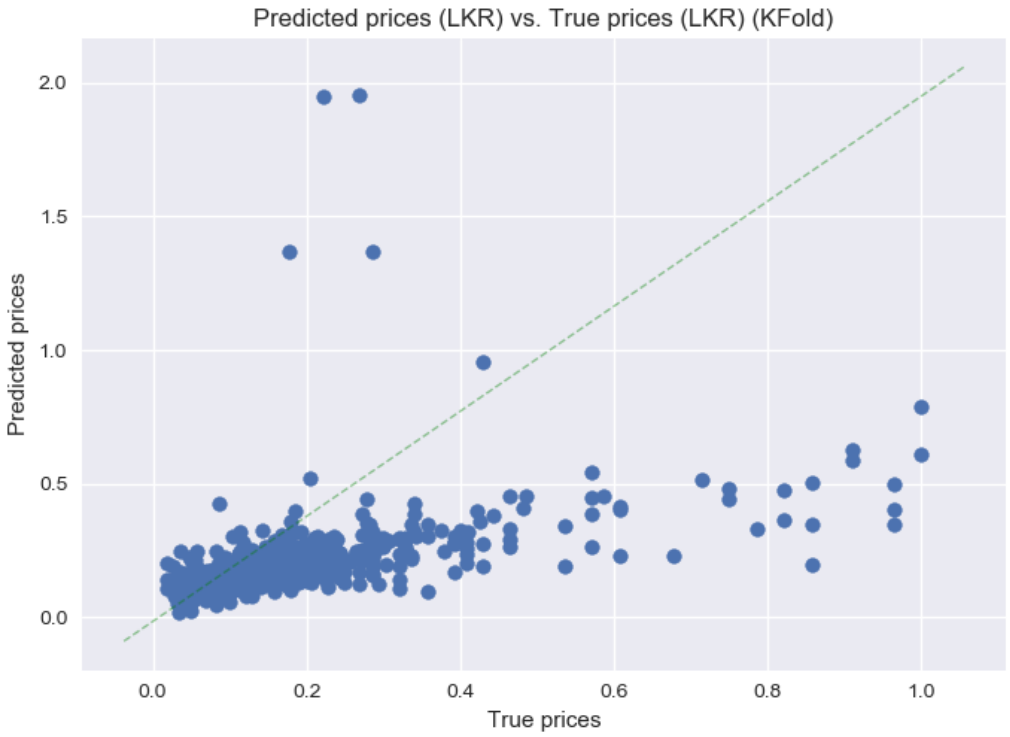
Features	**All Houses**	**Price < 60 million)**
Land Size	0.40	0.27
Baths	0.34	0.50
Beds	0.33	0.48
House Size	0.32	0.28
Luxury	0.22	0.26
Location	0.18	0.20
Pantry	0.13	0.26
Stories	0.07	0.029
Designed	0.048	0.20
Air Conditioner	0.085	0.16

The table 5.2 comparison identify correlation drop in house size and land size those houses price below 60 million at the same time other features in table shows increase of correlation with price.

5.1.2 Evaluation Results

We trained Linear Regression with different set of features(inclusion and exclusion of features) and normalized price, our experiment results in terms of R2 score always negative. The figure 5.3 shows one of experiments plots (R2 score close to zero).

Fig. 5.3 Linear Regression R2 Score: -0.0128



The negative R2 score and incorrect prediction in plot tells us there is no linear relation ship between price and house features.
The table 5.3 shows best results with SVR along with considered features set.

Table 5.3 SVR Model Results

Features	**R2 score**	**RMSE**
f1	0.49	0.11
f2	0.45	0.15
f3	0.48	0.14
f4	0.48	0.123

The f1, f2, f3, f4 indicate features set. They are below. f1 = (beds, baths, location, house size, land size, stories) , f2 = (beds, baths, location, house size,land size, price <= 60 million), f3 = (beds, baths, location, pantry, luxury, house size, land size, price <= 60 million), f4 = (beds, baths, location, pantry, luxury,house size, land size).

The figure 5.4 shows plot of predicted vs true price for the feature set f1.

Fig. 5.4 SVR R2 Score: 0.49

The results table 5.3 show an fact (pantry is highly correlated with price where price below 60 million). Exclusion of pantry in features f3 results relative low R2 score.

The table 5.4 shows best results with K-NN along with considered features set.

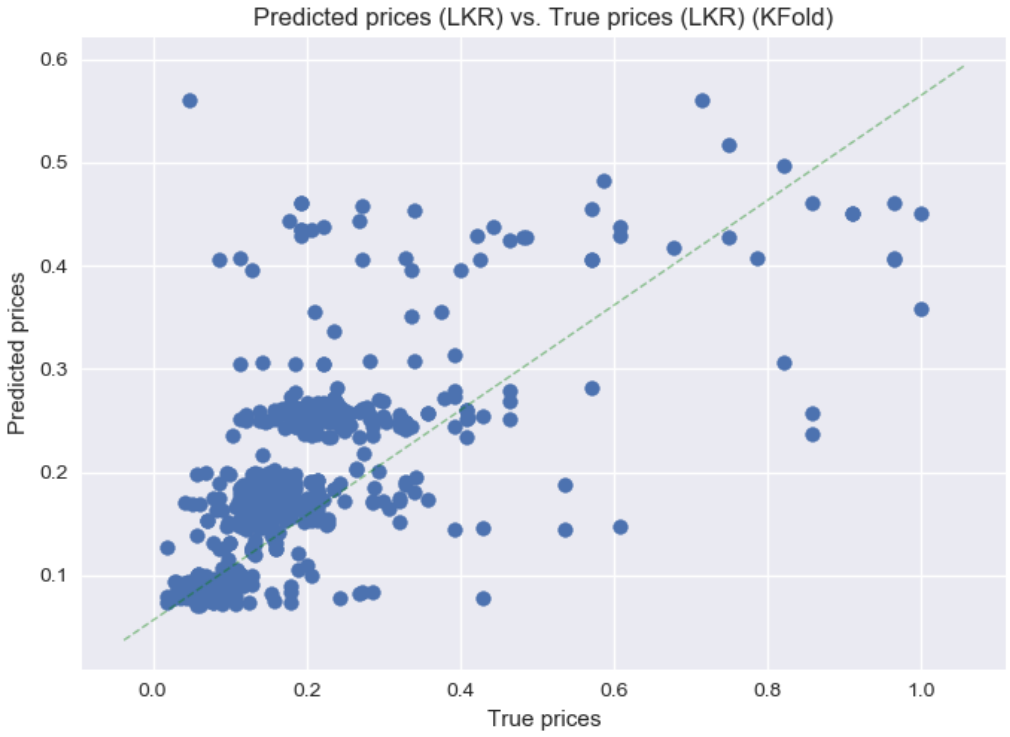
Table 5.4 K-NN Model Results

Features	**R2 score**	**RMSE**
f1	0.439	0.120
f2	0.402	0.157
f3	0.438	0.120

The f1, f2 and f3 indicate features set. They are below. f1 = (beds, baths, location, pantry, luxury, house size, land size) , f2 = (beds, baths, location, pantry, luxury, house size, land size, price <= 60 million), f3 = (beds, baths, location, house size, land size, Stories),

The figure 5.5 shows plot of predicted vs true price for the feature set f1.

Fig. 5.5 K-NN R2 score: 0.439



The table 5.4 results shows models built with K-NN score high (R2) when considering all houses and drop in R2 for house price below 60 million. One thing we can conclude that price below 60 million data set is more noisier than all house data set.

The table 5.5 shows best results with MLP along with considered features set.

Table 5.5 MLP Model Results

Features	**R2 score**	**RMSE**
f1	0.430	0.129
f2	0.487	0.151
f3	0.462	0.149

The f1, f2 and f3 indicate features set. They are below. f1 = (beds, baths, location, pantry, luxury, house size, land size) , f2 = (beds, baths, location, pantry, luxury, house

size, land size, price <= 60 million), f3 = (beds, baths, location, house size, land size, Stories), The figure 5.6 shows plot of predicted vs true price for the feature set f2.

Fig. 5.6 MLP R2 Score: 0.487

The results table 5.5 shows the MLP perform little better in terms of R2 score when considering features highly correlated with price (price below 60 million). This follows expected result (small increase in R2) from correlation matrices.

The table 5.6 shows best results with Random Forest along with considered features set.

Table 5.6 Random Forest Model Results

Features	**R2 score**	**RMSE**
f1	0.632	0.097
f2	0.541	0.138

The f1 and f2 indicate features set. They are below. f1 = (beds, baths, location, house size, land size) , f2 = (beds, baths, location, house size, land size,price <= 60 million),

The figure 5.7 shows plot of predicted vs true price for the feature set f1.

Fig. 5.7 Random Forest R2 Score: 0.632

The results of models built with Random forest show high R2 score and least RMSE, but when considering only houses price below 60 million, the R2 score dropped and RMSE increased. One of the possible reason to this behaviour is decrease of data instances. Other reason might be some other features also high correlation with price (air conditioner,pantry and luxury).

Chapter 6

Conclusions and Future Works

This study has presented set of processes to build a suitable ML house price prediction model by comparing prediction accuracy and performance. We were successful in predicting house prices in Colombo. This is the first machine learning house price prediction for Sri Lanka.

This research has conducted experiment using five regression algorithms which are LR, SVR, RF, MLP and K-NN. Results were compared using evaluation metrics R2 and RMSE. we have concluded Random Forest performs better than other algorithms with the evaluation matrices.

Land Size has been the factor highly affecting the Price with high correlation value with Price. Baths, Beds, House Size, Luxury, Location, Pantry are also have relatively high correlation with Price than other features.

We have got reasonable accuracy for the prediction. But comparing other researches it is relatively low. We have also got low accuracy after removing the outliers from the data.It can be improved by analysing feature interactions more and increasing number of data points in future works.

We tested five algorithms during short semester. In future more algorithms can be used out to carry out the experiment and Ensemble model can be built by combining various algorithms. Also, since the number of data points are small, it can be increased by automating the feature identification in data preprocessing stage since the data from ikman.lk contain more information and feature identification was done manually in this research.

References

[1] J. Y. Wu, "Housing price prediction using support vector regression," 2017.

[2] "Lanka Property Web." https://www.lankapropertyweb.com/.

[3] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso, "Identifying real estate opportunities using machine learning," *arXiv preprint arXiv:1809.04933*, 2018.

[4] R. Manjula, S. Jain, S. Srivastava, and P. Rajiv Kher, "Real estate value prediction using multivariate regression models," in *Materials Science and Engineering Conference Series*, vol. 263, p. 042098, 2017.

[5] J. Żak, "14th meeting of the euro working group on transportation (ewgt)-in quest for advanced models, tools and methods for transportation and logistics. editorial," 2011.

[6] O. Kitapci, Ö. Tosun, M. F. Tuna, and T. Turk, "The use of artifi cial neural networks (ann) in forecasting housing prices in ankara, turkey," *Journal of Marketing and Consumer Behaviour in Emerging Markets*, no. 1 (5), pp. 4–14, 2017.

[7] L. Yu, C. Jiao, H. Xin, Y. Wang, and K. Wang, "Prediction on housing price based on deep learning," *International Journal of Computer and Information Engineerin*, vol. 12, no. 2, pp. 90–99, 2018.

[8] A. Ng and M. Deisenroth, "Machine learning for a london housing price prediction mobile application," in *Imperial College London*, 2015.

[9] V. Limsombunchai, "House price prediction: hedonic price model vs. artificial neural network," in *New Zealand Agricultural and Resource Economics Society Conference*, pp. 25–26, 2004.

[10] N. Shinde and K. Gawande, "Survey on predicting property price," in *2018 International Conference on Automation and Computational Engineering (ICACE)*, pp. 1–7, IEEE, 2018.

[11] W. Tan and T.-N. Chou, "Combine grey relational analysis and weighted synthesis for housing price prediction," 2016.

[12] D. S. D. Dhvani Kansara, Rashika Singh, "Improving accuracy of real estate valuation using stacked regression."

[13] J. Wang, S. Hu, X. Zhan, Q. Luo, Q. Yu, Z. Liu, T. P. Chen, Y. Yin, S. Hosaka, and Y. Liu, "Predicting house price with a memristor-based artificial neural network," *IEEE Access*, vol. 6, pp. 16523–16528, 2018.

[14] A. Nguyen, "Housing price prediction," 2018.

[15] J. Mu, F. Wu, and A. Zhang, "Housing value forecasting based on machine learning methods," in *Abstract and Applied Analysis*, vol. 2014, Hindawi, 2014.

[16] P. Picchetti, "Hedonic residential property price estimation using geospatial data: a machine-learning approach," *Instituto Brasileiro de Economia*, vol. 4, 2017.

[17] R. E. Lowrance, *Predicting the market value of single-family residential real estate*. PhD thesis, Citeseer, 2015.

[18] J. Oxenstierna, "Predicting house prices using ensemble learning with cluster aggregations," 2017.

[19] I. S. H. Bahia, "A data mining model by using ann for predicting real estate market: Comparative study," *International Journal of Intelligence Science*, vol. 3, no. 04, p. 162, 2013.

[20] J. Frew and G. Jud, "Estimating the value of apartment buildings," *Journal of Real Estate Research*, vol. 25, no. 1, pp. 77–86, 2003.