FORECASTING LAND PRICES IN COLOMBO DISTRICT

A research project presented by

W.G.K.D. WEERASINGHE - (S/17/509)

to the

DEPARTMENT OF STATISTICS AND COMPUTER SCIENCE

FACULTY OF SCIENCE

In partial fulfillment of the requirement for the award of the degree of

BACHELOR OF SCIENCE IN DATA SCIENCE

UNIVERSITY OF PERADENIYA

SRI LANKA

2024

DECLARATION

I do hereby declare that the work reported in this project report was exclusively carried out by me under the supervision of Prof. Roshan D. Yapa. It describes the results of my own independent work except where due reference has been made in the text. No part of this project report has been submitted earlier or concurrently for the same or any other degree.

....

Date W.G.K.D. Weerasinghe (S/17/509)

Certified By:

1. **Supervisor:** Prof. Roshan D. Yapa
   Department of Statistics and Computer Science,
   Faculty of Science,
   University of Peradeniya.

   ....

   Date Signature

2. **Head of the Department:** Dr. Sachith Abeysundara
   Department of Statistics and Computer Science,
   Faculty of Science,
   University of Peradeniya.

   ....

   Date Signature

ABSTRACT

FORECASTING LAND PRICES IN COLOMBO DISTRICT

W.G.K.D. WEERASINGHE

Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Sri Lanka.

s17509@sci.pdn.ac.lk

Land prices are an important indicator of the economic and social development of a country. However, forecasting land prices is a challenging task due to the complexity of the factors that influence them. This study investigates the application of machine learning and geospatial analysis for forecasting land prices around 50 main cities within the Colombo district of Sri Lanka. Utilizing customer submitted data taken from 'Lanka Property Web' in the period of 2018 to 2023. Spatial network analysis was done and using shape files obtained from the Survey Department of Sri Lanka, allowing for the creation of new variables based on distances from specific locations to land plots.

Comprehensive model fitting was conducted to evaluate the performance of various Machine Learning algorithms, including Linear Regression, Gradient Boosting, XGBoosting, Random Forest and Artificial Neural Network. XG boosting emerged as the most accurate model with an accuracy of 62.23%, significantly exceeding the accuracy of linear regression (8.29%), GradientBoosting (60.01%), Random Forest (58.54%) and ANN (48.01%). To extend the forecast, a Seasonal Autoregressive Integrated Moving Average (SARIMA) model is fitted to predict land prices up to the second quarter of 2027.

The final product of this research is a user-friendly dashboard made for individual land sellers and buyers, offering a practical tool to stay informed about dynamic land market values. The dashboard empowers users with data-driven insights, facilitating well-informed decision-making in the ever-changing landscape of property transactions.

Keywords: Land Price Forecasting, Machine Learning, Spatial Analysis, Sri Lanka, Colombo District, Real Estate

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor, Prof. Roshan D. Yapa, Senior Lecturer at Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya. His guidance and invaluable advice were instrumental in the successful completion of this research project. During challenging times, his motivation and assistance were pivotal, contributing significantly to the overall achievement of the project. I extend my sincere appreciation to my external supervisor, Mr. Chamara Thewarapperuma, Senior research Analyst at Lanka Property Web and I would like to thank Mr. Tharindu Jayarathna, Head of Business Consultancy and Research at Lanka Property web. Their experience and expertise in the real estate sector gives an added advantage to this project.

Additionally, I would like to acknowledge all my lecturers and temporary academic staff members for imparting a strong foundation in statistics, providing comprehensive understanding, and offering guidance throughout the research journey. My heartfelt thanks go to my family and friends for their unwavering support. Without their encouragement, successful completion of this project would not have been possible. Lastly, I extend my gratitude to everyone who played a role in motivating and assisting me during this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ARIMA	Auto Regressive Integrated Moving Average
BPN	Back Propagation Neural Network
CBR	Case-Based Reasoning
CMA	Chennai Metropolitan Area
DT	Decision Tree
EM	Ensemble Methods
GBR	Gradient Boosting Regressor
GIS	Geographic Information System
IQR	Interquartile Range
KOSPI	Korean Stock Price Index
LPW	Lanka Property Web
MC	Municipal Council
MLP	Multilayer Perceptron
MSE	Mean Squared Error
MV	Max Voting
NN	Neural Networks
OOB	Out Of Bag
RBF	Radial Basis Function
RCBR	Regression Case-Based Reasoning
SARIMA	Seasonal Auto Regressive Integrated Moving Average
SVM	Support Vector Machine
WA	Weighted Averaging

Chapter 01:

INTRODUCTION

1.1. BACKGROUND OF THE STYDY

Sri Lanka has witnessed urban development in recent decades. This progress has been particularly evident in areas like Colombo, leading to a real estate market. The fluctuating prices of land as an aspect of the industry bring both opportunities and challenges for investors, developers, policymakers, and the general public. As the commercial center of Sri Lanka, the Colombo District has played a role in this real estate revolution. Its advantageous location, range of activities and cultural significance have made it an attractive destination for real estate investments. The increasing demand for land in the Colombo District has raised concerns about the sustainability of this growth and the necessity for making decisions in the real estate industry.

In partnership with “Lank Property Web”, which is the leading platform for selling real estate in Sri Lanka, this research project aims to tackle these urgent matters by focusing on predicting land prices in the Colombo District. Accurate predictions of land prices are crucial for purposes such as planning, investment analysis and policy development. Moreover, evolving real estate environment these forecasts empower individuals and organizations to make considered choices that can maximize returns and minimize risks.

Lanka Property Web has established itself as the primary platform for real estate selling in Sri Lanka. It boasts a vast repository of data, offering insights into property listings, market trends, and customer behavior. This collaboration provides us with a unique advantage in our endeavor to forecast land prices in Colombo District. Leveraging Lanka Property Web’s wealth of data and industry expertise, our research can benefit from real-time information, ensuring that our predictions are grounded in the most current market dynamics.

1.2. JUSTIFICATION OF THE PROBLEM

The importance of forecasting land prices in Colombo District is not limited to the real estate industry. It has more implications for economic stability, urban development, investor confidence, government policy, socioeconomic equity, and market transparency. Accurate land price predictions influence investment decisions and strategies. Colombo District, as the center of urban development in Sri Lanka, is particularly affected by these dynamics. Accurate land price forecasts boost investor confidence. Investors, whether local or foreign individuals, businesses, or institutions, often commit significant resources to real estate ventures. Reliable predictions empower them to make informed decisions, thereby minimizing risks and maximizing returns.

Fluctuating land prices can disrupt revenue projections and budget planning. An accurate understanding of land price trends is essential for ensuring stable revenue streams. Real estate markets play a vital role in determining property ownership to housing, impacting socioeconomic equity. In addition, accurate land price predictions can help identify specific bubbles in the real estate market, enabling early intervention to prevent market crashes. The collaboration with Lanka Property Web adds an essential layer of transparency to the real estate market. It encourages fair and competitive pricing, reduces information asymmetry between buyers and sellers, and contributes to market integrity. Ultimately, this research project, by the collaboration with Lank Property Web, stands at data-driven decision-making. It represents a timely and vital initiative for the betterment of Sri Lanka’s real estate landscape, contributing to its broader economy and the sustainable growth and development of Colombo District.

In addition to its economic implications, this research project underscores the social significance of accurate land price predictions. The equitable distribution of resources, accessibility to housing, and overall well-being are intricately linked to the stability and affordability of land in urban areas. As this navigates the complexities of Colombo District's real estate dynamics, collaboration with Lanka Property Web ensures that this analysis extends beyond financial considerations to broader societal impacts. This approach positions this research as a valuable resource for investors committed to fostering inclusive and sustainable development within the region.

1.3. OBJECTIVES OF THE STUDY

Build a Land Value Prediction Model that predicts land prices when customers enter their land details.

This project aims to make it easy for people to predict the value of land in their respective land which is going to be sold. With our user-friendly Land Value Prediction Model, they can enter details about a specific piece of land and get an accurate estimate of its price right away. This innovative tool will help individuals and investors make informed decisions, get more involved in the local real estate market, and feel confident that they're making the right choices. Also, this will bring more transparency to the decision-making process.

Develop a time series model for the prediction of future land prices in each city within the Colombo District

This research aims to create an advanced time series model dedicated to forecasting future land prices within every city in Colombo District. By analyzing historical data and using analytical techniques, this model will contribute valuable insights to urban planning, investment decisions, and policy formulation in the region's real estate sector.

Chapter 02:

LITERATURE REVIEW

2.1. SRI LANKAN REAL ESTATE PRICE INDEX

Real Estate can be categorized into commercial and residential properties, with varying definitions across countries. The types of property investments and their financial characteristics differ globally. Investing in Sri Lanka's exciting real estate scene beckons with promising returns, but navigating its price fluctuations can be a daunting task. The vital tool missing – a dedicated property price index – leaves investors, developers, and even homeowners adrift in a sea of uncertainty. Such an index would act as a trusty compass, charting the rise and fall of property values over time. Its absence shrouds the market in ambiguity, forcing decisions to be based on whispers and gut feelings rather than the guiding light of hard data.

This lack of clarity has tangible consequences. Investors may unknowingly overspend on their dream home, jeopardizing their financial security. Homeowners struggle to gauge their property's true value, potentially leaving money on the table. Even the government is hampered, lacking crucial insights for crafting effective policies and regulations. Fortunately, crafting this missing compass is not an impossible feat. A toolbox of methods exists, each with its own strengths and weaknesses, patiently awaiting deployment:

- **The Simple Mean/Median approach**, a straightforward method of averaging or finding the middle ground of all property prices, offers readily understandable data but can be easily swayed by outliers.

- **Stratification/Mix** Adjustment meticulously sorts similar properties into groups, tracking price changes within each category. This more nuanced approach, however, demands a wealth of detailed data.

- **Repeat Sales** meticulously tracks the price of a specific property over time, providing a precise measurement but limited in its scope.

- **Appraisal-Based methods** leverage the expertise of professional valuations, capturing market sentiment but potentially introducing subjectivity.

- **Hedonic Regression** unleashes the power of statistics, analyzing how individual property features influence price, though its complexity demands deeper expertise.

Ultimately, the chosen method hinges on the available data and the unique tapestry of Sri Lanka's real estate market. By weaving this vital tool into the fabric of the sector, we can transform blind speculation into informed decisions, guiding the bustling marketplace into a thriving ecosystem where clear, reliable information empowers everyone. Hedonic Regression Methods, in particular, focus on property attributes and their impact on prices. These methods involve identifying important characteristics, such as locational, structural, and neighborhood variables, and estimating their influence on Real Estate prices. However, the hedonic regression method has its challenges, including data consumption, misspecification due to missing variables, and the need for experienced judgment in model building.

2.2. FORECASTING LAND PRICES USING NEURAL NETWORKS

A project focuses on analyzing factors influencing land prices in the rapidly developing Chennai Metropolitan Area (CMA). Past data analysis reveals a non-linear trend, necessitating the development of a non-linear model for accurate forecasting. With substantial growth in population and economic activities, the study emphasizes the urgency of predicting land prices for various stakeholders. Previous research, employing diverse models such as time-series regression and Artificial Neural Networks, underscores the complexity of forecasting in the real estate market. The project aims to contribute a comprehensive non-linear model to assist decision-makers in navigating the evolving real estate landscape in CMA.

The application of Neural Networks (NN) in modeling land price trends is described, with economic and social indicators being utilized. The NN architecture, specifically a five-layered Back Propagation Neural Network (BPN) with three hidden layers and one output layer, is developed using the Levenberg-Marquardt algorithm for training. Thirteen indicators, including GDP, crude oil cost, inflation rate, and others, are used as input parameters, with unit land price as the output parameter. The normalization process is applied to input values, ensuring they are within the range of 0 to 1.

A total of 204 sets of exemplars are generated on a monthly basis from 1997 to 2013. The chosen NN architecture, depicted in Fig.2, consists of hidden layers with 20, 13, and 13 neurons, and one neuron in the output layer. The activation functions used are tan sigmoid for hidden layers and pure linear for the output layer. The training involves efficiently updating weights based on the 204 exemplars until error convergence falls below 0.01%.

To validate and forecast land prices, the trained network's updated weights are copied, allowing for implementation in subsequent years. Validation is performed on 24 exemplars for the years 2012 and 2013, while forecasting involves 24 exemplars for 2014 and 2015. The trained network is run again with the updated weights to predict land prices, showcasing the potential of NN.

2.3. CASE BASE REASONING FOR LAND PRICE PREDICTION

The proposal of Regression Case-Based Reasoning (RCBR) is characterized by the application of varying weights to independent variables in the domain of financial forecasting. In RCBR, a weight vector is determined through regression analysis to minimize prediction errors. The original independent variables are then transformed into weighted independent variables, and traditional Case-Based Reasoning (CBR) is subsequently performed. Data from the Korean Stock Price Index (KOSPI) were collected to assess the performance of three models: random walk, standard CBR, and regression CBR, as presented in 2006.

The approach involves searching for the top k cases that are closest to the new case, utilizing a suggested algorithm. The proposed solution is derived from the average value of the dependent variable among the searched cases. In this research, the top 3 cases are employed to provide a solution. The performance comparison of the models is facilitated through the examination of their abilities to predict financial outcomes based on historical data.

The weighted independent variables, determined through regression analysis, play a pivotal role in refining the traditional CBR machine's predictive capabilities. This methodology introduces a different approach to financial forecasting by assigning different weights to independent variables based on their significance in minimizing prediction errors. The proposed algorithm for selecting the top k closest cases ensures

a comprehensive evaluation of potential solutions, fostering an effective decision-making process in financial forecasting scenarios.

2.4. PRICE PREDICTION USING MULTIPLE REGRESSION

The rapid development observed in Gununganyar District is attributed to the high demand for housing. Over time, the transformation of land in Gununganyar, originally consisting of rice fields, ponds, and vacant lots, has resulted in the emergence of residential areas and apartments. This shift has not only influenced property prices but has also made the district an appealing investment destination.

Data collection involves on-site assessments utilizing land valuation methods such as market data comparison, cost approach, and income approach. The chosen methodology combines these methods, particularly emphasizing the comparison of market data and the cost approach, tailored to the specific conditions in the field. Market data comparison entails field surveys to acquire physical and legal data on assessed land or property. This process also includes obtaining market data for comparison purposes, focusing on objects with similar physical characteristics, legal aspects, and comparable environmental factors. The collected survey data are then adjusted to estimate the market value of the assessed object. In addition to market data comparison, the study employs the cost approach method to determine property values, encompassing both land and buildings. The estimation of land value utilizes the market data comparison approach, complemented by calculations of building costs. The cost approach also accounts for depreciation estimates, especially for structures that are not new, taking into consideration both physical and functional deterioration.

Sample selection considers the characteristics of villages or urban villages, ensuring a proportional representation of residential, commercial, and agricultural land use. Selected samples consist of vacant land plots, aligning with the base map used as the working reference. During the sample selection process, the identification of appropriate respondents is crucial. Respondents, serving as the primary data source, are chosen based on their ability to provide reliable descriptions and information about transaction prices for land transactions, including buying, selling, or leasing. Potential respondents include landowners involved in recent transactions, those intending to sell or lease their land, real estate agents, developers, and land or property tenants. If

primary respondents are unavailable, alternatives such as notaries, village heads, or other officials are considered as reliable sources of land market price information.

2.5. ENSEMBLE LEARNING FOR MODEL BUILDING

The decision tree (DT) is characterized as a tree structure resembling a flow chart, employing a branching technique to get potential outcomes of a decision. Its interpretability, simplicity, low computational cost, and graphical representation have led to increased utilization for classification tasks. The information gain approach is employed to determine the suitable property for each node in the generated tree. The entropy of the dataset is estimated to guide the decision tree's operation, considering the number of classes and instances proportion for each class.

The support vector machine (SVM) is a supervised machine learning tool for regression and classification tasks. It serves as a linear separator between two data nodes, detecting different classes in a multidimensional environment. SVM's implementation involves mapping the function of vectors in higher-dimensional space and finding a linear separating hyperplane with the best margin. The radial basis function (RBF) kernel is adopted for this study.

Neural networks (NN) are networks of interconnected components that accept input, actuate, and forward it to the next layer. In this paper, the Multilayer Perceptron (MLP) is employed as a supervised machine learning algorithm. The MLP learns a function by training on a dataset using the Adam optimizer. The logistic sigmoid activation function is utilized, and the backpropagation algorithm is employed for training.

Ensemble methods (EMs) are widely used in machine learning and statistics, combining multiple single classifiers or predictors to form a committee for improved decision-making. Two types of EMs, cooperative and competitive ensemble classifiers, are defined. Basic ensemble techniques, including Weighted Averaging (WA), Max Voting (MV), and Averaging, are discussed. Advanced ensemble techniques, such as Stacking (STK), Blending (BLD), Bagging (BAG), and Boosting (BOT), are also introduced. These advanced techniques aim to enhance predictive power and reduce bias, utilizing various strategies for combining individual models.

2.6. SPATIAL NETWORK ANALYSIS

Network accessibility in Turkey's Mersin Municipal Area, particularly concerning primary, middle, and high schools, as well as health facilities, has been analyzed. (Özer & Özlem, 2017) Primary education, mandatory for ages 6-14, is provided by the state within specified reach distances. However, the rural areas of Mersin often fall short of these standards due to the challenging topography of the Taurus Mountains.

For elementary schools, the study employs driving times instead of static distances, revealing that, despite coverage in driving distances of 5, 10, and 20 minutes, a significant population, particularly in mountainous regions, lacks access to elementary education. Similar challenges are observed for middle schools, with accessibility issues persisting despite the regulation-specified 1000m reach distance. High schools, concentrated in central areas, face less stringent regulations (2500m maximum distance) but still exhibit accessibility challenges, especially in more remote settlements.

The network accessibility analysis extends to health facilities, including hospitals, university hospitals, and small-scale health facilities. The study notes the dominance of location over the number of facilities in ensuring access. While central areas boast extensive coverage, outlying settlements face challenges, particularly in reaching small health facilities within 30 minutes.

Spatial analyses using road center lines indicate a polycentric urban structure in Mersin, shaped by topographical constraints. The road density is higher in central areas, creating a linear hub connecting Tarsus and Silifke. Global integration analysis highlights a linear integration core along the sea, with Silifke exhibiting higher integration than Mersin City Centre in certain lines. Local integration analyses reveal prominent local centers, emphasizing the significance of understanding the polycentric structure for effective urban development.

Chapter 03:

METHODOLOGY

3.1. DATA COLLECTION

The dataset used in this study was given by Lanka Property Web. It contains information from people who wanted to sell land on LPW. All the information is about lands in the Colombo district and was collected between January 2018 and September 2023. The original dataset had 185 columns and had almost everything about a land, like where it is, how much it costs, how big it is, its shape, Geo coordinates, and more. However, many of the columns had null values. The combined dataset has 58,352 entries and also has details about the people who put up the ads. Because of the availability of data for certain areas, this research looks at around 50 main cities in the Colombo district. This study mainly looks at three variables, 'main city', the 'ad posted date', and the 'price of the land'. Along with those columns, other few columns were selected.

Table 3.1: Description of the main columns of the dataset

Column Name	Description
ad_id	Advertisement unique ID number
UID	User ID Number
street	Street where particular land is located
main_city	Main City name which the land belongs
heading	Heading of the advertisement
description	Advertisement description
posted_date	Advertisement posted date
price_land_pp	Price of the land per perch
size_land	Size of the land in perches
lat	Latitude of the exact location of the land
lng	Longitude of the exact location of the land

3.2. DATA PREPROCESSING

In the data preprocessing phase, columns containing all null values or featuring less than 5% of data were systematically excluded. The Interquartile Range (IQR) method was employed to identify and eliminate outliers in the land prices. Specifically, only bare land plots were selected, aligning the dataset with the targeted emphasis on data points. This process aimed to improve the overall quality and reliability of the dataset, establishing a robust foundation for subsequent analyses and insights in the research.

3.3. K-PROTOTYPES CLUSTERING

The goal of the k-prototype clustering algorithm is to categorize the dataset (X) which has both categorical and numerical columns into k clusters by minimizing the cost function E, given by the formula:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} \times d(x_i, Q_i) \tag{3.1}$$

The k-prototype algorithm proceeds through the following steps:

1. Selection of Initial Prototypes:
   - Choose k initial prototypes for k clusters from the dataset X.
2. Allocation of Data Objects:
   - Assign each data object in X to the cluster whose prototype is closest, based on the dissimilarity measure.
   - Update the prototype of the cluster after each allocation.
3. Retesting Similarity:
   - After all data objects have been assigned to a cluster, reevaluate the similarity of data objects against the current prototypes.
   - If a data object is found whose nearest prototype belongs to a different cluster than its current one, reallocate the data object to that cluster and update the prototypes of both clusters.
4. Iteration:
   - Repeat the similarity retesting in Step 3 until no data object changes clusters after a full cycle test of X.

The k-prototype clustering algorithm iteratively refines the clusters by updating prototypes and reallocating data objects until stability is achieved, ensuring a comprehensive and effective grouping of the dataset into k clusters. (ji, et al., 2013)

3.4. SPATIAL NETWORK ANALYSIS

3.4.1. FUNDAMENTAL CONCEPT

Distance analysis constitutes various Geographic Information System (GIS) applications, providing valuable insights into spatial relationships and connectivity. In its simplest form, distance represents the measure of how far one point is from another. The fundamental premise of distance analysis lies in determining the separation between two points. The Euclidean distance, calculated as the straight-line distance between points in a Cartesian plane, serves as the baseline metric. However, this simplicity is augmented by the consideration of real-world complexities.

3.4.2. CONNECTIVITY ANALYSIS

The Distance Accumulation tool integrates various modifying parameters to refine straight-line distance calculations. This tool, along with the Distance Allocation tool, allows researchers to explore optimal paths, accounting for barriers, surface texture, and traveler characteristics. (Distance Analysis, 2023)

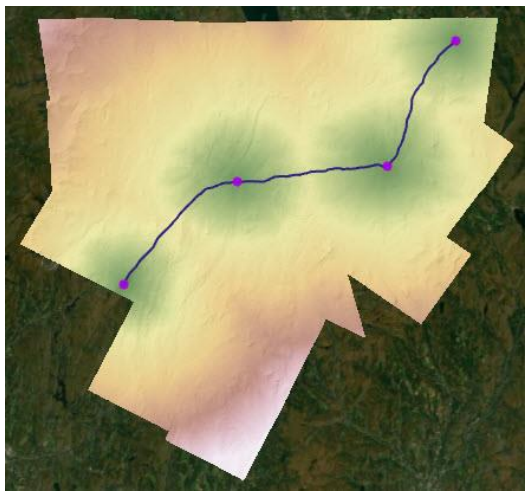


Figure 3.1: Least cost paths between the points

Connectivity analysis further extends the utility of distance analysis by exploring optimal paths and corridors between specific locations. Whether connecting regions with the optimal network, tracing optimal paths between specific locations, or

establishing corridors based on cost thresholds, these applications enhance GIS capabilities.

3.4.3. ARCGIS

In ArcGIS 8, two indispensable functions are provided by the Buffer Wizard and the Geoprocessing Wizard. The Buffer Wizard facilitates the creation of rings around features at specified distances, a technique employed in generating 0.5-mile buffers around selected points. This functionality is also present in ArcView but necessitates defined map units for proper processing.

The Geoprocessing Wizard in ArcGIS 8 have various operations, all of which can be executed passively:

- Features can be dissolved based on a specified attribute, aggregating those with identical attribute values.
- Layers can be merged, appending features from multiple layers into a single layer while retaining attributes with the same name.
- One layer can be clipped based on another, akin to a cookie cutter operation, without altering the input layer's attributes.
- The intersection of two layers results in an output layer containing features with attribute data from both layers.
- Union of two layers combines features from an input layer with polygons from an overlay layer, producing an output layer with attributes and full extent from both layers.

ArcView, chosen for its compatibility with the Point & Polyline Tools v1.2 extension, played a pivotal role in the study. The tools and extensions were actively utilized to isolate and clean data, enabling comprehensive connectivity measure calculations and analysis, detailed further in the study. The evaluation of connectivity measures necessitates Network Analyst, an extension unavailable in ArcGIS 8. Despite the availability of ArcGIS newer versions, the latest mapping program from ESRI, during

the research calculations, ArcGIS 8 was preferred for its robust Geoprocessing features and user-friendly display and interface. (Tresidder, 2005).

3.5. MACHINE LEARNING

3.5.1. COST FUNCTION

During model training in, a training dataset composed of input features and labeled outcomes is presented to the chosen algorithm. The model iteratively refines its internal parameters through cross validation and gradient boosting, aiming to minimize the error between its predictions and the actual labels in the training data. This error is quantified by a loss function, a mathematical measure of error. Common loss functions include mean squared error for regression tasks and cross-entropy for classification tasks. Model training happens in the following way (Aziz, et al., 2020)

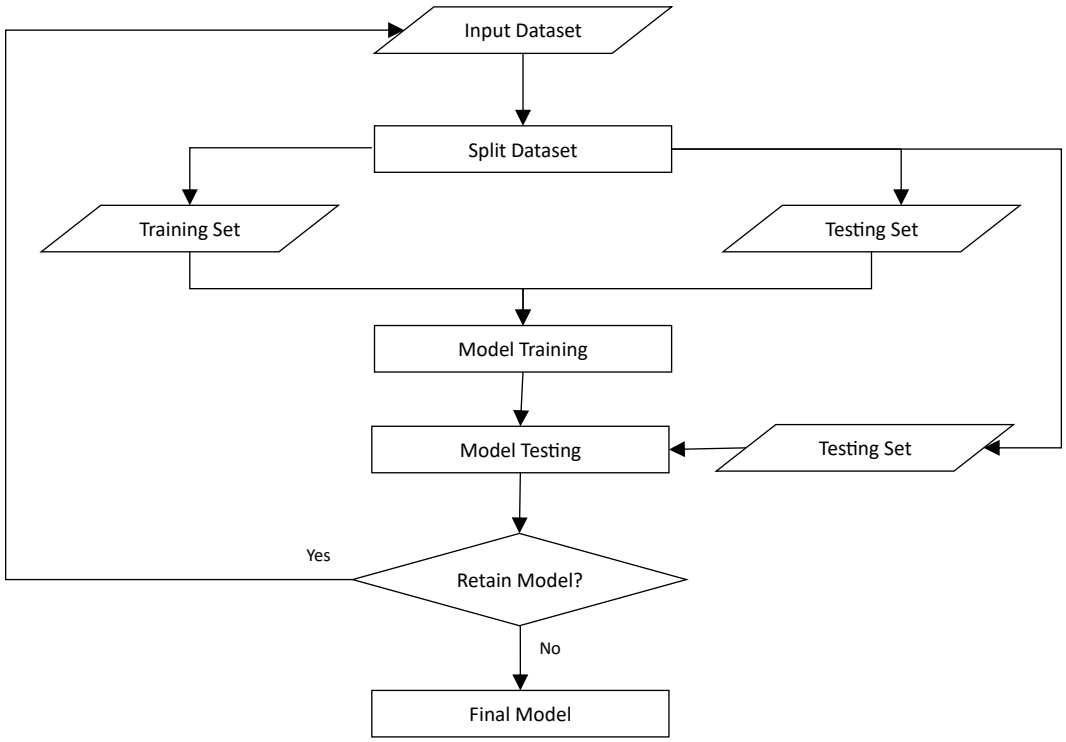


Figure 3.2: Model Training in Machine Learning

The cost function, that used in this project, is Mean Squared Error (MSE). The mean squared error serves as a quantitative measure assessing the proximity of the predicted response value for a given observation to the corresponding true response value (Hodson, et al., 2021). This assessment is encapsulated by the formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{3.2}$$

In this expression, y_i denotes the true response value for the ith observation and $f(x_i)$ represents the prediction yielded for the same observation. The utilization of Mean Squared Error is integral for measuring the accuracy of the learning algorithm.

The prediction of $f(x_i)$ for the ith observation becomes a pivotal element, allowing the evaluation of the model's accuracy. A smaller MSE conveys that the predicted response closely aligns with the true response variable. Conversely, a higher MSE implies a notable disparity between the predicted and true response variables, indicative of reduced model accuracy.

3.5.2. RANDOM FOREST

The random forest regressor is a powerful ensemble learning algorithm used for both classification and regression tasks. The algorithm operates by creating an ensemble of decision trees, each trained on a different bootstrap sample of the original data. In the context of regression, the goal is to predict continuous numerical outcomes. The random forest algorithm follows a specific set of steps to build a robust and accurate predictive model (Liaw, Andy, Wiener, & Matthew, 2001).

Firstly, the algorithm draws *n* bootstrap samples from the original data. Each bootstrap sample serves as a training set for an individual decision tree in the ensemble. During the construction of each tree, a modification is introduced at each node. Instead of choosing the best split among all predictors, a random subset of predictors is sampled, and the best split is chosen from among those variables. This process introduces an element of randomness, enhancing the diversity of the individual trees in the ensemble. The prediction for new data is made by aggregating the predictions of all the trees in the ensemble. For regression tasks, the predictions are averaged across the trees. This

ensemble approach helps mitigate overfitting and increases the model's generalization ability.

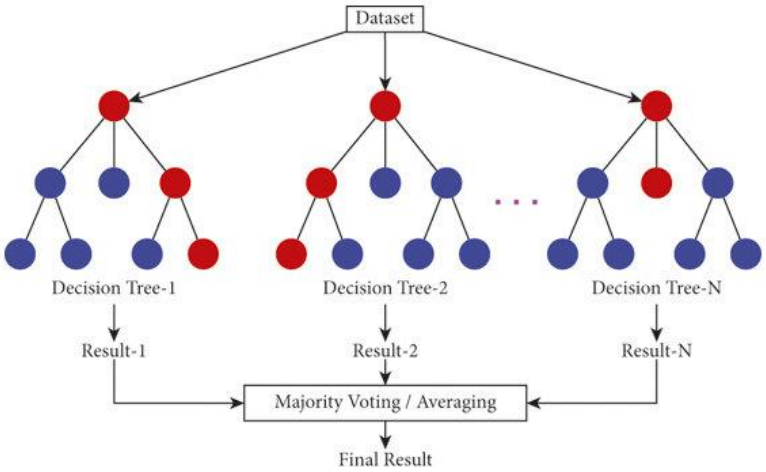


Figure 3.3: Illustration of Random Forest Trees

Source: (Khan, et al., 2021)

To estimate the error rate of the model, the out-of-bag (OOB) data, which is not included in the bootstrap sample for each tree, is utilized. (Breiman, 2001) At each bootstrap iteration, the data not in the bootstrap sample is predicted using the corresponding tree, and the OOB predictions are aggregated. The OOB estimate of the error rate is then calculated based on these predictions. It has been observed that the OOB estimate is quite accurate when a sufficient number of trees have been grown.

The random forest regressor shares commonalities with other ensemble methods that involve generating random vectors to govern the growth of each tree. Examples include bagging, where a random selection without replacement is made from the training set for each tree, and random split selection, where the split at each node is chosen randomly from the K best splits. The random forest method stands out for its ability to handle both classification and regression tasks seamlessly.

Variable importance and proximity estimates are stabilized with a large number of trees, although the ranking of importance remains relatively stable. In classification problems with highly unbalanced class frequencies, adjusting the prediction rule may be necessary. For instance, using the type='prob' argument in the predict method and setting a threshold based on class probabilities can be effective, especially in scenarios where classes are imbalanced.

To optimize memory usage, users can set the argument keep.forest=FALSE when running the random forest algorithm if prediction of test data is not required. This approach keeps only one tree in memory at any given time, potentially saving substantial memory and execution time, especially for large datasets or a high number of trees. Given its embarrassingly parallel nature, the random forest algorithm can be distributed across multiple machines, and the votes from each forest can be aggregated to obtain the final result. This parallelization strategy offers scalability and computational efficiency in large-scale applications. (Segal & Mark, 2004).

3.5.3. LINEAR REGRESSION

Within the framework of the multivariable regression model, the dependent variable is characterized as a linear function of independent variables, denoted as,

$$Y = a + b_1 \times X_1 + \cdots\cdots\cdots + b_n \times X_n \tag{3.3}$$

This model facilitates the determination of a regression coefficient b_i for each independent variable X_i. The coefficient of determination determines the overall association between the independent variables and the dependent variable Y. It corresponds to the square of the multiple correlation coefficient, representing the correlation between Y and $b_1 \times X_1 + \cdots\cdots\cdots + b_n \times X_n$

This method enables the examination of multiple independent variables simultaneously, adjusting their regression coefficients for more specific effects between variables. However, this approach poses challenges as the number of observations often falls short of the model's requirements, with a general guideline suggesting at least 20 times more observations than variables.

Furthermore, including too many irrelevant variables in the model can lead to overadjustment. This entails that some of the irrelevant independent variables may exhibit a bad effect by chance. While such inclusion may enhance the fit within the studied dataset, its applicability beyond the dataset is compromised due to random effects (Schneider, Hommel, & Blettner, 2010).

3.5.4. GRADIENT BOOSTING REGRESSOR

The Gradient Boosting Regressor (GBR) is another ensemble model that is an iterative collection of sequentially arranged tree models so as the next model learns from the error of the former model (Otchere, Ganat, Ojero, Tackie-Otoo, & Taki, 2022). This machine learning model makes predictions using "boosting" of the ensemble of weak prediction models, often decision trees, to form a more robust model. A GBR with M number of trees can be stated as;

$$f_M = (x_j) = \sum_{m}^{M} \gamma_m h_m(x_j) \tag{3.4}$$

where h_m is a weak learner that performs poorly individually, γ_m is a scaling factor adding the contribution of a tree to the model. GBR uses the gradient descent loss function to minimise errors by updating the initial estimation with the new estimation. Thus, a final model is created with the combination of all preliminary estimations with suitable weights. The GBR model implemented in this study is from the *GradientBoostingRegressor()* method provided in Scikit-learn package.

3.5.5. XG BOOSTING REGRESSOR

XGBoost is a powerful machine learning algorithm designed for high scalability. It falls under the category of decision tree ensembles and operates on the principle of gradient boosting, where the objective function is incrementally built by minimizing a specified loss function. This methodology is traditional gradient boosting, but XGBoost introduces various optimizations for improved efficiency.

$$L_{xgb} = \sum_{i=1}^{N} L(y_i, F(x_i)) + \sum_{m=1}^{M} \Omega(h_m) \tag{3.5}$$

$$\Omega(h_m) = \gamma T + \frac{1}{2}\lambda||\omega||^2 \tag{3.6}$$

Where T is the number of leaves of the tree and ω are the output scores of the leaves. In XGBoost, decision trees act as the fundamental classifiers, and the algorithm controls tree complexity through a variation of the loss function. The split criterion of decision trees incorporates the loss function, resulting in a pre-pruning strategy. By adjusting a parameter denoted as γ, one can influence the minimum loss reduction gain

required to split an internal node. Higher values of lead to simpler trees, promoting model simplicity.

XGBoost introduces a shrinkage parameter, which reduces the step size in the additive expansion of the objective function. Strategies such as controlling tree depth will decide the complexity. This reduction in complexity not only enhances model interpretability but also accelerates the training process and diminishes storage requirements. About overfitting concerns and enhance training speed, XGBoost employs randomization techniques. These include random subsampling of data instances and column subsampling, which introduce variability into the training process. The introduction of randomness helps combat overfitting and contributes to a more robust and generalized model. In terms of computational efficiency, XGBoost implements methods that go beyond optimizing ensemble accuracy. The algorithm reduces the computational complexity associated with finding the best split, a typically time-consuming task.

3.5.6. ARTIFICIAL NEURAL NETWORK

In the analysis process, neural networks prove to be a valuable tool. It is well-established to handle nonlinear problems. Artificial Neural Network (ANN), a technology for valuation, surpasses the neural network of the human brain. It consists of numerous processing elements that can be stored in local memory. Mainly it uses the model of the human brain. The fundamental components of ANN are the nodes(neurons), which are simple processing units, and the edges connecting them. These nodes collectively form a neural network, and they are organized into input, hidden and output layers (Abidoye, Rotimi, Chan, & Albert, 2017).

Neural networks consist of processing nodes and the connections between them. Each node in a layer is connected to every node in the next layer through these connections. The neural network's structure typically includes an input layer, one or more hidden layers, and an output layer. This configuration enables the neural network to efficiently process large amounts of data, providing accurate values for predictions.

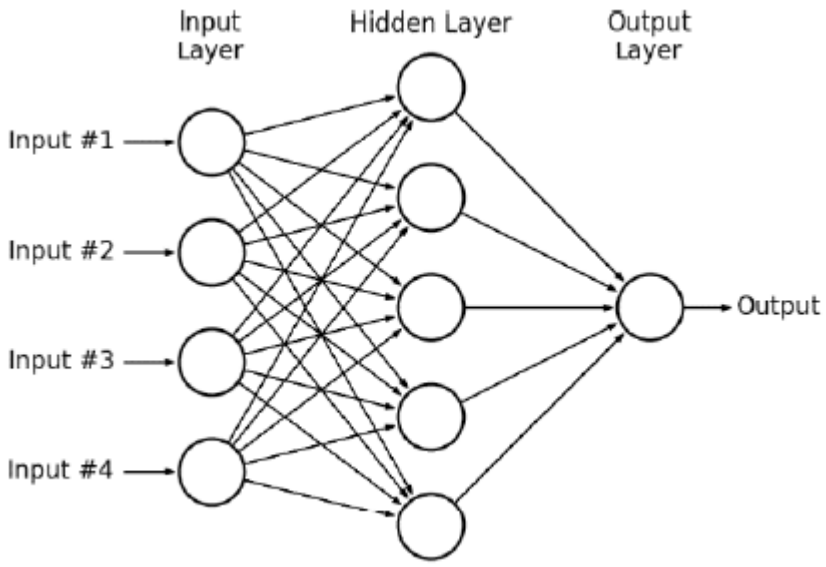


Figure 3.4: Neural Network Architecture

Source: (Velumani, Nampoothiri, & Kavithra, 2019)

3.6. SEASONAL AUTO REGRESSIVE INTEGRATED MOVING AVERAGE

To model time series, the traditional statistical models, including moving average, and SARIMA, can be employed. These models are characterized by linearity, as the future values are constrained to be linear functions of past data. Time series forecasting models, primarily used for predicting demand, were investigated under an autoregressive moving average hypothesis. It was calculated the seasonal variation of demand using historical data and validated the models by examining forecast performance.

In an earlier model, it incorporated seasonal factors to enhance forecasting accuracy, with the seasonal factors derived from a multiplicative model. The classical ARIMA approach becomes impractical, and in many cases, impossible to determine a model when the seasonal adjustment order is high, or its diagnostics fail to indicate that the time series is stationary after seasonal adjustment. In such cases, the static parameters of the classical ARIMA model are considered the principal constraint for forecasting high-variable seasonal demand. Additionally, a limitation of the classical ARIMA approach is its requirement for a large number of observations to determine the best-fit model for a data series.

An ARIMA model is denoted as an ARIMA model (p, d, q), where:

- p is the number of autoregressive terms,
- d is the number of differences, and
- q is the number of moving averages.

The autoregressive process assumes that Y_t is a linear function of preceding values, as expressed by equation: (Fattah, et al., 2018)

$$Y_t = \alpha_1 Y_{t-1} + \epsilon_1 \tag{3.7}$$

Each observation in this equation comprises a random component (random shock, ε) and a linear combination of previous observations. The self-regression coefficient in this equation is represented by α_1.

Chapter 04:

RESULTS AND DISCUSSION

4.1. DATA PREPROCESSING

In the initial dataset, it had 185 columns and 58,352 rows, but approximately 120 columns were found to be entirely null. These columns were eliminated from the dataset. Following this, a filtration process was implemented to retain information about bare land advertisements. A manual inspection was conducted, resulting in the removal of columns that contained data related to house advertisements, as they were irrelevant for the current study. Further preprocessing was carried out by eliminating columns that did not contain at least 5% of data, thereby focusing the analysis on more meaningful and informative variables relevant for this study. Duplicated columns were also identified and removed. Additionally, columns which had the same data in different formats were excluded.

A new column was added to calculate the land price per perch for all rows, providing a standardized metric for comparison. To enhance the dataset's robustness, outliers in the land prices were identified and removed using the interquartile range method grouped by each city. This resulted dataset had of 27,000 rows with more reliable and representative price data. The standardization of main city names was imperative due to the inclusion of customer-added data in various formats. Ultimately, 50 main cities were selected depending on the data availability. To improve easiness of analysis, the ad posted date was converted to a date-time format. The geographic coordinates (latitude and longitude) of land locations were retained, adding spatial context to the dataset.

4.2. EXPLORATORY DATA ANALYSIS

4.2.1. AVERAGE PRICES CALCULATION

Over the years from 2018 to 2023, the number of advertisements posted for each year was calculated, with a focus on grouping them by city to get an idea about available data. The few cities were removed those with fewer than 20 entries.

Table 4.1: Number of Advertisements posted in each year.

Main City	2018	2019	2020	2021	2022	2023	Row Total
Homagama	171	189	176	195	2322	1611	4664
Piliyandala	103	156	135	317	682	1980	3373
Padukka	49	28	36	44	1457	958	2572
Battaramulla	352	293	322	501	512	459	2439
Kaduwela	72	68	82	392	861	611	2086
Pannipitiya	129	104	98	114	1010	372	1827
Nugegoda	227	237	265	371	352	265	1717
Thalawathugoda	257	245	189	220	429	276	1616
Dehiwala	191	232	194	347	362	244	1570
Malabe	145	132	128	217	329	345	1296
Colombo 5	155	189	174	263	255	160	1196
Kahathuduwa	0	10	13	28	228	882	1161
Athurugiriya	95	105	74	160	521	206	1161
Rajagiriya	124	134	181	142	169	209	959
Nawala	136	157	152	207	173	123	948
Kottawa	99	111	100	110	333	172	925
Kesbewa	19	33	27	257	296	261	893
Maharagama	87	84	84	103	272	194	824
Mount Lavinia	112	104	64	166	157	152	755
Colombo 3	81	76	157	115	143	132	704
Pita Kotte	64	78	90	118	135	116	601
Moratuwa	102	82	53	96	160	102	595
Colombo 6	95	83	88	126	105	90	587
Colombo 10	14	22	20	34	175	293	558
Boralesgamuwa	72	76	94	107	101	107	557
Polgasowita	10	14	18	29	81	392	544
Watareka	6	5	3	2	314	212	542
Hanwella	28	34	31	44	192	131	460
Colombo 8	78	88	57	102	70	53	448
Rathmalana	60	57	48	86	70	84	405
Ethul Kotte	50	24	57	98	99	76	404
Colombo 7	76	92	79	66	39	50	402
Hokandara	49	60	40	40	119	80	388
Colombo 4	51	62	35	87	68	70	373
Avissawella	13	21	20	43	70	45	212
Angoda	25	38	15	18	34	20	150
Colombo 2	15	26	13	17	34	29	134
Kolonnawa	18	36	21	17	25	16	133
Colombo 15	18	18	21	25	19	17	118
Peliyagoda	13	12	8	23	25	26	107
Colombo 13	12	15	4	15	24	12	82
Colombo 9	4	6	7	19	12	30	78
Wellampitiya	16	13	12	16	11	10	78
Thalahena	17	10	9	22	9	6	73
Kosgama	8	11	7	18	12	9	65
Colombo 14	3	5	6	15	10	10	49
Thalangama	2	2	2	8	11	19	44
Colombo 12	4	2	9	20	6	1	42
Thalagala	1	4	3	0	10	4	22
Wattala	1	1	0	1	3	3	9
Kadawatha	0	1	2	0	2	2	7
Colombo 11	1	2	1	2	0	0	6

The table shows the number of cases in different main cities of Sri Lanka from 2018 to 2023. The total column shows the sum of cases for each city over the six years. The city with the highest number of entries in 2023 was Piliyandala, with 1980 entries,

followed by Homagama, with 1611 cases. The city with the highest total number of entries over the six years was Homagama, with 4664 entries, followed by Piliyandala, with 3373 entries, and Padukka, with 2572 entries. Wattala, Kadawatha and Colombo 11 were omitted because of the lack of data.Following this a table of metrics, including average price were computed for each city. This allowed for an analysis between each city.

Table 4.2: Price metrics of each city

Main city	Min price	Max price	Average price	Std price	Median price	25% percentile	75% percentile
Angoda	200000	1160000	697277.7	223988.7	698484.5	538928.5	867500
Athurugiriya	250000	1300000	836564.9	248781.4	850000	650000	1025000
Avissawella	73529	240000	173325.5	36193.13	173611	138888	200000
Avissawella	85000	210000	156497.1	40058.96	149444	138888	192500
Battaramulla	800000	4000000	2459332	618228.7	2500000	2000000	2950000
Boralesgamuwa	750000	2800000	1692748	377578.7	1700000	1400000	1950000
Colombo 1	250000	685000	496213.6	139888.9	550000	380000	600000
Colombo 10	340000	704166	617810.8	49454.02	637804	600000	650000
Colombo 12	7500000	8500000	7628947	277546.7	7500000	7500000	7500000
Colombo 13	2600000	7083333	4773906	1150138	5000000	3800000	5966667
Colombo 14	3000000	9000000	6157751	1561705	6500000	4925000	7225000
Colombo 15	1650000	4716981	3457112	627951.4	3500000	3000000	4000000
Colombo 2	12000000	25000000	17190647	2612710	16500000	15020450	18750000
Colombo 3	9000000	28000000	18733459	3291340	19500000	16500000	20454545
Colombo 4	5434782	30000000	14705200	3104619	15000000	13000000	16000000
Colombo 5	3000000	13500000	6914172	1782343	6500000	5537500	8143750
Colombo 6	3200000	14285714	8309393	1882244	8450000	7000000	9619354
Colombo 7	9500000	25000000	16162624	2517027	16000000	14090909	18000000
Colombo 8	3300000	12000000	7634377	1746715	7900000	6500000	9000000
Colombo 9	2100000	6065217	4496390	972750.5	4619565	3970588	5163043
Dehiwala	1300000	5700000	3442399	680364.9	3500000	3000000	3950000
Ethul Kotte	900000	4500000	2989868	710503.4	3000000	2400000	3500000
Hanwella	47692	400000	301316.7	61350.52	325000	285000	335000
Hokandara	450000	1850000	1067634	318982.2	1000000	850000	1352174
Homagama	120689	883534	557194.7	106031.5	570000	510000	639508
Kaduwela	165000	565000	404782.5	67049.33	395000	365000	475000
Kahathuduwa	180000	800000	519515.6	108785.7	550000	430000	600000
Kesbewa	265000	1200000	812591	184580.5	900000	700000	900000
Kolonnawa	350000	1675000	943966.2	281838.5	950000	712500	1082576
Kosgama	75000	275000	182996.8	52289.48	200000	145000	225000
Kosgama	550000	700000	616666.7	76376.26	600000	575000	650000
Kottawa	375000	1875000	1085662	304793.6	1062500	825000	1350000
Maharagama	520000	3000000	1788811	406689.2	1925000	1500000	2000000
Malabe	460000	2500000	1321009	346393.8	1350000	1075000	1550000

Malabe	722222	833333	797222	52410.96	816666.5	780555.5	833333
Moratuwa	520000	2000000	1223163	293344	1200000	1000000	1468750
Mount Lavinia	1300000	4642857	2881325	558767	2900000	2500000	3300000
Nawala	2000000	6500000	4146727	788386.9	4000000	3500000	4750000
Nugegoda	0	5272727	2840019	726873.9	2800000	2400000	3500000
Padukka	80000	450000	268233.4	40671.67	260000	250000	290000
Pannipitiya	420000	2700000	1830013	313326.2	1961290	1850000	2000000
Piliyandala	240000	1450000	812220.3	232480.8	875000	650000	975000
Pita Kotte	1150000	4305555	2681087	636406.2	2600000	2200000	3200000
Polgasowita	180000	950000	419558.6	81634.74	430000	380000	430000
Rajagiriya	0	7000000	3581343	1145920	3543293	2800000	4500000
Rathmalana	1000000	3500000	1960801	466134.9	1900000	1626016	2300000
Thalagala	500000	661538	602207.9	63389.46	650000	550000	650000
Thalahena	800000	2100000	1427653	297582	1433333	1200000	1687419
Thalangama	1700000	2700000	2163462	278503.3	2050000	2000000	2300000
Thalangama	1100000	1900000	1546429	289550.1	1575000	1375000	1750000
Thalawathugoda	500000	3000000	1823085	415209.1	1825000	1500000	2000000
Watareka	275000	550000	417032.1	35158.91	407692	400000	426576
Wattala	1500000	1500000	1500000	0	1500000	1500000	1500000
Wellampitiya	400000	1100000	769000.9	173018.1	750000	625000	900000

The table shows the descriptive statistics of the bare land prices in different main cities in Colombo District. The minimum and maximum prices vary significantly across different areas. The average prices give an indication of the central tendency of the data. Colombo 3 has the highest average price (18,733,459), while Avissawella has a lower average price (173,325.5). Colombo 12 has a high standard deviation (277,546.7), indicating a wider range of property prices, whereas Avissawella has a much lower standard deviation (36,193.13). Colombo 2, Colombo 3, Colombo 7, and Colombo 8 stand out with high maximum prices, reflecting premium real estate in these areas. Avissawella and Kosgama have relatively low prices and lower variability, making them potentially more affordable and stable markets. It can be clearly seen in the below visualization.

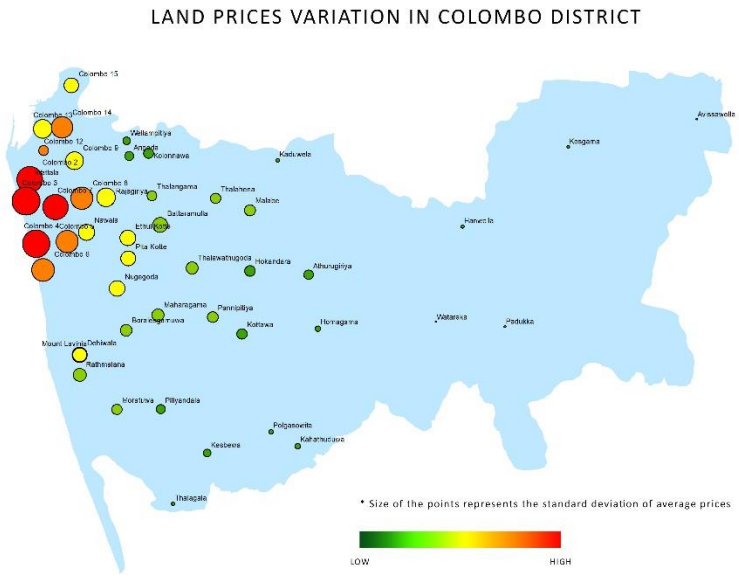


Figure 4.1: Deviation of prices in main cities

In this figure it shows the average prices in colors and standard deviation of price in size of the circles. It can be clearly seen that urban areas near Colombo city have higher prices and standard deviation compared to other places.

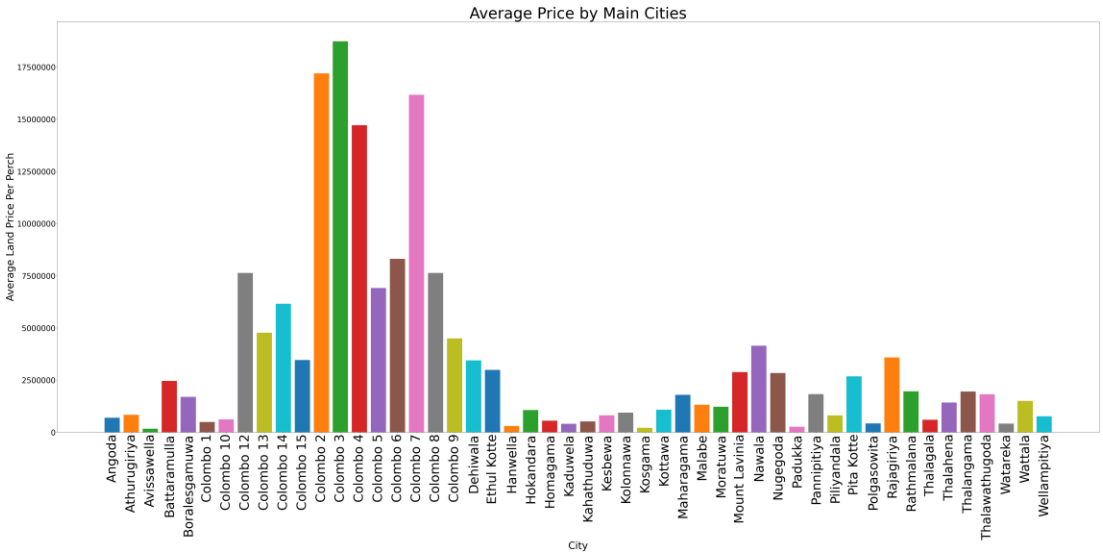


Figure 4.2: Average price by city from 2018 to 2023

This is a bar graph that shows the average price by main cities in Colombo District. Some cities have very high average prices, such as Colombo 1 to Colombo 15 while others have very low average prices, such as Avissawella and Kosgama. Colombo 1 to Colombo 15 is close to the commercial heart of Sri Lanka. So, this leads to the higher average prices in those areas.

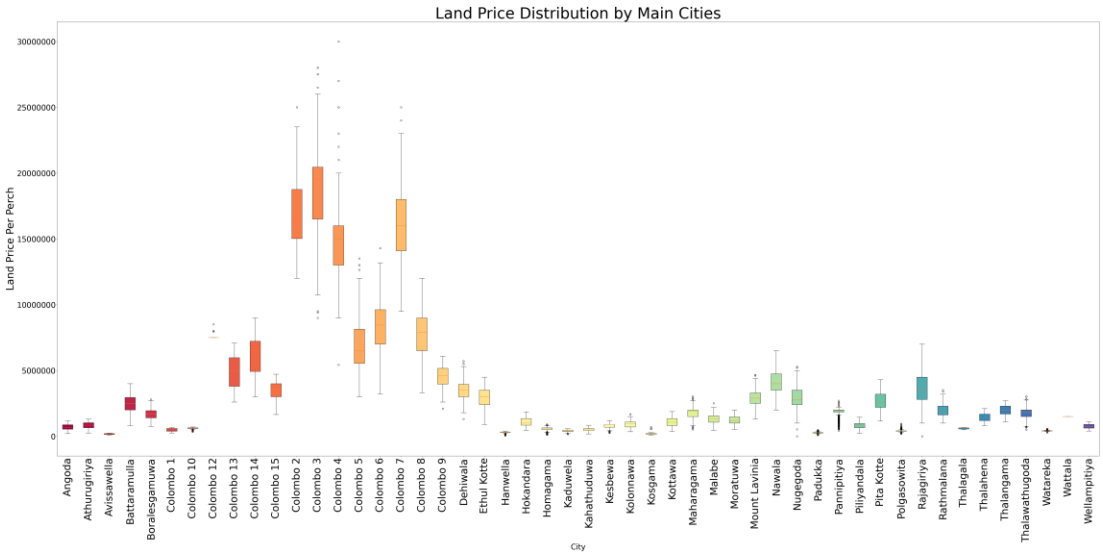


Figure 4.3: Land price distribution in main cities

In the above boxplot, it shows how the land price is distributed within the main cities. Furthermore, analysis was applied by delving into year-wise average price calculations.

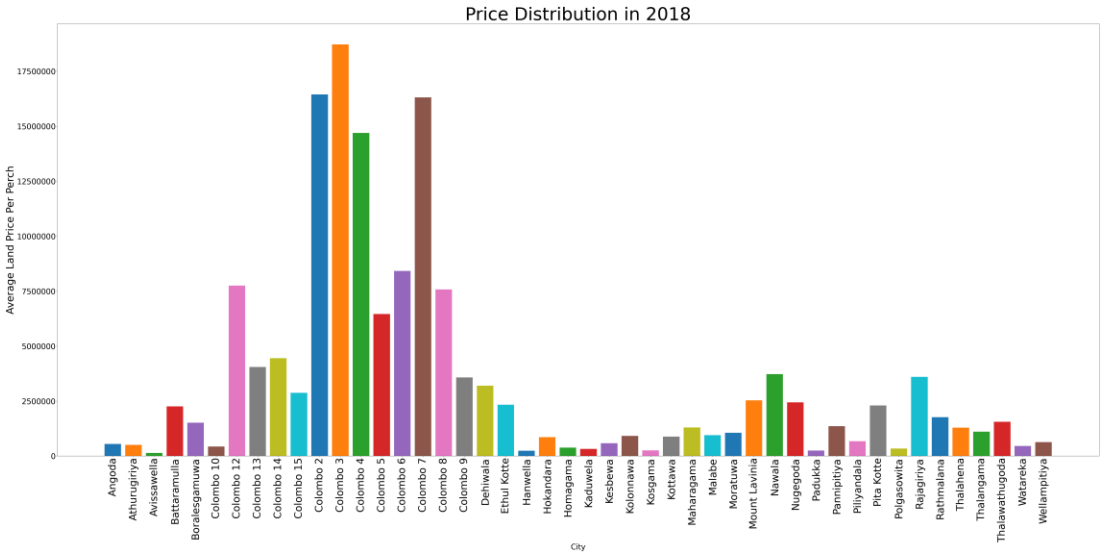


Figure 4.4: Land Price Distribution in 2018

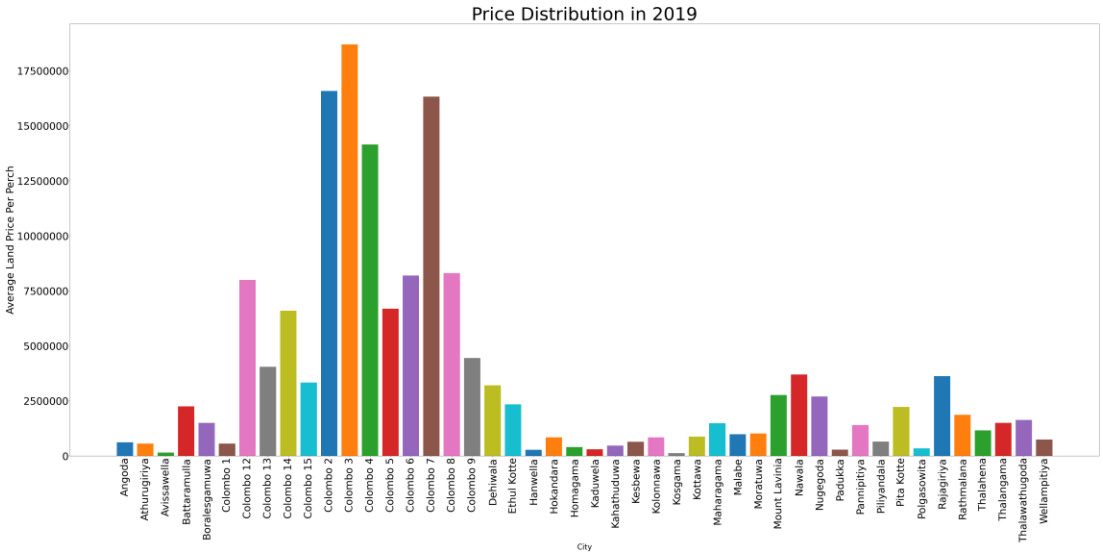


Figure 4.5: Land Price Distribution in 2019

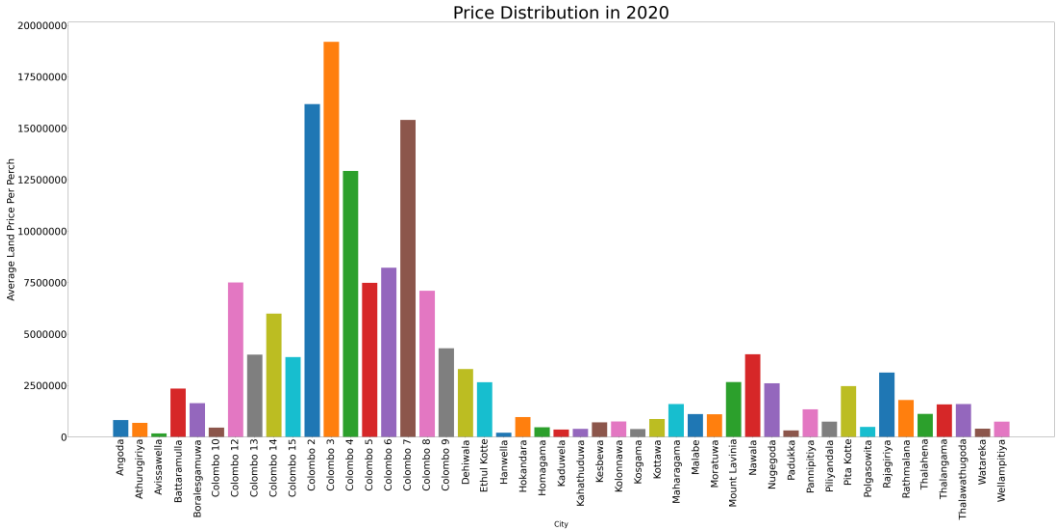


Figure 4.6: Land Price Distribution in 2020

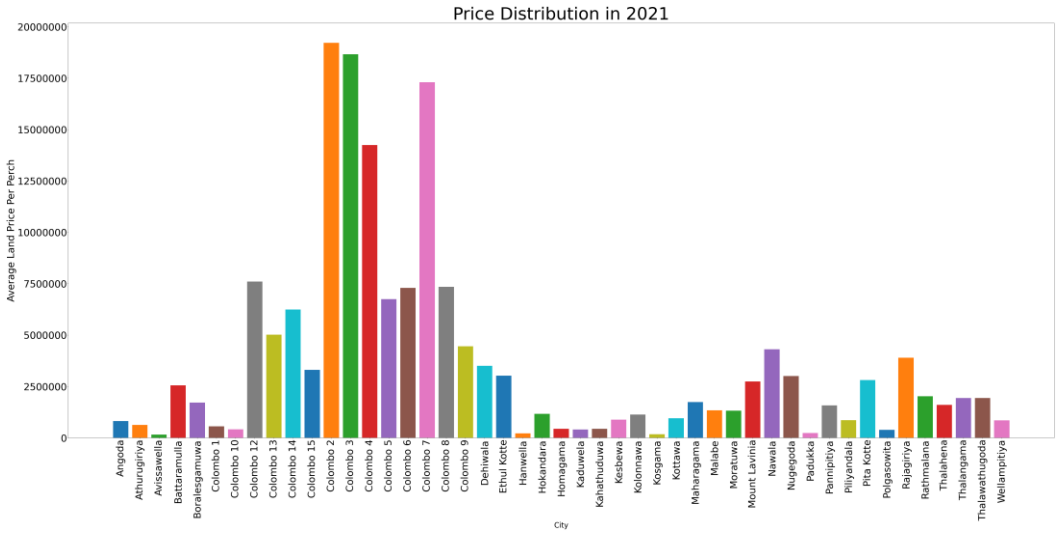


Figure 4.7: Land Price Distribution in 2021

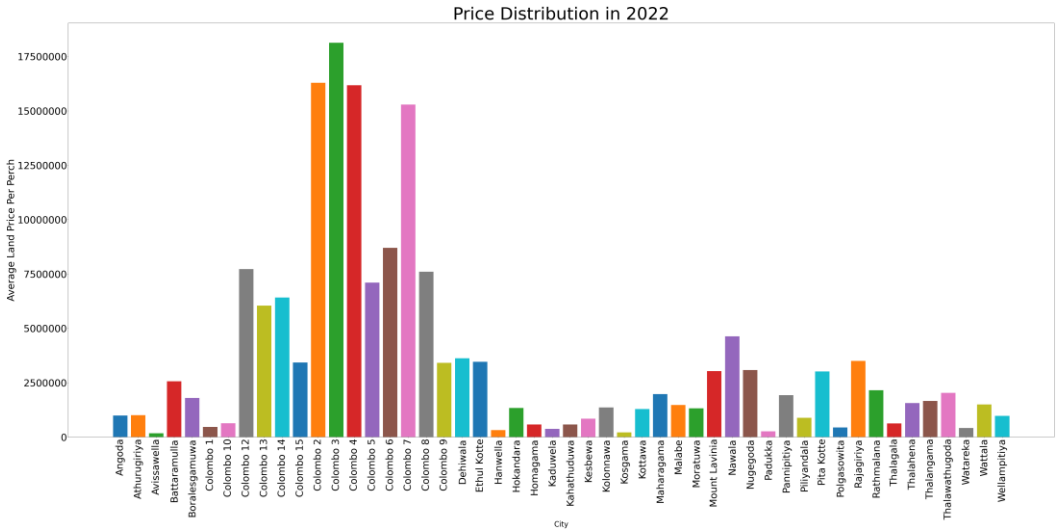


Figure 4.8: Land Price Distribution in 2022

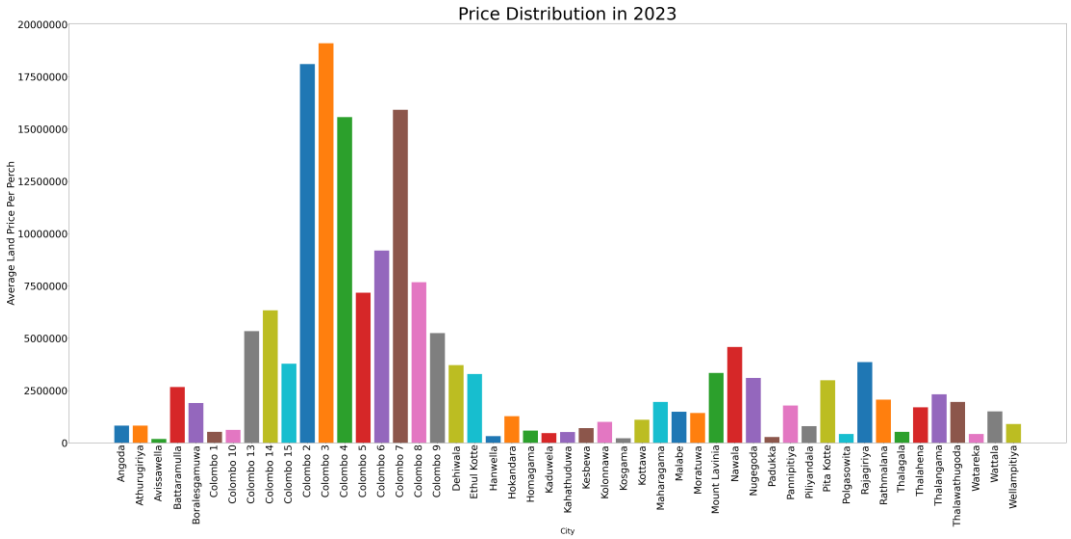


Figure 4.9: Land Price Distribution in 2023

In all the above years from 2018 to 2023, there is a clear indication that Colombo 1-15 areas were highly priced. After that some areas like Nawala, Nugegoda, Rajagiriya, Ethul Kotte and Mount Lavinia had a middle price range. This can be shown in the map of Colombo district like below.

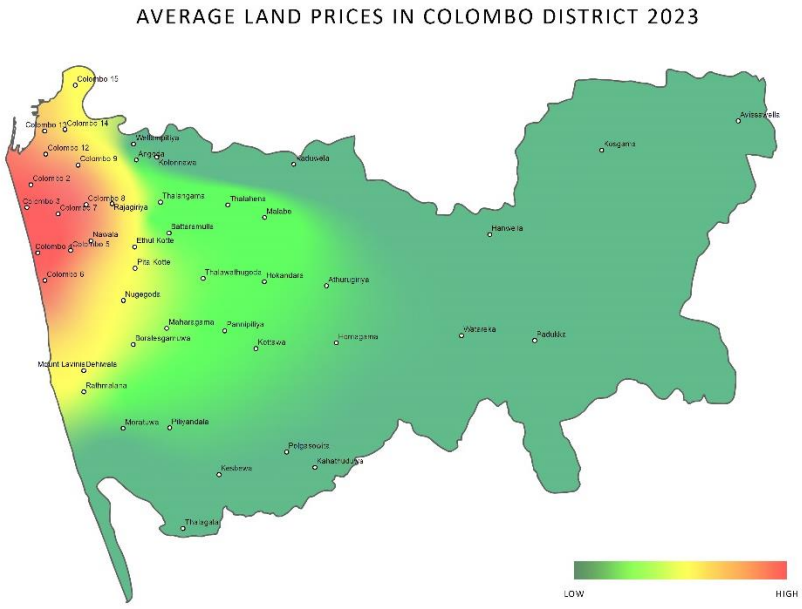
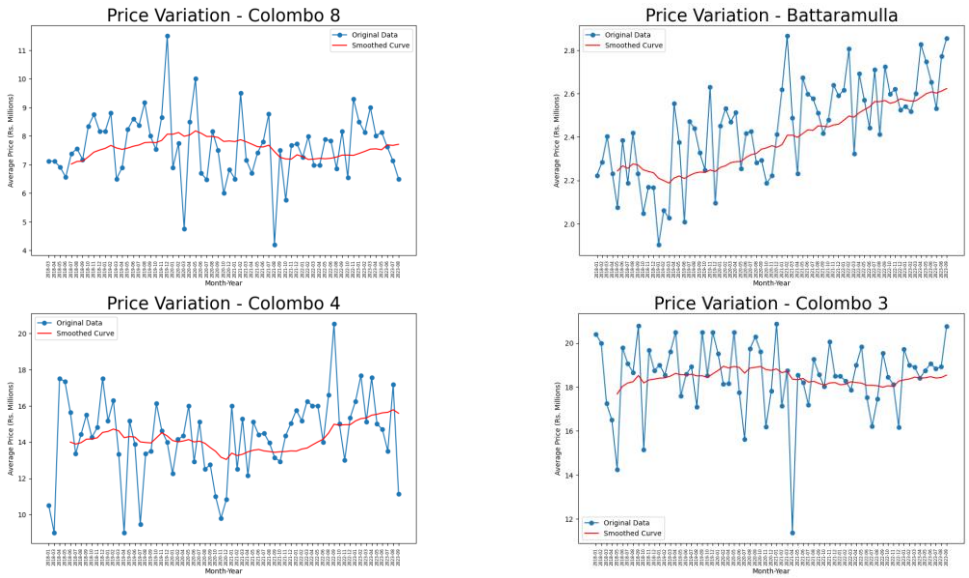


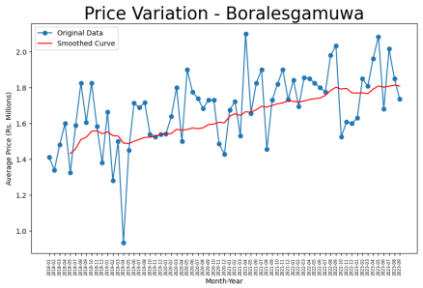
Figure 4.10: Land price variation in Colombo District

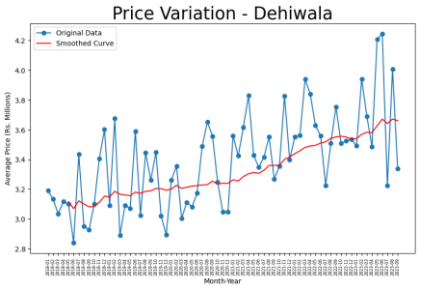
This map clearly indicates that Colombo 1-15 areas has higher land prices. When moving away from the city limits, it decreases gradually.

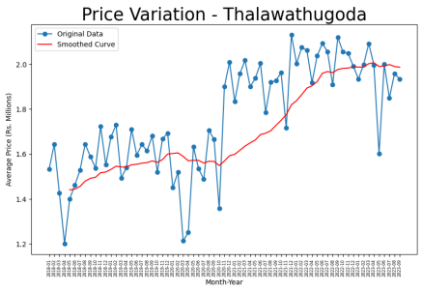
4.2.2. TIME SERIES ANALYSIS

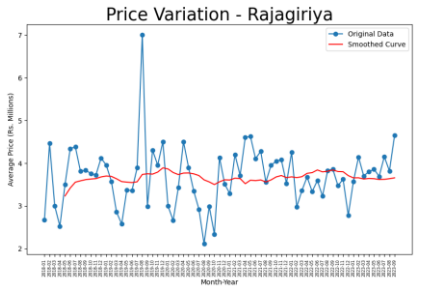
A time series analysis was done to get an idea about to the prices of main cities behave in last 6 years. To do that, Average price of each month grouped by main city were taken and plotted.

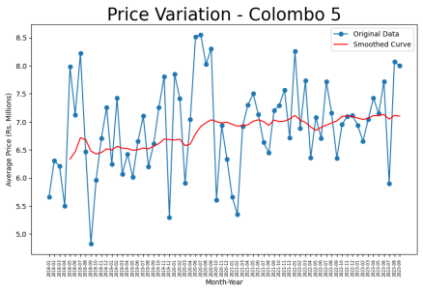


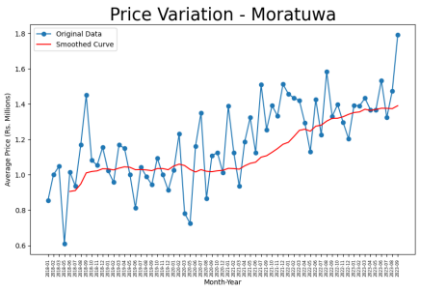

Price Variation - Boralesgamuwa
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

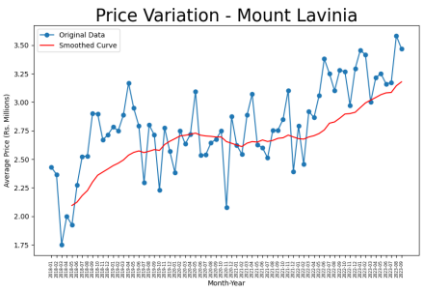

Price Variation - Dehiwala
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

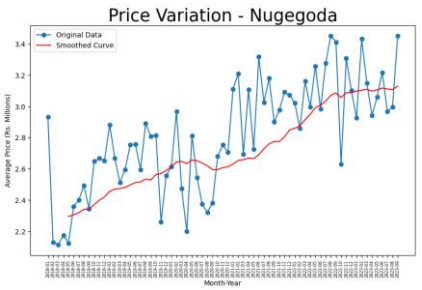

Price Variation - Thalawathugoda
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

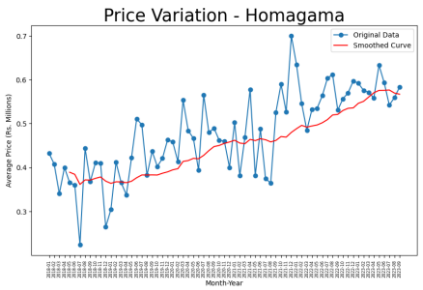

Price Variation - Rajagiriya
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

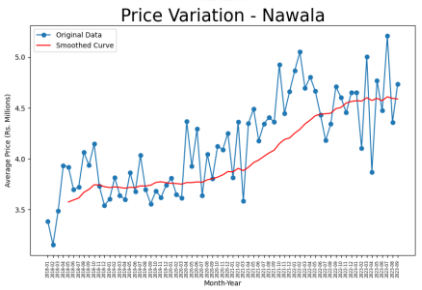

Price Variation - Colombo 5
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

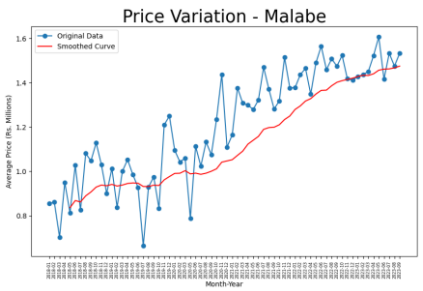

Price Variation - Moratuwa
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

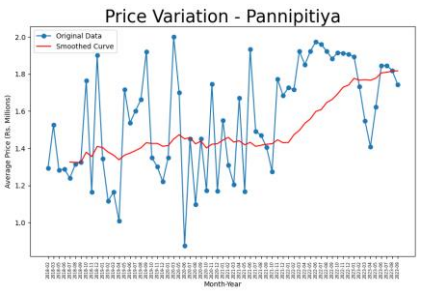

Price Variation - Mount Lavinia
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

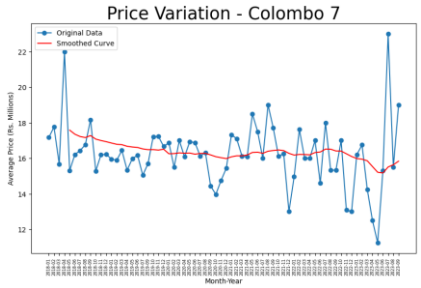

Price Variation - Nugegoda
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

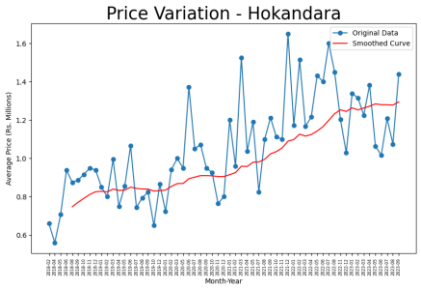

Price Variation - Homagama
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

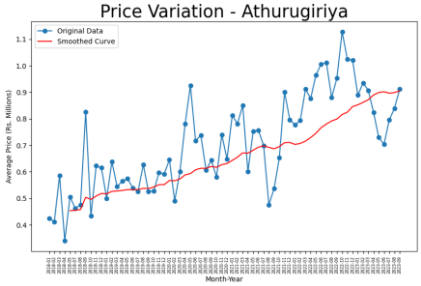

Price Variation - Nawala
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

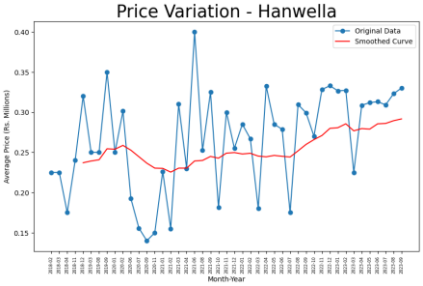

Price Variation - Malabe
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

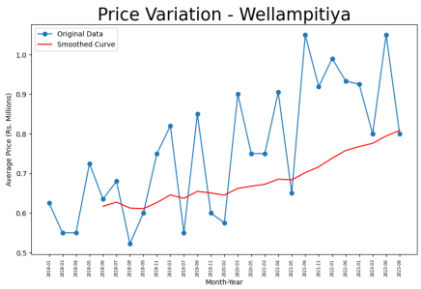

Price Variation - Pannipitiya
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

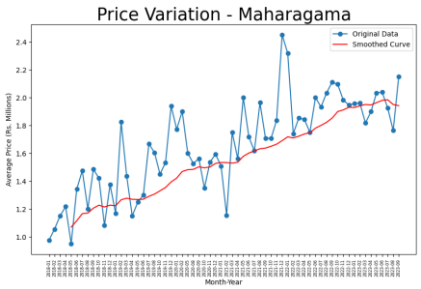

Price Variation - Colombo 7
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

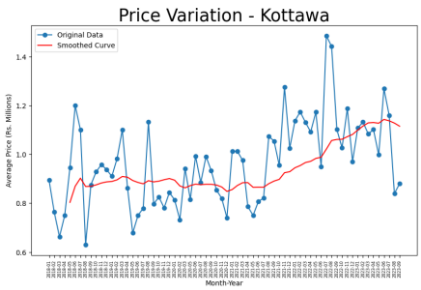

Price Variation - Hokandara
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

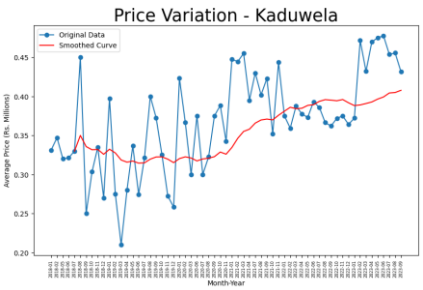

Price Variation - Athurugiriya
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

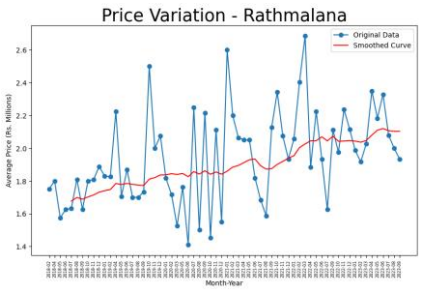

Price Variation - Hanwella
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

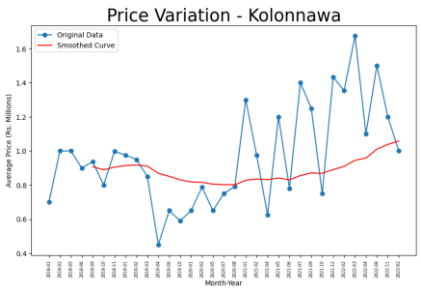

Price Variation - Wellampitiya
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

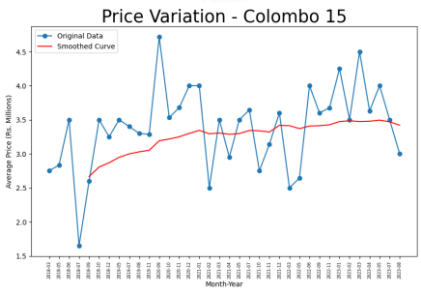

Price Variation - Maharagama
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

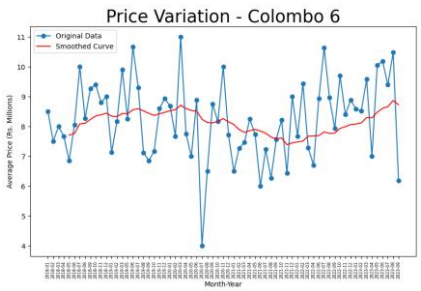

Price Variation - Kottawa
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

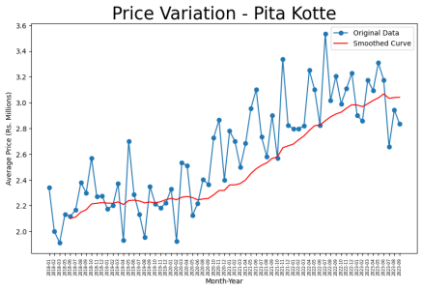

Price Variation - Kaduwela
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

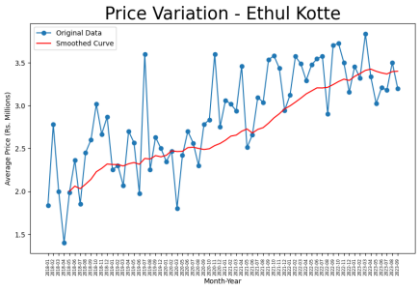

Price Variation - Rathmalana
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

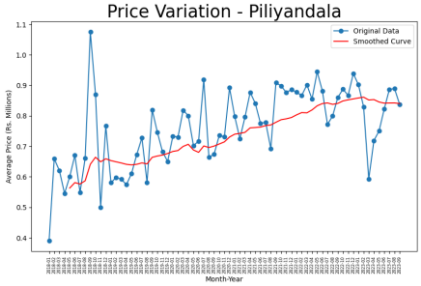

Price Variation - Kolonnawa
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

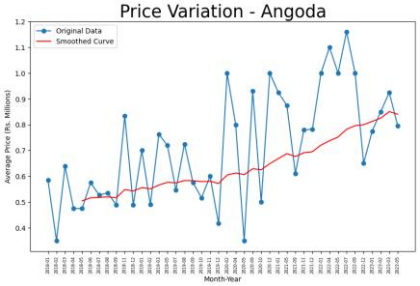

Price Variation - Colombo 15
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

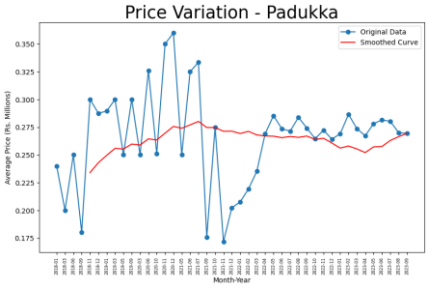

Price Variation - Colombo 6
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

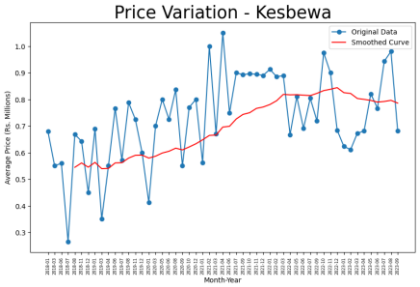

Price Variation - Pita Kotte
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

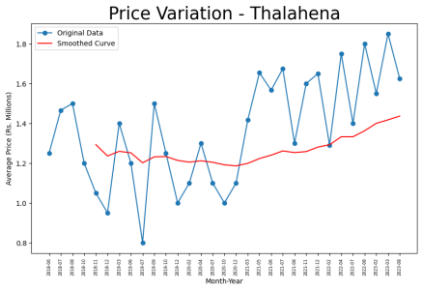

Price Variation - Ethul Kotte
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

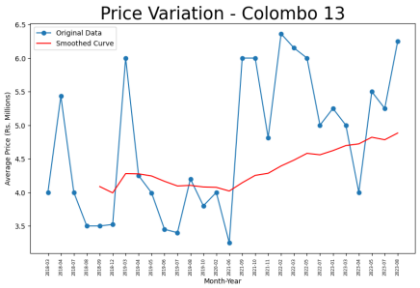

Price Variation - Piliyandala
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

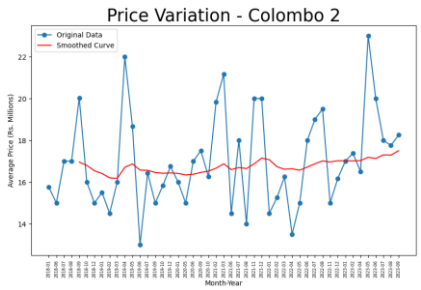

Price Variation - Angoda
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

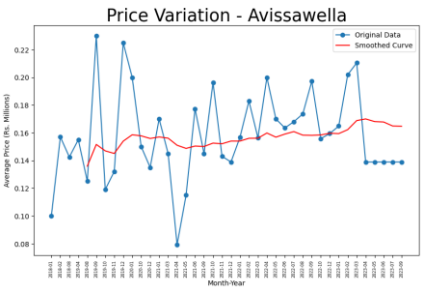

Price Variation - Padukka
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

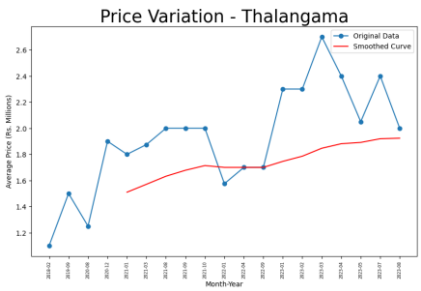

Price Variation - Kesbewa
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

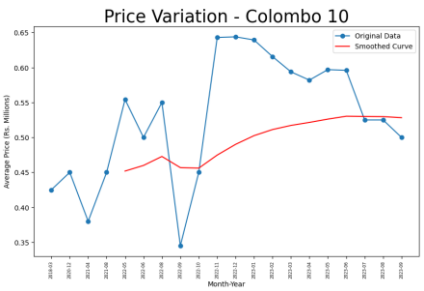

Price Variation - Thalahena
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year


Price Variation - Colombo 13
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year


Price Variation - Colombo 2
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year


Price Variation - Avissawella
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year


Price Variation - Thalangama
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year


Price Variation - Colombo 10
Original Data
Smoothed Curve
Average Price (Rs. Millions)
Month-Year

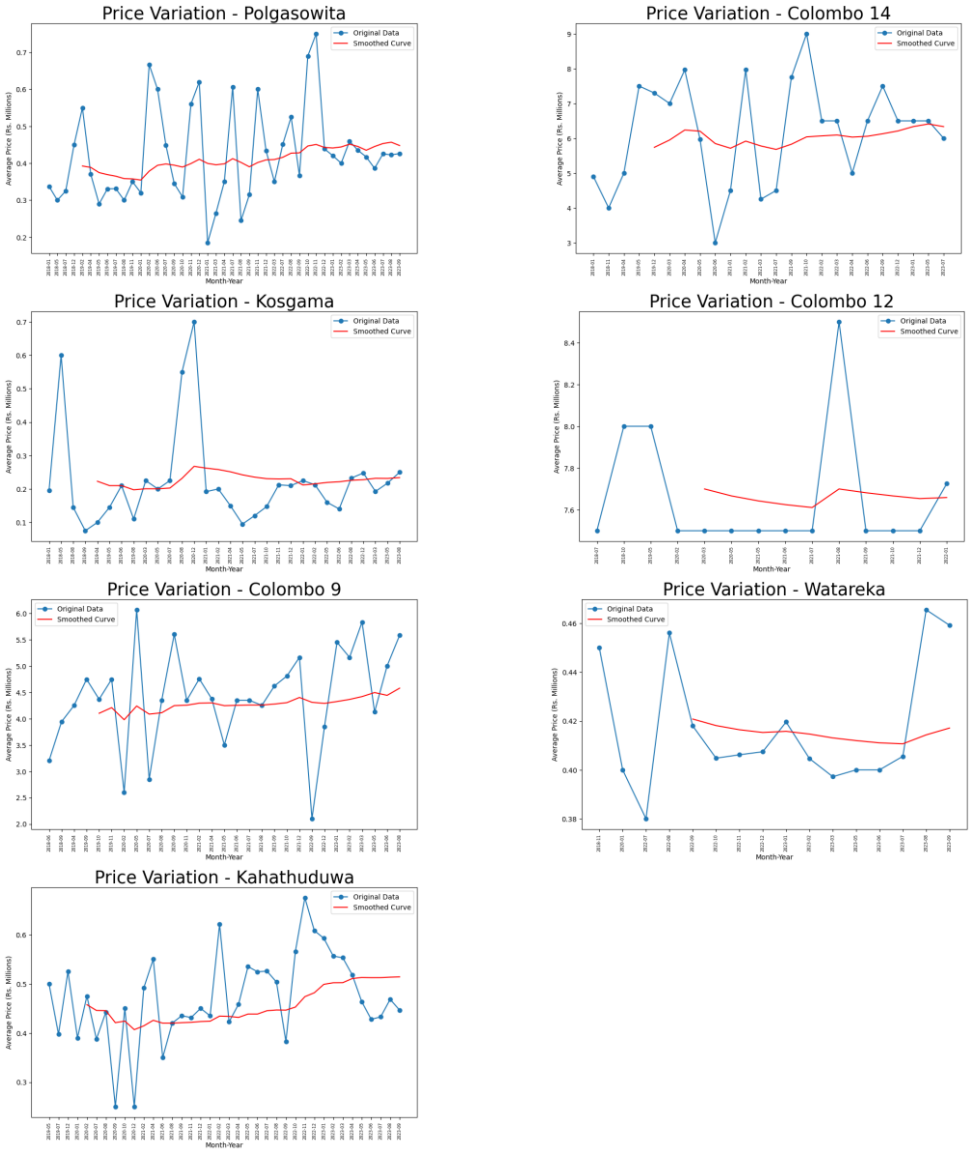


Figure 4.11: Time Series analysis of land price in each city

In the above grid of plots, blue line connects the monthly average of cities and red line is the smoothed curve. There is a clear indication that land prices in most of the cities have increased in the last 6 years. Also, in some cities. There is a price decrement in the period of 2020 - 2021. So, an assumption can be made that this price decrement is because of covid. After 2022, land prices started to rise again in almost all cities.

4.3. SPATIAL NETWORK ANALYSIS

The preprocessed dataset lacked variables necessary for constructing a machine learning model. To generate these variables, a spatial analysis was conducted, using the shape files bought from the Survey Department of Sri Lanka. Those shape files consist of geographical features such as roads, Municipal Council (MC) areas, water

streams, and buildings. Within the original dataset, the actual geographical coordinates of land plots were available.

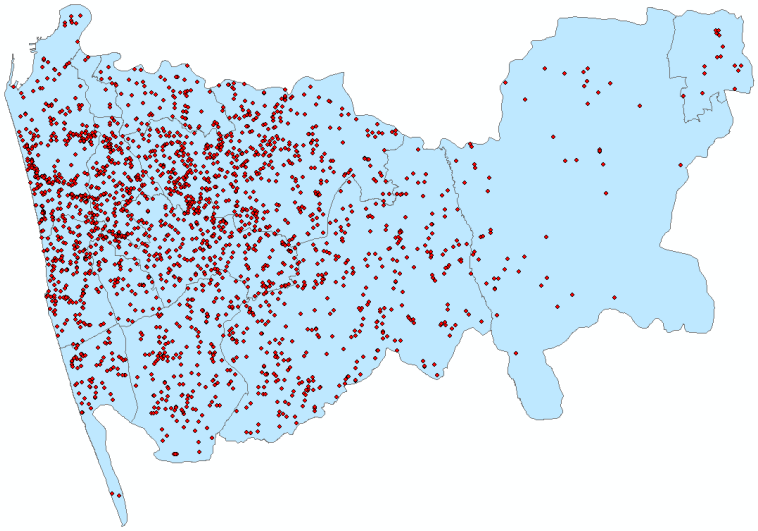


Figure 4.12: Visualization of land plots in Colombo district

These data leads to calculate the nearest distances from the lands to key amenities and infrastructure like nearest hospital, national school, post office, bus stop, main road, railway station, and police station. This spatial analysis, facilitated by the shape files, provided valuable insights into the geographic context of the land plots.

4.4. MODEL BULDING

4.4.1. LAND PRICE PREDICTION

After generating the variables from spatial data, a dataset was finalized for model building. It consists of 8 variables which are add posted date, main city, land price per perch, distance to nearest hospital, distance to nearest school, distance to nearest railway, distance to nearest bus stop distance to nearest main road.

Without add posted date, all the other variables were used to build five machine learning models. There was a slight correlation between those numerical variables.

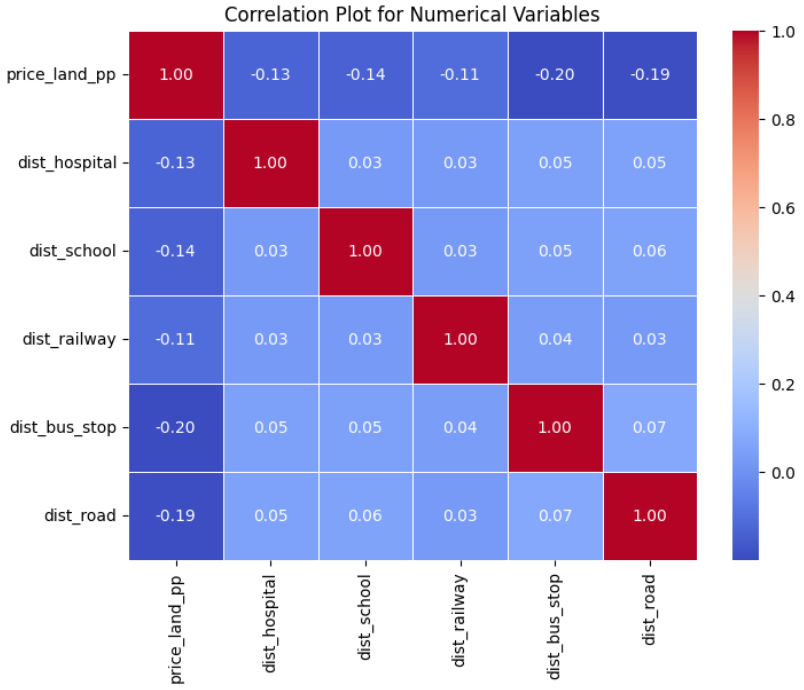


Figure 4.13: Correlation between numerical variables

The variable price_land_pp has a low negative correlation with all the other variables which means that land prices tend to decrease as the distance from those places increases. By taking these variables, RandomForestRegressor(), GradientBoostingRegressor(), LinearRegression(), XGBRegressor() and Artificial Neural network Models were fitted to the dataset.

A linear regression model was fitted to the dataset. It did not give the expected results. It had the training accuracy of 15.62% and testing accuracy of 8.29%.

Random Forest Regressor was fitted with tuned hyperparameters with 5-fold cross validation. The number of estimators were '100' and max depth was 'None'. The training accuracy of 90.26% and a testing accuracy of 58.54% was given in random forest regressor.

Gradient boosting Regressor was fitted with tuned hyperparameters with 5-fold cross validation. Model learning rate was '0.2' and number of estimators were '50'. The training accuracy of 88.48% and a testing accuracy of 60.01% was given in Gradient Boosting regressor.

XG boosting Regressor was fitted with tuned hyperparameters with 5-fold cross validation. Model learning rate was '0.1' and number of estimators were '50'. The

training accuracy of 91.38% and a testing accuracy of 62.23% was given in XG Boosting regressor.

After tuning the hyperparameters with GridSearchCV, a neural network with two hidden layers and an output layer consist with ‘relu’ output function was fitted to the dataset. Sizes of hidden layers are 64 and 32 respectively. The tuned model alpha was 0.001 and maximum iterations was ‘10000’. The training accuracy of 76.23% and a testing accuracy of 48.01% was given in the fitted artificial neural network.

Table 4.3: Summary of Price prediction models

Model Name	Training Accuracy	Testing Accuracy
Linear Regression	15.62%	8.29%
Artificial Neural Network	76.23%	48.01%
Random Forest	90.26%	58.54%
Gradient boost	88.48%	60.01%
XG boost	91.38%	62.23%

4.4.2. TIME SERIES FORECASTING

Time series forecasting model was build using a SARIMAX (Seasonal Auto Regressive Integrated Moving Average with eXogenous factors) model on a dataset. The 'posted_date' column is converted to a datetime format, and additional features such as 'year' and 'quarter' are extracted.

The training data is selected based on a cutoff date ('2023-09-01'), wherein records with 'posted_date' earlier than this date are considered as the training set. The user is then prompted to input a main city as a keyboard input, which filters the dataset and creates a subset specific to the chosen city. The SARIMAX model is then run and fitted to the training data. Model parameters, including the order (p, d, q) and seasonal order (P, D, Q, S), are specified for this purpose. The number of lag observations included in the model is 1 suggests an autoregressive order of 1. The number of times the raw observations are differenced to achieve stationarity. Value = 1 implies first-order differencing. Size of the moving average window order is specified as 1. Number of time steps in a seasonal period indicates a seasonal order of 4, suggesting that the seasonality occurs every 4 quarters.

The model is employed to forecast land prices per perch for the period spanning '2024-01-01' to '2027-03-31'. For visualization, the Matplotlib library is employed, with the resulting predictions using a red dashed line. The end of the historical data is marked with a vertical dashed line on the plot. The following plot is an example for the output of the model.

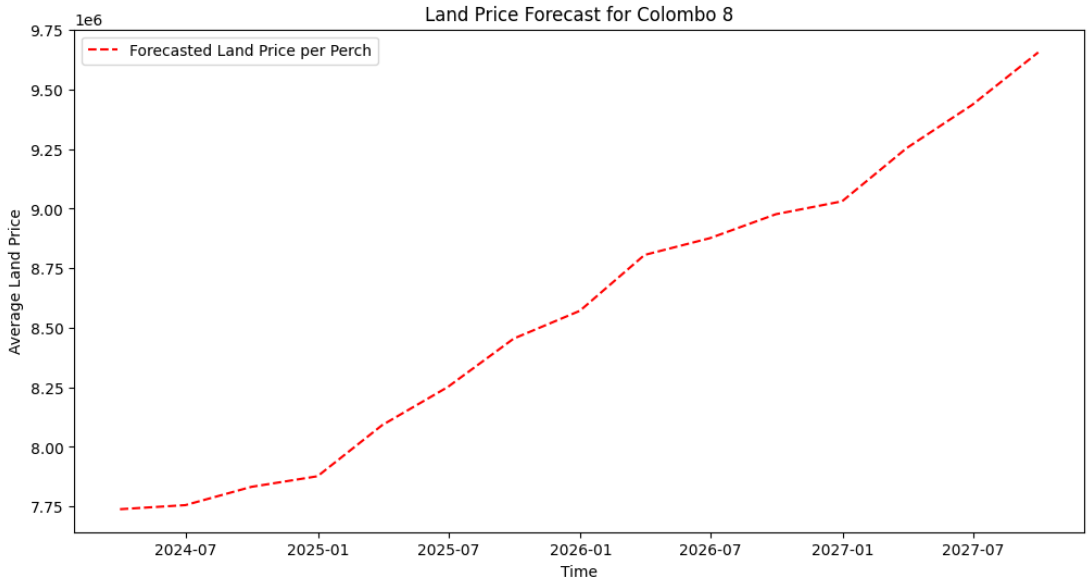


Figure 4.14: Forecasted prices of Colombo 8 (Example for forecasting model)

The resulting plot shows how the prices of Colombo 8 varies in next 3 years. City name was taken as a user input and plot was generated.

Chapter 05:

CONCLUSION

In the examination of land prices within the Colombo district of Sri Lanka, this study aimed to provide a comprehensive understanding of the economic and social landscape through the application of machine learning and geospatial analysis. Customer-submitted data from 'Lanka Property Web' spanning the years 2018 to 2023 was utilized, with a special pre-processing approach that involved outlier removal. The integration of spatial network analysis, shape files from the Survey Department of Sri Lanka, and machine learning techniques resulted in the creation of a user-friendly dashboard for individual land sellers and buyers.

Descriptive statistics of bare land prices in different main cities within the Colombo district revealed significant variations in minimum and maximum prices. Colombo 3 emerged with the highest average price, while Avissawella exhibited a lower average. Colombo 12 demonstrated a high standard deviation, indicating a wider range of property prices, whereas Avissawella had a lower standard deviation, suggesting that urban areas near to Colombo city limits have higher average land prices and standard deviation of price.

Time series analysis spanning the years 2018 to 2023 revealed a general increase in land prices across most cities, with a noticeable decline in 2020-2021, potentially attributed to the COVID-19 pandemic. Spatial network analysis provided valuable insights, including distances to key amenities and infrastructure. This analysis contributed to the generation of variables essential for constructing machine learning models. In the model-building phase, five machine learning models were applied to predict land prices: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XG Boosting Regressor, and Artificial Neural Network. The models were evaluated based on their training and testing accuracies. Among them, XG Boosting Regressor was selected as the best model for this data because of its training Accuracy of 91.38% and testing Accuracy of 62.23%

A time series forecasting model using SARIMAX was developed to predict land prices for the period 2024-2027. In conclusion, the user-friendly dashboard created in this analysis of the combination of machine learning, geospatial analysis, and time series forecasting has improved the understanding of land prices in the Colombo district, aiming investors and landowners in making informed decisions in a rapidly changing property landscape. The ultimate outcome, a user-friendly dashboard, stands as a valuable resource for individuals involved in land transactions, providing them with data-driven insights for well-informed decision-making.

As a continuation of this study, this model can be developed for the all the district in Sri Lanka. Since the original dataset contained only the data from Colombo district and there are very few data entries in other districts, this study had limited only to Colombo. But this can be developed to other districts with their significant variables. For example, altitude is not a significant factor for Colombo, but it will be significant for Kandy district. Apart from that, this study can drill down into another smaller level of location. City is the smallest unit in this study. As a future work, this can be developed up to street level.

REFERENCES

Abidoye, Rotimi, Chan, & Albert. (2017, 09). Artificial neural network in property valuation: application framework and research trend. *Property Management, 35*. doi:10.1108/PM-06-2016-0027

Aziz, Akhir, NorShakirah, Emelia, Jaafar, Izzatdin, . . . Ahmad. (2020, 10). A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems. 11-16. doi:10.1109/ICCI51257.2020.9247843

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32. doi:https://doi.org/10.1023/A:1010933404324

Distance Analysis. (2023). (Esri) Retrieved from Esri ArcGIS: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/distance-analysis.htm

Fattah, Ezzine, J. &., Aman, L. &., Moussami, Z. &., Lachhab, H. &., & Abdeslam. (2018, 10). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management*. doi:10.1177/1847979018808673

Hodson, Timothy, O., Over, Thomas, M., Foks, & Sydney, S. (2021). Mean Squared Error, Deconstructed. *Journal of Advances in Modeling Earth Systems, 13*(12). doi:https://doi.org/10.1029/2021MS002681

ji, Pang, J. a., Zhou, W. a., Han, C. a., Wang, X. a., & Zhe. (2013, 11). A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems, 30*, 129–135. doi:10.1016/j.knosys.2012.01.006

Khan, Qayoom, M. Y., Nizami, A. a., Siddiqu, M. a., M. S., Syed, S. a., & Raazi, K.-U.-R. (2021, 09). Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity*. doi:10.1155/2021/2553199

Liaw, Andy, Wiener, & Matthew. (2001, 11). Classification and Regression by RandomForest. *Forest*.

Otchere, D. A., Ganat, T. O., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering, 208*, 109244. doi:https://doi.org/10.1016/j.petrol.2021.109244

Özer, & Özlem. (2017). Accessibility of Spatial Networks: Using ArcGIS network analyst and space syntax to investigate accessibility to urban facilities.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis. *Deutsches Arzteblatt international, 107*(44), 776–782. doi:https://doi.org/10.3238/arztebl.2010.0776

Segal, & Mark, R. (2004, 04). Machine Learning Benchmarks and Random Forest Regression. *UCSF: Center for Bioinformatics and Molecular Biostatistics*. Retrieved from https://escholarship.org/uc/item/35x3v9t4

Tresidder, M. (2005). Using GIS to Measure Connectivity: An Exploration of Issues. *Field Area Paper*, 14-16.

Velumani, P., Nampoothiri, N., & Kavithra, P. (2019, 12). Predicting the Land Value using Regression Techniques and Artificial Neural Network. *International Journal of Engineering and Advanced Technology, 9*(1S4). doi:10.35940/ijeat.A1021.1291S419

APPENDIX

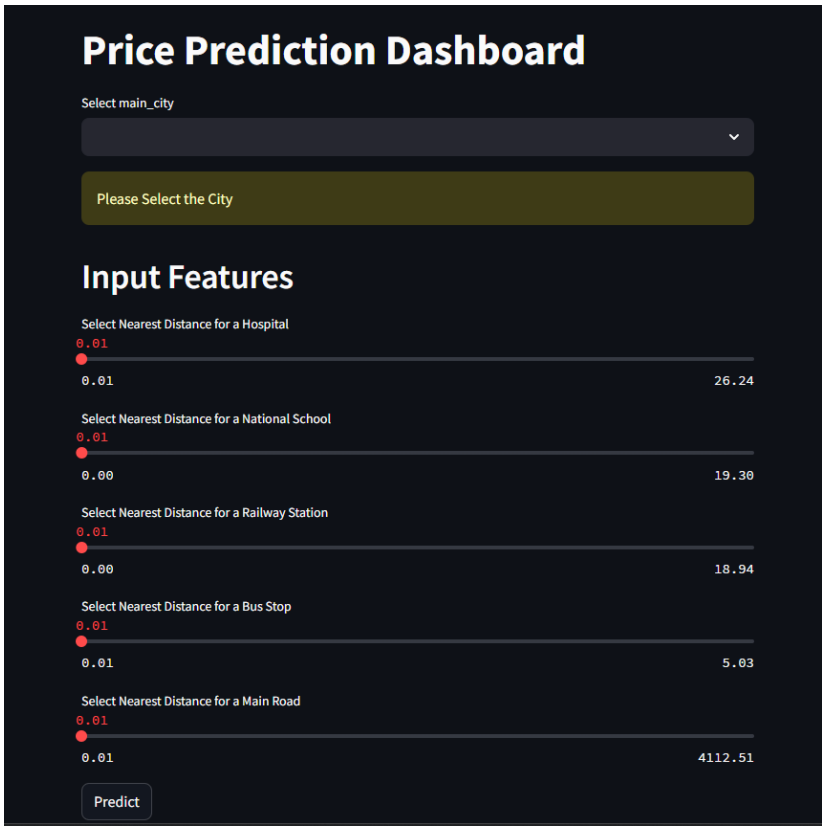
Head of the dataset used for modeling.

	A	B	C	D	E	F	G	H
1	posted_date	main_city	price_land_pp	dist_hospital	dist_school	dist_railway	dist_bus_stop	dist_road
2	1/2/2018	rajagiriya	1600000	5.300061226	8.213339348	4.10277854	1.420663172	1958.924652
3	1/13/2018	dehiwala	3000000	5.070247524	10.2529478	11.01605486	2.915772445	1498.262124
4	1/12/2018	homagama	285000	2.349761667	4.48532974	0.574572606	2.29851817	1863.774927
5	2/1/2018	malabe	1000000	1.435860704	4.531727032	10.16508972	3.153121111	1572.588438
6	1/24/2018	nugegoda	3250000	16.41869422	0.3429556	0.620786259	1.956450553	1170.203885
7	3/14/2018	wellampitiya	550000	9.59095031	9.674941655	2.006051075	1.884921789	1300.996911
8	2/7/2018	kottawa	650000	8.194594196	0.25013862	8.351094588	1.877165127	1051.927374
9	2/4/2018	homagama	360000	2.7294161	2.433243878	3.018347969	3.555231759	1143.689219
10	4/2/2018	thalawathugoda	1100000	7.826282961	7.520937987	9.380591943	1.798284783	2265.620557
11	3/11/2018	malabe	280000	6.251171877	7.239862448	5.100306427	1.977907403	1545.843457
12	3/4/2018	maharagama	1350000	16.89629865	5.563518016	0.863069035	2.267455354	1778.709386
13	5/15/2018	pannipitiya	1000000	11.74330808	4.60707077	5.203904719	2.179024052	1506.312884
14	3/2/2018	homagama	120689	5.250845798	7.171368474	8.838385136	1.731663845	2342.753953
15	5/16/2018	colombo 6	6700000	0.01359	9.192583229	8.040015353	2.655377375	874.8093985
16	5/15/2018	pannipitiya	1000000	8.545144693	6.036429433	3.578711805	2.421524283	3243.072631
17	4/3/2018	battaramulla	2300000	7.362596845	9.41577087	5.988651022	0.921395713	1993.79243

Github Repository with codes and Datasets

https://github.com/Dheelaka/Forecasting_Land_Prices_in_Colombo_District.git

Dashboard



Link - https://landpriceforecastingapp-sbfssfrle7epsae42m7xxo.streamlit.app/