

turnitin

22085617_Shamel_Rai.pdf



Document Details

Submission ID

trn:oid::29984:79761958

Submission Date

Jan 21, 2025, 11:22 PM GMT+8

Download Date

Jan 21, 2025, 11:23 PM GMT+8

File Name

22085617_Shamel_Rai.pdf

File Size

1.9 MB

40 Pages

6,301 Words

39,781 Characters





7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

Match Groups

-  **33 Not Cited or Quoted 7%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 3%  Internet sources
- 1%  Publications
- 6%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 33 Not Cited or Quoted 7%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 3% Internet sources
- 1% Publications
- 6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	islingtoncollege on 2024-12-24	1%
2	Submitted works	islingtoncollege on 2024-12-24	<1%
3	Internet	www.mdpi.com	<1%
4	Submitted works	University College London on 2023-08-06	<1%
5	Submitted works	Coventry University on 2023-04-11	<1%
6	Submitted works	University of Hertfordshire on 2024-08-28	<1%
7	Submitted works	University of Southampton on 2023-09-29	<1%
8	Internet	etd.repository.ugm.ac.id	<1%
9	Submitted works	islingtoncollege on 2025-01-16	<1%
10	Submitted works	National Institute of Development Administration on 2015-07-06	<1%

11	Submitted works	Icon College of Technology and Management on 2023-06-12	<1%
12	Submitted works	Colorado School of Mines on 2013-02-11	<1%
13	Publication	Jagadesh Kumar Jatavallabhula, Tshepo Gaonnwe, Sibusiso Nginda, Vasudeva Ra...	<1%
14	Submitted works	University of Hertfordshire on 2025-01-06	<1%
15	Submitted works	University of South Australia on 2020-06-13	<1%
16	Submitted works	islingtoncollege on 2025-01-16	<1%
17	Internet	www.researchgate.net	<1%
18	Submitted works	91336 on 2015-04-06	<1%
19	Submitted works	Australian National University on 2021-06-01	<1%
20	Submitted works	BPP College of Professional Studies Limited on 2024-09-25	<1%
21	Submitted works	University of Liverpool on 2024-09-16	<1%
22	Submitted works	islingtoncollege on 2024-12-22	<1%
23	Submitted works	islingtoncollege on 2024-12-24	<1%
24	Internet	www.coursehero.com	<1%

25	Submitted works	Liverpool John Moores University on 2023-11-21	<1%
26	Submitted works	islingtoncollege on 2025-01-16	<1%
27	Internet	www.open.edu	<1%
28	Submitted works	Liverpool John Moores University on 2023-04-19	<1%
29	Submitted works	islingtoncollege on 2024-12-24	<1%



Islington college
(इस्लिंग्टन कलेज)

Module Code & Module Title

Level 6 – Artificial Intelligence

Assessment Type

Semester

2024/25 Autumn

Student Name: Shamel Rai

London Met ID: 22085617

College ID: np01cp4s230135@islingtoncollege.edu.np

Assignment Due Date: Thursday, January 9, 2025

Assignment Submission Date: Friday, January 10, 2025

Submitted To: Alish Kc

Word Count (Where Required): 6167

I confirm that I understand my coursework needs to be submitted online via MST Classroom under the relevant module page before the deadline in order for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.

22085617_Shamel_Rai 3.docx

Islington College, Nepal

Document Details

Submission ID

trn:oid::3618:78536324

Submission Date

Jan 9, 2025, 11:37 AM GMT+5:45

Download Date

Jan 9, 2025, 11:38 AM GMT+5:45

File Name

22085617_Shamel_Rai 3.docx

File Size

33.9 KB

34 Pages

5,329 Words

29,172 Characters

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

- 20 Not Cited or Quoted 3%**
Matches with neither in-text citation nor quotation marks
- 3 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 1% Internet sources
- 2% Publications
- 1% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag

Table of Contents

Table of Figure.....	3
1. Introduction.....	1
1.1. Problem.....	1
1.2. Significance of the problem.....	2
1.3. Scope of Study.....	2
1.4. Machine Learning and Supervised Learning.....	2
1.5. Algorithm.....	3
1.6. Motivation.....	4
1.7. Aims And Objectives.....	4
2. Background.....	6
2.1. Related Work Review.....	6
2.1.1. System Architecture/ Block Diagram.....	10
2.2. Algorithm.....	11
2.3. Evaluation Metrics.....	14
2.4. Comparison to previous studies.....	14
2.5. Dataset Description.....	15
2.5.1. Key Features.....	16
2.6. Data source.....	16
2.6.1. Initial Observation and Data Preprocessing.....	17
2.7. Exploratory Data Analysis (EDA).....	17
3. Proposed Solution.....	19
3.1. Introduction.....	19
3.2. Data Preprocessing.....	19
3.3. Features Selection.....	19
3.4. Model Implementation.....	20
3.4.1. Linear Regression.....	20
3.4.2. Random Forest.....	20
3.4.3. XGBoost.....	21
3.5. Justification for algorithm choice.....	22
3.6. Pseudocode for algorithm.....	22
3.6.1. Linear regression.....	22

3.6.2.	Random Forest	23
3.6.3.	XGBoost	23
3.7.	Flowchart	24
3.8.	Diagrammatic Representation.....	25
3.8.1.	Data Flow Diagram.....	25
3.8.2.	Algorithm Workflow	26
3.9.	Data Pre-Processing	27
3.10.	Detail process	28
4.	Conclusion	34
	References	35

 19

 24

Table of Figure

Figure 1:	Research paper on Linear Regression	7
Figure 2:	Research paper random forest	8
Figure 3:	Research paper on XGBoost	9
Figure 4:	System Architecture Diagram.	10
Figure 5:	Linear Regression.....	11
Figure 6:	Random Forest	12
Figure 7:	XGBoost.....	13
Figure 8:	Dataset of the real estate property	15
Figure 9:	Linear Regression Formula	20
Figure 10:	Loss function Formula.....	21
Figure 11:	Regularization term	21
Figure 12:	Flowchart Diagram.....	24
Figure 13:	Data Flow Diagram	25
Figure 14:	Algorithm Workflow	26
Figure 15:	Importing all the modules and loading the data set.	28
Figure 16:	Dropping Unnecessary column and handling missing values	29
Figure 17:	Encoding Categorical Variables	29
Figure 18:	Data Visualization	30
Figure 19:	Correlation Matrix	31
Figure 20:	Training Feature-target split	31
Figure 21:	Training the Test Split and Standardizing the numeric features	32
Figure 22:	Training the model	32
Figure 23:	Displaying the result.....	33

1. Introduction

1.1. Problem

It is very essential determining the prices of property in the real estate market because this will very directly influence the individual and the business. It will give precise guidelines to the investors, houseowners, and the developers of real estates how their investments should be put under proper pricing strategy and good forecasting. This may come into great complex prediction to trace a given right-size value of property as this size is brought in the larger the increase becomes. It is lead into several factors such as the location, features in properties, and economic conditions of the market conditions and other economic indicator.

Such factors as proximity to urban regions, availability of educational institutions and health facilities, crime rates, and the general level of development in the neighborhood can significantly contribute to the value of a property (Kamba, 2021). Under such conditions, it could be quite a difficult task to forecast the property price since so many variables have to be accounted for, most of which would also be interrelated to each other, complicating the assessment of both market value and market trends.

This demand of forecasting is very important for the real estate agents' developers and for the investors, who use these predictions in order to invest their money in a correct and accurate manner, which will try to make the life of the buyer of the real estate market much easier. The methodology of real estate price prediction has gone through a sea change, moving from the conventional to the adoption of modern data driven machine learning methodologies. This involves using large dataset where historical data is analyzed to identified patterns that informed predictions. This strategy is often expected to make correct predictions but raises the difficulty of choosing the right algorithm for generalization when faced with new or unclear data, and it is sometimes hard to cope with the lacking details.

1.2. Significance of the problem

It is very important to solve the problem of property price prediction since it will benefit not only the investor but many other stakeholders. For example, the real estate investor will predict which properties are going to increase their price and which ones are likely to help him maximize his portfolio status by optimizing the returns on that appreciation. Due to this information, all the stakeholders connected will be in a position to make proper well-informed decisions regarding the decision of when to sell and when to buy within the trends that exist in the market so that all the parties connected can decrease the risk involved. This can help the government and other policymakers make better decisions while making policies concerning urban developments, investments, housing, and other infrastructure. (Topraklı, 2024) From the view of technology, machine learning is considered a game changer for practically all the industries, while from the prediction point of view, there is nothing different about the real estate industry.

1.3. Scope of Study

This coursework is based upon the predicting the property prices along with demand of the real estate market using machine learning algorithms. For this course, three machine learning algorithms which are going to be implemented and applied, and then from the outcome, their monitoring of the performance will reflect in getting knowledge regarding the effectiveness of the properties predicted by their price. All kinds of data, such as location, bedroom, parking, etc., are going to be applied for this course from the given CSV file.

The project will be based on these three algorithms in the context of real estate price prediction. Although unsupervised learning, deep learning, and many other learning techniques may show good potential aspects, the core focus of this project is to properly interpret machine learning models used widely for such tasks.

1.4. Machine Learning and Supervised Learning

Machine learning is also called a type of Artificial Intelligence (AI) that provides the ability to the machine for learning from an abundance of data and enhancing its performance just like a human does without programming them in obvious ways (ÇELİK, 2018). The machine learning task is very broad for property price prediction. The real estate industry, in general,

provides an outstanding context with which one can engage with large amounts of historical data, including a complex set of factors influencing the prices of the properties. Prediction of real estate prices can be considered within the supervised learning problem

In supervised learning, the algorithm learns from labeled data in which inputs features are associated with the correct outputs labels (Emma Oye, 2024). The technique is how to learn a mapping of inputs to target outputs, so in case some new inputs come to be given to model, this will provide a proper prediction corresponding to the new inputs, which can take place from the prediction of the real estate prices since these are the continuous learnings. Therefore, upon using historic information to develop the models for predicting a variable that remains continuous, there occurs some number of features.

It adopted a supervised learning paradigm and regression model types; those could apply on output continuous variable like in-house pricing models. All three models in application, namely linear regression, random forest, and XGBoost, are popular choices algorithms proven known for each differences in accuracy and complexity perspectives against a background of the regression task

1.5. Algorithm

In this coursework, the most appropriate algorithm for the property price prediction are linear regression, random forest and XGBoost.

1. Linear regression:

It is a very intuitive algorithm interplaying between the target variable in this case being the price of a property and input characteristics captured using a linear equation. One of the basic underlying assumptions upon which linear regression rests is that of linearity of features with the target variable, hence the method is very easy to understand and computationally efficient. This is basically the baseline model against which the performances of other complex algorithms can be compared. (Gupta, 2020)

2. Random Forest:

Random forest is probably the most frequently used ensemble learning technique. This is the type of decision tree that features multiple decision trees. Aggregating

their predictions improves on predictive accuracy. That means it can capture non-linear relationships within the data, and therefore alleviate overfitting. Due to that, it becomes a really good choice for such complicated datasets as real estate data. (Smith, 2021)

3. XGBoost:

Gradient boosting is one of the interesting algorithms which build decision trees sequentially, one after another and improve errors by the previously built tree.

XGBoost performs well in outperforming most other algorithms at prediction since it manages complex relationships, missing data, and outliers pretty well but is computationally efficient. Each of the algorithms comes with its bags of strengths and weaknesses. This forms the crux of this study: how each of the above will work in property price prediction. (Lee, 2022)

1.6. Motivation

This has led to the growing importance of property price prediction in real estate due to urbanization changes and rapidly changing economies. A forecast of property prices becomes of utmost consequence not only to the buyer and seller but also to policy makers and financial institutions. In several regions where no robust methods of prediction prevail, it leads to a situation in which stakeholders suffer significant challenges that are largely remonstrative of financial instability and poor decisions.

Machine learning is one way of solving the problem—by which the right means could be offered in building a model that will predict property prices more accurately considering different dimensions. The project is an effort towards the same to resolve these existing problems and contribute in building a decision support tool which would be robust in nature.

1.7. Aims And Objectives

Design and implement a machine-learning-based solution to accurate prediction of real estate prices that would empower stakeholders with reliable data-driven insights.

- To analyze the best existing methods for predicting real estate prices through machine learning

- To develop a rich dataset capturing relevant features with respect to the price of a property
- To develop and evaluate predictive models based on linear regression, random forests, and XGBoost for determining the most effective algorithm
- To justify and stay relevant by using standard performance evaluation metrics like R^2 , MSE, and RMSE

2. Background

2.1. Related Work Review

Economic, Social and Environmental changes have the real estate sector and therefore, this property price forecasting is of high importance not only to the buyers and sellers but also to policy makers and financial institutions (Ren, 2022). Questions have brought previous techniques of traditional methods of valuation, largely based on heuristics or limited statistical models, into present-day data-driven approaches that can handle big datasets and complicated relationships suitably. In fact, it helps the buyer in making a well-informed decision, compels the seller to quote sensible prices for property, and gives better opportunities to investors. On the other hand, urban planners and policy makers can take advantage of knowledge from the same to design better housing policies.

The most direct applications of machine-learning techniques apply to the prediction of price. Real estate, finance, and urban planning present these applications to give buyers, sellers, and policy makers appropriate price predictions of properties with an indication of understanding into market trends and valuation. Other studies realized with previous studies have also been carried out on the same issue with the help of a number of machine learning techniques, e.g., Linear Regression, Random Forest, and XGBoost. The following are some associated studies of the relevance of these algorithms to property prices.

a. Linear Regression in Property Price Prediction:

Linear Regression is one of the most apparently simple and widely used baseline models that predict the price of any property. In 2020, Gupta et al. carried out a study where the researchers trained such a model on a dataset of features that include square footage, location, and number of rooms; it gives a supporting R-squared value of 0.72. An important point is that the model will not capture nonlinear interactions. (Gupta, 2020)

Machine Learning based Predicting House Prices using Regression Techniques

Manasa J

Department Of Mathematics
Dayananda College Of Engineering-RC
Visveshwariah Technological University
Bengaluru, Country
manasa.chandan@gmail.com

Radha Gupta

Professor and Head,
Department Of Mathematics
Dayananda College Of Engineering
Bengaluru, Country
radha.gaurav.gupta@gmail.com

Narahari N S

Professor
Department Of IEM
RV College Of Engineering
Bengaluru, Country
naraharins.rvce.edu.in

Abstract— Predictive models for determining the sale price of houses in cities like Bengaluru is still remaining as more challenging and tricky task. The sale price of properties in cities like Bengaluru depends on a number of interdependent factors. Key factors that might affect the price include area of the property, location of the property and its amenities. In this research work, an analytical study has been carried out by considering the data set that remains open to the public by illustrating the available housing properties in machine hackathon platform. The data set has nine features. In this study, an attempt has been made to construct a predictive model for evaluating the price based on the factors that affect the price. Modeling explorations apply some regression techniques such as multiple linear regression (Least Squares), Lasso and Ridge regression models, support vector regression, and boosting algorithms such as Extreme Gradient Boost Regression (XG Boost). Such models are used to build a predictive model, and to pick the best performing model by performing a comparative analysis on the predictive errors obtained between these models. Here, the attempt is to construct a predictive model for evaluating the price based on factors that affects the price.

Keywords—house price, lasso regression, ridge regression, regression methods

I. INTRODUCTION

Modeling uses machine learning algorithms, where machine learns from the data and uses them to predict a new data. The most frequently used model for predictive analysis is regression. As we know, the proposed model for accurately predicting future outcomes has applications in economics, business, banking sector, healthcare industry, e-commerce, entertainment sports etc. One such method used to forecast

importantly the house price. Multiple linear regression is one of the statistical techniques for assessing the relationship between the (dependent) target variable and several independent variables. Regression techniques are widely used to build a model based on several factors to predict price. In this study, we have made an attempt to build house price prediction regression model for data set that remains accessible to the public in Machine hackathon platform. We have considered five prediction models, they are ordinary least squares model, Lasso and Ridge regression models, SVR model, and XGBoost regression model. A comparative study was carried out with evaluation metrics as well. Once we get a good fit, we can use the model to forecast monetary value of that particular housing property in Bengaluru. The paper is divided into the following sections: Section 2 addresses previous related work, Section 3 explains the description of the data set used, pre-processing of data and exploratory analysis of data before regression model is built. Section 4 presents a summary of the regression models developed in the comparison study and the evaluation metrics is used. Section 5 sums up the models and concludes with the future scope of the proposed work. Section 6 lists the applicability of the model.

II. PREVIOUS RELATED WORK

Pow, Nissan, Emil Janulewicz, and L. Liu [11] used four regression techniques namely Linear Regression, Support Vector Machine, K-Nearest Neighbors (KNN) and Random Forest Regression and an ensemble approach by combining KNN and Random Forest Technique for predicting the property's price value. The ensemble approach predicted the prices with least error of 0.0985 and applying PCA didn't

Figure 1: Research paper on Linear Regression

b. Best Performance by Random Forest:

It is an ensemble learning method famous for being robust against non-linear relationships and feature interaction. A case in point is one Research Paper done by Smith and Zhang, 2021. The authors, in particular, had more than 15,000 records related to real estate and applied the Random Forest algorithm to them. The metrics of this model was more accurate than linear regression which helps it consider to be one of the better solutions of predicting the price of the property. (Smith, 2021)



Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 806–813



The 8th International Conference on Information Technology and Quantitative Management
(ITQM 2020 & 2021)

House Price Prediction using Random Forest Machine Learning Technique

Abigail Bola Adetunji^a, Oluwatobi Noah Akande^{*b}, Funmilola Alaba Ajala^a, Ololade Oyewo^a, Yetunde Faith Akande^c, Gbenle Oluwadara^b

^aDepartment of Computer Science, Faculty of Computing and Informatics, Ladoke Akintola University of Technology, Nigeria

^bComputer Science Department, College of Pure and Applied Sciences, Landmark University, Nigeria

^cAccounting Department, College of Business Sciences, Landmark University, Nigeria

Abstract

Predicting a price variance rather than a specific value is more realistic and attractive in many real-world applications. Price prediction can be thought of as a classification issue in this situation. However, the House Price Index (HPI) is a common tool for estimating the inconsistencies of house prices. Since housing prices are closely correlated with other factors such as location, city, and population, predicting individual housing prices needs information other than HPI. The HPI is a repeat-sale index that tracks average price shifts in repeat transactions or refinancing of the same assets. Therefore, HPI is ineffective at predicting the price of a single house because it is a rough predictor based on all transactions. This study explores the use of Random Forest machine learning technique for house price prediction. UCI Machine learning repository Boston housing dataset with 506 entries and 14 features were used to evaluate the performance of the proposed prediction model. A comparison of the predicted and actual prices predicted revealed that the model had an acceptable predicted value when compared to the actual values with an error margin of ± 5 .

© 2021 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0>)

Peer-review under responsibility of the scientific committee of the The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

Keywords: Sales forecasting; House Price Prediction; Machine Learning; Random Forest Algorithm

1. Introduction

Housing is one of the integral components that can be used to measure how successful the economy of a nation is. As the economy increases, people tend to migrate from the urban to rural areas which results to an increase in

Figure 2: Research paper random forest

c. More advanced predictions from XGBoost:

This is a new framework in gradient boosting, created not so long ago. It has become extremely popular in prediction modeling due to high efficiency and holding the most accurate models. Input factors were used with the trend of historical prices. The model gives an RMSE of \$8000 accuracy standing as the best result when other models in its genre are compared. The most significant highlight identified from the study presented shows how XGBoost handled missing values and features regularization for preventing overfitting. (Lee, 2022)




Real Estate Price Prediction

Rabia Naz^{1*}, Bushra Jamil², Humaira Ijaz²
¹Department of Software Engineering, University of Sargodha, Sargodha, Pakistan.
²Department of IT, University of Sargodha, Sargodha, Pakistan.
***Correspondence:** rabianaz935@gmail.com

Citation | Naz. R, Jamil. B, Ijaz. H, "Real Estate Price Prediction", IJIST, Vol. 6 Issue. 2 pp 1031-1044, July 2024
Received | July 05, 2024 **Revised |** July 23, 2024 **Accepted |** July 24, 2024 **Published |** July 25, 2024.

Real estate price predictions are critical for stakeholders, including investors and developers, because they have a considerable impact on investment decisions and market stability. In order to fill in the shortcomings in earlier approaches, this work presents a novel methodology by utilizing Deep Learning (DL) and Machine Learning (ML) techniques to improve real estate price forecast accuracy. We used the "House Prices 2023 Dataset" from Kaggle, which contains 168,000 entries of Pakistani property data. Our methodology included extensive data preparation, feature engineering, and the use of various algorithms, including Linear Regression, Gradient Boosting, Random Forest, Convolutional Neural Networks (CNN), and K-Nearest Neighbors (KNN). The models were tested using MSE, RMSE, R-squared, and accuracy. KNN outperformed the other models, with a lower RMSE of 13.79 and a higher R-squared value of 0.85, indicating improved predictive accuracy. RF also produced impressive results, with an accuracy of 80%. Handling complicated feature interactions, guaranteeing model scalability, and controlling hardware resources were all challenges that suggested possibilities for future improvement. As a result, our research offers a solid foundation for raising forecasting accuracy in fluctuations in the market and emphasizes the possibility of utilizing ML approaches for better real estate price prediction.

Keywords: Real Estate; Machine Learning; Deep Learning; Market Dynamics; Investment Analysis













July 2024 | Vol 6 | Issue 3
Page | 1031

OPEN ACCESS

International Journal of Innovations in Science & Technology

Introduction:

Figure 3: Research paper on XGBoost

2.1.1. System Architecture/ Block Diagram

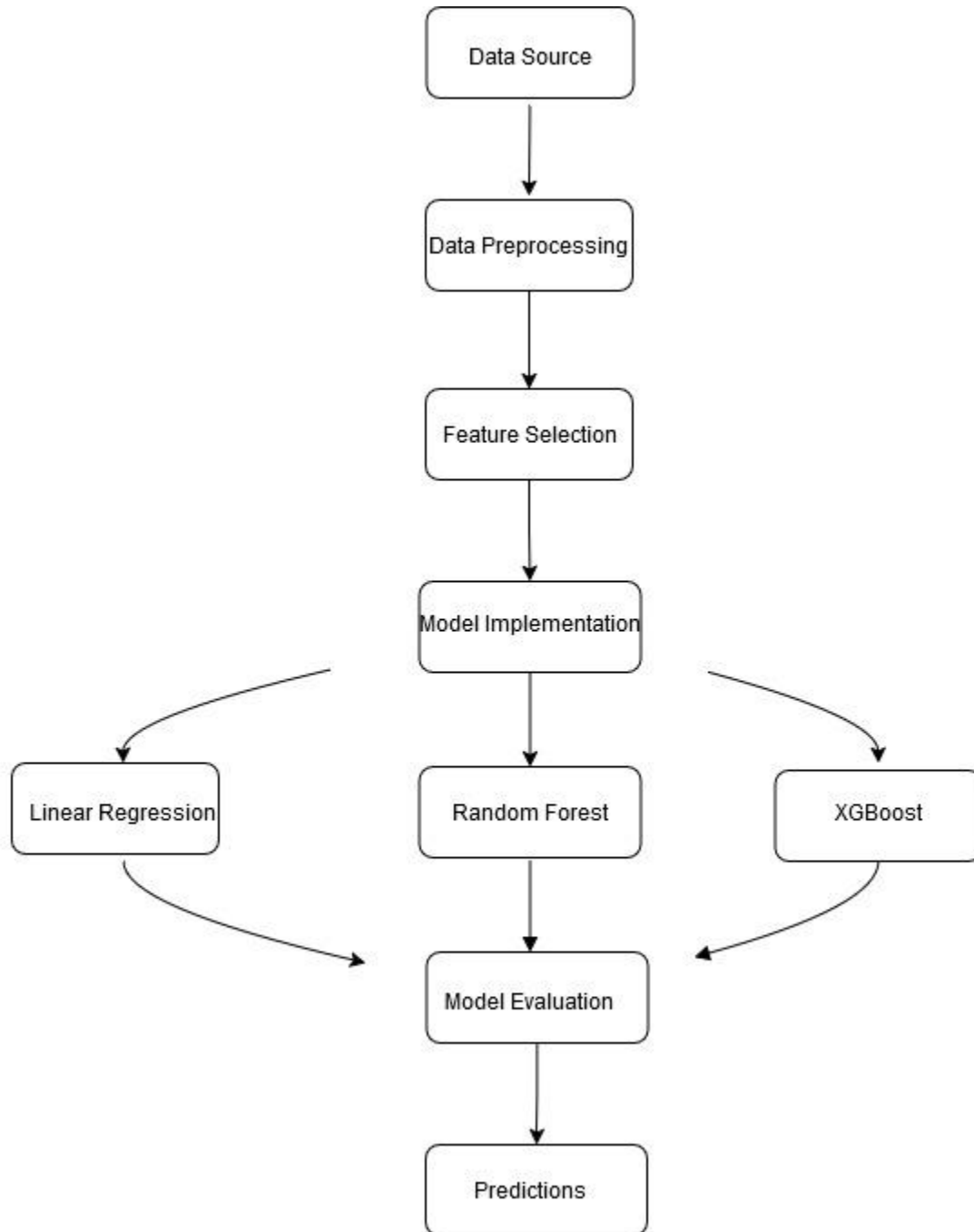


Figure 4: System Architecture Diagram.

2.2. Algorithm

The algorithms that are used in those studies are:

1. Linear Regression:

Gupta also took linear regression in 2020 as an instance. The author used it back then with linear regression. It implies the prices of houses are predictable by location, number of rooms, square footage, etc. He explicitly stated in the study an R-squared value of 0.72; therefore, this model can explain around seventy-two percent of the variation in the values of real estate properties. (Gupta, 2020)

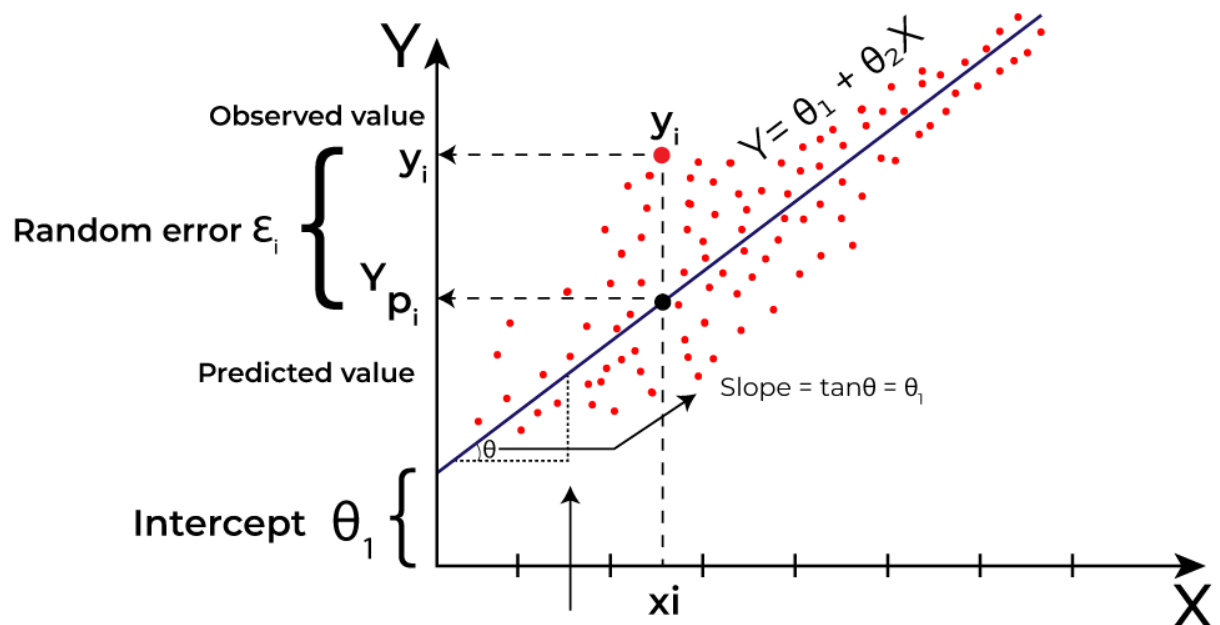


Figure 5: Linear Regression

2. Random Forest:

Smith and Zhang in 2021 published a paper where forest random algorithms were applied to a real estate database of over 15,000 properties. The model had completely dominated many of the conventional models including linear regression models through a complete Mean Absolute Error of lesser values which was approximately 10,000 lower than the slope twitches via linear regression. They had also discussed in their study how the algorithm could handle missing values and the complex nature of the dataset. (Smith, 2021)

Random Forest

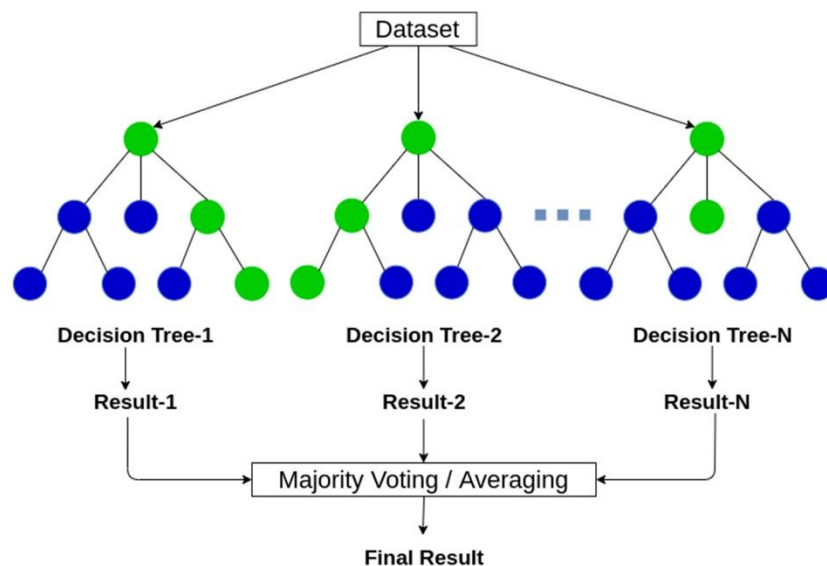


Figure 6: Random Forest

3. XGBoost:

In a recent paper, Lee has put to use XGBoost to predict the value of property, almost similar to what was done in the coursework. Among these are facilities in the neighborhood, transport, and historical price trends. In the random forest case, the error in the root mean square jumped all the way up near \$10,000. Therefore, by this time, the root mean square error that resulted from this algorithm was \$8,000, which thus neutralized both linear regression and the random forest algorithm. This algorithm aids in solving the problem of missing values and removing all possible ways of overfitting. (Lee, 2022)

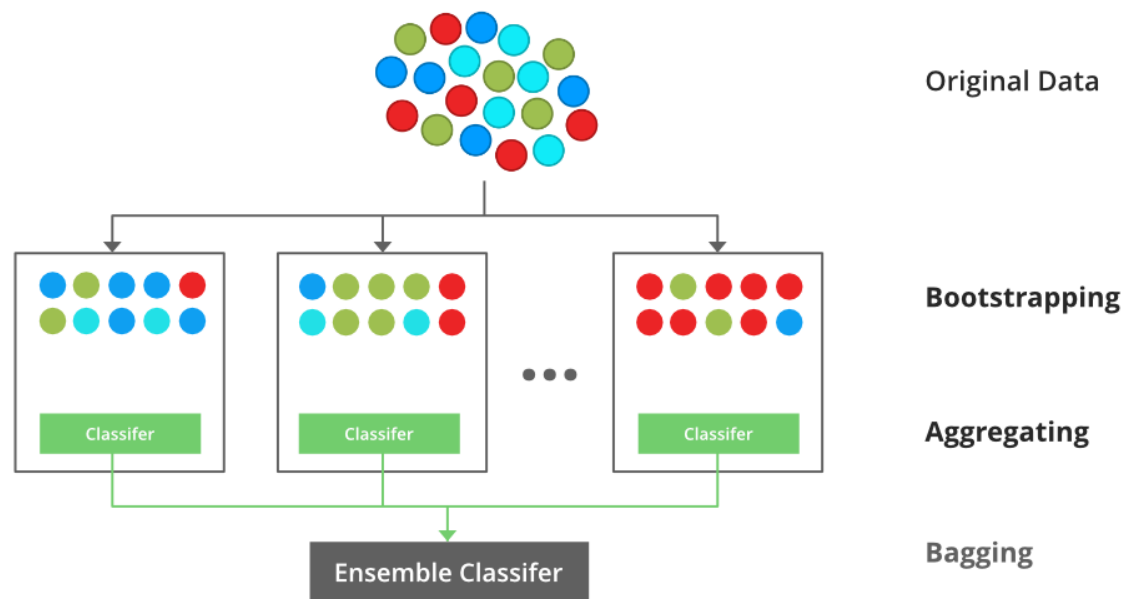


Figure 7: XGBoost

2.3. Evaluation Metrics

In machine learning there are several evaluation metrics that exists and some of them are often used to judge the performance of models for property pricing.

- R-square: This measures the variance in the dependent elements or factors which is predicted by the independent factors. It also gives an overall fit measure, but it is not necessarily predicted accurately every single time.
- Mean Absolute Error (MAE): This would calculate the average magnitude of the errors in the prediction which gives the untaught impression of the prediction accuracy.
- Root Mean Square Error (RMSE): It indicates a measure sensitive to large errors, thus appropriate for high-stake prediction making.
- Mean Squared Error (MSE): This is exactly like the RMSE, the only difference it has is it's without the square root and is more sensitive to the outliers

2.4. Comparison to previous studies

1. Study by Gupta et al. (2020)

- Algorithm: Linear Regression
- Dataset: Real estate data with features like area, location, and room count.
- Metrics: 0.72 for R-square
- Insights: The linear regression model was chosen model for this but it lacked to capture the relationship between the factors that weren't linear.

2. Study by Smith and Zhang (2021)

- Algorithm: Random Forest
- Dataset: Large dataset with over 15,000 entries.
- Metrics: This resulted in a Mean absolute error (MAE) of 1.90 and 0.90 R-squared.
- Insights: The Random Forest model could surpass Linear Regression mainly by very well treating feature interactions and non-linearities.

3. Study by Lee et al. (2022)

- Algorithm: XGBoost
- Dataset: It had features such as neighborhood amenities and historical price trends.
- Metrics: 0.79 for R-square
- Insights: XGBoost ran the best of all the previous model mentioned above. This was attributed to the fact that it could handle the missing values and regularized features well.

2.5. Dataset Description

The dataset used in this study contains 12,685 entries, reflecting individual property listings. 145 features engineered in aggregate represent a broad spectrum of property attributes reflecting clear diversity in the real estate market. To capture as much information as possible on the property, physical specifications, location, developer, and amenities both sets of features, categorical and numerical, are used. This is a very rich dataset, and hence forms the backbone for any in-depth analysis of factors influencing property prices.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD		
ID	Possessible Availability Floor No			Commercial	Developer	Approved	Units	Aval Price	Price (Eng)	Flooring	T Electricity	Maintenat	Maintenat	Booking A	Landmark	Covered A	Project Ne	sqft Price	Carpet Area	Area Nam	Property U	Unit of Cal	Society	Ownership	furnished	Bathroom	Parking	Facing	Ar		
1	12685	Under Cor	Dec 25	5	N	NA	KDMC	1	3150000	31.5 Lac	Vitrified	No/Rare P	Per sq. Un	3	100000	Kalyan Wt	635	NA	4960	375	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Covered East	Ge	
2	12684	Ready to h	NA	20	Y	TATA Hou	TMC	10	6300000	63 Lac	Vitrified	No/Rare P	Per sq. Un	3	100000	Rajoli nak	579	Tata Amar	10880	579	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Covered East	Ma	
3	12683	Ready to h	NA	18	N	Sai Satyan	KDMC	0	5400000	54 Lac	Vitrified	No/Rare P	Monthly	1200	100000	This propi	850	Sai Satyan	6332	585	Kalyan Wt	East Facir	Sq-ft	Y	Freehold	Unfurnish	2	1	Open East	Ge	
4	12682	Under Cor	Dec 25	5	N	Birla Estat	KDMC	70	9000000	90 Lac	Vitrified	No/Rare P	Monthly	3200	100000	Shahad is	1050	Birla Vany	8571	815	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Covered East	Pe	
5	12681	Under Cor	Dec 24	8	Y	Godrej Prc	NA	NA	4950000	49.5 Lac	NA	NA	NA	NA	500000	majiwada	561	Godrej Nlr	8824	419	Kalyan Wt	NA	Sq-ft	Y	Freehold	Semi-Furr	2	NA	East	Ge	
6	12680	Under Cor	Mar 23	16	N	Tycoons G	ABAAut	2	6570000	65.7 Lac	Vitrified	No/Rare P	Monthly	2	50000	Close to tr	1067	Tycoons S	6157	667	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Covered North	East	Ge
7	12679	Ready to h	NA	1	N	NA	KDMC	4	5500000	55 Lac	Vitrified	No/Rare P	Monthly	2000	51000	Tilak Chov	725	NA	7586	550	Kalyan Wt	NA	Sq-ft	N	Co-operat	Unfurnish	2	NA	East	Ge	
8	12678	Ready to h	NA	4	N	NA	TMC	1	4200000	42 Lac	Vitrified	No/Rare P	Monthly	1500	51000	Agra Roac	650	NA	6462	585	Kalyan Wt	NA	Sq-ft	N	Freehold	Semi-Furr	1	NA	East	Ma	
9	12677	Under Cor	Dec 25	16	N	Godrej Prc	ABAAut	15	3449000	34.5 Lac	Vitrified	No/Rare P	Monthly	3	50000	Within 4 n	592	Godrej Riv	5826	370	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	1	NA	North	East	Ge
10	12676	Ready to h	NA	5	N	NA	NA	NA	5000000	50 Lac	NA	NA	NA	NA	100000	NA	644	NA	7764	430	Kalyan Wt	NA	Sq-ft	N	NA	Unfurnish	1	NA	West	NA	
11	12675	Under Cor	Dec 24	10	N	NA	KDMC	25	2990000	29.9 Lac	Vitrified	No/Rare P	Monthly	1200	27000	near don i	550	NA	5436	370	Kalyan Wt	NA	Sq-ft	Y	Freehold	Furnished	2	1	Open East	Ge	
12	12674	Under Cor	Dec 24	10	N	NA	KDMC	25	4190000	41.9 Lac	Vitrified	No/Rare P	Monthly	1500	27000	near don i	850	NA	7618	570	Kalyan Wt	NA	Sq-ft	Y	Freehold	Semi-Furr	2	1	Covered East	Ge	
13	12673	Ready to h	NA	9	Y	Tharwani	NA	NA	4900000	49 Lac	Vitrified	NA	NA	NA	NA	kalyan we	595	Rosalie	8235	NA	Kalyan Wt	NA	Sq-ft	N	Co-operat	Unfurnish	1	NA	North	East	Ma
14	12672	Ready to h	NA	2	N	NA	ABAMNC	1	2187500	21.9 Lac	Vitrified	No/Rare P	Monthly	1000	51000	Yash Hote	595	NA	3676	450	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	1	1	Covered North	East	Ge
15	12671	Ready to h	NA	16	N	Gurukrupi	NA	0	4700000	47 Lac	Vitrified	NA	Per sq. Un	3	100000	Near Yogi	695	Guru Atm	6763	462	Kalyan Wt	NA	Sq-ft	N	Co-operat	Unfurnish	1	NA	North	West	Ma
16	12670	Under Cor	Jan 26	9	N	NA	ABADev	10	10450000	1.04 Cr	Vitrified	No/Rare P	Monthly	1600	51000	Godrej Hill	NA	NA	10674	980	Kalyan Wt	NA	NA	Y	Freehold	Unfurnish	2	1	Covered North	East	Ge
17	12669	Under Cor	Nov 25	18	N	NA	KDMC	NA	12000000	1.20 Cr	Vitrified	No/Rare P	Monthly	2500	100000	khadihak	1350	NA	8889	927	Kalyan Wt	East Facir	Sq-ft	N	Power Of	Unfurnish	3	1	Open East	Ge	
18	12668	Under Cor	Mar 23	9	N	Gurukrupi	KDMC	3	7000000	70 Lac	Vitrified	No/Rare P	Monthly	2500	51000	Atm 500m	960	Guru Atm	7608	650	Kalyan Wt	NA	Sq-ft	N	Freehold	Unfurnish	2	1	Open North	East	Ge
19	12667	Under Cor	Dec 23	9	N	Rutu Grou	KDMC	7	7000000	70 Lac	Vitrified	No/Rare P	Monthly	0	100000	agrawal c	1010	Riverview	6930	761	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Open North	East	Ge
20	12666	Ready to h	NA	4	N	Ajmera Re	KDMC	1	4800000	48 Lac	Vitrified	No/Rare P	Per sq. Un	4	51000	Yogidham	711	Ajmera Ne	6731	460	Kalyan Wt	NA	Sq-ft	N	Freehold	Unfurnish	2	NA	North	East	Ma
21	12665	Under Cor	Aug 25	20	N	Birla Estat	KDMC	50	9500000	95 Lac	Ceramic T	No/Rare P	Monthly	4500	100000	near by sh	1000	Birla Vany	9500	661	Kalyan Wt	NA	Sq-ft	Y	Freehold	Semi-Furr	2	1	Covered North	West	Ge
22	12664	Under Cor	Dec 24	18	N	Ajmera Re	KDMC	10	7499000	75 Lac	Ceramic T	No/Rare P	Monthly	0	45000	KalyanMu	709	One Kalya	12415	709	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Covered North	East	Ge
23	12663	Ready to h	Dec 22	5	N	Ajmera Re	KDMC	2	4500000	45 Lac	Vitrified	No/Rare P	Monthly	2200	51000	Yogidham	675	Ajmera Bli	6666	450	Kalyan Wt	East Facir	Sq-ft	Y	Freehold	Unfurnish	2	1	Open East	Ge	
24	12662	Under Cor	Jun 25	18	N	Godrej Prc	TMC	23	4300000	43 Lac	Vitrified	No/Rare P	Yearly	0	50000	Bhumi Wc	590	Godrej Up	7288	367	Kalyan Wt	East Facir	Sq-ft	N	Freehold	Unfurnish	2	1	Covered East	Ge	
25	12661	Under Cor	Jan 24	11	N	Tycoons G	ABAAut	NA	3950000	39.5 Lac	Vitrified	No/Rare P	Monthly	1200	51000	Tycoons S	NA	Tycoons S	NA	390	Kalyan Wt	East Facir	NA	Y	Freehold	Furnished	2	1	Covered East	Ge	
26	12660	Ready to h	Jan 23	21	N	Raunak G	KDMC	NA	3600000	36 Lac	Vitrified	No/Rare P	Monthly	NA	45000	Near Don	645	Raunak U	5581	467	Kalyan Wt	Near Don	Sq-ft	Y	Freehold	Semi-Furr	2	1	Open North	East	Ge
27	12659	Ready to h	NA	5	N	Vaishnavi	ABAMNC	NA	5580000	55.8 Lac	Vitrified	NA	Monthly	1000	100000	Pari Hosp	650	Vaibhavi C	5508	460	Kalyan Wt	East Facir	Sq-ft	N	Freehold	Unfurnish	1	NA	East	Ge	
28	12658	Under Cor	Dec 24	7	N	NA	ABAMNC	2	14796000	1.47 Cr	Granite,M	No/Rare P	Monthly	7000	100000	Kalyan Wt	2160	NA	6850	1385	Kalyan Wt	East Facir	Sq-ft	Y	Freehold	Unfurnish	3	2	Covered East	Ge	
29	12657	Under Cor	Dec 25	3	Y	NA	KDMC	12	4300000	43 Lac	Vitrified	No/Rare P	Per sq. Un	3	100000	Kalyan Wt	854	NA	5035	505	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Covered East	Ge	
30	12656	Under Cor	Mar 26	10	N	NA	KDMC	12	4185500	41.9 Lac	Vitrified	No/Rare P	Per sq. Un	3	100000	Kalyan Stz	384	NA	10899	384	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Covered East	Ge	
31	12655	Under Cor	Mar 23	16	N	Tycoons G	ABAAut	2	6570000	65.7 Lac	Vitrified	No/Rare P	Monthly	2	50000	Close to tr	1067	Tycoons S	6157	667	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	2	1	Covered North	East	Ge
32	12654	Ready to h	NA	1	N	NA	KDMC	4	5500000	55 Lac	Vitrified	No/Rare P	Monthly	2000	51000	Tilak Chov	725	NA	7586	550	Kalyan Wt	NA	Sq-ft	N	Co-operat	Unfurnish	2	NA	East	Ge	
33	12653	Ready to h	NA	4	N	NA	TMC	1	4200000	42 Lac	Vitrified	No/Rare P	Monthly	1500	51000	Agra Roac	650	NA	6462	585	Kalyan Wt	NA	Sq-ft	N	Freehold	Semi-Furr	1	NA	East	Ma	
34	12652	Under Cor	Dec 25	16	N	Godrej Prc	ABAAut	15	3449000	34.5 Lac	Vitrified	No/Rare P	Monthly	3	50000	Within 4 n	592	Godrej Riv	5826	370	Kalyan Wt	NA	Sq-ft	Y	Freehold	Unfurnish	1	NA	North	East	Ge
35	12651	Ready to h	NA	5	N	NA	NA	NA	5000000	50 Lac	NA	NA	NA	NA	100000	NA	644	NA	7764	430	Kalyan Wt	NA	Sq-ft	N	NA	Unfurnish	1	NA	West	NA	
36	12650	Under Cor	Dec 24	10	N	NA	KDMC	25	2990000	29.9 Lac	Vitrified	No/Rare P	Monthly	1200	27000	near don i	550	NA	5436	370	Kalyan Wt	NA	Sq-ft	Y	Freehold	Furnished	2	1	Open East	Ge	
37	12649	Under Cor	Dec 24	10	N	NA	KDMC	25	4190000	41.9 Lac	Vitrified	No/Rare P	Monthly	1500	27000	near don i	850	NA	7618	570	Kalyan Wt	NA	Sq-ft	Y	Freehold	Semi-Furr	2	1	Covered East	Ge	

Figure 8: Dataset of the real estate property

2.5.1. Key Features

Some of the important factors included in this dataset that can help to access the determination of the price of property are as follows:

- **Possession Status:** It is a categorical variable that shows whether the property is ready to move into or still under construction. Important variable because the present marketability of a property with the possibility of immediate occupation can result in high values as compared to those still under construction.
- **Price:** It is the target factor for the dataset and refers to the price of the property. It is continuous numerical value that shows the sale price of an item, which is the variable result the analysis is trying to forecast from other property features.
- **Floor Number:** There is a categorical feature that gives the idea of the level at which the property is situated. Relevant mostly to apartments and multi-story structures in general, higher floors provide better views and a quieter living environment. Therefore, they can influence their price.
- **Developer:** Developer is a categorical feature representing the builder or company responsible for the construction of the property. The reputation and history of the developer can strongly affect the perceived value of the property, because more reputable developers might automatically be associated with houses of better quality and hence premium pricing.
- **Land Area/Covered Area:** These are numeric variables indicating land area, which is the whole area of land that a property takes, and covered area which is within the structure of the property. Large size of the property will denote higher value, therefore it tends to be higher priced.

2.6. Data source

The current dataset being used in this coursework is on property listings and is an open source of data. The data that is at hand is diversified to such an extent that it takes care of different kinds of properties across various locations. In all consideration, it was more or less all-inclusive with respect to everything that might have an influence, direct or indirect, on real estate prices, including the geographic location, size of the property, amenities, and also the developer's repute. In short, the dataset is chosen to be varied in nature in representing real-life

complexities regarding the housing market, wherein a gamut of reasons finally causes the price of a property.

2.6.1. Initial Observation and Data Preprocessing

- **Missing Values:** The dataset provided some of the missing or null values in the coursework. The imputations were done using techniques such that missing entries could be catered for, or rows or columns were dropped whenever needed based on relevance and importance of the feature in question. Imputations are done on such features that show low proportions of missing values; otherwise, rows or columns having too many missing data were dropped.
- **Highly Skewed Distribution Price:** A few property prices were extremely right-skewed distributed, with most properties clustered around an average low price, and very few were of high value. One correct way to treat these kinds of skewness issues is the log transformation of the feature Price. This then normalizes the distribution to come out very good for modeling, especially enhancing the performance of machine learning algorithms.
- **High Cardinality Features:** Developer is high cardinality because there are many unique values. High cardinality features are very cumbersome to process in such a way that they do not overfit the model and do not add many dimensions. Whenever such a situation arises, the categorical variables were converted into numerical representations by label encoding to let the model process it.

2.7. Exploratory Data Analysis (EDA)

In the preliminary data preparation phases, Exploratory Data Analysis (EDA) can be touted as one of the very crucial tasks, allowing the data scientist to have an overview of their dataset and possibly unveil any patterns, anomalies, or relationships of interest later in the modeling phase. There are a large number of such attributes in this dataset that are anticipated to be influential in determining the way things work, such as property price, location, floor number, possession status, and developer details.

The implementation of the first stage EDA commenced with treating missing values. An initial check was run to observe how many of the features had null values; for instance, missing developer names and floor number.

Missing values were then replaced by the appropriate values wherever possible, or rows that contained very large data misses were dropped in a manner that does not harm the data integrity.

The analysis then zeroed in on the distribution of numerical variables centered on property price and land area. The distribution of features of property price distribution was found to be right-skewed, so that in most cases, only a few properties were high-valued and hence distortion on the model prediction. In this case, log transformation was applied as one of the possibilities in normalizing the distribution and making it work better for the model.

This included high-dimensional categorical variables in few categories like possession status and developer names. It encoded dimensionality of developer names after creating a new category 'other', where less frequent values were combined. The major point was to let this pattern's relationship be in line with the target variable: property prices. Subsequently, these were checked for relationships in numerical variables by using them to carry out correlation analysis. The plotted correlation matrix explained that land area and the number of bedrooms had a strong positive correlation with property prices; for some features, this correlation was either close to zero or zero.

This is something dynamic attribute selection based on high prediction power: land area and property price, after which box plots were used for outlier detection. Outliers will be detected by techniques. Extreme values will be identified using outlier detection techniques and capped or dropped in such a way that no skewed data is trained on. This will ultimately assist in establishing the trend and relationship of data in the dataset with the help of scatter plots and histograms. A scatter plot for land area against property price will be illustrative enough. This is non-linear in nature, which can well be captured by something like Random Forest and XGBoost during their workings. The visualizations are highly informative and gave us a total view of what influences property prices in the dataset. The steps carried out in EDA have perfectly laid down the foundation for modeling and evaluation. What will make the model robust is actually this comprehensive EDA.

3. Proposed Solution

3.1. Introduction

The primary goal here is to come up with a predictive model of property prices using the machine learning algorithms. This will be very critical not only to buyers but also to sellers, and to real estate agencies that require pin-point decisions in executing their business. The present problem is dealt with a dataset having a large number of features related to properties in terms of location, size, the number of rooms, amenities provided, etc., and correctly predicting the price of a property with efficient models. To solve the problem that has all been mentioned very efficiently three of these powerful algorithms has been considered. Those are linear regression, random forests, and XGBoost.

3.2. Data Preprocessing

Datasets made by humans are never perfect and can contain a lot of mistakes. So to solve this problem data preprocessing is done since it is the one of the critical starting phase to make sure that the dataset is correct and high quality (Praise Peace, 2024). It basically investigates missing values, outliers, and inconsistencies within the dataset. Then, applicable techniques are used for imputation of such missing values or dropping them to eschew any kind of bias. Normalisation of numerical features are done for consistency scaling and categorical features making them ready for machine learning models which are encoded in a form. This dataset processing will ensure that the dataset provided in this coursework is clean and available for training and testing purposes. This will also split the dataset into two parts, one for training and another for testing purposes which will help the model to validate and evaluate.

3.3. Features Selection

Feature Selection is one of the boosting concepts in fine-tuning machine learning models. It essentially works on provided variables required and filters out noises caused by irrelevant features. (Menon, 2024) Correlation analysis is done with the features to determine which feature is strongly correlating with the target variable, which in our case is property prices. Redundant features should be removed during this process of feature selection for computational efficiency. After the features are selected, only these selected features will be used

to train the models according to its requirement to learn about the different meaningful patterns and their relationship from the data.

3.4. Model Implementation

Three algorithms have been used in this coursework to solve the problem mentioned above: Linear Regression, Random Forest, and XGBoost. These were implemented because of unique characteristics and suitability based on different angles of a problem. Finally, once with their performance so far thoroughly examined, each model is implemented, step by step.

3.4.1. Linear Regression

Linear Regression is applied to set up the baseline to model the initial performance metric, where it assumes a linear relationship between the dependent variable, property prices, and independent variables or features. (Anon., 2024) It then minimizes the cost function for error made between their predicted values against their actual values. The cost function is given by:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$$

Figure 9: Linear Regression Formula

Where: m is number of data points, h_{θ} is the hypothesis function, θ is the parameters to optimize. Linear Regression gives an exact idea as to how each individual feature is related to property prices; it may well serve as a very good starting point for analysis.

3.4.2. Random Forest

It's an ensemble learning method, which at training time builds a number of decision trees; it then combines their output for greater accuracy. Each tree is built on a bootstrap sample of data, with a random subset of features considered at each node. This dramatically reduces overfitting and increases generalization of the results. Picking up non-linear relations and feature interactions are Random Forest's forte. (Revathy2, 2024)

Therefore, it also allows for determining important features, which we will use later to zero in on important variables in the prediction of property prices. The prediction from the Random

Forest model is an average prediction from all individual trees. It provides fairness as well as soundness of prediction. However, the expense of computation for building many trees will be expensive, mostly for large datasets.

3.4.3. XGBoost

XGBoost is an optimization of gradient boosting. It improves predictions iteratively through minimization of the loss function in the real space by adding a gradient regularized term. As the formulation of XGBoost includes adding the regularization terms for controlling overfitting, it is very effective and accurate. (Yi Chen, 2024) The XGBoost consists of two components:

- a. Loss function (L)

$$\text{Objective} = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Figure 10: Loss function Formula

- b. Regularization term (Ω)

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

Figure 11: Regularization term

To handle the large dataset very efficiently it uses a parallel processing ability. Balancing the trade-off between bias and variance, it comes in very handy for complicated relationships in the data. It has advanced features and afforded scalability, due to which it delivers high performance even in hard scenarios

3.5. Justification for algorithm choice

For this coursework the algorithms that were chosen are linear regression, random forest, XGBoost which are one of the best algorithms to train the model to predict the prices of the properties. Since linear regression is interpretable and simple, it is very suitable for developing baseline models and learning linear relationships within the data. It is very important to understand how impactful a single attribute can be to determine the price of the property.

While the random forest captures non linear patterns and also the interaction of the features. Other than that, Random Forest comes with an in-built feature importance ranking that will help give very useful insights into what contributes to the most in the prices of the property. In that respect, it still remains suitable for any dataset with complex relations and fluctuations in feature importance.

XGBoost was the most efficient model used among all state-of-the-art boosting frameworks in gradient boosting, throwing all these efficiencies and accuracies with scalability. The model would really generalize well with new data if it had provisions to work well with very large datasets by leveraging the regularization methods. It was computationally efficient within its structure, paralleling various stages, quite advanced technologically with the techniques used in tree pruning, learning rate optimization, and others. Combining all these algorithm can create a balanced solution to handle simple to and advance modelling

3.6. Pseudocode for algorithm

3.6.1. Linear regression

- Step 1: Input the training dataset (X, y)
- Step 2: Define the Hypothesis function
- Step 3: Initialize the parameters
- Step 4: Define the cost function
- Step 5: Cost function is minimized by Gradient Descent
- Step 6: Check for the convergence

 26

Step 7: Evaluate the model

Step 8: Make **the** predictions

3.6.2. Random Forest

Step 1: Input the dataset (X, y)

Step 2: Specify the number of the trees (N) in the forest.

Step 3: Randomly choose from the data and features for every tree in the forest

Step 4: Making the decision tree using the selected data and features.

Step 5: Aggregate predictions from all trees using majority voting for classification or averaging for regression

Step 6: Metric to evaluate model is using MSE, RMSE, and R^2 .

Step 7: Predict for unseen and test data.

3.6.3. XGBoost

Step 1: Begin by adding the training dataset.

Step 2: Formulate the model by baseline prediction.

Step 3: Residuals from the previous predictions are computed.

Step 4: Fit a decision tree to predict the residuals.

Step 5: Update predictions by adding weighted residuals using the learning rate.

Step 6: Apply regularization; this will avoid an overfitting condition using the objective function.

Step 7: The point at which the improvement brought by it to the loss function is below a predefined threshold is what we should be watching out for.

Step 8: Evaluate the model with different metrical values or parameters using the test data for the model, like MSE, RSME, R^2 , and others

Step 9: Make predictions on the unseen data.

3.7. Flowchart

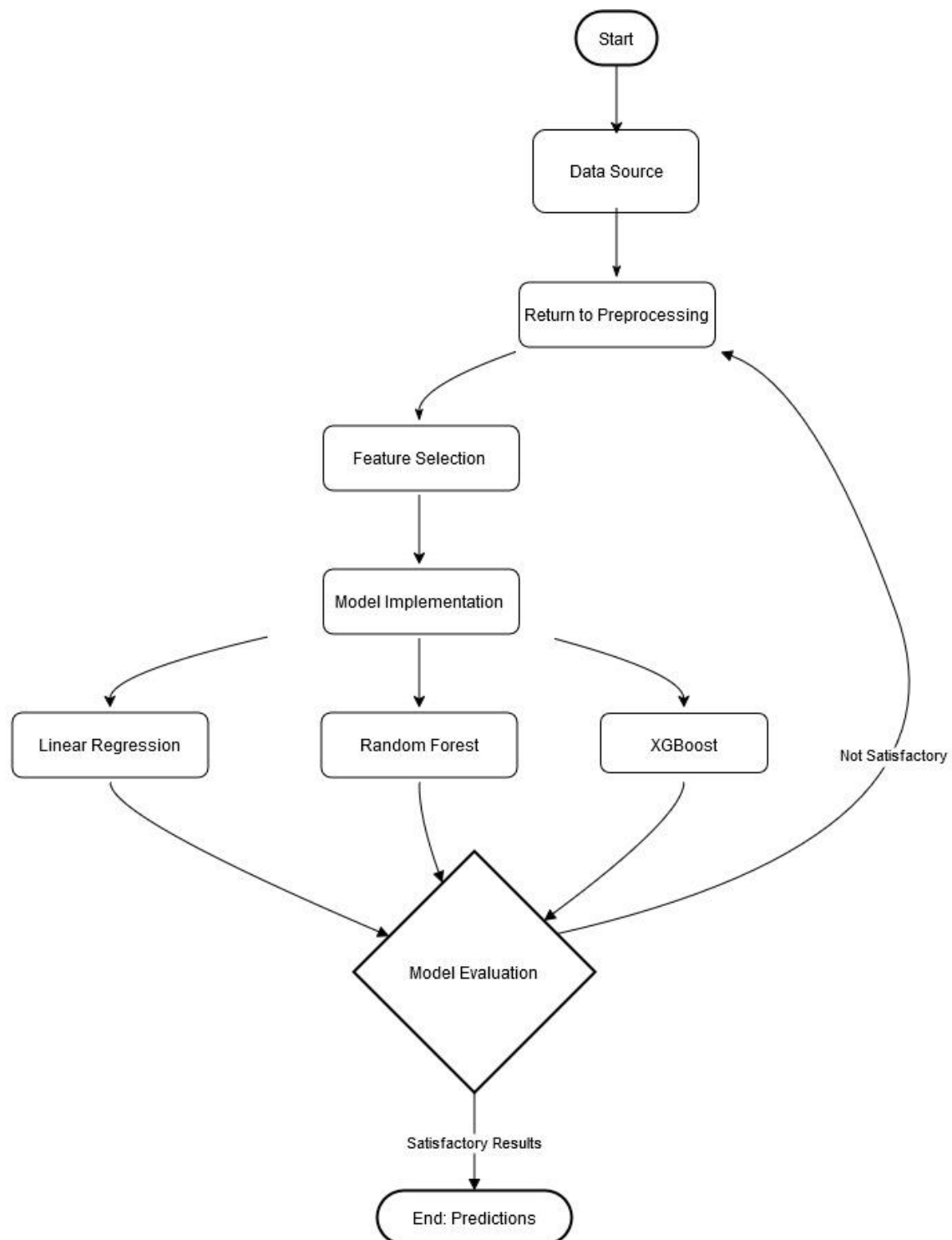


Figure 12: Flowchart Diagram

3.8. Diagrammatic Representation

3.8.1. Data Flow Diagram

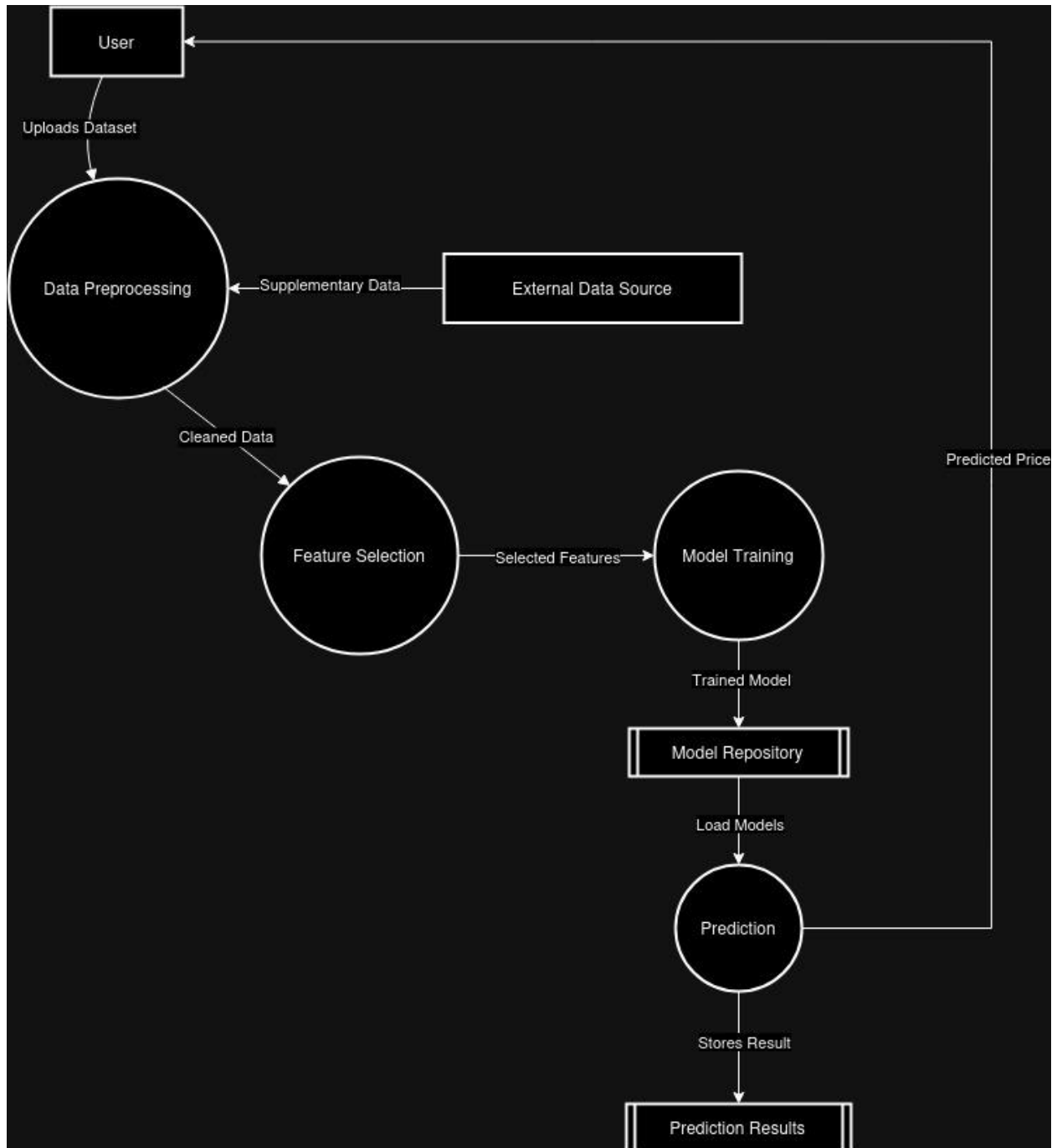


Figure 13: Data Flow Diagram

3.8.2. Algorithm Workflow

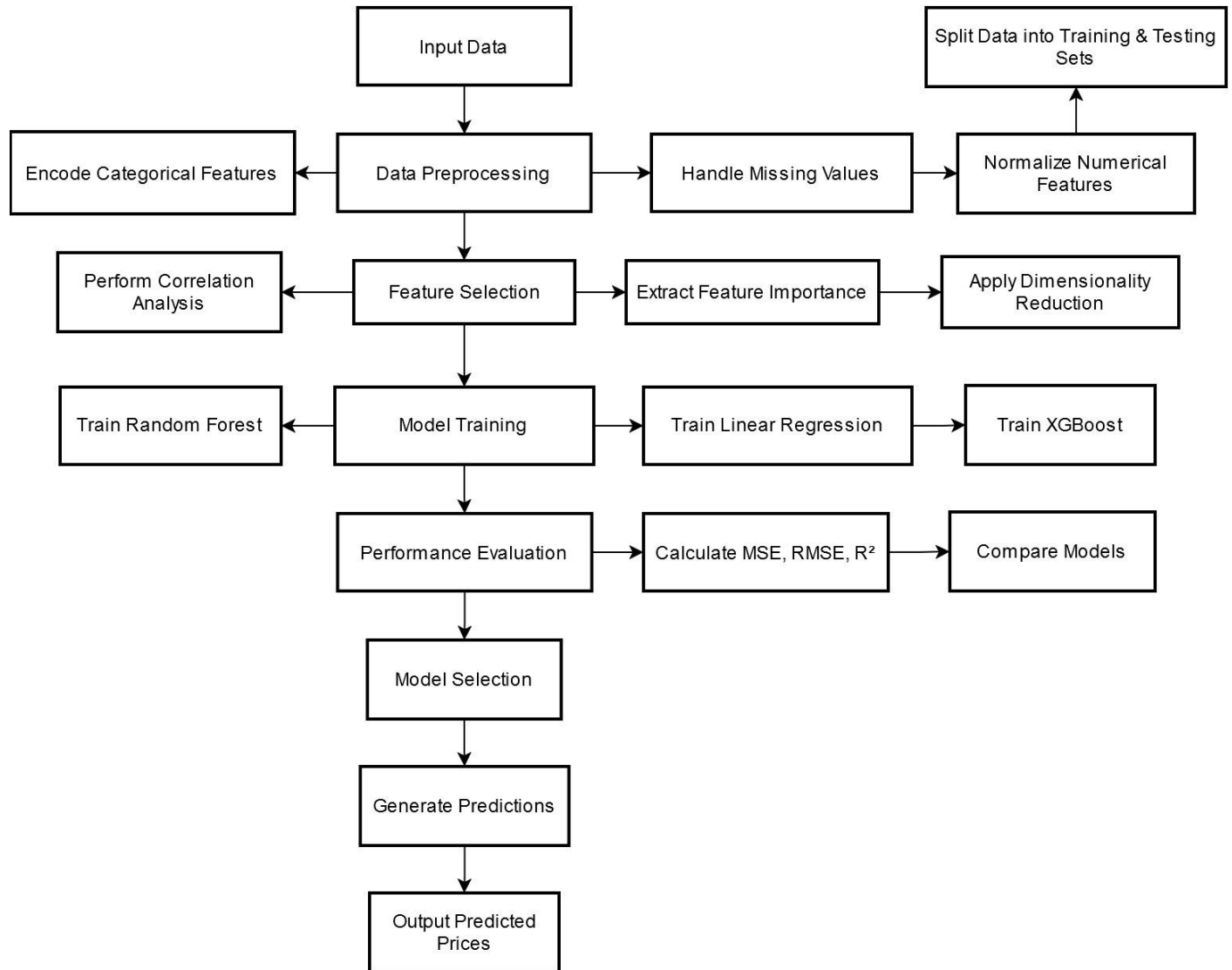


Figure 14: Algorithm Workflow

3.9. Data Pre-Processing

An exhaustive cleaning exercise is conducted on the dataset and transformed, thus making it ready for model training. Get rid of all irrelevant columns such as 'ID,' 'Availability Starts From,' as they offer no relevance to the analysis. Missing values of numerical columns are imputed with the median value so that there's no bias introduced. Missing values for the categorical columns are again treated as a label "Unknown"; thereby, no information would be lost. Prior to encoding, all columns in the data were converted to strings in order to have uniform data types.

Further breaking down of the dataset—features (X) and target variable ('price')—80% of the data was taken for training and 20% was left for testing. Finally, StandardScaler is applied to make numerical features normal. These preprocessing techniques make the data ready for model training effectively.

3.10. Detail process

```
# Importing Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.metrics import mean_squared_error, r2_score

# Loading the Dataset
file_path = 'properties.csv'
data = pd.read_csv(file_path)
print("Dataset Shape:", data.shape)
print("First few rows of the dataset:")
print(data.head())
```

Dataset Shape: (12685, 145)
First few rows of the dataset:

	ID	Possession Status	Availability Starts From	Floor No	Commercial	\
0	12685	Under Construction	Dec '25	5	N	
1	12684	Ready to Move	NaN	20	Y	
2	12683	Ready to Move	NaN	18	N	
3	12682	Under Construction	Dec '25	5	N	
4	12681	Under Construction	Dec '24	8	Y	

	Developer	Approved Authority	Name	\
0	NaN	NaN	KDMC	
1	TATA Housing Development Company Ltd.	NaN	TMC	
2	Sai Satyam Developers	NaN	KDMC	
3	Birla Estates	NaN	KDMC	
4	Godrej Properties	NaN	NaN	

	Units Available	Price	Price (English)	...	Rentable	CommuniPfty	Space	\
0	1.0	3150000.0	31.5 Lac	1	
1	10.0	6300000.0	63 Lac	1	
2	0.0	5400000.0	54 Lac	1	
3	70.0	9000000.0	90 Lac	1	
4	NaN	4950000.0	49.5 Lac	1	

	Retail Boulevard (Retail Shops)	Cycling & Jogging Track	\
0	1	1	
1	1	1	
2	1	1	
3	1	1	
4	1	1	

Figure 15: Importing all the modules and loading the data set.

```
# Dropping unnecessary columns
data = data.drop(columns=['ID', 'Availability Starts From'], errors='ignore')

# Handling missing values
for col in data.select_dtypes(include=['float64', 'int64']).columns:
    data[col] = data[col].fillna(data[col].median()) # Filling numerical missing values with median

for col in data.select_dtypes(include=['object']).columns:
    data[col] = data[col].fillna("Unknown") # Filling categorical missing values with "Unknown"

print("\nDataset Info After Cleaning:")
print(data.info())
```

Dataset Info After Cleaning:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12685 entries, 0 to 12684
Columns: 143 entries, Possession Status to Pantry Type
dtypes: float64(12), int64(92), object(39)
memory usage: 13.8+ MB
None

Figure 16: Dropping Unnecessary column and handling missing values

```
#Converts all categorical columns to strings to ensure consistent type
for col in data.select_dtypes(include=['object']).columns:
    data[col] = data[col].astype(str)

label_encoders = {}
for col in data.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le
```

Figure 17: Encoding Categorical Variables

```
if 'Price' in data.columns:
    # Histogram of the target variable (Price)
    data['Price'].hist(bins=20)
    plt.title('Price Distribution')
    plt.xlabel('Price')
    plt.ylabel('Frequency')
    plt.show()

    # Correlation heatmap
    plt.figure(figsize=(10, 8))
    sns.heatmap(data.corr(), annot=True, cmap='coolwarm')
    plt.title('Correlation Matrix')
    plt.show()
```

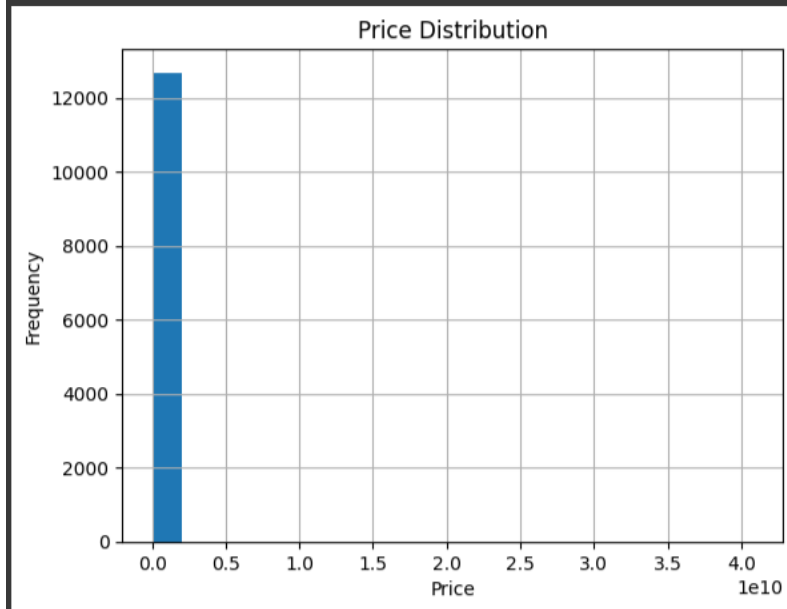


Figure 18: Data Visualization

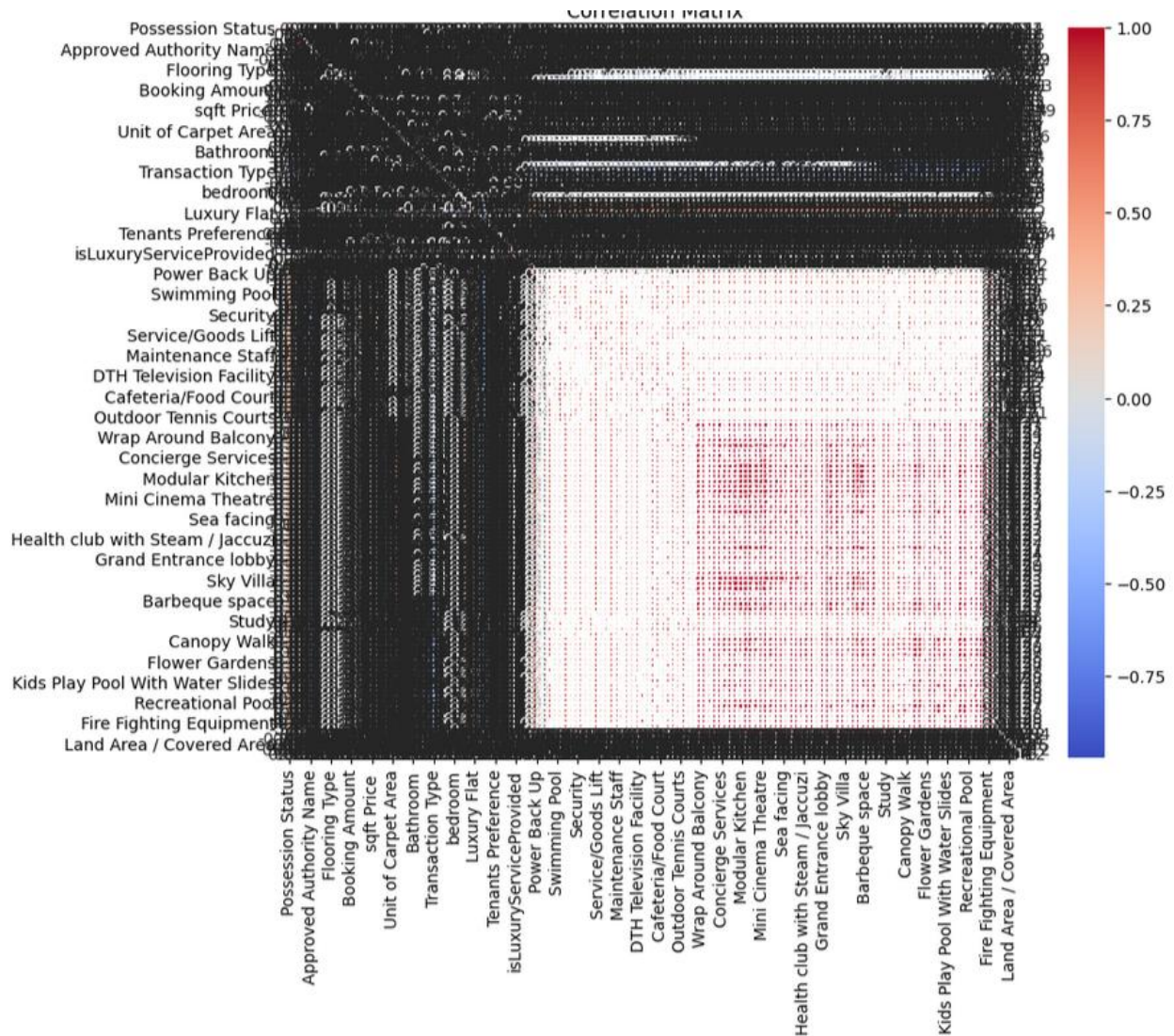


Figure 19: Correlation Matrix

```
#feature-target split
if 'Price' in data.columns:
    X = data.drop(columns=['Price'], errors='ignore')
    y = data['Price']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 20: Training Feature-target split

```
[78] #Training the Test Split
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

      #Standardizing the Numerical Features
      scaler = StandardScaler()
      X_train = scaler.fit_transform(X_train)
      X_test = scaler.transform(X_test)
```

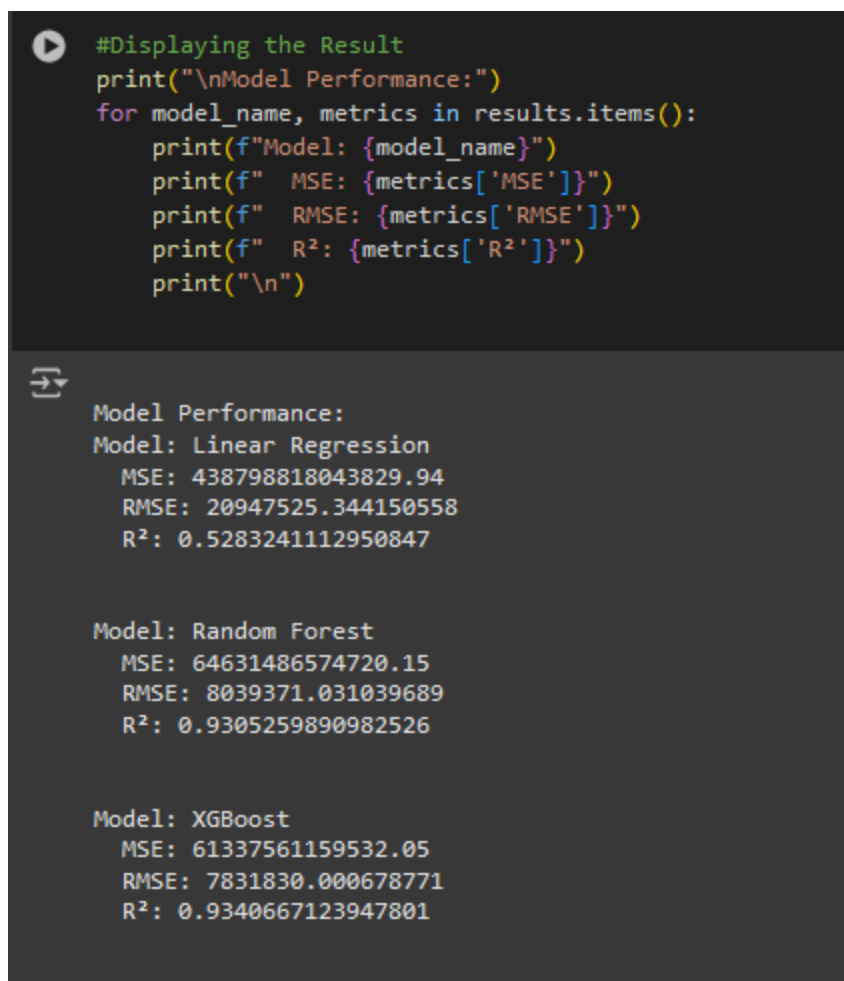
Figure 21: Training the Test Split and Standardizing the numeric features

```
[79] # Traning the model
      models = {
          "Linear Regression": LinearRegression(),
          "Random Forest": RandomForestRegressor(n_estimators=100, random_state=42),
          "XGBoost": XGBRegressor(n_estimators=100, random_state=42)
      }

      results = {}
      for name, model in models.items():
          print(f"Training {name}...")
          model.fit(X_train, y_train)
          y_pred = model.predict(X_test)
          mse = mean_squared_error(y_test, y_pred)
          rmse = np.sqrt(mse)
          r2 = r2_score(y_test, y_pred)
          results[name] = {"MSE": mse, "RMSE": rmse, "R²": r2}
```

```
↔ Training Linear Regression...
   Training Random Forest...
   Training XGBoost...
```

Figure 22: Training the model



```

#Displaying the Result
print("\nModel Performance:")
for model_name, metrics in results.items():
    print(f"Model: {model_name}")
    print(f"    MSE: {metrics['MSE']}")
    print(f"    RMSE: {metrics['RMSE']}")
    print(f"    R²: {metrics['R²']}")
    print("\n")

```

Model Performance:
 Model: Linear Regression
 MSE: 438798818043829.94
 RMSE: 20947525.344150558
 R²: 0.5283241112950847

 Model: Random Forest
 MSE: 64631486574720.15
 RMSE: 8039371.031039689
 R²: 0.9305259890982526

 Model: XGBoost
 MSE: 61337561159532.05
 RMSE: 7831830.000678771
 R²: 0.9340667123947801

Figure 23: Displaying the result.

4. Conclusion

17 The recent project is about the prediction of property rates by utilizing three machine learning techniques--Linear Regression, Random Forest, and XGBoost. Property valuation is one of the most potent tools in real estate which shall enable the buyers, sellers, investors to transact wisely. The solution fared as promised by means of proper data preparation, feature selection, and advanced machine learning methods responsible for accuracy and adaptability within results.

The final model was an amalgamation of contributions. The linear regression built a very clean and simple baseline, whereas Random Forest added robustness and importance measures of each feature. Whereas XGBoost is very due to its high performace nature. There is one reason why, even with every added prediction, it would come out as the most accurate: its ability to solve complex patterns within data.

Considering that MSE, RMSE, R^2 are the Drivers of such High-Speed Performance, harmony, and reliability between these models may be guaranteed. However, there are a few shortcomings in the project. The major concern for this is that the model becomes heavily on the training data. Secondly, XGBoost is very powerful but quite computationally expensive; thus, it may not be as applicable for a real-time system or resource-constrained system.

5 This is probably one of the most interesting coursework that has come up, given that A.I. and Machine Learning are very powerful in almost every industry within which one is working. It is exactly like the real estate market is trying to predict the prices of properties for one of their objectives.

The developed solution has now been implemented in such a way that so many developed properties come up, not only very accurate and effective but also very robust in further development and wide implementation specifically regarding the real estate industry.

References

- Anon., 2024. Application and interpretation of linear-regression analysis. *Medical Hypothesis Discovery & Innovation in Ophthalmology*, 13(3), pp. 151-159.
- ÇELİK, Ö., 2018. A Research on Machine Learning Methods and Its Application. *Journal of Educational Technology & Online Learning*, 1(3).
- Emma Oye, E. F. J. O., 2024. *Unsupervised vs. Supervised Learning*, s.l.: Emma Oye.
- Gupta, P. S. R. & K. S., 2020. Predicting Housing Prices Using Linear Regression. *Journal of Machine Learning Applications*, 12(3), pp. 45-56.
- Kamba, H. R. D. D., 2021. The Influence of Internal and External Factors on the Stock Price of Property & Real Estate Companies. *Journal of Economics, finance and Management Studies*, 7(7), pp. 4706-4713.
- Lee, H. K. J. & P. S., 2022. Advanced Gradient Boosting for Real Estate Price Prediction. *Proceedings of the AI in Real Estate Symposium*, 15(7), pp. 78-89.
- Menon, K., 2024. *Simplilearn.com*. [Online] Available at: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning> [Accessed 20 12 2024].
- Praise Peace, J. C. L. v., 2024. *Data Preprocessing for AI Models*, s.l.: Praise Peace.
- Ren, J. L. · W. F. · Y. S. · C., 2022. Assessing economic, social and environmental impacts on housing prices in Hong Kong: a time-series study of 2006, 2011 and 2016. *Journal of Housing and the Built Environment*, Volume 37, pp. 1433-1457.
- Revathy2, T. G., 2024. EMLARDE tree: ensemble machine learning based random de-correlated extra decision tree for the forest cover type prediction. *Signal, Image and Video Processing*, 18(2).
- Smith, J. & Z. Y., 2021. Enhancing Property Price Prediction Using Random Forests. *International Conference on Real Estate Analytics*, 9(4), pp. 123-135.
- Topraklı, A. Y. g., 2024. AI-driven valuation: a new era for real estate appraisal. *Journal of European Real Estate Research*.
- Yi Chen, Y. D. W. L., 2024. Prediction of Credit Default based on the XGBoost Model. *Proceedings of the 2nd International Conference on Machine Learning and Automation*, 96(1), pp. 85-92.