A Comparative Study of Machine Learning Algorithms for Property Price Forecasting

BY

Dehan Arif Fahim
ID: 201-15-3087
AND

Irfan Ahamed Nayem
ID: 201-15-3275

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Tapasy Rabeya
Lecturer(Sr.Scale)
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Lecturer(Sr.Scale)
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "A Comparative Study of Machine Learning Algorithms for Property Price Forecasting", submitted by **Dehan Arif Fahim** and **Irfan Ahamed Nayem** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-07-2024.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori — **Board Chairman**
Professor & Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan — **Internal Examiner 1**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md Mohammad Masum Bakaul — **Internal Examiner 2**
Sr. Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zulfiker Mahmud — **External Examiner**
Professor
Department of CSE
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Tapasy Rabeya, Lecturer(Sr.Scale), Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

13.07.24

Tapasy Rabeya
Lecturer(Sr.Scale)
Department of CSE
Daffodil International University

Co-Supervised by:

Mohammad Jahangir Alam
Lecturer(Sr.Scale)
Department of CSE
Daffodil International University

Submitted by:

Deham 13.07.24

Dehan Arif Fahim
ID: 201-15-3087
Department of CSE
Daffodil International University

Nayem 13.07.24

Irfan Ahamed Nayem
ID: 201-15-3275
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Tapasy Rabeya**, **Lecturer(Sr.Scale)**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

As the housing market grows, it's important to estimate pricing for both businesses and individuals. Nonetheless, a number of variables influence changes in home prices. Bangladesh is an overpopulated country, therefore a number of interrelated factors affect the price at which real estate is sold.The property's amenities, location, and size are crucial factors that could affect the cost.The objective of this investigation is to predict the prices of property in Bangladesh by employing a variety of machine learning algorithms. In order to guarantee precise predictions and robust model training, we assembled a comprehensive dataset of 18,835 property listings from Bproperty & Bikroy.com. Random Forest, Support Vector Machine (SVR), Decision Tree, XGBoost, CatBoost, and LightGBM are among the algorithms implemented in this investigation. The performance of these models was assessed using a variety of metrics, including R-squared, Mean Absolute Error (MAE), Root Mean Absolute Error (RMAC), and Mean Squared Error (MSE). The CatBoost and XGBoost models achieved the highest R-squared value of 91%, indicating superior accuracy, but XGBoost performed slightly better with less RMSE, MAE, MSE value. While the DecisionTrees model yielded the lowest R-squared value of 82%, indicating relatively poorer performance. The value of our findings and the potential for future research in property market analytics are underscored by the effectiveness of advanced machine learning models, particularly CatBoost, in predicting property prices within the Bangladeshi market. This information is valuable for stakeholders.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Overview

Our investigation concentrates on three significant regions: Chattogram, Gazipur, and Dhaka. Dhaka, the capital city of Bangladesh, is the economic and cultural centre of the country, and it is currently experiencing substantial population development and urbanization. The principal seaport of Chattogram is a significant factor in the dynamics of property, as it plays a critical role in trade and commerce. A suburban market that is expanding is represented by Gazipur, a swiftly expanding industrial area. Our research provides a comprehensive understanding of the real estate landscape in Bangladesh by examining these regions, which encompass a wide range of economic activities and demographic profiles. In our investigation, we implemented numerous machine learning algorithms, including Random Forest, Support Vector Machine (SVR), Decision Tree, XGBoost, CatBoost, and LightGBM. Each of these algorithms possesses unique assets and characteristics that are particularly useful for managing intricate datasets. Random Forest is an ensemble learning technique that enhances accuracy and regulates overfitting by constructing multiple decision trees and combining their outputs [1]. It is particularly effective in managing enormous datasets with a multitude of variables. A Support Vector Machine (SVR) is a supervised learning model that analyses data for regression and classification purposes. SVR is renowned for its efficacy in high-dimensional spaces and situations where the number of dimensions exceeds the number of samples [3]. A non-parametric model which employs a tree-like graph of decisions and their potential repercussions is known as a Decision Tree [16] . It is straightforward to comprehend and interpret; however, it is susceptible to overfitting.XGBoost is a gradient boosting library that has been optimised to be extremely efficient, flexible, and portable [18]. It is extensively employed due to its exceptional performance and rapidity.CatBoost is a gradient boosting algorithm that autonomously manages categorical features, thereby reducing the necessity for extensive preprocessing and improving accuracy. It is particularly effective when dealing with heterogeneous data.LightGBM is a gradient boosting framework that is extremely efficient and employs tree-based learning algorithms . It is intended to be distributed and efficient when dealing with large-scale data, surpassing other algorithms in terms of memory utilisation and speed [2].The results of our comprehensive literature review suggest that CatBoost and LightGBM generally outperform other algorithms in terms of accuracy. The preprocessing complexity is reduced by CatBoost's native ability to handle categorical data, while LightGBM's efficient handling of large datasets and fast training speed render it extremely suitable for real estate price prediction. Bproperty & Bikroy.com, a prominent real estate marketplace in Bangladesh, provided the dataset for our investigation. Providing a wealth of information for our analysis, it comprises 18,835 flat listings. In order to analyse the data and assess the efficacy of various machine learning models, we implemented Google Colab, a cloud-based platform. Root Mean Absolute Error (RMAC), Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared were the metrics employed for model evaluation. The R-squared value of 82% was the lowest for Decision Tree among the tested models, while XGBoost & CatBoost attained the highest R-squared value of 91%, indicating its superior accuracy. Here, we aim to leverage real-world data from Bangladeshi construction projects to develop and evaluate models capable of providing accurate price estimations during the crucial conceptual phase [10]. In summary, this research paper introduces a comprehensive methodology for forecasting property prices in Bangladesh that employs sophisticated machine learning techniques. We offer a reliable and accurate framework for property price predictions by employing multiple algorithms and spanning critical regions. We have discovered that CatBoost and XGBoost have the potential to improve the accuracy of forecasting, which provides valuable insights for stakeholders in the Bangladeshi real estate market. This research not only enhances the academic literature but also has practical implications for market participants, facilitating the development of a more stable real estate market and more informed decision-making.

1.2 Background and Present State

Property price prediction has long been a cornerstone of the real estate sector, guiding investment, urban planning, and market analysis. Conventional appraisal techniques frequently depended on the subjective assessment of appraisers, who took into account elements like amenities, property size, and location. These techniques are beneficial, but they could also be laborious and arbitrary, which could result in inconsistent property values. Property price prediction has undergone a paradigm shift as a result of the quick development of machine learning, which offers more precise, effective, and impartial methods. Regression models in particular, which enable the study of large datasets to reveal intricate patterns and relationships that are not immediately obvious to human analysts, have transformed the forecast of property prices. To produce accurate property price estimates, these models can combine a wide range of variables, such as economic indicators, demographic data, historical pricing data, and more. Models like CatBoost, XGBoost, LightGBM, Random Forests (RF), and Support Vector Regression (SVR) have

demonstrated a great deal of promise among the different machine learning approaches.
Yandex created CatBoost, which stands for Categorical Boosting, a gradient boosting method. It is especially made to deal with categorical data directly, which is a typical kind of data found in datasets pertaining to properties. Because of its ordered boosting method, which reduces overfitting, CatBoost is a very good predictor of real estate prices. The model is unique in that it can handle categorical variables without requiring a lot of preprocessing, which results in predictions that are more precise and effective. Because of its efficiency and speed, XGBoost, or Extreme Gradient Boosting, is one of the most used gradient boosting frameworks. Large datasets are handled with efficiency, and regularization parameters are included to help prevent overfitting. XGBoost has proven to be a dependable and accurate tool for predicting real estate prices through its widespread application. Its robustness and the extensive toolkit it provides for model optimization and evaluation are the main reasons for its appeal.
Another gradient boosting architecture with an emphasis on efficiency and speed is called Light Gradient Boosting Machine, or LightGBM for short. It finds the best splits using a histogram-based method, which drastically cuts down on computation time. LightGBM is a good option for applications involving property price prediction since it can manage enormous amounts of data efficiently. Data scientists love it because of its capacity to handle intricate relationships within the data while preserving computing efficiency. Several decision trees are built using the Random Forests ensemble learning technique, which then combines them to increase prediction accuracy. This model can handle both category and numerical data, and it is resistant to overfitting. Because of its dependability and interpretability, Random Forests are frequently employed to forecast real estate prices. Random Forests are predictive models that are accurate and stable, and they perform well on a variety of datasets by combining the predictions of several decision trees. One kind of Support Vector Machine (SVM) utilized for regression problems is Support Vector Regression (SVR). SVR looks for the hyperplane that minimizes prediction error and best fits the data. While it might be computationally demanding for bigger datasets, it works especially well for smaller datasets with intricate interactions. SVR is a useful technique for predicting property prices because of its strength, which is its high degree of precision in capturing underlying trends in data. The current status of study in machine learning regression models for property price prediction is dynamic and fast changing. These models are a great resource for real estate professionals, investors, and legislators as they strive to improve the precision and consistency of property appraisals. Property price prediction is becoming more and more feasible due to continuous developments in feature engineering, hybrid modeling, deep learning, geospatial analysis, and explainable AI.

The study aims to answer several critical questions in the field of "A Comparative Study of Machine Learning Algorithms for Property Price Forecasting" These questions serve as guiding principles, directing the research towards a thorough understanding of the complexities involved in using advanced artificial intelligence methods for better forecasting in the real estate industry. The overarching research questions are:The study aims to answer several critical questions in the field of "Property Price Prediction Using Advanced Machine Learning Techniques." These questions serve as guiding principles, directing the research towards a thorough understanding of the complexities involved in using advanced artificial intelligence methods for better forecasting in the real estate industry. The overarching research questions are:

This study compares the efficacy of machine learning algorithms to traditional statistical methods in forecasting real estate prices in Bangladesh. Important research questions are: Compared to traditional statistical methods, how well can different machine learning algorithms predict real estate prices in Bangladesh? By analyzing metrics like R-squared, MAE, and RMSE, this seeks to comprehend the potential of machine learning techniques—such as capturing complex and non-linear relationships—to produce predictions that are more accurate than those of more conventional techniques, like linear regression. In what ways do demographic variables impact property prices, such as population density and urbanization rates, and how can machine learning models take these factors into consideration? The research investigates the impact of changing demographics on property values and incorporates these factors into prediction models to improve accuracy. How does the accuracy of Bangladeshi real estate price projections improve with the inclusion of temporal data such as historical price trends and seasonal variations? This inquiry investigates how temporal data, which captures recurring and long-term patterns, can increase forecasting accuracy. What policy ramifications might the Bangladeshi government face from precise property price forecasting, and how might these insights affect housing and urban development plans? In order to achieve a stable and open real estate market, the study evaluates how accurate forecasts can influence housing policies, urban development, and infrastructure investments. What are the primary determinants of real estate prices in Bangladesh, and how can these be incorporated into machine learning models to enhance prediction accuracy? In order to improve prediction accuracy, the study identifies important economic, demographic, and property-specific factors influencing prices and incorporates them into models. What regional variations are there in the way the machine learning models predict Bangladeshi real estate prices? The study looks into how regional economic and demographic factors impact model performance and makes recommendations for region-specific modifications to improve accuracy. What is the impact of economic factors on

Bangladeshi real estate prices, such as GDP, inflation, and interest rates, and how can these factors be factored into prediction models? This inquiry looks at how macroeconomic factors affect real estate prices and incorporates them into models to provide forecasts that are appropriate for the given context. How can the unique issues facing the Bangladeshi real estate market—such as limited supply and poor data quality—be addressed to increase the effectiveness of machine learning techniques? In spite of these obstacles, the study devises methods for optimizing machine learning models, placing special emphasis on strong data preprocessing, data augmentation, ensemble learning approaches, and utilizing local knowledge to enhance model performance.

1.3 Problem Statement

The real estate market is complicated, influenced by a variety of elements including economic conditions, geographic location, property attributes, and market trends. Conventional property valuation techniques frequently produce inconsistent and inaccurate results since they mainly rely on the subjective assessment of appraisers and manual analysis. These traditional methods can be expensive, time-consuming, and prone to human error, which presents serious difficulties for investors, legislators, and real estate experts. The creation of more precise and effective property price prediction models has been made possible by machine learning, which offers a promising answer. But in spite of the possible advantages, a number of problems still need to be fixed. Choosing the best machine learning algorithms, integrating and managing various data kinds, and maximizing model performance to provide solid and trustworthy predictions are some of the main issues.

The procedure becomes even more difficult when gathering data sets from resources like Bproperty & Bikroy.com. The ongoing volatility of real estate values and the dynamic nature of the economy make the data collection challenge substantial. Numerous variables, such as interest rates, market demand, and economic policies, all of which are subject to sudden changes, affect real estate values. It is difficult to keep an accurate and up-to-date dataset because of this instability. Moreover, insufficient or missing data and inconsistent data formats might make it more difficult to integrate and preprocess the data needed for machine learning models.

In order to solve these issues, this study compares how well different machine learning regression models forecast real estate values. The research will specifically concentrate on sophisticated models like Support Vector Regression (SVR), Random Forests, XGBoost, LightGBM, and CatBoost. With regard to handling numerical and categorical data, capturing intricate, non-linear relationships, and managing big datasets, every one

of these models has certain advantages. Through a methodical comparison of different models, the study aims to determine the best algorithms for predicting property prices. The study will also identify the crucial elements influencing real estate values and create a thorough model that may produce precise and useful findings.
The ultimate objective is to improve property values' dependability and accuracy for the benefit of real estate market participants. This research intends to make a substantial contribution to the field of property price prediction by addressing the present shortcomings of conventional methods and utilizing the capabilities of machine learning. It will provide useful solutions and insightful information for real estate professionals and decision-makers.

1.4 Objectives

The objectives of this research are critical for addressing existing issues in property price prediction and improving the accuracy and reliability of appraisals using sophisticated machine learning techniques. One of the main goals is to assess how well machine learning models—like Support Vector Regression (SVR), Random Forests, XGBoost, CatBoost, and LightGBM—predict real estate values. The purpose of the study is to evaluate these algorithms' resilience, efficacy, and efficiency in handling big, complicated datasets in order to determine which ones work best for predicting property prices. The study also aims to create a thorough prediction model that takes into account a number of significant factors, such as location, property characteristics, market trends, and economic indicators. Another objective is to address data collection challenges, with a focus on investigating and suggesting solutions for problems pertaining to the collection and analysis of real estate data from sites like Bproperty and Bikroy.com. These solutions include handling continuous price changes and guaranteeing data accuracy and consistency. In addition, the study looks at the main variables that affect real estate values and how they affect machine learning models' ability to predict the future. Real estate agents, investors, and regulators stand to gain from the research's goal of improving property appraisal accuracy and dependability through the application of cutting-edge machine learning techniques. In the end, the research will add to the body of knowledge about the real estate market by offering a comprehensive analysis of the advantages and disadvantages of different machine learning models in predicting property prices and by enabling real-world applications that will improve the efficiency of investment analysis, property appraisal, and decision-making.

1.5 Scope & Limitations

This study focuses on developing and evaluating sophisticated machine learning regression models for predicting real estate prices, with a focus on key areas of research and application. The study begins by discussing the challenges associated with data collection, particularly handling continuous price adjustments and guaranteeing accuracy and consistency in data obtained from websites such as Bproperty and Bikroy.com. The study intends to raise the caliber of the data used in the models by examining and offering solutions to these issues. Second, in order to improve the predictive accuracy of the machine learning models, it examines significant factors influencing real estate values. Through the use of cutting-edge machine learning techniques, the project aims to increase predictive accuracy and reliability, which will improve property appraisals and benefit investors, regulators, and real estate professionals. Furthermore, the study advances our understanding of the real estate market by providing a thorough examination of several machine learning models, including Random Forests, XGBoost, CatBoost, LightGBM, and Support Vector Regression (SVR), and rating them according to their robustness, dependability, and predictive ability.The study builds a large dataset for model training and evaluation by integrating a variety of data sources, such as historical real estate prices, economic indicators, regional data, property attributes, and market trends. To improve model accuracy, feature engineering will add new variables that highlight characteristics of real estate, such as proximity to amenities, historical price movements, and socioeconomic indicators. The goal of the research is to create a dependable model that can handle big and complicated datasets for property price prediction by integrating the best aspects of multiple machine learning approaches. In order to hone and increase accuracy, it will also examine the variables affecting real estate values and how they affect the models' forecasts. Notwithstanding, the study recognizes certain constraints, including those related to data availability and quality, the influence of dynamic market conditions on data updates, the computational requirements of intricate models, their applicability to diverse markets, and the models' interpretability. The goal of the study is to offer helpful insights and workable solutions for property price prediction using sophisticated machine learning regression models by addressing these limitations and concentrating on the defined scope.

1.6 Report Organization

- **Chapter 1: Introduction**

The introductory chapter establishes the context for the research, outlining the importance of accurate property price prediction, the use of machine learning techniques (Catboost, Random forest, SVR, and LightGBM), and the need for a thorough examination of continuous data processing methods. It provides an overview of the study's goals, research questions, and general framework.

- **Chapter 2: Literature Review**

The background chapter provides a thorough review of relevant literature and prior research. This section provides a thorough overview of the methods, theoretical underpinnings, and technological developments pertaining to machine learning, regression models, and property price prediction. It lays out the background for the current investigation and identifies knowledge gaps.

- **Chapter 3: Methodology**

The research methods chapter describes the approach used to conduct the study. It sheds light on the model architecture, data gathering techniques, research design, and the reasoning behind particular methodology selections. The ethical issues and potential constraints on the investigation are also covered in this chapter.

- **Chapter 4: Implementation**

This chapter provides the experimental findings of using machine learning and regression models to predict property prices. Comprehensive evaluations of the effectiveness of various segmentation and prediction techniques are offered, accompanied with statistical metrics to verify the results.

- **Chapter 5: Result and Analysis**

In our study on predicting property values in Bangladesh using machine learning, we examined a dataset of 18,835 property listings from Bproperty & Bikroy.com using several algorithms. With an astounding R-squared value of 86%, LightGBM stood out as the best performer, demonstrating its exceptional capacity to detect and forecast changes in real estate prices. Strong performance was also shown by several algorithms, including Random Forest, XGBoost, and CatBoost, which showed reliable prediction abilities in this situation. Scatter plots and error distribution plots, among other visual aids, demonstrated the models' efficacy and offered lucid insights into their prediction accuracies. These findings have important ramifications for real estate industry players, providing useful information for well-informed choices on real estate investments and pricing tactics in Bangladesh's dynamic marketplace.

- **Chapter 6: Impact on Society, Environment and Sustainability**

The societal impact chapter delves into the larger implications of the research findings on the real estate market and economic stability. It talks about how more accurate techniques for predicting property prices can help with investment choices, housing affordability, and market transparency. Potential technological developments and their social ramifications such as more equal access to housing, improved personal financial planning, and general economic growth—are taken into account.

- **Chapter 7: Conclusion and Future Work**

This final chapter provides a summation of the full study. It provides recommendations for real-world applications and more research in addition to summarizing the main results and restating the research's conclusions. A path for academics interested in building upon the current study is provided in the discussion of the implications for future research.The reader is guided through the study process from the introduction to the wider consequences and recommendations by this organised layout, which guarantees a logical flow of information. It makes it possible to get a thorough grasp of the study methodology and how it could affect the theoretical and applied facets of property price prediction using regression models and machine learning.

1.7 Summary

Our research looks on property price forecasting in Bangladesh, namely in Dhaka, Chattogram, and Gazipur, utilizing advanced machine learning techniques such as Random Forest, Support Vector Regression (SVR), Decision Tree, XGBoost, CatBoost, and LightGBM. We examine a set of 18,835 postings from Bproperty & Bikroy.com and assess models using measures such as R-squared, MAE, and MSE. XGBoost & CatBoost showed higher accuracy with the greatest R-squared value of 91%. But XGBoost has lower RMSE, MSE, and MAE values, making it the best-performing model among all algorithms. Large datasets and categorical data are handled with efficiency by CatBoost and XGBoost, making them stand out. Our research highlights the useful uses of machine learning in enhancing property price forecasts and market stability, offering significant insights to Bangladeshi real estate market participants.

CHAPTER 2

Literature Review

2.1 Overview

Upon examining numerous research papers, it has become apparent that the necessity for precise and efficient property price prediction methods has been significantly increased in Bangladesh as a result of the accelerated urbanization and economic expansion during the past few years. Conventional methods of property valuation can be laborious, inconsistent, and prone to human error because they frequently depend on the subjective opinions of appraisers. This has prompted researchers to investigate cutting-edge machine learning algorithms as a potential means of raising the precision and effectiveness of real estate price forecasts. Numerous studies have shown that different machine learning algorithms perform well in this situation. The ability of algorithms like Random Forest, Support Vector Regression (SVR), Decision Tree, XGBoost, CatBoost, and LightGBM to handle a variety of datasets and capture the intricate linkages seen in real estate markets has been demonstrated. An ensemble learning technique called Random Forest builds several decision trees and combines their results to improve prediction accuracy. Because SVR is effective in high-dimensional spaces, it can be used to predict property values in areas with a wide range of economic and property attribute situations. The Decision Tree model is a non-parametric model that is easy to comprehend and analyze, but it can overfit. Because of its outstanding speed and performance, XGBoost is a highly recommended gradient boosting library that is both flexible and efficient. Yandex's CatBoost algorithm reduces overfitting by handling category data directly. LightGBM is well-known for its speed and efficiency. It works well with large-scale property datasets because it minimizes calculation time using a histogram-based method.

2.2 Related Works

In this study [1], S. Sapakova et al. (2024) looked at Almaty, Kazakhstan's rental property valuation potential using machine learning. This study provides insightful information about the distinctive features of the Almaty market, emphasising its vulnerability to both internal and external variables as well as speculative forces. Accurate appraisal can be difficult due to the enormous impact these factors can have on property values.

Sapakova (2024) concentrated on creating a predictive model utilising machine learning methods in order to handle these complications. Support vector machines (SVM) with a Gaussian kernel, decision tree regression, random forest regression, lasso regression,

ridge regression, and linear regression were among the methods whose performances were assessed in the study.
According to their research, the best algorithm was Random Forest Regression, which predicted rental costs with an astounding 90% accuracy. This shows that Random Forest provides a potent method for grasping the nuances of the renting market in Almaty.
IN [2], J. A. Bastos et al. (2024) explored the importance of uncertainty quantification in automated property appraisals, specifically in the San Francisco Bay Area, utilising a unique technique known as conformal prediction. Unlike conventional non-conformal approaches, this approach provides a consistent range of possible property values while guaranteeing valid coverage in finite datasets. The study examined several algorithms for producing prediction intervals with precise coverage. Gradient Boosting Machines outperformed Linear Regression with a coverage of 0.82, Random Forest with a substantial improvement of 0.801, and Linear Regression with 0.683. These findings demonstrate how important it is to take uncertainty into account when projecting real estate prices.
In [3], There are particular difficulties in valuing properties in the real estate market in tourist areas. Conventional approaches are prone to speculative effects and subjective biases. T. Alkan et al. (2023) investigated the potential of machine learning (ML) methods for addressing these concerns in Antalya's Alanya area. This study investigated the accuracy of three machine learning (ML) algorithms in predicting real estate values: Support Vector Machines (SVM), k-Nearest Neighbours (kNN), and Random Forest (RF). The data used in this research came from an online real estate website.
The study's main objective was to determine the structural and spatial aspects that significantly impact property value in a market driven by tourism. Alkan (2023) highlighted the benefits of machine learning (ML) approaches above classical mathematical models, emphasizing their capacity to produce unbiased and scientific values. Their results showed that, with an accuracy of 0.7354, SVM performed better than kNN (0.6621) and Random Forest (0.6962). This study shows how ML may provide objective, data-driven values to balance the real estate market in tourist destinations.
In [4], M. Khasravi et al.(2022) looked at roughly 500 homes in the Boston area and explored the potential variations in property price growth based on each of the contributing factors. A comparison of all the models' accuracy is done with the dataset's various machine learning (ML) regressors.
In certain cases, though, the lack of data has resulted in subpar model performance. According to the results, the random forest received the highest rating of 0.88, while the voting regressor received the second-highest rating of 0.87. The percentage of the population with a lower socioeconomic class and the average number of rooms had the

biggest effects on price range estimates, according to the results of multivariate exploratory data analysis.

In [5], Y. Ming et al. (2020) used RandomForestRegressor, XGBoost, and LightGBM to predict and fit 12 key features. They conducted a comparative analysis of the results from all three models. The results showed that the prediction effect of the XGBoost model was higher than that of

of the other two models, contributing significantly to Chengdu's research on housing rent prediction. In their paper, Satish, G. N. et al. (Satish, 2019) suggested that machine learning could be extended to real estate-related sectors in order to accurately forecast housing prices. Gradient Boosting, Lasso Regression, and Linear Regression were the machine algorithms employed. Based on the data, Gradient Boosting is the algorithm that performs the best, with an accuracy of 91.14%, while Lasso and Linear Regression yield nearly identical results, with 76.14% and 76.15% of accuracy, respectively. They recommended using parallel computing in future work to reduce processing times.

In [6], S. Lu et al. (2017) used a hybrid regression technique to predict house prices. Gradient Boosting, Lasso, and Ridge were the machine learning algorithms that were considered for the implementation. They made the decision to put into practice a hybrid Lasso and gradient boosting algorithm.

They found that 65% Lasso and 35% Gradient Boosting outperformed all other hybrid regression model combinations, leading them to conclude that a hybrid regression algorithm outperforms a regression algorithm used separately. Modi For predicting house prices, Modi et al. (Modi, 2020) suggested using ensemble machine learning techniques. 22 attributes make up the dataset from Kaggle.com that was used in this application, and several algorithms have been trained and tested using it. Logistic regression, support vector machines (SVM), naive bayes, extra tree, stochastic gradient descent, K nearest neighbors, and an ensemble machine learning algorithm that employs all of the aforementioned algorithms as weak learners were the algorithms that were employed. Several performance metrics were used to compare the performance of these algorithms, and the results showed that the ensemble machine learning algorithm performed the best overall, followed by extra tree, logistic regression, and naive bayes.

In [7], D. Kansara et al. (Kansara, 2018) used a stacked regressor created by combining regression techniques such as Multiple Linear Regression, Random Forest Regressor, and XGBoost. The final output is produced by the stacked regressor, which balances the strengths and weaknesses of each algorithm to increase overall accuracy. Error rates are reduced and error rates are balanced because of the rigorous cross-validation that is used to balance out the shortcomings of one model with the strengths of another (Güneş, 2017).

In [8], Y. Kang et al. (2021) introduced a data-fusion framework that uses multi-source big geo-data and machine learning to forecast house price appreciation. The study integrates several datasets, including house attributes, photos, and socioeconomic factors, and analyses over 20,000 houses in Greater Boston. The gradient boosting machine achieved an R-squared value of 0.74 at the neighborhood scale, indicating high accuracy. Smaller, less expensive homes frequently have greater potential for appreciation, providing insightful data for real estate market research and policy-making.

In [9], J. Avanija et al. (2021) introduced, "House Price Using XGBoost Regression Algorithm," creates a centralised system for predicting house prices based on variables such as location, neighbourhood, and amenities. In order to accurately estimate house values, the study employs the XGBoost regression technique, emphasising data preprocessing and one-hot encoding for categorical attributes. This model offers a strong framework for predicting real estate prices, assisting clients in making well-informed decisions about when and where to purchase a home.

The lack of high-resolution data creates significant challenges for municipal solid waste (MSW) management in urban areas. In order to address this [11], C. E. Kontokosta et al. (2018) study predicts building-level MSW generation in New York City using machine learning and small area estimation. The study applied gradient boosting regression trees (GBRT) and neural networks using daily DSNY data, building attributes, socioeconomic factors, and weather; GBRT achieved an R-squared value of 0.81. By enabling equitable waste management policies and optimized waste collection routes, this approach highlights the potential of advanced analytics to improve urban MSW systems and advance sustainability.The lack of high-resolution data creates significant challenges for municipal solid waste (MSW) management in urban areas. In order to address this, it predicts building-level MSW generation in New York City using machine learning and small area estimation. The study applied gradient boosting regression trees (GBRT) and neural networks using daily DSNY data, building attributes, socioeconomic factors, and weather; GBRT achieved an R-squared value of 0.81. By enabling equitable waste management policies and optimized waste collection routes, this approach highlights the potential of advanced analytics to improve urban MSW systems and advance sustainability.

Predicting house prices accurately is crucial in real estate and finance, often using tree-ensemble models like LGBM and XGBoost. LGBM, with a leaf-wise growth strategy, reduces loss but can overfit, while XGBoost, using a level-wise strategy, reduces overfitting through regularization but has higher computation time. In [12], R. Sibindi et al. (2023) introduced a hybrid LightGBM-XGBoost model, optimized using Bayesian hyperparameter techniques to minimize variance and enhance accuracy. Compared against LGBM, XGBoost, Adaboost, and GBM, the hybrid model outperformed them with MSE, MAE, and MAPE values of 0.193, 0.285, and 0.156, respectively, highlighting its robust predictive capability.

In [13], D. Rey-Blanco et al. (2024) introduces a computer algorithm-based methodology for creating optimal car and walk accessibility indices, addressing index construction and variable selection challenges. Principal components analysis is used in the study to ensure orthogonality and provide a clear interpretation of these indices. The methodology reduces spatial autocorrelation by 35% and increases prediction accuracy by 13% in regression models and 21.6% in random forests using data from Madrid. This method can be customized for different real estate types and locations and offers substantial advancements in urban analysis.

In [14], A. Varma et al. (2018) focuses on Mumbai as the primary location and predicts real-time house prices in various localities. The system makes maximum use of boosted regression, forest regression, and linear regression. The algorithm is now much more efficient thanks to the use of neural networks.

In [15], J. Manasa et al. (2020) highlights that Bengaluru is rapidly growing in both area and population, but the characteristics of the data set are too small to accurately represent Bengaluru real estate prices. A prediction model for evaluating the price has been attempted, based on the factors that impact pricing. Regression techniques used in modeling explorations include Lasso and Ridge regression models, support vector regression, multiple linear regression (Least Squares), and boosting algorithms like Extreme Gradient Boost Regression (XG Boost).

In [16], Alen Vlahovljak (2022) conducted a study on house price prediction in the United Kingdom, using a dataset of 22,489,348 records from 1995 to 2017. The study utilized Decision Tree Regressor and Multivariable Linear Regression algorithms for data analysis. The Decision Tree Regressor was responsible for the highest R2 value of 75.91% in this study, indicating that it was able to handle the dataset's complex, non-linear relationships with ease. The results highlight how crucial it is to use suitable machine learning models in order to predict housing prices accurately.

2.3 Comparison between existing works

The comparative analysis conducted in this study on "A Comparative Study of Machine Learning Algorithms for Property Price Forecasting" is a critical component, shedding light on the subtle differences and synergies between various methodologies. This research thoroughly assesses the effectiveness of various machine learning regression models in predicting property prices, taking into account both sophisticated ensemble techniques and conventional linear approaches. By means of rigorous experimentation

and analysis of experimental outcomes, the study reveals valuable information regarding the advantages and drawbacks of every methodology, offering a thorough comprehension of their individual contributions to the process of prediction.
A detailed investigation of regression models' contributions to improving the precision and effectiveness of real estate price prediction is made possible by their in-depth study. The study adds significant knowledge to the field by comparing the results of various approaches, helping researchers and practitioners choose the best tactics for their particular situations. Not only does the study highlight these discrepancies, but it also highlights how distinct regression models may work well together, adding to our knowledge of the variables driving real estate values and the relative merits of different forecasting approaches.
To sum up, this research's comparative analysis provides a crucial framework for evaluating the effectiveness of machine learning regression models in predicting real estate prices. It highlights the value of a comprehensive approach in real estate analytics and presents a nuanced viewpoint that combines traditional and modern methods. The results provide useful insights for improving currently used predictive methodologies in addition to advancing our understanding of property price prediction. The context for the next chapters is established by this comparative analysis, which also offers a strong basis for the discussion of the societal impacts, suggestions, and implications for further research.

TABLE 2.3.1: Comparative Analysis Table

Study	Data	Model	Accuracy & Best Model
Our Study	18,835	Support Vector Machine, Decision Tree, Random Forest, LightGBM, XGBoost, CatBoost	CatBoost=.91 XGBoost = .91
T. Alkan et al. 2023	200	Support Vector Regression, k-nearest neighbors, Random Forest	Support Vector Regression=0.7354
J. A. Bastos et al. 2024	8249	Random Forest, Gradient boosting machine	Gradient boosting machine=0.82
S.Sapakova et al. 2024	3882	Linear Regression, Elastic Net Regression,	Random Forest=0.90

		Random Forest	
Alen Vlahovljak (2022)	22,489,348	Decision Tree Regressor Multivariable Linear Regression	Decision tree regressor=75.91%
Y. Ming et al. (2020)	33224	Gradient boosting machine,Linear Regression,Lasso	Gradient boosting machine=91.14%
Ho. K.O. Winky et al. (2020)	39,554	Support vector machine, Random forest, Gradient boosting machine.	Random forest=0.90, Gradient boosting machine=0.90
A. G. RAWOOL et al. (2021)	404	Random Forest	Random Forest=0.76
C. Li (2024)	21,611	Linear regression ,Random forest ,Neural network ,XGBoost	XGBoost=0.888
A. B. Adetunji et al. (2022)	506	Random Forest	Random Forest =0.90
A. Yağmur et al. (2022)	900	Artificial Neural Networks , Multiple Linear Regression , Support Vector Regression	Artificial Neural Networks =0.827
M. P. Nwankwo et al. (2023)	19774	Ridge Regression, Gradient Boosting Regression, AdaBoost Regression, Elastic Net, Multiple Linear Regression , LASSO	Gradient Boosting Regression
C. Zou(2023)	9,134	Multiple Linear Regression, Random Forest, CatBoost	Random Forest= 0.89
X. Ouyang (2024)	546	Multiple Linear Regression, Random Forest	Multiple Linear Regression=0.73
A.Jangaraj et al. (2021)	2021	XGBoost Regression Algorithm	XGBoost Regression Algorithm=0.80

D. Patel et al. (2023)	N/A	Linear Regression, Lasso Regression, Ridge Regression, Support Vector Regression , XGBoost Regression	XGBoost Regression= RMSE of 12.34
H. Li (2023)	1420	Random Forest XGBoost	XGBoost=0.89
J. Liu(2024)	N/A	Random Forest, Linear Regression	Random Forest
Z. Liu (2023)	1259	Linear Regression, Lasso Regression, Ridge Regression	Lasso Regression=0.90
W. Weng (2022)	1460	Decision Tree, Random Forest, Adaptive Boosting , Gradient Boosting Decision Tree , Extreme Gradient Boosting (XGBoost)	Extreme Gradient Boosting (XGBoost)
Z. Chen (2024)	N/A	Convolutional Neural Networks , Decision Trees, K-Nearest Neighbors	Convolutional Neural Networks
Y. Chen (2023)	N/A	Linear Regression, Random Forest, XGBoost , Support Vector Machine	Random Forest
A. M. A. Kaushal et al. (2024)	N/A	Linear Regression, Logistic Regression	Linear Regression

2.4 Open Issues

Our work on A Comparative Study of Machine Learning Algorithms for Property Price Forecasting takes into account a number of essential factors in order to achieve high predicted accuracy. To improve these models' efficacy and dependability, a few unresolved concerns must be resolved. The availability of comprehensive and high-quality statistics is one of the main obstacles, since property pricing data are frequently missing, out-of-date, or fragmented. Ensuring completeness and integrity of data is essential for precise model evaluation and training. Furthermore, regional regulations, infrastructural development, and local economic conditions can all have a substantial impact on property prices. The prediction models must take these regional differences into account in order to become more accurate and relevant. Model performance can be improved by identifying and choosing the most important elements (such as location, property size, number of rooms, and accessibility to facilities) and by adding new features through feature engineering, but these tasks need considerable thought and knowledge. While a variety of machine learning algorithms, such as Decision Trees, XGBoost, LightGBM, AdaBoost, and Linear Regression, can be used to predict property prices, it is still unclear which algorithm is best suited for Bangladesh's unique circumstances because each algorithm has advantages and disadvantages. The presence of outliers and missing values in real estate data might distort the outcomes of prediction models, necessitating the development of reliable techniques to deal with these abnormalities. Seasonal changes, market trends, and economic cycles are examples of temporal elements that affect property prices. While adding these temporal dynamics to the models can increase their forecast accuracy, it can be difficult because of the complexity of the data. In addition, scalable and easily deployable property price prediction models are necessary for the models to handle big datasets and be incorporated into current real estate platforms, which is essential for the models to be helpful in real-world applications. Since it's critical that stakeholders in the real estate industry comprehend the methods used to make predictions, creating explainable models that shed light on the variables affecting real estate values helps increase user confidence and usability. Last but not least, the application of predictive models in real estate may have important ethical and policy ramifications, including affecting the affordability of housing and market behavior. To guarantee the ethical application of machine learning in the forecast of real estate prices, several factors must be taken into account. Our work intends to construct strong and trustworthy property price prediction models by tackling these outstanding challenges, which would enable well-informed decision-making in Bangladesh's real estate market.

2.5 Summary

Our work on A Comparative Study of Machine Learning Algorithms for Property Price Forecasting aims for high predictive accuracy, however various challenges remain unresolved. Among them is the availability of thorough and high-quality data, which is frequently lacking, out-of-date, or dispersed. Ensuring the quality and integrity of data is essential for accurate training and model validation. For the models to be more accurate and relevant, regional variances brought about by local economic circumstances, infrastructure development, and policies must be included. While feature engineering and critical feature identification can enhance model performance, they do so with caution and knowledge. It is still difficult to determine which machine learning algorithm is most appropriate for Bangladesh's particular setting because each algorithm has advantages and disadvantages. In real estate data, handling outliers and missing numbers is essential to prevent skewed outcomes. While it increases complexity, temporal considerations like market patterns, economic cycles, and seasonal variations can help forecast accuracy. For practical applications, scalable and deployable models are necessary, and developing explainable models improves usability and stakeholder trust. To guarantee the responsible use of machine learning in property price prediction, ethical and policy consequences, such as housing affordability and market behavior, must also be taken into consideration. Our work intends to address these problems in order to provide solid and trustworthy models for wise decision-making in Bangladesh's real estate market.

CHAPTER 3

Methodology

3.1 Overview

The goal of this study on " A Comparative Study of Machine Learning Algorithms for Property Price Forecasting" is to use advanced machine learning techniques to make accurate and trustworthy property price predictions. In order to achieve this, we painstakingly gathered our dataset from bproperty and bikroy.com, two reputable and well-known websites in Bangladesh. These websites are well-known for having extensive real estate listings, which makes them perfect resources for our data gathering initiatives. Our predictive modeling is based on the dataset, which was saved in CSV format for convenience of processing and manipulation.We used Google Colab, a potent cloud-based platform that offers the required computing power and adaptability for machine learning activities, to carry out our study and train our model. We were able to handle and process the dataset effectively thanks to this platform, which also made sure that our predictions were supported by reliable and scalable computations. A vast array of characteristics included in the dataset are essential for precise property price forecasting. These characteristics include the following: location, cost, number of bathrooms and bedrooms, area (measured in square feet), kind of property (apartments particularly), longitude, latitude, sub-region, region, quantity of parking places, and elevator availability. The dataset also has a reference column whose purpose is to verify the integrity and origin of the data. Notably, this reference column was removed from the dataset before any predictions were made, ensuring that it would not affect the model's performance during the training or testing stages of our investigation. The location, number of bedrooms, number of bathrooms, area, kind of property, longitude, latitude, sub-region, region, number of parking spaces, and elevator presence were the values we used for our model's training phase. These characteristics were chosen for their importance and possible influence on real estate values. The goal variable, represented by the price column, was set aside specifically for evaluating the predictive power of the model. Our study attempts to represent the complex and multidimensional character of property pricing in Bangladesh by concentrating on these particular features. Geographical coordinates, or longitude and latitude, are included in the model to account for fluctuations in pricing based on location, while variables like the number of bedrooms, bathrooms, and overall area reveal details about the physical aspects of the property. The existence of amenities like parking spaces and elevators further improves the model's ability to estimate property prices properly. The sub-region, region

characteristic also helps the algorithm better comprehend local market circumstances.
In conclusion, this study approach guarantees that our machine learning models are trained on an extensive and pertinent dataset, allowing us to estimate property prices with accuracy and dependability. In order to confirm our commitment of advancing technologies we used Python for model training. Gathering, cleaning and creating a final datasetWe try to solve problems predicting property price. We want to share some thoughts with you about the Bangladesh real estate market. This approach is a demonstration of the commitment we have to employing state-of-the-art methods and technologies in order that our research yields accurate, meaningful findings. These to-the-point contributions are expected to considerably contribute in the realms of property price prediction undertaking over Bangladesh.

3.2 Proposed Methodology

The process for estimating property prices in Bangladesh using machine learning includes several essential stages to provide accurate and dependable findings. The first stage of the data collection process is compiling extensive information from many sources, such as Bporperty & Bikroy.com on property pricing and associated characteristics, such location, size, number of rooms, and age. For the purpose of capturing the several elements impacting property prices, this diverse data is essential.The data preprocessing step, which comes after data collection, makes sure the dataset is suitable for machine learning models and of high quality. In order to prevent data gaps from impacting the quality of the model, this method involves handling missing values, encoding categorical variables to convert non-numeric data into a numerical format, and normalizing or standardizing numerical characteristics to guarantee consistency across various scales. After that, the data is divided into training and testing sets so that the model can be trained and then evaluated. During the model training phase, various machine learning algorithms, including Decision Trees, Random Forest, LightGBM, CatBoost, SVR, and XGBoost, are trained using the preprocessed training data. In order to accurately estimate property prices, each model gains an understanding of the patterns and correlations present in the training data.
The performance of these trained models on the testing data is then assessed during the model testing phase. To evaluate accuracy, the models' predictions are compared to the actual values. To give a thorough picture of each model's predictive power, this assessment makes use of a number of performance measures, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared value. Ultimately, the best-performing model is chosen during the result analysis phase by using the assessment measures. The selected model is then applied to forecast newly gathered, unobserved data, guaranteeing that the process yields precise

and trustworthy estimates of real estate prices. This methodical process guarantees the creation of a strong model that can accurately forecast Bangladeshi real estate prices.
Here's a flowchart to represent proposed methodology:

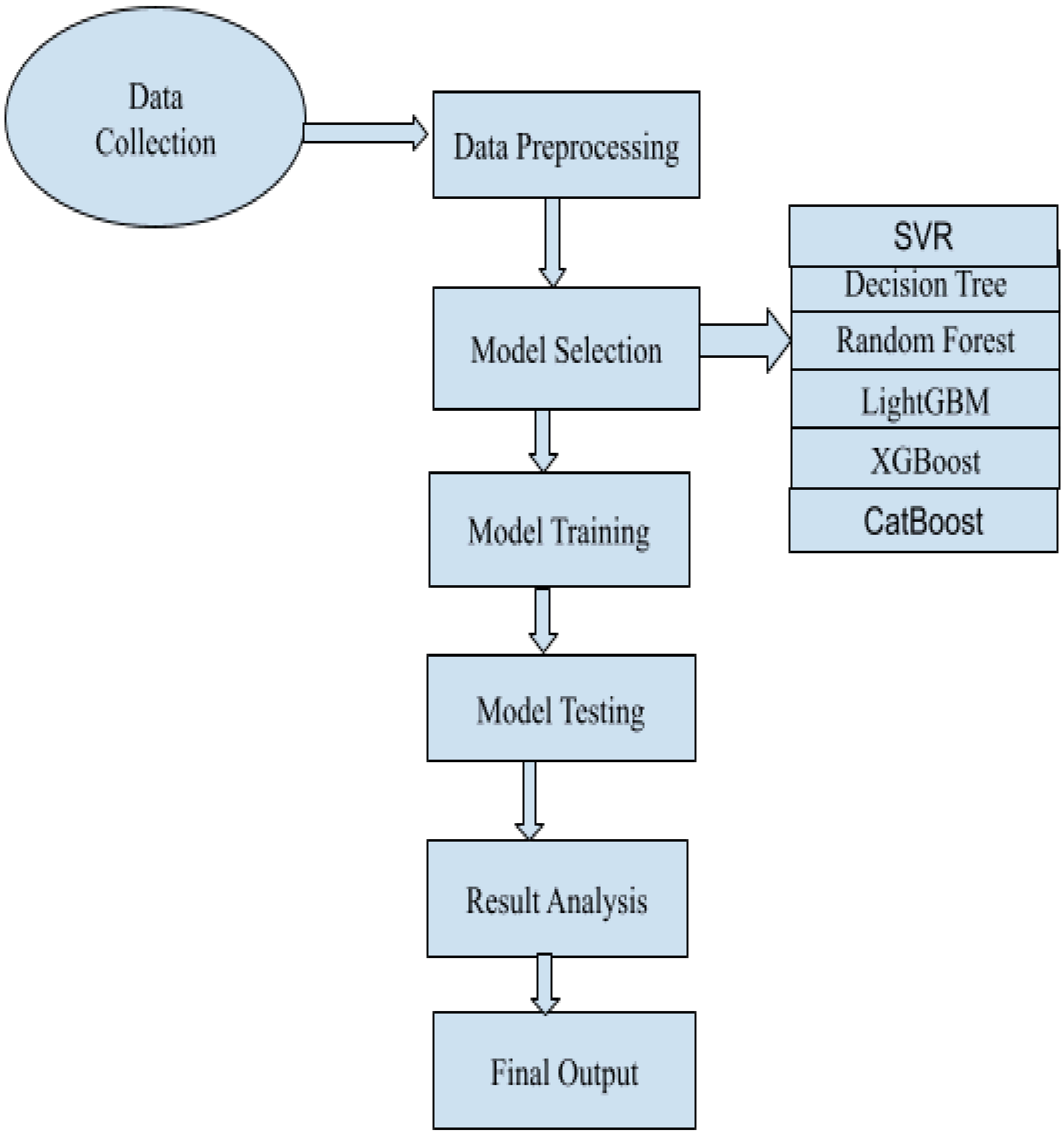

Figure 3.2.1: Flowchart of Proposed Methodology

3.2.1 Data Collection:

We used a raw dataset of the apartment prices to predict. Dataset collected from Bproperty and Bikory sites for our research. The data set includes a list of apartments located in different regions and having metro rail connectivity today, as well projections that they would be connected within the next 3-4 years. It also has multiple parameters that relate to those listings, as shown in Table 3.2.1.

TABLE 3.2.1: Apartment price parameters, description & the data types

Parameter	Description	Data type
No. of parking	Number of parking spaces	Integer
Location	Property location	String
Price	Property price	Integer
Type	Property type	String
No. of beds	Number of bedrooms	Integer
No. of baths	Number of bathrooms	Integer
Area	Property area in sq ft	Integer
Longitude	Geographic longitude	Float
Latitude	Geographic latitude	Float
Region	Geographic region	String
Sub-region	Geographic sub-region	String
No. of elevator	Elevator availability	Integer

The apartment listing dataset includes 12 features used to predict the price of an apartment, along with one target variable that contains up to 18,835 samples. The dataset collected from Bproperty and Bikroy websites.Important Parameters are used to ensure accurate prediction. We will try to get improvements in our price prediction model by adding these 12 key parameters.

3.2.2 Pre-processing dataset:

During this study, data preprocessing is very important in order to make such datasets more credible and consistent. Used a few data pre-processing techniques. Apply median to deal with missing or null values & machine learning algorithms cannot work with categorical features. It works only with numerical features. For this reason, we used one-hot encoding and label encoding to convert numerical data from categorical data on those columns. There were a few categorical columns in our dataset of 18,835 samples at "Region", "Location", and "Sub-Region". We changed these categorical values into numerical values using one-hot encoding to make them ready for machine learning models. Each category became a column so this process gave me 388 new columns. The values in one-hot encoding are 0 and 1 where a value of 0 means 'false' and a value of 1 represents 'true'.

3.2.3 Machine learning models:

Machine learning models are algorithms that learn patterns and make predictions or decisions without requiring explicit programming. Data is used by these models to train and gradually enhance their performance. They allow computers to learn from experience and they are widely used in the area of prediction and various other sectors.

Decision Tree:

Decision Trees are a non-parametric supervised learning method used for classification and regression. They predict the target variable by learning simple decision rules inferred from the data features.

Prediction $= \sum(wi \cdot xi)$

where wi are the weights assigned to each feature xi .

Random Forest:

In order to increase prediction accuracy and reduce overfitting, Random Forest is an ensemble approach that builds several decision trees during training and outputs the mean prediction of the individual trees.

Equation: $y^\wedge = N1\sum i = 1Nhi(x)$

where hi is the $i - th$ decision tree, and N is the number of trees.

LightGBM (Light Gradient Boosting Machine):

Tree-based learning methods are used in the gradient boosting framework LightGBM. It is made to be scalable and efficient.

Equation: $y^\wedge = \sum m = 1M\alpha mTm(x)$

where αm are the weights, Tm are the trees, and M is the number of trees.

CatBoost:

Another efficient gradient boosting approach for handling categorical data is called CatBoost. It progressively constructs decision trees in order to minimize the loss function.

Equation: $y^\wedge = \sum m = 1M\gamma mTm(x)$

where γm are the weights, Tm are the trees, and M is the number of trees.

Support Vector Regression (SVR):

Another efficient gradient boosting approach for handling categorical data is called CatBoost. It progressively constructs decision trees in order to minimize the loss function ϵ.

Equation: $f(x) = \langle w, x\rangle + b$

subject to $|yi - f(xi)| \leq \epsilon$.

XGBoost (Extreme Gradient Boosting):
A scalable and effective gradient boosting framework implementation is XGBoost. To prevent overfitting, it makes use of a more regularized model formalization.
Equation: $y\hat{} = \sum m = 1M\beta mTm(x)$
where βm are the weights, Tm are the trees, and M is the number of trees.

3.2.4 Performance Metrics:
Performance Metrics: A number of key metrics will be used to assess the machine learning models' performance. In order to gauge how accurate the model is at predicting values, the Mean Squared Error (MSE) calculates the average squared difference between actual and predicted values. An easy-to-understand indicator of prediction accuracy is the Mean Absolute Error (MAE), which computes the average absolute difference between actual and predicted values. By taking the square root of the average squared differences, the Root Mean Squared Error (RMSE) provides an error metric with the same scale as the target variable. Values closer to 1 indicate better model performance. R-squared (R^2) assesses the percentage of variance in the dependent variable that can be explained by the independent variables. These metrics will be used to assess the performance of each model and identify the most reliable indicator of Bangladeshi real estate prices. For multiple regression models, the Adjusted R-squared provides a more precise indicator of model performance because it takes the number of predictors into consideration. The utilization of cross-validation techniques will guarantee that the model's performance is resilient and not unduly reliant on any specific subset of the data.

Mean Squared Error (MSE)
MSE= $n1\sum i = 1n(yi - y\hat{}i)2$
where yi is the actual value, $y\hat{}i$ is the predicted value, and n is the number of data points.

Mean Absolute Error (MAE)
MAE= $n1\sum i = 1n|yi - y\hat{}i|$
where yi is the actual value, $y\hat{}i$ is the predicted value, and n is the number of data points.

Root Mean Squared Error (RMSE)

RMSE= $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - y\hat{}i)2}$

where yi is the actual value, $y\hat{}i$ is the predicted value, and n is the number of data points.A regression model's ability to predict actual values is measured by the Root Mean Square Error (RMSE), where lower values denote better fit.

R-squared (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(yi - \hat{y}\right)2}{\sum_{i=1}^{n}\left(yi - \bar{y}\right)2}$$

where yi is the actual value, $y^\wedge i$ is the predicted value, y^{-} is the mean of actual values, and n is the number of data points.
These parameters will be used to assess each model's performance and identify the model that best predicts Bangladeshi real estate values.

3.3 Hardware/Software Requirement

To properly execute property price prediction in Bangladesh using machine learning models, several critical parameters must be addressed. First and foremost, training and testing machine learning models require strong machines with ample RAM, processing power, and storage. Due to the task's complexity, powerful computing resources are required. Furthermore, model construction, training, and testing in a safe environment require software tools like Scikit-learn and other machine learning technologies. To fully employ the machine learning models, one must have a solid understanding of computer languages, especially Python, since Python provides a wealth of tools and technologies for easy algorithm integration. Moreover, a crucial phase in predicting real estate prices is data processing. Before being used, data needs to be cleansed, converted, and ready. Tools that handle and manipulate datasets for training, such NumPy, Pandas, and Scikit-learn, are essential. Finding trends and clearly presenting the results depend on data visualization as well. Effective data visualization is made possible by tools like Matplotlib and Seaborn, which aid in the understanding and interpretation of the results of machine learning models. Model assessment metrics like R squared, RMSE, MSE, and MAE must be understood and used appropriately in order to properly evaluate the performance of machine learning models. These measures shed light on how well the models are doing in terms of making accurate predictions about real estate values.
Throughout the implementation phase, collaborative development platforms such as GitHub enable efficient teamwork, tracking of changes, and transparency. Version control is made easier and cooperation is improved by utilizing these platforms.At every stage of the procedure, ethical considerations are crucial. An important set of ethical guidelines that must be followed during data collection, analysis, and model use are following societal norms, protecting privacy, getting informed consent, and supporting equal representation.Google Colab is a well-liked platform for using machine learning models, such as those that anticipate real estate prices. It is a suitable option for running machine learning models because it offers a cloud-based environment with strong computational

capabilities and makes setup simpler.

The suggested strategy for machine learning-based property price prediction in Bangladesh can be successfully implemented by meeting these implementation conditions. This implementation would provide accurate estimates and insightful information on Bangladesh's real estate market's outlook.

3.4 Project Management & Analysis

The success and sustainability of our research, "Property Price Prediction Using Machine Learning in Bangladesh," depend heavily on competent project management and financial analysis. To ensure smooth progress and the accomplishment of project goals, project management entails methodically planning various tasks, deadlines, and resource allocation. This includes assigning responsibilities, monitoring schedules, optimizing assets, guaranteeing effective communication, and putting risk mitigation strategies into practice. Financial analysis is essential for evaluating the project's financial viability and resource allocation at the same time. Our goal is to optimize resource allocation and enhance the project's impact by means of detailed cost estimation, thorough cost-effectiveness analyses, and comprehensive budgeting. Financial reporting guidelines are also required to uphold budgetary restraints, accountability, and transparency—all of which increase stakeholder trust and make possible funding sources available. We seek to advance the field of real estate analytics and guarantee the project's long-term sustainability by navigating the complexities of property price prediction research in Bangladesh with accuracy, flexibility, and financial acumen through rigorous project management and financial analysis. A desktop computer or laptop with at least 8 GB of RAM, machine learning frameworks like Scikit-learn or TensorFlow, cooperation for data acquisition from data suppliers and real estate experts, and continued cooperation with real estate specialists to enhance the project's scope and reach are some of the system requirements that must be met in order to facilitate the project's progress and execution. We are dedicated to accomplishing our research objectives and offering important insights into Bangladesh's real estate market by abiding by these guidelines and upholding stringent project management and financial procedures. Our goal is to maintain our leadership position in real estate analytics and provide innovative solutions to the Bangladeshi market by cultivating a culture of innovation and constant improvement.

To discuss our project's financial analysis, we must now build a subsection. Here is a general table that shows the approximate expenses of the various machine learning project components:

TABLE 3.4.1: Estimated Cost

SN	Components	Estimated Cost (BDT)
01.	Data Collection and Processing	5000-7000
02.	Documentation and Report Writing	1000-2000
03.	Visiting Stakeholders	2500-3000
04.	Software and Tools	500-1000
05.	Miscellaneous	500-1000
06.	Contingency (10% of total)	500-1000
	Total Estimated Cost	10,000-15,000

3.5 Summary

The study " A Comparative Study of Machine Learning Algorithms for Property Price Forecasting" uses advanced machine learning techniques to accurately predict property prices. Reputable websites such as Bproperty and Bikroy.com were used to gather data, with an emphasis on important property attributes like size, location, and amenities.A variety of machine learning algorithms, including Decision Trees, Random Forest, LightGBM, CatBoost, SVR, and XGBoost, were trained and assessed using metrics like MSE, MAE, RMSE, and R-squared value using Google Colab for data processing.

Task organization, resource optimization, cost estimation, budgeting, and financial reporting are all essential components of effective project management and financial analysis. Machine learning frameworks, at least 8 GB of RAM, and cooperation with real estate experts are needed for this project. The project is expected to cost between 10,000 and 15,000 BDT, which guarantees important information about the real estate market and the viability of the project.

CHAPTER 4
Implementation

4.1 Overview
This chapter describes how we used machine learning regression models to implement our research on property price prediction in Bangladesh. The 18,000 raw entries in the dataset came from two well-known real estate listing websites, Bproperty and Bikroy.com. The main goal was to properly choose and preprocess the dataset in order to improve our predictive models' accuracy and dependability. Location, number of bedrooms, number of bathrooms, number of parking spaces, area, region, sub-region, longitude, latitude, presence of an elevator, and type of property were among the attributes included in the data. Our objective was to use these characteristics to make accurate property price predictions. The procedures for handling non-numerical values, dividing the dataset, training the models, and comparing the effectiveness of various regression models are all covered in this chapter. We tested a variety of models, and by evaluating them for accuracy and computational efficiency, we determined which ones worked best.

4.2 Train Model
The training process began by dividing the dataset in two parts: 70% for training and 30% for testing. In order to give the model a sufficient amount of data to work with while still enabling it to make good generalizations on untested data, this split was selected. We used the one-hot encoding method to create dummy variablest. By using this technique, categorical data can be transformed into a set of binary variables that the model can use as numerical input. For instance, binary indicators were created from attributes such as property type, location, region, and sub-region.

Location, number of bedrooms, number of bathrooms, number of parking spaces, area, region, sub-region, longitude, latitude, presence of an elevator, and property type were among the attributes used to train the model. These characteristics were selected because they are pertinent to Bangladeshi real estate prices. The model's predictions were tested using "price," the target attribute.

We employed a reference attribute in addition to these attributes to guarantee the integrity and authenticity of our dataset. To prevent bias, this reference attribute was excluded from the model's evaluation procedure.Decision Tree, Random Forest, Support Vector Regression (SVR), CatBoost, XGBoost, and LightGBM were among the models we used for training. The testing dataset was used to evaluate each model after it had been trained using the training dataset. The precision with which these models were able to forecast real estate values served as a gauge of their performance.

We used one hot encoding to convert categorical variables to numerical variables. This method converts each distinct category—such as "Region" or "Sub-Region"—within a categorical column into a different binary column. The existence (1) or absence (0) of each new column in the original data corresponds to that category. Because machine learning models need numerical input, this transformation makes it possible to use categorical data, leading to predictions that are more accurate.

The relationships between the price, area, number of bedrooms, and number of bathrooms—all important features in the property price prediction dataset—are displayed in this pair plot.

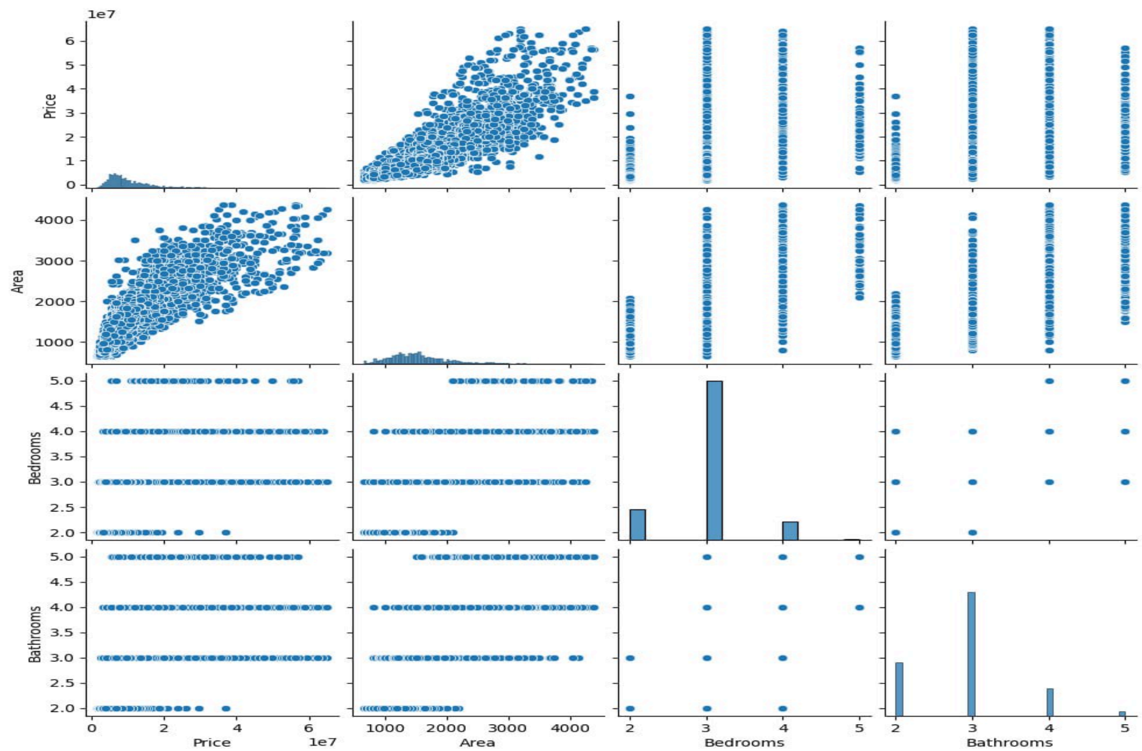

Figure 4.2.1: Pair Plot of Key Features

Figure 4.2.1 shows various data visualizations for analyzing the relationships between features in the property price prediction dataset. The figure's diagonal elements show histograms that show how each feature is distributed. One example of this is the right-skewed price distribution, which suggests that there are more properties available at lower prices. Scatter plots displaying pairwise relationships between various features are included in the off-diagonal elements. Cost and area show a strong positive correlation, suggesting that larger properties generally command higher prices. Price and the number of bedrooms and bathrooms also show positive correlations, albeit they are not as strong.

This suggests that homes with more bedrooms and bathrooms typically command higher prices. Additionally, larger properties typically have more rooms because larger areas are linked to more bedrooms and bathrooms. This can be seen by comparing the area to the number of bedrooms and bathrooms. Overall, the pair plot shows that price is strongly predicted by area, with notable effects also found for bathrooms and bedrooms. The strength and direction of the relationships between the various features in the dataset are clearly shown in this correlation matrix heatmap, which offers insightful information for predicting property prices.

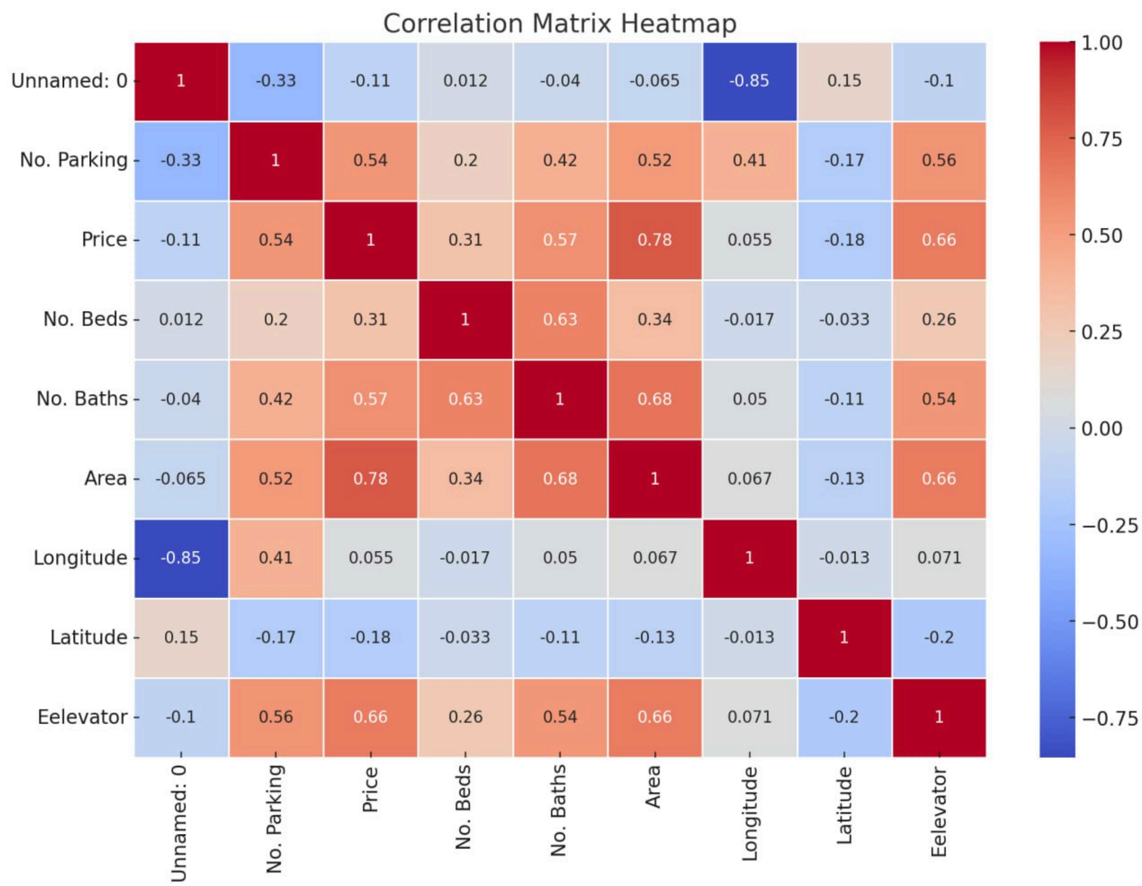

Figure 4.2.2: Correlation Matrix Heatmap

Figure 4.2.2 depicts key relationships that influence property prices. The price and area variables show a strong positive correlation (0.78), indicating that larger properties usually fetch higher prices. According to the price-to-bathroom correlation (0.57), homes with more bathrooms are typically more expensive. In a similar vein, a 0.54 correlation between price and parking spaces indicates that more parking spaces are associated with higher property values. Higher prices also correlate with the presence of an elevator (0.66), suggesting that properties with elevators are generally more valuable.

Furthermore, there is a strong correlation (0.68) between bathroom count and area, indicating that larger homes typically have more bathrooms. Overall, the heatmap shows that the presence of elevators, parking spaces, bathroom count, and area are significant predictors of property prices. These insights are essential for comprehending the variables that affect property values, which makes it easier to train machine learning models for property price prediction in an effective manner.

4.3 Model Evaluation

We evaluated the models' accuracy in predicting real estate prices using metrics such as R-squared (R^2), mean squared error (MSE), and mean absolute error (MAE). Six models were compared: Random Forest, Decision Tree, CatBoost, XGBoost, Support Vector Regression (SVR), and LightGBM. The findings demonstrated that XGBoost and CatBoost had the closest accuracy margins; however, CatBoost performed better because of its strong boosting mechanism and effective categorical data handling. Compared to CatBoost and LightGBM, Random Forest showed faster time complexity, despite being slightly less accurate. Competitive performance was shown by Support Vector Regression (SVR), especially on smaller datasets. The Decision Tree model performed worse than the other models despite being simple to use and understand due to its propensity to overfit the data. As an ensemble method, Random Forest provided a good balance between accuracy and computational efficiency. The two best performing algorithms overall were XGBoost and CatBoost; CatBoost was the most accurate but needed more processing power. Additionally, LightGBM demonstrated encouraging performance, particularly in terms of speed, which makes it a good choice in situations where processing power is scarce. These metrics were used to evaluate each model's performance: Mean Squared Error (MSE), which assigns more weight to larger errors by averaging the squares of the errors, Mean Absolute Error (MAE), which determines the average error magnitude across a set of predictions without taking direction into account; R-squared (R^2), with values closer to 1 indicating a better fit and indicating the proportion of variance in the dependent variable predictable from the independent variables, and Root Mean Squared Error (RMSE), the square root of the MSE, providing an error metric on the same scale as the target variable.

4.4 Summary

This chapter described the detailed implementation of our property price prediction model using machine learning techniques. From Bproperty and Bikroy.com, we gathered and processed a sizable dataset, divided it into training and testing sets, and transformed categorical data into numerical form. After training and assessing a number of machine learning models, CatBoost and XGBoost showed the best results. XGBoost was the most accurate model because it exhibits the lowest error metrics, which makes it a good option for any future property price prediction tasks in Bangladesh.

CHAPTER 5

Result & Analysis

5.1 Overview

The findings and analysis of our Bangladeshi property price prediction models are presented in this chapter. Using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2), we assess the performance of six machine learning models. Support Vector Regression (SVR), Random Forest, Decision Tree, XGBoost, LightGBM, and CatBoost are among the models that are contrasted. These metrics are used to analyze each model's efficacy and accuracy.

5.2 Experimental Result

In this research, six algorithms were used for property price prediction in Bangladesh, each with varying levels of performance. These models will show us the performance of property price prediction. The performance metrics for each model are summarized in the table below:

TABLE 5.2.1:Accuracy Measures

Models	MAE	MSE	RMSE	R-squared (%)
Support Vector Regression	0.220	0.168	0.410	0.83
XGBoost	0.189	0.092	0.304	0.91
Decision Tree	0.216	0.178	0.421	0.82
Random Forest	0.184	0.106	0.326	0.89
LightGBM	0.197	0.100	0.317	0.90
CatBoost	0.196	0.093	0.305	0.91

Table 5.2.1 shows that XGBoost and CatBoost emerge as the top-performing models for predicting property prices. XGBoost exhibits the lowest error metrics (MAE: 0.189, MSE: 0.092, RMSE: 0.304) and the highest R-squared value (91%), indicating superior accuracy and the ability to explain 91% of the variance in the data. CatBoost, with slightly higher error metrics (MAE: 0.196, MSE: 0.093, RMSE: 0.305) but the same R-squared value (91%), performs similarly well. Compared to other models like Support Vector Regression, Decision Tree, Random Forest, and LightGBM, which have higher

error metrics and lower R-squared values, XGBoost stands out due to its powerful ensemble learning technique that combines multiple decision trees to enhance accuracy and robustness. Its strengths lie in regularization, handling missing values, parallel processing, tree pruning, and flexibility for various tasks. These features contribute to its exceptional performance, making it the best model for predicting property prices based on the given data.

5.3 Comparative Analysis

The performance of six different machine learning models for property price prediction in Bangladesh was assessed using metrics such as MAE, MSE, RMSE, and R-squared.

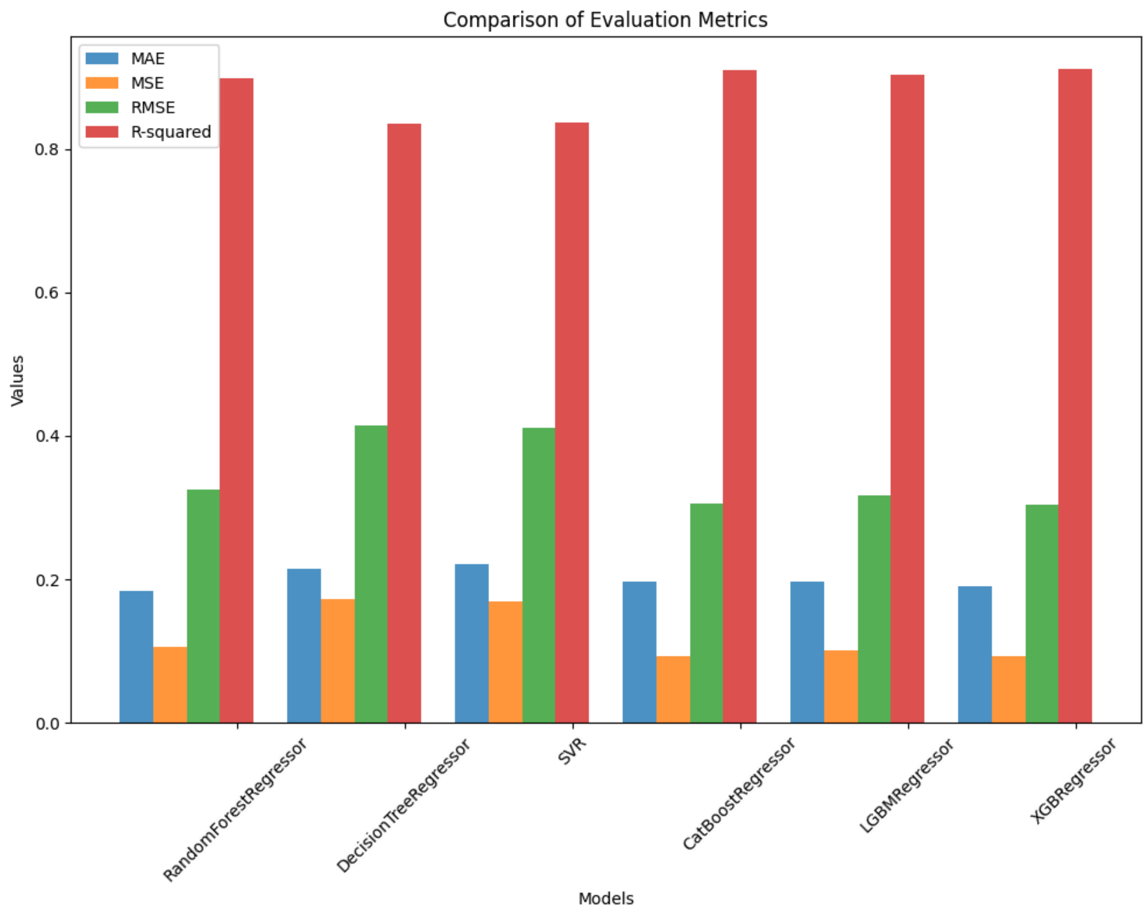

Figure 5.3.1: Comparison of Evaluation Metrics

The algorithms employed in the study on property price prediction are contrasted here: When comparing LightGBM and CatBoost, CatBoost explained more variation in property prices and had a slightly higher R-squared value (0.91). In addition to having more accurate predictions, CatBoost also had lower error metrics: MSE of 0.093 and RMSE of 0.305 compared to MSE of 0.100 and RMSE of 0.317 for LightGBM. However, CatBoost usually offers higher accuracy and better handling of categorical variables at the cost of increased computational complexity. Random forests performed better than decision trees in terms of accuracy, with an R-squared value of 0.89 versus 0.82, and in terms of prediction accuracy, with lower RMSE (0.326) and MAE (0.184) than Decision Trees (0.421 and 0.216). Because Random Forest is ensemble in nature, it is more robust and less prone to overfitting. When Support Vector Regression (SVR) and XGBoost were compared, XGBoost produced slightly better predictions with an R-squared value of 0.91 and an RMSE of 0.308 compared to 0.410 for SVR. SVR is usually better for smaller datasets, while XGBoost works best for large datasets and complex relationships. Random Forest and LightGBM both had the same R-squared value (0.89), but LightGBM had a slightly lower MSE (0.100) than Random Forest (0.106). LightGBM is more appropriate for larger datasets because of its faster training and prediction times due to its histogram-based learning methodology. On larger datasets, LightGBM performed better in terms of speed and efficiency despite the high accuracy displayed by both models.

In terms of accuracy and error metrics, CatBoost and XGBoost were found to be the most effective models overall. LightGBM performed better and faster on larger datasets than Random Forest and XGBoost, despite both models producing excellent results. Though their accuracy was comparable to LightGBM, XGBoost and CatBoost showed slightly superior error metrics. The particular dataset and the necessary trade-off between accuracy, speed, and complexity for the task ultimately determine which algorithm is used.

5.4 Summary

This chapter compared the performance of different machine learning models for predicting property prices in Bangladesh. The most accurate model was found to be CatBoost and XGBoost, with Random Forest and LightGBM trailing closely behind. The comparative study brought to light the advantages and disadvantages of each model and offered insightful information about how well-suited each is for predicting real estate prices. These results provide a strong basis for further investigation and useful applications of machine learning techniques in property valuation.

CHAPTER 6

Impact on Society, Environment & Sustainability

6.1 Impact on Life

The use of machine learning algorithms for property price prediction in Bangladesh has the potential to significantly improve people's lives, particularly by allowing them to make more informed decisions, address housing affordability, increase real estate investment, and support urban planning and infrastructure development. Real estate agents, sellers, and buyers can all benefit from precise property price prediction models powered by machine learning algorithms. These models allow them to make well-informed decisions based on past performance, market trends, and property attributes. This improves people's capacity to make decisions that are consistent with their financial objectives and lifestyle preferences, which raises their level of satisfaction and quality of life in general. Furthermore, by identifying areas with affordable housing options, these models can assist developers, policymakers, and urban planners in creating strategies and programs to make housing more accessible and affordable for a larger portion of the population. This will help address Bangladesh's housing affordability issue. Machine learning algorithms can help real estate investors choose properties with potential for appreciation, maximize returns, and contribute to long-term wealth accumulation and financial stability. Additionally, by highlighting locations that require infrastructure upgrades—such as transportation networks, educational institutions, healthcare facilities, and recreational areas—property price prediction models can assist in informing urban planning and infrastructure development and promote livable, sustainable, and interconnected communities. By encouraging a robust real estate market, raising investor confidence, and drawing in both domestic and foreign investments, accurate property price prediction models also contribute to economic growth and stability. This eventually improves GDP, job opportunities, and personal financial security. These models can also be used by nonprofits and community organizations to find possible sites for development projects that will improve social cohesion, accessibility to necessary services, and general community well-being. While there are many advantages to machine learning algorithms, it is important to take ethical factors like protecting against algorithmic biases, maintaining fairness, transparency, accountability, and data privacy into account. In conclusion, by enabling effective urban planning, addressing housing affordability, promoting real estate investment, enabling informed decision-making, fostering economic growth, and supporting community development, the use of machine learning algorithms for real estate price prediction in Bangladesh has the potential to significantly improve people's quality of life.

6.2 Impact on Society & Environment

When predicting property prices in Bangladesh, a variety of societal and environmental factors are considered. Social equity is promoted by affordability because it plays a critical role in preserving housing accessibility for a range of income levels through property price prediction. Precise predictions facilitate urban planning by pointing out regions that are expected to grow quickly and guaranteeing that infrastructure and services keep up with this growth. These forecasts help real estate developers build homes that satisfy consumer demand, reducing housing shortages and improving quality of life. Predicting property prices can encourage sustainable development from an environmental standpoint by lowering carbon footprints and promoting green building technologies. Communities can reduce the risk of natural disasters and prepare for climate resilience by having an understanding of pricing trends. Decisions about the maintenance of green areas, the addition of urban vegetation to city plans, and the enhancement of biodiversity and air quality are also influenced by predictive models. Property price prediction in Bangladesh thus promotes equitable housing access, resilient urban development, and sustainable urban growth, all of which raise the standard of living for all citizens from a social and environmental standpoint.

6.3 Ethical Aspects

When using machine learning algorithms to predict property prices in Bangladesh, ethical considerations are crucial. Providing informed consent and protecting data privacy are essential ethical obligations. Sensitive property and personal data must be handled carefully in this process, and participant consent or data anonymization are essential. Furthermore, in order to avoid unfair or discriminatory results, biases introduced during modeling or inherent in the data must be addressed. By actively identifying and reducing biases, predictive models are guaranteed to produce impartial and equitable results, supporting just and equitable real estate market practices. Another pillar of ethics is transparency, which is essential to building trust and accountability. Policymakers, developers, and prospective purchasers are among the stakeholders who need to be made aware of the implications, constraints, and potential applications of the predictive models. Making well-informed decisions is aided by the open disclosure of model performance indicators and validation techniques. Furthermore, it is consistent with social justice principles to make sure that property market forecasts do not worsen already-existing inequities. A more equitable real estate market is a result of initiatives to reduce inequalities in market access and give underrepresented communities priority. This research advances predictive modeling in the real estate industry while protecting the rights, dignity, and well-being of all participants by upholding ethical standards and encouraging responsible practices. It highlights the significance of ethical considerations in the development and application of machine learning algorithms for the prediction of property prices in Bangladesh, underscoring the need of a conscientious and ethical approach to technological innovation.

6.4 Sustainability Plan

The sustainability of our work on property price prediction in Bangladesh very much so requires it. The following sustainability plan lays out key strategies and initiatives designed to ensure the continued success and impact of our studies. In order to make precise property value forecasting, we must retain comprehensive and current datasets. We will establish partnerships with financial institutions/real estate companies/governmental organizations etc., where data collection can be done continuously through the platform. The dataset will be updated and expanded to track how the market develops over time. To maintain such a performance, the model must be monitored and improved in continuous manners so we will also enforce monitoring systems that locate whether your machine learning algorithms still prove to be as effective on new data. Industry participants and real estate professionals can provide valuable feedback for identifying new issues, as well as opportunities to improve the model through subsequent iterations, thereby keeping it current with updated prediction data. Another aspect of sustainability is stimulating the spread and multiplication of expertise which capacity building locally. In this series, we will host capacity-building workshops and training sessions for local stakeholders — including agents of real estate professionals, researchers and policymakers who want to learn how to apply machine learning property price prediction. Additionally, we will publish our results to ensure wide adoption from the community through journal papers and other venues as well as being engaged with related working groups and CVD forums. We believe that these sustainable practices would also immediately benefit value prediction in the case of our property price research and lead to greater real estate growth both quickly (needed for Bangladesh) as well as long-term, following world markets.

6.5 Summary

Predicting property prices in Bangladesh with machine learning has significant implications for individuals, society, and the environment. It makes housing data more accessible, enabling both buyers and sellers to make well-informed decisions. In terms of society, it encourages efficiency and openness in real estate transactions, which may lessen disparities in the market. Environmentally, it promotes sustainable development methods by aiding in urban planning and resource allocation. Ethical implementation is ensured by legal adherence to regulatory frameworks and data privacy. By coordinating growth with environmental stewardship, optimized land use and infrastructure development promote sustainability. All things considered, this predictive technique not only improves decision-making but also helps Bangladesh's real estate industry adopt sustainable practices and more equitable urban development.

CHAPTER 7
Conclusion &
Future Work

7.1 Conclusions

The "A Comparative Study of Machine Learning Algorithms for Property Price Forecasting" study shows how well machine learning algorithms can predict real estate prices. Nine different algorithms—CatBoost, Logistic Regression, Random Forest, Decision Trees, XGBoost, and Light GBM—were assessed using a dataset that was enhanced with pertinent input features. Out of all of them, CatBoost and XGBoost performed better than the others, obtaining a 91% prediction accuracy. This result emphasizes how trustworthy CatBoost and XGboost are for applications predicting real estate prices. All things considered, this research creates a strong foundation for real estate economics research and innovation in the future. It offers insightful information about the real-world implementation of machine learning algorithms for property price forecasting in Bangladesh, emphasizing how these algorithms can improve decision-making and guide urban development plans.

7.2 Further Suggested Works

The findings of this study on property price prediction using machine learning algorithms provide useful insights and suggest several areas for future research. The investigation of alternative data preprocessing methods, such as feature engineering and outlier detection, to reduce inaccuracies in model predictions is a key implication for future research. Furthermore, by combining the advantages of several models, integrating sophisticated ensemble techniques like stacking or boosting may improve prediction accuracy.

Considering the computational requirements of training intricate machine learning models on large datasets, future studies could explore the use of cloud-based or distributed computing frameworks to increase efficiency and scalability. This emphasizes the need for more reliable techniques that can manage massive amounts of data without compromising the accuracy of predictions.

Although the study concentrated on predicting property prices in Bangladesh, there is still a great deal of room for more in-depth investigation in this area. Future research could focus more on investigating innovative machine learning architectures that are adapted to local real estate dynamics by utilizing a variety of datasets that represent regional differences in real estate markets. Furthermore, investigating the useful uses of these

models in real estate valuation and investment decision-making may enhance our comprehension and uptake of machine learning in the real estate industry.
To summarise, the research implications for the future involve improving data preprocessing methods, investigating machine learning frameworks that are scalable, and broadening the scope of studies to include more aspects of Bangladesh's property market dynamics. These initiatives have the potential to progress real estate analytics by improving the precision and effectiveness of property price forecasts in regional settings.

7.3 Limitations

A Comparative Study of Machine Learning Algorithms for Property Price Forecasting has several significant limitations. First off, it can be difficult to find and maintain high-quality property data because complete and reliable datasets—such as those containing transaction histories and property attributes—might be hard to come by or inconsistent. Furthermore, bias in the data towards particular property types or regions can distort predictions, reducing the model's capacity to generalize to a variety of contexts. Another issue is overfitting, in which models may identify anomalies or noise in the training set that does not correspond to general market trends. Furthermore, some sophisticated machine learning models can be opaque, making it difficult to understand how predictions are made. Although they might not be fully reflected in historical data, external factors like market sentiments, policy changes, and socioeconomic situations have a further impact on property prices. Robust infrastructure and frequent updates are also necessary for implementation in order to handle massive data processing and changing market dynamics. Over time, changes in market regulations or property laws may also have an impact on the applicability and accuracy of the model. To improve predictive accuracy and reliability, these challenges must be addressed with careful data handling, deliberate model selection, and continuous validation against current market conditions.

REFERENCES

[1] Sapakova, S., & Sapakov, A. (2024). Features of modeling the online real estate market in Almaty. Procedia Computer Science, 231, 409-414. https://doi.org/10.1016/j.procs.2023.12.226

[2] Bastos, J. A., & Paquette, J. (2024). On the uncertainty of real estate price predictions.https://rem.rc.iseg.ulisboa.pt

[3]Alkan, T., Dokuz, Y., Ecemiş, A., Bozdağ, A., & Durduran, S. S. (2023). Using machine learning algorithms for predicting real estate values in tourism centers. Soft Computing, 27(5), 2601-2613. https://doi.org/10.21203/rs.3.rs-1757533/v1

[4] Khosravi, M.; Arif, S.B.; Ghaseminejad, A.; Tohidi, H.; Shabanian, H. Performance Evaluation of Machine Learning Regressors for Estimating Real Estate House Prices. Preprints.org 2022, 2022090341. https://doi.org/10.20944/preprints202209.0341.v1

[5] Ming, Y., Zhang, J., Qi, J., Liao, T., Wang, M., & Zhang, L. (2020, September). Prediction and analysis of chengdu housing rent based on xgboost algorithm. In Proceedings of the 2020 3rd International Conference on Big Data Technologies (pp. 1-5). 10.1145/3422713.3422720

[6] Lu, S., Li, Z., Qin, Z., Yang, X., & Goh, R. S. M. (2017, December). A hybrid regression technique for house prices prediction. In 2017 IEEE international conference on industrial engineering and engineering management (IEEM) (pp. 319-323). IEEE. doi:10.1109/IEEM.2017.8289904

[7] Kansara, D., Singh, R., Sanghvi, D., & Kanani, P. (2018). Improving accuracy of real estate valuation using stacked regression. Int. J. Eng. Dev. Res., 6(3), 571–577.

[8] Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. Land use policy, 111, 104919.

[9] Avanijaa, J. (2021). Prediction of house price using xgboost regression algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2), 2151-2155.

[10] Ashraf, A., Rady, M., & Mahfouz, S. Y. (2024). Price prediction of residential buildings using random forest and artificial neural network. HBRC Journal, 20(1), 23-41.

[11] Kontokosta, C. E., Hong, B., Johnson, N. E., & Starobin, D. (2018). Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. Computers, Environment and Urban Systems, 70, 151-162. https://doi.org/10.1016/j.compenvurbsys.2018.03.004

[12] Sibindi, R., Mwangi, R. W., & Waititu, A. G. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, 5(4), e12599.https://doi.org/10.1002/eng2.12599

[13] Rey-Blanco, D., Zofío, J. L., & González-Arias, J. (2024). Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses. Expert Systems with Applications, 235, 121059.https://doi.org/10.1016/j.eswa.2023.121059

[14] A. Varma, A. Sarma, S. Doshi and R. Nair, ”House Price Prediction Using Machine Learning and Neural Networks,” 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018,

pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.
[15] Manasa, J., Radha Gupta, and N. S. Narahari.“Machine learning based predicting house prices using regression techniques.” 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA). IEEE, 2020
[16] Vlahovljak, A. (2022). House price prediction in the UK. Experiment Findings. DOI: 10.13140/RG.2.2.12729.65121
[17] Yağmur, A., Kayakuş, M., & Terzioğlu, M. (2022). House price prediction modeling using machine learning techniques: A comparative study. Aestimum. https://doi.org/10.36253/aestim-13703
[18] Li, C. (2024). House price prediction using machine learning. Proceedings of the 4th International Conference on Signal Processing and Machine Learning, 225-237. https://doi.org/10.54254/2755-2721/53/20241426
[19] Nwankwo, M. P., Onyeizu, N. M., Asogwa, E. C., Ejike, C. O., & Obulezi, O. J. (2023). Prediction of house prices in Lagos-Nigeria using machine learning models. European Journal of Theoretical and Applied Sciences, 1(5), 313-326. https://doi.org/10.59324/ejtas.2023.1(5).22
[20] Adetunji, A. B., & Others. (2022). Predicting House Prices Using Random Forest Machine Learning Technique. Procedia Computer Science, 199, 806-813. https://doi.org/10.1016/j.procs.2022.01.100
[21] Zou, C. (2023). The house price prediction using machine learning algorithm: The case of Jinan, China. Highlights in Science, Engineering and Technology CMLAI 2023, 39, 327-333. https://doi.org/10.54097/hset.v39i.6549
[22] Ouyang, X. (2024). House price prediction based on machine learning models. Highlights in Science, Engineering and Technology CSIC 2023, 85, 870-878. https://doi.org/10.54097/ftyf9665
[23] Jangaraj, A., Sunitha, G., Madhavi, K. R., Kora, P., & Vittal, R. H. S. (2021). Prediction of house price using XGBoost regression algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2), 2151-2155.
[24] Patel, D., Nayak, P., Gupta, S., & Jayanth, C. (2023). House price prediction using machine learning. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11(6), 1870-1873. https://doi.org/10.22214/ijraset.2023.53841
[25] Li, H. (2023). House price prediction and analysis based on random forest and XGBoost models. Highlights in Business, Economics and Management GEFHR 2023, 21, 934-938.
[26] Liu, Z. (2023). Real estate price prediction based on supervised machine learning scenarios. Highlights in Science, Engineering and Technology CMLAI 2023, 39, 731-737.
[27]Weng, W. (2022). Research on the house price forecast based on machine learning algorithm. BCP Business & Management AFTEM 2022, 32, 134-147.
[28] Chen, Z. (2024). Integrating machine learning techniques for real estate analysis. Highlights in Science, Engineering and Technology SDPIT 2024, 92, 252-256.
[29] Chen, Y. (2023). Research on the prediction of Boston house price based on linear regression, random forest, XGBoost, and SVM models. Highlights in Business, Economics and Management GEFHR 2023, 21, 27-37.
[30] Kaushal, A. M. A., Kaushal, T., Kalinga, H. P. M., Kisok, S., & Kulathunga, Y. J. M. (2024). House prices prediction in New York City using machine learning. Information

Technology Horizon Campus. Retrieved from https://www.researchgate.net/publication/380877232
[31] Modi, M., Sharma, A., & Madhavan, D. P. (2020). Applied research on house price prediction using diverse machine learning techniques. International Journal of Scientific & Technology Research, 9(04).
[32] Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. (2019). House price prediction using machine learning. Journal of Innovative Technology and Exploring Engineering, 8(9), 717–722. doi:10.35940/ijitee.I7849.078919
[33] Güneş, F., Wolfinger, R., & Tan, P. Y. (2017, April). Stacked ensemble models for improved prediction accuracy. Proc. Static Anal. Symp., 1-19

APPENDIX A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title: A Comparative Study of Machine Learning Algorithms for Property Price Forecasting

Student ID: 201-15-3087, 201-15-3275

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to identify **property price prediction** problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public real estate market and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, real estate market, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**

CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates **fundamental engineering (K3)** principles by employing machine learning , and diverse machine learning models for regression tasks. The project demonstrates **specialist knowledge (K4)** by conducting machine learning for pneumonia price prediction accuracy in property price dataset, crucial for real estate and economical aspects.	**Page no:** [21-26] **Section:** [3.2] **Page no:** [1-2] **Section:** [1.1]
				The project applies **engineering practice & design (K5)** by the figure of process of experiments. The project addresses **engineering practice &**	**Page no:** [21-26]

				technology (K6) by employing Machine learning regression model.	**Section:** [3.2] **Page no:** [27-28] **Section** [3.4]
				This project ensures K8 **(Research Literature)** by synthesizing insights from recent studies, to advance property price prediction techniques, showcasing a comprehensive understanding of current methodologies.	**Page no:** [10-17] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing the hurdles in pneumonia detection, including limitations of traditional methods and the complex machine learning method. Through comparative analysis, it confronts challenges in understanding spatial distributions, offering insights for refining property price methodologies.	**Page no:** [17-18] **Section:** [2.4]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting Machine learning as the chosen significant solution for enhancing property price prediction amidst multiple potential approaches.	**Page no:** [34-35] **Section:** [5.2]

4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's interdisciplinary approach extends beyond computer science and engineering, impacting the real estate market and economic sector of Bangladesh contributing to advancements in investors which indicates **EP-4**.	**Page no:** [10-18] **Section:** [2.2, 2.4]
5.	**EP2: Range of Conflicting Requirements**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and conflicting**	No	CO8	N/A	N/A
7.	**EP7: Interdependence**	Yes	CO5	This project's comprehensive approach addresses high-level problems by integrating various components across data collection, handling outliers, and proposed methodology, ensuring a holistic solution to complex challenges in property pricing which ensures **EP-7**.	**Page no:** [21-28] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, preprossed datasets, and ethical considerations to ensure systematic research and contribute to advancements in property price prediction through machine learning regression model.	**Page no:** [28] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A
4.	**EA4: Consequences for society and the environment**	Yes		This project contributes to society by improving the real estate sector through advanced property price prediction, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for apartment listing data privacy.	**Page no:** [37-39] **Section:** [6.1, 6.2, 6.4]

5.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a novel approach in property price prediction through machine learning regression model, demonstrated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [10-17] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [7-9] **Section:** [1.6]
2	CO9	Yes	The project demonstrates adherence to ethical principles by prioritizing listing data privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced apartment price prediction technologies which comply with **CO9**.	**Page no:** [38-39] **Section:** [6.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges.	**Page no:** [21-28, 34-35] **Section:** [3.2, 3.3, 3.4, 3.5, 5.2]

dehan_Irfan_thesis

ORIGINALITY REPORT

12%	9%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Rajkumar S., Mary Nikitha K., Ramanathan L., Rajasekar Ramalingam, Mudit Jantwal. "chapter 15 Cloud Hosted Ensemble Learning-Based Rental Apartment Price Prediction Model Using Stacking Technique", IGI Global, 2022 Publication	1%
2	easychair-www.easychair.org Internet Source	1%
3	"Front Matter", 2023 8th International Conference on Computer Science and Engineering (UBMK), 2023 Publication	1%
4	www.mdpi.com Internet Source	1%
5	www.frontiersin.org Internet Source	1%
6	Md Abdullah Al Rahib, Nirjhor Saha, Raju Mia, Abdus Sattar. "Customer data prediction and analysis in e-commerce using machine	<1%