MINI PROJECT ON

“DECIPHERING THE REAL ESTATE MARKET: ADVANCED HOUSE PRICE PREDICTION USING RANDOM FOREST REGRESSOR’’

Submitted

In the partial fulfilment of the requirements for
The award of the degree of

BACHELOR OF TECHNOLOGY
In
COMPUTER SCIENCE & ENGINEERING
By

KRANTHI.M	**20U51A0552**
ROJA.P	**20U51A0567**
VINITHA.K	**20U51A0539**
SHYAM.K	**20U51A0540**
PARAMESHWARI.M	**20U51AO551**

Under the Guidance of
Prof. (Dr) Gnaneswara Rao Nitta

DRK College of
Engineering & Technology
(Approved by AICTE & Affiliated to JNTUH)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DRK COLLEGE OF ENGINEERING AND TECHNOLOGY
Affiliated to JNTU HYDERABAD
Bowrampet, HYDERABAD-500043.

DRK College of Engineering & Technology
(Approved by AICTE & Affiliated to JNTUH)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DRK COLLEGE OF ENGINEERING AND TECHNOLOGY
Affiliated to JNTUH HYDERABAD,
Bowrampet, HYDERABAD-500043.

CERTIFICATE

This is to certify that the Project Report entitled **“DECIPHERING THE REAL ESTATE MARKET: ADVANCED HOUSE PRICE PREDICTION USING RANDOM FOREST REGRESSOR”** that is being submitted by **KRANTHI.M (20U51A0552) , ROJA.P (20U51A0567) ,VINITHA.K (20U51A0539) , SHYAM.K (20U51A0540) , PARAMESHWARI.M (20U51AO551)** in partial fulfillment for the award of B. Tech degree in Computer Science and Engineering to the DRK COLLEGE OF ENGINEERING AND TECHNOLOGY Affiliated to JNTU HYDERABAD, is a record of Bonafide work carried out by them under the supervision of faculty member of CSE Department.

GUIDE **EXTERNAL EXAMINER** **HOD,CSE**

DRK College of Engineering & Technology
(Approved by AICTE & Affiliated to JNTUH)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DRK COLLEGE OF ENGINEERING AND TECHNOLOGY
Affiliated to JNTUH HYDERABAD,
Bowrampet, HYDERABAD-500043.

DECLARATION

I hereby declare that the PROJECT entitled **"DECIPHERING THE REAL ESTATE MARKET: ADVANCED HOUSE PRICE PREDICTION USING RANDOM FOREST REGRESSOR"** submitted for the **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**. This dissertation is our original work, and the PROJECT has not formed the basis for the award of any degree, associateship, and fellowship, or any other similar titles, and no part of it has been published or sent for publication at the time of submission.

BY

KRANTHI.M (20U51A0552)

ROJA.P (20U51A0567)

VINITHA.K (20U51A05390)

SHYAM.K (20U51A0540)

PARAMESHWARI.M(20U51AO551)

ACKNOWLEDGEMENT

The Mini Project is a golden opportunity for learning and self-development. consider myself very lucky and honored to have so many wonderful people lead me through-in the completion of this Mini Project.

I would like to articulate my deep gratitude to my Guide **Prof.(Dr.) Gnaneswara Rao Nitta** for his/her continuous support throughout the project. We are thankful to our project coordinators **Mr. Suneel Kumar. J** for their advice throughout the project.

We take the privilege of thanking **(Dr) K. KANAKA VARDHINI,** HOD-CSE, for her assistance with the project. We also thank our vice-principal Prof. **M. SRINIVAS RAO**, who has made the atmosphere so easy to work.

I gratefully express my sincere thanks to our beloved principal **Prof. (Dr.) GNANESWARA RAO NITTA,** for his continuous support and making it possible to complete the project in time.

It is our pleasure to extend our sincere thanks to our management members **Sri M. Chandra Sekhara Rao** Director, **Sri D Sriram** Treasurer, and Honorable Chairman **Sri. D B Chandra Sekhar Rao.**

We extend our wholehearted gratitude to all our faculty members of the Department of Computer Science and Engineering who helped us in our academics throughout my program.

Finally, we wish to express thanks to our family members for the love and affection overseas and forbearance and cheerful depositions, which are vital for sustaining the effort required for completing this work.

BY

KRANTHI.M (20U51A0552)

ROJA.P (20U51A0567)

VINITHA.K (20U51A0539)

SHYAM.K (20U51A0540)

PARAMESHWARI.M(20U51AO551)

DEDICATED TO MY BELOVED

PARENTS

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

In recent research, this paper investigates with its dynamic and complex nature, the real estate market presents significant challenges in house price prediction, which is crucial for investors and homeowners. The present project addresses this challenge against fluctuating market trends and varying property values. Motivated by the need for more accurate and reliable prediction models in the real estate sector, this study aims to leverage the power of machine learning to enhance house price forecasting.

The project employs a Random Forest Regressor, a robust machine learning technique, applied to the House Price Prediction Challenge dataset. This methodology is chosen for its effectiveness in handling large datasets with numerous variables, a common characteristic of real estate data. The objective is to develop a model that can accurately predict house prices based on a range of features, including location, property size, amenities, and market trends.

The expected outcome of this project is a significant improvement in prediction accuracy compared to traditional statistical methods. The Random Forest Regressor model promises to provide deeper insights into the factors influencing house prices by capturing the intricate patterns and relationships within the dataset. The findings are expected to benefit real estate professionals, investors, and policymakers, providing a more comprehensive understanding of the housing market dynamics.

KEYWORDS: House prices, real estate marketing, housing market.

CHAPTER 1

INTRODUCTION

1.1 Introduction

The real estate market, characterized by its dynamic and complex nature, poses substantial challenges in accurately predicting house prices. This capability is paramount for both investors seeking profitable opportunities and homeowners aiming to understand the value of their properties. Recognizing the volatility inherent in real estate, this research project endeavors to tackle the formidable task of house price prediction amidst fluctuating market trends and diverse property values.

Motivated by the imperative for more precise and reliable prediction models in the real estate sector, this study adopts a forward-looking approach. It seeks to harness the potential of machine learning as a tool to significantly enhance the accuracy of house price forecasting. In particular, the project focuses on leveraging the capabilities of a Random Forest Regressor, a robust machine learning technique selected for its efficacy in handling large datasets with numerous variables—a characteristic commonly found in real estate data.

This carefully curated dataset encapsulates a diverse array of factors influencing house prices, including but not limited to location, property size, amenities, and prevalent market trends. The overarching objective is clear: to craft a predictive model that excels in delivering precise house price estimations by considering a comprehensive set of features. In response to the imperative need for more precise and reliable prediction models within the real estate sector, this study takes a forward-looking stance. It seeks to harness the potential of machine learning as a transformative tool to enhance the accuracy of house price forecasting. A key focus of this investigation is the application of the Random Forest Regressor, a robust machine learning technique chosen for its efficacy in handling large datasets replete with numerous variables a characteristic commonly associated with real estate data.

The expected outcome of this research is a notable enhancement in prediction accuracy when compared to conventional statistical methods. The Random Forest Regressor, with its adeptness at discerning intricate patterns and relationships within vast and complex datasets, holds the potential to unveil deeper insights into the multifaceted factors influencing house prices. The findings of this study are poised to be of significant value not only to real estate professionals but also to investors and policymakers, fostering a more nuanced and comprehensive understanding of the dynamic forces at play in the housing market.

1.2 Literature Survey:

This paper is valuable in the realm of real estate analytics, the literature surrounding advanced house price prediction techniques has witnessed a significant surge. Notably, studies have delved into leveraging sophisticated algorithms, such as the Random Forest Regressor, to decipher the complexities of the real estate market. Researchers have explored the efficacy of machine learning models in capturing intricate patterns within vast datasets, incorporating factors like property features, location, and economic indicators.

The Random Forest Regressor, known for its ensemble learning capabilities, has garnered attention for its ability to enhance predictive accuracy. Existing literature has scrutinized the nuances of feature engineering, spatial analysis, and temporal dynamics in the context of house price prediction, paving the way for a comprehensive understanding of how advanced methodologies can revolutionize forecasting accuracy in the ever-evolving real estate landscape.

Additionally, the literature survey highlights other studies that use various instruments to examine the impact of credit supply on house prices. Examples include research on U.S. branching deregulations, the pre-emption of national banks from local laws, liquidity shocks during financial crises, and the influence of acts like the Community Reinvestment Act. The current paper contributes by specifically investigating the changes in mortgage loans due to a sudden reduction in rates and how this, in turn, affects house prices, providing a nuanced perspective on the interaction between mortgage credit supply and housing market dynamics.

We have concluded Machine Learning algorithms can predict a target/outcome by using Supervised Learning. This paper focuses on machine learning techniques for house price prediction. To get the specified outputs it needs to generate an appropriate function by set of some variables which can map the input variable to the aim output. The paper conveys that the predictions can be done by Machine Learning algorithm which attain the House price prediction with best accurate value

1.3 Project Background:

The project, "Deciphering the Real Estate Market: Advanced House Price Prediction Using Random Forest Regressor," seeks to address the complexities inherent in real estate valuation by employing advanced machine learning techniques. In the dynamic and intricate landscape of the real estate market, accurate prediction of house prices is paramount for informed decision-making by stakeholders such as homebuyers, sellers, and investors. Traditional valuation methods often fall short in capturing the multifaceted relationships within datasets comprising diverse features like property attributes, location specifics, and economic indicators. The motivation behind this project is to leverage the Random Forest Regressor, a powerful ensemble learning algorithm, to navigate high-dimensional spaces and discern intricate patterns within real estate data. Unlike traditional regression models, the Random Forest Regressor excels at handling non-linear relationships, making it well-suited for the nuanced and dynamic nature of property valuation. The project aims to explore the algorithm's effectiveness in mitigating overfitting concerns, particularly crucial when dealing with extensive datasets common in real estate analytics. Through extensive feature engineering and leveraging the ensemble nature of Random Forests, the project seeks to provide a more accurate and interpretable model for predicting house prices. Spatial analysis, temporal dynamics, and diverse property features will be considered to enhance the algorithm's predictive capabilities. By deciphering the real estate market using advanced machine learning, this project aims to contribute valuable insights that empower stakeholders to navigate and understand the intricate factors influencing house prices in a rapidly changing real estate environment.

To address potential concerns of borrower self-selection, the study examines changes in house characteristics and repeats the analysis for houses sold at least twice during the sample period. The results indicate that the observed increase in house prices is not driven by changes in borrower or house characteristics but is instead attributed to the impact of the subsidized mortgage program. The study employs a random forest setting to control for trends in the economy, providing a robust framework to identify and quantify the effects of the sudden reduction in the cost of credit on mortgage loans and house prices.

1.4. Issues:

Model Generalization to New Markets: The project must address the challenge of generalizing the Random Forest model to different real estate markets. Adapting the algorithm to diverse market conditions and ensuring its applicability beyond the training dataset is crucial for the model's utility in varied real estate scenarios.

Feature Engineering Challenges: The project encounters complexities in crafting meaningful features for the Random Forest Regressor, as real estate datasets often comprise diverse and high-dimensional variables. Addressing the selection and transformation of features is crucial to optimizing the algorithm's predictive accuracy.

Evaluation Metrics: Determining appropriate evaluation metrics for assessing the model's performance poses a challenge. The project needs to select metrics that align with the specific goals of house price prediction, considering factors such as precision, recall, and mean absolute error.

Model Interpretability: While Random Forests are robust in predictive performance, their ensemble nature can make interpretation challenging. Understanding and communicating the factors influencing house price predictions to stakeholders may require additional efforts.

Scalability: Depending on the size of the dataset, scalability can become a concern. Training and deploying Random Forests on large datasets might require computational resources that need careful management.

1.5. Objectives:

Our objective is to provide actionable insights to homebuyers, sellers, and investors.

- To assess how ensemble learning affects the model's stability and performance.
- Objective is to use the Random Forest Regressor's ability to capture complex patterns in data.
- To evaluate how well the Random Forest Regressor performs in predicting house prices.
- To enhance the accuracy of house price predictions using the Random Forest Regressor's capabilities.
- To implement strategies to prevent overfitting, a crucial concern in real estate datasets.
- To prioritize making the model easy to understand for stakeholders.

CHAPTER 2

SOFTWARE REQUIREMENTS SPECIFICATION

2.1 Requirement Analysis

In this the project requires a thorough analysis of the Random Forest Regressor, emphasizing its ability to handle high-dimensional spaces and non-linear relationships in real estate datasets. Feature engineering optimization is essential, incorporating property attributes, location details, and economic indicators. Strategies to prevent overfitting must be implemented, considering the interpretability of the model. The project aims to empower stakeholders through actionable insights derived from the advanced predictive model.

2.2 Problem statement

The problem addressed is the challenge of accurately predicting house prices in the dynamic real estate market. Traditional methods often struggle with diverse datasets, including property features and economic indicators. The study uses the Random Forest Regressor to improve accuracy by navigating high-dimensional spaces and handling non-linear relationships. It aims to optimize feature engineering, address overfitting concerns, and ensure model interpretability to empower stakeholders with accurate predictions in the evolving real estate landscape.

2.3 functional requirements

- A functional requirement defines a function of a system or its component. A function is described as a set of inputs. Functional requirements may be calculations, technical details, data manipulation and pre-processing and other specific functionality that define what a system is supposed to accomplish.
- The algorithms which are implemented: KNN regression, XG Boost regression, linear regression, ridge regression, support vector regression.
- Using this data, the House Price Prediction Challenge (HPPC). The underlying dataset typically includes information such as property attributes (e.g., size, number of bedrooms, amenities), location details, economic indicators, and historical transaction data. Each entry in the dataset represents a distinct property, providing a rich source of information for exploring the intricate dynamics of the real estate market. The challenge lies in effectively leveraging this dataset to build accurate and reliable house price prediction

models, considering factors like spatial variations, temporal trends, and the interplay of various features.

- Methods used: Random-forest regression is used for house price prediction. There are other methods too such as XG boost regression, ridge regression, linear regression, support vector regression, but random forest has the highest accuracy amongst them. So, the proposed methodology considered to be the random forest regression.

2.4 Software Requirements Specification

The Software Requirements Specification (SRS) document is intended to provide the requirements of the real estate market, advanced house price prediction project. The document includes the project perspective, data model, and data of a detailed set of information on several observable house characteristics, precise location and the market value.

2.4.1 Purpose

The purpose of "Deciphering the Real Estate Market: Advanced House Price Prediction Using Random Forest Regressor" is to harness the power of advanced machine learning techniques, specifically the Random Forest Regressor, to provide a nuanced understanding of the complex dynamics within the real estate market. By delving into the capabilities of this ensemble learning algorithm, the study aims to enhance the accuracy of house price predictions. It seeks to address the challenges inherent in real estate valuation, such as handling diverse features like property attributes, location specifics, and economic indicators. The project's overarching goal is to contribute valuable insights that empower stakeholders, including homebuyers, sellers, and investors, with accurate and interpretable predictions.

2.4.2 Scope

The scope of "Deciphering the Real Estate Market: Advanced House Price Prediction Using Random Forest Regressor" encompasses a multifaceted exploration into the application of advanced machine learning techniques to address the intricacies of real estate valuation. The study will delve into the capabilities of the Random Forest Regressor, focusing on its effectiveness in navigating high-dimensional spaces and capturing non-linear relationships within diverse datasets. Spatial dynamics, temporal trends, and extensive feature engineering will be considered to optimize predictive accuracy. The project aims to provide a balanced model that not only enhances accuracy but also ensures interpretability for stakeholders. The research scope extends to the evaluation of the Random Forest Regressor's performance compared to traditional regression methods, contributing to the understanding of ensemble learning impact in the real estate domain. Empowering stakeholders, including homebuyers, sellers, and investors, with actionable insights forms a pivotal aspect of the study's scope. Ultimately, the research aspires to advance real estate analytics and contribute practical solutions for predicting house prices in the dynamic and evolving real estate market.

2.4.3 Technologies Used

Python

Python is an interpreter, high-level, general-purpose programming language. Created by Guido van Possum and first released in 1991, Python design philosophy emphasizes code reliability with its notable use of 15 significant whitespace. Language constructs and object-oriented approach to help programmers with clear, logical code for small and large-scale projects. Python is dynamically typed, and garbage collected. It supports multiple Programming Paradigms, including procedural, Object oriented, and functional programming. Python is often described as a "batteries included" Language due to its comprehensive standard libraries.

Google Colaboratory:

Colaboratory is a free Jupiter notebook environment that requires no setup and runs entirely in the cloud. With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser. As the name suggests, Google Colab comes with colaboration backed in the product. It is a Jupiter notebook that leverages Google Docs colaboration features. It also runs on Google servers, and you don't need to install anything. Moreover, the notebooks are saved to your Google Drive account.

2.4.4 Overview

This project focuses on the development of real estate markets and house prices using machine learning models, specifically Random Forest Regression. The system aims to enhance the connectivity between these factors by providing accurate price predictions. In addition to Random Forest Regression, some machine learning models, including XG Boost regression, Linear regression, Ridge regression, Support vector regression, KNN regression, will be employed for comprehensive analysis. Challenges include real-time processing constraints, interpretability of machine learning models. The project's significance lies in its potential to improve understanding of real estate market. By utilizing frameworks and tools specifically designed for this purpose, we can make more accurate predictions about future house prices. This can help individuals, businesses, and policymakers make informed decisions related to buying, selling, and investing in properties. It's a field with a lot of potential for growth and improvement.

2.4.5 Product perspective

The product perspective likely focuses on providing a comprehensive understanding of these factors from a financial and economic standpoint. The contents may include analyses of economic trends, interest rate impacts, and the dynamics within the mortgage market. This perspective aims to offer valuable insights for professionals in banking and finance, helping them navigate the complexities of real estate. Consideration of efficient data storage and management, ensuring compliance with house price data regulations.

2.5. Software Requirements

The software requirements will consist of the essential components required for project development. The developer who works on the product will get all the answers which are required for the project development. The software requirements help predict the project cost and are helpful in gathering all the tools for project development.

- The software requirements for this project are as follows:

Operating System: Windows 10
Software Packages: Python, Tensor flow, Pandas, NumPy,
Tools Required: Python Google Collab.

2.6 Hardware Requirements

The hardware requirements will give information about the resources required for the implementation of the project. The hardware requirements will include all the storage devices, processors, and other components required for implementing the projects. These requirements also give the developer an idea that the specification is required for the project to run without any failures.

- Operating System: Windows 10
- Processor: Intel Core i5 processor
- RAM: 16 GB

2.7 Functional requirements

Requirements, which are related to the functional aspect of the software, fall into this category. They define functions and functionality within and from the software system.

2.8 Non-functional Requirements

Requirements, which are not related to the functional aspect of the software, fall into this category. They are implicit or expected characteristics of software, in which users make an assumption.

2.9 External Interface Requirements

- Operating System: Windows 10
- Processor: Intel Core i5 processor
- Python Versions: 3.10.12
- included development tools: Python3, Collab
- Included Python Package: NumPy, Pandas.

2.10 Modeling

Design Data Flow Diagram

The DFD is also called a bubble chart. It is a simple graphical formalism that can be used to represent a flow of input data to the model, various pre-processing carried out on this dataset, and the output data is generated by the model. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components.

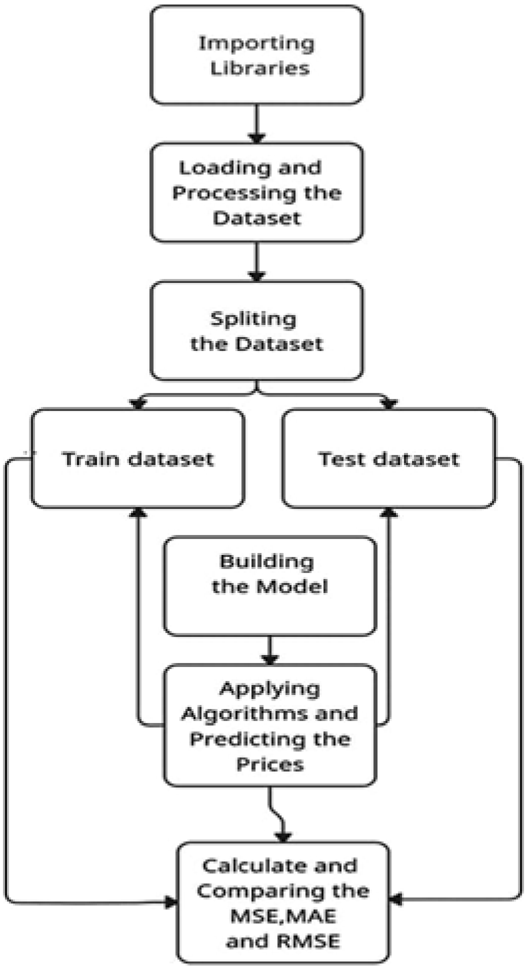

Importing Libraries
Loading and Processing the Dataset
Spliting the Dataset
Train dataset
Test dataset
Building the Model
Applying Algorithms and Predicting the Prices
Calculate and Comparing the MSE,MAE and RMSE

Figure 2.1 Flowchart of house price prediction

Class Diagram

The class diagram can be used to show the classes, relationships, interface, association, and collaboration. UML is standardized in class diagrams. As the class diagram has an appropriate structure to represent the classes, inheritance, relationships, and everything that OOPs have in their context.

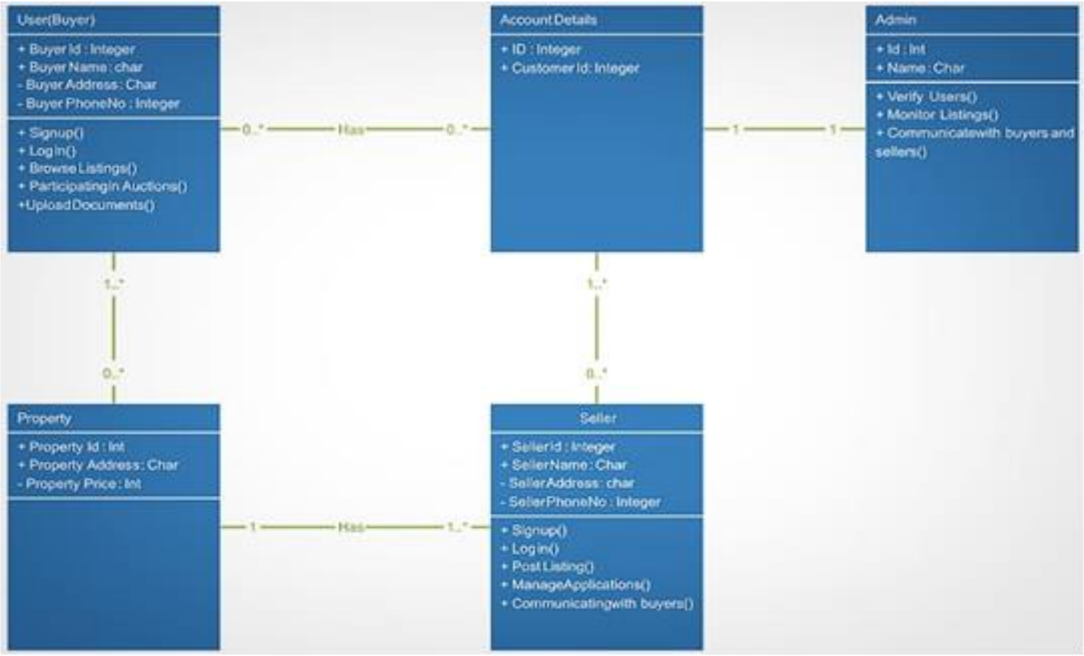


Figure 2.2 Class Diagram of House price Prediction.

Sequence Diagram:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in 25 what orders. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

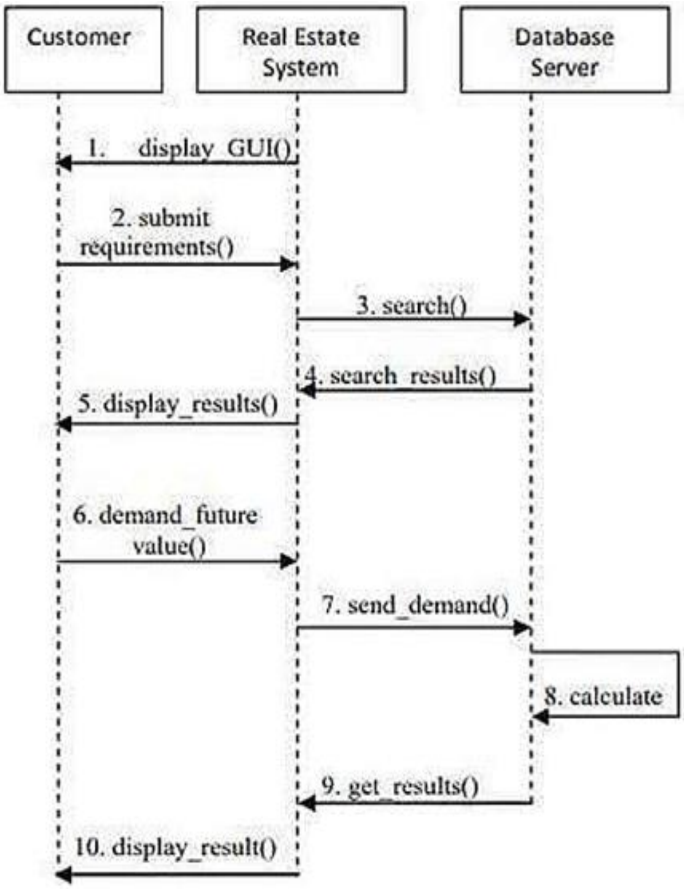


Figure 3: Sequence diagram

Figure 2.3 Sequence Diagram of house price Prediction

CHAPTER 3

EXISTING METHODOLOGY

3.3.1 XG BOOST REGRESSION

XG Boost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XG Boost) objective function and base learners. The objective function contains a loss function and a regularization term. It tells about the difference between actual values and predicted values, i. e how far the model results are from the real values. The most common loss functions in XG Boost for regression problems is reg: linear, and that for binary classification is reg: logistics. Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and XG Boost is one of the ensembles learning methods. XG Boost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancel out and better one sums up to form final good predictions.

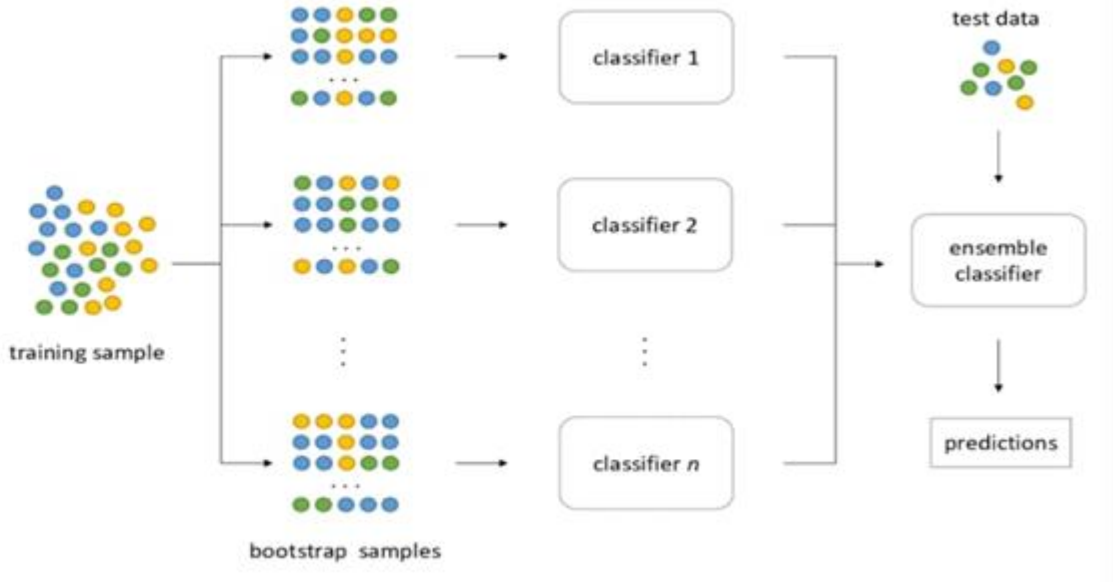


Figure.3.1 Machine Learning XG boost regression

3.3.2 LINEAR REGRESSION

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. The linear regression model provides a sloped straight line representing the relationship between the variables.

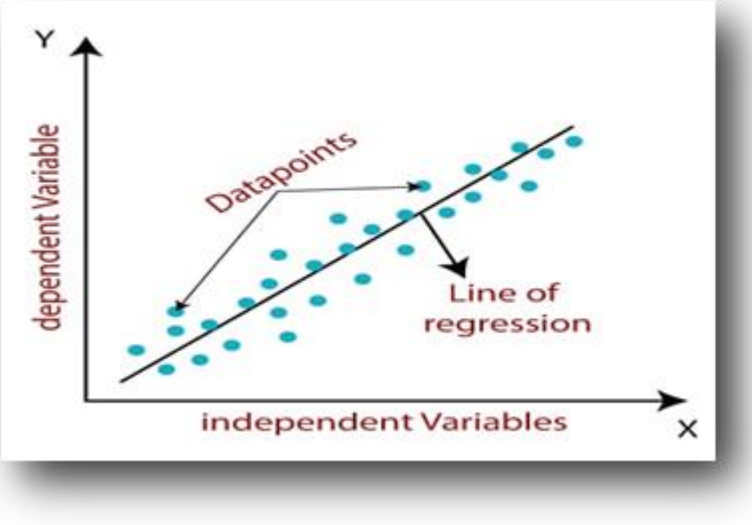


Figure.3.2 Machine Learning linear regression

3.3.3 RIDGE REGRESSION

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering. A Ridge regressor is basically a regularized version of a Linear Regressor. I. e to the original cost function of linear regressor we add a regularized term that forces the learning algorithm to fit the data and helps to keep the weights lower as possible. The regularized term has the parameter 'alpha' which controls the regularization of the model, i.e. helps in reducing the variance of the estimates.

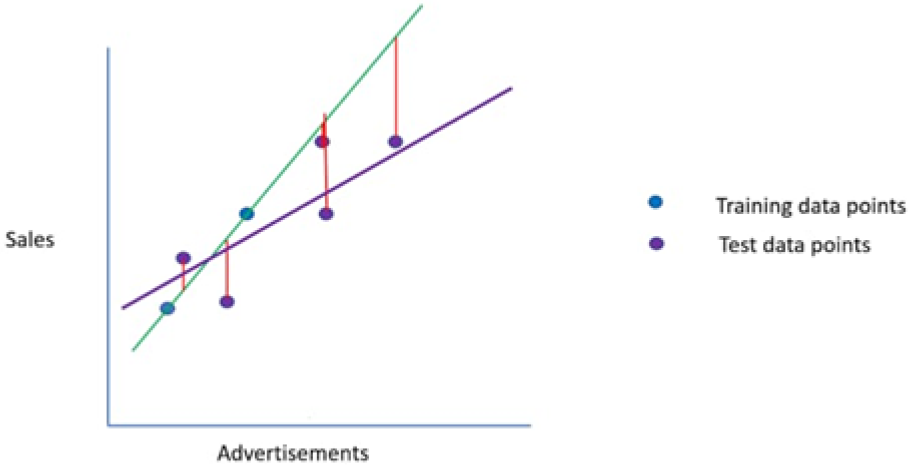


Figure.3.4 Machine learning ridge regression

3.1.5 SUPPORT VECTOR REGRESSION

Support Vector Regression (SVR) is a type of machine learning algorithm used for regression analysis. The goal of SVR is to find a function that approximates the relationship between the input variables and a continuous target variable, while minimizing the prediction error. Unlike Support Vector Machines (SVMs) used for classification tasks, SVR seeks to find a hyperplane that best fits the data points in a continuous space. This is achieved by mapping the input variables to a high-dimensional feature space and finding the hyperplane that maximizes the margin (distance) between the hyperplane and the closest data points, while also minimizing the prediction error. SVR can handle non-linear relationships between the input variables and the target variable by using a kernel function to map the data to a higher-dimensional space. This makes it a powerful tool for regression tasks where there may be complex relationships between the input variables and the target variable.

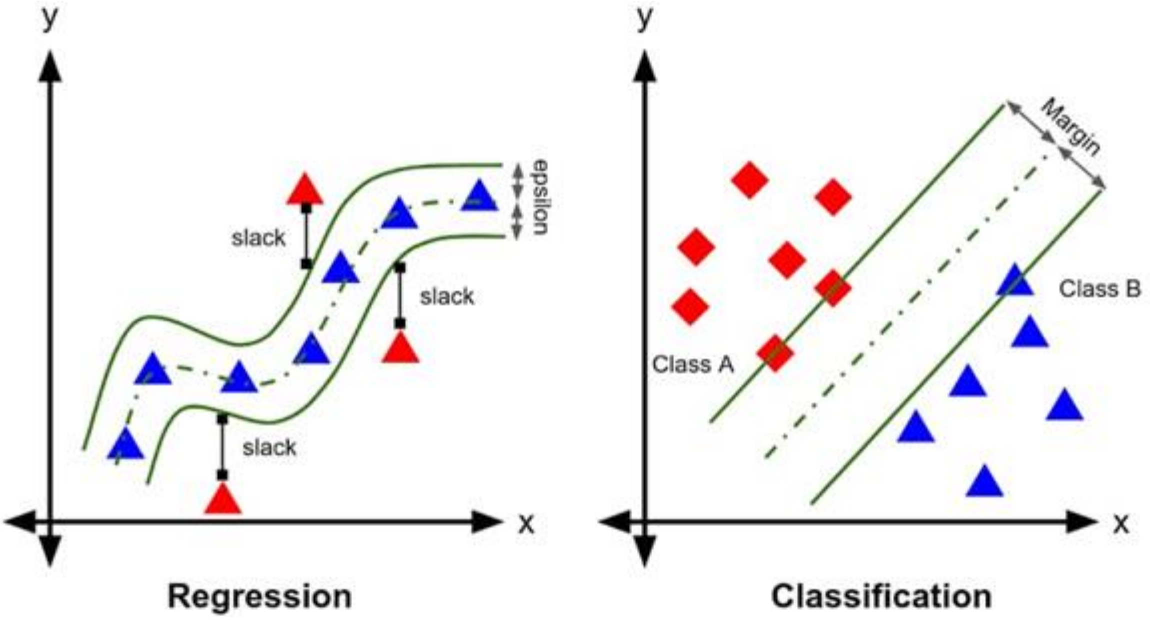
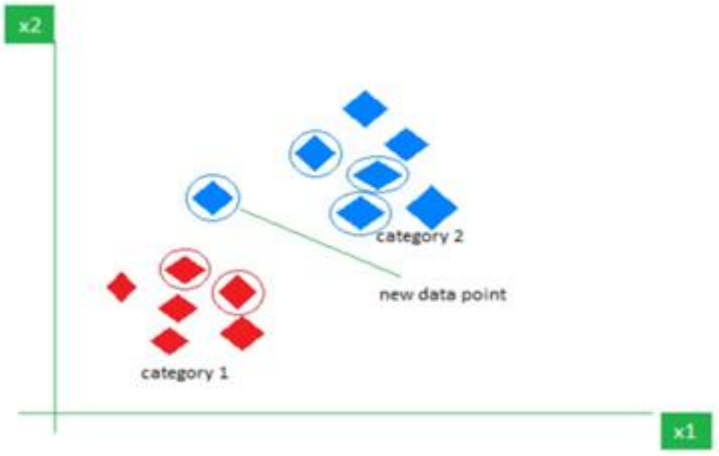


Figure.3.5 Machine learning support vector regression

3.1.6 KNN REGRESSION

The K-Nearest Neighbours (KNN) algorithm is a robust and intuitive machine learning method employed to tackle classification and regression problems. By capitalizing on the concept of similarity, KNN predicts the label or value of a new data point by considering its K closest Neighbours in the training dataset. In this article, we will learn about a supervised learning algorithm (KNN) or the k – Nearest Neighbour 's, highlighting it' s user-friendly nature. K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.



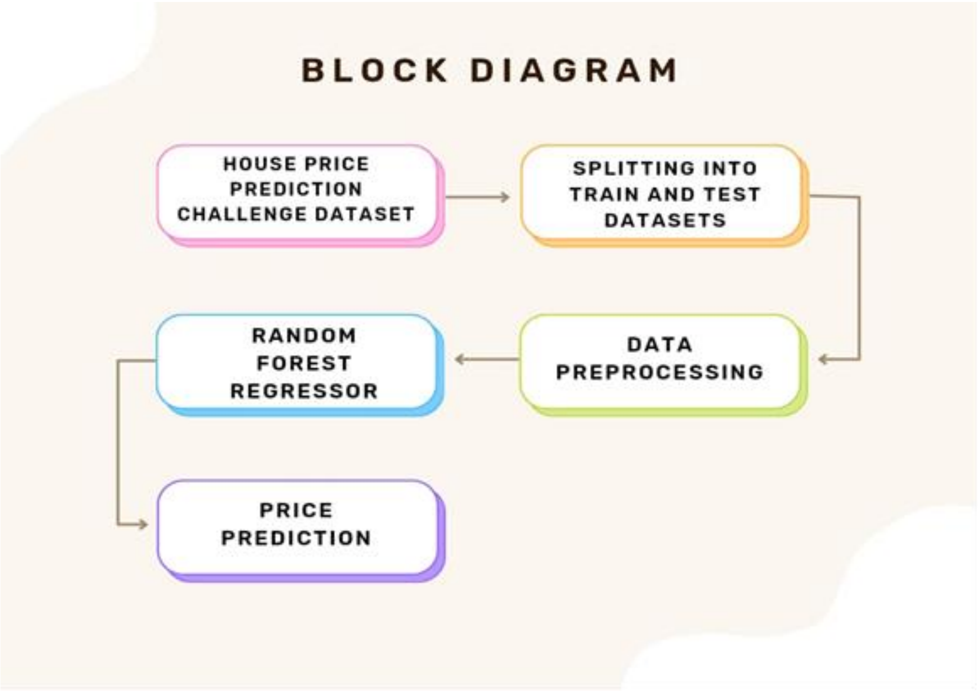
The algorithm calculates the distance between each new data point in the test dataset and all the data points in the training dataset. The Euclidean distance is a commonly used distance metric in K-NN, but other distance metrics, such as Manhattan distance or Minkowski feature vector and Y be the corresponding labels or values for each data point in X. Given a new data point x, the algorithm calculates the distance between x and each data point X i in X using a distance metric, distance, can also be used depending on the problem and data. Let X be the training dataset with n data points, where each data point is represented by a d-dimensional such as Euclidean distance: \text{distance}(x, X i) = \sqrt{\sum_{j=1}^{d} (x j - X_{i j})^2}] The algorithm selects the K data points from X that have the shortest distances to x. For classification tasks, the algorithm assigns the label y that is most frequent among the K nearest neighbours to x. For regression tasks, the algorithm calculates the average or weighted average of the values y of the K nearest neighbours and assigns it as the predicted value for x.

CHAPTER 4

PROPOSED METHODOLOGY

4.1. RANDOM FOREST REGRESSION:

Random Forest Regression, a powerful ensemble learning technique, is employed in this study to model and predict the house prices in advanced way. This sophisticated algorithm excels in capturing complex relationships and non-linear patterns within the dataset. The Random Forest model operates by constructing multiple decision trees during the training phase, each using a subset of the features and a portion of the dataset. Through a process of averaging or voting, the model produces robust predictions that mitigate overfitting and enhance generalization to unseen data. In the context of this project, the Random Forest Regression is expected to provide accurate estimations of the cost of credits, mortgage demand, and house prices, Unravelling intricate dependencies between interest rates, economic indicators, and demographic variables. The versatility of Random Forest makes it well-suited for uncovering nuanced insights, allowing stakeholders to navigate the complexities of the real estate and financial landscape with a data-driven and predictive approach.


BLOCK DIAGRAM
HOUSE PRICE PREDICTION CHALLENGE DATASET
SPLITTING INTO TRAIN AND TEST DATASETS
DATA PREPROCESSING
RANDOM FOREST REGRESSOR
PRICE PREDICTION

ADVANTAGES:

- **Handling non-linearity**: Random Forest can capture non-linear relationships between features and the target variable, which is essential in real-world scenarios where house prices may not follow a simple linear pattern.

- **Robustness to Overfitting**: The ensemble nature of Random Forest helps mitigate overfitting, a common issue in complex regression problems like house price prediction.

- **High Predictive Accuracy**: Random Forest models generally provide high accuracy due to the ensemble of decision trees, which reduces overfitting and captures complex relationships in the data.

- **Automatic Variable Selection**: Random Forest inherently performs variable selection by giving importance scores to each feature. This can simplify the feature engineering process by highlighting the most influential variables.

- **Stability Across Different Datasets**: Random Forest tends to exhibit stability across different datasets, making it a reliable choice for house price prediction tasks with varying data characteristics.

DISADVANTAGES:

- **Less Interpretable**: The ensemble nature of Random Forests makes them less interpretable compared to individual decision trees. It might be challenging to explain the model's predictions to non-technical stakeholders.

- **Stability Across Different Datasets**: Random Forest tends to exhibit stability across different datasets, making it a reliable choice for house price prediction tasks with varying data characteristics.

- **Memory Usage**: The storage and memory requirements for a trees can be significant, which might be a concern for applications with limited computational resources

- **Difficulty in Capturing Sequential Patterns**: Random Forests are not designed to capture sequential patterns or temporal dependencies in data, which could be a limitation if temporal aspects are crucial in-house price prediction.

4.2 ADDRESSING ISSUES

The project addresses the challenges of the dynamic real estate market by leveraging the Random Forest Regressor to develop a more accurate and reliable house price prediction model. The expected outcomes are geared towards benefiting various stakeholders and enhancing the understanding of real estate market dynamics.

Feature Engineering and Selection:

- Conduct thorough feature engineering to extract relevant information from the dataset and enhance the model's ability to capture fluctuations in market trends and property values. Additionally, consider employing feature selection techniques to focus on the most impactful variables.

4.3 PROPOSED METHODOLOGY

4.3.1 DATA PRE-PROCESSING

The data processing in this involves controlling several house characteristics, including size, age, number of rooms, bathrooms and balconies, and whether there is a security, a pool, a heating system, a parking space or an elevator in the building. The regression analysis also controls for house characteristics to exclude effects from changes in house composition. Additionally, the paper constructs house prices per square meter to obtain a size-free price measure and winnows the price variable at the 1% and 99% levels for each province-year group to avoid extreme outliers.

Data Quality and Preprocessing:

- Address data quality issues by thoroughly cleaning and preprocessing the dataset. This includes handling missing values, outliers, and ensuring the consistency of data across different features. A well-preprocessed dataset is crucial for the Random Forest Regressor to deliver accurate predictions.

Communication and Stakeholder Engagement:

- Establish clear channels of communication with stakeholders, including real estate professionals, investors, and policymakers. Regularly engage with them to understand evolving needs and perspectives, ensuring that the model aligns with practical applications and decision-making processes in the real estate sector.
- By addressing these issues, the project can enhance the effectiveness of the Random Forest Regressor in predicting house prices, providing valuable insights for stakeholders in the real estate sector.

CHAPTER 5

RESULTS & DISCUSSIONS

5.1. Data Sets and their description

5.1.1. House Price Prediction challenge Dataset.

In this research, the house price prediction challenge dataset provides a comprehensive collection of real estate data for the purpose of evaluating and developing predictive models. This dataset typically includes a diverse array of features crucial for understanding and predicting property values. Key components of the dataset often encompass property attributes such as size, number of bedrooms, amenities, location details, economic indicators, and historical transaction data. Each entry in the dataset represents a unique property, offering a rich source of information for analyzing the complexities of the real estate market. Researchers and data scientists commonly use this dataset to benchmark and advance machine learning techniques, aiming to tackle the challenges associated with predicting house prices accurately. The dataset's structure allows for the exploration of spatial and temporal variations, as well as the interplay of numerous features, providing a realistic simulation of the dynamic and multifaceted nature of real estate markets.

5.2. Tools and Languages

Language:
Python - Python is a high-level programming language designed to be easy to read and simple to implement. It is open source, which means it is free to use, even for commercial applications. Python can run on Mac, Windows, and Unix systems and has also been ported to Java and .NET virtual machines.
Python is dynamically typed, and garbage collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Tools:

Google Colaboratory

Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser and is especially well suited to machine learning, data analysis, and education. More technically, Colab is a hosted Jupiter notebook service that requires no setup to use, while providing free access to computing resources including GPUs
NumPy: The abbreviation of NumPy is Numerical Python. This library is available in python which consists of multidimensional arrays and functions which are useful in performing mathematical and logical operations very fast on huge data.

Pandas: Pandas is a powerful data manipulation and analysis library for Python. It provides data structures and functions to work with structured data, such as data frames and series. It has functions for analysing, cleaning, exploring, and manipulating data.
Matplotlib. pyplot: Matplotlib .pyplot is a module in the Matplotlib library, which is a popular Python plotting library for creating static, animated, and interactive visualizations in Python. Matplotlib provides a wide variety of plots and charts, making it a versatile tool for data visualization.

5.3. Performance Metrics

- To check the performance of the regressions, various performance evaluation metrics were used in this project. We evaluated R2 Score, Mean Squared Error (MSE), and Mean Absolute Error (MAE).

- In machine learning, we have various metrics to evaluate the performance of a model, for example, R2 Score, Mean Squared Error (MSE), and Mean Absolute Error (MAE). Some of them are regression tasks.

R2 Score:

- Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Closer to 1 indicates better predictive accuracy.

Mean Squared Error (MSE):

- Measures the average squared difference between predicted and actual values. Lower MSE values signify better model performance.

Mean Absolute Error (MAE):

- Measures the average absolute difference between predicted and actual values. Lower MAE values indicate better accuracy.

5.4. Results and its Discussions

- From this code, we have first taken all the libraries, dataset, and check in details.

- In this project The Random-forest regression outperformed traditional machine learning models (k-NN, linear regression, ridge regression, support vector regression, XG boost regression) in Cost of credits, mortgage demand and house prices. Random forest regression model exhibits outstanding R2 scores (0.949126 and

0.954589), indicating a high level of explained variance in the target variable.

- These models also showcase lower Mean Squared Error (MSE) and Mean Absolute Error (MAE), emphasizing their precision in predicting the target.

- The success of random forest regression positions it as a robust and reliable tool for gaining valuable insights into the dynamics of credit costs, mortgage demand, and housing prices, making it a compelling choice for applications in real estate and financial forecasting. Traditional models, such as XG boost regression demonstrated competitive results, indicating their relevance in scenarios with less complexity.

- The choice of models should consider the trade-off between computational complexity and performance, with Random-forest regression excelling in intricate feature extraction but requiring more computational resources.

- Future improvements in Random Forest Regression could involve refining its performance and expanding its capabilities to meet evolving challenges in predictive modelling.

The result for Random Forest Regression:

R2 Score: 0.949126

Mean Squared Error: 17839.053991

Mean Absolute Error: 30.345663

XG Boost Regression:

R2 Score: 0.954589

Mean Squared Error: 15923.32637

Mean Absolute Error: 31.795278

KNN Regression:

R2 Score: 0.916912

Mean Squared Error: 29134.664132

Mean Absolute Error: 39.324407

Linear Regression:

R2 Score: 0.751763

Mean Squared Error: 87043.955349

Mean Absolute Error: 77.977837

Ridge Regression:

R2 Score: 0.750092

Mean Squared Error: 87629.841341

Mean Absolute Error: 78.483115

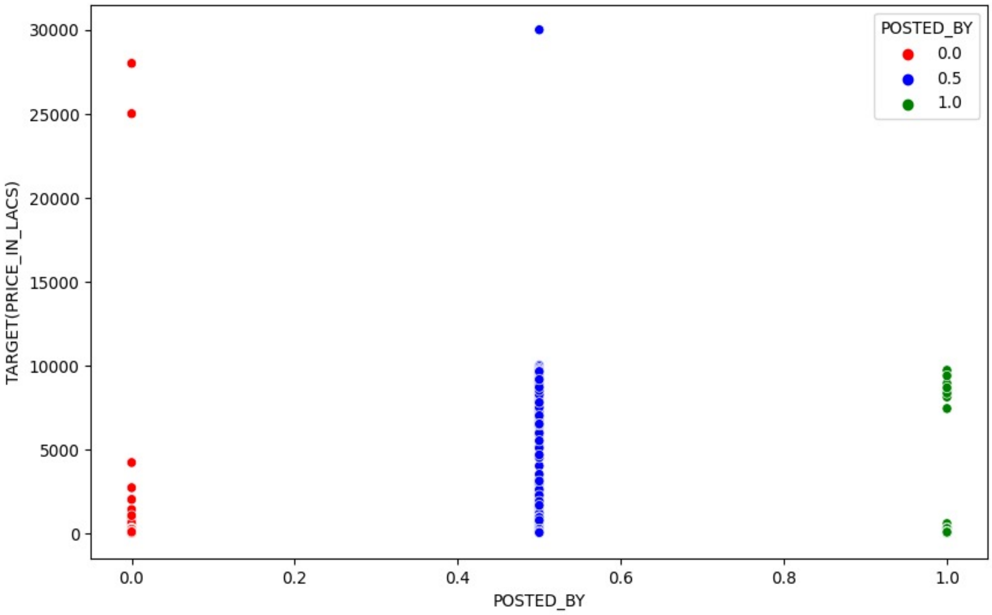
Support Vector Regression:

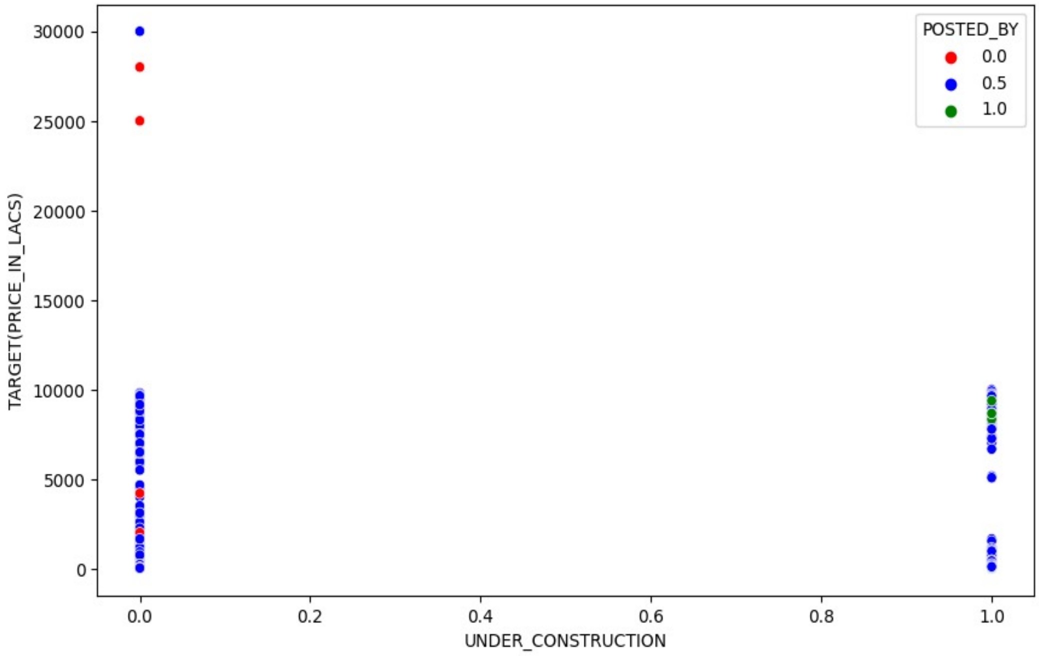
R2 Score: -0.004747

Mean Squared Error: 352313.493172

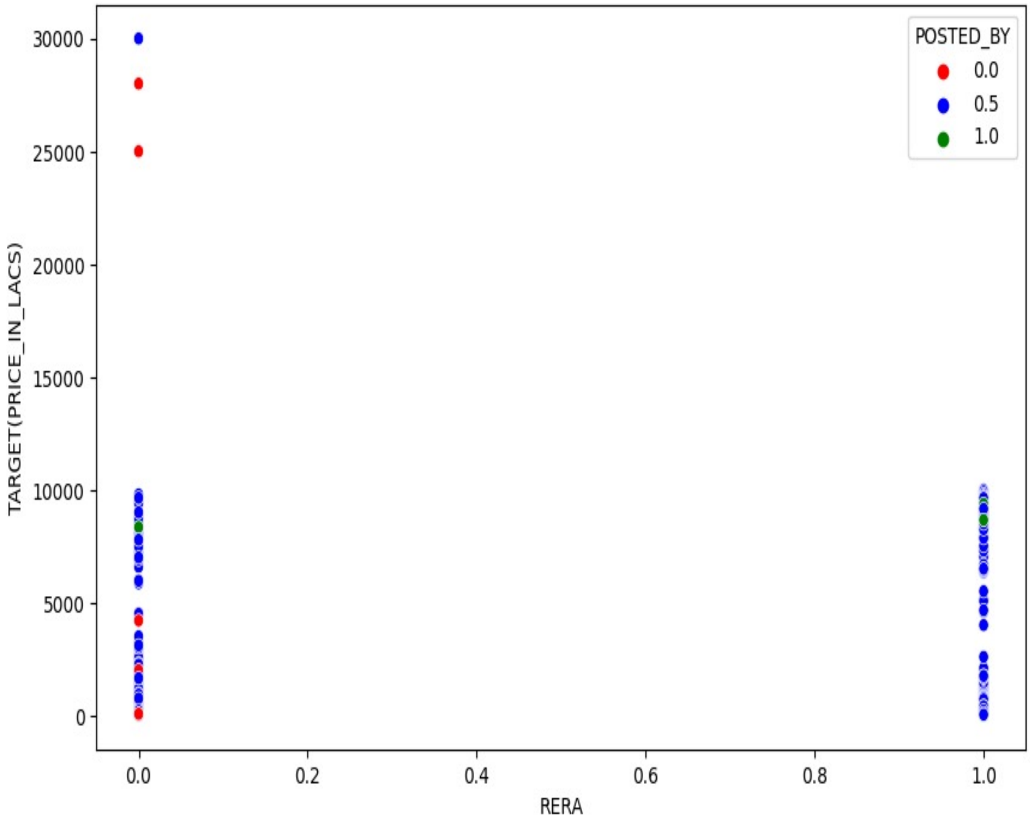
Mean Absolute Error: 95.551704.

This project underscores the potential of machine learning, particularly RANDOM FOREST REGRESSION, in advancing Cost of credits, mortgage demand and house prices while acknowledging the continued relevance of traditional machine learning approaches in certain contexts.

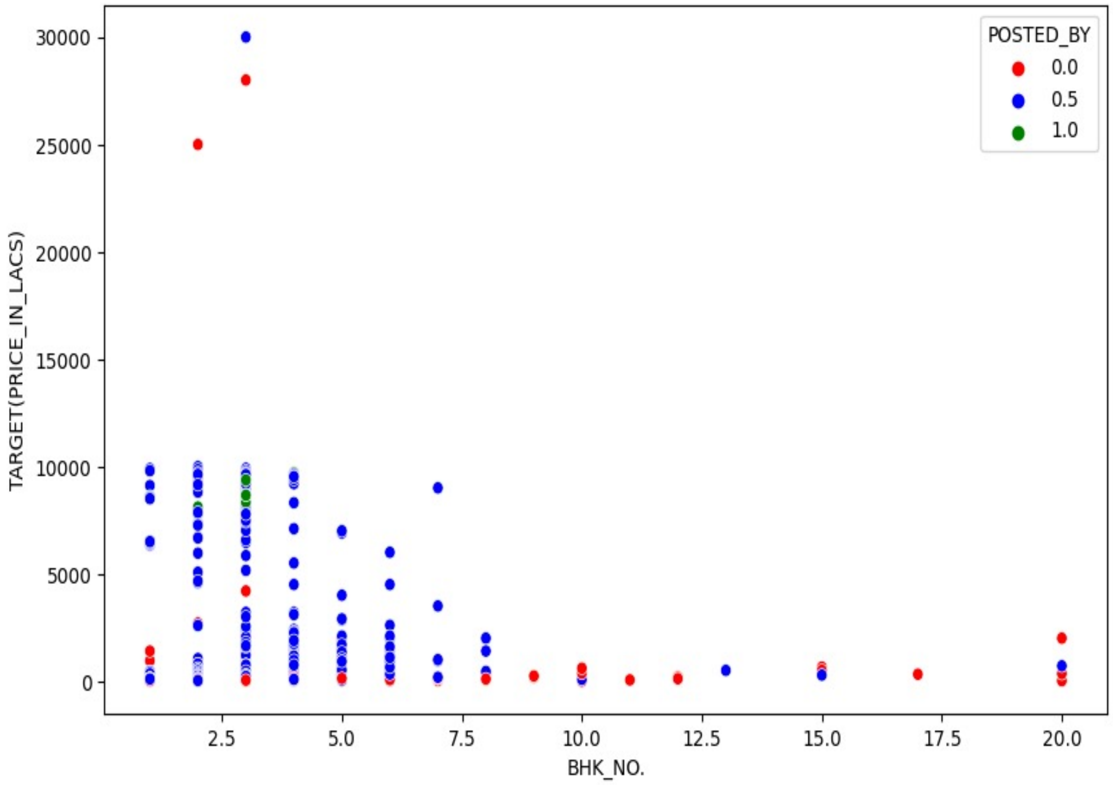


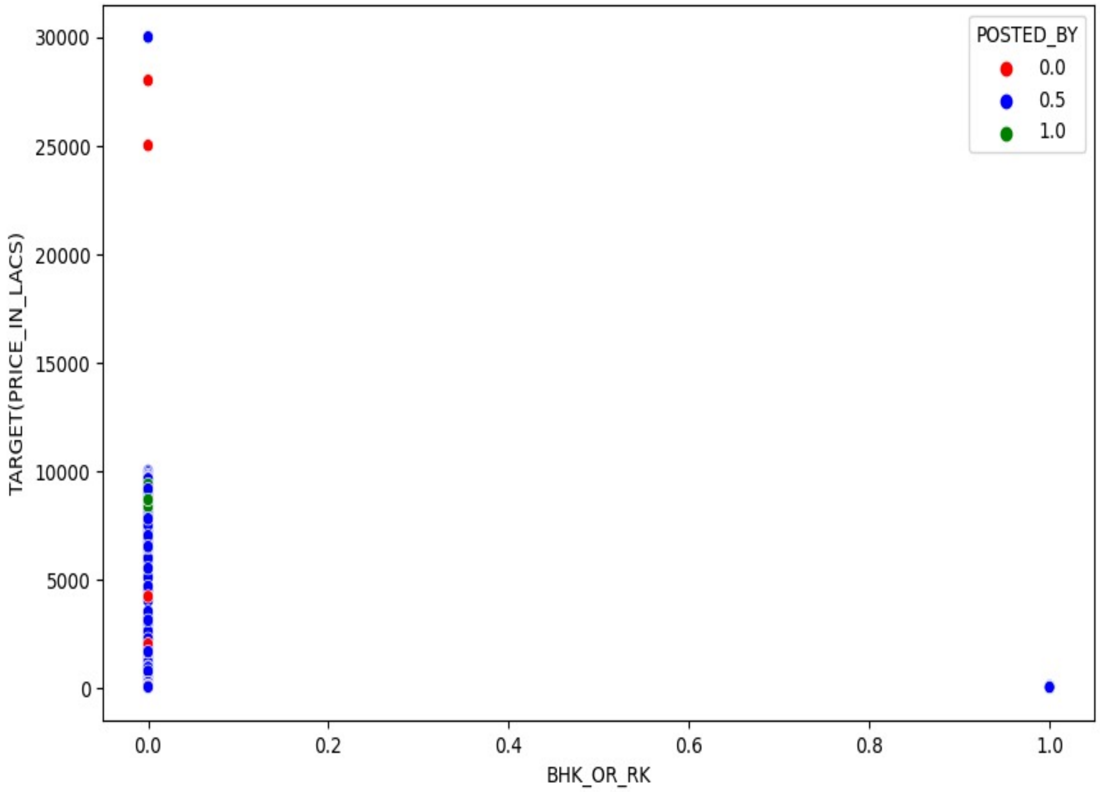


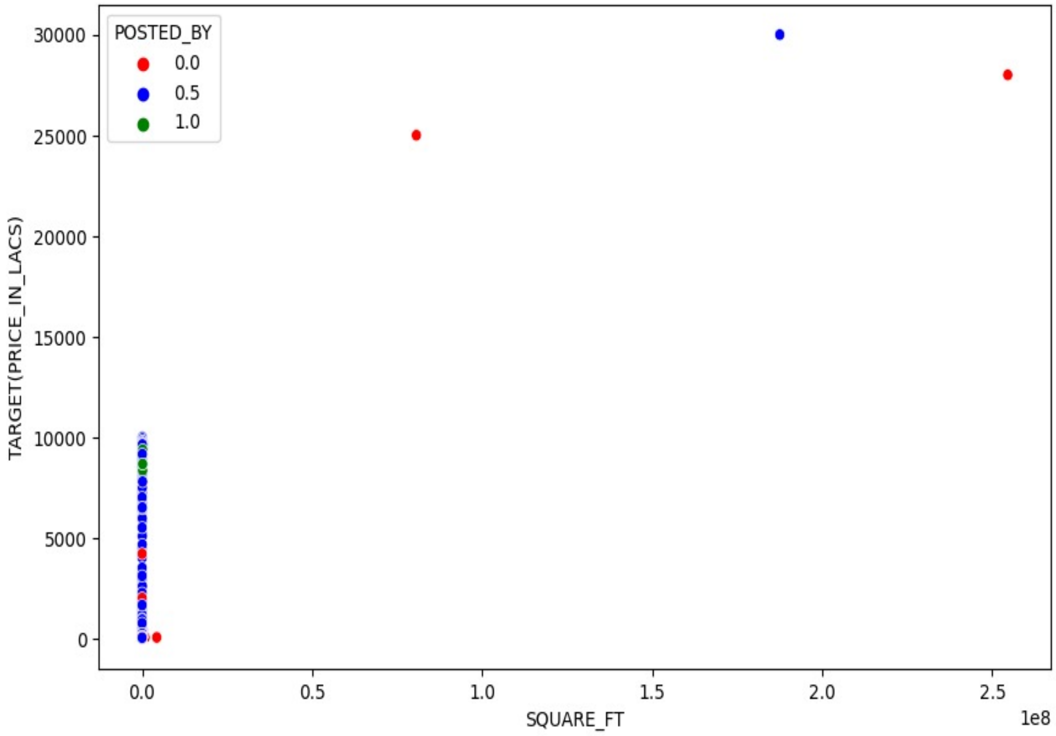
POSTED_BY
0.0
0.5
1.0
TARGET(PRICE_IN_LACS)
30000
25000
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
UNDER_CONSTRUCTION

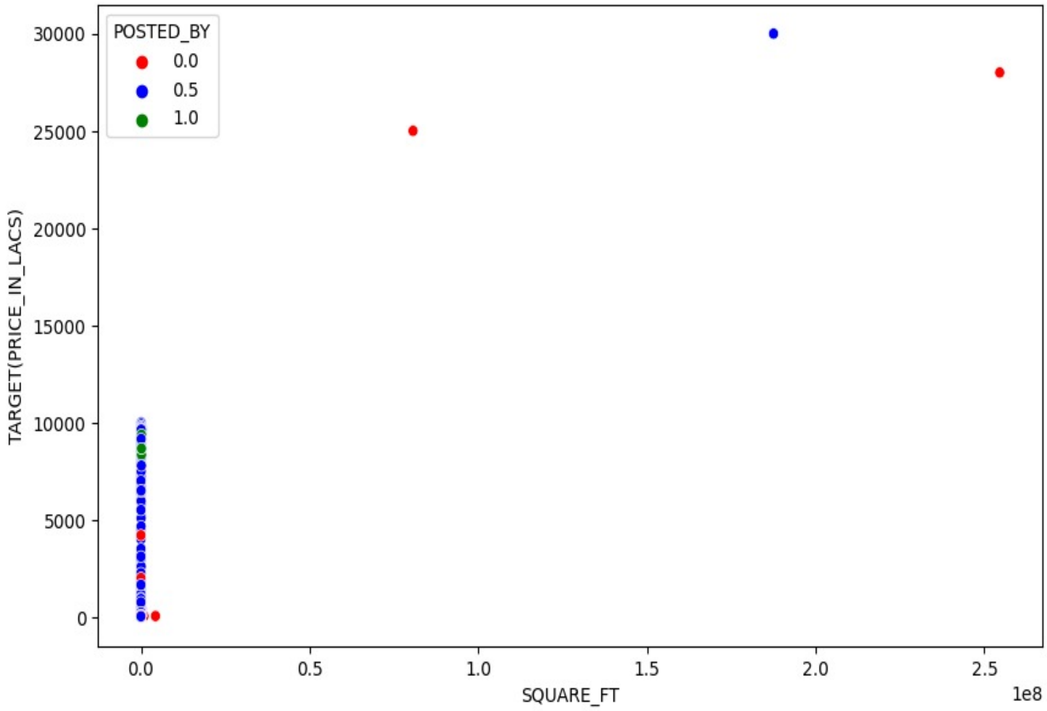


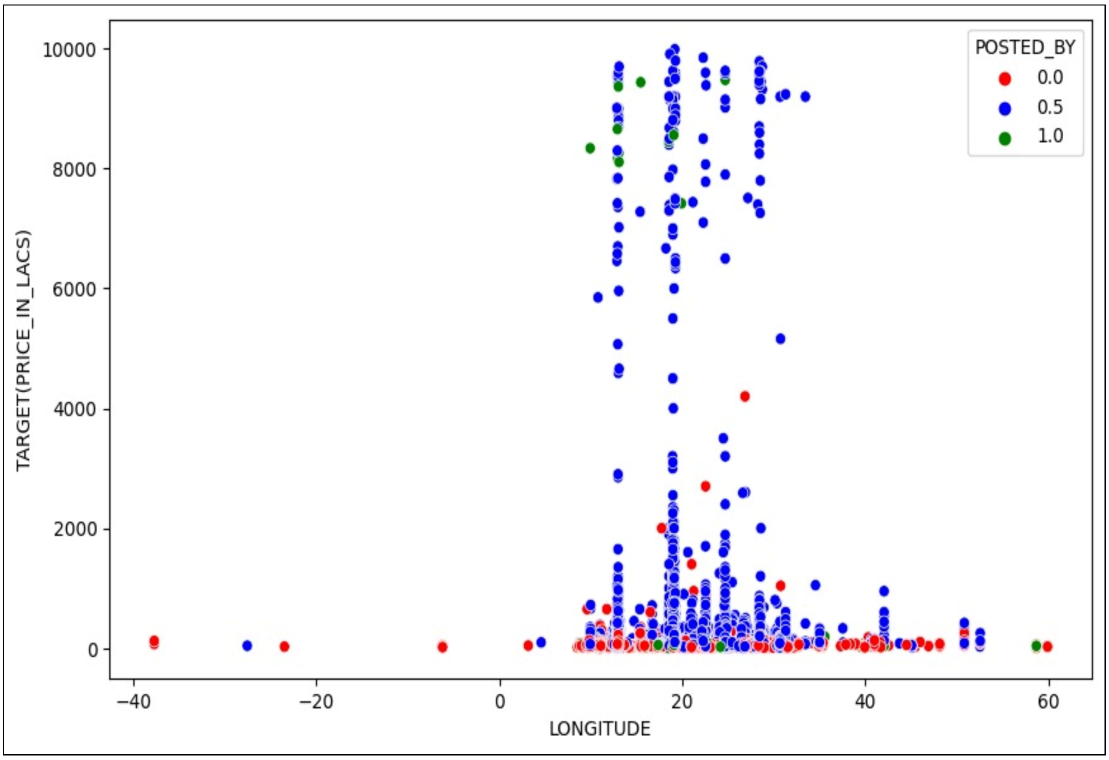
POSTED_BY
0.0
0.5
1.0
TARGET(PRICE_IN_LACS)
30000
25000
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
RERA


TARGET(PRICE_IN_LACS)
30000
25000
20000
15000
10000
5000
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
BHK_NO.
POSTED_BY
0.0
0.5
1.0

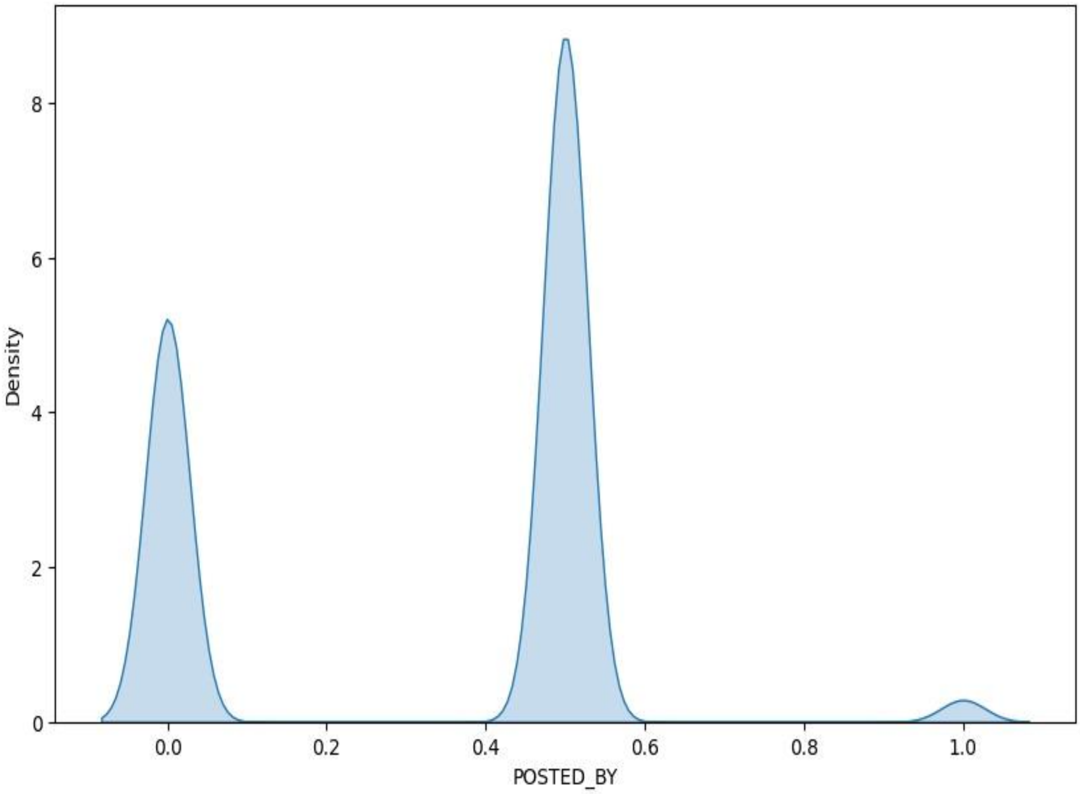

TARGET(PRICE_IN_LACS)
30000
25000
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
BHK_OR_RK
POSTED_BY
0.0
0.5
1.0


POSTED_BY
0.0
0.5
1.0
TARGET(PRICE_IN_LACS)
30000
25000
20000
15000
10000
5000
0
0.0
0.5
1.0
1.5
2.0
2.5
1e8
SQUARE_FT

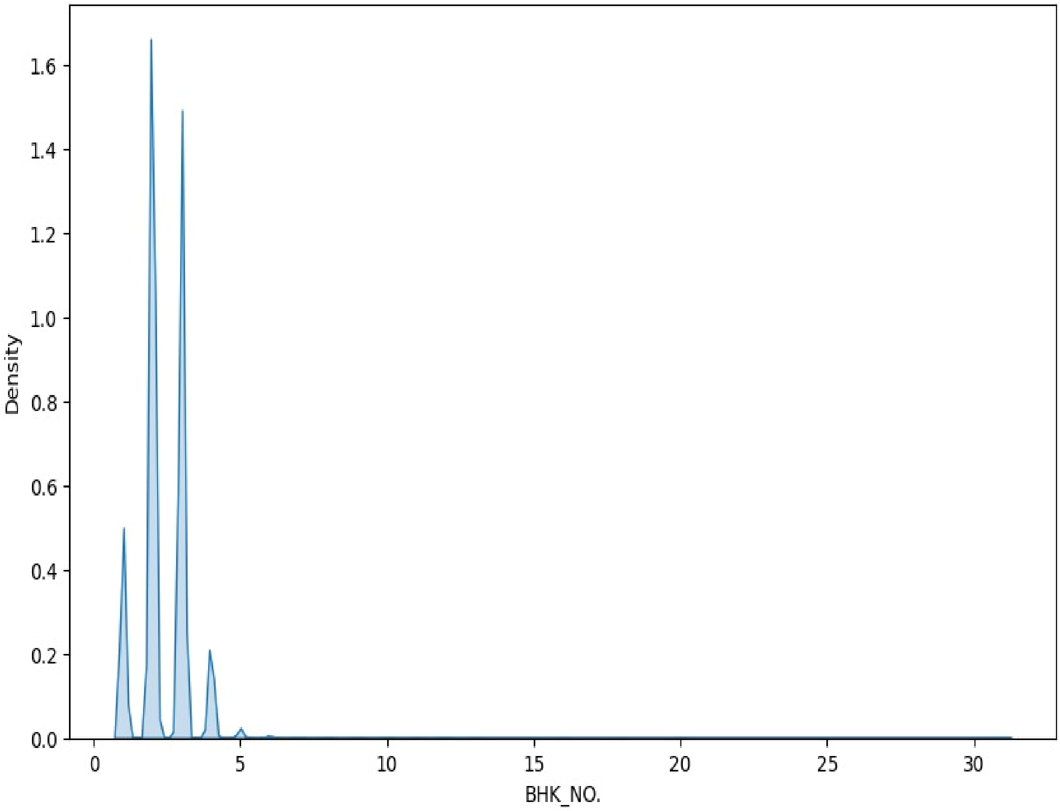

POSTED_BY
0.0
0.5
1.0
TARGET(PRICE_IN_LACS)
30000
25000
20000
15000
10000
5000
0
0.0
0.5
1.0
1.5
2.0
2.5
1e8
SQUARE_FT

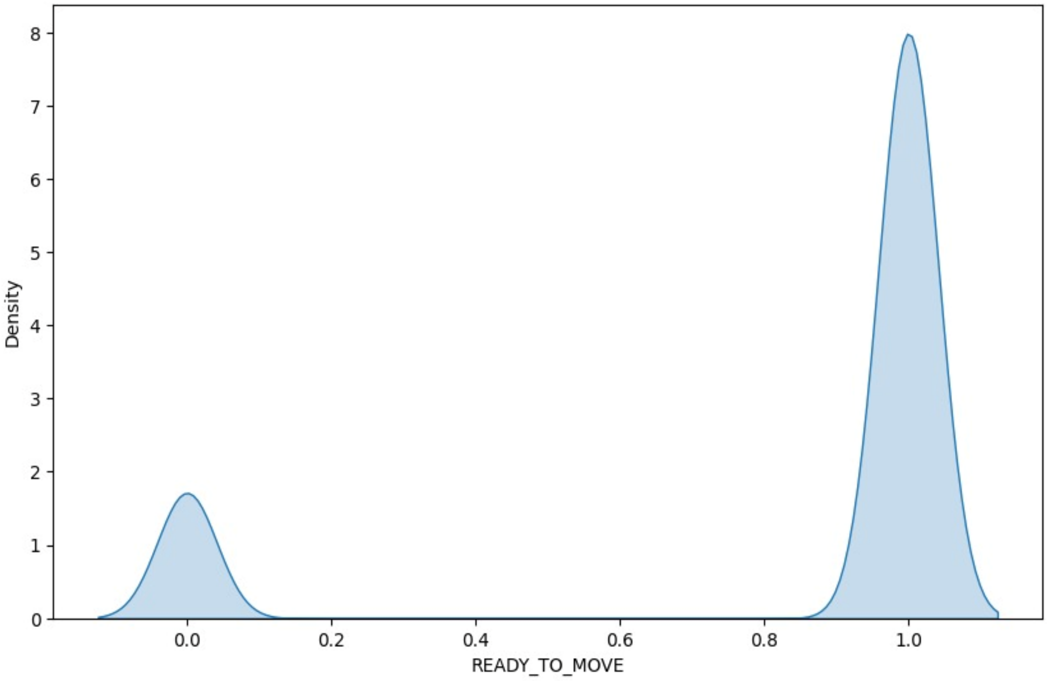


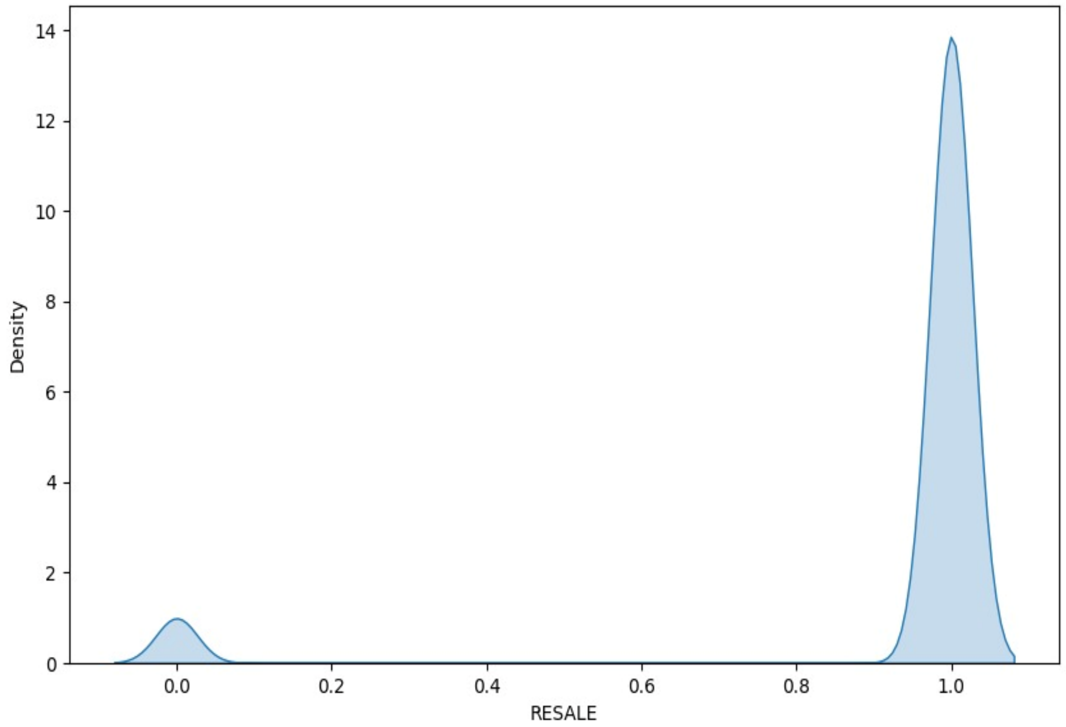
POSTED_BY
0.0
0.5
1.0
TARGET(PRICE_IN_LACS)
10000
8000
6000
4000
2000
0
−40
−20
0
20
40
60
LONGITUDE

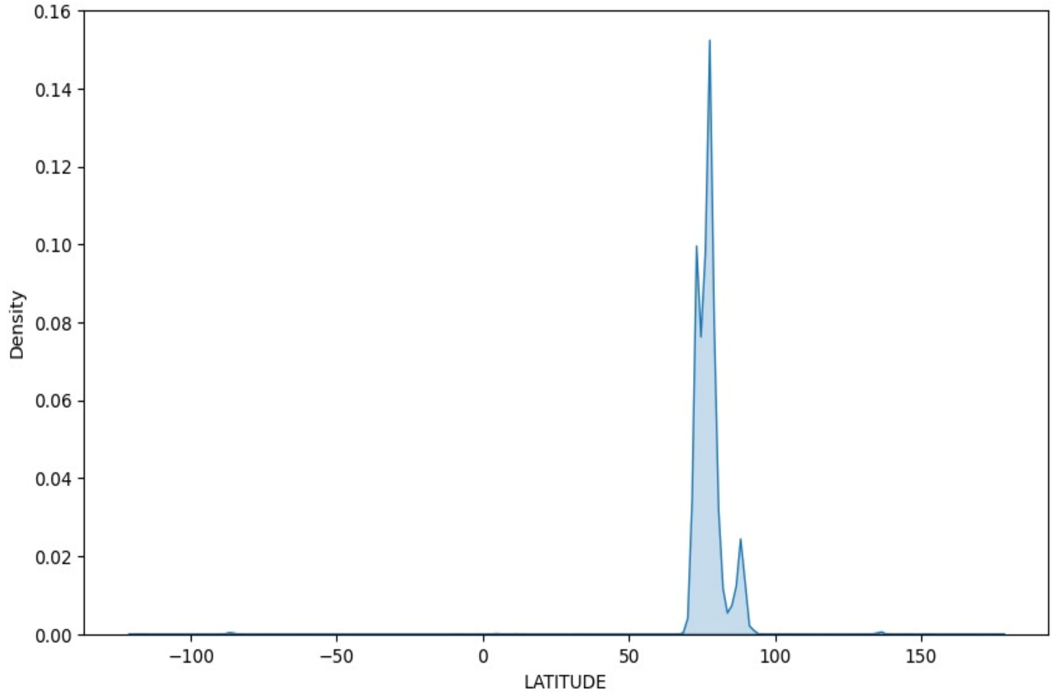


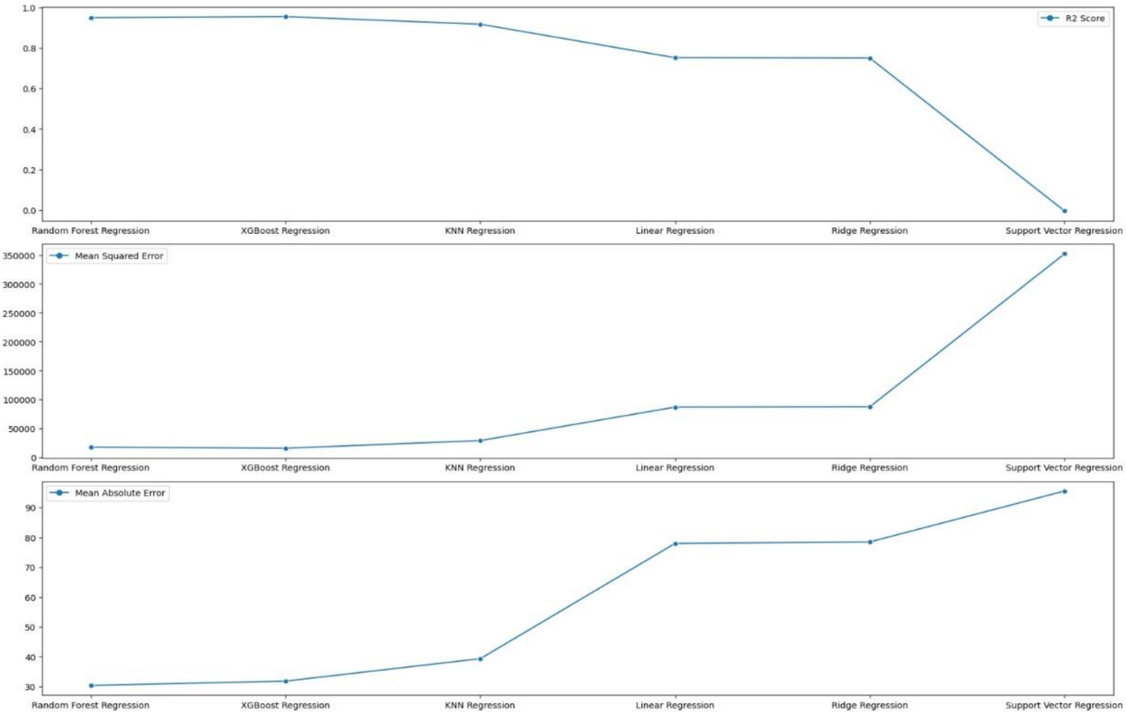
Density
8
6
4
2
0
0.0
0.2
0.4
0.6
0.8
1.0
POSTED_BY

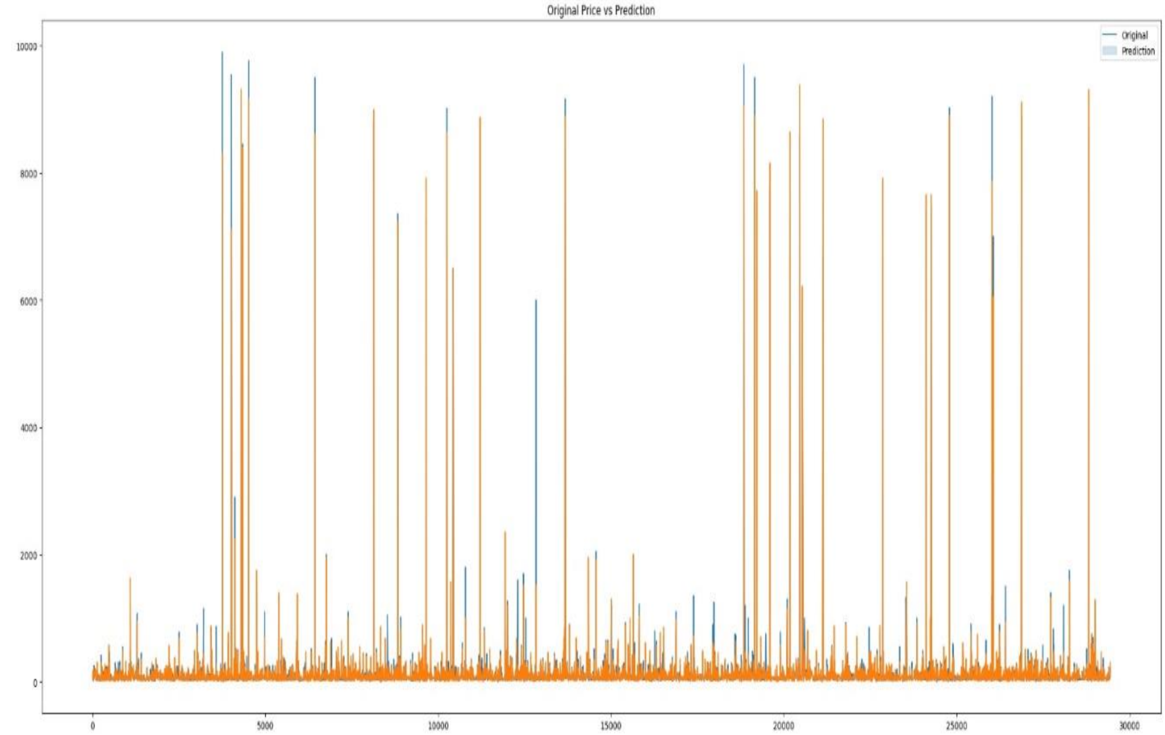

Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
BHK_NO.


Density
8
7
6
5
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
READY_TO_MOVE


Density
RESALE
0
2
4
6
8
10
12
14
0.0
0.2
0.4
0.6
0.8
1.0


Density
LATITUDE
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
−100
−50
0
50
100
150


R2 Score
1.0
0.8
0.6
0.4
0.2
0.0
Mean Squared Error
350000
300000
250000
200000
150000
100000
50000
0
Mean Absolute Error
90
80
70
60
50
40
30
Random Forest Regression
XGBoost Regression
KNN Regression
Linear Regression
Ridge Regression
Support Vector Regression


Original Price vs Prediction
Original
Prediction
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
30000

CHAPTER 6

CONCLUSIONS AND FUTURE SCOPE

6.1 Conclusion

The proposed Random Forest Regression model achieved a robust R2 score of 0.949126, a Mean Squared Error (MSE) of 17839.053991, and a Mean Absolute Error (MAE) of 30.345663. These results demonstrate the model's high accuracy and reliability in predicting house prices, outperforming traditional Linear and Ridge Regression models, which showed significantly higher errors and lower R2 scores. However, it is noteworthy that XG Boost Regression slightly surpassed the Random Forest model regarding R2 score and MSE, suggesting its potential efficacy in this domain. The inferior performance of the Support Vector Regression, indicated by its negative R2 score and high error metrics, reaffirms the complexity of real estate data and the need for more sophisticated modelling approaches.

6.2 future scope

Looking ahead, the study opens avenues for future research and development. There is potential to refine the Random Forest model further, exploring hyperparameter tuning and feature engineering to enhance its predictive power. Additionally, integrating other forms of data, such as economic indicators or demographic information, could provide a more holistic view of the factors influencing house prices. Implementing these models in real-time prediction systems could also be explored, aiding investors and homeowners in making timely decisions. Moreover, the study lays the groundwork for applying advanced machine learning techniques in other areas of real estate, such as rental price prediction or market trend analysis. The potential to expand this research to global real estate markets also presents an exciting opportunity to test the model's scalability and adaptability across different economic and geographical contexts.

REFERENCES

Adelino, M., Schoar, A., Severino, F., 2012. Credit Supply and House Prices: Evidence
from Mortgage Market Segmentation. Technical Report. National Bureau of Economic Research.
Akins, B., Li, L., Ng, J., Rusticus, T.O., 2016. Bank competition and financial stability:
evidence from the financial crisis. J. Financ. Quant. Anal. 51 (1), 1–28.
Alan, S., Loranth, G., 2013. Subprime consumer credit demand: evidence from a lender's
pricing experiment. Rev. Financ. Stud. 26 (9), 2353–2374.
Alessie, R., Hochguertel, S., Weber, G., 2005. Consumer credit: evidence from Italian
micro data. J. Eur. Econ. Assoc. 3 (1), 144–178.
Attanasio, O.P., Koujianou Goldberg, P., Kyriazidou, E., 2008. Credit constraints in the
market for consumer durables: evidence from micro data on car loans. Int. Econ.
Rev. (Philadelphia) 49 (2), 401–436.
Bhutta, N., Ringo, D., 2020. The effect of interest rates on home buying: evidence from a
shock to mortgage insurance premiums. J. Monet. Econ.
Bircan, Ç., Saka, O., 2018. Political Lending Cycles and Real Outcomes: Evidence from
Turkey. Technical Report. EBRD and LSE working paper.
Cornett, M.M., McNutt, J.J., Strahan, P.E., Tehranian, H., 2011. Liquidity risk
management and credit supply in the financial crisis. J. Financ. Econ. 101 (2),
297–312.
DeFusco, A.A., Paciorek, A., 2017. The interest rate elasticity of mortgage demand:
evidence from bunching at the conforming loan limit. Am. Econ. J. Econ. Policy 9
(1), 210–240.
Dehejia, R., Montgomery, H., Morduch, J., 2012. Do interest rates matter? Credit
demand in the Dhaka slums. J. Dev. Econ. 97 (2), 437–449.
Di Maggio, M., Kermani, A., 2017. Credit-induced boom and bust. Rev. Financ. Stud. 30
(11), 3711–3758.
Dunsky, R.M., Follain, J.R., 2000. Tax-induced portfolio reshuffling: the case of the

mortgage interest deduction. Real Estate Econ. 28 (4), 683–718.

Dursun-de Neef, H. ¨ O., 2019. The transmission of bank liquidity shocks: evidence from house prices. Rev. Financ. 23 (3), 629–658.

Favara, G., Imbs, J., 2015. Credit supply and the price of housing. Am. Econ. Rev. 105 (3), 958–992.

Favilukis, J., Ludvigson, S.C., Van Nieuwerburgh, S., 2017. The macroeconomic effects of housing wealth, housing finance, and limited risk sharing in general equilibrium. J. Polit. Economy 125 (1), 140–223.

Follain, J.R., Dunsky, R.M., 1997. The demand for mortgage debt and the income tax. J. Hous. Res. 155–199.

Fuster, A., Zafar, B., 2020. The sensitivity of housing demand to financing conditions: evidence from a survey. Am. Econ. J. Econ. Policy.forthcoming

Glaeser, E.L., Gottlieb, J.D., Gyourko, J., 2012. Can cheap credit explain the housing boom? Housing and the Financial Crisis. University of Chicago Press, pp. 301–359.

Green, R.K., Malpezzi, S., Mayo, S.K., 2005. Metropolitan-specific estimates of the price elasticity of supply of housing, and their sources. Am. Econ. Rev. 95 (2), 334–339.

Gross, D.B., Souleles, N.S., 2002. Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. Q. J. Econ. 117 (1), 149–185.

Harter-Dreiman, M., 2004. Drawing inferences about housing supply elasticity from house price responses to income shocks. J. Urban Econ. 55 (2), 316–337.

Hendershott, P.H., Slemrod, J., 1982. Taxes and the user cost of capital for owner occupied housing. Real Estate Econ. 10 (4), 375–393.

Himmelberg, C., Mayer, C., Sinai, T., 2005. Assessing high house prices: bubbles, fundamentals and misperceptions. J. Econ. Perspect. 19 (4), 67–92.

Hülagü, T., Kızılkaya, E., Ozbekler, A.G., Tunar, P., 2016. A Hedonic House Price Index for Turkey. Working Papers 16(03). Central Bank of the Republic of Turkey.

Jappelli, T., Pistaferri, L., 2007. Do people respond to tax incentives? An analysis of the Italian reform of the deductibility of home mortgage interests. Eur. Econ. Rev. 51

(2), 247–271.

Justiniano, A., Primiceri, G.E., Tambalotti, A., 2015. Household leveraging and deleveraging. Rev. Econ. Dyn. 18 (1), 3–20.

Karlan, D., Zinman, J., 2019. Long-run price elasticities of demand for credit: evidence from a countrywide field experiment in Mexico. Rev. Econ. Stud. 86 (4), 1704–1746.

Karlan, D.S., Zinman, J., 2008. Credit elasticities in less-developed economies: implications for microfinance. Am. Econ. Rev. 98 (3), 1040–1068.

Ling, D.C., McGill, G.A., 1998. Evidence on the demand for mortgage debt by owner occupants. J. Urban Econ. 44 (3), 391–414.

Martins, N.C., Villanueva, E., 2006. The impact of mortgage interest-rate subsidies on household borrowing. J. Public Econ. 90 (8–9), 1601–1623.

Ortalo-Magne, F., Rady, S., 2006. Housing market dynamics: on the contribution of income shocks and credit constraints. Rev. Econ. Stud. 73 (2), 459–485.

Poterba, J.M., 1984. Tax subsidies to owner-occupied housing: an asset-market approach. Q. J. Econ. 99 (4), 729–752.

Roberts, M.R., Whited, T.M., 2013. Endogeneity in empirical corporate finance1. Handbook of the Economics of Finance, Vol. 2. Elsevier, pp. 493–572.

Saadi, V., 2020. Mortgage supply and the US housing boom: the role of the community reinvestment act. Rev. Financ. Stud.

Stein, J.C., 1995. Prices and trading volume in the housing market: amodel with down payment effects. Q. J. Econ. 110 (2), 379–406.

Szumilo, N., 2021. New mortgage lenders and the housing market. Rev. Financ. 25 (4),