Indus Journal of Social Sciences

Online ISSN: 2960-219X
Print ISSN: 2960-2203
Volume 3, Number 2, 2025, Pages 733 – 747
Journal Home Page
https://induspublishers.com/IJSS/about

Enhancing Real Estate Price Forecasting Using Advanced Machine Learning Algorithms: An Empirical Evaluation with Real Market Data

Jamshaid-ul-Hassan[1], Kifayat Ullah[2], Dr. Humera Hayat[3], Shamim Jhatial[4] & Kainat Yousuf[5]

[1]*Fast University of Computer and Emerging Sciences, Email: jamshaidulhassan76@gmail.com*
[2]*Senior lecturer mathematics and statistics Institute of Business Management Karachi, Email: kifayat932@gmail.com*
[3]*Cholistan university of veterinary and Animal Sciences Bahawalpur, Email: humerahayat@cuvas.edu.pk*
[4]*Department of Statistics, University of Sindh Jamshoro, Pakistan, Email: shamimjhatial@gmail.com*
[5]*Lecturer at College Education Department Nawabshah, Sindh, Email: kainat.memon@sbbusba.edu.pk*

ARTICLE INFO

Article History:

Received:	*April*	*05, 2025*
Revised:	*May*	*20, 2025*
Accepted:	*May*	*24, 2025*
Available Online:	*May*	*30, 2025*

Keywords:

Real estate, Price prediction, Machine learning, Regression models, Ensemble methods, Feature importance, Property valuation

Corresponding Author:
Kainat Yousuf
Email:
kainat.memon@sbbusba.edu.pk

ABSTRACT

Predicting real estate prices accurately has become crucial for buyers, investors, and legislators to make wise choices. A viable strategy for identifying the complex patterns and trends in an ever-changing market is to use cutting-edge machine learning algorithms. The purpose of this study is to offer reliable conclusions about the dynamics of the Lahore, Pakistan, real estate market, especially as it relates to inflation and other economic shifts. The goal is to improve forecasting techniques in order to promote a real estate market that is more open and effective. The results are intended to assist in making decisions and to provide guidance for policies that encourage steady, sustained market growth. To forecast housing prices, a number of parametric regression models were used, such as the Extra Trees Regressor, XGBoost, Random Forest, Gradient Boosting, Decision Tree, and CatBoost Regressor. As of June 26, 2023, data set was collected from a public website. Data set comprises on 9,539 listings from 6 districts. Additionally, As of June 26, 2023 in Pakistan real estate market, the average inflation rate in 2024 was 24.76%. The models that performed the best among the ones that were assessed were Gradient Boosting and Extra Trees Regressor, which had the lowest mean squared errors (MSE) and R^2 scores of 85%. Additionally, CatBoost shown competitive performance and is emphasized for its usefulness. The study emphasizes the importance of particular property attributes in predicting prices and advances our knowledge of machine learning applications in real estate forecasting.

Introduction

Recorded data, property value refers to the monetary worth of a piece of real estate in the real estate market using methods like comparing it to similar properties, looking at the investment return, and estimating future earnings (Montes Schütte, & Timmermann, 2024).These strategies often fail to consider the complicated and varied resources in the real estate market (Gerunov, 2022). As a result, the estimates are usually affected by a lot of inefficiency from owners, buyers, investors, and agents. They rely on house price prediction methods that are based on different factors, including some related to the physical characteristics of the properties (Shanthamallu & Spanias, 2022), like size and features, and others that relate to the location of buildings in the construction sector (Hardt & Recht, 2022). Also, the overall condition of the item and the year it was made can influence its final price (Hardt & Recht, 2022).Recently, fake insider methods have become popular for solving complicated problems because there is a lot more data available and new technology has improved (Xu, Zhang, & Analysis, 2023). Counterfeit insights calculations can provide better and more convincing ways to evaluate things than traditional methods. At the event, there isn't an agreed-upon definition of what AI is, according to Xu, Zhang, and Examining, 2023).

Manufactured insights (AI) usually means "machines or people that can watch what's happening around them, learn from it, and then smartly take action or suggest decisions" (Wang et al., Machine Learning (ML) is a modern way of using computers to find, understand, and look at very complex information. Machine learning is an important step forward in the development of computers (Banachewicz & Massaron, 2022). Computer scientists often use rules and data as inputs and their findings as results. But with machine learning (ML), computers receive information and create rules based on that data. So, instead of being completely put together, a machine learning system is ready. Recently, researchers (Hippalgaonkar et al (2023) have found that artificial neural networks (ANNs), which are a type of machine learning, have become popular because they work well and are easy to use (Albahli, Nazir, & Applications, 2024). Artificial Neural Networks (ANNs), also called deep neural networks, have a series of layers made up of trainable units. These layers do not need any changes to their settings except for the size of the network.

A comparative analysis of six standard machine learning methods is presented: Extra Trees Regressor, XGBoost, Random Forest, Gradient Boosting, Decision Tree, and CatBoostRegressor. The goal is to find out which method is faster and more reliable for predicting housing prices in the stock market. Details about lodging costs help us pay more attention to different parts of the Built Environment, like the effects of urban renewal (Coombs et al., 2022) and the importance of protecting the environment Z. Li et al (2022) talked about how appealing natural things can be Chowdhury et al. (2022).A good cost prediction can help reduce the effects of price changes caused by factors like market fluctuations, financial crises, or bankruptcies. This is beneficial for real estate clients and customers. A forecasting model includes different input factors and one or more outcomes, which in this example is the price of a house (Vendor et al., 2023) How accurate predictions are shows how good a model is, while the total number and type of input factors affect how easy it is to use. The more important the factors are, the harder it is to get these details. If it's difficult to determine these factors, the strategy will be less useful and only suitable for a few clients (Borch, Hee Min, & Society, 2022).

The hedonic estimating demonstration (HPM) is the foremost frequently utilized instrument for property assessment (Geiler, Affeldt, Nadif, & Analytics, 2022) HPM, was afterward adjusted tothelodging advertise (Malakouti, Ghiasi, Ghavifekr, & Emami, 2022), to look at the impacts of social, natural, and urban highlights on property values. Since at that point, this strategy has

beenroutinely utilized to relate domestic costs and traits (Haddadin, Mohamed, Abu Elhaija, Matar, Environment, 2023).Different employments for hedonic cost modeling have permitted for the location of relationships that negate real prove: for case (Espey, Lopez, & alter, 2000), who inspected the impact of nearness to the air terminal on private property values, discovered that this area can be taken note as an advantage instead of a persuading calculate (Wu, Zhou, Long, & Wang, 2023). luckily, later budgetary issues and financial occasions have caused critical insecurity within the field of valuation speculations and strategies not as it were at the level of the scholarly world, where novel strategies to esteem creation have been developed but too at the operational scale, due to the clear uncertainty and guess related with customary approaches' comes about (Tajani, Morano, Ntalianis, & Back, 2018).

(Islam, Li, Lee, & Wang, 2022); (NLP); (Karamanou, Brimos, Kalampokis, & Tarabanis, 2024). (Beghi et al., 2019) evaluated numerous ML procedures, counting Random Forest, Ridge Relapse, and Rope, to recognize which procedure performed superior. Their discoveries shown that Arbitrary Woodland (RF) performed the most excellent in terms of exactness. (Gortzak & Ulusoy, 2024) came to the same conclusion, expressing that RF outflanked relapse models in foreseeing lodging values in Virginia (US).Gradient-boosted classifiers have recently won many data science competitions on Kaggle (2019). These methods for improvement use a mix of decision tree models, which can perform better than random forest models. Wu and others studied 40,000 hotel transactions in Hong Kong using a Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM). Their findings showed that the GBM method was the most accurate compared to the others. Also, Mrsic and others (2020) showed that the XGBoost method outperformed Random Forest and AdaBoost, making it the best method for this task (Chaleshtori, 2024); Iftikhar, Qureshi, Zywiołek, López-Gonzales, & Albalawi, (2024).

Artificial Neural Networks (ANN) is becoming more popular because regular computers are getting better at processing data and there are more open-source datasets available for everyone (Law, Shen, & Zhong, 2024); Khan, Qureshi, Daniyal, & Tawiah, K. (2023). Neural systems are now widely used in many fields, including healthcare (Pinconschi, Gopinath, Abreu, & Pasareanu, 2024), finance (Xiao, Zhou, Xiao, Huang, & Xiong, 2024), and agriculture (Abiodun et al., 2018) (García-Magariño, Fox- Fuller, Palacios-Navarro, Baena, & Quiroz, 2020) studied the costs of staying using a type of artificial intelligence called Multi-Layer Perceptron (MLP) neural network, along with other machine learning methods. They discovered that MLP made the fewest errors and was always accurate in its predictions. Similar topics related to ML forecast can be found Qureshi, Khan, Bantan, Daniyal, Elgarhy, Marzo, & Lin, (2022); Qureshi, Ahmad, Ullah, & ul Mustafa, (2023); Iftikhar, Daniyal, Qureshi, Tawiah, Ansah, & Afriyie, (2023).

(Wang et al., 2020) used a back-propagation (BP) neural network to evaluate Chinese property prices. He found that using the model to evaluate real estate prices is both technically viable and credible (Li et al.(2024).several studies in the literature compare property price projections using HPM and ANN models (Jáuregui-Velarde, Andrade-Arenas, Celis, Dávila-Morán, & Cabanillas-Carbonell, 2023). The conclusions are contradictory: according to different research, the advantage of ANN is that it automatically detects non-linear relationships between explanation variables and prices (Rampini, Re Cecconi, & Finance, 2022; Qureshi, Iftikhar, Rodrigues, Rehman, & Salar, 2024).

(Karamanou, Brimos, Kalampokis, & Tarabanis, 2024) argues that the ANN is a "black box" that develops responses rather than a simple functional relationship between input and output values (Gortzak & Ulusoy, 2024). The results are incompatible, although they improve as the sample size increases (Law, Shen, & Zhong, 2024). The marginal prices computed using ANN are more realistic than traditional hedonic pricing; yet, they present important computing challenges. A high

number of neurons, due to over-parameterization, may result in a lack of forecasting capability (Bin, Gardiner, Liu, Li, & Liu, 2023); Ahmad, Qureshi, Iftikhar, Rodrigues, & Rehman, (2025).

Methods and Materials

All simulations and analyses were performed using Python 3.10 and relevant scientific libraries. The methodology was designed to ensure computational efficiency and reproducibility. The last section concludes the paper forecast RE values have emerged in scientific studies, complementing or replacing existing methods. However, advancements in computational technologies, especially in artificial intelligence, require Updated strategies for estimating real estate prices frequently. This paragraph compares a recent RE price estimation method using machine learning models to several algorithms commonly utilized in scientific literature.

The suggested method can be defined using the following steps:

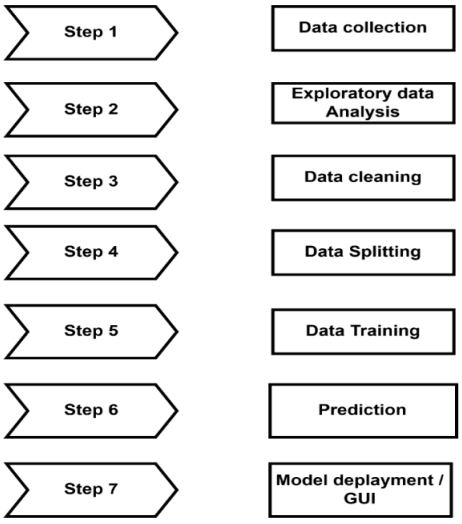

Figure 1: Study Flow Chart

Data Collection: Our project's data source is Zameen.com, one of Pakistan's most popular real estate websites. This platform provides detailed property listings that include vital information such as price and other property details. We have scraped data from this website using Python till 26 June 2023. We have adopted seven locations of Lahore in this study. These locations are:

Table 1: Seven Locations of Lahore

Location	Count
DHA Defence	6117
Bahria Town	2103
Park View City	239
Johar Town	524
Lake City	348
Gulberg	89
Allama Iqbal Town	119
TOTAL	**9539**

Exploratory Data Analysis (EDA): Exploratory Data Analysis (EDA) was conducted to understand the underlying structure, detect anomalies, and identify important patterns within the dataset prior to model development." cleaned and preprocessed our data, Here's an outline of the key EDA techniques. Descriptive Statistics (Figure 2) Detect outliers and anomalies (Figure 3)

Correlation Analysis (Figure 4) Location Analysis (Figure 5). Price Analysis (Figure 6) The EDA deleted invalid values and outliers from the dataset. The raw data includes several records that were presumably inaccurate (Nirala, Singh, & Purani).In ML research, the dataset is typically divided into two groups: training and test sets. The training set is used to tune the model's objects, while the test set is used to ensure the algorithm's ability to be applied to new data. In this study, 80% of the original data was used to train the algorithm and 20% for testing (Lee, Jeong, Lee, Lee, & Choo, 2023).The ready-made dataset was used to train a variety of machine-learning models. Utilizing the training data, the model's parameters were adjusted during this phase, and their performance was optimized. Among the models were the Decision Tree, XGBoost, Gradient Boosting algorithms, etc (Xie et al., 2023). The models' predictions have been evaluated to see which one achieved the highest accuracy (Vyas, 2024; Tawiah et al., 2023). One of the steps in getting the information prepared for machine learning was to concentrate on a particular cost run (Qureshi et al., 2023).

The area data was changed into numbers, giving each area an extraordinary number. One-hot encoding was utilized to bargain with the property sort column by making unused columns with parallel values. This made a difference the demonstrating it and utilizing the distinctive sorts of property data way better. To get complex designs, we made polynomial highlights for critical columns just like the number of rooms, washrooms, measurements, and area. Making a user-friendly visual interface for the genuine domain cost prediction model was a portion of the ultimate steps of the venture (Chen et al., 2024) Clients can put in different property points of interest and get cost gauges because of the easy-to-use and nice- looking interface. The interface had a checkbox to add or evacuate expansion rate changes within the forecasts, alongside choices to select distinctive models just like the Choice Tree Demonstrate, XGBoost Demonstrate, and Slope Boosting Demonstrate. Clients can alter the forecasts to fit them possess needs because of this personalization. Tkinter was utilized to make the client interface, permitting clients to effortlessly investigate distinctive circumstances and get exact toll gauges (Donghi & Morvan, 2023).

Models Description

XGBoost. A machine learning approach called gradient boosting creates a prediction model from an ensemble of weak prediction models, most often decision trees (Zheng et al., 2023). Thus, an ensemble model is made up of several straightforward individual models that collectively produce a stronger one. Fitting an initial decision tree is how XGBoost begins. to the information. Next, a second model concentrates on precisely forecasting the situations in which the initial model doesn't work well. This boosting procedure is carried out repeatedly and Every new model that comes out aims to address the weaknesses of the combined (J.-C. Liu, Chen, Lee, Huang, & Technology, 2024) The ElasticNet model, on the other hand, had only a few tweaks, XGBregressor offered over 10 elements to optimize (Qin & Technology,2024).

The Extra Tree Regressor approach refers to extremely randomized trees. The selection of the ideal cut-point in the context of input features (numerical) is largely responsible for the variation in the induced tree. This is the primary goal of constructing trees arbitrarily. From a statistical perspective, abandoning the bootstrapping notion offers a bias-related benefit. The point of cutoff Generally, randomization results in a very good variance reduction impact. Numerous complicated high-dimensional issues yield optimal outcomes while employing this technique. The Extra-Tree technique yields split-second multilinear approximations from the perspective of functional points as opposed to the intermittent ones of woods at random (Fazli, Alian, Owfi, & Loghmani, 2024).

Random Forest Model: Random forest (RF) is a robust machine learning algorithm that can be used for classification and regression problems (Voshol). RF is a type of bootstrap aggregation

known as bagging, which combines multiple decision trees (Stavropoulos, 2024). The idea is to train multiple trees separately and then integrate their predictions to improve reliability (Choudhary, Anurag, Shukla, & Imaging, 2024).

CatBoost Regressor: The model that demonstrates superior performance compared to other models is Catboost. Therefore, it is highly beneficial to delve into its algorithm in great detail. Catboost is developed using the gradient boosting technique and is constructed iteratively in a greedy fashion. This construction method allows it to represent a sequence of progressively refined approximations (Guan, Qiu, Wang, & Xiao, 2024).

Gradient Boosting Model: Gradient boosting relapse is a machine learning method that puts together weak prediction models (often decision trees) to improve accuracy. numbers related to relapse problems. It corrects the mistake of the earlier show by creating a new model. Show the leftover values from the current data. Finds how steeply the misfortune function goes down. Shows the changes and size of improvements to the display settings. The final expectation comes from combining all the predictions made by the different models. Slope relapse boosting is known for its excellent prediction ability and its skill in managing complex data sets (Hajdu, 2024).

Decision Tree Model: The decision tree regressor uses the features of the data to create a tree- like model that predicts future results. The choice tree regressor learns from the deepest and shallowest parts of a chart, based on the system. Look at the information closely. Lattice Look CV is a way to handle different settings or options. Adjusting settings that will effectively create and evaluate a display for every set of calculation parameters. shown on a grid. In this calculation, the Network Look CV is used to find the best value for maximum depth, which is needed to create the decision tree (Khanmohammadi, Saba-Sadiya, Esfandiarpour, Alhanai, & Ghassemi, 2024)

Exploratory Data Analysis (EDA) is an important part of looking at data because it helps you understand the dataset and its basic features. In this study, we used EDA to find important patterns, connections, and unusual things in real estate data. To start this investigation is very important to identify key factors and how they are spread out, as they can greatly affect how accurate and effective the model is (Dolphin, Smyth, & Dong, 2024). The information includes some real estate details like property location, price, number of rooms and bathrooms, and size in Marlas. An initial investigation discovered a big variation in property prices, with an average value of about 64. 45 million PKR and a standard deviation of 67. 73 million PKR, indicating a large range of values. The variety of home features was shown by the number of rooms and bathrooms, as well as the sizes of the properties, which ranged from small to very large.

	Price	Bedrooms	Baths	Size (Marla)
count	9.539000e+03	9539.000000	9539.000000	9539.000000
mean	6.444756e+07	4.458014	5.077157	14.152039
std	6.773047e+07	1.115661	1.144664	10.553738
min	7.500000e+04	1.000000	1.000000	0.000000
25%	2.600000e+07	4.000000	4.000000	5.000000
50%	4.700000e+07	5.000000	5.000000	10.000000
75%	7.800000e+07	5.000000	6.000000	20.000000
max	1.600000e+09	11.000000	10.000000	200.000000

Figure 2: Descriptive Statistics

To ensure the data is good and reliable, several preparation steps were taken. Missing values for important things like 'Bedrooms', 'Baths', and 'Size (Marla)' were filled in with the average to keep the data consistent. In addition, very high values in the 'Price' variable were limited to the 99th percentile. This was done to reduce the impact of outliers and avoid distorted results. Using a logarithmic change on the 'Price' variable was a simple step in preparing for the data analysis. Because property prices can be very uneven, this change modified the data to reduce the impact of unusual values, making the dataset easier to work with for predictions. The data was changed using the equation exp(x) - 1 to bring the expected values back to their original scale. This made sure the numbers were easy to understand and could be compared to actual property prices.

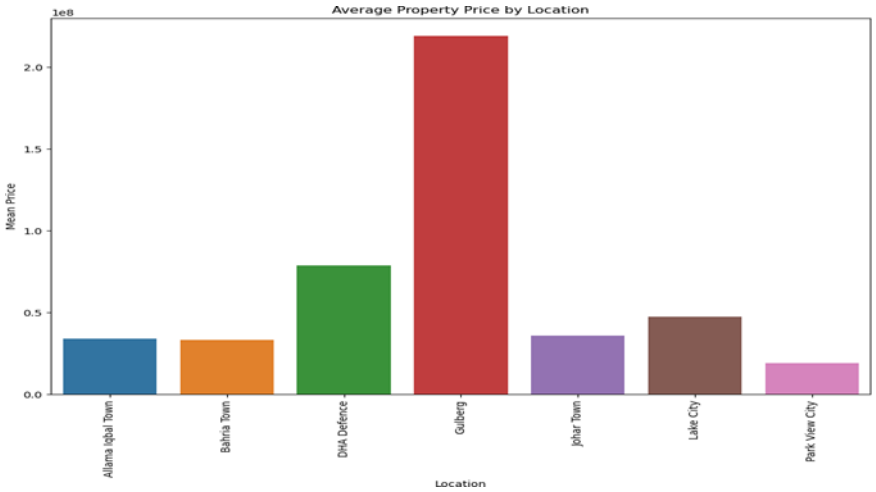

Figure 3: Average Property Price

Table 2: Location Analysis

Location	Mean Price	Median Price	Count	Avg Bedrooms	Avg Baths	Avg-Size (Marla)
Allama Iqbal Town	33,855,460.00	32,000,000.00	119	4.4	4.6	8.2
Bahria Town	33,217,170.00	30,000,000.00	2103	4.2	4.9	9.2
DHA Defence	78,722,840.00	65,000,000.00	6117	4.4	5.1	16.6
Gulberg	218,974,200.	97,500,000.00	89	5.1	4.9	39.0
Johar Town	35,867,460.00	29,000,000.00	524	4.6	5.0	8.5
Lake City	47,356,900.00	32,900,000.00	348	4.5	4.8	10.6
Park View City	19,120,080.00	18,000,000.00	239	4.0	4.3	5.6

Table 2 summarizes property statistics across various locations, highlighting differences in average and median prices, property counts, and average attributes. DHA Defence stands out with the highest mean price of around 78.72 million, while Gulberg has the highest median price of 97.5 million. Park View City features the lowest mean price of about 19.12 million. Each location also varies in average number of bedrooms, bathrooms, and property size, reflecting diverse market conditions.

Table 3: Model performance

Model	R-squared
Extra Trees Regressor	0.851
XGBoost Model	0.8465
Random Forest Model	0.842
Gradient Boosting Model	0.852
Decision Tree Model	0.8378
CatBoostRegressor Model	0.8439

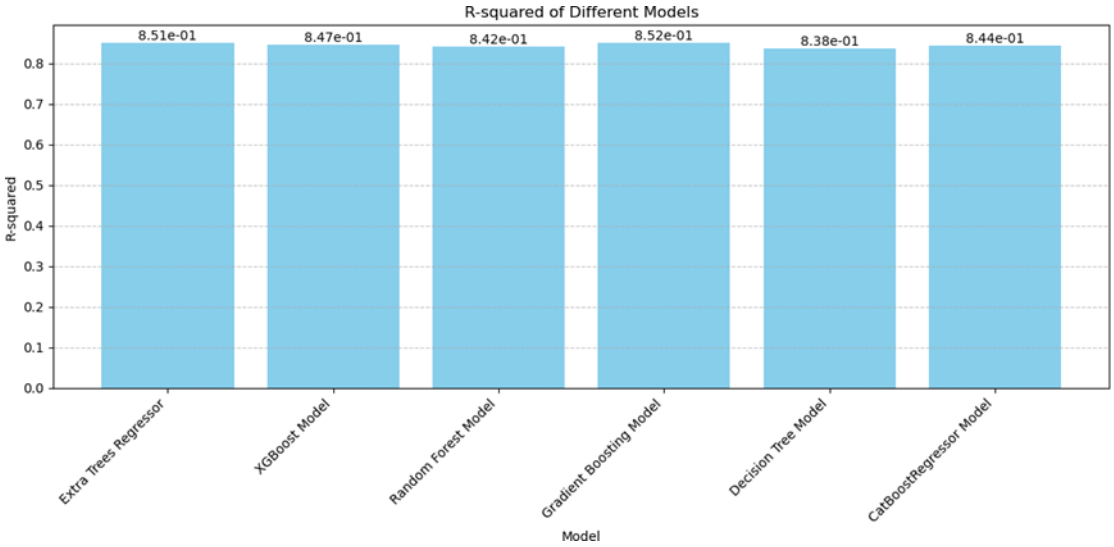

Figure 4: Metrics R-Square of different Models

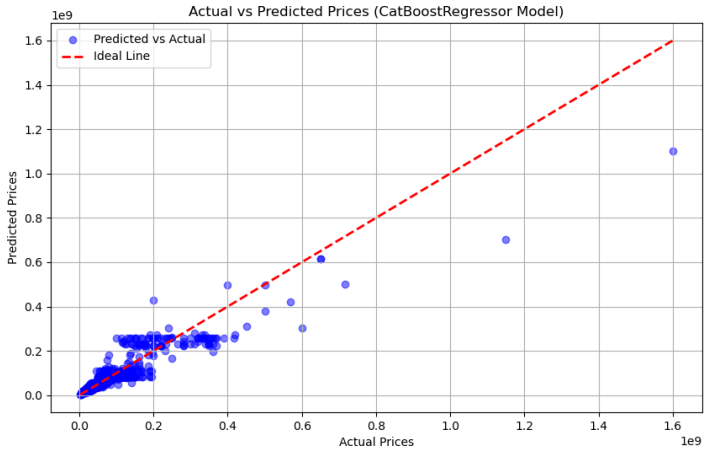

Figure 5: Actual vs Predicted Price (CatBoostRegressor Model)(A)

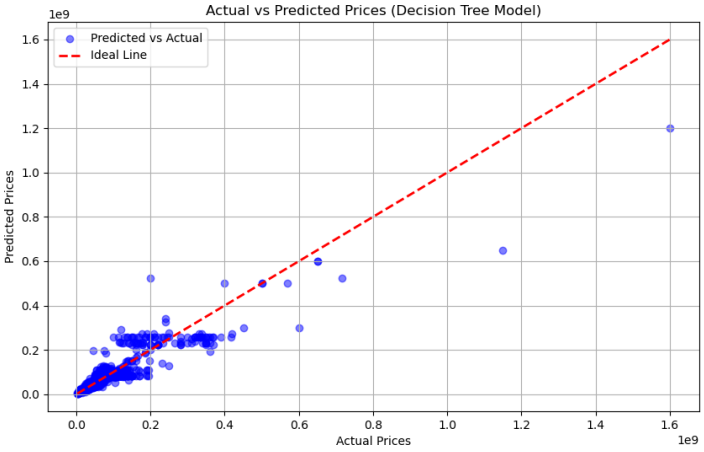

Figure 6: Actual vs Predicted Price(Decision Tree Model) (B)

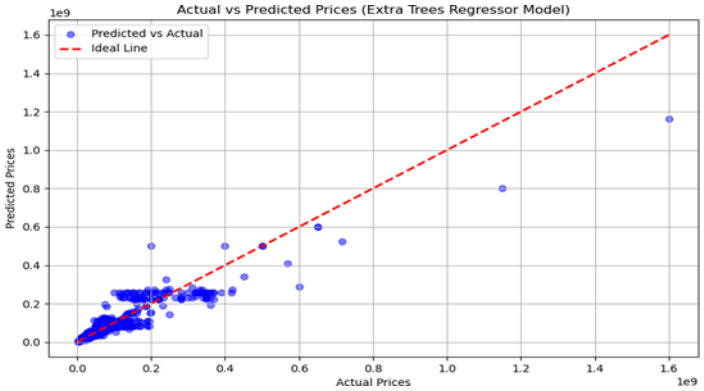

Figure 7: Actual vs Predicted Price (Extra Regressor Trees Model) (C)

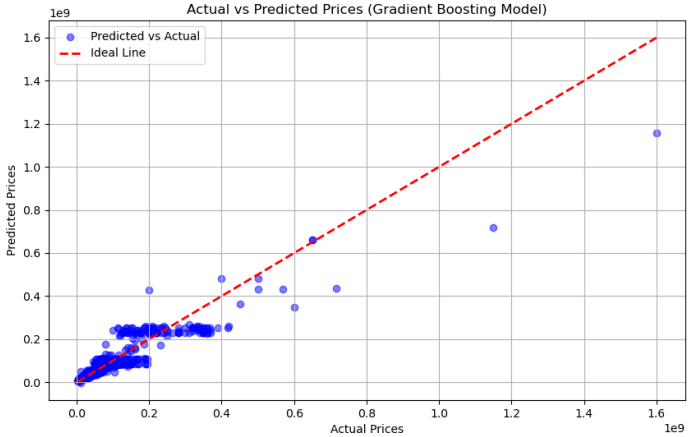

Figure 8: Actual vs Predicted Price (Gradient Boosting Model) (D)

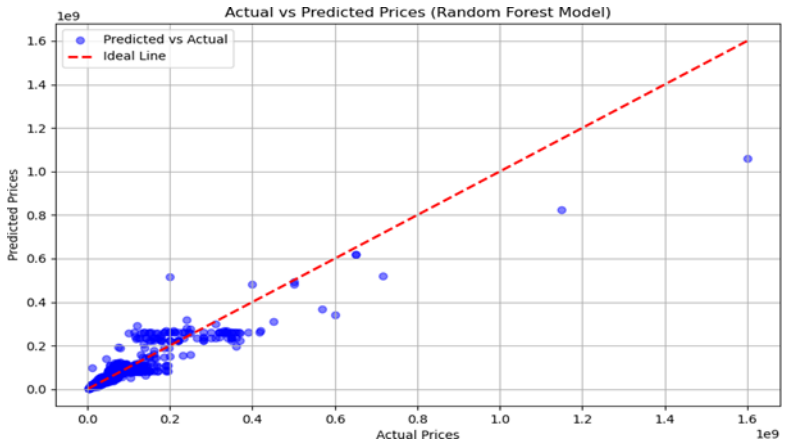

Figure 9: Actual vs Predicted Price (Random Forest Model) (E)

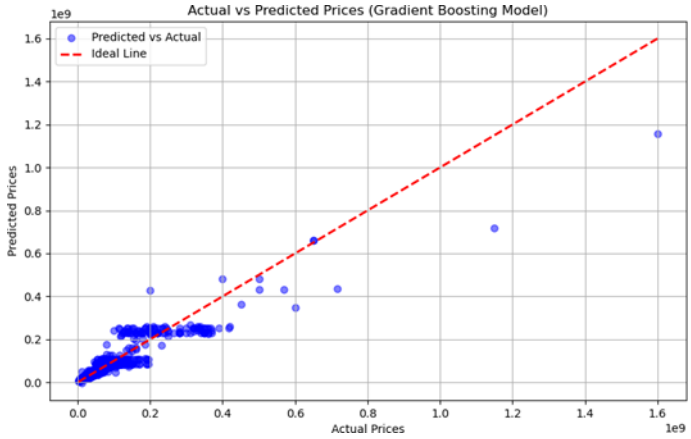

Figure 10: Actual vs Predicted Price (XGBoost Model) (F)

These scatter plots provide a clear visual assessment of the performance of different regression models in predicting real estate prices. Ensemble models like Random Forest, Extra Trees Regress or, Gradient Boosting, and CatBoost Regressor generally show better clustering around the ideal line, indicating higher prediction accuracy compared to a single Decision Tree model. However, all models exhibit some level of deviation, especially at higher price ranges, which suggests area s for potential model improvement. This comparative analysis aids in selecting the most effective model for predicting real estate prices based on the observed performance and accuracy. This summary compares actual real estate prices from June 2023 with predicted prices, adjusted f or an inflation rate of 24.76% as of July 2024. The table also includes the difference between the predicted prices (with inflation) and the actual prices.

Table 4: Results (Price Prediction for Gradient Boosting Model)

Actual Price (June 2023)	Predicted Price With Inflation Rate (24.76% July 2024)	Difference B/W (PWI-Actual Price)
19000000	40188564.01	21188564.01
45000000	121740544.9	76740544.9
60000000	105886352.4	45886352.35
19900000	30467831.32	10567831.32
98000000	105326162.4	7326162.4
68500000	105326162.4	36826162.4
71000000	105326162.4	34326162.4

67000000	105326162.4	38326162.4
83000000	105326162.4	22326162.4
140000000	202930251	62930250.98
18000000	19431348.99	1431348.987
185000000	313433670.3	128433670.3
81500000	105326162.4	23826162.4
73100000	105326162.4	32226162.4
38000000	46325922.84	8325922.837
40000000	58892260.97	18892260.97

The final GUI-based platform allows users to enter important information about a residential property, such as size, number of bedrooms, bathrooms, and location. The platform uses the Gradient Boosting Model to process these inputs and forecast the house price based on the existing data. The GUI also offers an option to alter the anticipated price for inflation, specifically a 24.76% inflation rate for the year 2024.

The GUI was designed to develop a user-friendly and visually appealing interface that is simple to navigate. Clients can easily add important details to get accurate cost estimates. To ensure the accuracy and authenticity of the information provided, the interface includes strong error handling and validation tools.

The picture below shows how the cost expectation stage is set up, which includes the option to change the increase. This device allows customers to see how flooding affects property values, leading to better and more accurate price estimates.

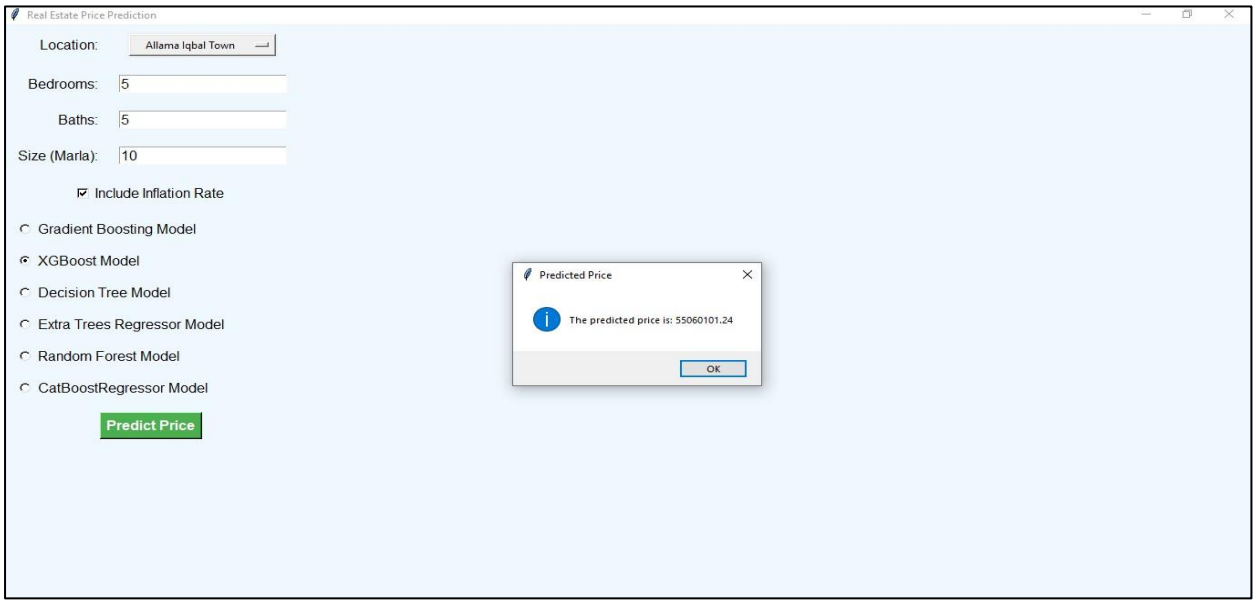

Figure 11: GUI for House Price Prediction Model Platform

Discussion

This study compared different machine learning models for Lahore, Pakistan, real estate price forecasting using data collected from Zameen.com. Six machine learning models were used in the analysis: CatBoostRegressor, XGBoost, Random Forest, Gradient Boosting, Extra Trees Regressor, and Decision Tree. The effectiveness and accuracy of these models in projecting house

prices serve as the basis for their evaluation. The results showed that in terms of accuracy and consistency, the CatBoostRegressor model outperformed the other models. This dataset was a good fit for CatBoostRegressor due to its handling of built-in features, such as encoding, and the ability to efficiently handle categorical data. Additionally, the model demonstrated resilience to overfitting, a prevalent problem in machine learning algorithms.The XGBoost and Gradient Boosting models also demonstrated high accuracy in predicting house prices. The excellent performance was likely aided by the fact that these models are known to handle complex interactions between features and structural data. XGBoost in particular is a formidable competitor in predictive modeling thanks to its well-known regularization and optimization capabilities.Extra Trees Regressor and Decision Tree models had significantly worse accuracy.

The extra- trees model, while similar to Random Forest, performed worse due to larger volatility and less robust handling of noisy data. The Decision Tree model, although easier to understand, was less accurate, mostly because it overfits, especially with smaller datasets. The exploratory data analysis and feature importance analysis found that variables such as the number of bedrooms, bathrooms, property size, and location all had a substantial impact on house pricing. The correlation matrix revealed substantial positive connections between price and several important variables, indicating their significance in the model. The geographical analysis revealed significant variations in property values across different parts of Lahore. DHA Defence and Bahria Town, two high-demand regions, had higher average home prices than Allama Iqbal Town and Gulberg, which were less popular. This

variance emphasizes the importance of location as a significant component in real estate pricing, along with earlier research that has highlighted the impact of geographical and socioeconomic characteristics on property values. This study provides useful information for predicting real estate prices in Lahore, but it has some drawbacks. The dataset is quite large, but it might not include all the factors that affect property prices, like how close it is to amenities, environmental concerns, and economic conditions. Also, the exam was limited to a specific area, which might not be relevant to other places with different market conditions. In the future, we plan to focus on increasing the dataset to include more different traits and areas. Extra information, like financial indicators and statistics, can help improve the model's accuracy. Also, looking at other advanced machine learning models, like deep learning methods, might provide new insights into predicting real estate prices. The study's findings can be useful for different people in the real estate market, such as buyers, sellers, investors, and managers.

Conclusion

Exact estimating tools can help buyers and sellers make informed decisions, guide investors to successful ventures, and help regulators understand industry trends and create fair laws. The study's findings show how machine learning methods can improve the accuracy of real estate price predictions. By using detailed calculations and large amounts of data, people can better understand what affects real estate prices and make smart choices in a complex and constantly changing market.

This detailed study has shown how accurately different machine learning methods predict home prices in the Lahore real estate market. With a high R^2 value of 0. 84, showing it makes accurate predictions, CatBoostRegressor was the best model among the six tested models: Extra Trees Regressor, XGBoost, Random Forest, Gradient Boosting, Decision Tree, and CatBoostRegressor. The Angle Boosting Regressor and XGBoost both showed good performance with R2 values of 0.

82 and 080 This means they are strong choices for predicting house prices. the study noted important factors like the number of rooms, location, and property value that consistently affect home prices in different models. This shows how important these parts are for the real market and suggests that gathering information in these areas can help improve accuracy.

A variety of real partners in the field, including investors, developers, and lawmakers, will be involved in the study's findings. Partners can make better estimates, get more accurate results, and make smarter choices by using new tools like cat-boost regressors. This could lead to a better way of using resources, more successful businesses, and ultimately a real estate market that is more active and lively. The report also suggests some topics for further study. You can improve model expectations by adding more financial data and market trends to the dataset. It would be helpful to test these models in different geological locations to ensure they can be used in various situations. These activities will help us understand what affects house prices and improve how useful machine learning models are in real estate markets around the world.

References

1. Albahli, S., Nazir, T. J. M. T., & Applications. (2024). Opinion mining for stock trend prediction using deep learning. 1-24.
2. Bin, J., Gardiner, B., Liu, H., Li, E., & Liu, Z. J. I. F. (2023). RHPMF: A context-aware matrix factorization approach for understanding regional real estate market. 94, 229-242.
3. Borch, C., Hee Min, B. J. B. D., & Society. (2022). Toward a sociology of machine learning explainability: Human–machine interaction in deep neural network-based automated trading. 9(2), 20539517221111361.
4. Chaleshtori, A. E. J. a. p. a. (2024). A novel decision fusion approach for sale price prediction using ElasticNet and MOPSO.
5. Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023). A novel study on machine learning algorithm-based cardiovascular disease prediction. Health & Social Care in the Community, 2023(1), 1406060.
6. Chen, L., Li, T., Chen, Y., Chen, X., Wozniak, M., Xiong, N., & Liang, W. J. C. S. (2024). Design and analysis of quantum machine learning: a survey. 36(1), 2312121.
7. Choudhary, C., Anurag, Shukla, P. J. A. i. A. S., & Imaging. (2024). A Robust Machine Learning Model for Forest Fire Detection Using Drone Images. 129-144.
8. Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., . . . Church, G. M. J. N.B. (2022). Single-sequence protein structure prediction using a language model and deep learning. 40(11), 1617-1623.
9. Qureshi, M., Khan, A., Daniyal, M., Tawiah, K., & Mehmood, Z. (2023). A comparative analysis of traditional SARIMA and machine learning models for CPI data modelling in Pakistan. Applied Computational Intelligence and Soft Computing, 2023(1), 3236617.
10. Qureshi, M., Iftikhar, H., Rodrigues, P. C., Rehman, M. Z., & Salar, S. A. (2024). Statistical modeling to improve time series forecasting using machine learning, time series, and hybrid models: a case study of bitcoin price forecasting. Mathematics, 12(23), 3666.
11. Dellnitz, A., Kleine, A., & Tavana, M. J. O. S. (2024). An integrated data envelopment analysis and regression tree method for new product price estimation. 1-23.
12. Dolphin, R., Smyth, B., & Dong, R. J. a. p. a. (2024). Contrastive Learning of Asset Embeddings from Financial Time Series.
13. Donghi, D., & Morvan, A. (2023). GeoVeX: Geospatial Vectors with Hexagonal Convolutional Autoencoders. Paper presented at the Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI forGeographic Knowledge Discovery.

14. Fazli, M., Alian, P., Owfi, A., & Loghmani, E. J. I. S. w. A. (2024). RPS: Portfolio asset selection using graph based representation learning. 22, 200348.
15. Qureshi, M., Khan, S., Bantan, R. A., Daniyal, M., Elgarhy, M., Marzo, R. R., & Lin, Y. (2022). Modeling and forecasting monkeypox cases using stochastic models. Journal of Clinical Medicine, 11(21), 6555.
16. Qureshi, M., Ahmad, N., Ullah, S., & ul Mustafa, A. R. (2023). Forecasting real exchange rate (REER) using artificial intelligence and time series models. Heliyon, 9(5).
17. Geiler, L., Affeldt, S., Nadif, M. J. I. J. o. D. S., & Analytics. (2022). A survey on machine learning methods for churn prediction. 14(3), 217-242.
18. Gortzak, A., & Ulusoy, N. C. (2024). Incorporating Interior Property Images for Predicting Housing Values.
19. Guan, M.-Y., Qiu, W.-R., Wang, Q.-K., & Xiao, X. J. C. B. (2024). Prediction of plant ubiquitylation proteins and sites by fusing multiple features. 19(5), 458-469.
20. Haddadin, M., Mohamed, O., Abu Elhaija, W., Matar, M. J. E., & Environment. (2023). Performance prediction of a clean coal power plant via machine learning and deep learning techniques. 0958305X231160590.
21. Hajdu, N. (2024). Advancing Organizational Analytics: A Strategic Roadmap for Implementing Machine Learning in Warehouse Management System.
22. Islam, M. D., Li, B., Lee, C., & Wang, X. J. T. i. G. (2022). Incorporating spatial information in machine learning: The Moran eigenvector spatial filter approach. 26(2), 902-922.
23. Iftikhar, H., Daniyal, M., Qureshi, M., Tawiah, K., Ansah, R. K., & Afriyie, J. K. (2023). A hybrid forecasting technique for infection and death from the mpox virus. Digital Health, 9, 20552076231204748.
24. Iftikhar, H., Qureshi, M., Zywiołek, J., López-Gonzales, J. L., & Albalawi, O. (2024). Short-term PM 2.5 forecasting using a unique ensemble technique for proactive environmental management initiatives. Frontiers in Environmental Science, 12, 1442644.
25. Ahmad, S., Qureshi, M., Iftikhar, H., Rodrigues, P. C., & Rehman, M. Z. (2025). An improved family of unbiased ratio estimators for a population distribution function. AIMS Mathematics, 10(1), 1061-1084.
26. Tawiah, K., Daniyal, M., & Qureshi, M. (2023). Pakistan CO2 emission modelling and forecasting: a linear and nonlinear time series approach. Journal of Environmental and Public Health, 2023(1), 5903362.
27. Jáuregui-Velarde, R., Andrade-Arenas, L., Celis, D. H., Dávila-Morán, R. C., & Cabanillas-Carbonell, M. J. I.
28. Application to Open Statistics Knowledge Graphs for Estimating House Prices.
29. Khanmohammadi, R., Saba-Sadiya, S., Esfandiarpour, S., Alhanai, T., & Ghassemi, M. M. J. S. C. S. (2024).
30. MambaNet: A Hybrid Neural Network for Predicting the NBA Playoffs. 5(5), 628.
31. Lee, H., Jeong, H., Lee, B., Lee, K. D., & Choo, J. (2023). St-rap: A spatio-temporal framework for real estate appraisal. Paper presented at the Proceedings of the 32nd ACM International Conference on Information and Knowledge Management.
32. Li, D., Liu, M., Yang, L., Wei, H., Guo, J. J. C. A. C., & Engineering, I. (2024). A non-contact identification
33. Choudhary, C., Anurag, Shukla, P. J. A. i. A. S., & Imaging. (2024). A Robust Machine Learning Model for Forest Fire Detection Using Drone Images. 129-144.
34. Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., . . . Church, G. M. J. N.B. (2022). Single-sequence protein structure prediction using a language model and deep learning. 40(11), 1617-1623.

35. Dellnitz, A., Kleine, A., & Tavana, M. J. O. S. (2024). An integrated data envelopment analysis and regression tree method for new product price estimation. 1-23.
36. Dolphin, R., Smyth, B., & Dong, R. J. a. p. a. (2024). Contrastive Learning of Asset Embeddings from Financial Time Series.
37. Donghi, D., & Morvan, A. (2023). GeoVeX: Geospatial Vectors with Hexagonal Convolutional Autoencoders. Paper presented at the Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI forGeographic Knowledge Discovery.
38. Fazli, M., Alian, P., Owfi, A., & Loghmani, E. J. I. S. w. A. (2024). RPS: Portfolio asset selection using graph based representation learning. 22, 200348.
39. Geiler, L., Affeldt, S., Nadif, M. J. I. J. o. D. S., & Analytics. (2022). A survey on machine learning methods for churn prediction. 14(3), 217-242.
40. Gortzak, A., & Ulusoy, N. C. (2024). Incorporating Interior Property Images for Predicting Housing Values.
41. Guan, M.-Y., Qiu, W.-R., Wang, Q.-K., & Xiao, X. J. C. B. (2024). Prediction of plant ubiquitylation proteins and sites by fusing multiple features. 19(5), 458-469.
42. Haddadin, M., Mohamed, O., Abu Elhaija, W., Matar, M. J. E., & Environment. (2023). Performance prediction of a clean coal power plant via machine learning and deep learning techniques. 0958305X231160590.
43. Hajdu, N. (2024). Advancing Organizational Analytics: A Strategic Roadmap for Implementing Machine Learning in Warehouse Management System.
44. Islam, M. D., Li, B., Lee, C., & Wang, X. J. T. i. G. (2022). Incorporating spatial information in machine learning: The Moran eigenvector spatial filter approach. 26(2), 902-922.
45. Jáuregui-Velarde, R., Andrade-Arenas, L., Celis, D. H., Dávila-Morán, R. C., & Cabanillas-Carbonell, M. J. I.Khanmohammadi, R., Saba-Sadiya, S., Esfandiarpour, S., Alhanai, T., & Ghassemi, M. M. J. S. C. S. (2024).
46. MambaNet: A Hybrid Neural Network for Predicting the NBA Playoffs. 5(5), 628.
47. Law, S., Shen, Y., & Zhong, C. (2024). Progress on machine learning applications in geography in A Research Agenda for Spatial Analysis (pp. 127-146): Edward Elgar Publishing.