ISSN E: 3006-1466
ISSN P: 3006-1458

CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW

Vol.03 No.03 (2025)

TRANSFORMING LAHORE'S REAL ESTATE MARKET WITH MACHINE LEARNING-DRIVEN PRICE PREDICTIONS

BILAL AHMAD
Department of Statistics,
University of Okara, Punjab, Pakistan.
Email: bilaldataset97@gmail.com
DR. RASHID MAQBOOL
Incharge Main Library,
University of Okara, Punjab, Pakistan.
Email: rashid_gee@yahoo.com
DR. ZAHID IQBAL
Assistant Registrar,
University of Okara, Punjab, Pakistan.
Email: zahidiqballak@gmail.com

ABSTRACT

This study aims to provide reliable information about the dynamics of the Lahore City, Pakistan real estate market and its response to economic changes, especially inflation. The study seeks to enhance forecasting methods for real estate prices in Lahore, Pakistan, to create a more transparent and efficient market. The research findings will guide decision-making processes and policy decisions aimed at promoting steady, long-term growth in Lahore's real estate market. The selling price of homes in Lahore, Pakistan was forecasted using various parametric regression models including the Extra Trees Repressor Model, XG-Boost Model, Random Forest Model, Gradient Boosting Model, Decision Tree Model, and Cat-Boost Regressor models. Data sourced from Zameen.com as of June 26, 2023, was utilized for this purpose. The dataset included 9539 homes located in prominent districts such as DHA Defense, Bahria Town, Johar Town, Park View City, Lake City, and Allama Iqbal Town. The average inflation rate in Pakistan's real estate market in 2024 was 24.76%. As graded by the R-square value and MSE, the Gradient Boosting (85%), and Extra Trees Regressor (85%) regression model emerged on top regarding the two metrics. In this dataset, it suffices to say that the Gradient Boosting and Extra Trees Regressor models are the way to go in predicting house prices.

Keywords: - *Price prediction, Real estate, Machine learning, Models, Regression.*

1. INTRODUCTION: -

In the past, property value was determined in the real estate market using methods like comparing it to similar properties, looking at the investment return, and estimating future earnings (Møller, Pedersen, 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.

This text talks about comparing six common machine learning methods: Extra Trees Regressor, XGBoost, Random Forest, Gradient Boosting, Decision Tree, and Cat-Boost-Regressor. 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 and others (2022) talked about how appealing natural things can be (Chowdhury and others, 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 rest of paper is organized as follows. Section 2 discusses relevant works on the subject. Section 3 describes the process used to collect and clean data, as well as the models used. Section 4 gives an exploratory data analysis that focuses on the dataset's features and entries. Section 5 analyzes the findings and compares the performance of the models. Section 6 is devoted to discussion, and Section 7 closes the study.

2. LITERATURE REVIEW: -

The hedonic estimating demonstration (HPM) is the foremost frequently utilized instrument for property assessment (Geiler, Affeldt, Nadif, & Analytics, 2022) HPM, was afterward adjusted to the lodging 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 been routinely utilized to relate domestic costs and traits (Haddadin, Mohamed, Abu Elhaija, Matar, & Environment, 2023).Moreover, 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). However, 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).

Similarly, based on information AI calculations are presently expanding in acknowledgment over numerous industries (Lu et al., 2020). A few investigate have utilized ML frameworks to expect domestic prices (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). Additionally, 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). Likewise, using 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). 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.

Thus, ,(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). However, 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). On the other hand, several researches (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).

3. METHODOLOGY: -

Finally, the last section concludes the paper. New ways to 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:

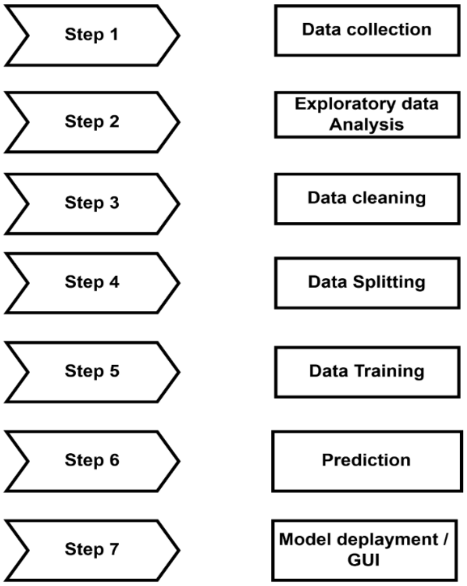


Figure 1: Study Flow Chart

3.1.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 may gain valuable insights into the Lahore real estate market by using data from Zameen.com. 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 Defense	6117
Bahria Town	2103
Park View City	239
Johar Town	524
Lake City	348
Gulberg	89
Allama Iqbal Town	119
TOTAL-7	**9539**

3.2.Exploratory Data Analysis (EDA): -

Having cleaned and preprocessed our data, we can now delve into exploratory data analysis (EDA). This crucial step allows us to gain insights into the data's distribution, identify patterns and relationships between variables, and assess its suitability for building our real estate price prediction model (Zhang et al., 2023). Here's an outline of the key EDA techniques we will employ:

- Descriptive Statistics (Figure 2)
- Detect outliers and anomalies (Figure 3)
- Correlation Analysis (Figure 4)
- Location Analysis (Figure 5)
- Price Analysis (Figure 6)

3.3.Data cleaning: -

The EDA deleted invalid values and outliers from the dataset. The raw data includes several records that were presumably inaccurate (Nirala, Singh, & Purani).

3.4.Data Splitting: -

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).

3.5.Model Training: -

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).

3.6. *Prediction: -*

The models' predictions have been evaluated to see which one achieved the highest accuracy (Vyas, 2024).

3.7.Model Deployment/GUI: -

One of the steps in getting the information prepared for machine learning was to concentrate on a particular cost run. 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 their 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).

3.8.Models Description: -

Following are the details descriptions of the model that are used in this study.

3.8.1. 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). enhanced ensemble of every model used before. The ElasticNet model, on the other hand, had only a few tweaks, XGBregressor offered over 10 elements to optimize (Qin & Technology, 2024).

3.8.2. Extra Tree Regressor: -

This 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).

3.8.3. 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).

3.8.4. 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).

3.8.5. 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).

3.8.6. Decision Tree Model: -

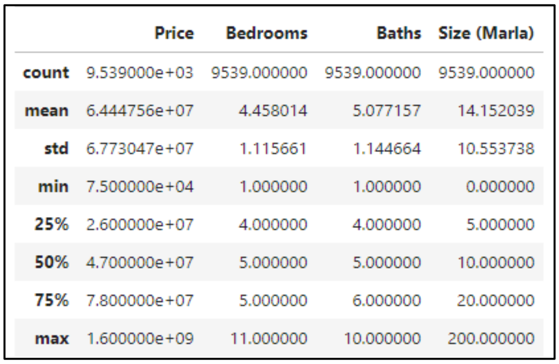
The decision tree regressor uses the features of the data to create a tree-like model that predicts future results. The choice tree regressor learns from the deepest and shallowest parts of a chart, based on the system. Look at the information closely. Lattice Look CV is a way to handle different settings 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).

4. Results an Analysis: -

4.1. Exploratory data analysis(EDA)

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.

In Figure 3: Finding unusual cases in the data has provided important information about the features of the properties. In the Cost column, there are 533 unusual entries, suggesting that many properties have prices that are very different from the usual market range. This shows that very tall or very cheap properties nearby might affect the overall assessment. The dataset shows 456 unusual cases where properties have either a very high or low number of rooms compared to usual. Changing room numbers can impact the study of property value trends. The Showers column shows 4 exceptions, which means that most properties usually have a regular number of bathrooms, with only small differences. Finally, for Measure (Marla), there are 55 exceptions. This means that some properties are either much bigger or much smaller than the average size. These discoveries show the differences in property characteristics, and we need to deal with these differences to ensure a more accurate investigation.

Therefore, to remove unusual results from the data, use different methods to filter them out. The Interquartile Range (IQR) method is a popular way to find outliers by looking at how far they are from the average range of data. Exceptions are data points that are outside the usual range set by the lower and upper limits based on the first and third quartiles. After identifying these high values, you will remove them from the dataset, keeping only the data points that fall within the acceptable range. This method makes sure that the data is more representative of normal values and reduces the negative impact of outliers

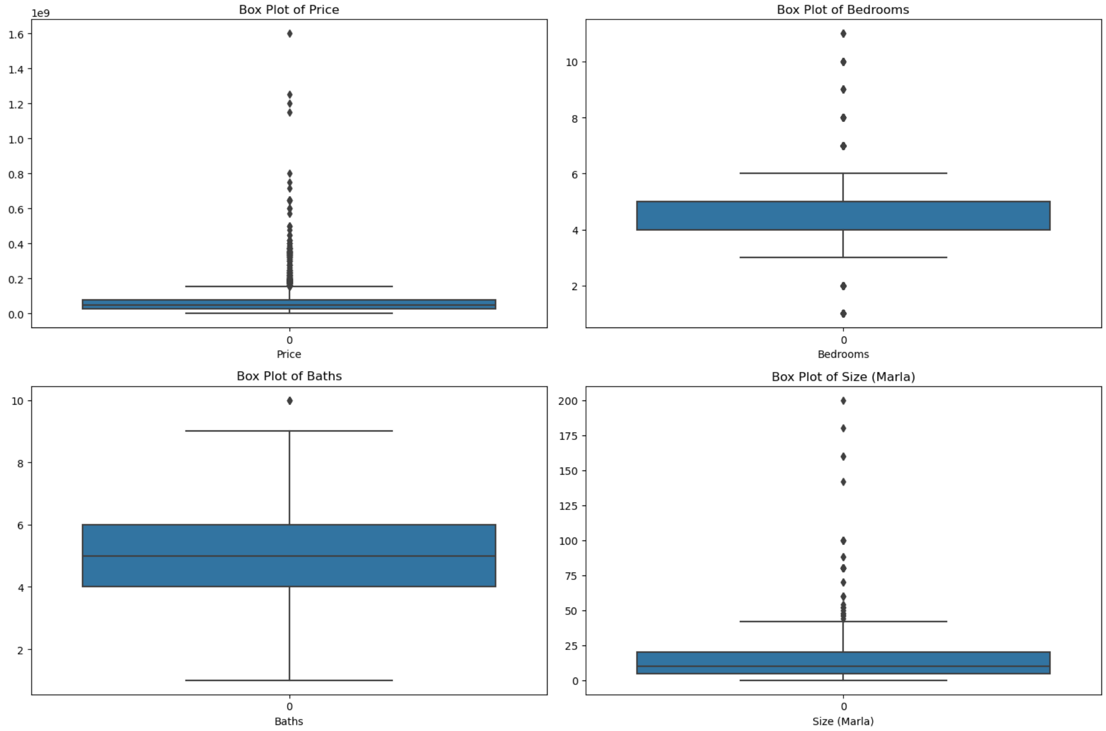


Figure 3: Box Plot (Detection Of Outliers)

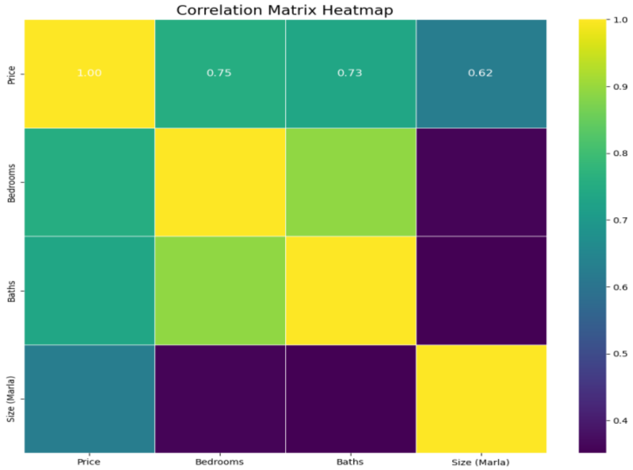


Figure 4: correlation matrix heatmap

The correlation matrix heatmap shows how different numbers in the dataset are connected. Each box in the heatmap shows the relationship between two numbers. For this situation, the link between cost and rooms is 0. 752 This shows a strong positive connection, meaning that as the number of rooms goes up, the cost usually increases too. In simple terms, Cost and Showers have a connection of 0. 735, suggesting that having more bathrooms is linked to higher costs. The relationship between Cost and Measure (Marla) is 0. 624, showing a strong positive connection. On the other hand, the connection between Rooms and Measure (Marla) is 0. 357, which shows a weaker positive relationship. The connection between Baths and Measure (Marla) is weak, with a score of 0. 351 The heatmap uses a range of colors to show these relationships. Darker colors mean weaker or negative relationships, while brighter colors indicate stronger positive relationships. The explanations inside the cells provide specific numbers, making it easy to see how exactly the different factors are connected and in what way.

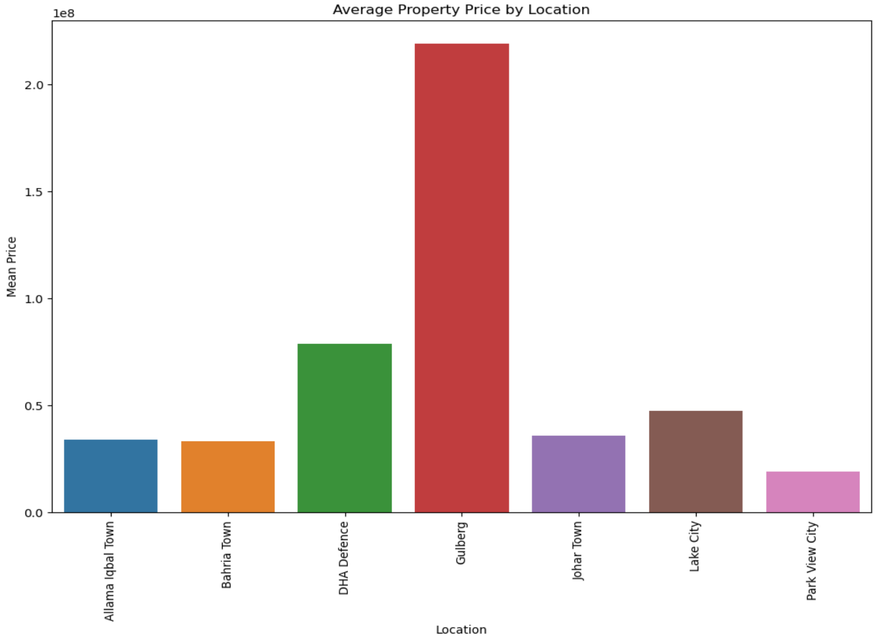


Figure 5: Average Property Price by Location

Table. 2: Location analysis

Location	Mean Price	Median Price	Coun	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

The table 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.

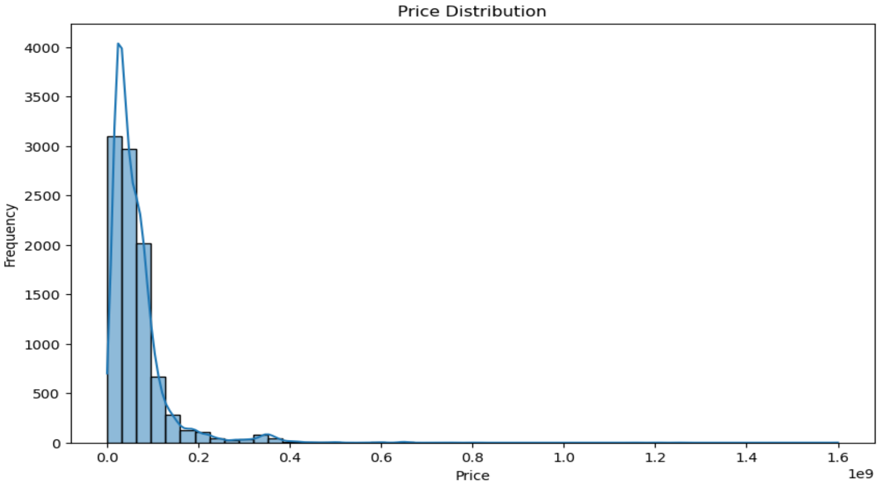


Figure 6: Price Distribution (A)

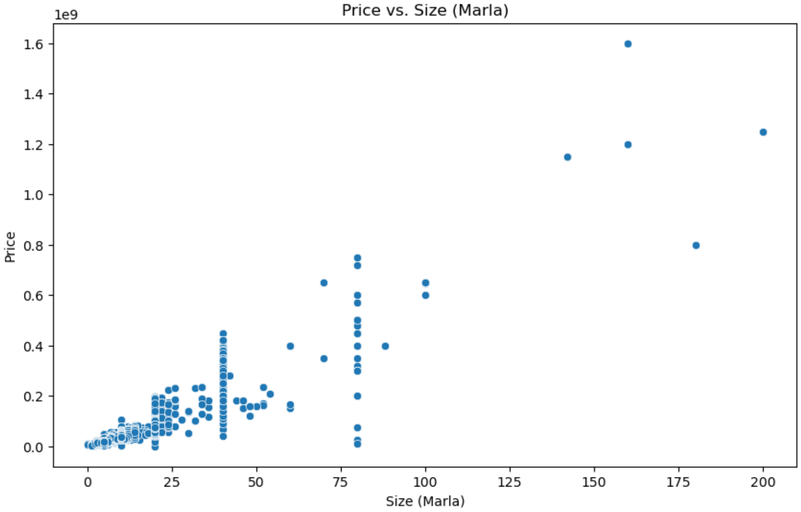


Figure 7: Price vs.Size(Marla) (B)

4.1.1. Price Distribution: -

Price distribution ndicating that it represents the distribution of property prices in the dataset. X-axis (Price): The x-axis represents the range of property prices. The prices are likely in some monetary unit, potentially in millions or billions, as indicated by the values shown on the axis (e.g., 0.0, 0.2, 0.4, etc., up to approximately 1.6e9). Y-axis (Frequency): The y-axis represents the frequency of properties within each price range. The frequency counts the number of properties that fall within each bin of the histogram. Distribution Shape: The histogram shows a right-skewed distribution, indicating that most property prices are concentrated towards the lower end, with fewer properties at higher price ranges. A high frequency of properties at lower price levels gradually decreases as the price increases. This plot helps visualize the spread and concentration of property prices, suggesting that most properties are priced lower, with a long tail towards the higher price ranges, which is typical for real estate price distributions.

4.1.2. The scatter plot: -

The scatter plot "Price vs. Size (Marla)" explains how property size is measured. and price The x-axis shows the property size in Marla, and the y-axis shows the property value. expense the graph shows a good connection, indicating that larger properties usually have higher values. Prices can vary a bit, especially for larger sizes.

4.1.3. Data cleaning: -

We fixed the missing values in the data by replacing them with the median. In Particularly, For the Bedrooms column with 355 missing entries and the Baths column with 414 missing entries, the missing information was filled in using the average middle values. This method makes sure that filling in missing values shows the average for each variable, keeping the dataset accurate and not adding much bias. The Price and Size (Marla) columns didn't have any missing information, so we didn't need to do any extra cleaning for these parts. This method helps get the data ready for more analysis and modeling, making sure it is complete and consistent.

4.1.4. Data Splitting: -

In the 'Data Splitting' part of the study, the dataset was divided into two parts: one for training and one for testing. This was done to evaluate how well the prediction models work. Using the train_test_split function from sklearn. model_selection, the dataset was divided into features (X) and target values (y). X includes all columns except for the Cost, which is saved in y. The information was then divided into two parts: 80% was used for training and 20% was used for testing. This part makes sure that the show is based on a large amount of information and tested on a smaller, separate group, making it reliable in its predictions. The random_state setting was changed to 42 to make sure we get the same results every time.

Table. 4: Model performance by Metrics

Model	Mean Squared Error	Mean Absolute Error	R-squared
Extra Trees Regressor	927,064,209,459,840.00	14,317,025.63	0.851
XGBoost Model	955,063,108,819,127.00	14,283,271.80	0.8465
Random Forest Model	983,486,924,430,755.00	14,469,853.60	0.842
Gradient Boosting Model	922,698,555,799,466.00	14,439,494.03	0.852
Decision Tree Model	1,009,576,165,920,510.00	14,567,447.20	0.8378
CatBoostRegressor Model	971,412,583,652,685.00	14,407,970.05	0.8439

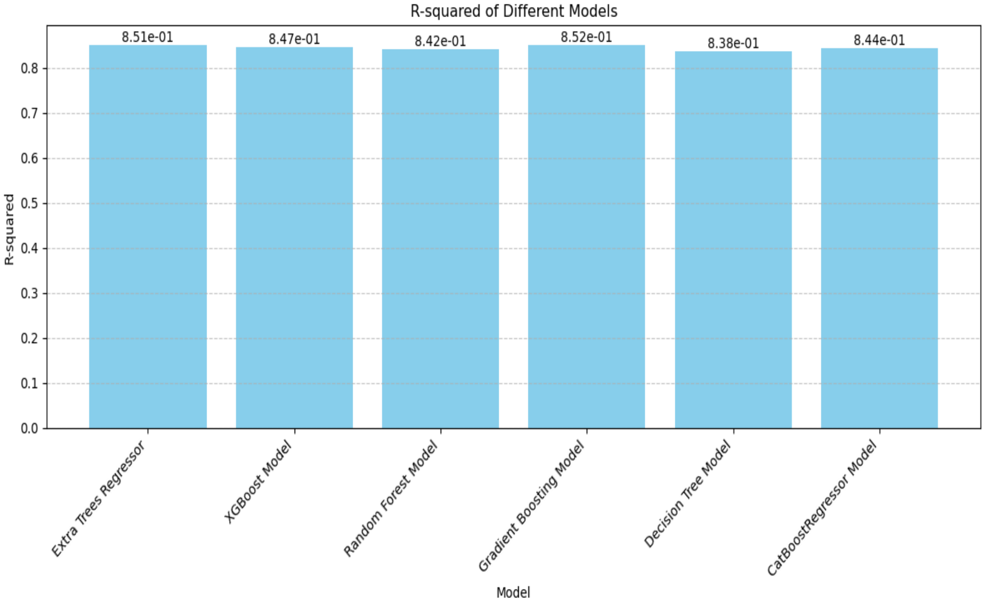


Figure 8: Metrics R-Square of different Models

Based on the evaluation metrics of several machine learning models, the Gradient Boosting Model has demonstrated the best performance in predicting real estate prices. It achieved the lowest Mean Squared Error (MSE) of 922,698,555,799,466.00 and a competitive Mean Absolute Error (MAE) of 14,439,494.03, with the highest R-squared value of 0.852 among all the models. This indicates that the Gradient Boosting Model provides the most accurate predictions with a high degree of correlation between the predicted and actual values, making it the most reliable choice for this real estate price prediction task.

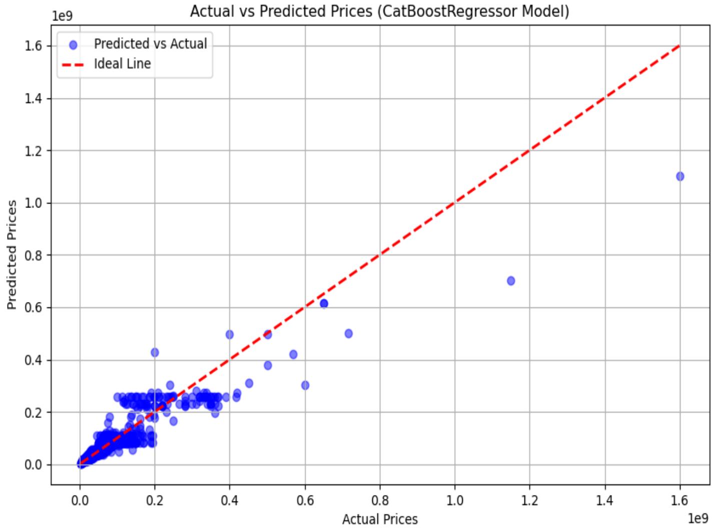


Figure 9: Actual vs Predicted Price(CatBoostRegressor Model) (A)

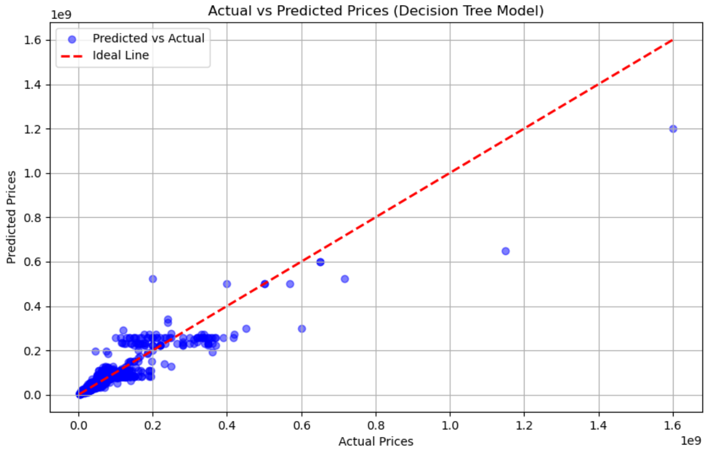


Figure 10: Actual vs Predicted Price (Decision Tree Model) (B)

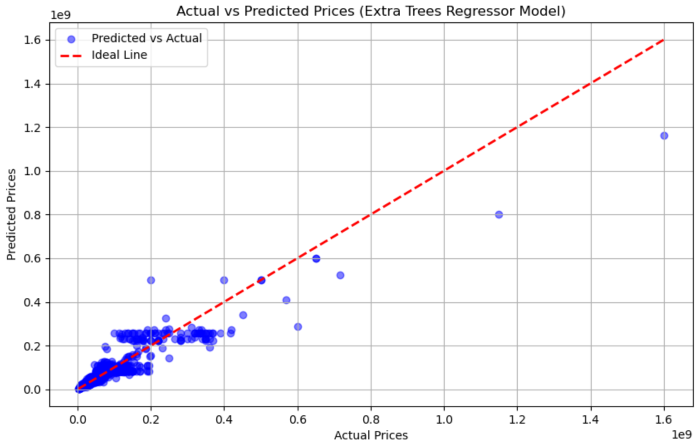


Figure 11: Actual vs Predicted Price (Extra Regressor Trees Model) (C)

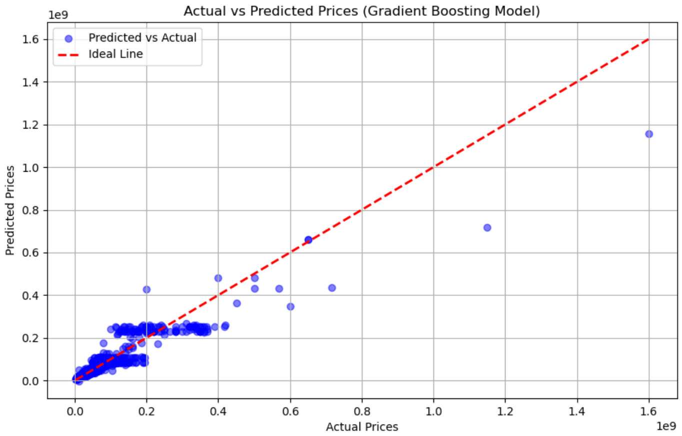


Figure 12: Actual vs Predicted Price (Gradient Boosting Model) (D)

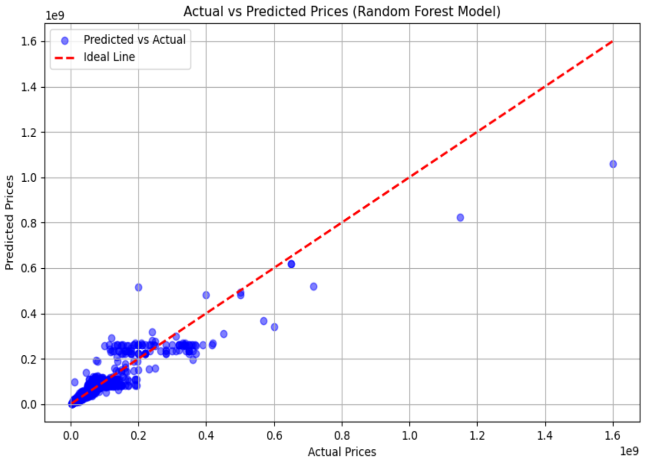


Figure 13: Actual vs Predicted Price (Random Forest Model) (E)

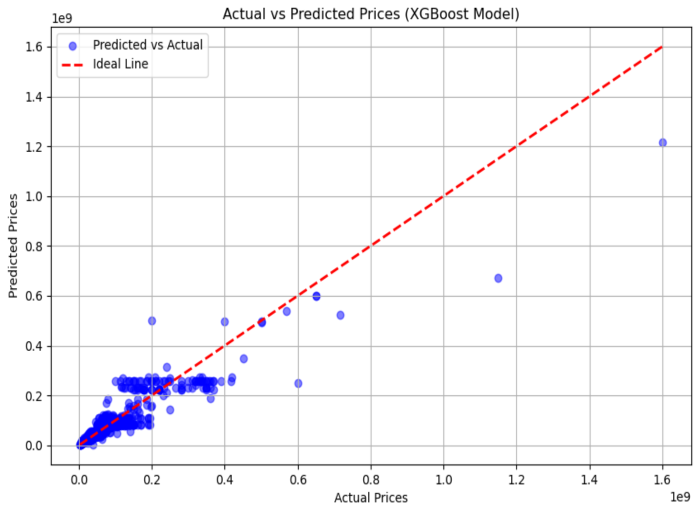


Figure 14: Actual vs Predicted Price (XGBoost Model) (F)

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 Regressor, 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 areas 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.

4.2. Price Prediction: -

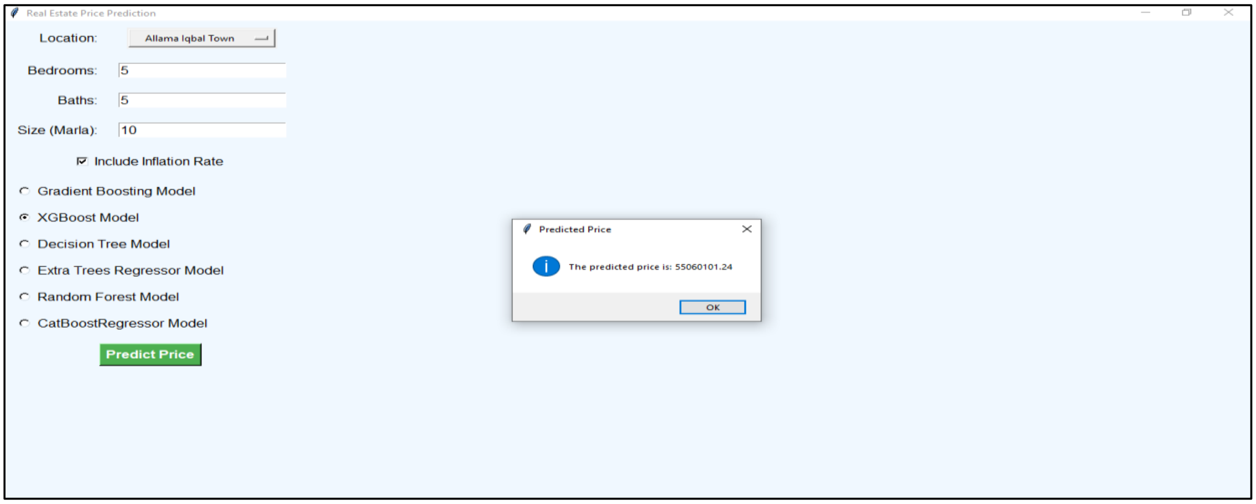
This summary compares actual real estate prices from June 2023 with predicted prices, adjusted for an inflation rate of 24.76% as of July 2024. The table also includes the difference between the predicted prices (with inflation) and the actual prices.

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.



Real Estate Price Prediction

Location: Allama Iqbal Town

Bedrooms: 5

Baths: 5

Size (Marla): 10

☑ Include Inflation Rate

○ Gradient Boosting Model
● XGBoost Model
○ Decision Tree Model
○ Extra Trees Regressor Model
○ Random Forest Model
○ CatBoostRegressor Model

Predicted Price

The predicted price is: 55060101.24

Figure 15: GUI for House Price Prediction Model Platform

5. Discussion and Conclusion: -

This detailed study has shown how accurately different machine learning methods predict home prices in the Lahore real estate market. With a high R² 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. Also, 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.

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.

- **Model Performance:** 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.
- ***Feature Importance and Insights****:* 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.

5.1. Limitations and Future Work:

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.

5.2. Practical Implications:

The study's findings can be useful for different people in the real estate market, such as buyers, sellers, investors, and managers. 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.

REFERENCES: -

Albahli, S., Nazir, T. J. M. T., & Applications. (2024). Opinion mining for stock trend prediction using deep learning. 1-24.

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.

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.

Chaleshtori, A. E. J. a. p. a. (2024). A novel decision fusion approach for sale price prediction using ElasticNet and MOPSO.

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.

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.

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.

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.

Dolphin, R., Smyth, B., & Dong, R. J. a. p. a. (2024). Contrastive Learning of Asset Embeddings from Financial Time Series.

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.

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.

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.

Gortzak, A., & Ulusoy, N. C. (2024). Incorporating Interior Property Images for Predicting Housing Values.

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.

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.

Hajdu, N. (2024). Advancing Organizational Analytics: A Strategic Roadmap for Implementing Machine Learning in Warehouse Management System.

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.

Jáuregui-Velarde, R., Andrade-Arenas, L., Celis, D. H., Dávila-Morán, R. C., & Cabanillas-Carbonell, M. J. I.

J. o. I. M. T. (2023). Web Application with Machine Learning for House Price Prediction. *17*(23). Karamanou, A., Brimos, P., Kalampokis, E., & Tarabanis, K. (2024). Explainable Graph Neural Networks: An

Application to Open Statistics Knowledge Graphs for Estimating House Prices.

Khanmohammadi, R., Saba-Sadiya, S., Esfandiarpour, S., Alhanai, T., & Ghassemi, M. M. J. S. C. S. (2024).

MambaNet: A Hybrid Neural Network for Predicting the NBA Playoffs. *5*(5), 628.

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.

Lee, H., Jeong, H., Lee, B., Lee, K. D., & Choo, J. (2023). *St-rap: A spatio-temporal framework for real estateappraisal.* Paper presented at the Proceedings of the 32nd ACM International Conference on Information and Knowledge Management.

Li, D., Liu, M., Yang, L., Wei, H., Guo, J. J. C. A. C., & Engineering, I. (2024). A non-contact identification

method of overweight vehicles based on computer vision and deep learning.

Liu, J.-C., Chen, C.-T., Lee, C., Huang, S.-H. J. A. T. o. I. S., & Technology. (2024). Evolving Knowledge Graph Representation Learning with Multiple Attention Strategies for Citation Recommendation System.*15*(2), 1-26.

Liu, R., Liu, H., Huang, H., Song, B., & Wu, Q. J. P. R. (2024). Multimodal multiscale dynamic graph convolution networks for stock price prediction. *149*, 110211.

Malakouti, S. M., Ghiasi, A. R., Ghavifekr, A. A., & Emami, P. J. W. E. (2022). Predicting wind power generation using machine learning and CNN-LSTM approaches. *46*(6), 1853-1869.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. J. N. (2023). Scaling deeplearning for materials discovery. *624*(7990), 80-85.

Mhlongo, N. Z., Falaiye, T., Daraojimba, A. I., Olubusola, O., Ajayi-Nifise, A. O. J. W. J. o. A. R., & Reviews. (2024). Artificial intelligence in stock broking: A systematic review of strategies and outcomes. *21*(2), 1950-1957.

Nirala, K. K., Singh, N. K., & Purani, V. S. A Counter-Propagation Based Neuro Solution Model for Categorization and Fee Fixation of Engineering Institutions. In *Robotics and Automation in Industry 4.0* (pp. 380-395): CRC Press.

NLP, P. H. P. U. *STORYTELLING REAL ESTATE.* tilburg university,

Pinconschi, E., Gopinath, D., Abreu, R., & Pasareanu, C. S. J. a. p. a. (2024). Evaluating Deep NeuralNetworks in Deployment (A Comparative and Replicability Study).

Qin, H. J. I. J. o. C. T., & Technology. (2024). Revolutionizing Cryptocurrency Operations: The Role ofDomain-Specific Large Language Models (LLMs). *72*(6), 101-113.

Stavropoulos, C. (2024). Deep Learning, Streaming Data: A Study. Voshol, M. Temporal Causal Discovery with Machine Learning.

Vyas, T. K. J. a. p. a. (2024). Deep Learning with Tabular Data: A Self-supervised Approach.

Wu, H., Zhou, H., Long, M., & Wang, J. J. N. M. I. (2023). Interpretable weather forecasting for worldwidestations with a unified deep model. *5*(6), 602-611.

Xiao, C., Zhou, J., Xiao, Y., Huang, J., & Xiong, H. (2024). ReFound: Crafting a Foundation Model for UrbanRegion Understanding upon Language and Visual Foundations.

Xie, L., Ouyang, Y., Chen, L., Wu, Z., Li, Q. J. I. T. o. V., & Graphics, C. (2023). Towards Better Modeling WithMissing Data: A Contrastive Learning-Based Visual Analytics Perspective.

Zeng, W.-F., Zhou, X.-X., Willems, S., Ammar, C., Wahle, M., Bludau, I., . . . Mann, M. J. N. C. (2022). AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *13*(1), 7238.

Zhang, Q., Huang, C., Xia, L., Wang, Z., Li, Z., & Yiu, S. (2023). *Automated spatio-temporal graph contrastivelearning.* Paper presented at the Proceedings of the ACM Web Conference 2023.

Zheng, X., Liu, Y., Bao, Z., Fang, M., Hu, X., Liew, A. W.-C., & Pan, S. J. a. p. a. (2023). Towards data-centricgraph machine learning: Review and outlook.

Abidoye, R. B., Chan, A. P., Abidoye, F. A., Oshodi, O. S. J. I. j. o. h. m., & analysis. (2019). Predicting propertyprice index using artificial intelligence techniques: Evidence from Hong Kong. *12*(6), 1072-1092.

Abidoye, R. B., Chan, A. P. J. I. J. o. H. M., & Analysis. (2018). Achieving property valuation accuracy in developing countries: the implication of data source. *11*(3), 573-585.

Abidoye, R. B., & Chan, A. P. J. P. M. (2017). Artificial neural network in property valuation: application framework and research trend. *35*(5), 554-571.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. J. H. (2018). State-of-the-art in artificial neural network applications: A survey. *4*(11).

Baldini, G., & Phelan, K. D. J. J. o. E. (2019). The melanocortin pathway and control of appetite-progress and therapeutic implications. *241*(1), R1-R33.

Beghi, E., Giussani, G., Nichols, E., Abd-Allah, F., Abdela, J., Abdelalim, A., . . . Alahdab, F. J. T. L. N. (2019). Global, regional, and national burden of epilepsy, 1990–2016: a systematic analysis for the GlobalBurden of Disease Study 2016. *18*(4), 357-375.

Breiman, L. J. M. l. (2001). Random forests. *45*, 5-32.

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., . . . Lee, W.-H. J. N. a. r. (2018).
miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions.
46(D1), D296-D302.

Cummings, J., Lee, G., Mortsdorf, T., Ritter, A., Zhong, K. J. A. s., Research, D. T., & Interventions, C. (2017).

Alzheimer's disease drug development pipeline: 2017. *3*(3), 367-384.

Espey, M., Lopez, H. J. G., & change. (2000). The impact of airport noise and proximity on residential property values. *31*(3), 408-419.

Friedman, J. H. J. A. o. s. (2001). Greedy function approximation: a gradient boosting machine. 1189-1232.

Gao, S., Tang, G., Hua, D., Xiong, R., Han, J., Jiang, S., . . . Huang, C. J. J. o. M. C. B. (2019). Stimuli-responsivebio-based polymeric systems and their applications. *7*(5), 709-729.

García-Magariño, I., Fox-Fuller, J. T., Palacios-Navarro, G., Baena, A., & Quiroz, Y. T. J. R. d. n. (2020). Visualworking memory for semantically related objects in healthy adults. *71*(8), 277.

Guresen, E., Kayakutlu, G., & Daim, T. U. J. E. s. w. A. (2011). Using artificial neural network models in stockmarket index prediction. *38*(8), 10389-10397.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112):Springer.

Jiang, Y., & Jiang, Z.-P. (2017). *Robust adaptive dynamic programming*: John Wiley & Sons.

Keskin, B. J. I. J. o. S. P. M. (2008). Hedonic analysis of price in the Istanbul housing market. *12*(2), 125-138.

Lancaster, K. J. J. J. o. p. e. (1966). A new approach to consumer theory. *74*(2), 132-157. LeCun, Y., Bengio, Y., & Hinton, G. J. n. (2015). Deep learning. *521*(7553), 436-444.

Lenk, M. M., Worzala, E. M., Silva, A. J. J. o. P. V., & Investment. (1997). High-tech valuation: should
artificial neural networks bypass the human valuer? *, 15*(1), 8-26.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., . . . Zhu, N. J. T. l. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *395*(10224), 565-574.

Luttik, J. J. L., & planning, u. (2000). The value of trees, water and open space as reflected by house pricesin the Netherlands. *48*(3-4), 161-167.

Mrsic, D., Smajlovic, J., Loncar, D., Avdic, S., Avdagic, M., Smajic, E., . . . Jahic, A. J. M. S.-m. (2020). Risk factors in patients with non-ST segment elevation myocardial infarction. *32*(3), 224.

Park, S.-J., Ahn, J.-M., Kim, Y.-H., Park, D.-W., Yun, S.-C., Lee, J.-Y., . . . Park, S.-W. J. N. E. J. o. M. (2015).
Trial of everolimus-eluting stents or bypass surgery for coronary disease. *372*(13), 1204-1212.

Rafiei, M. H., Adeli, H. J. J. o. C. E., & Management. (2016). A novel machine learning model for estimationof sale prices of real estate units. *142*(2), 04015066.

Rahadi, R. A., Wiryono, S. K., Koesrindartoto, D. P., Syamwil, I. B. J. I. J. o. H. M., & Analysis. (2015). Factorsinfluencing the price of housing in Indonesia. *8*(2), 169-188.

Rampini, L., Re Cecconi, F. J. J. o. P. I., & Finance. (2022). Artificial intelligence algorithms to predict Italianreal estate market prices. *40*(6), 588-611.

Rosen, S. J. J. o. p. E. (1976). A theory of life earnings. *84*(4, Part 2), S45-S67.

Tajani, F., Morano, P., Ntalianis, K. J. J. o. P. I., & Finance. (2018). Automated valuation models for real estate portfolios: A method for the value updates of the property assets. *36*(4), 324-347.

Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., . . . Tan, Y. J. N. a. r. (2020). Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *48*(D1), D1031-D1041.

Wu, Y., Ho, W., Huang, Y., Jin, D.-Y., Li, S., Liu, S.-L., . . . Wang, Q. J. T. L. (2020). SARS-CoV-2 is an appropriatename for the new coronavirus. *395*(10228), 949-950.

Yin, A.-h., Peng, C.-f., Zhao, X., Caughey, B. A., Yang, J.-x., Liu, J., . . . Liu, H.-l. J. P. o. t. N. A. o. S. (2015).
Noninvasive detection of fetal subchromosomal abnormalities by semiconductor sequencing of maternal plasma DNA. *112*(47), 14670-14675.

Zurada, J. M., Ward, K. M., & Grossman, M. E. J. J. o. t. A. A. o. D. (2006). Henoch-Schönlein purpura associated with malignancy in adults. *55*(5), S65-S70.