



Real Estate House Price Prediction Using Extreme Gradient Boosting

²Sandeep Kumar, ³Rahul Yadav, ⁴Jwala Yadav, ¹Silviya D'Monte,

¹Assistant Professor, ²⁻⁴UG Student,

¹Department Of Computer Engineering,

¹Universal College of Engineering, Mumbai, India

Abstract : House price prediction is a critical task in the real estate industry, influencing various stakeholders such as buyers, sellers, and investors. Traditional regression techniques and machine learning algorithms have been extensively employed for this purpose. In recent years, Extreme Gradient Boosting (XGBoost) has gained popularity due to its high performance and robustness in handling complex datasets. This research paper presents a comprehensive study on the application of XGBoost for house price prediction. We explore the effectiveness of XGBoost in capturing nonlinear relationships and handling large-scale datasets, thus enhancing the accuracy of price predictions. Our experimental results demonstrate the superior predictive performance of XGBoost compared to traditional regression models and other machine learning algorithms. Additionally, we discuss important factors such as feature selection, hyperparameter tuning, and model evaluation techniques to optimize the XGBoost model for house price prediction. The findings of this study contribute to advancing the state-of-the-art in real estate analytics and provide valuable insights for stakeholders involved in property valuation and investment decision-making.

Keywords- House Price, Real Estate, Extreme Gradient Boosting, hyperparameter tuning, model evaluation.

I. INTRODUCTION

The real estate market is a dynamic and multifaceted sector that plays a crucial role in the global economy. House price prediction, a fundamental aspect of real estate analysis, holds significant importance for various stakeholders, including buyers, sellers, investors, and policymakers. Accurate predictions of house prices are essential for making informed decisions regarding property transactions, investment strategies, and economic policy formulation. However, predicting house prices accurately poses several challenges due to the complex and nonlinear nature of real estate data. One such technique that has gained prominence in the field of machine learning is Extreme Gradient Boosting (XGBoost). XGBoost is an ensemble learning method that combines the predictions of multiple weak learners, typically decision trees, to build a robust and accurate predictive model. Its popularity stems from its ability to handle large-scale datasets efficiently, capture complex interactions among features, and produce highly accurate predictions. These characteristics make XGBoost particularly well-suited for regression tasks, including house price prediction. The contributions of this research paper are multi-faceted. Firstly, we provide empirical evidence on the performance of XGBoost in the context of house price prediction, comparing it with traditional regression models and other machine learning algorithms. Secondly, we identify key factors that influence predictive accuracy and analyze the relative importance of different features in predicting house prices. Thirdly, we explore various techniques for optimizing the XGBoost model, including feature selection, hyperparameter tuning, and model evaluation strategies. Lastly, we offer insights into the practical implications of using XGBoost for house price prediction and its potential applications in real-world scenarios. By advancing the understanding of XGBoost in real estate analytics, this research paper aims to provide valuable guidance for real estate professionals, researchers, and policymakers involved in property valuation, investment analysis, and market forecasting. Furthermore, the findings and insights derived from this study can contribute to the development of more accurate and reliable models for house price prediction, thereby enhancing decision-making processes and promoting efficiency in the real estate market. The main objective of this project is to predict the house price using the Extreme Gradient Boosting algorithm. It is used to predict the price using the city, area, type, luxury category, location, BHK etc., The XGBoost is measured for accuracy and performance against several algorithms.

II. LITERATURE SURVEY

Anand G. Rawool, Dattatray V. Rogye [1] is presented in the paper "House Price Prediction Using Machine Learning" discusses the use of machine learning techniques to predict house prices. The authors highlight the importance of machine learning in various fields, such as speech command recognition, product recommendation, and healthcare, and recognize its potential in improving customer services and enhancing automobile safety. The focus of the paper is on applying machine learning to address the challenges in the real estate market, where housing prices are highly competitive and volatile. The primary challenge addressed in the paper is the difficulty of predicting house prices accurately, considering the ever-changing market trends. The existing systems often fail to provide necessary predictions about future market trends, leading to potential losses. The authors propose a machine learning-based solution to predict house prices with greater accuracy.

Vibha B.Bhor, Mohini S. Gaikwad, Prachi S. Zende [2] presented the paper discussing the use of machine learning for housing price prediction. It outlines three key stages: initial data exploration, data cleaning, and transformation; data modeling; and data analysis using four models: Linear Regression, Random Forest, Gradient Boosting Regressor, and XGBoost Regressor. The study emphasizes the importance of accurately predicting housing prices to aid homebuyers and real estate agents. The authors recommend that linear regression increases prediction accuracy and mitigates investment risks. Overall, the paper offers insights into employing machine learning techniques for real estate price predictions.

Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, and Gbenle Oluwadara [3] is presented the paper presented by, This research, presented at the 8th International Conference on Information Technology and Quantitative Management, explores the use of a Random Forest machine learning technique for house price prediction. The study aims to address the challenge of predicting house prices, which are influenced by various factors such as location, physical condition, and concepts. The authors argue that conventional methods like the House Price Index (HPI) are not sufficient for accurate predictions in the 21st century. The research employs the UCI Machine Learning Repository Boston housing dataset, containing 506 entries and 14 features. Data exploration and preprocessing are carried out to prepare the dataset for model development. The Random Forest algorithm is used to build a predictive model. Key performance evaluation metrics include Mean Absolute Error (MAE), R^2 (Coefficient of Determination), and Root Mean Square Error (RMSE).

Winky K.O. Ho, Bo-Sin Tang, and Siu Wai Wong [4] presented the paper exploring the application of three machine learning algorithms (Support Vector Machine - SVM, Random Forest - RF, and Gradient Boosting Machine - GBM) for property price appraisal. The study analyzes a dataset of over 40,000 housing transactions spanning 18 years in Hong Kong and compares the predictive performance of these algorithms. Results indicate that RF and GBM outperform SVM in terms of predictive power, as evidenced by lower Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). However, the study acknowledges that SVM remains useful for producing reasonably accurate predictions in situations with tight time constraints. The research suggests that machine learning can offer a promising alternative for property valuation and price prediction, particularly in the context of property research.

Chen Chee Kin Zailan Arabee Bin Abdul Salam, Kadhar Batcha Nowshath [5] presented the paper, titled "Machine Learning based House Price Prediction Model," introduces a model that leverages machine learning methods to forecast house prices. While the precise details are not provided, the authors likely an approach encompassing the methodology, data sources, and machine learning algorithms employed to create this predictive model. This research likely contributes to the growing field of real estate and property price prediction using advanced data-driven techniques, and potential of machine learning in this domain.

Sheng Bin [6] presented the research paper focuses on the prediction and analysis of the real estate market in China, which is closely tied to economic development and societal stability. The authors emphasize the importance of forecasting real estate prices to facilitate macroeconomic control and aid real estate investors in forming strategies. The study employs a multiple linear regression model to analyze various factors, including policy, economic, and housing supply, affecting real estate prices. It also introduces three key innovations: the use of entropy and information gain to identify primary influencing factors, the construction of a real estate price trend model, and the development of a multiple linear regression model for predicting prices. The paper highlights the significance of empirical research on real estate price determinants, filling a gap in the literature, and contributes to data-driven decision-making in the real estate market.

Summary

In this above paper they mainly focus on focuses on predicting and analyzing the real estate market in China, using a multiple linear regression model to identify primary influencing factors and develop a trend model for price prediction,

contributing to data-driven decision-making in the real estate market focuses on predicting and analyzing the real estate market in China, using a multiple linear regression model to identify primary influencing factors and develop a trend model for price prediction, contributing to data-driven decision-making in the real estate market. outline stages and models for data exploration, cleaning, and modeling, suggesting that linear regression improves prediction accuracy for housing prices. But they don't have the feature of analytics with the Power BI and with analytics diagram.

III. PROPOSED SYSTEM

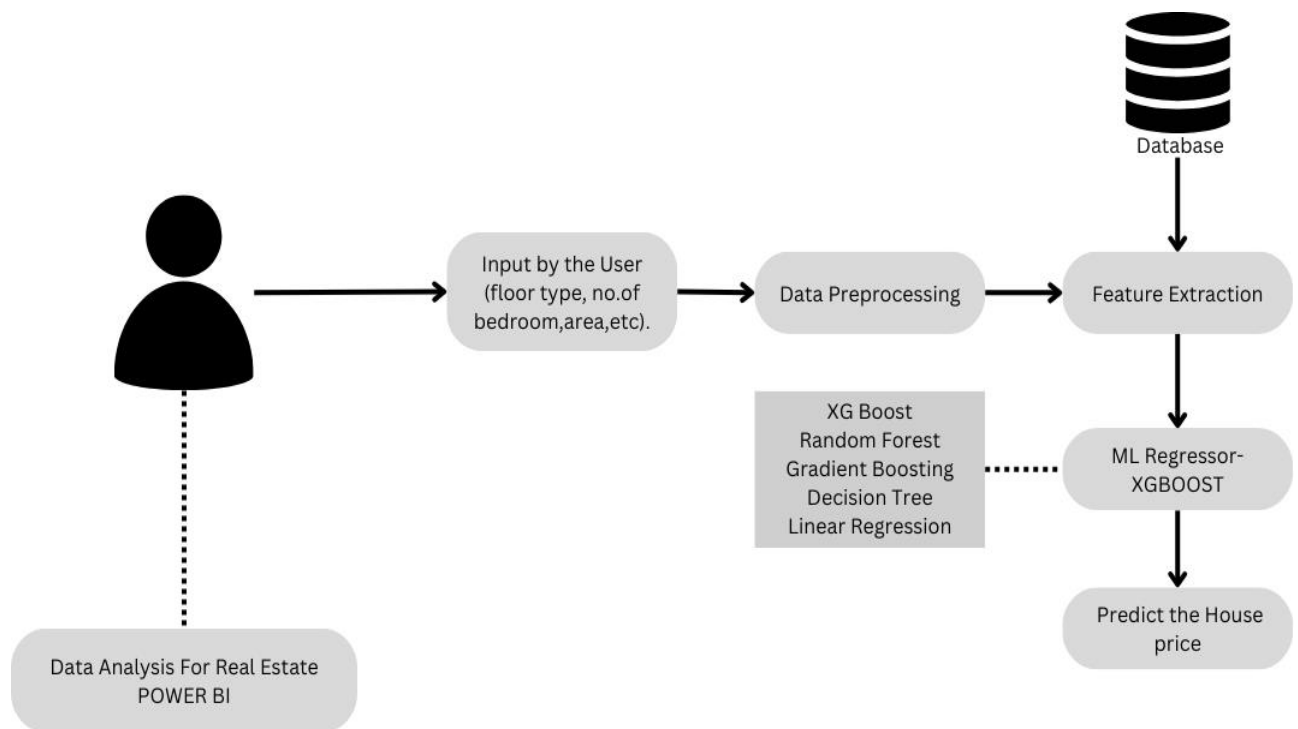


Fig 2.1. System Architecture

The user can access the price prediction module from the main page. user has to search first the city for which he/she want to predict the house price. after entering this city user has to provide the are(sqft), no of bedroom he/she want, how many balcony he want, whether the user want the luxury room or in budget. after providing this information user has to click on predict button, and the price will be predicted for the house. here after providing the input our system will preprocess it, it will apply the feature extraction technique and then it will predict the price using the xgboost model which is trained. Here we have used several machine learning algorithm like Linear Regression, Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting, Xgboost and here best performance is of xgboost so we have used xgboost for deployment.

The data analysis process begins with data preparation, where we gather and clean property data from multiple sources, including real estate listings, property databases, and market research reports. We then use Power BI to connect to the cleaned data and create interactive dashboards and reports.

One of the key features of Power BI is its ability to create dynamic visualizations that allow users to explore and interact with the data in real-time. We leverage this capability to create visualizations such as maps, charts, and graphs to represent property-related metrics such as property prices, location trends, property types.

IV. METHODOLOGY

Extreme Gradient Boosting : XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed environments (Hadoop, SGE, MPI) and can solve problems beyond billions of examples. Gradient Boosting Framework: XGBoost is based on the gradient boosting framework, which is an ensemble learning technique where a series of weak learners (usually decision trees) are combined to create a strong learner. In gradient boosting, each new model is trained to correct the errors made by the previous models. Tree Boosting Algorithm: XGBoost builds trees sequentially, with each new tree attempting

to correct the errors made by the previous ones. It adds trees one at a time, and each tree is trained on the residuals (the differences between the predicted and actual values) of the previous ensemble. Gradient Calculation: XGBoost computes the gradient of the loss function with respect to the predicted values of the model. This gradient provides information on how to update the model to minimize the loss function. XGBoost efficiently calculates gradients using second-order derivatives to improve convergence speed.

Algorithm

Initialization:

STEP 1: Given training information from the instance space $S=(x_1,y_1)\dots(m,y_m)$ where $x_i \in X$ and $y_i \in Y$

$Y=-1,+1$

STEP 2: Start the distribution off. $D_1(i)=\frac{1}{n}$.

Using the algorithm for $t=1,\dots,T$, do

Utilizing the distribution D_t , train a weak learner $h_t: X \rightarrow R$.

Z_t is a normalization factor chosen to create the distribution of D_{t+1} , and its formula is

$D_{t+1}(i) = D_t(i) e^{-\eta y_i h_t(x_i)} Z_t$.

$f(x) = \sum_{t=1}^T h_t(x)$ and $H(x) = \text{sign}(f(x))$ are the final scores.

General Parameter

eta [default=0.3, alias: learning_rate] :

Step size shrinkage used in update to prevent overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative.

range: [0,1]:

gamma [default=0, alias: min_split_loss]

Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be.

max_depth [default=6]:

Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. 0 indicates no limit on depth. Beware that XGBoost aggressively consumes memory when training a deep tree. exact tree method requires a non-zero value.

min_child_weight [default=1]:

Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning. In a linear regression task, this simply corresponds to the minimum number of instances needed to be in each node. The larger min_child_weight is, the more conservative the algorithm will be.

max_delta_step [default=0]:

Maximum delta step we allow each leaf output to be. If the value is set to 0, it means there is no constraint. If it is set to a positive value, it can help make the update step more conservative. Usually this parameter is not needed, but it might help in logistic regression when class is extremely imbalanced. Set it to a value of 1-10 might help control the update.

subsample [default=1]:

Subsample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees. and this will prevent overfitting. Subsampling will occur once in every boosting iteration.

lambda [default=1, alias: reg_lambda] :

L2 regularization term on weights. Increasing this value will make the model more conservative.

V. MODEL BUILDING

When we are building any machine learning model two datasets are required : one for training and one for testing.now we have only one so let's divide it into the training and testing part:here we have used 80:20 ratio i.e 80% data for

training and 20% data for testing .now after splitting the dataset our dataset is partitioned into the $x_{train}, y_{train}, x_{test}, y_{test}$. after splitting the dataset we have trained the model using the Extreme Gradient Boosting algorithm i.e we have call the fit method to train x_{train}, y_{train} dataset.Xgboost is one of the most potent technique used in machine learning for solving both regression and classification problem as our problem is belonging to the regression class so we have used Extreme Gradient Regressor.

Rank	Model	Function	R2	MAE(Mean Absolute Error)
1	XGBoost	from xgboost import XGBRegressor	0.904798	0.447518
2	Extra Trees	from sklearn.ensemble import ExtraTreesRegressor	0.901308	0.451269
3	Random Forest	from sklearn.ensemble import RandomForestRegressor	0.901166	0.453624
4	Gradient Boosting	from sklearn.ensemble import GradientBoostingRegressor	0.889234	0.509266
5	Decision Tree	from sklearn.tree import DecisionTreeRegressor	0.826898	0.549944
6	MLP	from sklearn.neural_network import MLPRegressor	0.586050	0.619448
7	AdaBoost	from sklearn.ensemble import AdaBoostRegressor	0.816854	0.698479
8	Linear Regression	from sklearn.linear_model import LinearRegression	0.829522	0.713011
9	Ridge	from sklearn.linear_model import Ridge	0.829536	0.713523
10	SVR	from sklearn.svm import SVR	0.782917	0.818851
11	Lasso	from sklearn.linear_model import Lasso	0.059434	1.528906

Evaluation and Measure: In order to assess the performance of our predictive model for house price estimation in the Real Estate domain, we employed two commonly used evaluation metrics: the R2 score and Mean Absolute Error (MAE).

R2 Score:

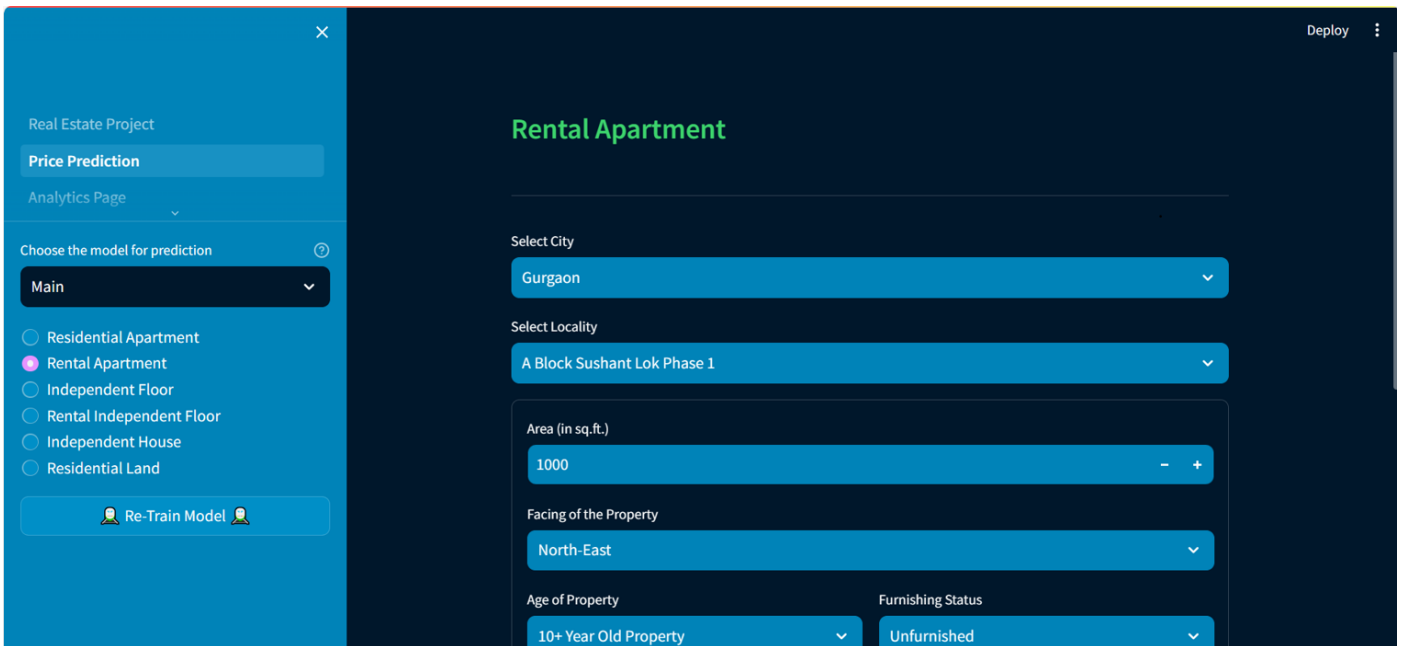
The R2 score, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable (house prices) that is predictable from the independent variables (property attributes, location, etc.) in our model. A higher R2 score indicates a better fit of the model to the data, with values ranging from 0 to 1. A score of 1 indicates a perfect fit, while a score of 0 indicates that the model does not explain any of the variance in the target variable better than a horizontal line.

Mean Absolute Error (MAE):

The Mean Absolute Error (MAE) provides a measure of the average magnitude of errors between the actual and predicted house prices. It is calculated as the average of the absolute differences between the predicted and actual prices for each property in the dataset. The MAE is particularly useful for evaluating the accuracy of our model in terms of the absolute deviation of predictions from the true values.

In this experiment analysis the accuracy of Extreme Gradient Boosting,extra trees,random forest is 90 % i.e almost the same but the mean absolute error of xgboost is less as compared to others that's why we have deployed the model using Xgboost.

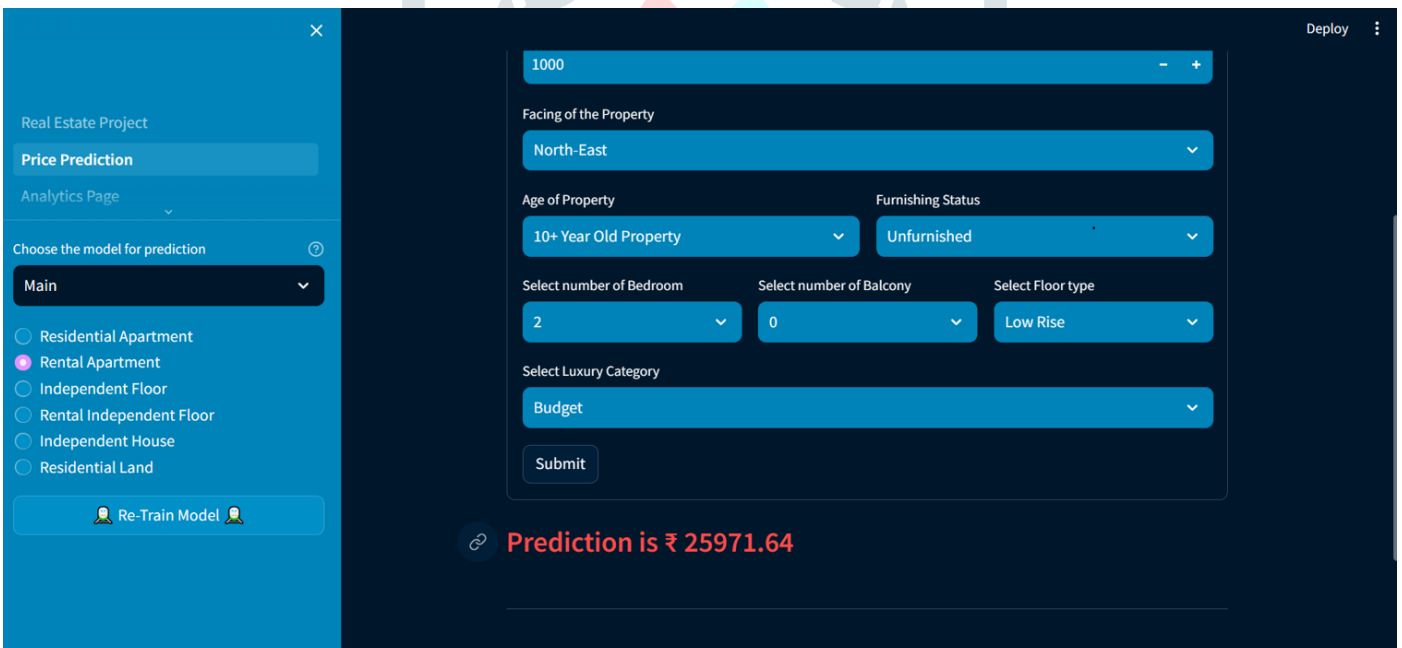
VI. RESULT AND DISCUSSION



The screenshot shows a web application interface for predicting rental apartment prices. On the left, there is a sidebar with a 'Real Estate Project' section containing 'Price Prediction' and 'Analytics Page'. Below this is a 'Choose the model for prediction' dropdown set to 'Main'. A list of property types includes 'Residential Apartment', 'Rental Apartment' (selected), 'Independent Floor', 'Rental Independent Floor', 'Independent House', and 'Residential Land'. A 'Re-Train Model' button is at the bottom of the sidebar. The main area is titled 'Rental Apartment' and contains several input fields: 'Select City' (Gurgaon), 'Select Locality' (A Block Sushant Lok Phase 1), 'Area (in sq.ft.)' (1000), 'Facing of the Property' (North-East), 'Age of Property' (10+ Year Old Property), and 'Furnishing Status' (Unfurnished). A 'Deploy' button is in the top right corner.

Fig 3.1. Input by user in order to predict the price

In figure 3.1 we are specifying the criteria by entering the City, Locality, Area Facing, Age of property and furnished.



This screenshot shows the same interface as Figure 3.1, but with additional input fields and the final prediction result. The 'Area' field is now 1000. Below the 'Facing of the Property' field, there are three more dropdowns: 'Age of Property' (10+ Year Old Property), 'Furnishing Status' (Unfurnished), 'Select number of Bedroom' (2), 'Select number of Balcony' (0), and 'Select Floor type' (Low Rise). A 'Select Luxury Category' dropdown is set to 'Budget'. A 'Submit' button is located below these fields. At the bottom of the main area, a red text box displays the prediction: 'Prediction is ₹ 25971.64'. The 'Deploy' button remains in the top right corner.

Fig 3.2. Prediction of Rental Apartments

In the figure 3.2 we are specifying the criteria by entering the Number of bedroom, Balcony, Floor type, Luxury Category.

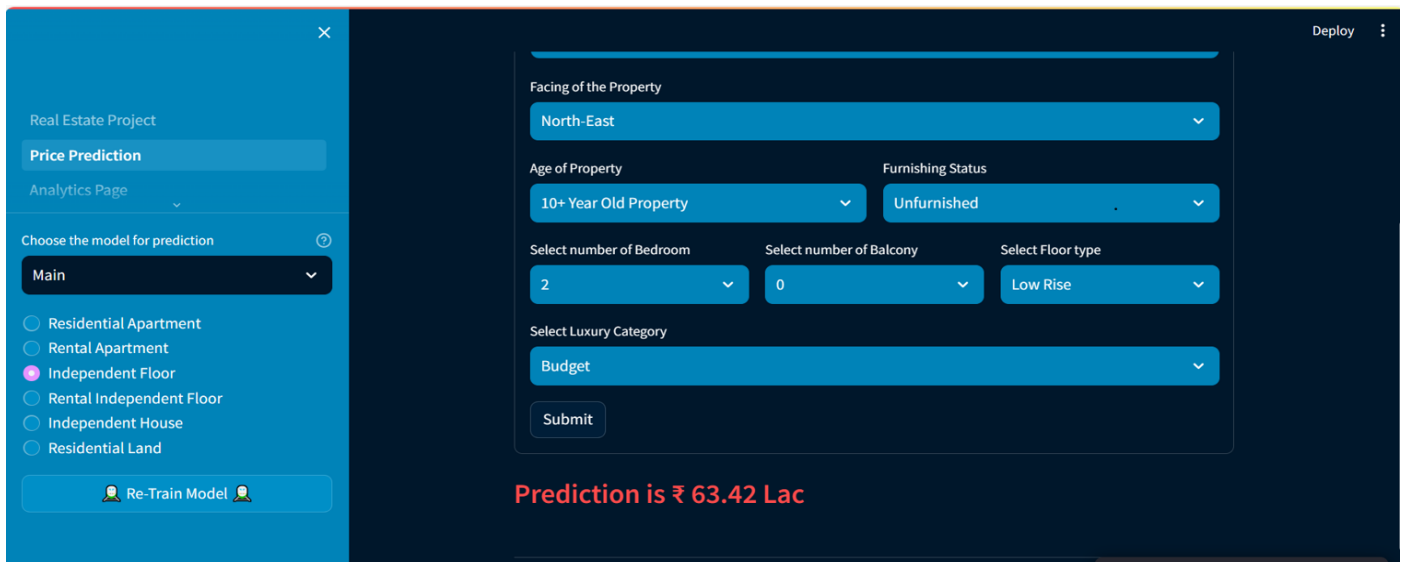


Fig 3.3. Prediction of Independent Floor

In the figure 3.3 we are getting the final price prediction.



Fig 3.4 Home Page of Dashboard

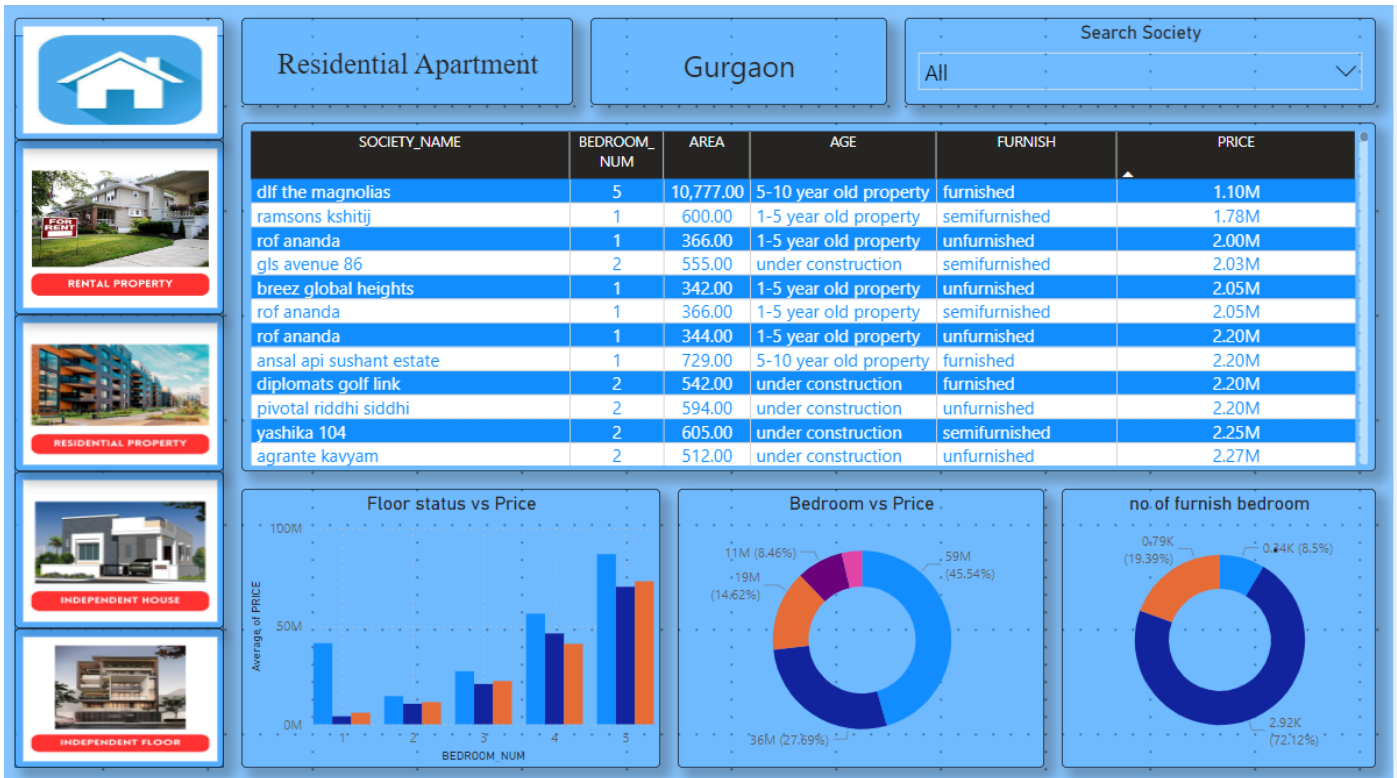


Fig 3.5 Analysis Of Residential Apartment of Gurgaon City



Fig 3.6 Analysis Of Rental Apartment of Thane City

Here user can select the city and can visualize the property of the different city.

VII. CONCLUSION AND FUTURE SCOPE

This study makes an important contribution to the field of Real Estate by creating the price prediction model which predict the price of the house so it is useful for the homeseeker.the model can accurately predict the price of the house on the input provided by the user .Here the xgboost performing the best.Our findings reveal that XGBoost outperforms traditional regression models and other machine learning algorithms in terms of predictive performance. This superiority can be attributed to XGBoost's ability to ensemble weak learners sequentially, effectively capturing complex patterns and interactions within the data.Overall, the results of this study contribute to advancing the state-of-the-art in real estate analytics and offer valuable insights for stakeholders involved in property valuation and investment decision-making. Moving forward, further research could explore additional enhancements to XGBoost or investigate the application of other advanced machine learning techniques in the domain of real estate prediction and we have added some city further city can be added that include all india city data.

Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm widely used for predictive modeling, including real estate house price prediction. Visualizations: Power BI offers a wide range of visualization options. Integration with XGBoost models can enable the creation of visualizations such as trend analysis, geographical heat maps showing price variations, and scatter plots comparing predicted vs. actual prices. Interactive Dashboards: Power BI allows the creation of interactive dashboards that can display real-time or periodic updates on house price predictions based on XGBoost models. Users can interact with these dashboards, filtering data based on location, property type, features, etc., to gain insights into specific market segments.

VIII. REFERENCES

- [1] Anand G. Rawool, Dattatray V. Rogye, Sainath G. Rane, Dr. Vinayk A. Bharadi, "House Price Prediction Using Machine Learning", International Review of Education – Journal of Lifelong Learning (IRE), 2021.
- [2] Vibha B.Bhor,Mohini S.Gaikwad,Prachi S. Zende, "Implementation of Housing Price Prediction", International Journal of Innovative Research in Technology(IJIRT), 2020.
- [3] Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, and Gbenle Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique", ELSEVIER, 2021.
- [4] Winky K.O. Ho, Bo-Sin Tang, and Siu Wai Wong, "Predicting property prices with machine learning algorithms", Journal of Property Research, 2021.
- [5] N. Ragapriya,T. Ananth Kumar,R. Parthiban,P. Divya,S. Jayalakshmi & D. Raghu Raman, "Machine Learning Based House Price Prediction Using Modified Extreme Boosting", Asian Journal of Applied Science and Technology (AJAST), 2023.
- [6] Sheng Bin, "Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model", Hindawi, 2022.
- [7] Tianqi Chen,Carlos Guestrin "XGBoost: A Scalable Tree Boosting System", June 2016.