Ensemble-Based House Price Prediction using Boosting Regressors

Senthil Pandi S
Department of CSE
Rajalakshmi Engineering College
Chennai, India
mailtosenthil.ks@gmail.com

Madhumitha K
Department of CSE, REC
Chennai, India
210701141@rajalakshmi.edu.in

Mohana Prasath G
Department of CSE, REC
Chennai, India
210701163@rajalakshmi.edu.in

Abstract— Prospective homeowners and real estate investors are always on the lookout for solutions that deliver precise, data-driven, and highly customized insights into property prices, especially in the dynamic and ever-fluctuating real estate market. This project centers on the development of a powerful machine learning (ML) website that forecasts home values based on a wide range of user-specified parameters, including location (latitude and longitude), property size (square footage), and configuration (BHK - Bedrooms, Hall, Kitchen). The platform utilizes a vast dataset that not only encompasses geographic data and real estate listings but also integrates additional external data sources, such as current market trends, local amenities, neighborhood features, economic conditions, and historical pricing patterns. By combining these rich data sources with advanced machine learning algorithms, the project aims to create a highly effective tool for optimizing property assessment, improving real estate market insights, and empowering users to make well-informed, strategic decisions in buying or investing in property.

***Keywords*— House Price Prediction, Ensemble Technique, Boosting**

I. Introduction

Propoerty price prediction is a important aspect of the real estate market, influencing decisions made by homebuyers, investors, developers, and policymakers. Accurate prediction models can provide valuable insights into market trends, helping stakeholders make informed decisions. Given the complexity of real estate data—characterized by a multitude of factors such as location, economic conditions, and property features—developing reliable predictive models is challenging yet essential.

This project focuses on leveraging advanced machine learning techniques to predict house prices with greater accuracy. By putting forward novel ideas that increase prediction accuracy and provide useful applications for the real estate sector, we hope to support the ongoing efforts in the field.

Accurately estimating property values in the quickly changing real estate market continues to be a major obstacle for potential homeowners and investors. Conventional property valuation techniques frequently lack accuracy and don't take into consideration the intricate interactions between many elements that affect property values. Developing a predictive model that successfully combines various data sources and cutting-edge machine learning algorithms to produce accurate and customized property price forecasts is the main difficulty. This initiative is driven by the growing need in a turbulent real estate market for more precise and customized real estate information. This project intends to develop an efficient ML-based platform by utilizing a large dataset that contains real estate listings, geographic data, and additional external aspects like market trends and local amenities.

To develop an advanced machine learning (ML)-based website designed to accurately forecast property values by leveraging cutting-edge algorithms and comprehensive datasets. The research goal is to develop a prediction model by combining information from many sources, such as real estate listings, geographic information (latitude, longitude), property features (BHK - Bedrooms, Hall, Kitchen), and additional external factors such as market trends and local amenities. The website will have an easy-to-use design that lets visitors explore property alternatives, enter search parameters, and get forecasts that are specific to them. The ultimate goal is to deliver a valuable resource for homeowners, investors, and real estate professionals, enabling more precise property assessments.

II. Literature Survey

Anders Hjort et al.,[1] proposed a tree boosting model, Locally Interpretable Tree Boosting (LitBoost) that combines the strengths of Generalized Additive models(GAM) and Gradient Boosted Trees(GBT). It improves predictive performance by avoiding overfitting and also enhances predictability. It was observed that GAM displayed poor performance when the observations were small. The proposed method LitBoost outperforms GAM when the number of observations per group is small and performs better than GBT when specific interactions are included in the data1. The performance of Random Forest for house price prediction and have shown that the model is able to achieve a minor difference. Lu Wang et al. proposed a spatiotemporal model (FSTM) to predict the prices of houses using the spatiotemporal characteristics of small cities in China. The model was designed to handle the relationship between space and time reflecting the variation in house prices across different locations and time periods. A flexible structure that combined the long term trends with spatially correlated random factors enabled the model to predict accurately3.

Anders Hjort et al., [2] introduced an improvement in XGBoost by introducing a loss function Squared percentage error (SPE). The introduction of the SPE loss function improves model performance from 88.24% to 91.02% under the 22% measure. XGBoost-SPE reduces incorrect predictions by 4.9% compared to XGBoost-SE, particularly improving in lower price segments, though it performs slightly worse in higher price segments. Combining both models into a hybrid yields even better results, achieving 90.4% accuracy and further reducing errors by 9.3%, demonstrating the effectiveness of this combined approach.

Jun Liu and Zihan Ma [3] proposed a hybrid method by combining advanced neural network models with spatial analysis techniques. that includes spatial factors in house price predictions. This integration not only improves predictive accuracy but also provides a new perspective on incorporating geographical factors into real estate forecasting. To forecast the cost of used homes in Sanghai, Zhongyun Jiang et al., [4] suggested a multilayer feedforward neural network trained using the error inverse propagation approach. SCR is a hybrid approach that was suggested by Sureyya Ozogur Akyur et al., [5] that combines support vector regression, closest neighbor classification, clustering analysis, and linear regression. One method's output is used as another's input. The research will classify the houses for which the cluster is unknown, generate distinct housing clusters using the data at hand, and anticipate prices by developing distinct prediction models for each class in order to provide a hybrid method. Salim Lahmiri et al., [6] used Gaussian process regression, support vector regression, and ensemble regression trees to forecast housing prices in Taiwan. The experimental results showed that boosting ensemble regression trees performed better than Gaussian process regression and support vector regression. All three prediction algorithms fared better than artificial neural networks.

In 2024, Y Zhao et al.,[7] explored the application of multiple machine learning models, including MLP, SVM, Linear Regression, XGBoost, and Random Forest. Their focus was on evaluating the models' accuracy in predictive tasks, highlighting the effectiveness of combining various algorithms for improved performance. In 2019, Feng Wang et al.,[8] investigated the use of SVM and ARIMA models for predictive analysis. They evaluated the accuracy, aiming to enhance the reliability of their predictions in various applications. Wang et al.,[9] conducted a comprehensive study utilizing various machine learning models, including Random Forest Algorithm, Lasso Regression, Linear Regression, and Decision Tree. Advanced neural network architectures, including LSTM, CNN, and RNN, to enhance predictive accuracy in their study. Training Loss, providing valuable insights into their effectiveness and reliability in handling complex data patterns.

III. Methodology

The proposed model follows a structured workflow, the process involves Data Preprocessing, where the data is cleaned, transformed, and made suitable for modeling. In the Model Selection phase, different machine learning models are evaluated to determine the best fit for the problem. Once the model is chosen, Model Training takes place, allowing the model to learn from the data. This is followed by Validation and Testing to assess the model's performance and fine-tune it for optimal results. Finally, the trained model is used to Predict the Output, providing accurate forecasts for house prices based on user input parameters.

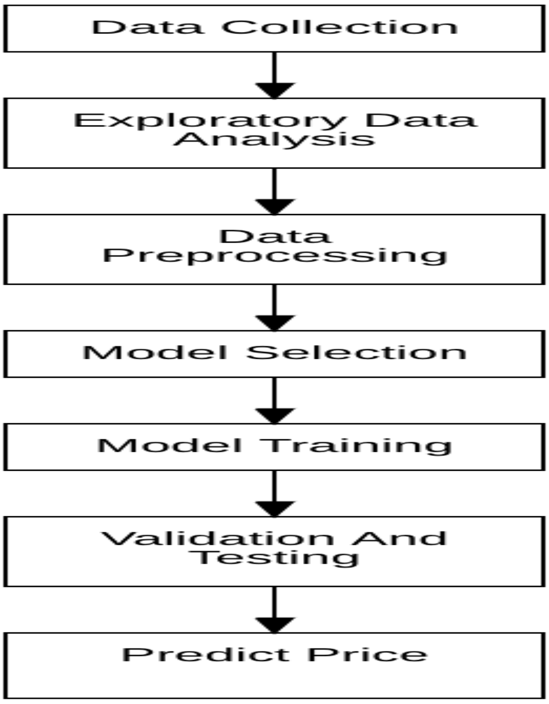


Figure. 1. Workflow model

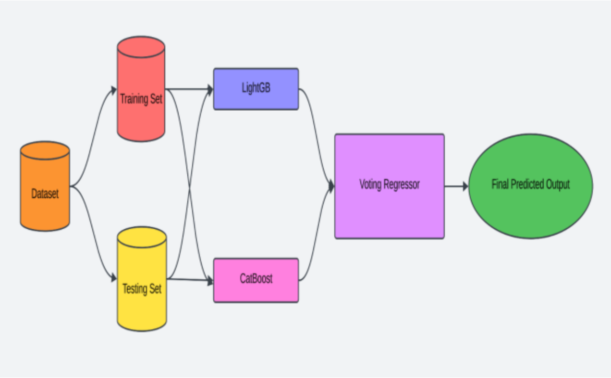


Fig. 2. Architecture Diagram

This architecture diagram presents a comprehensive workflow for predicting house prices using an ensemble of two powerful machine learning models - LightGBM and CatBoost - combined through a Voting Regressor. The system is structured to maximize predictive accuracy by integrating the strengths of both models within a carefully organized pipeline, which begins with data collection and proceeds through multiple stages of data preparation, model training, and prediction generation.

The first stage of the system is Data Collection, where a dataset is compiled with key property attributes, such as location (latitude and longitude), property size (in square feet or square meters), and configuration (number of bedrooms, hall, and kitchen). These features are essential for accurate house price prediction, as they capture both the physical characteristics of the property and its geographic location.

Once the data is prepared, the Model Training phase begins. Here, two machine learning models - LightGBM and CatBoost - are trained on the Training Set. LightGBM, or

Light Gradient Boosting Machine, is a high-performance gradient boosting algorithm that is optimized for speed and efficiency. It handles large datasets well and uses a leaf-wise growth strategy that can reduce computational costs, making it ideal for this application. CatBoost, short for Categorical Boosting, is another gradient boosting algorithm known for its ability to work effectively with categorical features and handle complex relationships within the data with minimal preprocessing. Both models independently learn patterns in the data to predict house prices.

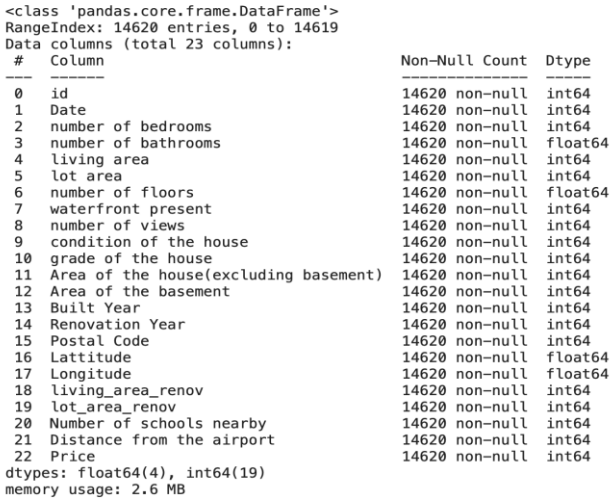
After individual training, the outputs of LightGBM and CatBoost are combined using a Voting Regressor, an ensemble technique designed to improve predictive accuracy. The Voting Regressor consolidates predictions from both models, either by averaging them or assigning weights to each model's output based on their individual performance. This approach capitalizes on the complementary strengths of LightGBM and CatBoost: while LightGBM's efficiency makes it strong in handling high-dimensional data, CatBoost excels in managing categorical variables and complex interactions. By integrating both models, the Voting Regressor reduces errors associated with single-model predictions and enhances the system's robustness.

In the final stage, the Voting Regressor outputs a Final Predicted Output—the house price estimate. This output represents the combined predictive capability of both LightGBM and CatBoost, offering an accurate and reliable estimate. Such a prediction is invaluable to users like prospective homeowners and real estate investors, who can use this information to make informed decisions based on current market trends and specific property characteristics. Additionally, the Voting Regressor's ensemble approach ensures that the final prediction is not only accurate but also generalizes well to new data, reducing the risk of overfitting.

In summary, this architecture leverages the strengths of LightGBM and CatBoost to provide a robust and accurate system for house price prediction. By combining these models in a Voting Regressor ensemble, the system benefits from a balanced approach that draws on the best qualities of both algorithms. This comprehensive and systematic design ensures the system's reliability, making it a powerful tool for real-world applications in real estate analytics. The architecture's structured workflow - encompassing data collection, model training, ensemble prediction, and output generation - demonstrates a thoughtful approach to solving complex regression problems in the domain of property price prediction.

IV. RESULTS & DISCUSSIONS

The dataset was gathered from Kaggle. The dataset used in this study came from Kaggle, a website that offers a wide variety of datasets for analysis and research. ID, are some of the input features that make up the dataset. "Price" is the label assigned to the target variable, or output feature, which represents the house's price. The dataset consists of 14620 rows and 23 columns, including the output feature shown in figure.3.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14620 entries, 0 to 14619
Data columns (total 23 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   id                                      14620 non-null  int64
 1   Date                                    14620 non-null  int64
 2   number of bedrooms                      14620 non-null  int64
 3   number of bathrooms                     14620 non-null  float64
 4   living area                             14620 non-null  int64
 5   lot area                                14620 non-null  int64
 6   number of floors                        14620 non-null  float64
 7   waterfront present                      14620 non-null  int64
 8   number of views                         14620 non-null  int64
 9   condition of the house                  14620 non-null  int64
 10  grade of the house                      14620 non-null  int64
 11  Area of the house(excluding basement)   14620 non-null  int64
 12  Area of the basement                    14620 non-null  int64
 13  Built Year                              14620 non-null  int64
 14  Renovation Year                         14620 non-null  int64
 15  Postal Code                             14620 non-null  int64
 16  Lattitude                               14620 non-null  float64
 17  Longitude                               14620 non-null  float64
 18  living_area_renov                       14620 non-null  int64
 19  lot_area_renov                          14620 non-null  int64
 20  Number of schools nearby                14620 non-null  int64
 21  Distance from the airport               14620 non-null  int64
 22  Price                                   14620 non-null  int64
dtypes: float64(4), int64(19)
memory usage: 2.6 MB
```

Figure.3. Dataset Description

Exploratory Data Analysis- Initially, the dataset should be analyzed to extract meaningful insights. The number of attributes, relationships between features, and the relationship between each input attribute and the output feature are evaluated to better understand the data. This step provides a foundation for model training. Here, techniques like the Confusion matrix, box plot and mutual information for regression are used to analyze the data. Confusion matrices show the relationship of each attribute in the dataset with every other attribute in the dataset.

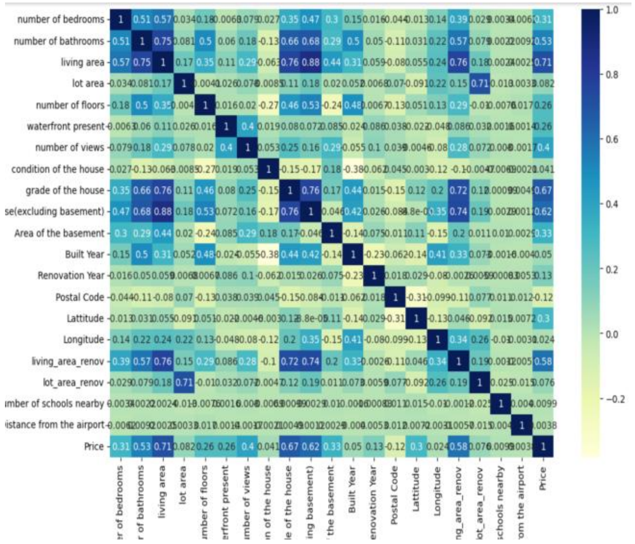


Fig. 4. Confusion Matrix

In the above confusion matrix in Fig 4, the output feature Price has a strong positive correlation with the living area. The price also has a strong positive correlation with grade of the house and area of the house (excluding basement). The Price has a weak negative correlation with Postal code.

A box plot helps to visualize the distribution of numerical data, highlighting any outliers, and can be used with both numerical and categorical features. It provides insights into data symmetry, spread, and skewness.

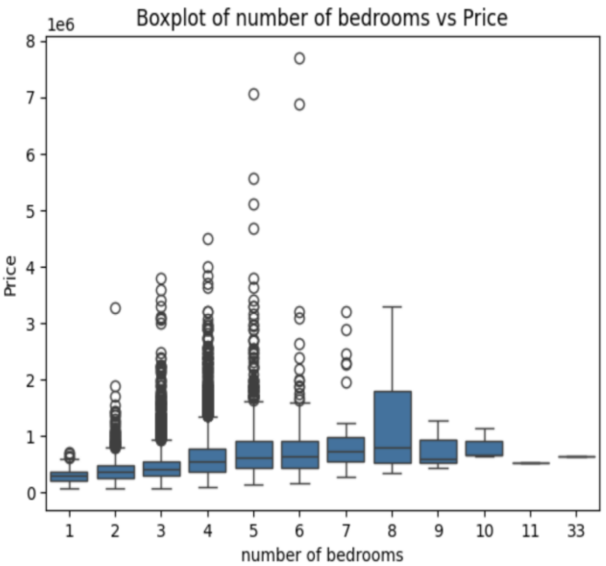


Fig. 5. Box Plot of Number of bedrooms Vs Price

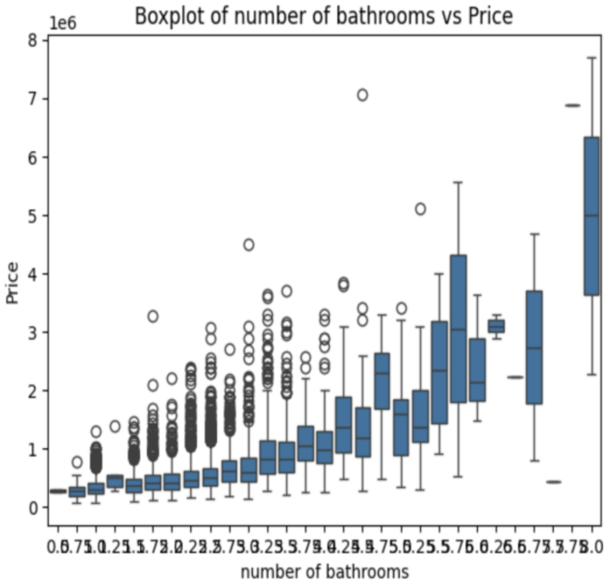


Fig. 6. Box Plot of Number of bathrooms Vs Price

Fig 5 represents the box plot of Price versus number of bedrooms. The price of the home somewhat rises as the number of bedrooms increases. Fig 6 represents the box plot of price versus number of bathrooms. The number of bathrooms and the price have a linear connection. The cost likewise rises linearly with the number of bathrooms.

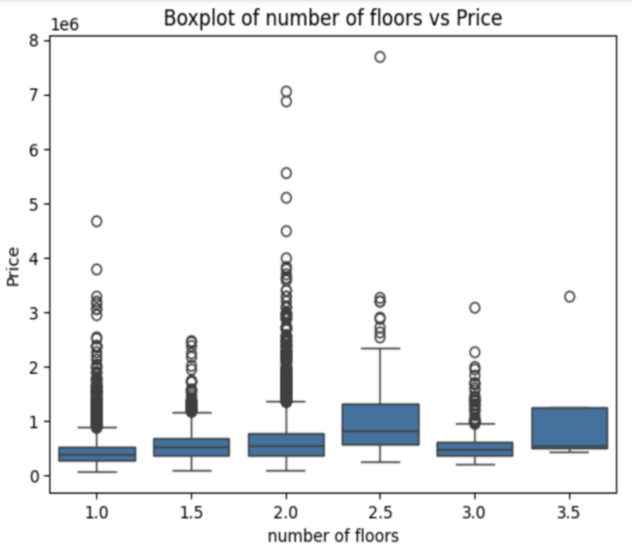


Fig. 7. Box Plot of Number of floors Vs Price

Fig 7 represents the box plot of Price versus number of floors. There is no relation between the number of floors and the price of the price.Fig 8 represents the box plot of condition of the house versus the price. House condition doesn't seem to have a strong linear relationship with the median price. Although better conditions (4 and 5) have more expensive houses, the price distribution is still concentrated towards lower values with a few outliers being much higher.

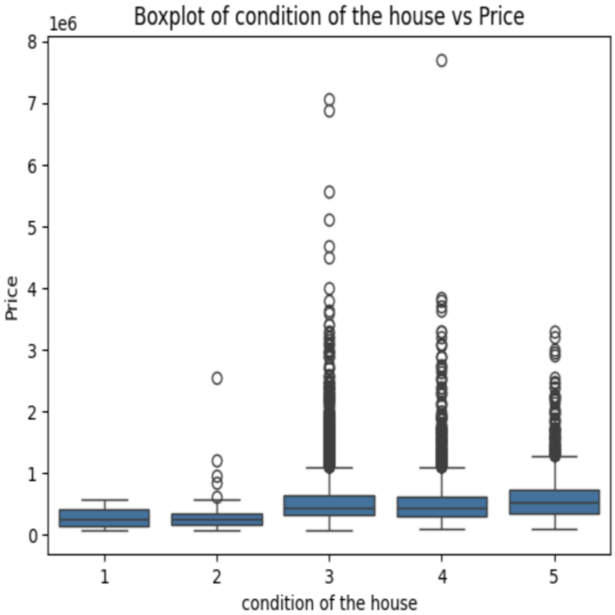


Fig. 8. Box Plot of Condition of the house Vs Price

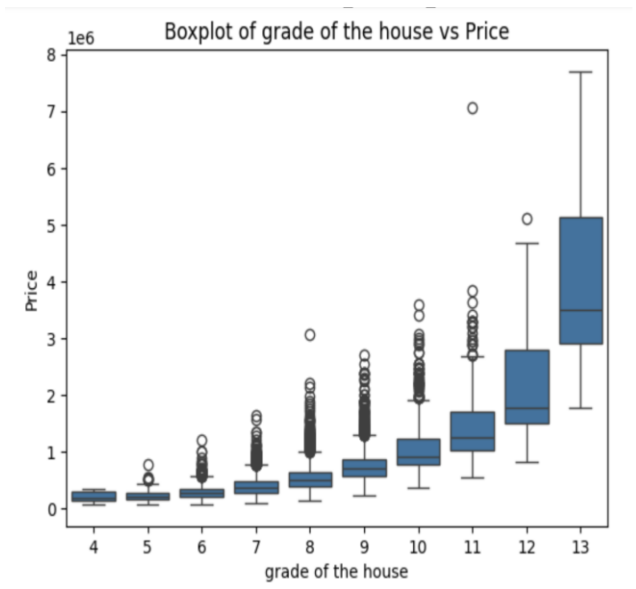


Fig. 9. Box Plot of Grade of the house Vs Price

The box plot showing the house's grade against price is shown in Fig. 9. The grade of the house and the property's price have a strong linear relationship. Mutual information measures the dependency between variables, giving an idea of how much information about the target is provided by each feature. Here postal code gives the high mutual information. Other strong mutual information are given by the features - Living area, Grade of the house, Latitude and living_area_renov.

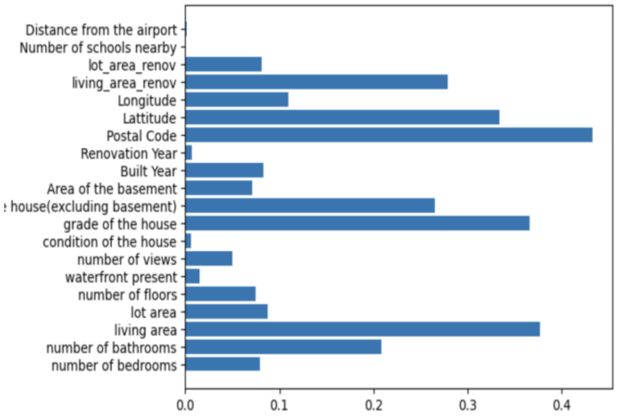


Fig. 10. Mutual Information

Dataset Preprocessing- Before proceeding with model training, it's essential to check for null values and categorical variables in the dataset. Null values can either be removed or filled with appropriate default values (e.g., mean, median, or mode). It is necessary to translate categorical variables into numerical values, frequently with the aid of methods like Label Encoding or One-Hot Encoding. This ensures that the data is clean and ready for the machine learning algorithms, which typically require numerical inputs.

Model Selection and Model Training- In the model selection phase, several machine learning algorithms were evaluated to determine the best approach for predicting house prices. The models considered included Random Forest, XGBoost, Gradient Boosting, Extra Trees, LightGBM, Hist Gradient Boosting, and CatBoost. These models were chosen due to their strength in handling complex, non-linear relationships, making them well-suited for a regression problem like house price prediction.

Each model was trained on the dataset after performing feature selection and preprocessing steps such as handling missing values and encoding categorical variables. The training process involved feeding the models with the input features and their corresponding house prices (target variable). Hyperparameter tuning was also conducted to optimize the performance of each model, ensuring that they could learn patterns in the data effectively.

After training, the performance of each model was evaluated based on the R-squared score, which measures how well the model predicts the variance in house prices. The results indicated strong performance across the board, with CatBoostand LightGBM standing out with the highest R-squared scores.

In the next step, an ensemble method using a Voting Regressor was employed to combine the predictions of the top-performing models—CatBoost and LightGBM. This technique allowed the model to take advantage of the strengths of both algorithms, resulting in improved predictive accuracy and robustness.

The final trained model was then validated using cross-validation, confirming that it generalized well to unseen data, making it suitable for real-world house price prediction tasks.

In the testing phase, several machine learning models were trained and evaluated based on their R-squared scores, which measure how well each model explains the variance in the house price data. The individual R-squared scores were as follows: 88% for Random Forest, 87% for XGBoost, 86% for Gradient Boosting, 88% for Extra Trees, 89% for LightGBM, 87% for Hist Gradient Boosting, and 90% for CatBoost.

These results show that while each model performed well in predicting house prices, CatBoost achieved the highest R-squared score of 90%, followed closely by LightGBM at 89%. Fig. 11 provides a visual representation of the R-squared scores of all the models tested, allowing for easy comparison of their relative performance.

To further improve prediction accuracy, an ensemble approach was considered. Based on the individual model performance, CatBoost and LightGBM were selected for combination using a Voting Regressor. This technique leverages the strengths of both models, leading to a combined R-squared score of 90% on average, showing that the ensemble approach slightly enhanced the performance over individual models.

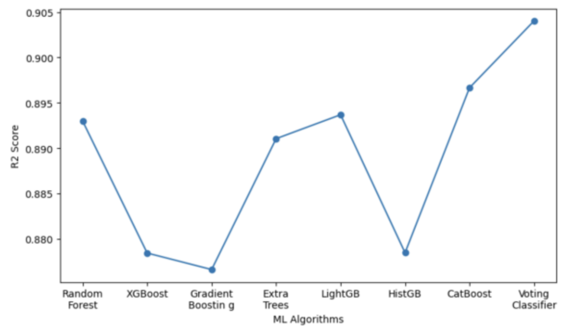


Fig. 11. R2 score of all models

During cross-validation, the dataset was split into multiple folds, and each model was trained and evaluated across these different folds to ensure that the results were not influenced by a particular split of the data. This process helps in estimating the model's generalizability to unseen data. The R-squared scores from cross-validation were plotted for all models, as shown in Fig. 12, to visualize their consistency and robustness across different subsets of the dataset.

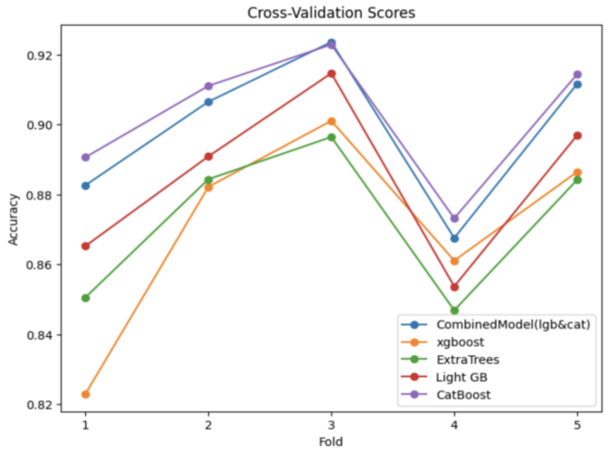


Fig. 12. Cross validation

The graph presented is a line chart depicting the cross-validation scores of various machine learning models across different folds. Each line represents a model: CombinedModel (an ensemble of LightGBM and CatBoost), XGBoost, Extra Trees, LightGBM, and CatBoost. The y-axis shows the accuracy of the models, while the x-axis represents the fold number in the cross-validation process, with values ranging from 1 to 5.

The graph illustrates the effectiveness of the CombinedModel (LightGBM & CatBoost) in achieving more consistent and higher accuracy across cross-validation folds. It suggests that the combination of these two models, leveraging their boosting capabilities, leads to improved performance compared to using them individually. The variability in performance across folds for individual models like LightGBM and XGBoost indicates that while they perform well on some folds, they may be less stable when faced with different subsets of the data, reinforcing the benefit of the ensemble approach for better generalization.

CombinedModel (LightGBM & CatBoost) consistently outperforms the other models in most folds, especially in Fold 2 and Fold 3, where its accuracy peaks at around 0.92, making it the most accurate model overall. It exhibits strong generalization, as its performance remains stable across all folds, except for a slight dip in Fold 4. LightGBM and CatBoost follow closely behind, with LightGBM peaking at around 0.91 in Fold 3 but showing more variation across folds, particularly a noticeable drop in accuracy in Fold 4. Extra Trees and XGBoost show slightly lower performance, with Extra Trees showing the most volatility, dropping below 0.84 in Fold 1 before recovering in subsequent folds.

The final ensemble model, which combined CatBoost and LightGBM, delivered an average R-squared score of 90%. This indicates that the model is able to explain 90% of the variance in house prices, making it a reliable choice for accurate price prediction.

This combination of boosting algorithms like CatBoost and LightGBM proved to be effective due to their ability to handle large datasets, capture nonlinear relationships, and optimize through iterations. The final model not only performed well on the test set but also showed stability during cross-validation.

V. Conclusion

By evaluating a diverse range of models—including Random Forest, XGBoost, Gradient Boosting, Extra Trees, LightGBM, Hist Gradient Boosting, and CatBoost—it was observed that while each model individually performed well, the best results were achieved through an ensemble of CatBoost and LightGBM. With an R-squared score of 90%, this ensemble approach—which capitalized on the complimentary boosting characteristics of both models—showed that the model was very trustworthy, explaining 90% of the variance in home prices. The Voting Regressor ensemble technique allowed for the effective combination of each model's predictive capabilities, leading to significantly improved performance compared to individual models. These results were further validated through cross-validation, where the model's stability, reliability, and consistency were confirmed across different data folds, demonstrating its capacity to effectively generalize to unknown data and perform well across varied scenarios. Additionally, the careful handling of data—including meticulous preprocessing, feature selection, and the use of advanced machine learning techniques—ensured that the model was both accurate and robust. This research demonstrates how effective group learning can be and sophisticated algorithms in tackling complex regression tasks, offering a highly effective, scalable solution for real-world house price prediction and personalized recommendation systems that can assist buyers and investors alike in making well-informed decisions.

References

[1] Hjort, Anders, Ida Scheel, Dag Einar Sommervoll, and Johan Pensar. "Locally interpretable tree boosting: An application to house price prediction." Decision Support Systems 178 (2024): 114106.

[2] Anders Hjort, Johan Pensar, Ida Scheel, and Dag Einar Sommervoll. "House price prediction with gradient boosted trees under different loss functions." Journal of Property Research 39, no. 4 (2022): 338-364.

[3] Jun Liu, and Zihan Ma. "Forecasting Housing Price Using GRU, LSTM and Bi-LSTM for California." In 2024 IEEE ICCECT, pp. 1033-1037. IEEE, 2024.

[4] Jiang, Zhongyun, and Guoxin Shen. "Prediction of house price based on the back propagation neural network in the keras deep learning framework." In 2019 ICSAI, pp. 1408-1412. IEEE, 2019.

[5] Süreyya Özöğür Akyüz, Birsen Eygi Erdogan, Özlem Yıldız, and Pınar Karadayı Ataş. "A novel hybrid house price prediction model." Computational economics 62, no. 3 (2023): 1215-1232.

[6] Salim Lahmiri, Stelios Bekiros, and Christos Avdoulas. "A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization." Decision Analytics Journal 6 (2023): 100166.

[7] Y. Zhao, J. Zhao and E. Y. Lam, "House Price Prediction: A Multi-Source Data Fusion Perspective," in Big Data Mining and Analytics, vol. 7, no. 3, pp. 603-620, September 2024.

[8] F. Wang, Y. Zou, H. Zhang and H. Shi, "House Price Prediction Approach based on Deep Learning and ARIMA Model," 2019 IEEE ICCSNT, Dalian, China, 2019.

[9] Wang, Lu, Guangxing Wang, Huan Yu, and Fei Wang. "Prediction and analysis of residential house price using a flexible spatiotemporal model." Journal of Applied Economics 25, no. 1 (2022): 503-522.

[10] Adetunji, Abigail Bola, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, and Gbenle Oluwadara. "House price prediction using random forest machine learning technique." Procedia Computer Science 199 (2022): 806-813.