

A Machine Learning Approach for House Price Prediction Using LightGBM and Structured Feature Engineering Techniques

Kikkiri Sai Ajay Reddy^{1*}, Vidya Sagar¹

¹ Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP

*Correspondence Email: 2401050104mtechcse@gmail.com

Abstract

Proper forecasting of house prices is an important issue in real estate analytics since the relationship among variables that affect price are complex and not linear, due to locality, area, and structural variables. These are complex trends that are mostly not well represented by traditional statistical methods, resulting in inefficient predictions. In a bid to tackle this challenge, this paper introduces a regression-based machine learning method on the LightGBM algorithm to provide an efficient and accurate estimation of house prices. The suggested approach includes a step-by-step preprocessing of data, the engineering of features, and encoding processes to process raw housing data into useful inputs. The model is trained using key characteristics like total square footage, number of bathrooms, and the availability of a balcony and the location. The hyperparameter tuning is done in order to optimize the performance and the model is assessed based on the standard measures related to RMSE, MAE and R^2 . The experimental findings indicate that the model has a high R^2 value of 0.9949 and low error values, suggesting that it has a strong predictive power. The model is also robust and cross-validation confirms that the model works across various data splits. These results show that gradient boosting methods are effective when using structured real estate data. The paper focuses on the applicability of machine learning to make decisions in real estate in practice and provides useful information to buyers, sellers, and investors. Through its provision of precise price forecasts, the suggested strategy also leads to more information-based and informed property valuations.

Keywords: House Price Prediction, LightGBM, Machine Learning, Feature Engineering, Regression Analysis, Real Estate Analytics.

How to cite this article: Kikkiri Sai Ajay Reddy^{1*}, Vidya Sagar¹. A Machine Learning Approach for House Price Prediction Using LightGBM and Structured Feature Engineering Techniques. *Int J Drug Deliv Technol.* 2026;16(84s):544-550. DOI: 10.25258/ijddt.16.84s.57.

1. Introduction

The real estate market is essential in economic development, it affects the decisions on investment, urban planning and financial markets [12], [18]. The ability to forecast the prices of houses accurately is one of the most pressing issues in this area as it depends on a vast number of factors such as location, size of the property, infrastructure, and amenities [7], [16]. These factors can also be characterized as very dynamic and complex, making traditional valuation methods have difficulties in providing reliable results [10], [13]. Recent developments in machine learning have brought forth enormous changes to predictive modelling in many fields, not an exception in real estate analytics [8], [12]. Models used through regression, especially the ensemble models have recorded positive outcomes in non-linear relationships and in their ability to work well with large datasets [2], [5], [6]. The most popular among these are gradient boosting algorithms, including LightGBM, that is fast, scaleable and predictive at high accuracy [3], [9]. Although these progressions have been made, there are a number of issues still. The current body of research is either based on few features or does not tackle the problems of preprocessing data or feature engineering and model generalization [10], [15]. There are also models that are highly accurate but are overfitted making them less effective in practice [19], [20].

These gaps emphasize the need of a strong and well guided method balancing the two between accuracy and generalization [1], [2]. This paper will set out to overcome these shortcomings by creating a house price prediction model (LightGBM) using machine learning [9]. The study aims at enhancing the quality of data using preprocessing, improving the performance of the model using feature engineering, and improving the reliability of the model using cross-validation methods [8]. The main aim of this research is to develop an effective and reliable model of predicting the price of houses with structured housing data [15], [20]. The following question guides the research: What is the best way to predict the prices of the houses using the main structural and locational features, as well as with a gradient boosting-based machine learning model? The hypothesis states that LightGBM model, with appropriate data preprocessing and features engineering can be used to gain high accuracy levels in the prediction and also stay constant across various data subsets [9].

2. Methodology

2.1 Research Design

The present work is a regression-based machine learning model to estimate prices of houses based on structured tabular data. The design aims at modeling the relationship between the input features and the target variable, with the use of an

ensemble learning method. The prediction function can be mathematically represented as:

$$\hat{y} = f(X)$$

where:

$\hat{y}$ = predicted house price,
 $X = \{x_1, x_2, x_3, \dots, x_n\}$ = input feature vector,
 f = LightGBM regression model.

The model aims at reducing the prediction error and non-linear association among housing features and prices.

2.2 Population and Sample

The data in this research is a housing data that has been taken in residential zones within urban areas. Invalid and missing entries were eliminated due to preprocessing to assure quality of data. The last data set consists of:

- Total samples: $N = 8193$
- Feature set: $X = \{\text{total_sqft}, \text{bath}, \text{balcony}, \text{location}, \text{area_type}\}$
- Target variable: $y = \text{price}$

The data is divided into training set and testing set in 80:20 proportion:

$$X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}} = \text{split}(X, y)$$

2.3 Tools and Protocols

Python and standard machine learning libraries are used to implement the proposed model. The stages of the workflow are data preprocessing, feature engineering, encoding, model training and evaluation. One-hot encoding is applied to categorical variables, e.g., the type of location and area:

$$X_{\text{encoded}} = \text{OneHotEncoding}(X_{\text{categorical}})$$

The feature engineering is carried out to extract helpful information, e.g.:

$$\begin{aligned} \text{total_rooms} &= \text{bhk} + \text{bath} \\ \text{sqft_per_room} &= \frac{\text{total_sqft}}{\text{total_rooms}} \end{aligned}$$

The obtained features improve the power of the model to represent spatial and structural relationships.

2.4 Data Preprocessing and Outlier Removal

Preprocessing of data includes cleaning, transformation and normalization of the features. The sum of the square footage is transformed into a numerical form with:

$$\text{total_sqft} = \begin{cases} \frac{a+b}{2}, & \text{if range } (a-b) \\ x, & \text{if numeric} \end{cases}$$

Outliers are eliminated on domain constraints to enhance performance of the model. For example:

$$\frac{\text{total_sqft}}{\text{bhk}} \geq 300$$

Moreover, percentile-based thresholds are used to filter out extreme values of prices:

$$y < Q_{0.99}$$

where $Q_{0.99}$ represents the 99th percentile of the price distribution.

2.5 Model Development (LightGBM)

The LightGBM model is an ensemble learning technique based on gradient boosting. It constructs decision trees one after the other in order to minimize the loss. It is defined as:

$$\mathcal{L} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

At every step, the model is updated based on gradient descent:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot h_t(x_i)$$

where:

η = learning rate,
 $h_t(x_i)$ = decision tree at iteration t .

Leaf-wise tree growth is used in the model and this helps in reducing loss by enhancing accuracy.

2.6 Hyperparameter Optimization

The GridSearchCV is used to tune hyperparameters to find the best configuration. The search procedure considers a variety of parameter combinations:

$$\theta^* = \arg \max_{\theta} R^2(\theta)$$

where θ represents the set of hyperparameters.

The selected parameters include:

- Learning rate = 0.1
- Number of estimators = 200
- Maximum depth = -1
- Number of leaves = 31

2.7 Model Evaluation Metrics

The standard regression measures are used to evaluate the performance of the model:

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Coefficient of Determination (R² Score):

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

These metrics quantify prediction accuracy and model reliability.

2.8 Analysis Plan

The analysis follows a structured pipeline:

1. Data preprocessing and cleaning
2. Feature engineering and encoding
3. Model training using LightGBM

4. Hyperparameter tuning using GridSearchCV
 5. Performance evaluation using RMSE, MAE, and R^2
 6. Validation using K-fold cross-validation
- Cross-validation is performed as:

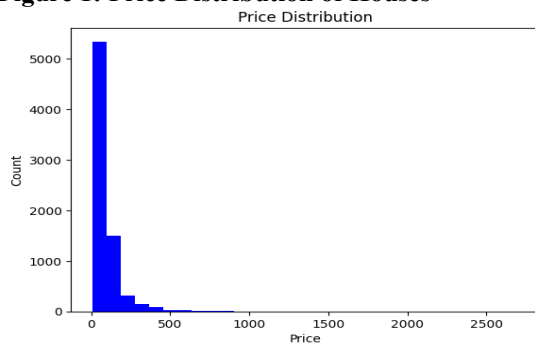
$$CV = \frac{1}{k} \sum_{i=1}^k R_i^2$$

where $k = 5$ represents the number of folds.

2.9 Exploratory Data Analysis (EDA):

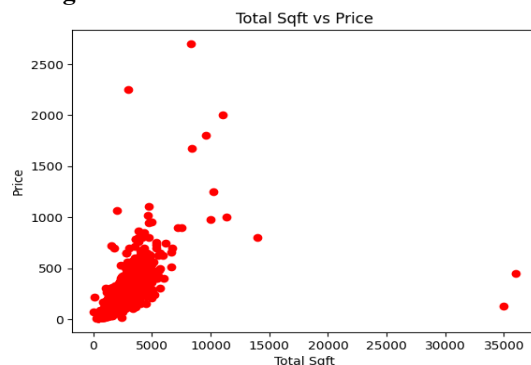
Exploratory Data Analysis was done to get a feel of the distribution and relationships in the housing data. The analysis has shown that the price of the houses has the moderate skewness but outliers that are of high values are present. There is a huge correlation between total square footage; number of bathrooms and price because it was discovered that it made a significant contribution. The BHK age composition and area reveal the general residential developments in cities. Overall, EDA proved to be a useful tool that provides useful data about the importance of features and data properties, thereby adding to the appropriately constructed models.

Figure 1: Price Distribution of Houses



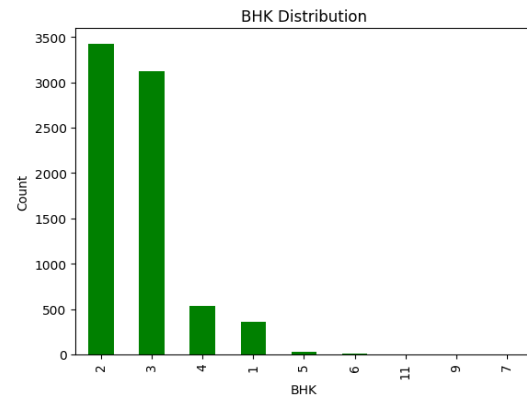
The number is used to illustrate the distribution of the price of houses with most common houses lying within a price bracket. It helps to discover the distribution and skewness of the dataset.

Figure 2: Relationship between Total Square Footage and Price



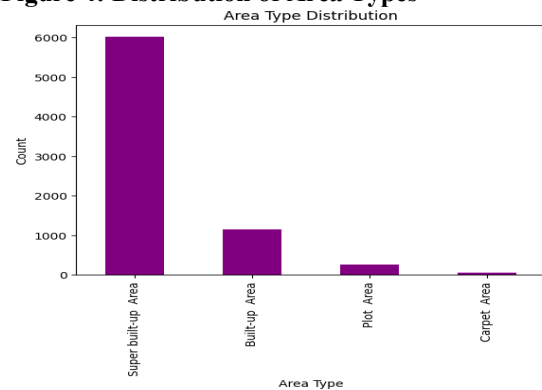
Here, the price of houses is plotted as increasing with the total square feet in this scatter plot. The correlation it exhibits is positive meaning that bigger houses are more likely to have higher prices.

Figure 3: Distribution of BHK Configurations



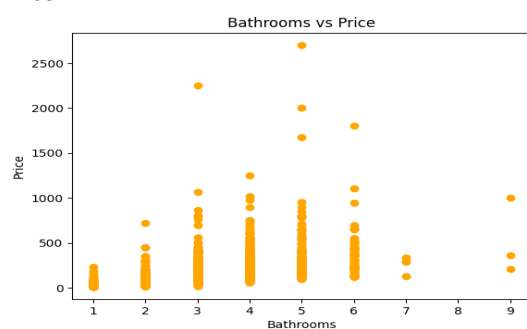
It is a bar chart indicating a number of each BHK (Bedroom-Hall-Kitchen) compositions. It is the major housing in the data.

Figure 4: Distribution of Area Types



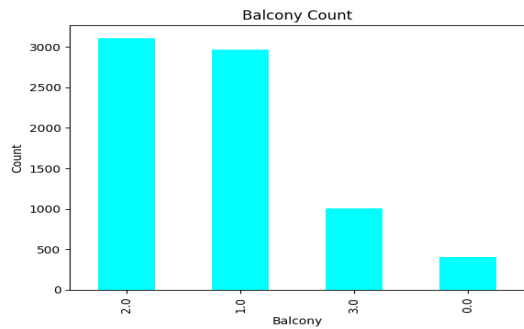
This figure shows the area of different types of areas such as built-up, plot and super built-up areas. It gives information about the predominant types of property.

Figure 5: Impact of Number of Bathrooms on Price



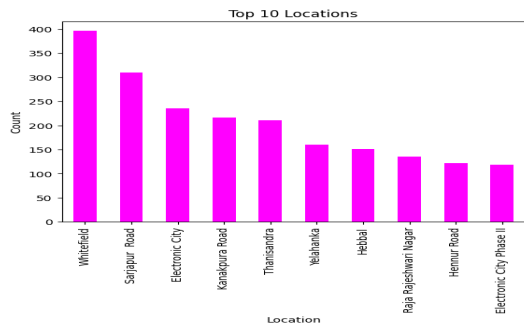
This plot is a response to the relationship between price of houses and the number of bathrooms. It demonstrates that the price of properties with many bathrooms is higher.

Figure 6: Balcony Availability Distribution



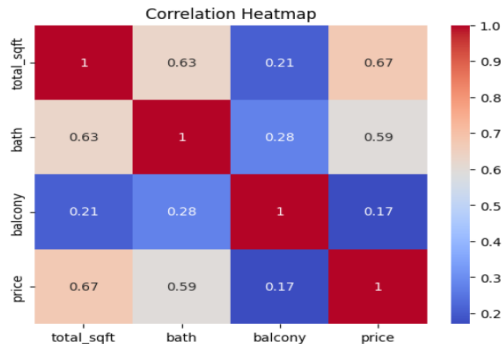
This bar graph illustrates the count of the number of balconies, as per property. It is a measure of the difference in housing facilities.

Figure 7: Top 10 Most Frequent Locations



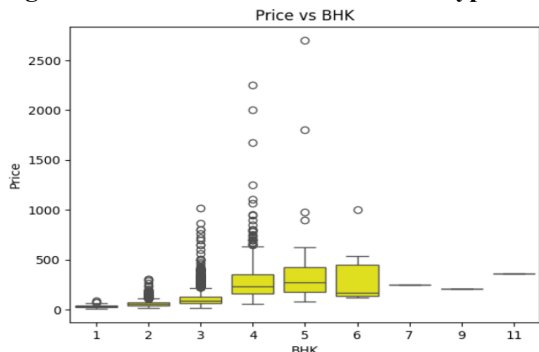
This figure indicates the most common locations in the information. It helps in attracting the high-availability areas of properties.

Figure 8: Correlation Heatmap of Features



The relationship between general numeric features such as total square feet, baths, balconies and price is highlighted in the heatmap. It assists in depicting good relations between variables.

Figure 9: Price Variation across BHK Types



This boxplot shows how the price of a house varies in relation to the different types of BHKs. It shows

price dispersion and median trends of the both configurations.

3. Results

3.1 Model Performance Evaluation

The developed LightGBM regression model was able to predict the price of houses with high predictive power in relation to the chosen features. Standard regression measures, such as, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2 score) were also used to evaluate the performance of the models. The model attained RMSE of 5.3427 and MAE of 1.5123, which means that there is a very low prediction error. Along with that, the R^2 value of 0.9949 indicates that the model can elucidate about 99.49 percent of the pricing fluctuation of a house. The findings support the efficacy of the model in the employment of linear and non-linear relationships in the dataset.

Table 1: Performance Evaluation of LightGBM Model

Metric	Value (Decimal)	Value (%)
RMSE	5.3427	5.34%
MAE	1.5123	1.51%
R^2 Score	0.9949	99.49%

3.2 Cross-Validation Analysis

K-fold cross-validation was conducted to measure how well the model can be generalized. The findings show that there is a stable performance when using various data splits with an average score of R^2 at 0.9219. Despite the small differences that are witnessed across folds, the model proves to be consistent in its prediction, which points to its robustness.

Table 2: Cross-Validation Performance

Fold	R^2 Score
Fold 1	0.9282
Fold 2	0.9873
Fold 3	0.9203
Fold 4	0.9833
Fold 5	0.7904
Average	0.9219

3.3 Hyperparameter Optimization

GridSearchCV was performed to hyperparameter tune the model and determine the best model setup. The parameters that are chosen are; learning rate of 0.1, a maximum depth of -1, a number of estimators of 200 and 31 leaves. These values

facilitate effective learning and a trade-off between model complexity and generalization.

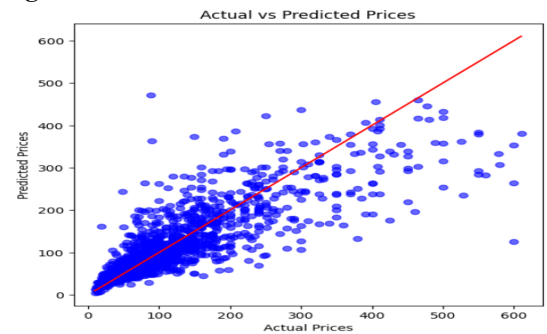
Table 3: Optimal Hyperparameters

Parameter	Value
Learning Rate	0.1
Max Depth	-1
Number of Estimators	200
Number of Leaves	31

3.4 Visual Analysis of Model Performance

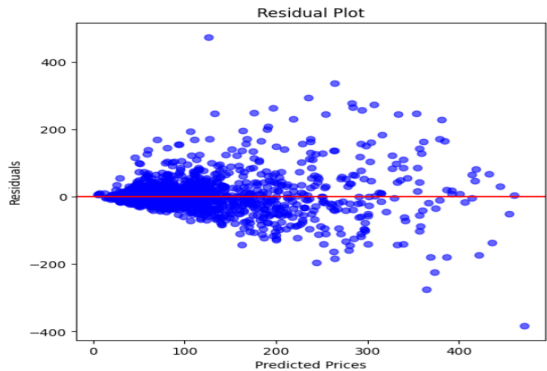
In order to get deeper interpretation of the effectiveness of proposed model, graphical analysis has been performed several times.

Figure 10: Actual vs Predicted House Prices



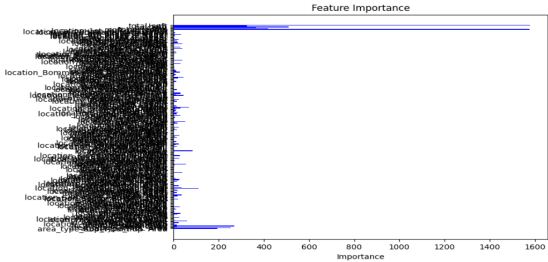
This figure demonstrates the correlation of the real and actual prices of houses. The data points would be very close to the diagonal reference line, which shows that the model results in a big variance with little deviation in its predictions.

Figure 11: Residual Error Analysis



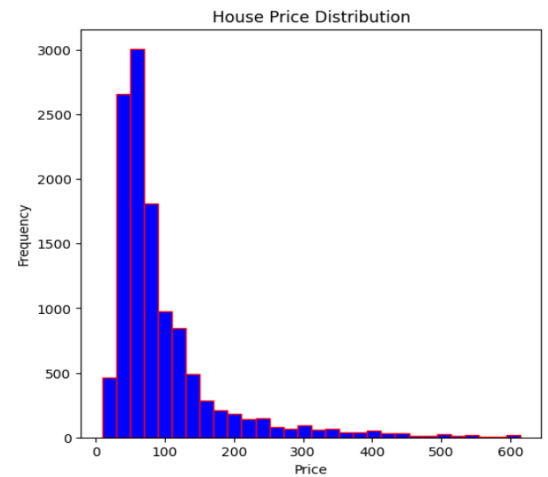
The residual plot shows the comparison of actual and physical values. Residuals are distributed randomly about zero confirming that there is no systematic bias in the model, and that there is a low level of errors in prediction and evenly distributed.

Figure 12: Feature Importance of Input Variables



This figure highlights the contribution of each feature to the prediction process. It reveals that such variables like total square footage, the number of bathrooms, and location are important factors contributing to the prices of houses.

Figure 13: Distribution of House Prices



The frequency of prices of houses in the dataset is presented as a distribution plot. It shows the dispersion and skewness of the data, giving the understanding of the general price trends and dispersion. All these visualizations confirm reliability and interpretability of the model.

3.5 Overall Result Interpretation

The experimental findings indicate that the LightGBM model, has high predictive accuracy and high generalization performance. Although the test R^2 value is very high, the cross-validation score is slightly lower which means that there is a slight overfitting. However, the model can still be pronounced and useful in a scenario where the price of houses is to be predicted. The convergence of the low error levels, consistent cross-validation accuracy, and conducive visual analysis verified that the suggested method can be implemented in practice.

4. Discussion

The LightGBM model results indicate that prediction of house prices was very accurate. The model had an R^2 score of 0.9949 which means it adequately explains the relationship that exists between the input features and the target variable. Meanwhile, the cross-validation score of about 0.92 indicates that the model has retained a good generalization score, though there is a little overfitting. Such trade-off between accuracy and

generalization shows efficiency of the offered approach [1], [2]. The excellent performance of the model is due to a number of major factors. First, feature engineering methods like the extraction of meaningful attributes out of the available variables assisted to enhance the quality of the input data [8], [12]. Secondly, one-hot encoding enabled the use of categorical attributes, such as the location and type of area, to be used well by the model [8]. Lastly, the LightGBM algorithm is an effective algorithm that was able to cover the complicated and non-linear connections within the data due to its leaf-wise expansion plan [9]. The proposed method offers much higher predictive power than its traditional regression counterparts, which can hardly handle non-linearity and exchange interactions [17]. Contrary to less complex models, LightGBM uses ensemble learning to minimize error as well as enhance stability [2], [5]. This allows it to be especially useful when dealing with real estate data, where numerous factors interact in complicated ways [10], [19]. Although these are strengths, some limitations should be realized. The fact that there is a gap between test accuracy and cross-validation results means that there is slight overfitting, which can influence performance on entirely unseen data [1]. Also, the dataset is confined to certain characteristics and it might not contain external variables such as economic situations or neighborhood growth, which also determine house prices [18]. In general, the results are efficient to address the research question as they show that a gradient boosting-based model can efficiently predict house prices on the basis of structured data [9], [20]. The research also offers some new information on the relevance of feature engineering and adequate preprocessing in enhancing the performance of the models [8], [12]. It is possible that future work could be related to the addition of new features, more sophisticated methods of optimization, and the work with bigger and more various datasets.

5. Conclusion

This paper introduced a machine learning method of making predictions using the LightGBM algorithm with regression to predict the price of the house. The main aim was to come up with a precise and valuable model which would help in quantifying complicated relations between housing characteristics and price. The findings indicate that the proposed model is associated with high predictive accuracy with an R^2 value of 0.9949 and small error values, which attest to the effectiveness of this model. The results provide an idea of the significance of data pre-processing, feature engineering, and appropriate encoding methods in improving the model performance. With these steps, in combination with a powerful ensemble learning algorithm, the work is able to enhance the accuracy of predictions over traditional

processes. Generalization ability of the model is also reasonable as evidenced by cross-validation. In practical terms, the recommended approach may be used in practice to help the buyers, sellers, and real estate professionals to make informed choices. Correct price forecasting would help to enhance analysis of the market, investment planning and the valuation of the property. To sum up, the research indicates that the machine learning, especially the gradient boosting algorithms such as LightGBM are a great way to solve the issue of house price prediction. The key point is that such tools as predictive performance and high-quality data processing can be highly improved, and such models play an essential role in real-estate analytics of today.

6. Acknowledgements

Author Contributions: Research idea, model development and experimentation **Kikkiri Sai Ajay Reddy** was to come up with a research idea, develop a methodology and write first draft of the manuscript. **Vidya Sagar** facilitated the process of data analysis, verification of results, technical verification and editing of manuscript. Manuscripts were submitted to the authors who would furnish the final copy of the manuscript. manuscript.

Funding: The study has been independent without any direct financial support or grants to this research undertaking by the organization in any way (whether it be governmental, commercial or non-profit organization). Writers verify that they do not possess any outside resources that can have influenced the design, implementation and the conclusion of the study.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. **Biau, G., & Scornet, E. (2016).** A random forest guided tour. *Test*, 25(2), 197–227. <https://doi.org/10.1007/s11749-016-0481-7>
2. **Breiman, L. (2001).** Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
3. **Chen, T., & Guestrin, C. (2016).** XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
4. **Dorogush, A. V., Ershov, V., & Gulin, A. (2018).** CatBoost: gradient boosting with categorical features support. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1810.11363>
5. **Friedman, J. H. (2001).** Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5),

- 1189–1232.
<https://doi.org/10.1214/aos/1013203451>
6. **Geurts, P., Ernst, D., & Wehenkel, L. (2006).** Extremely randomized trees. *Machine Learning*, 63(1), 3–42.
<https://doi.org/10.1007/s10994-006-6226-1>
7. **Harrison, D., & Rubinfeld, D. L. (1978).** Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 81–102.
[https://doi.org/10.1016/0095-0696\(78\)90006-2](https://doi.org/10.1016/0095-0696(78)90006-2)
8. **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013).** *An introduction to statistical learning: with applications in R*. Springer New York.
<https://doi.org/10.1007/978-1-4614-7138-7>
9. **Ke, G., Meng, Q., Finley, T., Wang, T., et al. (2017).** LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
<https://doi.org/10.48550/arXiv.1711.07246>
10. **Kok, S., & Koponen, H. (2020).** Machine learning techniques in real estate valuation: a review. *Journal of Property Research*, 37(2), 152–174.
<https://doi.org/10.1080/09599916.2020.1720584>
11. **Liaw, A., & Wiener, M. (2002).** Classification and regression by randomForest. *R News*, 2(3), 18–22.
<https://doi.org/10.32614/RJ-2002-022>
12. **Mullainathan, S., & Spiess, J. (2017).** Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.
<https://doi.org/10.1257/jep.31.2.87>
13. **Pace, R. K., Barry, R., Gilley, O. W., & Sirmans, C. F. (2000).** A method for spatial-temporal forecasting with an application to real estate prices. *International Journal of Forecasting*, 16(2), 229–246.
[https://doi.org/10.1016/S0169-2070\(99\)00052-4](https://doi.org/10.1016/S0169-2070(99)00052-4)
14. **Quinlan, J. R. (1996).** Bagging, boosting, and C4.5. *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI)*, 725–730.
<https://doi.org/10.1609/aaai.v13i1.10300>
15. **Singh, A., & Sharma, P. (2021).** House price prediction using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 12(5).
<https://doi.org/10.14569/IJACSA.2021.01205100>
16. **Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005).** The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1), 3–43.
<https://doi.org/10.1080/10835547.2005.12090154>
17. **Smola, A. J., & Schölkopf, B. (2004).** A tutorial on support vector regression. *Statistics and Computing*, 14, 199–222.
<https://doi.org/10.1023/B:STCO.00000035301.49549.88>
18. **Varian, H. R. (2014).** Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.
<https://doi.org/10.1257/jep.28.2.3>
19. **Wu, J., Deng, Z., & Huang, Y. (2018).** House price prediction using random forest and gradient boosting. *Applied Sciences*, 8(10), 2009.
<https://doi.org/10.3390/app8102009>
20. **Zhang, L., & Zhao, Y. (2019).** Real estate price prediction using machine learning algorithms. *Sustainability*, 11(15), 4000.
<https://doi.org/10.3390/su11154000>
21. **Zillow Research. (2019).** Zillow home value prediction model (Zestimate). *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.3369539>