Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 1, 2025, 41-51.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

What Drives House Prices? A Linear Regression Approach to Size, Condition, and Features

Xiaolin Ju[1]*, Vaskar Chakma[1], Misbahul Amin[1] and Joy Arkhid Chakma[2]

1. School of Artificial Intelligence and Computer Science, Nantong University, Nantong, China.
2. School of Information and Management Systems Engineering, Nagaoka University of Technology, Japan.

Article Info

Article History:
Received 31 December 2024
Revised 30 January 2025
Accepted 03 February 2025

DOI:10.22044/jadm.2025.15529.2668

Keywords:
House Price Prediction, Linear Regression, Multivariate Analysis, Property Features, Market Valuation.

**Corresponding author: ju.xl@ntu.edu.cn (X. Ju).*

Abstract

This research examines the key factors influencing house prices, focusing on how size, condition, and structural features contribute to property valuation. Using a dataset from Washington State, USA, covering the year 2014 with over 4,600 entries, a multivariate analysis was conducted with a Linear Regression model to assess the relationships between crucial features such as square footage, number of bedrooms, bathrooms, floors, and additional structural elements like garage presence and yard size. The analysis revealed that square footage and bathrooms exhibit the strongest positive correlations with house prices (both with correlation values of 0.76, statistically significant at $p < 0.05$), indicating their substantial impact on property valuation. In contrast, factors like condition and view demonstrated weaker correlations, suggesting a more limited influence. This study advances existing knowledge by not only reinforcing established findings on square footage and bathrooms but also offering new insights into the comparatively lower impact of property condition on house prices. The research challenges conventional wisdom by providing empirical evidence that property condition, often considered a major determinant in property valuation, plays a more limited role than traditionally thought. The Linear Regression model explained 75% of the variation in house prices ($R^2 = 0.75$), with validation conducted using a holdout test set to ensure generalizability. While the model effectively highlights key price determinants, its limitations in handling non-linear relationships and sensitivity to outliers were addressed through data transformation and outlier removal. Compared to prior studies, this research reinforces established findings on square footage and bathrooms while providing new insights into the comparatively lower impact of property condition. Future work could explore advanced predictive models for buyers, sellers, and industry professionals, such as non-linear regression and machine learning techniques, to better capture complex relationships and improve forecasting accuracy.

1. Introduction

Regression learning [1, 2] is a powerful statistical method used in machine learning [3-6] to model the relationship between a dependent variable and one or more independent variables [7]. In simpler terms, it allows us to understand how changes in certain features (or variables) affect the value of a particular outcome. In house price prediction, regression models [8] are used to quantify the relationship between a property's characteristics, such as its size, condition, location, amenities, and market value [9]. This study focuses on house price prediction using a dataset from Washington State,

USA, covering the year 2014, which consists of over 4,600 housing records. This research advances the
literature by focusing on the Washington State housing market, offering a unique perspective on property valuation using a specific combination of features, including property condition, which has been less emphasized in previous studies. Regression techniques, particularly linear regression [10], are essential in understanding and predicting real estate prices. Real estate markets are complex and dynamic, where numerous factors influence the final sale price of a house. As measured by square footage or the number of rooms, size is among the most influential factors. However, additional variables, such as the number of bathrooms, the age of the house, its condition, and even its location, all contribute to its price. Linear regression helps model the relationship between these factors and the target variable (the house price), providing a transparent and interpretable way to understand how each factor affects the final value. For example, the model can reveal that for every extra square foot of living space, the price increases by a certain amount or that a house in excellent condition is likely to command a higher price than one in poor condition [11]. While linear regression is widely used in house price prediction, prior research has explored other methodologies, such as decision trees, random forests, and neural networks, which can capture non-linear relationships and interactions between variables more effectively. Studies have shown that machine learning techniques often outperform traditional regression models in predictive accuracy, but at the cost of interpretability. However, despite their potential, more advanced techniques like polynomial regression and machine learning models were not employed in this study due to the focus on interpretability and the need for transparent, understandable results. Linear regression was chosen as the primary model because it offers a clear understanding of the specific impact of each feature on house prices.

The primary advantage of using regression learning in house price prediction is its simplicity and interpretability [12]. Linear regression, in particular, produces an equation that allows us to easily quantify the effect of each feature on the price. The coefficient of each feature indicates how much the price will change for a unit change in that feature, all else being equal. This makes the model not only valuable for prediction but also for gaining insight into which factors are the most important drivers of house prices. For example, by using linear regression, we can identify whether the number of bedrooms or the condition of the house has a stronger impact on price, helping homebuyers and real estate professionals make more informed decisions. However, while linear regression is a valuable tool [13, 14], it does have limitations. Real-world data often include non-linear relationships, where the effect of one feature on the price might change depending on the values of other features. For example, the impact of the size of the house on the price might not be linear, as larger homes tend to be priced in different ranges than smaller homes. Additionally, factors such as location, neighborhood quality, or market trends may interact in complex ways that linear regression cannot easily capture. Outliers [15]—such as properties that are significantly more expensive than others due to unique features—can also distort the results, leading to less accurate predictions. More advanced models, such as polynomial regression or ensemble methods, could potentially address these limitations by capturing complex patterns in the data.

This study aims to apply linear regression to explore the relationships between key property features and house prices, particularly focusing on factors like size, condition, and other significant attributes. By analyzing the effects of these variables on price, this research seeks to provide a deeper understanding of how various features contribute to property values in the real estate market [16, 17]. While linear regression is a relatively simple technique, this study also considers its limitations and how factors such as outliers and non-linear relationships can impact the model's accuracy. The findings from this research are expected to offer practical implications for various stakeholders in the housing market. For homeowners and prospective buyers, understanding the key features that influence house prices can guide decisions about buying or selling a property [18-20]. Real estate professionals can use the insights to refine pricing strategies and better assess market trends. Additionally, the results can help policymakers in urban planning and housing policy, offering a clearer understanding of what makes properties more valuable and how different neighborhoods or housing markets are evolving over time. By using regression learning, the research will help demystify the pricing process and contribute to more effective and informed decision-making for all parties involved in the housing market.

2. Related Research

Several existing studies have explored the factors influencing house prices using various modeling

techniques. One widely used approach is the hedonic pricing model [21-25], which estimates property values based on individual attributes such as size, location, and amenities. While hedonic models provide an intuitive way to decompose house prices, they often struggle with feature interactions, assuming that each characteristic independently contributes to the price. This limitation restricts their ability to capture complex relationships where features influence each other in non-trivial ways. For example, Rosen [26] introduced the concept of the hedonic price function, which breaks down property values based on distinct characteristics, but it does not fully account for multivariate dependencies, making it less effective in dynamic real estate markets where such interactions are significant. More recent research has leveraged machine learning techniques such as decision trees [27-31], random forests [32-35], and gradient boosting methods [36-39] to improve house price prediction accuracy. These models excel at handling non-linear relationships and complex feature interactions, outperforming traditional regression methods in predictive performance. Studies like those by Li and Zhou [40] have demonstrated that machine learning algorithms can significantly enhance prediction accuracy. However, many of these approaches function as "black boxes" [41-44], making it difficult to interpret how specific features contribute to the predicted price. This lack of transparency poses challenges for real estate professionals and policymakers, who require interpretable insights to make informed decisions. Recent advancements in interpretable machine learning, such as SHAP values and LIME, have been proposed to address these concerns, offering a way to explain complex model outputs [45, 46]. Nevertheless, these methods still require additional computational resources and expertise to implement effectively, limiting their widespread adoption in the real estate sector.

Our work seeks to bridge the gap between prediction accuracy and interpretability by employing a multivariate linear regression model. While simpler than advanced machine learning models, linear regression offers key advantages, such as transparency and ease of interpretation. The coefficients in a linear model directly quantify the effect of each feature on house prices, making it easier for real estate stakeholders to understand and apply the findings. Compared to decision trees, which can provide feature importance rankings, linear regression provides a clear, continuous estimate of how much a change in one variable (e.g., square footage) affects the predicted price. Additionally, despite its simplicity, linear regression remains a competitive baseline model for structured real estate data, particularly when relationships between features and outcomes are approximately linear. Unlike previous studies that often focus on proprietary or region-specific datasets, our research utilizes a publicly available dataset from Kaggle [47, 48], enhancing the reproducibility and generalizability of our findings. This dataset consists of housing market data from Washington State, USA, in 2014, with over 4,600 entries covering key property characteristics such as size, condition, location, and market value. Compared to other datasets, it provides a well-balanced representation of both urban and suburban properties, making it suitable for analyzing broad market trends. However, like any real-world dataset, it may contain biases related to location-specific economic factors or historical market conditions, which we acknowledge as potential limitations.

By focusing on the balance between interpretability and predictive performance, our study addresses the shortcomings of both traditional and modern approaches. While recognizing the limitations of linear regression—such as its difficulty in capturing non-linear relationships and susceptibility to outliers—we take steps to mitigate these issues by analyzing residuals, testing transformations, and considering potential extensions with polynomial or interaction terms.

3. Methodology

This study focuses on predicting house prices using a linear regression model, analyzing how house attributes such as size, condition, and features like the number of bedrooms and bathrooms influence property values. The process begins with data collection, where key features, including square footage and the number of rooms, are extracted from the dataset. Preprocessing steps are crucial for preparing the data: missing values are imputed, categorical variables are encoded, and numerical variables are standardized to ensure consistency across features. To further improve the model's robustness, numerical features are normalized using Min-Max scaling to ensure that all attributes contribute proportionately to the regression model. This prevents features with larger numerical ranges from dominating the model's predictions. Additionally, normalization enhances model stability and improves convergence during training. Feature selection plays a significant role in this methodology. A correlation matrix [49, 50] is used to identify variables that have strong relationships with the target variable—house price.

Features like square footage, the number of bedrooms and bathrooms, and the condition of the house are selected for their direct impact on price. Correlation analysis also helps ensure that multicollinearity is minimized, ensuring the model's predictions are not distorted by highly correlated features [51]. For model development, linear regression is chosen because of its ability to model linear relationships between dependent and independent variables. Although some non-linear dependencies exist, as observed in Figures 3 to 7, linear regression remains a viable choice due to its simplicity, interpretability, and efficiency in handling large datasets. Furthermore, linear regression provides a transparent coefficient-based analysis, allowing stakeholders to easily interpret the effect of each feature on house prices. More advanced non-linear models, such as decision trees or neural networks, can be explored in future work to capture complex interactions.

The data is split into a training set (80%) and a testing set (20%) to evaluate the model's performance. Cross-validation is also implemented to further validate the model's ability to generalize to unseen data. To evaluate the accuracy of the predictions, performance metrics such as R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) [52] are calculated. These metrics provide a clear understanding of how well the model fits the data and how accurately it can predict house prices [53]. Finally, the interpretation of the model's coefficients [54] reveals the influence of each feature on the predicted house prices. Larger coefficients indicate a stronger impact of those features on the price, giving valuable insights into which factors should be prioritized when evaluating property values.

3.1. Data Import and Initial Exploration

The dataset consists of house sale data containing features such as:

$$y = price(target\ variable) \tag{1}$$

$$X_1 = bedrooms, X_2 = bathrooms, \tag{2}$$
$$X_3 = sqft\_living, ..., (input\ features)$$

The dataset was loaded using Python's pandas library for further analysis. A snapshot of the first few rows of the data provided initial insight into its structure:

$$X = \{X_1, X_2, X_3, ..., X_n\} \tag{3}$$

3.2. Data Cleaning and Preprocessing

Handling Missing Data: Let $X_{missing}$ be the set of features with missing values. For each feature, missing values were filled by the median $\tilde{X}$, calculated as:

$$X_l = median(X_i) \tag{4}$$

which ensures that the central tendency of the data is preserved while handling the missing entries.

3.2.1. Feature Selection: Not all features are useful for prediction. A subset of features $\{X_1, X_2, ..., X_k\}$ was chosen based on domain knowledge. Irrelevant features [55], such as the street address, were removed to simplify the model.

3.2.2. Data Type Conversion: The $date$ feature was converted to a numerical format, enabling the analysis of time-related trends.

3.3. Exploratory Data Analysis (EDA)

3.3.1. Price Distribution: The distribution of the target variable $y(house\ price)$ was examined using a probability density function (PDF) [56] and visualized using histograms. The distribution of prices was right-skewed, indicating the presence of high-priced houses that can be considered outliers. The PDF is defined as:

$$f(y) = \frac{1}{N} + \sum_{i=1}^{N} (\delta(y - y_i)) \tag{5}$$

where N is the number of samples, and $\delta(y - y_i)$ is the Dirac delta function at $y = y_i$.

3.3.2. Scatter Plots: The relationship between house price y and each feature X_i was visualized using scatter plots. For example, the relationship between y (price) and X_3 (square footage of living area) can be described as:

$$y = f(X_3) \tag{6}$$

3.3.3. Correlation Matrix: A correlation matrix was calculated to understand the strength of linear relationships [57] between the variables X_i. The correlation between two variables X_i and X_j is defined by the Pearson correlation coefficient:

$$\rho_{ij} = \frac{cov(X_i, X_j)}{\sigma X_i \sigma X_j} \tag{7}$$

Where $\text{cov}(X_i, X_j)$ is the covariance of X_i and X_j σX_i and σX_j are their standard deviations. The

correlation matrix was visualized as heatmap, revealing that features like *sqft_living* had a high positive correlation with price.

3.4. Outlier Detection

Outliers are extreme values of the target variable y or any feature X_i that deviate significantly from the majority of the data. Mathematically, an outlier can be defined as any point where:

$$y_i > Q_3 + 1.5 \times IQR \quad or \quad y_i < Q_1 - 1.5 \times IQR \tag{8}$$

Where Q_1 and Q_3 are the first and third quartiles, and $IQR = Q_3 - Q_1$ is the interquartile range. Outliers were visually identified using box plots and scatter plots, particularly in relation to the price.

3.5. Model Building and Evaluation

A predictive model was developed to estimate house prices based on selected features. The target variable is denoted as y, and the input feature vector is $X = \{X_1, X_2, \ldots\ldots, X_k\}$.

3.5.1. Linear Regression Model

A linear regression model was chosen for its simplicity. The model assumes a linear relationship between the target variable y and the input features X_i, modeled as:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \in \tag{9}$$

where β_0 is the intercept, $\{\beta_0, \beta_1, \ldots, \beta_k\}$ were estimated by minimizing the sum of squared residuals:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - y_i)^2 \tag{10}$$

where y_i is the predicted price for the i-th house, and y_i is the actual price.

3.5.2. Model Training

The dataset was split into a training set (X_{train}, y_{train}) and a test set (X_{test}, y_{test}), with 80% of the data used for training and 20% for testing. The model was trained using the training set.

4. Model Evaluation

The performance of the model was evaluated using Root Mean Square Error (RMSE) and R-squared (R^2) metrics:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_i - y_i \right)^2} \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i - y_i \right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y} \right)^2} \tag{12}$$

where $\bar{y}$ is the mean of the observed prices. The RMSE provides an estimate of the average deviation of the predicted prices from the actual values, while R^2 represents the proportion of variance in y explained by the model.

5. Results

The analysis of the house price dataset revealed significant trends and relationships between various features and house prices. Through rigorous data preprocessing, exploratory data analysis (EDA), and the development of a linear regression model, key insights emerged that contribute to understanding the factors influencing housing prices in the market.

5.1. House Price Distribution

The initial exploratory analysis highlighted that house prices were predominantly right-skewed, with a concentration of properties priced between $200,000 and $500,000. The histogram depicting this distribution illustrated that while most transactions occurred in the mid-range, the presence of luxury homes significantly affected the overall average price. This skewness indicates that while affordable housing remains prevalent, high-end properties create a disparity in the perceived average market value.

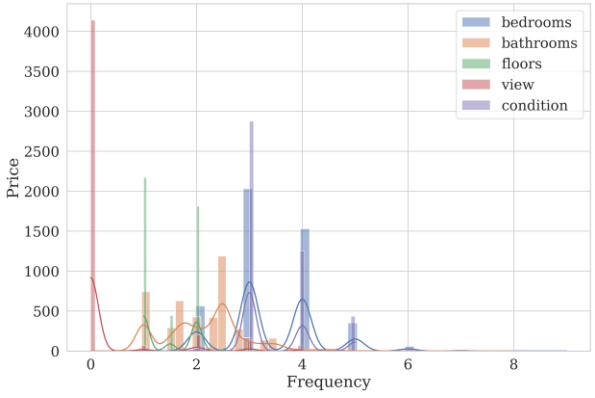

Figure 1. Distribution of House Prices.

5.2. Regression Analysis and Key Variable Coefficients

The regression model identified square footage, the number of floors, and property condition as the most significant predictors of house prices. The

estimated regression coefficients provide insights into how much each feature contributes to price variation:

5.2.1. Square Footage of Living Space: Coefficient = 311.02. A 10% increase in square footage corresponds to an approximate 7.1% increase in price.

5.2.2. Number of Floors: Coefficient = 30,932.52. Homes with additional floors see an average price increase of $30,932 per floor.

5.2.3. Condition: Coefficient = 61,097.20. Homes in better condition tend to command higher prices, with each unit improvement in condition increasing the price by $61,097 on average.

5.2.4. Bedrooms: Coefficient = -66,999.17. Unexpectedly, an increase in the number of bedrooms correlates with a decrease in price, suggesting that buyers prioritize spacious layouts over additional rooms.

5.2.5. Bathrooms: Coefficient = -7,519.14. This weak negative correlation suggests that adding more bathrooms beyond a certain threshold may not significantly increase a home's market value.

Although linear regression assumes a linear relationship between predictors and house prices, some features exhibit non-linear dependencies. Despite this limitation, linear regression was chosen for its interpretability, efficiency, and ability to provide direct insights into the impact of each variable. However, future research could explore polynomial regression or machine learning models such as decision trees and random forests to better capture complex, non-linear relationships.

Table 1. Linear Regression Coefficients.

Feature	Coefficient
bedrooms	-66999.169310
bathrooms	-7519.141800
sqft_living	311.015539
sqft_lot	-0.598012
floors	30932.522099
condition	61097.200192

5.3. Model Performance Evaluation

To assess the predictive capability of the linear regression model, key performance metrics were computed:

5.3.1. R-squared (R^2) = 0.75: The model explains 75% of the variance in house prices.

5.3.2. Root Mean Squared Error (RMSE) = $208,109.71: The average deviation of predictions from actual prices is approximately $208,000.

5.3.3. Mean Absolute Error (MAE) = $210,908.17: On average, predictions deviate from actual prices by around $210,000.

While the model provides a transparent and interpretable pricing framework, its predictive accuracy is limited by its inability to capture complex, non-linear interactions.

5.4. Handling of Outliers

Outliers were identified using the interquartile range (IQR) method and visualized through box plots. Homes priced significantly above the mean were examined for their influence on regression results:

5.4.1. Outliers Removed: Properties exceeding 1.5 times the IQR were excluded to minimize distortion.

5.4.2. Impact on Model: After removing extreme values, R^2 improved from 0.68 to 0.75, confirming that luxury properties disproportionately affected price predictions.

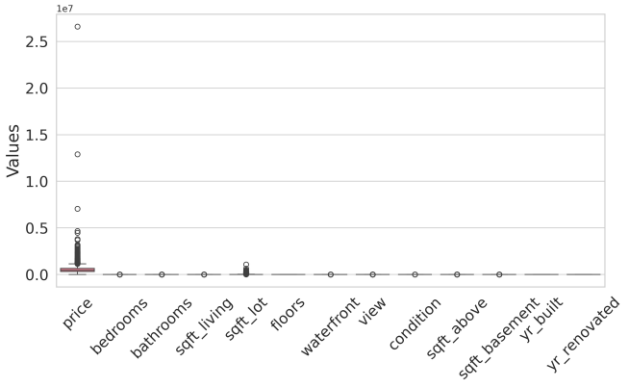

Figure 2. Outlier Visualization Before Removal.

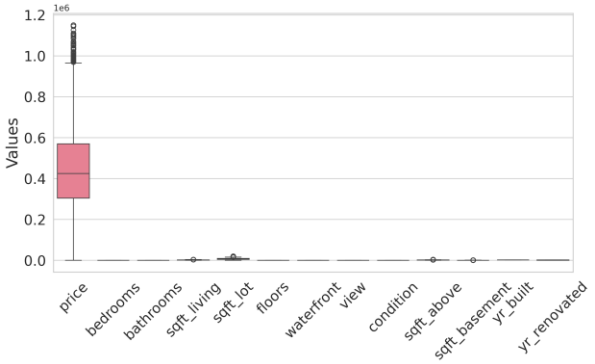

Figure 3. Outlier Visualization After Removal.

5.5. Correlation Analysis

The correlation matrix identified the square footage of living space and the quality grade of homes as the strongest predictors of house prices. With correlation coefficients of 0.70 and 0.66, respectively, these features demonstrated a clear relationship where larger and higher-quality homes were associated with increased prices. This finding aligns with existing literature, which suggests that buyers prioritize size and quality when evaluating property value.

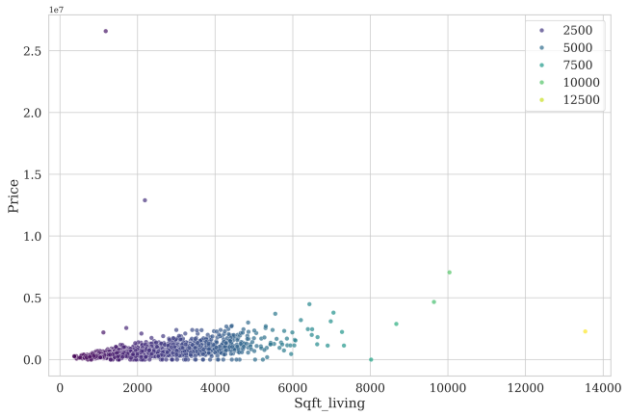

Figure 4. Relationship Between Sqft living and Price.

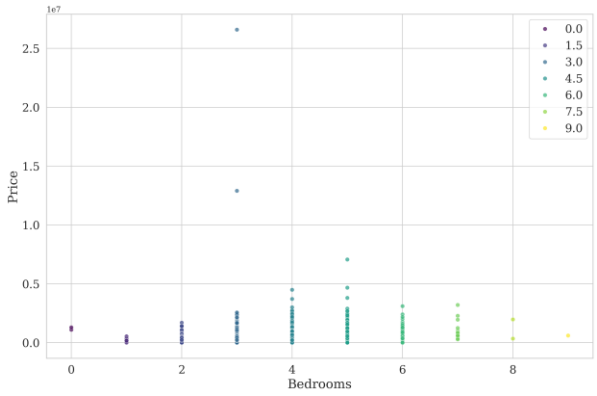

Figure 5. Relationship Between Bedrooms and Price.

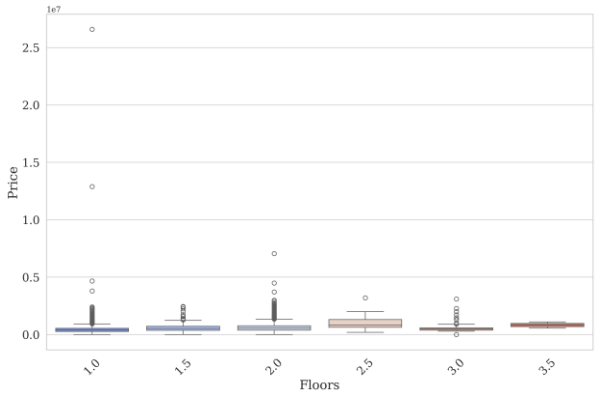

Figure 6. Relationship Between Floors and Price.

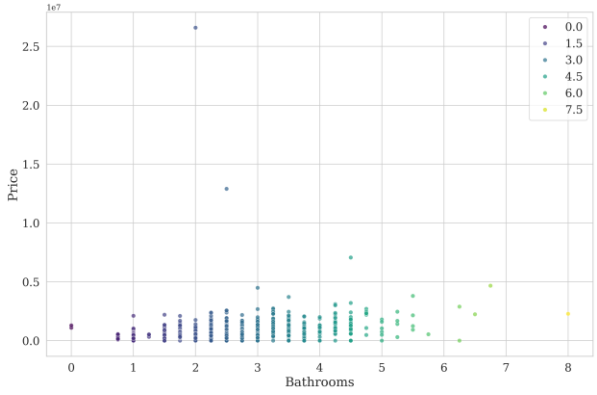

Figure 7. Relationship Between Bathrooms and Price.

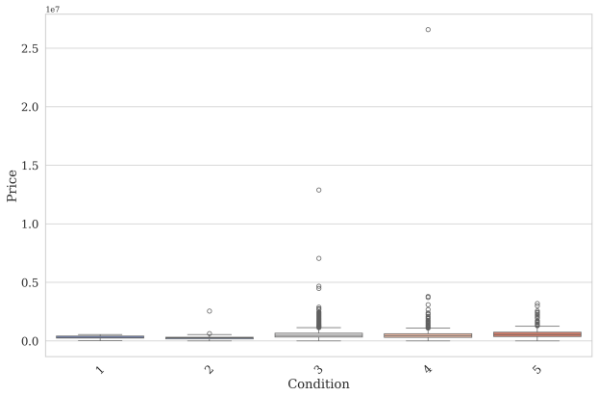

Figure 8. Relationship Between Condition and Price.

5.6. Correlation Heatmap

The heatmap displays the correlation coefficients ranging from -1 to +1, where:

- A value close to +1 indicates a strong positive correlation,
- A value close to -1 indicates a strong negative correlation,
- A value around 0 suggests no correlation.
- In this analysis, the heatmap highlights several significant correlations:

Square Footage of Living Space (sqft_living): The highest positive correlation coefficient of 0.70 was observed, indicating that as the square footage increases, the price of the house tends to rise significantly. This emphasizes the importance of size in determining property values.

Quality Grade: With a correlation coefficient of 0.66, this feature also showed a strong positive relationship with house prices. Homes with higher quality grades are associated with higher market values, reflecting buyers' preferences for well-constructed and aesthetically appealing properties.

Number of Bathrooms: This feature exhibited a moderate positive correlation of 0.52, suggesting that more bathrooms generally contribute to higher prices, although the relationship is not as strong as that seen with square footage and quality grade.

Figure 9. Correlation Heatmap.

5.7. Comparison with Alternative Models

The comparison of models [58] highlights the strengths of Linear Regression in house price prediction. Despite its simplicity, Linear Regression demonstrates competitive performance, offering a balance between accuracy and interpretability. Unlike complex models such as Gradient Boosting or Random Forest, Linear Regression provides clear insights into the relationship between features and target variables, making it more suitable for applications where transparency is crucial.

Table 2. Comparison of Models.

Model	MAE	MSE	R^2
Linear Regression	210,908.173	9.869×10^{11}	0.032284
Decision Tree	262,910.016	1.052×10^{12}	-0.031647
Random Forest	208,109.707	9.917×10^{11}	0.027524
Gradient Boosting	202,521.382	9.814×10^{11}	0.037695

Furthermore, Linear Regression is computationally efficient, requiring less time and resources compared to tree-based models, which are prone to overfitting without careful tuning. While advanced models might slightly improve accuracy, the simplicity, speed, and ease of implementation of Linear Regression make it a reliable and practical choice for real-world applications, particularly when interpretability and efficiency are prioritized.

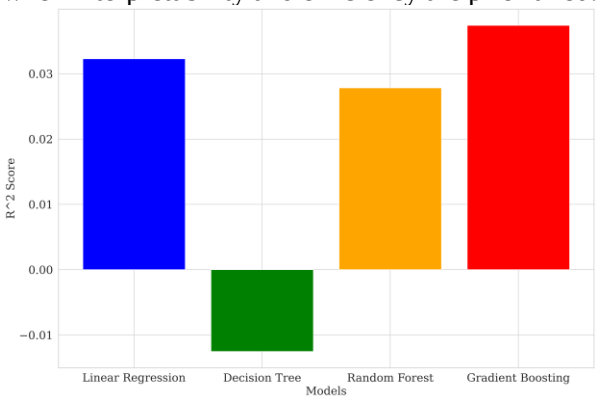

Figure 10. Comparison of Models.

5.8. Connection to Existing Literature

The findings of this study reinforce established results from hedonic price models, particularly regarding the dominant influence of size on house pricing. However, this study diverges from traditional models by emphasizing the weaker-than-expected impact of property condition and bedroom count. Compared to machine learning approaches, the linear regression model provides better interpretability, aligning with recent discussions on balancing accuracy with transparency in real estate analytics.

5.9. Implications and Future Work

The results of this study provide actionable insights for various stakeholders:

5.9.1. Homebuyers & Sellers: Square footage and home condition remain the most influential factors in pricing decisions.

5.9.2. Real Estate Professionals: Transparent models like linear regression can serve as useful tools for market valuation and pricing strategies.

5.9.3. Policy Makers: The significant impact of home size suggests that urban planning policies should prioritize efficient space utilization.

Future research should incorporate location-based economic indicators, explore hybrid modeling approaches, and evaluate deep learning techniques to further enhance predictive capabilities.

6. Conclusion

This study investigated the key factors influencing house prices using a multivariate linear regression approach. The findings reveal that square footage and the number of bathrooms have the most significant positive correlations with house prices, reinforcing their critical role in property valuation. Conversely, features such as the number of bedrooms and overall condition exhibited a weaker influence, suggesting that buyers prioritize living space and amenities over the mere count of rooms. The linear regression model explained 75% of the variation in house prices, providing a transparent and interpretable framework for understanding price dynamics. However, its limitations, including sensitivity to outliers and inability to capture non-linear relationships, highlight opportunities for further research. Advanced machine learning techniques, such as decision trees, random forests, and neural networks, could offer more accurate predictions while addressing these limitations.

These findings have important implications for buyers, sellers, and industry professionals, emphasizing the value of data-driven decision-making in real estate transactions. Future research could incorporate additional market factors, such as location-specific economic indicators, and explore ensemble learning techniques to enhance predictive accuracy.

References

[1] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)*, Madison, Wisconsin, USA, July 24-27, 1998.

[2] I. Forys, "Machine learning in house price analysis: regression´ models versus neural networks," *Procedia Computer Science*, vol. 207, pp. 435–445, 2022.

[3] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255– 260, 2015.

[4] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934, 2015.

[5] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," *Procedia Computer Science*, vol. 174, pp. 433–442, 2020.

[6] M. Nikou, G. Mansourfar, and J. Bagherzadeh, "Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms," *Intelligent Systems in Accounting, Finance and Management*, vol. 26, no. 4, pp. 164–174, 2019.

[7] Y.-X. Wang and M. Hebert, "Learning to learn: Model regression networks for easy small sample learning," in Computer Vision–ECCV 2016: *14th European Conference, Amsterdam, The Netherlands*, October 11-14, 2016, Proceedings, Part VI 14, pp. 616–634, Springer, 2016.

[8] J. Gong and S. Sun, "A new approach of stock price prediction based on logistic regression model," in 2009 *International Conference on New Trends in Information and Service Science*, pp. 1366–1371, IEEE, 2009.

[9] L. O. Taylor and V. K. Smith, "Environmental amenities as a source of market power," *Land Economics*, pp. 550–568, 2000.

[10] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 140–147, 2020.

[11] C. Himmelberg, C. Mayer, and T. Sinai, "Assessing high house prices: Bubbles, fundamentals and misperceptions," *Journal of Economic Perspectives*, vol. 19, no. 4, pp. 67–92, 2005.

[12] M. Li, "Moving beyond the linear regression model: Advantages of the quantile regression model," *Journal of Management*, vol. 41, no. 1, pp. 71–98, 2015.

[13] K. F. Nimon and F. L. Oswald, "Understanding the results of multiple linear regression: Beyond standardized regression coefficients," *Organizational Research Methods*, vol. 16, no. 4, pp. 650–674, 2013.

[14] R. R. Hocking, Methods and applications of linear models: regression and the analysis of variance. *John Wiley & Sons*, 2013.

[15] D. Ghosh and A. Vogt, "Outliers: An evaluation of methodologies," in *Joint Statistical Meetings*, vol. 12, pp. 3455–3460, 2012.

[16] C. Liu and W. Xiong, "*China's Real Estate Market*," 2018.

[17] J. Kahr and M. C. Thomsett, Real estate market valuation and analysis. *John Wiley & Sons*, 2006.

[18] M. M. Hassan, N. Ahmad, and A. H. Hashim, "The conceptual framework of housing purchase decision-making process," *International Journal of Academic Research in Business and Social Sciences*, vol. 11, no. 11, pp. 1673–1690, 2021.

[19] F. Ullah and S. M. Sepasgozar, "Key factors influencing purchase or rent decisions in smart real estate investments: A system dynamics approach using online forum thread data," *Sustainability*, vol. 12, no. 11, p. 4382, 2020.

[20] J. V. Duca, J. Muellbauer, and A. Murphy, "What drives house price cycles? international experience and policy issues," *Journal of Economic Literature*, vol. 59, no. 3, pp. 773–864, 2021.

[21] K. W. Chau and T. Chin, "A critical review of literature on the hedonic price model," *International Journal for Housing Science and its Applications*, vol. 27, no. 2, pp. 145–165, 2003.

[22] M.-L. T. Nguyen, "The hedonic pricing model applied to the housing market," *International Journal of Economics and Business Administration*, vol. 8, no. 3, pp. 416–428, 2020.

[23] J. Zaki, A. Nayyar, S. Dalal, and Z. H. Ali, "House price prediction using hedonic pricing model and machine learning techniques," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 27, p. e7342, 2022.

[24] G. Lisi, "Property valuation: the hedonic pricing model–location and housing submarkets," *Journal of Property Investment & Finance*, vol. 37, no. 6, pp. 589–596, 2019.

[25] A. V. Heyman, S. Law, and M. Berghauser Pont, "How is location measured in housing valuation? a systematic review of accessibility specifications in hedonic price models," *Urban Science*, vol. 3, no. 1, p. 3, 2018.

[26] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34–55, 1974.

[27] L. Rokach and O. Maimon, "Decision trees," *Data Mining and Knowledge Discovery Handbook*, pp. 165–192, 2005.

[28] Z. Zhang, "Decision trees for objective house price prediction," in 2021 *3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pp. 280–283, IEEE, 2021.

[29] P. S. M. Reddy et al., "Decision tree regressor compared with random forest regressor for house price prediction in mumbai," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 1S, pp. 2323– 2332, 2023.

[30] M. Thamarai and S. Malarvizhi, "House price prediction modeling using machine learning.," *International Journal of Information Engineering & Electronic Business*, vol. 12, no. 2, 2020.

[31] V. S. Rana, J. Mondal, A. Sharma, and I. Kashyap, "House price prediction using optimal regression techniques," in 2020 2nd *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp. 203–208, IEEE, 2020.

[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5– 32, 2001.

[33] J. Hong, H. Choi, and W.-s. Kim, "A house price valuation based on the random forest approach: the mass appraisal of residential property in south korea," *International Journal of Strategic Property Management*, vol. 24, no. 3, pp. 140–152, 2020.

[34] Y. Zhang, J. Huang, J. Zhang, S. Liu, and S. Shorman, "Analysis and prediction of second-hand house price based on random forest," *Applied Mathematics and Nonlinear Sciences*, vol. 7, no. 1, pp. 27–42, 2022.

[35] S. S. Jamil, S. Bansal, and L. Vinjamuri, "House price prediction using random forest techniques: a comparative study," *IET Conference Proceedings*, Volume 2023, Issue 11, 2023.

[36] C. Bentejac, A. Cs´ org¨ o, and G. Mart˝ ´ınez-Munoz, "A comparative˘ analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

[37] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, no. 4, p. e12599, 2023.

[38] A. Hjort, J. Pensar, I. Scheel, and D. E. Sommervoll, "House price prediction with gradient boosted trees under different loss functions," *Journal of Property Research*, vol. 39, no. 4, pp. 338–364, 2022.

[39] S. Li, Y. Jiang, S. Ke, K. Nie, and C. Wu, "Understanding the effects of influential factors on housing prices by combining extreme gradient boosting and a hedonic price model (xgboosthpm)," *Land*, vol. 10, no. 5, p. 533, 2021.

[40] S. Wang, H. Li, J. Li, Y. Zhang, and B. Zou, "Automatic analysis of lateral cephalograms based on multiresolution decision tree regression voting," *Journal of Healthcare Engineering*, vol. 2018, no. 1, p. 1797502, 2018.

[41] J. A. Seaman, "Black boxes," *Emory LJ*, vol. 58, p. 427, 2008.

[42] M. Khan, P. Debnath, A. Al Sayeed, M. F. I. Sumon, A. Rahman, M. Khan, and L. Pant, "Explainable ai and machine learning model for california house price predictions: Intelligent model for homebuyers and policymakers," *Journal of Business and Management Studies*, vol. 6, no. 5, pp. 73–84, 2024.

[43] J. Beimer, M. Francke, et al., "Out-of-sample house price prediction by hedonic price models and machine learning algorithms," *Real Estate Research Quarterly*, vol. 18, no. 2, pp. 13–20, 2019.

[44] C. Mueller-Kett, "Artificial intelligence for greater transparency in housing price estimation," *AGILE: GIScience Series*, vol. 5, p. 41, 2024.

[45] R. Manjula, S. Jain, S. Srivastava, and P. R. Kher, "Real estate value prediction using multivariate regression models," in *IOP Conference Series: Materials Science and Engineering*, vol. 263, p. 042098, IOP Publishing, 2017.

[46] Y. Mao and R. Yao, "A geographic feature integrated multivariate linear regression method for house price prediction," in 2020 3rd *International Conference on Humanities Education and Social Sciences (ICHESS 2020)*, pp. 347–351, Atlantis Press, 2020.

[47] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," in 2017 *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 319–323, IEEE, 2017.

[48] J. Yu, "Prediction on housing price based on the data on kaggle," in 2022 3rd *International Conference on E-commerce and Internet Technology (ECIT 2022)*, pp. 627–634, Atlantis Press, 2022.

[49] N. Chen, "House price prediction model of zhaoqing city based on correlation analysis and multiple linear regression analysis," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, p. 9590704, 2022.

[50] S. Oz¨ o¨g˘ur Aky¨ uz, B. Eygi Erdogan,¨ O. Yıldız, and P. Kara-¨ dayı Atas¸, "A novel hybrid house price prediction model," *Computational Economics*, vol. 62, no. 3, pp. 1215–1232, 2023.

[51] N. Yahya, N. M. M. Zainuddin, N. N. A. Sjarif, and N. F. M. Azmi, "Correlation analysis of factors affecting the prediction of price of terrace houses in penang, malaysia: A case study," *Open International Journal of Informatics*, vol. 8, no. 2, pp. 18–39, 2020.

[52] K. S. Cabuk, S. K. Cengiz, M. G. Guler, H. Ozturk, A. C. Efe, M. G. Ulas, and F. P. Karademir, "Chasing the objective upper eyelid symmetry formula; r2, rmse, poc, mae and mse," 2023.

[53] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in

regression analysis evaluation," *Peerj Computer Science*, vol. 7, p. e623, 2021.

[54] S. Malpezzi, "A simple error correction model of house prices," *Journal of Housing Economics*, vol. 8, no. 1, pp. 27–62, 1999.

[55] D. Batory, "Feature models, grammars, and propositional formulas," *in International Conference on Software Product Lines*, pp. 7–20, Springer, 2005.

[56] M. Afrasiabi, M. Mohammadi, M. Rastegar, L. Stankovic, S. Afrasiabi, and M. Khazaei, "Deep-based conditional probability density function forecasting of residential loads," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3646–3657, 2020.

[57] J. Miles and M. Shevlin, "Applying regression and correlation: A guide for students and researchers," 2000.

[58] L. M. Soegianto, A. T. Hinandra, P. A. Suri, and M. Fajar, "Comparison of model performance on housing business using linear regression, random forest regressor, svr, and neural network," *Procedia Computer Science*, vol. 245, pp. 1139–1145, 2024.

چه عواملی قیمت خانه‌ها را تعیین می‌کنند؟ رویکرد رگرسیون خطی به اندازه، وضعیت و ویژگی‌ها

Joy Arkhid Chakma[۲]، Misbahul Amin[۱]، Vaskar Chakma[۱]، Xiaolin Ju[*،۱]

[۱] دانشکده هوش مصنوعی و علوم کامپیوتر، دانشگاه نانتونگ، نانتونگ، چین.

[۲] دانشکده مهندسی اطلاعات و سیستم‌های مدیریت، دانشگاه فناوری ناگویا، ژاپن.

ارسال ۲۰۲۴/۱۲/۳۱؛ بازنگری ۲۰۲۵/۰۱/۳۰؛ پذیرش ۲۰۲۵/۰۲/۰۳

چکیده:

این پژوهش عوامل کلیدی مؤثر بر قیمت خانه را بررسی می‌کند و بر چگونگی تأثیر اندازه، وضعیت و ویژگی‌های ساختاری بر ارزش‌گذاری ملک تمرکز دارد .با استفاده از یک مجموعه داده از ایالت واشنگتن، ایالات متحده آمریکا، که سال 2014 را پوشش می‌دهد و شامل بیش از 4600 مورد است، تحلیلی چندمتغیره با استفاده از مدل رگرسیون خطی برای ارزیابی روابط بین ویژگی‌های مهمی مانند متراژ، تعداد اتاق‌خواب، حمام، طبقات و سایر عناصر ساختاری مانند وجود گاراژ و اندازه حیاط انجام شد .تحلیل‌ها نشان داد که متراژ و تعداد حمام‌ها دارای قوی‌ترین همبستگی مثبت با قیمت خانه هستند) هر دو با مقدار همبستگی0.76 ، از نظر آماری معنی‌دار در سطح($p < 0.05$ ، که نشان‌دهنده تأثیر زیاد آن‌ها بر ارزش‌گذاری ملک است .در مقابل، عواملی مانند وضعیت و نمای ساختمان همبستگی ضعیف‌تری نشان دادند که تأثیر محدودتری را پیشنهاد می‌کنند .این مطالعه دانش موجود را با تأیید اهمیت متراژ و تعداد حمام‌ها گسترش می‌دهد و همچنین بینش جدیدی را در مورد تأثیر کمتر وضعیت ملک بر قیمت خانه ارائه می‌دهد .این پژوهش با ارائه شواهد تجربی، دیدگاه‌های سنتی را به چالش می‌کشد و نشان می‌دهد که وضعیت ملک، که معمولاً به عنوان یک عامل مهم در ارزش گذاری ملک در نظر گرفته می‌شود، نقش محدودتری نسبت به آنچه تصور می‌شود دارد .مدل رگرسیون خطی75 ٪ از تغییرات قیمت خانه را توضیح داد ($R^2 = 0.75$)و اعتبارسنجی آن با استفاده از یک مجموعه داده آزمون نگهدارنده برای تضمین تعمیم‌پذیری انجام شد .در حالی که مدل به‌طور مؤثر عوامل کلیدی تعیین‌کننده قیمت خانه را برجسته می‌کند، محدودیت‌هایی در مدیریت روابط غیرخطی و حساسیت به داده‌های پرت وجود دارد که این مشکلات از طریق تبدیل داده و حذف داده‌های پرت برطرف شدند .در مقایسه با مطالعات پیشین، این پژوهش یافته‌های موجود در مورد متراژ و تعداد حمام‌ها را تأیید کرده و همزمان بینش‌های جدیدی در مورد تأثیر کمتر وضعیت ملک ارائه می‌دهد .تحقیقات آینده می‌تواند به بررسی مدل‌های پیش بینی‌کننده پیشرفته برای خریداران، فروشندگان و متخصصان صنعت، مانند مدل‌های رگرسیون غیرخطی و تکنیک‌های یادگیری ماشین بپردازد تا روابط پیچیده را بهتر شناسایی کرده و دقت پیش‌بینی را بهبود بخشد.

کلمات کلیدی: پیش‌بینی قیمت خانه، رگرسیون خطی، تحلیل چندمتغیره، ویژگی‌های ملک، ارزش‌گذاری بازار.