A Comparative Study of House Price Prediction Using Linear Regression and Random Forest Models

Yahan Fu*

Loudoun school for advanced studies, Ashburn, Virginia, 20147, USA

*Corresponding author: yahanfu@stu.sqxy.edu.cn

Abstract. The accurate prediction of house prices is crucial for stakeholders in the real estate sector, financial institutions, and urban planners. It not only informs investment decisions but also aids in policy formulation and market analysis. This research paper delves into the comparison of two prominent predictive analytics techniques-Linear Regression and Random Forest—to ascertain their effectiveness in forecasting house prices. Using a Kaggle dataset, this study analyzes key house price predictors such as building classification, living area size, construction year, and land area. The analysis shows Random Forest outperforms Linear Regression in accuracy, emphasizing the importance of building classification and living area in price prediction. Detailed visualizations, like feature importance graphs and scatter plots, offer clear insights into model performance. This research contributes significantly to real estate predictive analytics, offering insights to guide investment strategies and policy-making. It also opens avenues for exploring alternative machine learning approaches and socio-economic factors for a more comprehensive understanding of housing market dynamics.

Keywords: House price prediction; linear regression; random forest; real estate analytics.

1. Introduction

The realm of house price prediction has garnered increasing attention due to its profound implications for various sectors, including real estate, finance, and urban planning. Understanding housing market dynamics is crucial for investors, policymakers, and homeowners seeking profitable opportunities and effective strategies. Traditional methods of house price prediction often relied on simplistic approaches or expert judgment, lacking the sophistication required to capture the multifaceted nature of housing market dynamics. The advent of advanced techniques like Linear Regression and Random Forest has given researchers powerful tools for analyzing large datasets and extracting insights [1].

House price prediction occupies a paramount position at the confluence of probabilities and statistics, wielding considerable influence over real estate investments, market analyses, and economic projection is pursuit, Linear Regression and Random Forest emerge as two stalwart methodologies for modeling the dynamics of house prices [2]. Linear Regression, deeply rooted in classical statistical methodologies, operates on the principle of establishing a linear relationship between a dependent variable, such as house prices, and a set of independent variables, encompassing factors like square footage, location, and number of bedrooms [3]. Through regression analysis, Linear Regression endeavors to estimate coefficients that best fit the observed data points, thereby capturing the underlying linear trends inherent in the dataset. For instance, in the context of house price prediction, a linear regression model might reveal that for every additional square foot of living space, the price of a house tends to increase by a certain fixed amount [4].

Conversely, Random Forest, an ensemble learning method, uses multiple decision trees to provide robust predictions. During training, Random Forest constructs an ensemble of decision trees, with each tree offering its unique perspective on the data based on a subset of features and samples. These decision trees are then aggregated to produce a final prediction, typically by averaging or voting across the ensemble. The strength of Random Forest lies in its ability to handle complex, nonlinear relationships between variables [5]. For example, in the context of house price prediction, Random Forest can adeptly capture interactions between variables that may not follow a simple linear pattern, such as the combined influence of neighborhood characteristics, school district ratings, and proximity to amenities on house prices [6].

The significance of this comparative study lies in its quest to unravel the relative efficacy of Linear Regression and Random Forest models in the context of house price prediction [7]. By subjecting these models to rigorous evaluation on historical housing datasets and scrutinizing their performance using established metrics such as Mean Squared Error, R-squared value, and accuracy rates, the author endeavored to delineate the nuanced nuances of their predictive capabilities [8]. This research not only aids in elucidating the strengths and limitations inherent in each modeling approach but also empowers researchers and practitioners to make informed decisions regarding the selection of the most appropriate methodology for their specific house price prediction endeavors.

In general, this research serves as a beacon guiding the way forward in the realm of house price prediction, providing valuable insights into the comparative efficacy of Linear Regression and Random Forest models and paving the path for advancements in predictive analytics within the real estate domain.

2. Methods

2.1. Data Source

The data utilized in this study were sourced from Kaggle, a prominent platform for datasets and machine learning competitions. Specifically, the dataset used for this analysis comprises information on house prices and various attributes related to residential properties. Each observation in the dataset represents a unique property and includes details such as the date of sale, price, number of rooms, number of bathrooms, square footage, and other relevant features. These attributes provide valuable insights into the factors influencing house prices, allowing for a comprehensive comparative study of predictive models such as linear regression and random forest [9].

2.2. Variable Selection

The selection of variables is a critical step to ensure the robustness and accuracy of the linear regression model. A set of key indicators, commonly employed in real estate business analysis, was selected to investigate their relationship with housing prices. These indicators encompass various aspects of residential properties, including the number of bedrooms, bathrooms, floors, waterfront presence, scenic views, overall condition, and grade rating. Each of these variables contributes to the overall valuation and desirability of a property, making them essential factors to consider when analyzing housing market trends. The housing price, denoted as 'y', represents the target variable, indicating the price of the house under consideration. Table 1 provides a detailed overview of each selected variable along with their respective logograms and meanings:

Table 1. Variables Selected for House Price Prediction

Variable	Logogram	Meaning
price	y	Price of the house
bedrooms	X1	The number of bedrooms
bathrooms	X2	The number of bathrooms
sqft_living	X3	Square footage of the living space
sqft_lot	X4	Square footage of the lot
floors	X5	The number of floors
waterfront	X6	The number of waterfronts
view	X7	Scenes you can see through a window
condition	X8	Usage level
grade	X9	Rate of the house
sqft_above	X10	Square footage of the space above ground
sqft_basement	X11	Square footage of the basement
yr_built	X12	The year the house was built
yr_renovated	X13	The year the house was renovated
zipcode	X14	The ZIP code area
lat	X15	Latitude
long	X16	Longitude
sqft_living15	X17	Square footage of living space in 2015
sqft_lot15	X18	Square footage of the lot in 2015

2.3. Method Introduction

In this study, the author employed two distinct methodologies for house price prediction: Linear Regression and Random Forest. Linear Regression is a classical statistical technique that establishes a linear relationship between a dependent variable (house prices) and a set of independent variables (e.g., square footage, location attributes). Linear Regression aims to capture linear trends within the data by estimating coefficients through regression analysis. This method is particularly suitable for scenarios where straight lines can approximate relationships and offers simplicity and interpretability.

Random Forest represents an ensemble learning technique that harnesses the collective intelligence of multiple decision trees. Each decision tree within the Random Forest provides its unique perspective on the data, and predictions are aggregated to yield a final output. Random Forest excels in capturing complex, nonlinear relationships between variables, making it suitable for modeling intricate and multifaceted scenarios such as house price prediction. It offers resilience to overfitting and versatility in handling diverse datasets [10].

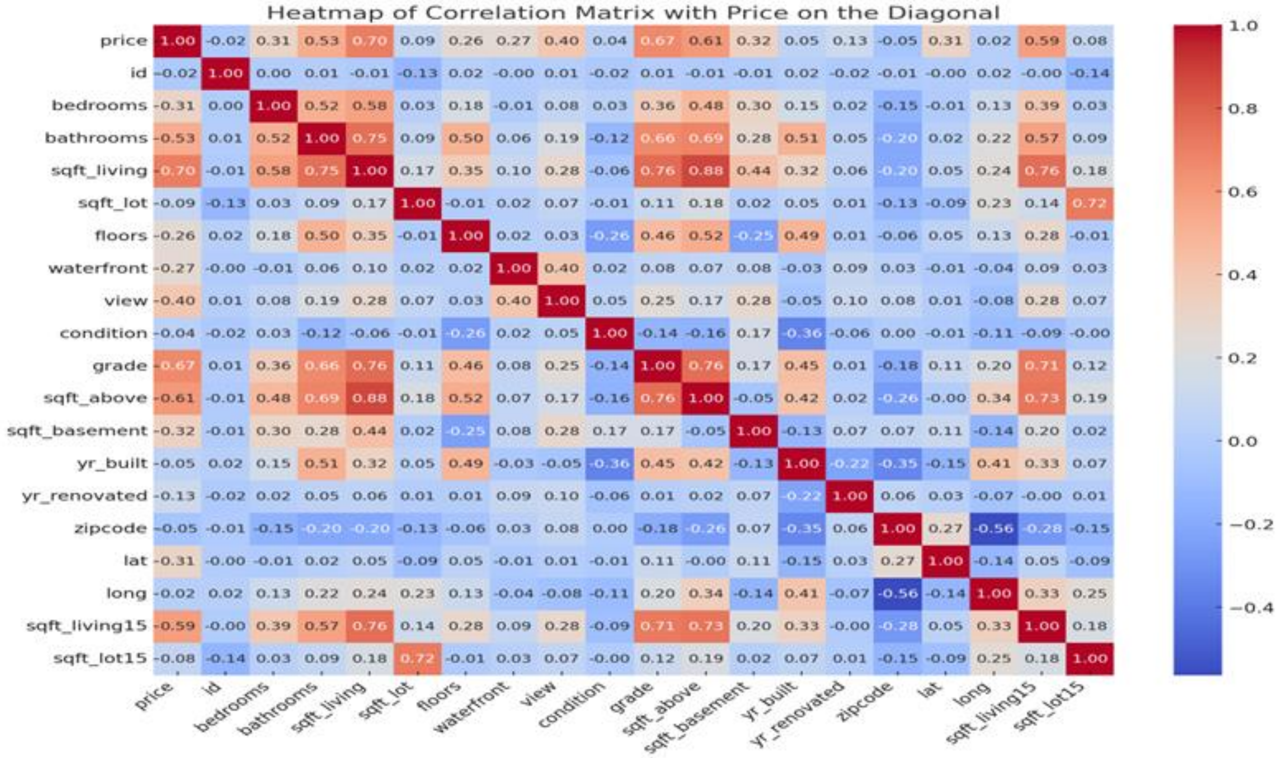
By comparing the performance of Linear Regression and Random Forest models using established metrics, this paper aims to elucidate their relative efficacy in house price prediction. This comparative analysis facilitates informed decision-making for researchers and practitioners involved in real estate analytics, guiding the selection of appropriate methodologies for specific prediction tasks.

3. Results and Discussion

3.1. Correlation Analysis

Firstly, the author commenced the analysis by examining a heatmap of the correlations among multiple variables in relation to the target variable, 'price'. Figure 1 is the Heatmap of Correlation

Matrix With Price on the Diagonal facilitated the identification of variables that exhibit significant linear relationships with 'price', either positively or negatively. Based on the correlation heatmap, variables demonstrating a high degree of correlation with 'price' were preliminarily selected for further analysis [3]. This approach allowed people to narrow down potential predictors from the dataset, prioritizing those with substantial relevance to house pricing. Subsequently, these selected variables were subject to a detailed linear regression analysis to quantitatively assess their impact on house prices. This methodology ensures that the predictive model is grounded on variables that are statistically significant and have practical implications for house price determinants, thereby enhancing the model's explanatory power and predictive accuracy.



	price	id	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
price	1.00	-0.02	0.31	0.53	0.70	0.09	0.26	0.27	0.40	0.04	0.67	0.61	0.32	0.05	0.13	-0.05	0.31	0.02	0.59	0.08
id	-0.02	1.00	0.00	0.01	-0.01	-0.13	0.02	-0.00	0.01	-0.02	0.01	-0.01	-0.01	0.02	-0.02	-0.01	-0.00	0.02	-0.00	-0.14
bedrooms	0.31	0.00	1.00	0.52	0.58	0.03	0.18	-0.01	0.08	0.03	0.36	0.48	0.30	0.15	0.02	-0.15	-0.01	0.13	0.39	0.03
bathrooms	0.53	0.01	0.52	1.00	0.75	0.09	0.50	0.06	0.19	-0.12	0.66	0.69	0.28	0.51	0.05	-0.20	0.02	0.22	0.57	0.09
sqft_living	0.70	-0.01	0.58	0.75	1.00	0.17	0.35	0.10	0.28	-0.06	0.76	0.88	0.44	0.32	0.06	-0.20	0.05	0.24	0.76	0.18
sqft_lot	0.09	-0.13	0.03	0.09	0.17	1.00	-0.01	0.02	0.07	-0.01	0.11	0.18	0.02	0.05	0.01	-0.13	-0.09	0.23	0.14	0.72
floors	0.26	0.02	0.18	0.50	0.35	-0.01	1.00	0.02	0.03	-0.26	0.46	0.52	-0.25	0.49	0.01	-0.06	0.05	0.13	0.28	-0.01
waterfront	0.27	-0.00	-0.01	0.06	0.10	0.02	0.02	1.00	0.40	0.02	0.08	0.07	0.08	-0.03	0.09	0.03	-0.01	-0.04	0.09	0.03
view	0.40	0.01	0.08	0.19	0.28	0.07	0.03	0.40	1.00	0.05	0.25	0.17	0.28	-0.05	0.10	0.08	0.01	-0.08	0.28	0.07
condition	0.04	-0.02	0.03	-0.12	-0.06	-0.01	-0.26	0.02	0.05	1.00	-0.14	-0.16	0.17	-0.36	-0.06	0.00	-0.01	-0.11	-0.09	-0.00
grade	0.67	0.01	0.36	0.66	0.76	0.11	0.46	0.08	0.25	-0.14	1.00	0.76	0.17	0.45	0.01	-0.18	0.11	0.20	0.71	0.12
sqft_above	0.61	-0.01	0.48	0.69	0.88	0.18	0.52	0.07	0.17	-0.16	0.76	1.00	-0.05	0.42	0.02	-0.26	-0.00	0.34	0.73	0.19
sqft_basement	0.32	-0.01	0.30	0.28	0.44	0.02	-0.25	0.08	0.28	0.17	0.17	-0.05	1.00	-0.13	0.07	0.07	0.11	-0.14	0.20	0.02
yr_built	0.05	0.02	0.15	0.51	0.32	0.05	0.49	-0.03	-0.05	-0.36	0.45	0.42	-0.13	1.00	-0.22	-0.35	-0.15	0.41	0.33	0.07
yr_renovated	0.13	-0.02	0.02	0.05	0.06	0.01	0.01	0.09	0.10	-0.06	0.01	0.02	0.07	-0.22	1.00	0.06	0.03	-0.07	-0.00	0.01
zipcode	-0.05	-0.01	-0.15	-0.20	-0.20	-0.13	-0.06	0.03	0.08	0.00	-0.18	-0.26	0.07	-0.35	0.06	1.00	0.27	-0.56	-0.28	-0.15
lat	0.31	-0.00	-0.01	0.02	0.05	-0.09	0.05	-0.01	0.01	-0.01	0.11	-0.00	0.11	-0.15	0.03	0.27	1.00	-0.14	0.05	-0.09
long	0.02	0.02	0.13	0.22	0.24	0.23	0.13	-0.04	-0.08	-0.11	0.20	0.34	-0.14	0.41	-0.07	-0.56	-0.14	1.00	0.33	0.25
sqft_living15	0.59	-0.00	0.39	0.57	0.76	0.14	0.28	0.09	0.28	-0.09	0.71	0.73	0.20	0.33	-0.00	-0.28	0.05	0.33	1.00	0.18
sqft_lot15	0.08	-0.14	0.03	0.09	0.18	0.72	-0.01	0.03	0.07	-0.00	0.12	0.19	0.02	0.07	0.01	-0.15	-0.09	0.25	0.18	1.00

Fig. 1 Heatmap of Correlation Matrix With Price on the Diagonal

3.2. Data Preprocessing

In this academic exploration of housing price determinants, this paper embarked on constructing a predictive model leveraging the Linear Regression algorithm, a cornerstone technique in statistical modeling. This endeavor necessitated a meticulously curated dataset, from which this paper sourced 'kc_house_data.csv', an expansive compilation of housing transaction records.

The initial phase of the analytical process involved the careful preparation of the dataset to align with the objectives of the study. Given the plethora of variables within the dataset, it was paramount to judiciously select those variables that were most pertinent to the analysis while excluding identifiers and temporal markers ('id', 'date') as well as 'zipcode', due to their nominal nature and potential to introduce noise into the predictive modeling process. The target variable for the model was designated as 'price', reflective of the housing transaction values this paper aimed to predict. The remaining variables, encapsulating a wide range of house attributes, were treated as independent predictors in the model.

With the dataset aptly prepared, this paper proceeded to the model training stage, employing the Linear Regression model from the sklearn.linear_model library-a choice motivated by the model's simplicity and interpretability. This stage involved fitting the model to the training data, which consisted of the selected predictors and the target variable. This process facilitates the model's learning of the underlying relationships between the house attributes (predictors) and their corresponding prices.

Upon the successful training of the model, the author extracted the model coefficients and intercept respectively. These coefficients represent the magnitude and direction of the relationship between each independent variable and the target variable 'price'. The intercept, on the other hand, signifies the expected value of 'price' when all predictors are held at zero.

3.3. Model Results

Finaly, this paper got the parameters for the regression equation as the following Table 2 Linear Regression Equation Parameters:

Table 2. Linear Regression Equation Parameters

Parameters	Value	P value
Interception	-36,862,616.47	0.023
bedrooms	-34151.66911	P<0.001
bathrooms	42161.70750	P<0.001
sqft_living	108.72297	P<0.001
Sqft_lot	0.12742	0.007
floors	760.68806	0.063
waterfront	587847.21524	P<0.001
view	49429.22770	P<0.001
condition	31031.80174	P<0.001
grade	97219.36921	P<0.001
Sqft_above	70.79085	P<0.001
Sqft_basement	37.93212	P<0.001
Yr_built	-2456.27635	P<0.001
Yr-renovated	21.53244	P<0.001
Lat	561060.48229	P<0.001
Long	-117020.85869	P<0.001
Sqft_living15	27.43120	P<0.001
Sqft_lot15	-0.39329	P<0.001

This equation encapsulates the predictive model this paper devised to forecast housing prices based on a variety of house attributes, thus providing a quantitative foundation for the subsequent analysis and discussions (Table 3).

Table 3. Predictive Performance of the Two Models

Models	RMSE	R^2
Linear Regression	214,472.76	0.70
Random Forest Model	148,428.13	0.85

Figure 2 shows how much each feature in the random forest model influences the prediction of house prices. This paper can see that `grade` (building class) and `sqft_living` (living area) are the two most important features, which means that they play a decisive role in predicting house prices. Other characteristics such as `yr_built` (year of construction) and `sqft_lot` (land area) also have some influence on house price prediction, but to a lesser extent. The results of these analyses provide an intuitive understanding of model performance and the importance of data features, which helps people to better evaluate and interpret model predictions.

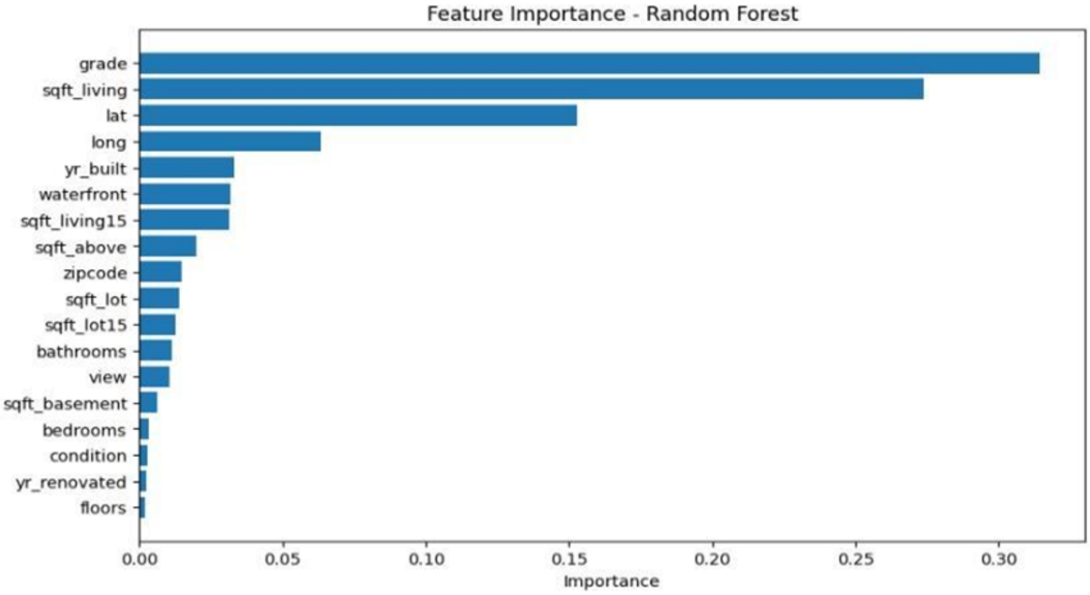


Fig. 2 Feature Importance-Random Forest

The scatter plots Figure 3 shows the comparison between the predicted results of the linear regression and random forest models and the actual house prices. In the chart, the ideal prediction result should be as close as possible to the black line (i.e., the line where the predicted value is equal to the actual value). The Random Forest model has a tighter distribution of points, indicating that its predictions are more consistent with the actual values.

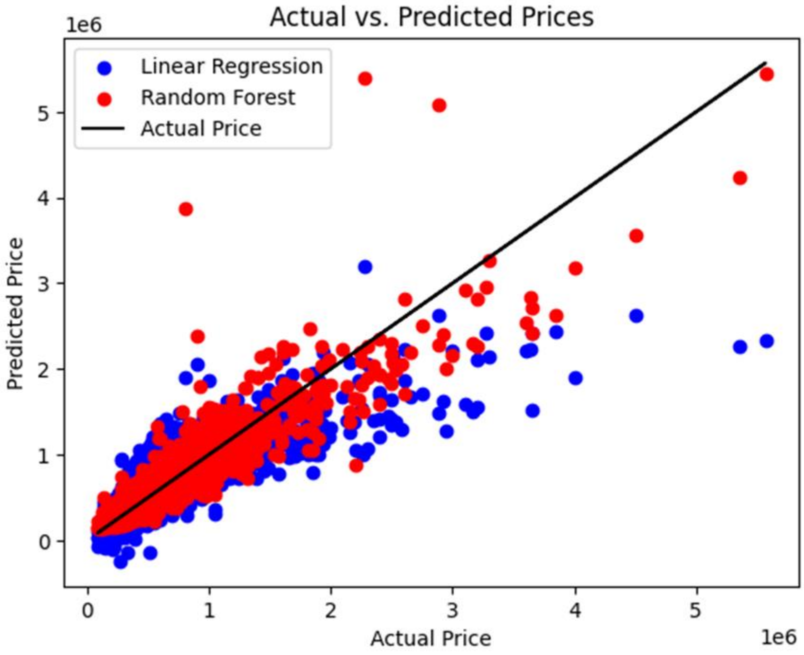


Fig. 3 Actual vs Predicted Prices

The analysis of feature importance conducted on the random forest model provides valuable insights into the factors driving house price predictions. Figure 2 illustrates the significance of various features in influencing the model's predictions. Notably, the 'grade' (representing the building class) and 'sqft_living' (indicating living area) emerge as the most influential features. This suggests that the quality and size of the property are crucial determinants of its price, aligning with common intuition in real estate valuation. Additionally, 'yr_built' (year of construction) and 'sqft_lot' (land area) exhibit some degree of influence on price prediction, albeit to a lesser extent. These findings highlight the importance of considering multiple aspects of a property when estimating its market value.

Furthermore, the scatter plots presented in Figure 3 offer a visual comparison between the predicted house prices generated by the linear regression and random forest models against the actual prices. Ideally, the predicted values should align closely with the black reference line, indicating accurate predictions. The tighter distribution of points around this line in the Random Forest model suggests a higher level of consistency between predicted and actual prices compared to the linear regression model. This improved performance underscores the effectiveness of the random forest approach in capturing complex relationships within the data and making more precise predictions.

Overall, these results provide valuable insights into the performance and interpretability of the models employed in predicting house prices. By understanding the relative importance of different features and comparing predictive accuracy, stakeholders can make more informed decisions in the real estate market, enhancing efficiency and confidence in property valuation processes.

3.4. Discussion

The comparison between linear regression and random forest models in prediction house prices reveals significant differences in their predictive performance. As highlighted in the results section, the Random Forest model demonstrates superior accuracy with a lower RMSE of and a higher R^2 value. This distinction underscores the Random Forest model's enhanced capability in capturing the complexity of the housing market, attributing to its ability to handle nonlinear relationships and interactions between features more effectively than linear models.

The analysis of model predictions against actual house prices, illustrated in Figure 2, further supports the superiority of the Random Forest model. The closer alignment of the predicted values to the actual prices, as demonstrated by the tighter distribution around the ideal prediction line, evidences the Random Forest model's consistency and reliability in prediction. This observation is critical for stakeholders in the real estate market, where prediction accuracy directly influences investment and decision-making processes.

Furthermore, the examination of feature importance in the Random Forest model provides valuable insights into factors driving house prices. The prominence of grade and sqft_living as key determinants of price highlights the significant impact of property quality and size on market value. Although other features such as yr_built and sqft_lot also contribute to the prediction model, their influence is comparatively minor, suggesting that buyers prioritize the quality and functional living space of properties.

4. Conclusion

In conclusion, this study delved into house price prediction, leveraging the power of Linear Regression and Random Forest methodologies. The comparative analysis aimed to elucidate the relative efficacy of these approaches in predicting housing prices, offering valuable insights for stakeholders in the real estate domain. Through feature importance analysis, the author identified key determinants of house prices, with factors such as building class ('grade') and living area ('sqft_living') emerging as pivotal influencers. These findings underscore the significance of property quality and size in shaping market valuations. Additionally, the scatter plots provided a visual representation of model predictions, highlighting the superior performance of the Random Forest model in capturing the nuances of housing market dynamics.

The study's significance lies in its contribution to advancing predictive analytics within the real estate sector. By comparing Linear Regression and Random Forest models, researchers and practitioners can make informed decisions regarding selecting appropriate methodologies for house price prediction tasks. Furthermore, the insights gleaned from this research can inform investment strategies, policy formulations, and decision-making processes in the real estate market.

Moving forward, future research could explore additional machine learning techniques, integrate alternative datasets, or investigate the impact of socio-economic factors on housing market dynamics. By continuing to refine predictive models and expand analytical frameworks, the author can enhance the understanding of housing market trends and empower stakeholders with actionable insights for navigating the complexities of real estate investment and management. Ultimately, this research serves as a stepping stone towards more accurate and reliable house price prediction, facilitating informed decision-making and driving innovation within the real estate industry.

References

[1] Wang Yige. House-Price Prediction Based on OLS Linear Regression and Random Forest. 2021 2nd Asia Service Sciences and Software Engineering Conference, 2021.

[2] Kokot Sebastian, Sebastian Gnat. Simulative Verification of the Possibility of Using Multiple Regression Models for Real Estate Appraisal. Real Estate Management and Valuation, 2019, 109-123.

[3] Levantesi S, Piscopo G. The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach. Risks, 2020, 8(4): 112.

[4] Priyatno Arif Mudi, et al. Comparison Random Forest Regression and Linear Regression For prediction BBCA Stock Price. Journal Teknik Industry Terintegrasi, 2023.

[5] Hong Jengei. A Mass Appraisal Model on Residential Property with Random Forest Algorithm. Journal of Real Estate Analysis, 2021, 1-28.

[6] Alfaro Navarro, José Luis, et al. A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems. Complexity, 2020, 112.

[7] Yilmazer Seckin, Sultan Kocaman. A Mass Appraisal Assessment Study Using Machine Learning Based on Multiple Regression and Random Forest. Land Use Policy, 2020.

[8] Zhao Lili, Jiao Jiwen. Grey correlation analysis of factors affecting housing prices. Statistics and Decision Making, 2007, 23: 2.

[9] Luo Yubo. Analysis of influencing factors on housing prices: quantile regression method. Statistics and Decision Making, 2011, 6: 2.

[10] Guo Bin, Wang Ying. A Study on the Factors Influencing Urban Housing Prices Based on Dynamic Econometric Models. Business Research, 2010, 4.