Predicting US Airbnb Listing Prices by Machine Learning Models

Yuchen Yang*

School of Economics, University of Edinburgh, United Kingdom

*Corresponding author: Y.Yang-218@sms.ed.ac.uk

Abstract. This paper addresses the prediction of Airbnb property prices using 2023 open data through the application of machine learning methodologies. In the context of the flourishing sharing economy, accurate price prediction within the short-term rental market holds great significance for hosts and users alike. Drawing on the 2023 Airbnb open dataset, the study employs three distinct models – Linear Regression, Random Forest, and XGBoost. Rigorous training, testing, and evaluation of these models reveal insights into their predictive capabilities. The focus centers on assessing model fit using essential evaluation metrics including R-squared, Mean Squared Error, and Root Mean Squared Error. Results demonstrate that the XGBoost model outperforms both Linear Regression and Random Forest. After parameter tuning, the best parameter for XGBoost regressor exhibits the lowest prediction error and highest R-squared value, showcasing its ability to capture intricate patterns within the data. This outcome underscores the potency of advanced ensemble learning techniques for precise property price predictions. The study's implications are substantial, offering hosts and potential guests improved decision-making insights.

Keywords: Airbnb, XGBoost, sharing economy.

1. Introduction

The emergence of the sharing economy has revolutionized various industries, including the hospitality industry. Over the past decade, Airbnb has rapidly evolved from a small online bed and breakfast product to a leading point-to-point hotel giant with a presence in 80,000 cities around the world [1]. Unlike traditional hotels, Airbnb offers a variety of accommodation options to suit the unique preferences and needs of travelers. Among them, pricing is a key factor that can significantly influence the traveler's decision and the owner's revenue [2]. Determining the best price for an Airbnb listing is a complex task that takes into account a variety of factors, such as the property's location, amenities, and prevailing market rates. As a result, owners are often challenged to set a competitive and attractive price that also reflects the value of their space. To address this challenge, this study aims to develop robust predictive models to estimate Airbnb listing prices in the United States. By predicting Airbnb listing prices using Linear Regression, Random Forest, and XGBoost, it provides homeowners with the tools to make informed decisions in the area of pricing. At the same time, travelers will gain transparency and fairness in pricing, ensuring the accommodation they choose meets their expectations.

2. Data Processing

The success of any predictive model depends on the quality of the data it relies upon. Thus, this section will discuss the steps taken to process and prepare the dataset for modeling.

2.1. Data Understanding

During the initial exploration of the dataset, two critical aspects were examined: the presence of missing values (NaNs) and the uniqueness of values within each column. These insights play a pivotal role in understanding the dataset's quality and potential implications for subsequent analyses. The analysis of missing values per column indicates that the majority of columns either have no missing values or a relatively small number of them. Notably, the "neighbourhood group" column exhibits

approximately 58.43% NaNs, signifying a substantial proportion of missing data. This observation prompts a deeper investigation into the role of "neighbourhood group" in sub sequent analyses, as the extent of missing data could impact its significance. The examination of unique values for each column highlights interesting patterns. Columns such as "id," "name," and "host id" exhibit a high degree of variability, which is expected given their role as identifiers and names associated with properties and hosts. Conversely, columns like "neighbourhood group" and "room type" have a limited number of unique values, suggesting their potential use as categorical variables. For instance, the "room type" column with only four unique values implies a classification of accommodation types. Furthermore, the presence of 30 unique values in the "neighbourhood group" column and 1412 unique values in the "neighbourhood" column reflects the granularity of geographical seg mentation within the dataset. This geographic information holds significance in understanding property pricing dynamics, as different neighborhoods and groups may exhibit distinct pricing trends. The findings align with the expectations of an Airbnb dataset, which comprises diverse property listings across various neighborhoods and groups. To succinctly summarize the insights gained from this analysis, the information is presented in Table 1.

Table 1. Summary of Missing Values and Unique Values per Column

Column Name	NaN Count	% of NaNs	Unique Values
id	0	0.00%	232146
name	16	0.00%	220164
host_id	0	0.00%	119582
host_name	13	0.00%	29368
neighbourhood_group	135647	58.43%	30
neighbourhood	0	0.00%	1412
latitude	0	0.00%	157966
longitude	0	0.00%	159988
room_type	0	0.00%	4
price	0	0.00%	2429
minimum_nights	0	0.00%	178
number of reviews	0	0.00%	861
last_review	49085	21.14%	3147
reviews_per_month	49085	21.14%	1367
calculated_host_listing_count	0	0.00%	149
availability_365	0	0.00%	27
city	0	0.00%	271

These insights provide a foundational understanding of the dataset's characteristics, influencing decisions on addressing missing data, selecting relevant features, and devising optimal model development strategies.

2.2. Outlier Removal

Outliers within a dataset refer to data points that exhibit notable deviations from the typical or average data points. It is crucial to thoroughly examine and potentially exclude these unusual patterns from the dataset in order to enhance the predictive model's accuracy [3]. In the context of this project, it is possible that Airbnb listing prices may encompass such outliers, potentially affecting the overall performance of the predictive model. Visualization using distribution line graphs allowed us to observe the impact of outlier removal on various factors like price, minimum nights, calculated host listings count, reviews per month, number of reviews, and last review. This study initiated its analysis by focusing on the numerical columns in the dataset, excluding the "id" column due to its identifier role. Kernel density estimation (KDE) plots were employed to visualize the distribution of values within each numerical feature. These plots provided a clear view of potential outliers, showcasing

distinctive patterns that deviated from the bulk of the data. Notably, features like “price,” “minimum nights,”, “calculated host listings count,” ”reviews per month,” ”number of reviews,” and ”last review” exhibited distribution shapes hinting at the presence of outliers (See Fig. 1).

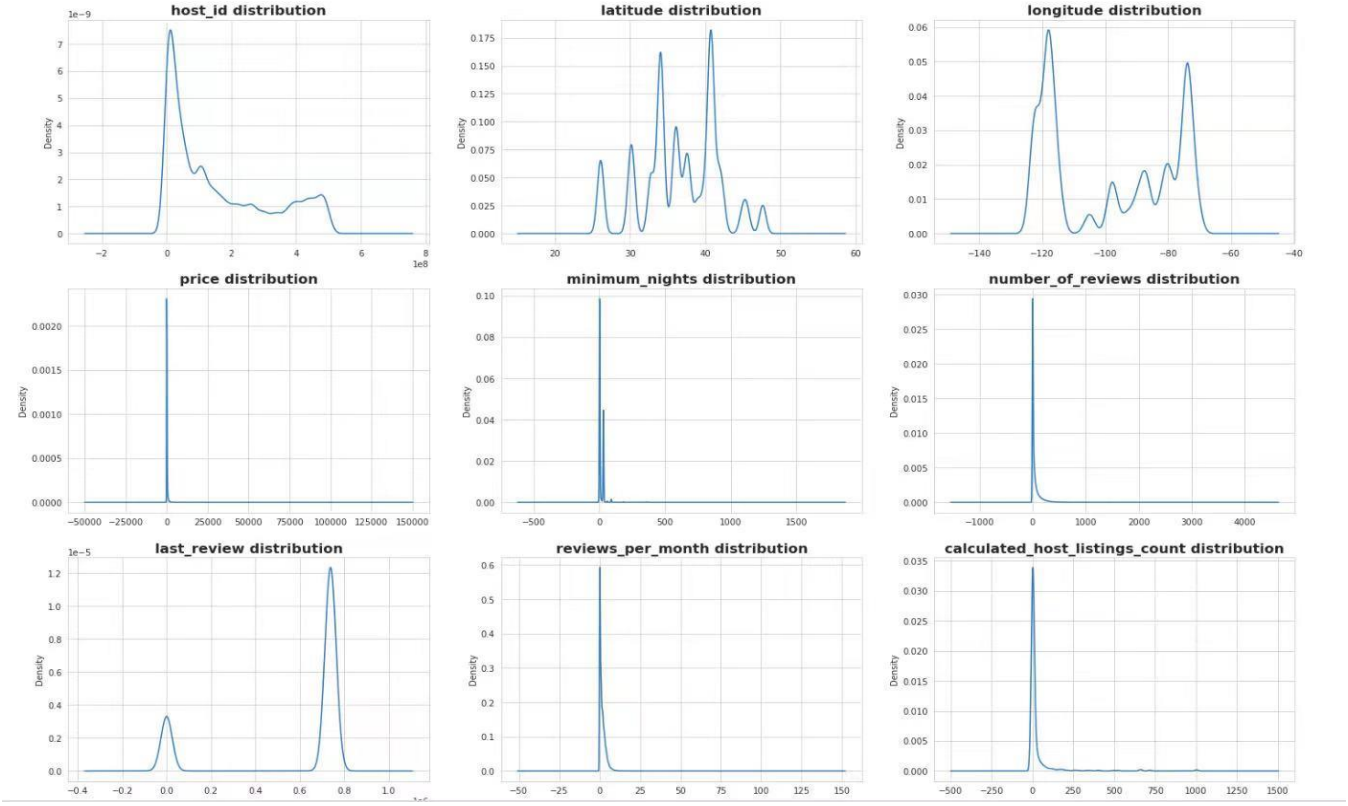

Fig. 1 Distribution of Factors Before Outlier Removal

To quantitatively address the outliers, the upper bounds are calculated for each feature’s values. These bounds, set at the 92nd percentile of the distribution, served as thresholds to identify and subsequently remove rows containing values beyond these limits. The impact of outlier removal was immediately evident in the KDE plots, which exhibited improved distributional characteristics (See Fig. 2).

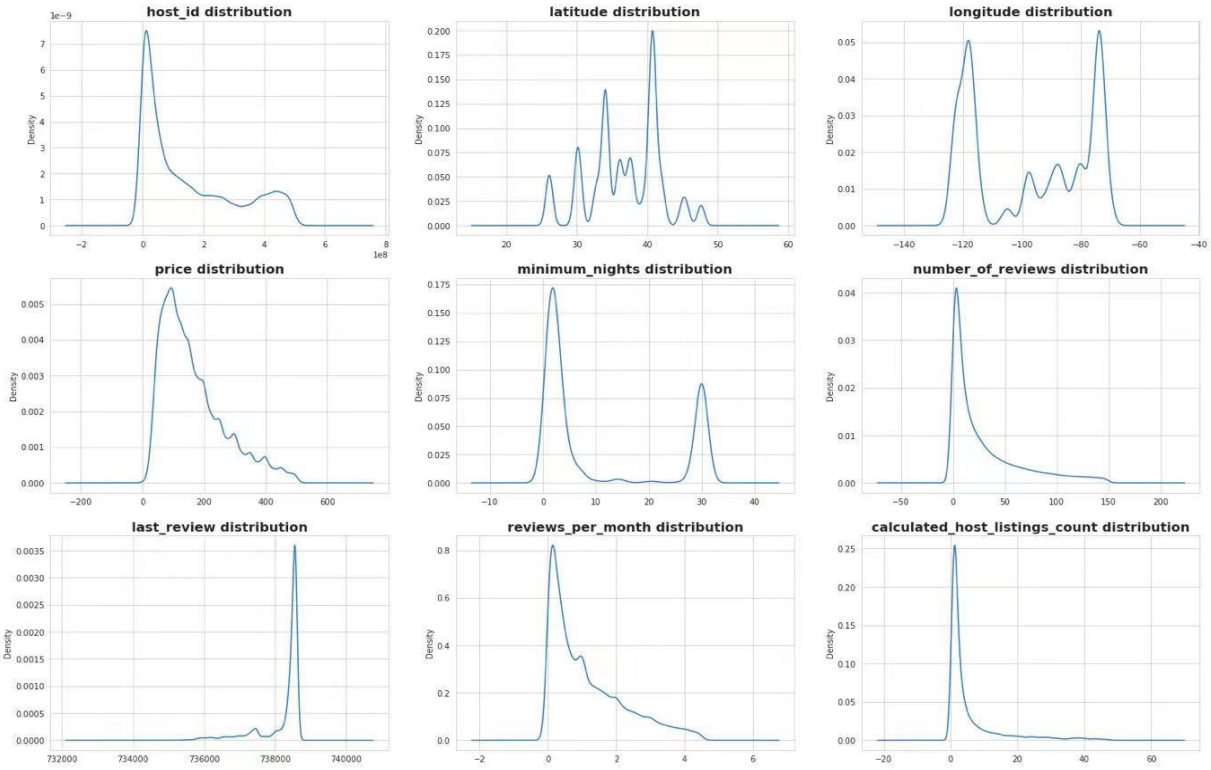

Fig. 2 Distribution of Factors After Outlier Removal

The elimination of outliers resulted in smoother, more interpretable distribution shapes that aligned with the expected patterns for each feature. This enhanced visualization demonstrated the effectiveness of the outlier removal process in restoring meaningful data representations.

In particular, the distributions of "number of reviews" and "reviews per month" features displayed exponential shapes after the removal of outliers. This observation potentially reflects the inherent nature of data in these categories. The comprehensive removal process fortified the dataset's quality, rendering it more suitable for robust predictive modeling.

By meticulously addressing outliers, the dataset has been refined for subsequent analyses, enhancing the integrity of this study's findings. This meticulous approach ensures the reliability of the predictive models and contributes to the accuracy of the conclusions and recommendations.

2.3. Feature Correlation

Understanding the relationships between different attributes is crucial for building accurate predictive models. Feature correlation analysis aims to identify patterns of association and dependency between variables. In this study, this exploration helps us select relevant features and uncover potential multicollinearity.

The correlation matrix of numerical attributes is visualized through a heatmap. It represents two-dimensional tables of numbers as shades of colors [4]. Notably, at tributes such as "host ID," "longitude", and "reviews per month" exhibit relatively strong correlations, suggesting potential interactions between them. Additionally, a moderate positive correlation between "calculated host listings count" and "number of reviews" indicates that hosts with more listings tend to accumulate more reviews.

This analysis aids in feature selection and model performance optimization, enabling us to construct predictive models that capture the complexities of Airbnb property pricing and market dynamics (See Fig. 3).

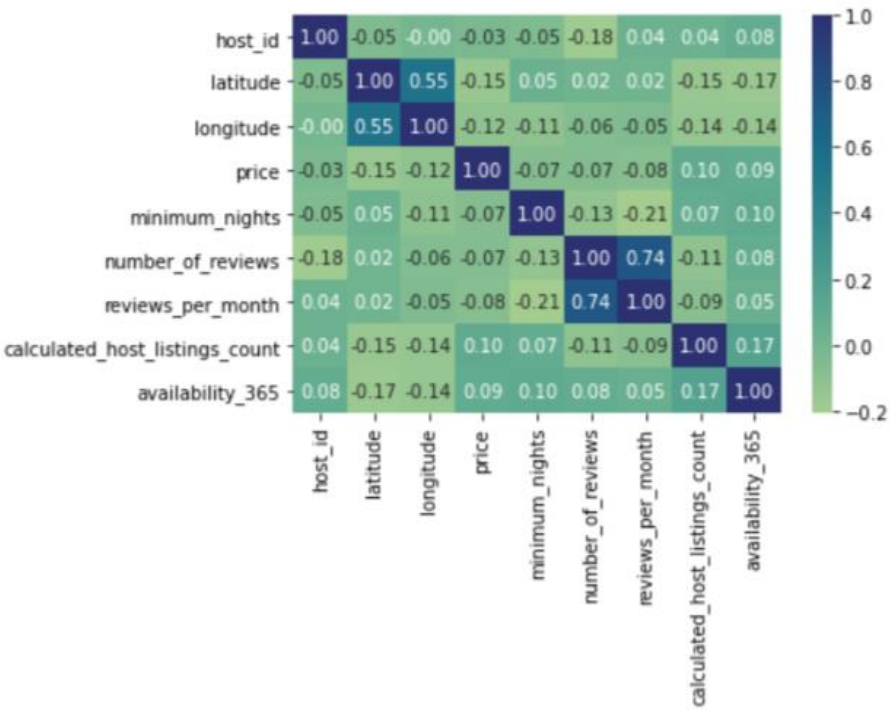

Fig. 3 Feature Correlation Heatmap

3. Modeling and Analysis

With a cleaned and processed dataset, the following step is to build the predictive models and analyze their performance.

3.1. Linear Regression

Linear Regression, a fundamental machine learning technique, forms the baseline for predicting Airbnb listing prices [5]. This method assumes a linear relationship between input features and the target variable. In this study, linear regression was employed as the baseline model. The assessment of the model's performance involved the utilization of metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R^2) score (See Fig. 4).

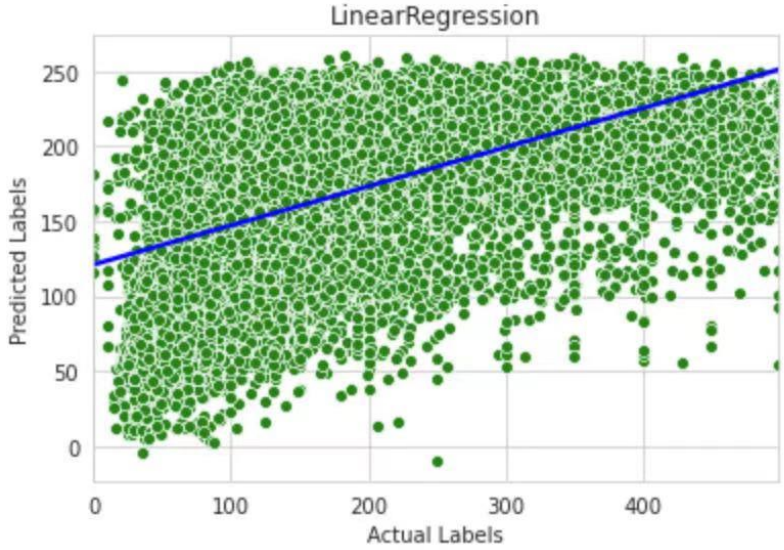

Fig. 4 Scatter Plot for Linear Regression

To evaluate the model's performance, the scatter plots and regression lines are employed. These visualizations illustrate how well the model's predictions align with the actual listing prices. Observing the scatter plot, it can be deduced whether the model captures the underlying trends and variations in the data. The regression line represents the model's best attempt to fit a linear relationship (See Table 2).

Table 2. Scatter Plot for Linear Regression

Metric	MSE	RMSE	R^2
Linear Regression	5969.93	89.27	0.265

Quantitative metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score provide deeper insights into the model's efficacy. MSE measures the average squared difference between predicted and actual values. RMSE, a variant of MSE, offers a more interpretable measure in the same unit as the target variable. The R^2 score gauges the extent to which the model elucidates the variation in the target variable. The evaluation of the Linear Regression model's performance metrics offers key insights into its predictive capabilities. With a calculated Mean Squared Error (MSE) of 5969.93, the model's predictions exhibit a substantial average squared difference from the actual listing prices, suggesting a struggle to capture the intricate pricing patterns in the dataset. The Root Mean Squared Error (RMSE) of 89.27 indicates a moderate level of prediction errors, implying relatively close predictions with some variability. The R-squared (R^2) score of 0.265 reveals that around 26.5% of the variance in listing prices is explained by the model's independent variables, highlighting limited explanatory power.

Moreover, from an economic standpoint, the US housing market is known to exhibit multifaceted pricing trends influenced by factors such as location, property characteristics, economic conditions, and demand-supply dynamics. The limited predictive accuracy of the Linear Regression model may be attributed to its inability to capture the non-linear relationships and interactions among these variables. Consequently, relying solely on this model for pricing decisions in the US housing market could lead to suboptimal results.

In summary, while Linear Regression serves as a baseline model, its performance in predicting Airbnb listing prices falls short in both the context of machine learning and the complex US housing market. This calls for the exploration of more advanced algorithms like Random Forest and XGBoost that can effectively capture the intricate pricing patterns and dynamics inherent to the domain.

3.2. Random Forest

A random forest is a collection of decision tree predictors in which the construction of each tree is influenced by a random vector from an identical distribution of all the trees in the forest and independently drawn [6]. The analysis of the Random Forest model's performance offers insights into its predictive accuracy and its significance within the context of machine learning and the US housing market. Reinforcing the Random Forest model's accuracy, the scatter and line graph visually represent its performance (See Fig. 5):

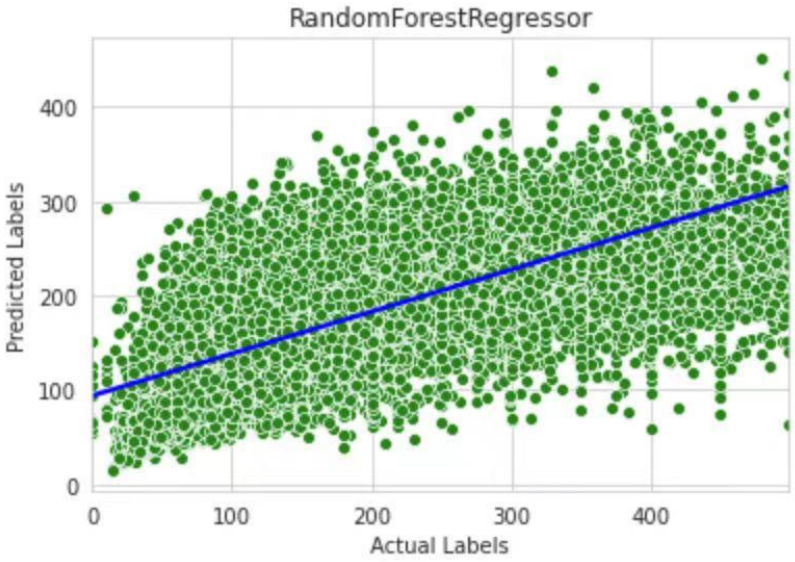

Fig. 5 Random Forest Model Visualization

The scatterplot shows that the data points are scattered and not tightly clustered around the regression line. This suggests that there is some degree of discrepancy between the random forest model's predictions and the actual labels. While the model's predictions may be relatively accurate for some samples, they may be significantly wrong for other samples. This dispersion may be caused by the failure of the model to capture certain features or complex relationships in the data, resulting in unstable predictions. Comparing the regression line with the diagonal line (y=x) again, there is a significant deviation rather than a complete overlap. This indicates a linear deviation of the model's predictions with respect to the true labels. Forecasts in some price ranges may be higher or lower than actual prices, indicating that the accuracy of the model varies across market segments. The calculated output metrics for the Random Forest model are as follows in Table 3:

Table 3. Performance of Random Forest Model

Metric	MSE	RMSE	R^2
Random Forest	4502.38	76.85	0.457

The Mean Squared Error (MSE) signifies the mean of the squared discrepancies between predicted and actual listing prices. A lower MSE indicates closer alignment between predictions and true prices, which in turn reflects improved model accuracy. Similarly, the Root Mean Squared Error (RMSE) offers an interpretable measure of prediction errors. In this case, the RMSE suggests moderate predictive accuracy. The R^2 score of 0.46 signifies that the Random Forest model captures approximately 45.7% of the variance in listing prices. This score indicates a stronger model fit compared to the Linear Regression model.

In the context of the US housing market, the Random Forest model's accuracy implies its effective incorporation of diverse economic factors, and market dynamics for accurate price predictions.

3.3. XGBoost

XGBoost, a powerful gradient boosting algorithm, was utilized for predictive modelling. By observing the scatter plot and regression line generated by the XGBoost model, notable conclusions can be drawn. It differs from a random forest in two ways. Firstly, Random Forest constructs individual trees independently, while XGBoost constructs one tree sequentially. Secondly, Random Forest aggregates the outcomes of all the trees at the conclusion, whereas XGBoost consolidates the results throughout its iterative process [7]. The scatter plot visually portrays the alignment between the predicted and actual listing prices. Notably, the data points are more closely clustered around the regression line, indicating a stronger agreement between the model's predictions and the true values.

Furthermore, the regression line closely follows the diagonal line (y = x), suggesting that the model's predictions generally exhibit strong concurrence with actual prices (See Fig. 6).

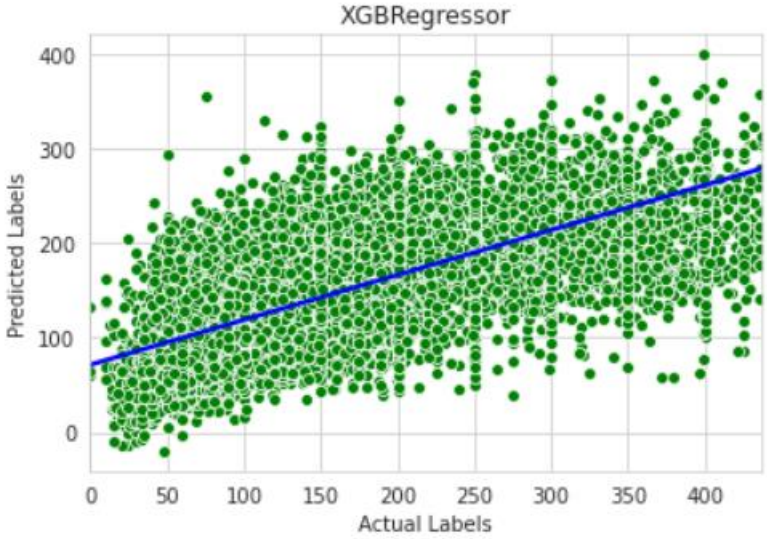

Fig. 6 XGBoost Model Visualization

These metrics offer quantitative insights into the model's predictive accuracy. The MSE and RMSE values reflect the average prediction errors, with lower values indicating better performance. The relatively low MSE and RMSE values for the XGBoost model signify that its predictions exhibit smaller deviations from the true prices, suggesting a higher level of predictive accuracy. The R^2 score of 0.473 indicates that the model explains around 47.3% of the variance in listing prices, highlighting its strong ability to capture the underlying patterns in the data. In the realm of machine learning, models with lower MSE, RMSE, and higher R^2 scores are considered more accurate and reliable. The XGBoost model's robust performance is indicative of its effectiveness in capturing complex data relationships, making it a suitable choice for predicting Airbnb listing prices. (See Table 4)

Table 4. Performance of XGBoost Model

Numble	MSE	RMSE	R^2
Linear Regression	4273.57	66.99	0.473

From an economic perspective, precise price prediction models carry substantial implications for the US housing market. Accurate pricing empowers property owners to set competitive rates aligned with market trends while providing potential buyers with transparent and fair pricing. The XGBoost model's ability to predict listing prices with minimal error contributes to a more stable and efficient housing market, fostering trust and confidence among both buyers and sellers. In summary, the XGBoost model's scatter plot, along with its MSE, RMSE, and R^2 scores, underscores its strong predictive capacity. This alignment with machine learning principles and its significance in the US housing market underscores its potential as a valuable tool for enhancing transparency and efficiency in the real estate landscape.

3.4. Parameter Tuning

To enhance the XGBoost model's performance, parameter tuning was performed using GridSearchCV and cross-validation. The XGBoost model performed better than the Linear Regression and Random Forest models: Among the three models, the XGBoost model obtained a lower RMSE and a higher R^2 score value, indicating that the XGBoost model has a relatively better prediction effect and can more accurately predict Airbnb prices [8]. The prediction of the Random Forest model is relatively unstable: Although the Random Forest model may perform well on some

samples, the dispersion of the scatter plot and the deviation of the regression line indicate that the model may have large prediction errors on other samples. This may be due to the complexity of the Random Forest model, which leads to overfitting of the training data and affects the generalization ability. The Linear Regression model performed the worst: the RMSE and R^2 score values of the linear regression model were high, indicating that the model had a large prediction error and a weak ability to explain the data, which may be because the linear regression model is not suitable for complex nonlinear data relationships and fails to capture the changes and trends in the data well [9] (See Fig. 7).

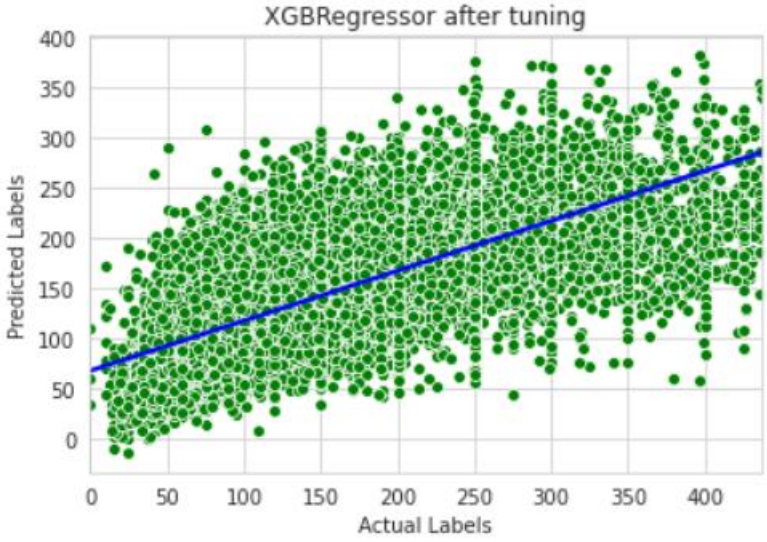

Fig. 7 XGBoost Model After Tuning Visualization

Adding to these, GridSearchCV and cross-validation are employed to identify the best parameters, resulting in the best combination. The data of the model test is shown in the Table 5 below The MSE of the XGBoost model is reduced to 3969.53, the RMSE is reduced to 62.27, and the R^2 score is increased to 0.51. The improvement of these indicators shows that the prediction accuracy of the model has improved, which can explain about 50.6% of the data variance, and the deviation from the real house price has also decreased.

Table 5. Tuned XGBoost Model Performance

Metric	MSE	RMSE	R^2
Linear Regression	3969.53	62.27	0.506

To sum up, after finding the optimal parameters by GridSearchCV and cross-validation, the XGBoost model performed best in predicting Airbnb housing prices in the United States, and it has better prediction accuracy and generalization ability than the Linear Regression and Random Forest models [10].

4. Conclusion

Airbnb is growing rapidly in the U.S. market, and price predictions are critical for both hosts and guests. In this study, three machine learning models are employed to explore the prediction of Airbnb listing prices in the US. The analysis revealed that XGBoost outperformed Linear Regression and Random Forest in predictive accuracy. After parameter tuning, the tuned XGBoost model, yielded the best results. However, this study only compared three machine learning models. Future research can try more complex models and methods, expand data set features, optimize parameters and feature engineering to improve prediction accuracy and interpretability.

References

[1] Jiao, J., & Bai, S. An empirical analysis of Airbnb listings in forty American cities. Cities, 2020, 99: 102618.

[2] Dhillon, J., Eluri, N. P., Kaur, D., Chhipa, A., Gadupudi, A., Eravi, R. C., & Pirouz, M. Analysis of Airbnb Prices using Machine Learning Techniques. In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference, 2021: 0297-0303.

[3] Kunlong, M. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[4] Rajaram, S., & Oono, Y. NeatMap-non-clustering heat map alternatives in R. BMC bioinformatics, 2010, 11(1): 1-9.

[5] Su, X., Yan, X., & Tsai, C. L. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 2012, 4(3): 275-294.

[6] Wang, H. Predicting Airbnb Listing Price with Different models. Highlights in Science, Engineering and Technology, 2023, 47: 79-86.

[7] Lektorov, A., Abdelfattah, E., & Joshi, S. Airbnb Rental Price Prediction Using Machine Learning Models. In 2023 IEEE 13th Annual Computing and Communication Workshop and Conference, 2023: 0339-0344.

[8] KESER, M. PREDICTING AIRBNB LISTING PRICES IN ISTANBUL USING MACHINE LEARNING AND SENTIMENT ANALYSIS (Doctoral dissertation, tilburg university).

[9] Cao, Y., Ashuri, B., & Baek, M. Prediction of unit price bids of resurfacing highway projects through ensemble machine learning. Journal of Computing in Civil Engineering, 2018, 32(5): 04018043.

[10] Zhu, A., Li, R., & Xie, Z. Machine learning prediction of new york airbnb prices. In 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), 2020: 1-5.