Proceedings of CONF-APMM 2025 Symposium: Simulation and Theory of Differential-Integral Equation in Applied Physics
DOI: 10.54254/2753-8818/2025.DL33617

Developing a Predictive Pricing Model for Airbnb Rentals Using Linear Regression Analysis

Zhipeng Luo[1*†], Zixuan Gao[2†]

[1]*Tsinglan International School, Dongguan, China*
[2]*Bashu Science City School, Chongqing, China*
**Corresponding Author. Email: Louis.Luo_26@tsinglan.org*
[†]*All the authors contributed equally to this work and should be considered as co-first author.*

Abstract. This research paper aims to identify the factors that address the pricing of Airbnb rentals and thereafter develop an interpretable predictive model that can increase the transparency and effectiveness of any peer-to-peer rental platform. We combine the openly available information from Inside Airbnb and Kaggle to create a linear regression model with log-transformed prices, whereby the skewed data and nonconstant variance across the observations will be stabilized. We will comprehensively analyze the key listing-level characteristics, which include room type, number of bedrooms, number of bathrooms, accommodates, cancellation policy, and review score, to see their impact on the rental prices. Through regression analysis of 37,872 listings, a model with an Adjusted R^2 of 0.506 has been derived, confirming the model's explanation. Research has led to the notion that the determinants of price in this sector are mainly property-level ones, whereby room type, and capacity exert more pressure than the review score itself. Multicollinearity and non-normal residuals, however, could be issues, but they are poor analytical solutions by the models. On the other hand, the data-driven insights generated by these models are useful for all property managers and guests alike. Through this, the hosts would have pricing strategies, the guests can tell that the right value is paid, and the platform to refine recommendation systems could be introduced. The value of interpretable regression models for democratizing data-driven pricing, despite their shortcomings in feature scope and model complexity, is highlighted in this research. Future additions to the model will consider information pertaining to geospatial, temporal, and text data. As well, these data will be processed to be analyzed by advanced machine learning systems to enhance the prediction accuracy of the market and to capture non-linear demand patterns.

Keywords: Predictive modeling, Linear regression, Log-transformation, Data-driven pricing

1. Introduction

An increase in data coming from geo-location information for dockless bike-sharing systems is evidence of a growing sharing economy. Consumption patterns and the social demographic reflect can be examined using these data sets [1]. Accordingly, platforms such as Airbnb have emerged as principal mates of this wave of change. By the construction of complex and decentralized short-term

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.DL33617

rental market, December they deeply challenge the conventional pricing procedure. The "3Vs" of big data—Volume, Variety, and Velocity—are presented as important analytical opportunities for studying and controlling contemporary urban areas that are characterized by such large-scale data sets [2]. Consequently, this situation has been augmented by AI tools whose objective is to increase the use of the internet for advanced analytical methods that enable data-based approaches. Moreover, flexibility for the market and pricing is improved [3]. A significant part of the marketplace for Airbnb is that unlike standard hotel and apparent accommodation pricing, the Airbnb marketplace is heterogeneous and has a dynamic mix of qualitative and quantitative components so that it is difficult to arrive at what the actual market value for the accommodation is. This exponentially worsens in the case of spatial determinants as the issue gets more puzzling when one considers the contradictory nature of the way neighborhood or city level characteristics or factors affect pricing exercises [3, 4]. Complexity comes with the market in which you as a host or a guest are essentially "information poor" and the market is largely inefficient because of this information asymmetry.

This project is aimed at showcasing the process in which predictive analytics can remove discrepancies and thus promote efficiency and value creation by doing the following:

• Understand and utilize the impact of various characteristics on the listings using the structured data from the listing to achieve this. These characteristics or variables include the room type, number of bedrooms and bathrooms, accommodation capacity, cancellation policy, and review scores.

• Use the log-price as the target variable to present the most common method that statistical modeling applies to address the issue of price skewness. This approach also allows for a wider range of/ or more stable variance across the observations.

According to this study, three research questions help in understanding the aim of the study: thus,

1. What listing properties explain most of the variation in rental cost for an Airbnb rental?

2. What can the unstable situation and limitations of simple linear regression in this situation be when the log transformation is applied?

3. The implications of the results for both hosts as well as guests and revenue management strategies are at a platform level.

The practical importance of the presented work is that it could lead to democratization of data-driven pricing, improving the fairness of markets in digital. This maximizes the yield for landlords, provides information on ideal fees and algorithmic recommendations for coverage platforms, and allows for better-informed and transparent policymaking, as city policymakers can understand dynamics in the short-term rental industry. This article is divided into the following sections. Section two brings about the review of the related literature on the predictive models and price determinants in the Airbnb sector. In Section three, the techniques and methods used are discussed; the process of gathering data, cost, and model methods. A brief overview in section four regarding the relevant results of the regression analysis is provided and all results are discussed in Section 5: this section emphasizes the real world understanding of results and their analytical weaknesses. In Section 6 of the report, the main conclusions have been summarized and suggestions for future research have been provided. The areas of future research that are mentioned include the integration of dynamic pricing and real-time external elements.

2. Literature review

The rise of Airbnb as a disruptive player within the industry of accommodations has brought the effect of market pricing mechanism, causes, and possibility of prediction to the attention of

DOI: 10.54254/2753-8818/2025.DL33617

researchers. The earliest outlines of urban governance systems driven by technology-stamped platforms accentuate big data and artificial intelligence (AI) as forms of urban governance and plan making [1-3]. Such works bring out the fact that the environmental information-rich systems facilitate quick decisions in market management such as short-term rental platforms. Airbnb, in this sense, has been proven an example, where the theory of homogeneity and centralized pricing is flawed, causing information asymmetries and market failure. Price setting on Airbnb has been one of the core research tracks in price setting. Location is the major feature, but its impact on price is different from expected despite study by different scientists. There are those studies which attribute the leveling of neighborhood amenities, transport availability, and socio-cultural characteristics mainly for raising listing prices [5]. High attention will go to the level features like property size, host reputation, and the reviews [6]. The disparity here actually accounts for the price units not being uniform across various peer-to-peer markets. Besides, as the scholarship has now realized the shortcomings of traditional econometrics, applications of machine learning and artificial intelligence are on the rise. Hedonic Regression and log-price transformations were widely used both to reduce skewness and promote interpretation in the analysis [7].

At the same time, not only has including significant public transport and neighborhood index in Istanbul led to more than a twenty percent increase in estimated accuracy level but also led to a more than a seventy percent fit of the data [8]. Therefore, the work also included calendar-level time series and sentiment analytic techniques to capture case fluctuations in Austin, Texas as having heterogeneous data sources as an advantage. Studies such as [9, 10] represent these frameworks by embedding sustainable ML models with sentiment analysis, as well as recommendation-enhanced architecture to choose the accurate and simple predictions that will be useful for both host and the entire ecosystem.

Among the various issues, there appears to be a ban on making models that try to mimic human decision-making abilities use AI-based models (e.g., Random Forest, XGBoost, Neural Networks) due to the advantages of predicting more accurately than the traditional models. Alternatively, well-established models such as linear orders and log transformation of price also hold significant importance in identifying statistically significant determinants of price. There is suggested a hybrid usage of interpretable and high-performing model to keep a reasonable balance between their accuracy and usability towards end-users [6]. Eventually, comparative markets and models are also researched in which there exist such similar blockages. We see implementation of machine learning techniques as well as real estate, agricultural products [11], and housing platforms such as Zillow [12] as not just enhancing the cross-domain applicability of machine learning models but also giving researchers application-oriented references for their Airbnb price prediction studies.

With this motivation, we are working on hybrid interpretable model for Airbnb price predication that will strengthen the current efforts to improve the quality of results. Through this approach, we are not only able to learn richly nonlinear structures but also retain the artifacts of complex system property. In doing so we would make a platform for people such as hosts, algorithms, and policymakers, as well as for proper stakeholders, to gain an understanding of the role of structured features, which include spatial context and latent features, in a way that is user-friendly. To accomplish this, we shorten the gap between accuracy and interpretability and make data processing with adjusted prices more democratic, as well as provide a knowledge base that is currently in existence to fill the gaps in Airbnb pricing investigation.

DOI: 10.54254/2753-8818/2025.DL33617

3. Methodology

3.1. Data description

This study uses data from the list level of Airbnb that are carried out in a manner that avoids scrapping. This has been accomplished by us making use of information that is readily available from specific repositories that are inside Airbnb and Kaggle. Most of these websites have been built in a way that enables the use of datasets that comprise property features in different cities. The range of the data set covers multiple years and provides also a heterogeneous property characteristics range. This includes the various type of data such as numeric [bedrooms, bathrooms, accommodates, reviews, prices] and categorical [Type of Room, Cancellation Policy, HC, Instant Bookable].

Since accommodation rent in peer-to-peer accommodation markets are steeped in present high-end assets and luxury listings too, we assign log prices as the dependent variable. Log price (log) transformation is not only dealing with skewness and downgrading single values influence, but the regression model including is gets also able to account for proportional relation. As shown in Figure 2, the original price distribution was highly right-skewed. After applying log transformation, Figure 1 demonstrates that the distribution became approximately normal, which improves model stability.

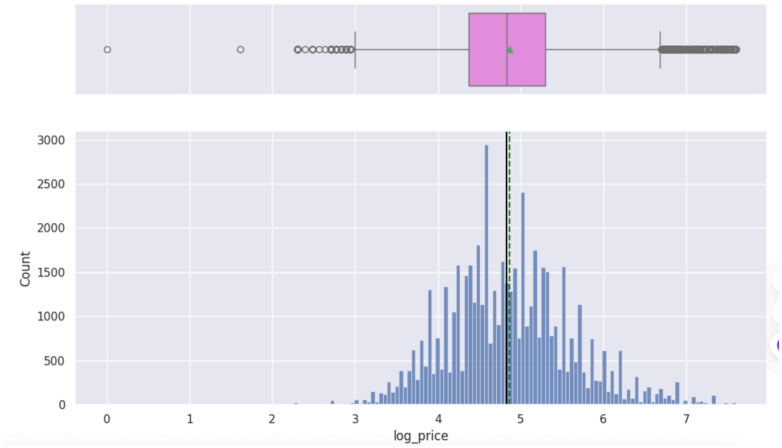

Figure 1. After log

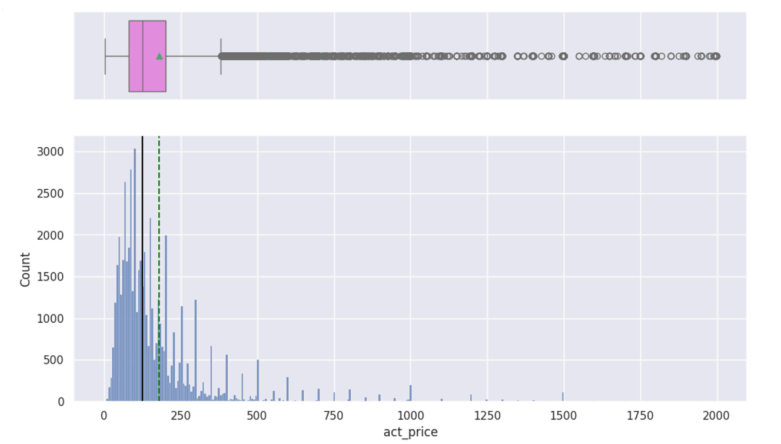

Figure 2. Before log

DOI: 10.54254/2753-8818/2025.DL33617

In addition, there are numerous variables including rent style information, property features, and occupancy level that account for differences among the same listings. As far as the algorithms are concerned, non-essential characteristics like the description, name or symbolic element are dismissed due to the minor explanation of prediction. The dataset following the data cleaning process comprehensively represents the situation concerning predictive analysis, but it is still simplified and manageable.

3.2. Data pre-processing and visualizations

The existing raw data file of Airbnb was imported to identify the type of data, which contains the number of rows, columns, and data types. By removing non-informative items such as the "id" column that has little to no analytics value, we create a clean analytical copy of the dataset to avoid inadvertently modifying the original. We have adopted and incorporated duplicate detection methods to eliminate repeat or over-representation of corresponding data while creating an updated copy of the dataset. We also examined carefully all missing values in order not to lose out on useful information. Analyzing the category variables, we ensured to prune the ones that have low explanatory ability while we made the imbalanced outliers of numerical variables exactly uphold their distribution of the data. To prepare the data later for the regressions, redundant or wrongly recorded information were distinguished and dropped off, all the variables were recoded into the same format for coherence. Outlier detection techniques relied on visual methods such as box plots that outlined outliers in relation to their leverage, while clipping techniques were employed to mitigate the influence of those outliers. Normalization was also applied to rescale features into a comparable range, and the price variable was log-transformed to correct skewness and stabilize variance. The role of data visualization was key in understanding the underlying patterns: we used the method of sampling random data set to spot inconsistencies, and the method of correlation he, or heat maps to show correlations among numerical predictors, and the results revealed a strong dependence on features such as bedrooms to bathrooms, both of which during accommodation time supplement each other. Additional histograms and distribution plots confirmed that with the application of log transformation, the price distribution became more symmetrical. Preprocessing and visualization stages accompanied each other to give the dataset an implicit form, as being of an understandable, standardized form. Hence, this reassignment the dataset into a structured, normalized, and interpretable set and sets up a reliable foundation for regression modeling and statistical inference.

3.3. Model: linear regression

Inevitably, a linear regression estimator is the first choice as the benchmark of modeling; it's transparent and interpretable in contrast to other methods of valuing heterogeneous properties, non-occupancy, like Airbnb rentals. The postulate in the framework is stated as follows: the dependent variable log (rental price) can be expressed as a linear function of structural property and host attributes, which can be stated mathematically by the following formula:

$$log(price_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i \quad (1)$$

where:

- $\log(\text{price}_i)$ is the dependent variable representing the logarithm of the listing's price,

DOI: 10.54254/2753-8818/2025.DL33617

- $x_1, x_2, \ldots, x_k$ are explanatory variables such as room type, accommodates, and review scores,
- β represents the regression coefficients that capture the marginal effect of each predictor,
- ϵ denotes the error term, assumed to follow a normal distribution with mean zero and constant variance.

The persistence of the linear regression depends upon whether stringent statistical conventions are observed or not as stipulated. This article admits and tries to confirm the hypotheses that underpin this theory. First, we assume that property features (independent variables) cause proportional changes in log-price (the dependent variable). This is interpreted to mean that the price of rent would be close to the current amount multiplied by the constant for each additional unit, e.g., a room or a bathroom, in line with how much the rate will go up each time an additional feature or property. This is simplified, no doubt, but real-life relationships maybe more complicated; however, this serves as a strong and intuitive basis for our pricing models. The model moreover presumes that there is a relatively low degree of multicollinearity when it comes to independent variables. Predictor variable's magnitude is not a matter to be heavily dependent on other predictors. Of course, this is to say there could be a correlation between bedrooms and accommodates.

The correlation heatmap in Figure 3 illustrates strong associations among several predictors, particularly between the number of bedrooms and accommodates. Such correlations indicate potential multicollinearity, which can reduce the stability and interpretability of the regression coefficients. Such problems were discovered when we performed an exploratory data analysis, where for each variable we calculated how unique each feature was to each price level. These findings are revealed in the correlation matrix. Particularly, for the second problem we were able to observe easily how the error increased for every level within each vine. We did this graphically through a plot of high residual values against predicted price and hoped for a scattered pattern of points with no obvious shape. Moreover, we are presuming that the residuals of the model have a normal distribution. This is very essential for the vagueness of any statistical inferences (like the calculation of confidence intervals). Another way to assess whether this normality assumption holds is through a histogram or a Q-Q plot; from those plots, we conclude that residuals are normally distributed (approximately). Therefore, this means that the error of the model is randomly distributed and unbiased. Third, it is necessary that the residuals be not autocorrelated, then, there is no association between them. This is implying the consequence which the errors are uncorrelated. In this context, less important for independent observations. However, when compared to time-series data, they do not predict directly to full dependence in a cross-sectional dataset such as this one. An example can be residual planes and Q-Q planes. These conditions can be met through the log transformation of the target variables boosted by the scaling of features.

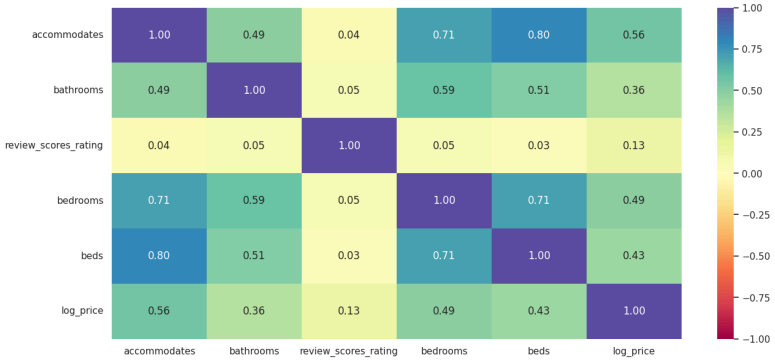

Figure 3. Heatmap

DOI: 10.54254/2753-8818/2025.DL33617

4. Results

4.1. Regression results

This section talks about the results from the ordinary Least Squares (OLS) regression analysis that was carried out to log price of the Airbnb listing. The diagnostic properties of the OLS regression model indicate the adequacy of the model fit which was based on the posterior checks of the linear regression assumptions. The OLS regression was carried out on a sample of 37,872 hosts. During this analysis, the best model is the one with R2 being .506 and the adjusted R2 being .506; this model also produced an F statistic of 3528.00, which at p-value .000 showed that the model had statistically significant existence and indeed explained large part of the variance.

4.2. Coefficient analysis

The coefficients shown in the result table provide evidence showing statistically significant effects of independent variables on log-price of an Airbnb listing. The implication of a positive and negative coefficient is considered that with the increase (or the decrease) of the variable's value, the log-price is inclined to increase (or to decline). Cancellation_policy_strict's coefficient in this case is statistically insignificant (p=0.00).Key findings from the coefficients include: Property-level features: Physical features, like room size and number of bedrooms, have significantly correlated with price level; so, we predicted at the end of our study that these factors will have an impact on log-price and they did have; moreover, the coefficient on bedrooms (value was 0.1572) reveals that an extra room in one's house would link a unit addition to a increase in rental price in total. For review_scores_rating, the reading is also positive and has the same positive coefficient as approval ratings leads to higher logs price.

Room Type: room_type can certainly affect house rental prices. However, to further illustrate that room_type is of great importance in the pricing predictability, this study found that some other room_types (like room_type_Private room and room_type_Shared room) are highly negative correlated with room_type (which was indeed room_type_Entire home/apt). Besides, like our study's outcome, we strongly believe that room_type is the most important variable for the housing price. Other Features: Having small negative coefficients (-0.0001) for cleaning fee implies that almost no effect of cleaning fee on the log-price. A similar correlation (-0.0440) can be derived from cleaning beds. Quite high results were obtained with instant_bookable being -0.0578, which holds true as those guests who are secured need to charge slightly higher rates. Cancellation_policy_moderate predicted coefficient is -0.0681, with Cancellation_policy_strict having no significant (p) meaning (0.388).

4.3. Validation of assumptions

The tables depicting the degree of the assumptions (which are yet discussed in the methodology section) made about the adoption of linear regression have been displayed in the results section. Multicollinearity: The Condition Number of 1.87e+03 is quite high, hence it may signal that overwhelmed collinearity or another problematic numerical issue that happen over there. Given this, it can be evidenced that some features may have high correlations amongst them which leads to a situation whereby the underlying coefficient estimate values are more likely to be affected by different data inputs and hence it may render the estimation process of the linear regression model and interpretation of the model output values imperfect, as detailed in the method section.

DOI: 10.54254/2753-8818/2025.DL33617

Homoscedasticity and Normality of Residuals: Through applying "Omnibus Test" statistic value of 2396.203 and p value is 0.00. JB yields a statistic of 3978.924 and corresponding p-value of 0.00. Consequently, this error implies that we should reject the idea that the residuals are normally distributed and hence revolt against the idea of normality. Moreover, Skew of 0.503 is very high whilst Kurtosis is at 4.229 in a very high range. The Durbin-Watson statistic is calculated as 2.014 close to 2 which points out the lack of autocorrelation of the residuals. The model exhibits signs of problematic multicollinearity and clear violations of the normality assumption for the residuals even though it leads to a very important and significant portion of the variance of airbnb prices explained in the model. A real-world relationship that is sometimes can't be solved by a linear approach. When carrying out valid research, one should test such assumptions to see whether they are violated using various test techniques. These techniques are dependence (test of residuals independence) and normality (test of residuals normality). The response variable, i.e., price has been transformed using the log function and the features have been scaled, subsequently to sometimes empirically proof these issues, but they are not overpassed.

5. Conclusion

The core of the study is building a predictive pricing model by means of linear regression analysis. The main objective is to establish the property features with most influence on rental prices together with the development of a model with good accuracy and outcome that can generate actionable advice for hosts/ guests/ airbnb platformers. Summary of Main Results Our linear regression model, amidst the numerous limiting conditions and constraints, helped us capture a rather steep trend and provided us with a data set that could help us in explaining Bhutanese prices. The model performance indicator particularly R-squared (0.506) and model RMSE suggested that the specified method had a status of a suitable predictor. The implication of this finding is that the physical property characteristics, as a variable, have a sizeable influence on the value. By in large, the type of room comprised the most significant feature of all, with lowest prices being found on the listings called 'Private room' and 'Shared room', and the highest being found in the listings for 'Entire home/apt'. Our model unfortunately also identified number of persons, bedrooms and bathrooms which the property serves as important variables for the rental price also. log_price the logarithmic model of the price (log_price) was the primary way in which it normalized its distribution. However, this was done to help the model meet more closely to the linear regression model's assumptions.

Practical Implications of Findings. For Hosts: The model also can give a purpose of an actuarial guide to data-driven pricing. The model be a host to compare their price with those of comparable objects, so they can be sure that they are beating the competition. This section of the report lists the features that are more highly valued by the host and subsequently allows the host to invest (adding an additional bed, if there is still room) to maximize the revenues that will be spent. For Guests: This study has laid a foundation for the other studies done in the area by providing a base with honesty. Due to the transparency that has come with the introduction of prices for properties having certain features, the audience have a better taste to know what a fair amount requires most properties, and those high-priced ones meant to offer great value-for-money. For Airbnb: They can be used to improve price recommendation algorithm on Airbnb for hosts. Ayima Certainly from these insights, we start understanding how the market works whoever because we can interrelate different behaviors, we also know when a behavior happens good or bad, we know the indicator to use to anticipate how things would go. Possessing such quality, airbnb is better able to get the market developments, inventory control, and these decisions probably would guide to shape the better strategies for its form.

DOI: 10.54254/2753-8818/2025.DL33617

Study Limitations Notwithstanding this strength, there are downsides to this investigation. The cases arise for the following two reasons: The first one is that there are only a few concerned features in the beginning of the model testing. The crucial factors of loss of exogenous variables: location detail (neighborhood, near transportation, and so on), seasonality, and availability (e.g. a pool or parking space) were excluded, despite their utmost necessity in the price definition. Secondly, linear regression is an inherently simple and unadjustable model, not realizing any non-linear correlation and relationship that exist within the data. Comprehensive methods of ML (e.g., Random Forest, XGBoost) used in previous chapters of the literature under consideration may get more powerful predictions. Ultimately, the information is only a snapshot in time. So, it will be very hard to say whether this model will be able to generalize the performance of other cities unlike its application to the city, or this application will be able to stand for other time periods. There are numerous ways to carry out an extended version of the current research. An important extension is for the global model to include geospatial data to measure the impact of the neighborhood and distance to popular links on price. Secondly, machine learning is capable of modeling more complicated and the more complex patterns. Hence, by doing this, the accuracy of prediction will probably be higher, though of course, it depends on the model's specification. Another angle we might want to explore is to do the price research in the historical context. Finally, NLP can be implemented to quantify the guest perspective provided in the reviews that potentially describe quality and perceived reputation of the host. That would be a potentially low-cost intervention based on retrospective research that could improve the final prediction of our model.

Acknowledgement

All the authors contributed equally to this work and should be considered as co-first author.

References

[1] Allam, Z. and Dhunny, Z.A. (2019) 'On big data, artificial intelligence and smart cities, ' Cities, 89, pp. 80–91. https: //doi.org/10.1016/j.cities.2019.01.032.
[2] Kandt, J. and Batty, M. (2020). Smart cities, big data and urban policy: Towards urban analytics for the long run. Cities, 109(102992), p.102992. doi: https: //doi.org/10.1016/j.cities.2020.102992.
[3] Yigitcanlar, T. et al. (2024) 'Artificial intelligence and the local government: A five-decade scientometric analysis on the evolution, state-of-the-art, and emerging trends, ' Cities, 152, p. 105151. https: //doi.org/10.1016/j.cities.2024.105151.
[4] Xu, F., Hu, M., La, L., Wang, J. and Huang, C. (2019). The influence of neighborhood environment on Airbnb: a geographically weighed regression analysis. Tourism Geographies, 22(1), pp.192–209. doi: https: //doi.org/10.1080/14616688.2019.1586987.
[5] Sainaghi, R. and Baggio, R. (2020). Substitution threat between Airbnb and hotels: Myth or reality? Annals of Tourism Research, 83(1), p.102959. doi: https: //doi.org/10.1016/j.annals.2020.102959.
[6] Camatti, N., Tollo, G. di , Filograsso, G. and Ghilardi, S. (2024). Predicting Airbnb pricing: a comparative analysis of artificial intelligence and traditional approaches. Computational management science, 21(1). doi: https: //doi.org/10.1007/s10287-024-00511-4.
[7] Wang, H. (2023). Predicting Airbnb Listing Price with Different models. Highlights in Science Engineering and Technology, 47, pp.79–86. doi: https: //doi.org/10.54097/hset.v47i.8169.
[8] Chapman, S., Mohammad, S. and Villegas, K. (2023) Predicting listing prices in dynamic short term rental markets using machine learning models. https: //arxiv.org/abs/2308.06929.
[9] Alharbi, Z.H. (2023). A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis. Sustainability, 15(17), pp.13159–13159. doi: https: //doi.org/10.3390/su151713159.
[10] ganta, N. and Krishna, M. (2025). Price Prediction and Recommendation of Airbnb Property Listings Using Machine Learning. SSRN Electronic Journal. [online] doi: https: //doi.org/10.2139/ssrn.5214669.

DOI: 10.54254/2753-8818/2025.DL33617

[11] Göktaş, Ö. (2023) 'Yield, yield components and fatty acid composition of Artemisia dracunculus L. local breeds under organic cultivation conditions, ' The Pakistan Journal of Agricultural Sciences, 60(04), pp. 583–592. https: //doi.org/10.21162/pakjas/23.33.

[12] Rolli, M., Chen, X., & Patel, R. (2020). Zillow homevalue prediction (Zestimate) by using XGBoost. In Proceedings of the International Conference on DataScience Applications. IEEE