Contents

[Literature review 1](#_Toc155824722)

[An overview of models for predicting Airbnb accommodation prices 2](#_Toc155824723)

[Linear regression 3](#_Toc155824724)

[Ensemble learning 4](#_Toc155824725)

[Neural Network 6](#_Toc155824726)

[Multi-modality data to predict rent prices 6](#_Toc155824727)

[Environmental noise and air quality on accommodation price 7](#_Toc155824728)

[Room prices and points of interest 8](#_Toc155824729)

[Conclusions 9](#_Toc155824730)

# Literature review

In recent years, the landscape of the sharing economy has undergone substantial growth, propelled by platforms like Airbnb that have revolutionized how people access accommodations. Within the sharing economy, individuals collaborate by sharing resources, goods, or services through online platforms or communities, optimizing the use of assets that might otherwise be underutilized. Airbnb, a key player reshaping the accommodation sector, introduces a noteworthy challenge due to the dynamic pricing nature. Unlike traditional lodging options, Airbnb prices can fluctuate based on factors such as location, demand, seasonal variations, and unique listing attributes. This inherent unpredictability poses a specific dilemma for hosts and guests as they seek to strike a balance between cost-effectiveness and value.

To grapple with this challenge, the integration of data science has assumed increasing importance. Researchers can leverage the wealth of historical data available on Airbnb to extract insightful information on pricing trends, customer preferences, and market dynamics. These insights serve as the groundwork for constructing accurate models capable of predicting Airbnb listing prices.

The accurate prediction of Airbnb listing prices carries significant implications for both hosts and guests immersed in the sharing economy. For hosts, precise price predictions empower them to fine-tune their listings by setting competitive prices that align with market dynamics and property attributes. This, in turn, maximizes their revenue potential and elevates overall business performance. Moreover, it furnishes hosts with insights into the factors influencing price fluctuations, aiding informed decisions regarding property upgrades, amenities, and other listing characteristics.

Conversely, accurate price predictions benefit guests navigating the sharing economy by fostering transparency and enabling them to make well-informed decisions based on their budget, preferences, and desired amenities. Within the sharing economy framework, where resources are shared among individuals, precise price predictions enable guests to select the most fitting accommodation options and plan their trips with efficiency. This transparency heightens their overall satisfaction and enriches their experience of engaging in the sharing economy through the Airbnb platform.

While predicting prices of accommodation has certainly been a well-researched area there still is room for improvement, as mentioned sharing economy prices are subject to fluctuations from different factors highlighting the need for well rounded approaches that cannot simply rely on accommodation-based modelling.

This research will focus on the concept of using different features mostly related to geolocation of the accommodations to create a reliable and accurate price prediction strategy.

## An overview of models for predicting Airbnb accommodation prices

Many papers have explored the prediction of room prices for both long- and short-term accommodation; many models were chosen for this objective.

### Linear regression

In this context linear regression can be used for modelling the relationship between a dependent variable and one independent variable by fitting a linear equation to the observed data. The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple independent variables) that minimizes the differences between the observed and predicted values.

Linear regression offers many advantages: it can perform better than other models when the relationship between the variable follows a linear pattern, it is easy to compute and easier to explain than more complex algorithms.

However, it needs to be handled with caution as linear regression is susceptible to overfitting; also, noise and outliers can dramatically reduce model’s performance, when handling linear regression, it is important to keep in consideration how to deal with multicollinearity (a state when independent features are correlated within themselves).

When it comes to utilizing linear regression for predicting accommodation prices in tourism or real estate several authors already approached the subject.

Yu et al. (2018) in a paper called "Real Estate Price Prediction with Regression and Classification" [1] developed a set of models and used linear regression as benchmark for forecasting the ideal sale price of a home by using a dataset of residential houses sold in Ames, Iowa from 2006 to 2010; the datasets consisted of 79 features and more than 1000 houses, the authors used Root mean squared error (RMSE) as a commonly used performance measure for evaluation of price predictions, overall the paper reached a score of 0.55 by using simple linear regression with no regularization.

To reduce loss the paper also applied two regularization methods: Lasso and Ridge: both regularizations help prevent overfitting enhancing the generalization of the model, Lasso works by adding a penalty term to the function discovered by the regressor therefore avoiding overfitting.

Lasso regression often shrinks down the coefficient of a particular variable down to zero therefore implying a feature selection technique were only the relevant features have a coefficient that is different from zero, Ridge regression also similarly to Lasso adds a penalty term and shrinks down the coefficient but it never reaches exactly zero.

Yu et al. Reached a similar RMSE score with ridge and lasso regression of 0.54 compared to 0.55 on simple linear regression, in another paper Chapman et al. “PREDICTING LISTING PRICES IN DYNAMIC SHORT TERM RENTAL MARKETS USING MACHINE LEARNING MODELS” [2] found similar scores using Lasso and ridge regression: the performances of the two models were extremely close regardless of regularization techniques; they found a R-squared of around 0.52 for Ridge and 0.51 for Lasso attributing the small difference to the ridge regularization factor handling better the multicollinearity of the features.

The paper also explored, among others, Elastic Net algorithm: Elastic Net can combine both L1 and L2 regularization techniques; the methodology of the paper revolved around comparing the results for each algorithm on two datasets, one with and another without feature selection. Elastic Net performed quiet similarly to Lasso regression for a standard dataset; however, it performed worse on the dataset when feature selection was applied, the authors identified Elastic Net as unnecessary complex leading to slightly worse performance than Lasso and Ridge.

Another paper by Ziyue Huang “Logistic Regression in Rental Price and Room Type Prediction Based on Airbnb Open Dataset” [3] confirmed previous results with very similar success using Lasso regression and hyper parameter turning showing a positive correlation between the accuracy or the model and the regularization parameter.

Overall, the literature seems to do frequent use of linear regression techniques, however, the algorithm is mostly used as a benchmark to test other algorithms which proven more effective than linear regression.

### Ensemble learning

When it comes to price predictions in accommodation random forest and other ensemble learning models are one of the most used; ensemble learning can be explained as a technique which takes in consideration various model and their prediction and combines them into a more robust prediction by compensating the weakness of a single prediction improving the overall performance.

Ensemble learning utilizes different methods to achieve an accurate prediction, the two main approaches are bagging and boosting: bagging is the process of train the same type of models on different batches of the same dataset, each model is trained independently and then the output of the models are combines into a single prediction; boosting instead works by developing models in a sequential fashion and compensating the mistakes made by the previous model by putting emphasis on the errors made by the previous model, again the prediction are then combined at the end for a final prediction.

Overall ensemble learning models results more generalizable than traditional models for the aforementioned reasons, over the years multiple papers have explored the use of ensemble learning models of the price prediction of accommodation.

Hu et al. in “Prediction and Analysis of Rental Price using Random Forest Machine Learning Technique” [4] explored the impact of different features when predicting the prices of rents for two major cities in China Wuhan and Shanghai, the dataset used was comprehensive of multiple features regarding location, amenities and quality of the accommodation.

The result of the paper indicates a good R-squared score, the model worked especially well for Wuhan compared to Shanghai, the researchers stressed the importance of selecting the correct features such as city district which was found as particularly important.

Adetunjia et al. in “House Price Prediction using Random Forest Machine Learning Technique” [5] also implemented a random forest model with good results, the data was collected from a dataset of 500 homes in Boston collected in 1978 with 14 features such as location, closeness to the main river, number of rooms.

The methodology applied consisted of using bootstrap as sampling method and 500 trees in total, as for previous literature the R-squared score was set at 0.9 showcasing a high degree of accuracy.

XGBoost, or extreme Gradient Boosting, stands out as a potent ensemble machine learning algorithm, widely recognized for its outstanding predictive capabilities. Functioning as a boosting algorithm, XGBoost amalgamates the outcomes of numerous weak learners, typically decision trees, resulting in a robust and precise model. It systematically constructs trees with a focus on rectifying errors from previous trees. XGBoost uses regularization techniques to prevent overfitting and Its efficiency is further elevated through parallel processing and hardware optimization.

In the already mentioned paper from Chapman et al. “PREDICTING LISTING PRICES IN DYNAMIC SHORT TERM RENTAL MARKETS USING MACHINE LEARNING MODELS” XGBoost even if comparing better than other models it falls short performing slightly worst than a random forest regressor used on the same dataset; the authors outlined how the iterative nature of the model made the predictions susceptible to noise and outliers reducing the generalization of the trained model to the testing data, however the ensemble learning regressors used in the paper outperformed linear regression.

### Neural NetworkTop of Form

In recent years neural networks have emerged as powerful tools in predicting financial market prices, revolutionizing the landscape of quantitative analysis and forecasting. These computational models draw inspiration from the human brain, consisting of interconnected nodes that process information and learn patterns from historical data. In the realm of predicting prices, neural networks excel at capturing complex relationships and non-linear dependencies within vast datasets.

The ability of neural networks to adapt and optimize their parameters training makes them particularly well-suited for predicting price movements. By ingesting historical price data and relevant features, neural networks can learn intricate patterns, identify trends, and adapt to changing market conditions. This adaptability is crucial in the dynamic and ever-evolving world of financial markets.

Neural networks are employed in various financial applications, ranging from stock and commodity markets to cryptocurrency and foreign exchange. Their capacity to analyse large datasets and discern subtle patterns enables them to uncover hidden correlations and exploit predictive signals that may elude traditional forecasting methods.

When it comes to price prediction for accommodations neural networks have been deployed with relative success in Airbnb or other rent datasets: Peng, et al. in “Leveraging Multi-Modality Data to Airbnb Price Prediction” [6] used a deep neural network with two hidden layers having nine and ten neurons each with ReLU activation function; the methodology of the paper had them using Principal component analysis and selecting six components as inputs for the algorithm, overall the model compared similarly to XGboost performing better than linear counterparts.

However, in another paper from Ahuja et al. “Predicting Airbnb Rental Prices Using Multiple Feature Modalities” [7] by using a similar approach on multi-modality data found how even with no hyper parameter tuning LightGBM and XGboost outperformed both Ridge regression and DNN.

## Multi-modality data to predict rent prices

Previous research has been optimistic on the use of external and multi-modality data to enhance and better predict the prices of an accommodation: Peng, et al. in “Leveraging Multi-Modality Data to Airbnb Price Prediction” [6] employed a wide range of data inputs for their predictions demonstrably reaching higher accuracy than single-type data; The findings of this study indicate that customer reviews, house features, and geographical data serve as effective predictive factors for Airbnb rentals.

Ahuja et al. followed a similar methodology in “Predicting Airbnb Rental Prices Using Multiple Feature Modalities” [7] this work comprehensive preprocessing steps for a rental price dataset, focusing on geospatial, visual, categorical, numerical, and temporal features. The exploration of geospatial features involves reverse-geocoding and hierarchical clustering, emphasizing the significance of neighbourhoods. Visual features, incorporating listing images, were explored using a neural network but were ultimately deemed non-informative for modelling. Categorical features were managed through one-hot and label encoding to prevent sparseness. Numerical features underwent scaling, outlier handling, and a log transformation for normalization. Temporal features were introduced to capture host experience and listing newness. This preprocessing framework serves as a valuable contribution to understanding and modelling rental prices, especially considering its emphasis on diverse feature types.

## Environmental noise and air quality on accommodation price

Many papers highlighted the effect (mostly negative) that environmental noise and air quality has on rent prices, however not much literature has been explored for investigating the impact of noise on tourism accommodation specifically, Kemiki et al. in “THE IMPACT OF NOISE AND DUST LEVEL ON RENTAL PRICE OF RESIDENTIAL TENEMENTS AROUND LAFARGE CEMENT FACTORY IN EWEKORO TOWN, NIGERIA” [8] utilizes a hedonic model to assess the influence of noise and dust levels on the rental prices of residential properties in Ewekoro, Nigeria, with a focus on the impact of a nearby cement factory. The findings indicate that both noise and dust levels significantly affect housing rents, The study suggests that tenement rents decline with increasing distance from the cement factory due to the severity of dust and noise; These findings underscore the importance of addressing environmental factors in housing markets and their broader implications for community well-being and development.

The impact on noise in rent and property prices have been validated all across the globe with different research for many cities: as expressed in “The Influence of Traffic Noise on Apartment Prices on the Example of a European Urban Agglomeration“ by Szczepanska et al [9] “Acoustic discomfort cannot be fully eliminated in urban areas, but noise levels vary in different locations and, consequently, influence property prices.”

When it comes to the relation between property prices and noise in cities most of the research shows that one of the highest contributors for devaluation is in fact traffic noise: different papers evaluated the impact of traffic noise on prices in percentage points which may different from city to city.

Other than traffic noise, previous literature highlighted how other source of noise can have a negative impact on property prices and rents: according to Jun et al. “Noise Pollution Loss Value Evaluation of Railway Transportation Based on Hedonic Price Method- The Case of Taiyuan City” [10] There is a notable adverse correlation between residential unit prices in proximity to railways and the noise generated by railway traffic. The economic impact of traffic noise pollution is quantified at 50.8 yuan/(m2 ·dB), and its influence extends to approximately 500 meters.

### Room prices and points of interest

In the dynamic realm of the tourism industry, the determination of accommodation prices is intricately tied to the surrounding points of interest. These points of interest, ranging from cultural landmarks and natural attractions to vibrant urban centers, serve as key influencers that shape the perceived value of accommodations. The proximity and accessibility of these points of interest often become pivotal factors for travelers when choosing where to stay. Whether it's the allure of iconic landmarks, the tranquility of scenic landscapes, or the vibrancy of local communities, these attractions contribute significantly to the overall experience for tourists. As a result, accommodations strategically positioned in close proximity to points of interest not only enhance the desirability of the stay but also command a premium in pricing. Understanding how points of interest intersect with the pricing dynamics of tourism accommodations is essential for both industry stakeholders and travelers seeking an enriched and tailored experience.

A study conducted in Sanya city from Hu and Song “Analysis of Influencing Factors and Distribution Simulation of Budget Hotel Room Pricing Based on Big Data and Machine Learning from a Spatial Perspective“[8] explained how the spatial landscape is indeed fundamental for determing hotels and accommodations prices; the research used different machine learning models (XGBoost, Linear regression, random forest and Multilayer perceptron) and identified three main categories for points of interest and their impact on accommodation prices:

1. Traffic related features: density of parking lots, density of coach stations, density of bus stops
2. Business related features: density of restaurants, density of markets.
3. Public Service related features: density of schools, density of parks.

It is important to note that every city is unique, with distinct characteristics and points of interest. As such, any research in this field must consider the specific points of interest and contextual factors relevant to each city. The dynamic interplay between accommodations and points of interest requires a nuanced understanding of the local environment to accurately assess and predict pricing dynamics.

### Conclusions

Different methodologies and machine learning models have been tested with success for predicting the price of a property in Airbnb, however only few researchers have explored the impact of features such as environmental noise, air quality and closeness to points of interest; research from other fields demonstrated how these features can be effective for enhancing the accuracy of a price predictor.