Predicting Airbnb Rental Prices Using Machine Learning

Harideep Aepuri
Marketing and Data Science
Regis University
haepuri@regis.edu

ABSTRACT

Airbnb has also emerged as one of the most esteemed outlets in the short-term accommodation industry that allows property owners to transfer accommodations to tourists all over the globe. Setting a reasonable nightly rate of their listings on the Airbnb platform is one of the most difficult tasks of Airbnb hosts. The prices are determined by various variables that depend on other variables like the property size, location, amenities, host credibility, demand by guests, and neighborhood factors. Non-correct pricing could lead to the loss of income or low booking rates. Hence, it is of great importance to create a robust price prediction model that is supported by data.

This practicum project aims at creating a machine learning regression model that can forecast the Airbnb rental price based on publicly available listing information. The project was performed in several steps such as data collection and preprocessing, exploratory data analysis, feature engineering and feature selection, model development and training, and model assessment and performance adjustment. All the stages have led to constituting a total predictive model that could examine intricate pricing trends.

A number of regressions algorithms had been introduced and tested, which were the Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regression, Support Vector Regression, and Gradient Boosting Regressor. Gradient Boosting was the best predictor of these models. The resulting model after hyperparameter optimization and cross-validation had an R2 score of about 0.55 and the necessary levels of prediction errors. The findings imply that the ensemble learning techniques can be successfully used to identify non-linear correlations in Airbnb pricing information and deliver useful ideas to pursue data-driven pricing decisions.

I. INTRODUCTION / BACKGROUND

The fast development of online rental market places has revolutionized the hospitality industry around the world. Of particular importance is the fact that Airbnb has become one of the most potent short-term accommodation service providers. By use of this platform, hosts are able to advertise their own private rooms, apartments, and houses to travelers who want to temporarily stay in them. Nevertheless, defining the best price at which one can rent per listing is also a complicated matter.

Airbnb properties do not have common features, size, amenities, and location, unlike the hotels that usually have standardized pricing strategies. Besides, the perceived value of a listing may be affected by the preferences of the guests, local demand, and host credibility. Consequently, the use of manual prices due to pure intuition or competitor value cannot lead to the best possible results.

In order to meet this challenge, this practicum project uses machine learning on the data available on Airbnb listings to predict price of renting. The project goes through a formal data science workflow in several phases, where the exploration, modeling, and validation of predictive algorithms are systematically performed. Through the implementation of sophisticated analysis tools, the research will determine what can be considered the most important/matter of pricing and come up with a valid model that can be used to make price-related decisions based on the available data [1].

II. PROBLEM STATEMENT

The main goal of this practicum project is to create a predictive machine learning model that can estimate Airbnb rental prices on the background of listing attributes. Short-term rental pricing is dictated by many interacting factors hence pricing is hard to estimate through the application of traditional statistical models.

Numerous Airbnb hosts depend on the manual comparisons with the nearby listings or the basic heuristic approaches to set prices of rent. Nonetheless, these methods usually overlook the latent correlation between variables like guest capacity, types of property, and reputation of the host. As a result, pricing choices can be not attached to market realities or property value.

This project helps solve this issue, and it analyzes the data of the Airbnb listing using a structured machine learning pipeline. The method combines preprocessing data, feature exploration, feature engineering, regression, and feature performance. The project will aim at determining an efficient predictive model that can produce estimates of rental price using the evaluation of various algorithms and model parameter optimization [10].

III. METHODOLOGY/PROPOSED APPROACH

DATA COLLECTION AND PREPROCESSING

In the project first step is to do data gathering and data preparation to facilitate analysis. The data has been taken by publicly available Airbnb listing websites, which offer a lot of details about the attributes of the property, the attributes of the host, the reviews of the guests, and availability.

The raw data viewed had many columns that could not necessarily be explicitly put into a predictive model, such as identifiers, URLs, images, and long descriptive texts. These were columns removed off in order to minimize noise and computational efficiency. The price column also needed to be further cleaned up since there were currency symbols and formatting characters contained here. These marks were eliminated, and the values were represented as numbers to allow math operations.

The other important measure that was classified under critical steps was to identify and eliminate unrealistic price entries. All the listings whose price was zero or negative were not included in the dataset, as such a value is not a legitimate rental transaction. The Interquartile Range (IQR) approach was also employed to solve outliers because the high or very low values of prices might skew the results of model training [6].

All missing values were addressed with the right imputation strategies. The median values were entered into numerical columns, and the placeholder categories were entered in the categorical columns when the need arose. The types of rooms and properties used were categorical in nature and turned into numeric values by use of one-hot encoding. Lastly, the use of standardization methods to scale the features was done to ensure a normal magnitude of the variables. The following processing steps enabled the raw data to be converted to a clean and structured format that can be analyzed with machine learning [5].

EXPLORATORY DATA ANALYSIS

The Exploratory Data Analysis was intended to learn how the dataset works and what its internal structure is EDA, which is important in the process of determining patterns, relationships, and possible anomalies in data prior to the development of models.

Histogram representations helped to understand the distribution of the target variable, price. The distribution analysis showed that the distribution was skewed right i.e. majority of the listings are concentrated in the moderate price range, and few luxury listing have high prices that are significantly higher.

A correlation analysis was carried out to understand the correlations between features and rental prices. The analysis has revealed that the variables accommodates and the number of bedrooms have positive relationship with price. This implies that the bigger the properties, the higher the number of guests it is able to accommodate, and the higher the rental price.

Additional data was analyzed to determine the impact of room type on pricing. Listings that were whole homes or apartments tended to have an average price that was relatively higher than that of private rooms and shared places. The analysis of property type depicted that the rental value of villas and large houses is high as compared to small apartments or studio housing [7].

There was also the investigation of host-related attributes. According to the findings, such features like the host response rate, host acceptance rate, and verification status appeared to have the moderate effect on pricing; thus, the host credibility can probably impact the perceived listing value. These thoughts were used to influence feature engineering and model development phases of the project [3].

FEATURE ENGINEERING AND FEATURE SELECTION

The predictive capability of the dataset using techniques of feature engineering and feature selection. The feature engineering is either changing existing variables or forming new features that are more representative of underlying patterns in the data.

Calculation of host experience was one of the engineered features. The initial dataset also had a column that pointed to the date a host had joined the platform. This data was translated into a numerical characteristic that is the number of years in which the host has been active. This was transformed to enable the model to reflect the impact of host experience on pricing.

The other engineered functionality consisted in determining the host verification count. The types of the verification done by the host were included in the column of the dataset. The entries were transformed into a numerical value to show the number of verification methods that each host was involved in [8].

The process of feature selection was also carried out to get rid of the unnecessary variables. Columns that had too many missing values were removed due to the fact that they may introduce noise and make the predictive performance less accurate. Also, the ones that contained raw textual data like amenities and descriptions were taken out because of their unstructured nature. These measures allowed to make sure that the end dataset included only the variables that are relevant and meaningful and can be used in the regression modeling [4].

MODEL DEVELOPMENT AND TRAINING

Training various regression models to forecast prices of Airbnb rentals. Supervised regression techniques were employed due to the nature of the target variable, which is a continuous numerical value.

The dataset has been split into training and testing data with 80 percent and 20 percent split, respectively. Predictive models were constructed using the training data, and the performance of the models was tested using the testing data.

Some of the regression algorithms were put in place in this phase, which include Linear Regression, ridge regression, Lasso Regression, decision tree regressor, random forest regression, support vector regression, and gradient boosting regressor. The prepared set of features was used to train each of the models and assessed by standard regression metrics.

Mean Absolute Error, root mean squared error, and the coefficient of determination (R 2) were used to measure model performance. The gradient boosting regressor recorded the highest R2 score of all models evaluated and it exhibited high predictive ability [2].

MODEL ASSESSMENT AND PERFORMANCE ADJUSTMENT

The assessment of model reliability and refining its performance by use of validation and tuning methods. There was also cross-validation, as the performance of the model relies not on train-test split.

The Gradient Boosting model was cross-validated 5 times and yielded the same R2 scores. This meant that the model would be generalized well to unseen data, and it does not outfitt severe overfitting [9].

Then, tuning using GridSearchCV was done on hyperparameters. The combination of a number of parameters, such as learning rate, number of estimators, and tree depth were tested to find the combination of parameters that gave the best predictive capability.

Optimized model got improved predictions and a lower value of error after tuning. This step proved that model performance can be considerably increased when care is taken to optimize the parameters of the models [9].

IV. EXPECTED OUTCOMES

The ultimate result of the given project is the creation of a predictive model that will be able to approximate the price of an Airbnb rental based on listing features. The model offers significant details on the factors that have the greatest effect on the prices of rentals, such as the capacity of the property, the type of room, and the type of property.

Besides coming up with a successful predictive model, the project has shown how machine learning techniques can effectively be applied in a business environment. The results can guide Airbnb hosts in their pricing decisions and will be valuable in the research about predictive analytics in the sharing economy on a wider scale.

V. PROJECT TIMELINE

The project had an implemented schedule taking place through multiple phases. Week 2: Data collection and preprocessing. Week 3: Exploratory data analysis. Week 4 was when feature engineering and selection were carried out. Week 5 was used to model development, and Week 6 was used to model assessment and optimization of performance. Last week was spent on the compilation of the findings and a drafting of the final practicum report.

Week	Planned Activity
Week 2	Data collection and preprocessing
Week 3	Exploratory Data Analysis and visualization
Week 4	Feature engineering and feature selection
Week 5	Model development and training
Week 6	Model assessment and performance adjustment
Week 7	Final report preparation and writing

VI. CONCLUSION

This practicum project has been able to develop a full machine learning pipeline to forecast Airbnb rental prices. The project proved the efficiency of the ensemble learning techniques in the process of including sophisticated all-datum trends of prices through its numerous stages of data preparation, data analysis, and value modeling.

Gradient Boosting Regressor appears to be the most efficient model because it demonstrated a moderate yet significant predictive accuracy. Further optimizations to model reliability and minimize the overfitting risk included cross-validation and optimization of hyperparameters.

Even though the model accounts for about 55 percent of the variation of the rental prices, improvement in the future may involve more variables like seasonal demand, local events, and sentiment of textual reviews. Although limited, the paper identifies that machine learning can be used to facilitate data-driven strategies in pricing in the short-term rental sector.

References

1. Alharbi, Z. H. (2023). A sustainable price prediction model for Airbnb listings using machine learning and sentiment analysis. *Sustainability, 15*(17), Article 13159. https://doi.org/10.3390/su151713159

2. Álvarez-Herranz, A., Macedo-Ruíz, E., & Quiroga, E. (2025). Determinants of the price of Airbnb accommodations through a weighted spatial regression model. *Sustainability, 17*(21), Article 9364.

   https://doi.org/10.3390/su17219364

3. Camatti, N., di Tollo, G., Filograsso, G., & Ghilardi, S. (2024). Predicting Airbnb pricing: A comparative analysis of artificial intelligence and traditional approaches. *Computational Management Science*.

   https://doi.org/10.1007/s10287-024-00511-4

4. Ganta, N., & Krishna, M. (2024). Price prediction and recommendation of Airbnb property listings using machine learning. *SSRN Electronic Journal*.

   https://doi.org/10.2139/ssrn.5214669

5. Kanakaris, N. (2023). Predicting prices of Airbnb listings via graph neural networks and document embeddings. *Procedia Computer Science, 219*, 705–712.

   https://doi.org/10.1016/j.procs.2023.01.342

6. Costa, J. R. G. A. (2025). Predicting Airbnb prices using machine learning algorithms (Doctoral dissertation, *Instituto Superior de Economia e Gestão)*.

   https://repositorio.ulisboa.pt/entities/publication/6de62c0f-df4e-406d-8a50-81fc3824426a

7. Pittala, T. S. S. R., Meleti, U. M. R., Vasireddy, H. (2024). The unwilling patterns in European Airbnb prices: a study of using machine learning techniques. *arXiv Preprint*.

   https://arxiv.org/abs/2407.01555

8. Safari, A. (2025). Airbnb reviews by Sentiment analysis: Exploring across multiple regions U.S. that impacts acceptance rates and prices. *arXiv Preprint*.

   https://arxiv.org/abs/2504.14053

9. Tan, H., Su, T., Wu, X., Cheng, P., Zheng, T. (2024). A sustainable rental price prediction model based on multimodal input and deep learning—Evidence from Airbnb.*Sustainability, 16(15), Article 6384.*

https://doi.org/10.3390/su16156384

10. Tang, J. (2024). Forecasting Airbnb prices through machine learning. *Managerial and Decision Economics, 45*(1), 148–160.

https://doi.org/10.1002/mde.3985