UNIVERSITY OF PISA

MSc in Artificial Intelligence and Data Engineering

Data Mining and Machine Learning

Data-Driven Exploration of Airbnb Listings Price

Technical Report

Christian Petruzzella

ACADEMIC YEAR 2023/2024

Contents

Chapter 1

Introduction

Airbnb has brought significant changes to the hospitality industry worldwide. Experiencing remarkable growth, it currently offers over six million listings in 191 countries across one hundred thousand cities. Airbnb has gained immense popularity among travelers seeking accommodations globally.

As the platform has grown, so has the complexity of its pricing system, with a multitude of factors influencing the listing price and understanding these factors is crucial for both hosts, seeking to maximize revenue, and guests, aiming to secure the best deal. Pricing rental properties on Airbnb still presents a challenge for owners, as it directly impacts customer demand.

Using publicly available data for listings in Barcelona, this study aims to develop a price estimation model using machine learning techniques to aid both property owners and customers with price evaluation given information about the property, including features of the rentals, owner characteristics and customer reviews.

The performance of different models is compared, based on metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Squared Mean Error (RMSE) and R^2 score. Additionally, feature importance analysis is carried out to highlight the factors that most contribute to the price variation, providing valuable information to hosts seeking to optimize their pricing strategies.

Chapter 2

Dataset

2.1 Data Collection

The data collected is taken from *insideairbnb.com*, a platform that provide tools and data to explore Airbnb usage worldwide. Specifically, the city of Barcelona has been selected for this study, and the data about *listings* (18925 entries and 75 features including lodging details, neighborhood information, host characteristics, availability) and *reviews* were downloaded. The two dataset provided include information up to date on 5 July 2024.

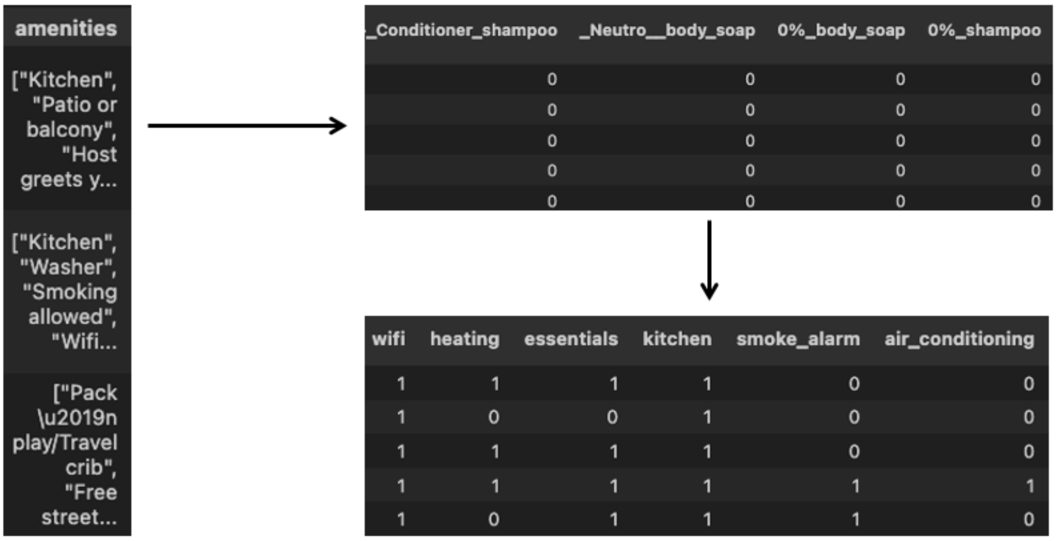
2.2 Preprocessing

Each feature in the *listings* dataset was carefully inspected:

- removed columns where all values were the same
- removed uninformative features like URLs (*listings_url, host_url*), host details (*host_id, host_name, host_about*)
- converted some features into binary format, such as *host_is_superhost, instant_bookable, host_has_profile_pic, host_has_identity_verified* (true/false values were converted into 1/0 representation)
- removed special characters like dollar signs and percentages from the *price* column and *host_response_rate*
- converted date columns such as *host_since* into numeric values representing the number of days since the date of reference (2024-07-04)
- handled categorical features using the *OneHotEncoder*, which converted them into binary columns
- filled missing values through *Linear Regression Imputation*: for each column with missing values a linear regression model was trained on the other available features to predict and fill the missing data.

2.2.1 Amenities Handling

The feature *amenities* had more than 2000 unique values: examination, cleaning and aggregation were performed and finally split into binary representation, indicating 1 if the particular amenity was included in the listing and 0 otherwise. At the end 25 amenities have been selected by leveraging insights from different research studies.



amenities
["Kitchen", "Patio or balcony", "Host greets y...
["Kitchen", "Washer", "Smoking allowed", "Wifi...
["Pack \u2019n play/Travel crib", "Free street...

_Conditioner_shampoo	_Neutro__body_soap	0%_body_soap	0%_shampoo
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

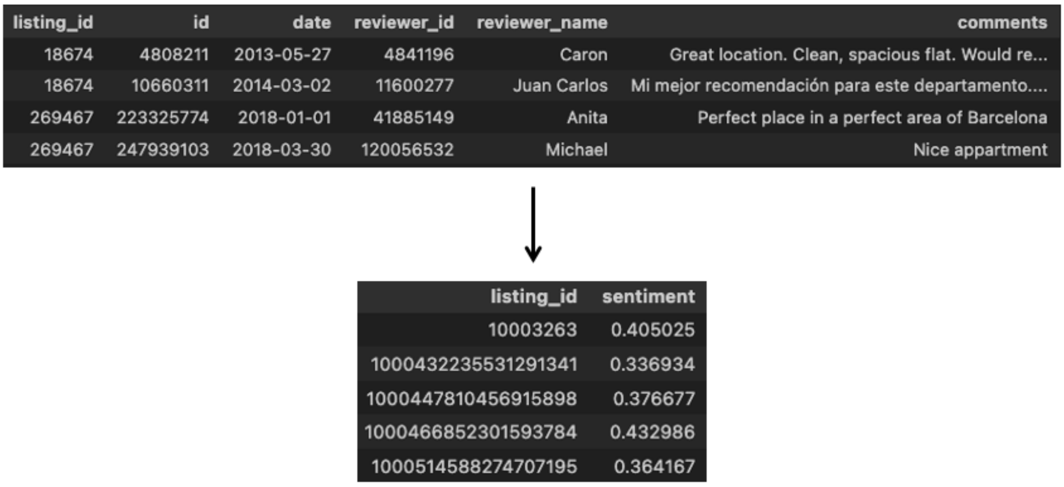
wifi	heating	essentials	kitchen	smoke_alarm	air_conditioning
1	1	1	1	0	0
1	0	0	1	0	0
1	1	1	1	0	0
1	1	1	1	1	1
1	0	1	1	1	0

2.2.2 Sentiment Analysis on the Reviews

The *review* dataset contained reviews for listings in Barcelona. The following initial operations were performed on it:

- unnecessary columns to the sentiment analysis were removed, such as *id, date, reviewer_ id, reviewer_ name*
- since the dataset contained reviews in multiple languages, English reviews were detected using the *langdetect* library, while all the others have been pruned
- Sentiment analysis was then performed using the *Textblob* library which allowed to calculate the *sentiment polarity* of each review: a value between -1 and 1, where negative values indicate negative sentiment and positive values indicate positive sentiment
- the *polarity* values have been grouped by *listing_ id* and the average sentiment score was calculated for each listing.

The final preprocessed review dataset has been merged with the main listings data.



listing_id	id	date	reviewer_id	reviewer_name	comments
18674	4808211	2013-05-27	4841196	Caron	Great location. Clean, spacious flat. Would re...
18674	10660311	2014-03-02	11600277	Juan Carlos	Mi mejor recomendación para este departamento....
269467	223325774	2018-01-01	41885149	Anita	Perfect place in a perfect area of Barcelona
269467	247939103	2018-03-30	120056532	Michael	Nice appartment

listing_id	sentiment
10003263	0.405025
1000432235531291341	0.336934
1000447810456915898	0.376677
1000466852301593784	0.432986
1000514588274707195	0.364167

Chapter 3

Experiments and Discussion

3.1 Regression Models Evaluation

A 10-fold cross-validation of the data has been performed with various regression approaches: Lasso, Ridge Regression, Bayesian Regression, K-Nearest Neighbors, Decision Tree, Random Forest and XGBoost.

To assess the performance of the regression models, different evaluation metrics have been calculated:

- *Mean Squared Error (MSE)* measures the variability of the residuals, indeed the distance between the data points and the regression line

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- *Mean Absolute Error (MAE)* measures the average of the absolute differences between each prediction and its corresponding true observation

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- *Root Mean Absolute Error (RMAE)* indicating how tightly the data points are clustered around the line of best fit

$$\text{RMAE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}$$

- *R^2 score* serves as indicator of how well the model fits the data

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

For MSE, MAE and RMSE, smaller the value and higher the prediction quality of the model, while R^2 score ranges between 0 and 1, with value closer to 1 implying a better fit.

3.2 Results

The following table shows the analysis of the effects of applying various regression algorithms to the preprocessed dataset.

Model	**MSE**	**MAE**	**RMSE**	$\mathbf{R^2}$
Lasso	0.00037	0.0079	0.01823	-0.0011
Ridge	0.00027	0.00500	0.01501	0.3635
Bayesian Regressor	0.00027	0.00502	0.01501	0.3627
K-Nearest Neighbors	0.00025	0.00475	0.01474	0.3651
Decision Tree	0.0002	0.00513	0.01342	0.3947
XGBoost	0.00018	0.00367	0.01211	0.5611
Random Forest	0.00016	0.00334	0.01163	0.5443

Lasso has the worst R^2, which means that the model fails to explain any variance in the data, likely due to its feature selection approach, which may have removed too much information.

Ridge seems to handle better the multicollinearity, capturing more variance (R^2 of 0.364) and the Bayesian regressor performs almost identically.

Tree-based ensemble models perform better than linear models suggesting that the relationships between the *price* and the other features are complex and non-linear: XGboost and Random Forest have the highest R^2 (0.561 and 0.544 respectively), indicating that ensemble methods are better at capturing complex patterns and interactions. However XGBoost provides a slight edge in explained variance.

3.2.1 Residuals Analysis

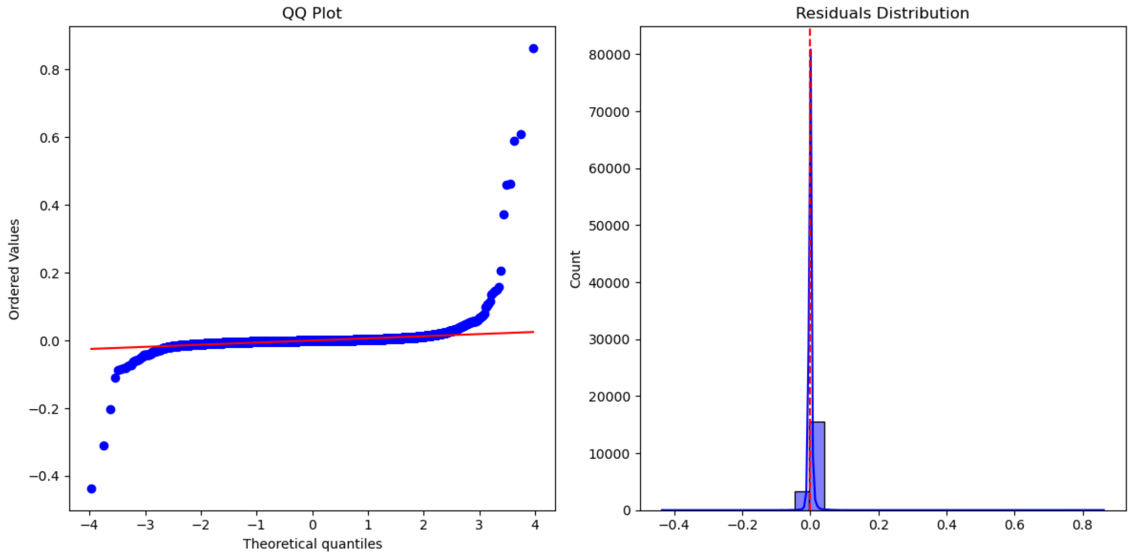


Figure 3.1: Residuals (XGBoost)

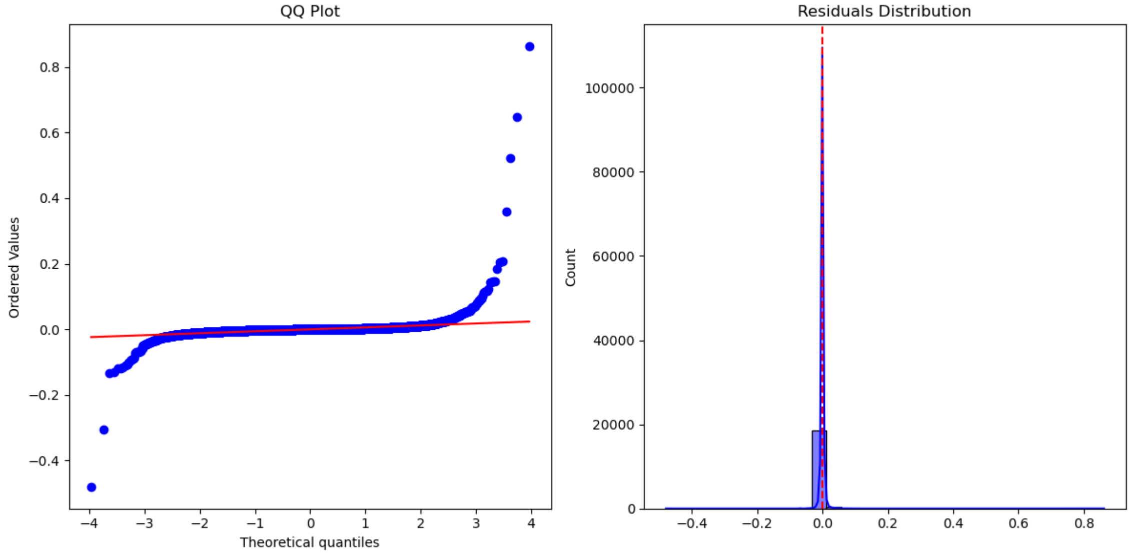


Figure 3.2: Residuals (Random Forest)

Both for the XGBoost and Random Forest models the QQ plot shows that unlike the center where the residuals are around the mean, there is a deviation at both left and right tails, indicating that the residuals do not follow a normal distribution, confirmed by the Shapiro-Wilk test.

The distribution of residuals is centered close to zero, suggesting that the predictions are mostly accurate. Although there are some outliers where the residuals deviate significantly from zero, the models well-behaved.

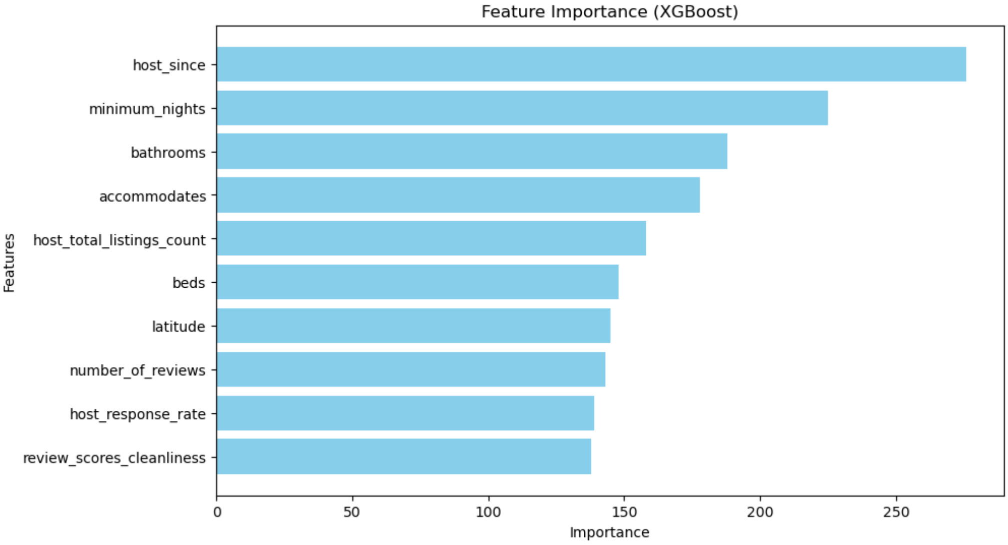
Moreover, the Wilcoxon test reveals that despite the slight difference between the performance of the models, the residual distribution varies significantly, indicating that their errors are not distributed in the same way: while Random Forest minimizes errors slightly better (lower MSE, MAE and RMSE), XGBoost explains more variance (higher R^2) indicating a better fit to the data.

3.2.2 Feature Importance Analysis

The importance of a feature reflects how much it contributes to the model's prediction, with higher values indicating greater importance.

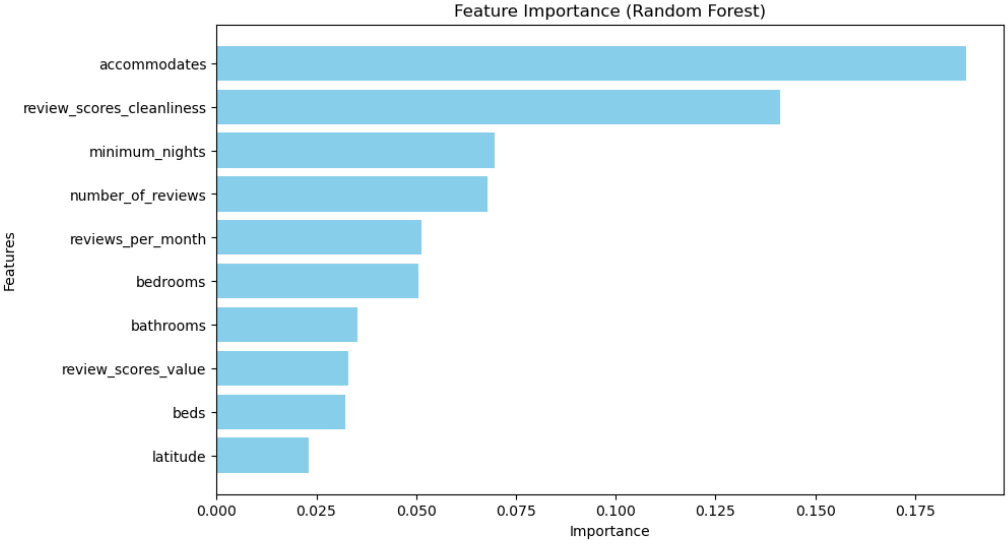
In the XGBoost feature importance analysis, the *weight* of the features has been used: it is an indicator of the frequency of how many times a feature is used to split the data across all trees in the model. The top features are:

- *host since* and *host total listings count*: since when the host is active and the number of listings managed, indicating potentially experience and reliability
- *minimum nights*: the longer minimum stay in terms of number of nights
- *beds* and *bathrooms* correlated with the property size
- *accommodates*: number of guests that a listing can host
- *number of reviews*, which reflects the popularity and trustworthiness of the listing
- *host response rate*: the promptness of the host in responding to booking requests and messages
- *review cleanliness score*: impacting the price as correlated with guest satisfaction.



Random Forest assigns importance based on variance reduction, meaning that features that reduce prediction errors across many trees are considered more important. Feature importance is calculated on how much each feature decreases the impurity across all trees in the model. The impurity is a measure of how well a decision tree split divides the data into groups with similar target values: the reduction in impurity is obtained by summing across all trees and normalizing the result in order to provide the importance score for each feature.

Accommodates is the most influential feature, indicating that it is the major factor in pricing, followed by the cleanliness rating. Beside the mutual important feature with the XGBoost model, other features that the most influence the price increasing are the *number of reviews per month*, and the *review value score*, which captures the guests perception of value for money.



In conclusion both models agree that the property size (accommodates, beds, bedrooms), guest feedback (number of reviews, review score) and minimum stay requirements emerge as the most important factors in determining listing prices.

Chapter 4

Conclusion

In this project various machine learning models have been explored in predicting Airbnb listing prices using features including property characteristics, host information and guest reviews, with the goal of assisting Airbnb hosts in determining the optimal price for their listings.

Through rigorous analysis and model evaluation based on metric such as MSE, MAE, RMSR and R^2 score, Random Forest and XGBoost emerged as the best models, with Random Forest model exhibiting slightly lower MSE and MAE, while XGBoost had a marginally better R^2 score. The results indicated that while both models performed well, however the Wilcoxon test revealed a statistically significant difference between the residuals of the two models.

Additionally, feature importance analysis has been carried out to highlight the key drivers of price variation, indeed the number of guests accommodated, the number of minimum nights and property cleanliness ratings. These insights provide valuable guidance for hosts looking to adjust their pricing strategies based on property features and guest reviews.