Decoding the Boston Airbnb Market

Leveraging Data Science to Optimize Pricing Strategies for Hosts and Travelers

David Mulas Zelenskaya
Data Science
Wentworth Institute of Technology
Boston, MA, USA
mulaszelenskayad@wit.edu

ABSTRACT

The main objective of this project is to find what sets the price of Airbnb rentals in Boston to understand the dynamics of modern economy. This study aims to identify key pricing drivers, like property characteristics, location and hosts.

Utilizing the dataset from InsideAirbnb the analysis uses Multiple linear regression and Random Forest Regression to predict prices the results suggest that location and room type are the main predictors, Randon Forest regression outperforms linear regression, explaining the complex markets of Real State.

KEYWORDS

Airbnb, Regression, Real State.

1 Introduction

Short-term rentals are one of the major pieces of modern economies, making it very important the pricing factors that take place when seeing an Airbnb rental.
For hosts understanding this factors is essential for maximizing revenue and for travelers understanding this factors can be crucial to find better deals.
By investigating factors like location, property features and who is hosting you at the rental, this project aims to answer the valuation of these rentals.

2 Data

https://data.insideairbnb.com/united-states/ma/boston/2025-09-23/data/listings.csv.gz

2.1 Source of dataset

The data is from a website called inside Airbnb, an independent dataset that provides data about Airbnb rentals in multiple cities around the world, their main objective is to find the impact of short-term rentals in communities.

2.2 Characters of the datasets

This data set is in a CSV file, it contains around 4000 rows(listings) and 75 columns(variables)\
For this project I cleaned the data to remove outliers and format the price as a continuous float, the main variables selected for this models are the following:
Target,: Price(Continuous USD)
Location: neighborhood (categorical)
Property: room type (Entire home, Private room)
Reputation number of reviews and if the host is a super host

Figure 1 Distribution of Airbnb Prices in Boston (Sept 2025).

As illustrated on figure 1 the prices in Boston range between 50$ to 1400$ with the majority being in the $100 to $250 range per night, I deleted extreme listing above $1500 to avoid any distortion.

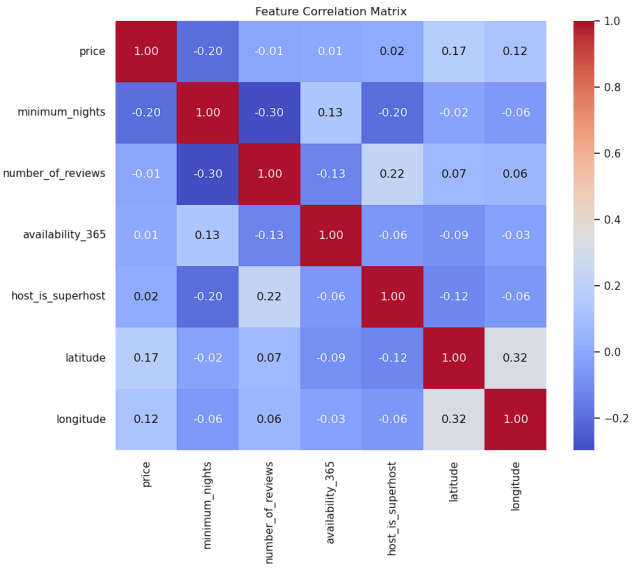

Figure 2 Correlation Matrix of Key Variables.

As shows in Figure 2 the correlation matrix shows key relationships. Weak positive relations between price and number of reviews suggesting that the price is not a factor in the number of bookings, also room type shows a strong negative relation with price suggesting that "Private Rooms" and "Shared Rooms" are cheaper than "Entire Homes.".

3 Methodology

To answer this question about pricing, I will be using quantitative modeling. Price is continuous making regression a decent approach. Two different algorithms will be used one for interpretability and another one to predict.

3.1 Multiple Linear Regression

MLR is a statistical technique that uses different variables to predict the outcome of a variable.

This model assumes linear relationships between variables while it offers high interpretability it may underfit non-linear data and is sensitive to outliers

3.2 Random Forest Regression

It averages multiple decision trees, assumes the training data is representative of the population.

This model captures non-linear but has lower interpretability.

I selected this model to improve predictive accuracy, this model unlike linear regression can capture better non linear relationships.

4 Results

The dataset was split into two sets one with 80% of the data(training set) and the other 20% in testing set to test performance.

4.1 Model performance

Both models were tested using RSME and R-squared.
The random forest model outperformed the linear one and here are the results:

```
--- Linear Regression Results ---
RMSE: $139.54
R-squared: 0.3107
--- Random Forest Results ---
RMSE: $126.77
R-squared: 0.4310
```

The random forest model was able to explain 43% of the variance in pricing, MLR was only able to explain 31% of the variance.
This confirms that Airbnb is complex and has a lot of non-linear interactions.

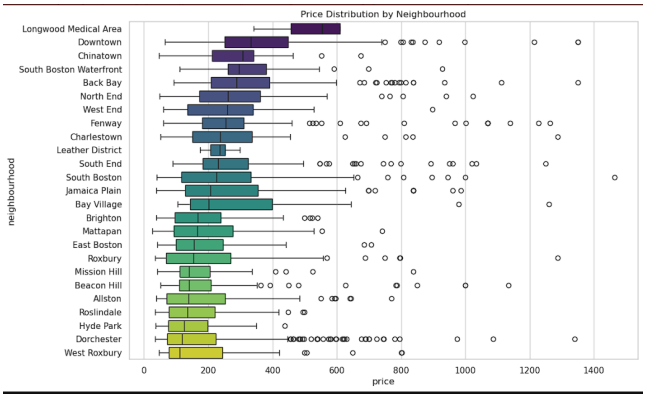

Figure 3 Price Range Variation by Neighborhood.

4.2 Feature importance

Random forest model reveals the most important drivers in price:

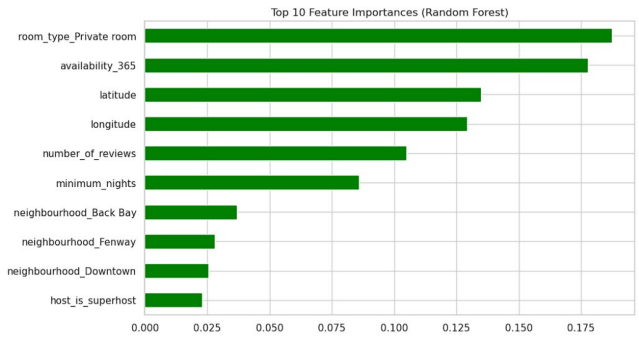

Figure 4 Top 10 Features Influencing Price Prediction.

As shows in Figure 3 room type, availability, location (neighborhood) and number of reviews are the top features**.**

5 Discussion

The Random Forest Regression was the superior model and this model suggested that "location" is not a simple variable but a complex one interacting with latitude, longitude and neighborhood that the linear model struggled to identify, the data showed that physical aspect like beds and room type sets the baseline price while the neighborhood is the main price factor that drives the price up or down

5.1 Limitations

Random Forest had an R-squared of 0.43 that tells me that over 50% of the price variations remains without an explanation, this could be intangible factors such as design, photo quality in the listings, or textual descriptions in the posting.

6 Conclusion

Thiis project helped me successfully identify the drivers in the Airbnb Boston prices using this datasets from September 2025.

The project proved that Random Forest Regression gives us a better estimator of the market value compared to linear methods.

The results suggest that for hosts, pricing depends heavily on location and room type, but to be able to maximize returns optimizing other factors that we cant measure with data like design and postings are even more important

REFERENCES

[1] Inside Airbnb. 2025. Boston Listings Data Snapshot: A Detailed Dataset of Short-Term Rentals in Boston. In *Inside Airbnb Data Repository*. Inside Airbnb, Boston, MA, USA. Retrieved December 7, 2025 from http://insideairbnb.com/get-the-data

[2] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, and Vincent Michel et al. 2011. Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research (JMLR)*, Vol. 12. Microtome Publishing, Brookline, MA, USA, 2825–2830.

[3] Leo Breiman. 2001. Random Forests: An Ensemble Learning Method for Classification and Regression. In *Machine Learning Journal*, Vol. 45, No. 1. Springer, New York, NY, USA, 5–32. DOI:https://doi.org/10.1023/A:1010933404324