Maximizing Revenue for Airbnb

Rebekah Sander   
*School of Data Science and Analytics*  
*Kennesaw State University*Kennesaw, United States of America  
rsander@students.kennesaw.edu

*Abstract*—This project aims to optimize the monthly revenue potential for Airbnb hosts with respect to their location. Utilizing a comprehensive dataset from 2017, encompassing 44,318 Airbnb listings situated in New York City, alongside 31 recorded features for each listing per month, the study employs variable clustering techniques to identify key predictors. An 80:20 data split facilitates the training and evaluation of three distinct regression models: multiple linear regression, Random Forest regression, and XGBoost for regression. The performance of each model is compared in terms of maximizing and minimizing root mean square error. XGBoost emerges as the most effective predictive model. This study underscores the importance of leveraging advanced analytical techniques to extract actionable insights, thus empowering hosts to make informed decisions in the hospitality industry.

Keywords—Airbnb, Revenue Optimization, Regression, Variable Clustering, Predictive Modeling

# Introduction

In the beginning stages of any business venture, the primary concern is centered on maximizing profit. In the hospitality space, particularly in Airbnb, the solution to this concern may lie in having the right proptery in the right location. However, is there more to increasing profit than just having the right location?

This study seeks to delve deeper into the factors that contribute to maximizing revenue for Airbnb hosts. While location undoubtedly plays a significant role, this analysis aims to explore whether hosts can take additional steps to enhance their earnings, regardless of their property’s location.

The goal of this study is to identify the most effective predictive model capable of accurately forecasting and optimizing monthly revenue for Airbnb hosts.

# Methods

In the initial phase of data exploration, the dataset comprised 44,318 observations and 31 features. Preprocessing procedures involved handling missing values, identifying duplicates, and evaluating data types for each observation. Additionally, to ensure uniformity and facilitate model training, a robust scaling technique was applied to normalize the data. Feature engineering procedures focused on assessing the importance and relevance of each feature. The resulting dataset consists of 23,595 observations and 6 features. This resulting dataset is then subjected to model training and evaluation using three distinct regression models: Multiple Linear, Random Forest, and XGBoost.

## Preprocessing

In the preprocessing phase, missing values, duplicates, and datatypes were identified. To address missing values, features with majority of missing instances were deemed unreliable and subsequently removed from the dataset. For the remaining missing values, rather than imputing values, observations containing any missing values within any feature were excluded from further analysis. Additionally, instances where the ‘availability\_30’ feature indicated that an Airbnb property was unoccupied for the entire month were filtered out to prevent skewing of the data. These approaches were deemed optimal given the size of the dataset. Duplicate entries were identified by evaluating entire rows or columns and subsequently removed to prevent duplication bias. Furthermore, discrepancies in data types, particularly instances where numerical values were classified as objects, were rectified to maintain consistency across the dataset. To mitigate the impact of outliers, a robust scaler was employed to normalize the data, thus enhancing robustness and stability in the modeling stage. Following these preprocessing steps, the dataset was refined to contain 23,595 observations, which proceed to the modeling stage.

## Feature Engineering

In the feature engineering phase, the target variable, 'relative\_revenue,' was constructed using three key variables--'price,' 'availability\_30,' and 'neighbourhood\_cleansed.' First, monthly revenue had to be calculated. Since ‘availability\_30’ provides information on the availability within a given month, these values were subtracted from 30 to determine the number of nights the Airbnb property was occupied. These values were then multiplied by the corresponding ‘price’ value to derive the monthly revenue.

Recognizing the potential impact of location on revenue, ‘relative\_revenue’ is created to standardize revenue relative to the neighborhood. This was achieved by calculating the mean monthly revenue for each neighborhood and then dividing each observation’s monthly revenue by its neighborhood’s mean monthly revenue. Subsequently, all other categorical features were omitted from further analysis.

To identify the most influential features for modeling, variable clustering was employed. Through this process, six key features were identified as significant contributors in predicting relative revenue. These six features were retained for modeling:

1. *‘relative\_revenue’;* The monthly revenue of an Airbnb relative to its location; (%)
2. *‘accomodates’;* How many people the Airbnb can accommodate; (people)
3. *‘number\_of\_reviews’;* The number of reviews the Airbnb received in a month; (reviews)
4. *‘review\_scores\_rating’;* The average review score rating for the Airbnb in a month; (%)
5. *‘host\_reponse\_rate’;* The percentage of times the host responds to inquiries out of all potential responses; (%)
6. *‘bathrooms’;* The number of bathrooms at the Airbnb; (bathrooms)

Finally, before proceeding to the modeling stage, the dataset was divided into training and testing sets to facilitate effective model evaluation. This analysis adopted an 80:20 ratio for data splitting, with 80% of the data allocated for training the models and the remaining 20% reserved for testing model performance.

## Modeling

In the modeling phase, predictive models were trained using both explanatory and response training data, followed by testing to assess their performance. Each model generated predictions for the explanatory test data, which were then compared to the actual test response data.

To identify the most effective predictive model, three distinct regression techniques were employed and analyzed using performance metrics of and the root mean square error. These models were selected for their versatility and capability to handle various types of data.

1. *Multiple Linear Regression;* This model was implemented using a package within the scikit-learn library, leveraging its robust regression capabilities to predict the target variable.
2. *Random Forest Regression;* This model was also implemented using a package within the scikit-learn library. This ensemble learning technique aggregates the predictions of multiple decision trees to produce more stable results.
3. *XGBoost Regression;* This model was implemented using a package within the XGBoost library. Known for its efficiency and performance, XGBoost employs gradient boosting algorithms to iteratively refine the predictions, resulting in improved accuracy.

# Results

## A screenshot of a graph Description automatically generatedCorrelations

Analysis of the correlation heatmap reveals that the variable 'accommodates’ has the highest correlation with the response variable ‘relative\_revenue.’ Specifically, there is a moderate, positive, linear correlation between how many guests an Airbnb can accommodate and its relative revenue. This finding underscores the significance of ‘accomodates’ in predicting revenue, suggesting that properties with greater capacity tend to generate higher revenue. with relative revenue.

## The Best Model

Table 1 presents the performance metrics of the three regression models. Based on the evaluation criteria of and root mean square error (RMSE), the XGBoost Regression model emerges as the optimal choice, exhibiting the highest and the lowest RMSE among the three models assessed.

It is noteworthy that while XGBoost outperforms the other models, the comparison of performance metrics across each model indicate similar predictive capabilities.

## A graph with blue dots and a line Description automatically generatedPredictions

Figure 2 displays the scatter plot of actual versus predicted values, providing insight into the predictive performance of the XGBoost model. The plot reveals a consistent trend where the predicted values tend to underestimate the actual values. This observation aligns with the performance metrics of the XGBoost model.

Several factors may contribute to this phenomenon, including the presence of high leverage or influential observations within the dataset. Further analysis may be warranted to identify and mitigate the impact of these observations, thereby enhancing the accuracy and reliability of the predictive model.

# Discussion

* Note that ‘neighbourhood\_cleansed’ was the only categorical variable used in this analysis. Though it was not directly in the model, using this variable to calculate ‘relative\_revenue’ was imperative. Location is everything when owning a business. Different locations in New York City will not only be worth different amounts, but also will have different demands. By adjusting monthly revenue based on neighborhood averages, we could better compare performance across locations.
* Initially, all available variables were considered to determine the best model with the right settings. However, the resulting models showed similar performance in not providing significant insights. With more time and expertise, a more comprehensive model considering all variables might have been more fruitful.
* One limitation of this analysis is its potential lack of generalizability to other locations. Different geographic regions exhibit unique characteristics and attractions that can profoundly influence revenue dynamics. To enhance the applicability of future studies, incorporating a more geographically assorted dataset may provide more useful insights.

# conclusion

This study aimed to investigate the factors influencing monthly revenue optimization for Airbnb hosts and identify the most effective predictive model for this purpose. Through comprehensive data analysis and modeling, key insights include:

* Significance of accommodation capacity: The analysis highlights the importance of the capacity of Airbnb property to accommodate guests in influencing monthly revenue. Properties with higher accommodation capacity tend to generate higher revenue. For a host, this could look like adding more beds, bigger beds, rollaway beds, and ideas of that nature.
* XGBoost, the optimal predictive model: The model’s superior performance highlights its potential as a valuable tool for Airbnb hosts seeking to optimize their revenue generation strategies.

In light of these findings, it is evident that predictive modeling offers valuable opportunities for Airbnb hosts to optimize their revenue generation strategies and adapt to the dynamic environment of the hospitality industry. By leveraging analytical techniques, hosts can make informed decisions to maximize their profitability and competitiveness in the market.

##### References

1. ivanovskia1, “Maximize value of Airbnb rental,” Kaggle, https://www.kaggle.com/code/ivanovskia1/maximize-value-of-airbnb-rental/input (accessed May 6, 2024)