

Rio De Janeiro – Airbnb Recommendation System

Sumair Rathore | COMP 4311 | April 17, 2024

## Abstract

**Background:** With the growing popularity of Airbnb as an alternative accommodation option, the need for effective recommendation systems tailored to specific locations has become paramount. In this project, I focus on building a recommendation system using information about Airbnb listings in Rio de Janeiro, Brazil. The goal is to provide personalized recommendations to users based on their preferences and requirements, thereby enhancing their overall experience and satisfaction with the platform. **Objectives**: My objectives include developing predictive models to recommend suitable Airbnb listings based on user preferences, historical booking patterns, and listing features. **Methods**: I utilize a dataset comprising over 70,000 Airbnb listings in Rio de Janeiro, encompassing various attributes such as property type, location, amenities, and pricing. Supervised learning techniques such as logistic regression, random forest, and neural networks are employed to train classifiers on the dataset. Additionally, I explore feature selection methods and evaluate model performance using accuracy, sensitivity, specificity, and area under the curve metrics. Results: My results demonstrate the effectiveness of logistic regression and random forest classifiers in predicting suitable Airbnb listings for users in Rio de Janeiro. **Conclusion**: The developed recommendation system shows promise in providing personalized recommendations to users, which will improve their overall satisfaction and experience with Airbnb in Rio de Janeiro. Further refinement and optimization of the system could lead to even more accurate and personalized recommendations in the future.

## **Introduction**

As global tourism continues to evolve, so too does the demand for personalized and enriching travel experiences. Among the countless accommodation options available, Airbnb has emerged as a leader, offering travelers a unique opportunity to immerse themselves in local culture and lifestyle. For adventurers, particularly those exploring vibrant destinations like Rio de Janeiro, the quest for suitable and safe accommodations becomes paramount. However, amidst the excitement of travel, concerns arise regarding the selection of an appropriate Airbnb listing that aligns with one's preferences and requirements. To address these concerns, this study aims to develop a recommendation system tailored to the unique context of Airbnb listings in Rio de Janeiro. By leveraging information about property attributes, location, and user preferences, the goal is to empower travelers with personalized recommendations that enhance their overall experience and safety during their stay in Rio de Janeiro.

## **Methods**

### **Data**

Airbnb has become a significant player in the hospitality industry, offering a vast array of accommodation options worldwide. For this study, I utilize a comprehensive dataset obtained from Airbnb listings in Rio de Janeiro, Brazil. The dataset comprises detailed information about over 70,000 Airbnb listings in Rio de Janeiro, including attributes such as property type, location, amenities, pricing, and user reviews. Additionally, the dataset contains demographic information about hosts and guests, such as host response time, host verification status, and guest review scores. The dataset was collected Kaggle.com and has been pre-processed to ensure consistency and accuracy. Each listing in the dataset represents a unique accommodation option available for booking through the Airbnb platform in Rio de Janeiro. This rich and diverse dataset serves as the foundation for developing my recommendation system, allowing us to analyze and extract valuable insights to enhance the travel experience for users exploring Rio de Janeiro.

## **Data Preprocessing**

* All data pre-processing and analyses were done in Python 3 (ipykernel). Firstly, I used just the Data Frame from April 2018 as an example to view the different data types and column names. Columns such as xl\_picture\_url, host\_neighbourhood, host\_listings\_count, and host\_total\_listings\_count were removed, as they provided no real useful information.

## **Outcome Operationalization**

In constructing my outcome variables for the Airbnb recommendation system, I encountered challenges due to the granular nature of the dataset and the low frequency of endorsement for specific behaviors or preferences. As such, I manually constructed outcome variables to capture relevant indicators that align with the objectives of my study.

**Objective 1:** Predicting Listing Suitability: The outcome variable for this objective is the suitability of an Airbnb listing for a given user. I consider various factors such as property type, location, amenities, pricing, and user reviews to determine the suitability of each listing. A listing is coded as 1 if it is deemed suitable for the user based on their preferences and requirements, and 0 otherwise.

**Objective 2:** Predicting User Satisfaction: The second outcome variable focuses on user satisfaction with Airbnb listings. User satisfaction is assessed based on factors such as guest review scores, host responsiveness, cleanliness, and overall experience. A listing is coded as 1 if it receives high satisfaction ratings from users, and 0 otherwise.

## **Feature Construction**

To select model features, I employed two approaches: manual construction and selection, and data-driven selection.

**Manual Feature Construction and Selection:** I examined various attributes of Airbnb listings, such as property type, location, amenities, and pricing, to identify relevant features that may influence listing suitability and user satisfaction. Features were manually constructed by grouping related attributes that contribute to the overall appeal and quality of a listing. For example, amenities related to safety and security (e.g., smoke detectors, fire extinguishers) were grouped together to create a composite feature. Additionally, features related to user reviews and host responsiveness were considered to gauge user satisfaction.

**Data-Driven Feature Selection:** In addition to manual feature construction, I utilized a data-driven approach to select optimal features for my recommendation system. I employed a recursive feature elimination (RFE) algorithm with logistic regression as the estimator to identify the most informative features that contribute to predicting listing suitability and user satisfaction. This approach prioritizes predictive performance and aims to optimize the accuracy of my recommendation system.

Overall, the combination of manual feature construction and data-driven feature selection allows us to leverage the richness of the Airbnb dataset and develop a robust recommendation system that effectively matches users with suitable and satisfying accommodation options in Rio de Janeiro.

## **Classification Techniques**

In my endeavor to develop a robust recommendation system for Airbnb listings in Rio de Janeiro, I compared three classification techniques for each objective: logistic regression, and random forest. Each technique offers unique advantages and insights, contributing to the overall effectiveness of my recommendation system.

**Logistic Regression:** Logistic regression is a well-established linear classification method that models the log-odds of a binary outcome as a linear combination of features. It handles categorical features well and provides interpretable output, making it suitable for my analysis.

**Random Forest:** Random forest is a powerful non-parametric tree-based classification method that excels in handling categorical features. It works by creating multiple decision trees during training and aggregates the results from these individual trees to make the final classification. This approach mitigates concerns of overfitting and high variance associated with single decision tree methods.

## **Results**

Sample Size:

The Kaggle dataset had a total of 758474 Airbnb listings. After the removal of duplicate listings from the df\_all, the new value of unique ID’s was brought down to a mere 73904.

## **Discussion**

In this study, I endeavored to develop a robust recommendation system for Airbnb listings in Rio de Janeiro by leveraging a diverse array of data science techniques and methodologies. My exploration encompassed various stages, from data preprocessing and feature engineering to model selection and evaluation. Here, I discuss my final thoughts on the insights garnered and the implications for the development of effective recommendation systems in the realm of Airbnb accommodations.

My model evaluation yielded promising results, with an overall accuracy of 0.62 on the testing set. The classification report further revealed insights into the model's performance, with precision, recall, and F1-score metrics indicating balanced predictive capabilities. Specifically, the precision and recall for both booking and non-booking classes underscored the model's ability to discriminate effectively between positive and negative instances. These findings highlight the potential of my recommendation system to assist users in making informed decisions regarding Airbnb accommodations.

The implemented Flask web application serves as a pivotal component of my recommendation system for Airbnb listings in Rio de Janeiro. The application seamlessly handles user interactions, database queries, and recommendation generation. Through intuitive routes and templates, users can effortlessly navigate the system, explore listings, and receive personalized recommendations based on their specified preferences. The integration of SQLite database ensures efficient data retrieval and management, facilitating the seamless delivery of Airbnb listings and recommendations to users. Additionally, the utilization of random selection logic enhances the user experience by offering diverse and unbiased recommendations tailored to specific neighborhoods. Overall, this Flask application plays a crucial role in empowering users with actionable insights and enhancing their decision-making process when selecting Airbnb accommodations in Rio de Janeiro.

**Data Preprocessing and Feature Engineering:** My initial step involved meticulous data preprocessing, including handling missing values, encoding categorical variables, and constructing relevant features. This process was crucial in ensuring the quality and integrity of the dataset for subsequent analysis. Additionally, I conducted manual and data-driven feature selection to identify informative predictors associated with booking likelihood.

**Model Selection and Evaluation:** I employed a range of classification techniques, including logistic regression, and random forest to develop predictive models for booking likelihood. Each technique offered unique advantages, with logistic regression providing interpretability, and random forest excelling in handling categorical features.

## **Limitations**

Despite my comprehensive approach, several limitations merit acknowledgment. The reliance on survey data and the assumptions made during modeling may introduce biases or limitations in generalizability. Additionally, the focus on Rio de Janeiro may limit the applicability of my findings to other regions or contexts. Future research could explore alternative data sources, such as user reviews and listing attributes, to enhance the predictive performance of recommendation systems.

A screenshot of a computer screen

Description automatically generatedA graph of a number of bedrooms and price

Description automatically generatedA graph showing a cluster distribution

Description automatically generatedA white screen with black text

Description automatically generatedA graph of a distribution of listing prices

Description automatically generatedA graph of a number of blue and black text

Description automatically generated with medium confidenceA graph of a number of items

Description automatically generated with medium confidence

## REFERENCES:

<https://www.kaggle.com/datasets/allanbruno/airbnb-rio-de-janeiro?select=fevereiro2020.csv>