Airbnb Room Price Prediction in Rio de Janeiro: A Machine Learning Approach

Introduction

Airbnb provides an innovative platform for individuals to offer lodging or rent short-term lodging in residences. In recent times, Rio de Janeiro has been a hotspot for Airbnb listings. Predicting the right price for such listings is crucial both for the host to maximize earnings and for Airbnb to ensure competitive pricing to drive more bookings. Given the intricacies and nuances in the pricing factors, from amenities to location proximity, predicting an accurate price poses a considerable challenge. This project employs machine learning regression models to predict Airbnb room prices in Rio de Janeiro, Brazil, based on a dataset spanning from August 2018 to May 2020.

Data Description

The foundation of any reliable machine learning model is its underlying data. For this project, we've sourced a comprehensive dataset covering Airbnb listings from August 2018 to May 2020. This dataset encompasses a total of 740K rows representing unique data for 65,920 listings, averaging 11 data points per listing. With an initial 108 variables, our data preprocessing steps were crucial. After rigorous cleaning - removing null entries, outliers, and irrelevant variables - we streamlined the dataset down to 400K rows and a manageable 9 columns. This refined data set forms the basis of our modeling.

Modeling & Analysis

Our aim was to implement and test various regression models, evaluate their performance, and select the most appropriate one based on different criteria:

1. **Model Accuracy:** The Gradient Boosting Trees Regression Model emerged as the top performer, boasting the lowest Root Mean Square Error (RMSE) value of 0.58.

2. **Execution Time:** Efficiency is key in real-world applications. Here, the Decision Tree Regression model proved superior, clocking the fastest execution time of just 6.95 seconds.

3. **Data Utilization:** Our analyses revealed that optimal predictions were achieved using only 20% of the data, emphasizing the quality of the data over quantity.

4. **Infrastructure Performance:** In terms of execution speed, Amazon EMR consistently outperformed local processing, highlighting the advantages of parallel processing capabilities.

5. **Data Consistency Across Nodes:** A striking observation was the consistent performance of the models across different nodes, underscoring the uniform density and quality of the data.

Conclusion

Our machine learning journey to predict Airbnb prices in Rio showcased the dominance of the Gradient Boosting Trees Regression Model in terms of prediction accuracy. However, if execution time is a priority, the Decision Tree Regression stands out. The experiment also highlighted the significant advantages of cloud computing platforms like Amazon EMR, especially in dealing with large datasets. Additionally, the data's consistent performance across various nodes emphasized its reliability and uniform quality. Overall, this study provides valuable insights and tools for hosts and Airbnb to optimize pricing strategies in Rio de Janeiro.