Analyzing Airbnb Listings in Seattle Using the CRISP-DM Framework

Abstract

This paper presents a data analysis of Airbnb listings in Seattle, WA, using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. The primary objective is to identify key factors that influence listing prices and provide insights to Airbnb hosts for pricing strategies. By utilizing machine learning models, specifically Linear Regression and Random Forest Regressor, we explore the relationships between features such as review scores, accommodations, and amenities, and their effect on pricing. The analysis reveals that features like bedrooms, bathrooms, and review scores significantly impact pricing, offering actionable insights for optimizing listing performance.

Introduction

In recent years, the short-term rental market has grown substantially, with platforms like Airbnb allowing property owners to list and manage rentals. For hosts, setting the right price is crucial for maximizing occupancy and revenue. This paper applies the CRISP-DM framework to analyze Airbnb listings in Seattle, Washington, to uncover the primary factors that affect listing prices. The insights derived from this study aim to help hosts make data-driven decisions on pricing and improve listing performance.

Methodology

The analysis follows the CRISP-DM methodology, which consists of the following phases:

1. Business Understanding

The objective is to identify factors that influence Airbnb listing prices. Our research questions are:

- - Which features have the most significant impact on listing prices?
- - How can hosts optimize these features to improve booking rates and profitability?

2. Data Understanding

We use three datasets to understand the listings and their characteristics:

- - Listings: Contains detailed information about each Airbnb listing, including prices, reviews, and amenities.
- - Calendar: Provides daily availability and price information for each listing.
- - Reviews: Contains reviews with details like comments, reviewer IDs, and dates.

By examining these datasets, we gain insights into the data structure, features available for analysis, and missing values. This step is essential for selecting relevant features and identifying potential issues with data quality.

3. Data Preparation

To prepare the data for analysis, we perform several cleaning and preprocessing steps:

1. 1. Price Conversion: We remove dollar signs ($) from the price columns and convert them to numeric values to allow quantitative analysis.
2. 2. Missing Values: We handle missing values in critical fields such as review scores and prices by filling them with median values where applicable.
3. 3. Feature Engineering: We create new features, such as extracting the month and year from the date, to allow for potential trend analysis.
4. 4. Merging Data: We merge the calendar and listings datasets to create a comprehensive dataset, facilitating a deeper analysis of how features relate to pricing.

4. Modeling

To explore how different features impact listing prices, we build predictive models using Linear Regression and Random Forest Regressor. These models help us understand relationships between features and provide a basis for assessing the impact of various attributes on price.

- - Linear Regression: A simple model that establishes a baseline by examining linear relationships between features and price.
- - Random Forest Regressor: A more complex model that captures non-linear relationships among features, providing deeper insights into feature importance.

These models are trained using key features such as review scores, number of accommodations, bedrooms, bathrooms, and the number of reviews. Each model is evaluated to assess its accuracy and reliability in predicting listing prices.

5. Evaluation

We evaluate the models using Mean Absolute Error (MAE) and R-Squared (R^2) metrics. The Random Forest model achieves a better MAE and R^2 score than Linear Regression, suggesting it captures more complex relationships between features and prices. The Random Forest model also provides feature importance scores, highlighting the most influential factors on listing prices.

Results

The Random Forest model outperforms Linear Regression in predicting Airbnb prices, as indicated by the MAE and R^2 scores.

Feature Importance

The most important features identified by the Random Forest model are:

- - Bedrooms: Listings with more bedrooms generally command higher prices due to their capacity to host more guests.
- - Bathrooms: The number of bathrooms is a significant factor as it directly impacts guest comfort, especially for larger groups.

- - Review Scores: Higher-rated properties are generally priced higher, reflecting guest satisfaction and host reliability.

These results suggest that hosts can focus on these key features when setting prices or improving their listings to attract more bookings.

Discussion

The findings provide several actionable insights for Airbnb hosts:

5. 1. Hosts can optimize pricing by focusing on critical features such as bedrooms, bathrooms, and positive reviews.
6. 2. Improving guest experience to boost review scores could allow hosts to charge higher prices, as listings with better reviews tend to attract more bookings.
7. 3. The model's performance could be further enhanced by including additional neighborhood-specific or property-type data, which could capture more localized pricing trends.

Conclusion

This analysis uses the CRISP-DM framework to uncover factors influencing Airbnb pricing in Seattle. Our findings reveal that features such as the number of bedrooms, bathrooms, and review scores are highly influential in determining listing prices. By focusing on these aspects, hosts can set more competitive prices and improve listing performance. Future research could incorporate more detailed neighborhood data to provide even more localized insights.

References

- - CRISP-DM 1.0. Step-by-step data mining guide. Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf
- - Airbnb Inside. Seattle Airbnb Open Data. Retrieved from https://www.kaggle.com/airbnb/seattle