A CRISP-DM Analysis of New York City Airbnb Pricing

Aishwarya Murahari

Abstract

This paper presents a comprehensive analysis of the New York City Airbnb dataset using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. We explore the factors influencing Airbnb listing prices and develop a predictive model to assist hosts in optimal pricing strategies. The study demonstrates the application of each CRISP-DM phase, from business understanding to deployment, highlighting the importance of a structured approach in data science projects.

1 Introduction

The sharing economy has revolutionized various sectors, with Airbnb leading the transformation in hospitality. Understanding the dynamics of pricing in this market is crucial for both hosts and the platform. This research applies the CRISP-DM methodology to analyze the New York City Airbnb dataset, aiming to uncover key price determinants and develop a predictive model for listing prices.

2 Methodology

We employed the CRISP-DM methodology, which consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase was meticulously executed to ensure a robust and comprehensive analysis.

2.1 Business Understanding

The primary objective was to identify factors influencing Airbnb listing prices in New York City and develop a model to predict these prices accurately. This information is valuable for hosts to optimize their pricing strategies and for Airbnb to provide data-driven recommendations.

2.2 Data Understanding

We utilized the New York City Airbnb Open Data dataset from Kaggle. Initial exploration revealed key features such as location, room type, minimum nights, and number of reviews. Descriptive statistics and visualizations were used to gain insights into the data distribution and relationships between variables.

2.3 Data Preparation

Data preparation involved handling missing values, removing outliers, and encoding categorical variables. Specifically:

- Missing values in 'reviews_per_month' were filled with 0.
- Extreme prices (above 99th percentile) were removed to mitigate outlier effects.
- Categorical variables like 'neighbourhood_group' and 'room_type' were one-hot encoded.
- Numerical features were scaled using Min-Max scaling.

2.4 Modeling

We implemented two models:

1. Linear Regression: As a baseline model to understand linear relationships.
2. Random Forest Regressor: To capture complex, non-linear relationships in the data.

The dataset was split into 80% training and 20% testing sets. We used scikit-learn for model implementation and evaluation.

2.5 Evaluation

Model performance was assessed using Mean Squared Error (MSE) and R-squared metrics. The Random Forest model outperformed Linear Regression, indicating the presence of non-linear relationships in the data. We also conducted residual analysis and feature importance evaluation to gain deeper insights into the model's behavior.

2.6 Deployment

The deployment phase focused on making the model accessible for business use. We demonstrated:

- Model serialization using joblib for easy loading in production environments.

- Creation of a Flask API for serving predictions.
- Development of a simple Streamlit web application for user interaction.

3 Results and Discussion

The Random Forest model demonstrated superior performance in predicting Airbnb listing prices. Key findings include:

- Location (neighbourhood_group) and room type were among the most influential factors in price determination.
- The number of reviews and minimum nights also played significant roles in pricing.
- The model achieved an R-squared value of [insert actual value], indicating a good fit to the data.

Residual analysis showed [insert findings about residual distribution], suggesting [insert implications].

4 Conclusion

This study demonstrates the effectiveness of the CRISP-DM methodology in analyzing the New York City Airbnb market. The developed model provides valuable insights for hosts to optimize their pricing strategies. Future work could explore incorporating more external data sources and implementing more advanced machine learning techniques to further improve predictive accuracy.