Airbnb Pricing Inc.

2487 – Machine Learning -2324_S2

61682 - Robert Helmut Münchau

60526 - Fabrizio Rigodanzo

60178 - Julia Antonioli

61678 - Kuba Bialczyk

59935 - Nicolò Mazzoleni

Submitted to Prof. Nuno André da Silva

Carcavelos, 03.04.2024

Executive Summary

Airbnb Pricing Inc. has embarked on a mission to provide Airbnb hosts in New York City with precise, data-driven pricing recommendations. Leveraging a dataset comprising 102,599 unique Airbnb listings, our analysis identifies critical price influencers and patterns essential for crafting sophisticated price-prediction models. This comprehensive study encompasses data preparation, exploratory analysis, and the application of advanced regression techniques, culminating in actionable insights and recommendations for optimizing Airbnb listing prices.

Our dataset underwent extensive preprocessing to ensure accuracy and relevance for machine learning applications. The exploratory analysis highlighted the dynamic pricing landscape of Airbnb listings, with significant variations across neighborhoods, room types, and operational strategies. Key findings include the potential for Airbnb price optimization especially in districts such as Queens, Bronx, and Staten Island. Also, the identification of factors such as listing availability, minimum nights, and room type has shown to be pivotal in determining pricing strategies.

Our methodology integrates a blend of linear and tree-based regression models, including Linear Regression, Decision Tree Regressor, and Random Forest Regressor, to forecast optimal pricing strategies accurately. The models were evaluated based on Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R2), alongside computational efficiency considerations.

The Random Forest Regressor emerged as the most effective model, offering a balance between accuracy and generalizability. Despite its relatively higher computational demands, this model significantly outperformed its counterparts, suggesting a robust framework for price prediction that accommodates the complex, multifaceted nature of Airbnb listings.

Airbnb Pricing Inc. is leveraging the Random Forest model to offer targeted pricing guidance to New York hosts, emphasizing factors like location, accommodation type, and guest satisfaction. Acknowledging the model's current limitations, the company is not ready to release the model to the public, and it plans to enhance its dataset with more detailed reservation and accommodation specifics, aiming to improve the precision of its pricing recommendations. Looking ahead, Airbnb Pricing Inc. intends to retrain its models with this enriched data to introduce a dynamic pricing feature within the Airbnb platform. This feature will support hosts in achieving various rental objectives, from securing quick bookings to optimizing for different rental durations. By integrating machine learning features into Airbnb's services, the company aims to simplify the listing process, provide strategic pricing insights, and strengthen Airbnb's market position by adding value for hosts.

1. Introduction

Airbnb, Inc. is a renowned American company that operates an online marketplace catering to short-term accommodations and experiences. Founded in 2008, it has by now solidified its position as the leading platform for short-term housing rentals, with more than 448 million nights and experiences booked in 2023. Airbnb's popularity stems from its affordability, prime locations, and authentic local experiences. Presently, over five million hosts offer their properties for rent over the Airbnb platform worldwide.

The majority of Airbnb hosts face the challenge of efficiently pricing their properties to maximize bookings and profits while considering competition rates. The mission of our company, Airbnb Pricing Inc., is to offer Airbnb hosts the service of providing best-pricing recommendations based on various robust price-prediction models. Initially, our startup will concentrate solely on New York City, given its status as the city with the most Airbnb listings in the United States, exceeding 40 thousand, and ranking among the top three cities globally.

At the beginning of the report, we will outline the dataset utilized by our company, detailing the steps involved in cleaning and processing it. Following this, we'll conduct an exploratory data analysis to uncover notable pricing patterns present within Airbnb listings. This analysis will be crucial in pinpointing the most impactful predictors for our models. In the methodology section, we will elaborate upon three distinct modeling techniques employed to ensure the highest possible accuracy in our predictions, thereby offering our customers high-quality service. Finally, we will compare the outcomes of the models and provide recommendations to our clients on how to best utilize our service, while also outlining our plans for optimizing model development in the future.

2. Data

This section delves into the origin of our dataset and the insights derived from our exploratory data analysis. Sourced from a network of experienced Airbnb hosts and property managers, recognized for their meticulous record-keeping and proactive integration of technology to enhance hosting, the dataset provides an exhaustive overview of Airbnb listings in New York City. It includes 102,599 unique entries, embodying distinct Airbnb properties, and spans 26 variables. These range from basic identifiers like 'id' and 'name' to intricate attributes such as 'host_identity_verified' and 'neighbourhood_group', equipping us with a deep understanding of market dynamics, hosting practices, and guest preferences.

Data Preparation Process

In preparation for analysis, the dataset underwent rigorous preprocessing to ensure its integrity and applicability for advanced machine learning techniques. This process addressed critical aspects of data quality, including completeness, consistency, correctness, uniqueness, relevance, contextualization, and trustworthiness. Through these comprehensive data management practices, we have curated a dataset poised for the deployment of sophisticated predictive models.

Exploratory Data Analysis Insights

1. **Price Dynamics**: Airbnb prices span from $50 to $1200, with nearly equal distribution across this range. To refine our price prediction model, 247 rows lacking values were removed from the dataset.
2. **Correlation and Segmentation**: No variables besides service fee displayed a direct correlation with price. Significant variations in price distribution across neighborhoods and construction years were unearthed using the Kruskal-Wallis H-test, though no substantial correlation was found with neighborhood group, instant bookability, or host identity verification.
3. **Geographical Insights and Market Potential**: Brooklyn and Manhattan are the epicenters of Airbnb activity, suggesting the vibrancy and appeal of these areas. Contrastingly, Queens, Bronx, and Staten Island offer untapped potential, with their lower listing counts hinting at different market dynamics and guest preferences. Employing 'neighbourhood_group' as a pivotal location feature streamlines our models, ensuring they capture critical geographical price influences.
4. **Review Dynamics**: Examining reviews alongside neighborhood popularity and pricing revealed median price points and variance based on review volume, with no clear correlation between the average review rate and price at the neighborhood level.
5. **Room Types, Cancellation Policies, and Construction Year**: Our analysis indicates notable regional differences in the impact of instant bookability and cancellation policies on pricing. Staten Island and Queens charge a premium for instant bookability compared to the other districts (Figure 1). Airbnb listings with flexible policies command higher prices in Queens, Brooklyn, and particularly Staten Island, while in Manhattan, the cancellation policy type does influence the price(Figure 2). As for lodging types, hotel rooms are generally the most expensive, while shared rooms in the Bronx and private rooms in Staten Island are the most budget-friendly (Figure 3). Lastly, the Bronx and Queens show a trend where newer properties are priced higher, while in Manhattan, Brooklyn, and Staten Island, property age does not significantly affect pricing (Figure 4).
6. **Host Listings and Market Equilibrium**: New York listings predominantly cater to short-term rentals, with significant borough-wise variations in availability. Despite differing durations, median prices remain stable. Minimum stay requirements reveal a diverse landscape, with outliers indicating longer-term options. Host listings exhibit a highly skewed distribution, suggesting a market dominated by individual hosts with a minority of professional management companies. Despite these variations, median prices demonstrate stability across different host listing counts.

3. Methodology

The further preparation of our dataset for predictive modeling involved removing irrelevant columns, addressing missing values in 'construction_year' through removal, and imputing median values for 'review_rate_number' and 'number_of_reviews'. Categorical variables were transformed into numerical formats using label encoders, optimizing the dataset for advanced analysis and resulting in a refined, model-ready dataset.

Our initial feature selection was informed by deep business insights, emphasizing factors such as trustworthiness (evidenced by host verification), location desirability, accommodation type, cancellation policy flexibility, and guest satisfaction (through review scores and volume). Operational aspects such as listing availability, host strategies (e.g., minimum nights), and construction year (indicating property modernity) were also considered. However, seeking to optimize our model's performance further, we employed Recursive Feature Elimination (RFE) to identify the most impactful features empirically. This process highlighted five key features: availability_365, number_of_reviews, neighborhood, construction_year, and minimum_nights, which led to an improved model performance.

In our modeling evaluation protocol, we meticulously assess the performance of regression models utilizing a comprehensive set of statistical metrics, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R2). These metrics collectively provide a robust framework for evaluating the accuracy and effectiveness of our predictive models. The evaluation process is further enriched by measuring the training duration, offering insights into the computational efficiency of each model. Additionally, we maintain a detailed record of the evaluation results in a CSV file, fostering a transparent and accessible repository of model performance data. This repository aids in the comparative analysis of different models, guiding strategic decision-making in model selection. Moreover, the visualization of actual versus predicted values through scatter plots serves as an intuitive tool for assessing model precision, enabling us to quickly identify areas for improvement. This comprehensive approach underpins our commitment to deploying highly accurate and efficient predictive models, as detailed in our modeling report.

Finally we explored three regression models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor. Linear Regression serves as a foundational model, prized for its simplicity and clarity in establishing linear relationships between features and the target variable. The Decision Tree Regressor, by contrast, captures complex, non-linear patterns through a hierarchical structure, though it carries the risk of overfitting. The Random Forest Regressor, an ensemble of decision trees, addresses this by averaging predictions across multiple trees, thereby enhancing predictive accuracy and stability. Each model was applied to data refined through a robust scaler and stratified train-test splits, demonstrating a balanced approach that scales from simplicity to complexity and from interpretability to enhanced predictive performance. We additionally ran these models twice: first with features chosen based on the business context, and following with features chosen using Feature Selection Techniques with Random Forest Regressor and Extra Trees Regressor. These two methods were chosen for their effectiveness in identifying and selecting the most relevant features from the dataset, thereby improving model performance and interpretability through robust feature selection techniques. This structured progression underscores our commitment to deploying sophisticated, highly accurate predictive models tailored to the nuanced dynamics of our dataset.

4. Results & Recommendations

When evaluating key performance metrics—MSE, MAE, RMSE, and R2—along with training duration, we aim to determine the most effective predictive model for our dataset.

The Linear Regression model showed consistent performance, with marginal improvements in MSE and RMSE, yet an R2 score that suggests limited predictive power. Its relatively unchanged R2 score after feature engineering indicates that Linear Regression may not be well-suited for datasets with complex, non-linear relationships.

The Decision Tree Regressor yielded disappointing results. High error scores and a negative R2 score reveal that this model may have difficulties with the dataset, suggesting it is either overfitting or not capturing the underlying patterns effectively.

In contrast, the Random Forest Regressor stands out with its lower MSE and RMSE and a positive R2 score. This model's ability to improve with feature engineering speaks to its robustness and indicates a superior capacity to generalize from the data without overfitting. When considering training times, the Random Forest Regressor required longer durations, a common trade-off for more sophisticated models. Despite this, its performance suggests a clear advantage in dealing with complex datasets.

Overall, while the Random Forest Regressor is the standout performer, especially with the refinements made through feature engineering, it's important to note that the metrics—though improved—are not ideal. The model, despite showing the best results with lower MSE and RMSE and a higher R2 score, doesn't fully capture the dataset's variability, which may limit its reliability. Thus, while it ranks as the best among the tested models, the moderate R2 score reflects room for further improvements and caution in relying too heavily on its predictions. This underscores the potential necessity for additional strategies to enhance model reliability further.

It's crucial to acknowledge the limitations of the current models' accuracy. To bolster the relevance and precision of our pricing advice, it's crucial to incorporate more detailed factors into our dataset, including granular reservation-level data and specific accommodation characteristics. This would enrich our understanding and enable more nuanced pricing strategies. If the data is enhanced, we will utilize the same pipeline to train multiple models and subsequently integrate this as a feature in the Airbnb product that will automatically suggest pricing ranges to hosts. Such a feature could be further developed to align with other hosting objectives, such as tailoring recommended prices to prioritize rapid rentals or optimizing for either long-term or short-term tenancies. The integration of ML-based features into Airbnb's offerings will not only streamline the listing process but also maintain Airbnb's competitive edge, as hosts will find greater value in the ease and strategic insight provided by the platform's rental services.

Appendix

Figure 1: Impact of Booking Type on Price by New York Districts

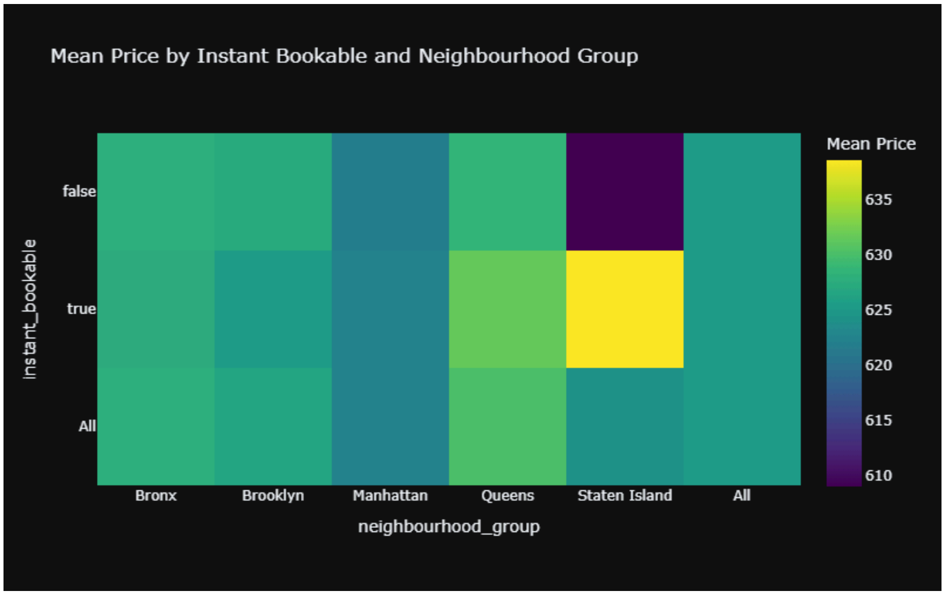


Figure 2: Impact of Cancellation Policy on Price by New York Districts

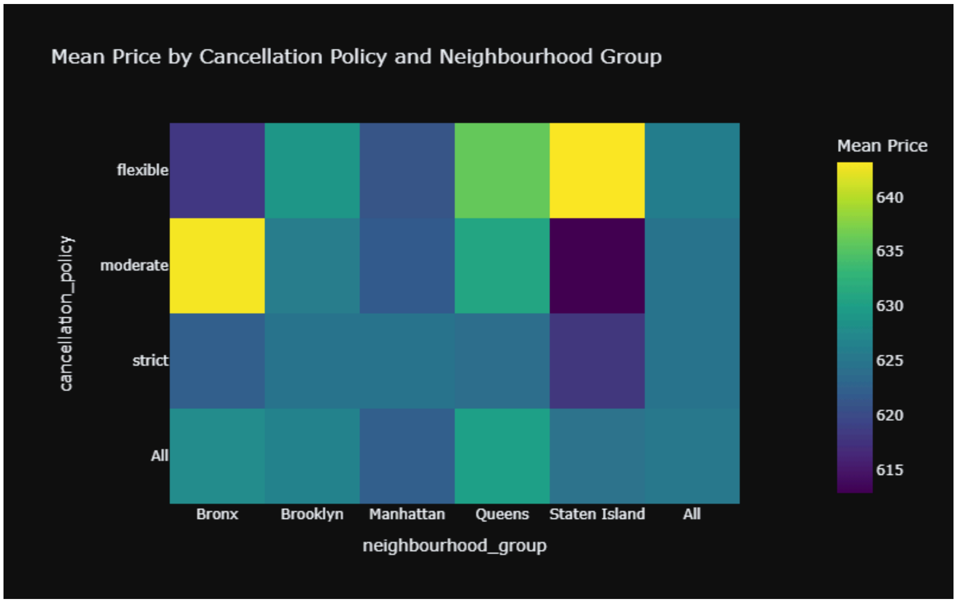


Figure 3: Impact of Room Type on Price by New York Districts

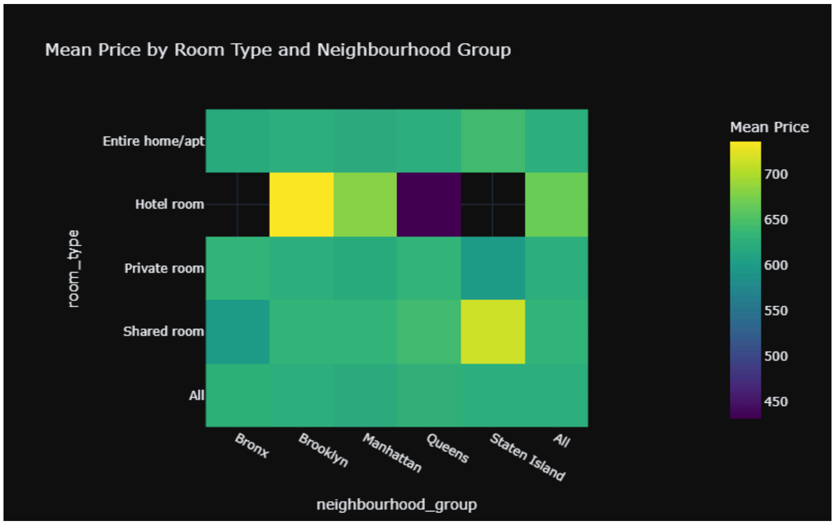
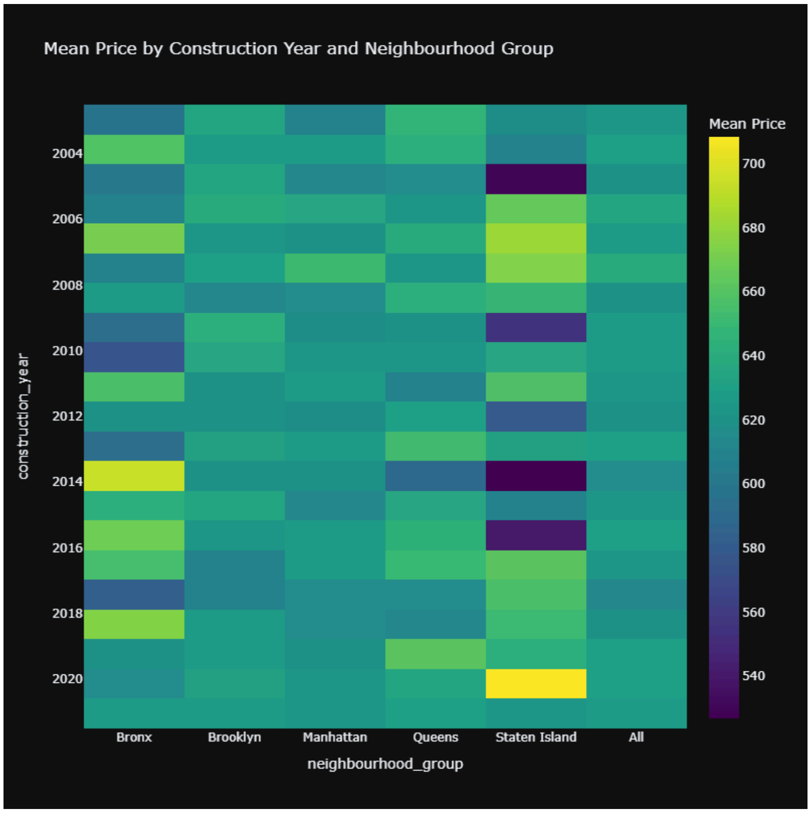


Figure 4: Impact of Construction Year on Price by New York Districts


Mean Price by Construction Year and Neighbourhood Group
construction_year
Bronx
Brooklyn
Manhattan
Queens
Staten Island
All
neighbourhood_group
Mean Price