BAIT 509 Final Report

Pricing Optimization for New York Airbnb Listings

Insights and Strategies from Machine Learning Analysis

February 17, 2024

BA 1 Group 1

Namith Mohan

Agnes Xi

Yihan Yang

Sissi Zhang

Table of Contents

Background, Business Question & Motivation

Background

Since its inception in 2008, Airbnb has significantly transformed the accommodation sector, introducing a novel paradigm in how people travel and stay in destinations around the world. In New York City, a beacon of tourism and culture, Airbnb has played a pivotal role in reshaping lodging options. By leveraging technology to connect homeowners with travelers seeking unique and personalized accommodation experiences, Airbnb has diversified the range of lodging options beyond traditional hotels. This platform offers everything from shared rooms to entire homes, catering to a wide spectrum of preferences and budgets. This transformation has not only provided travelers with greater flexibility and choice but also enabled property owners in New York City to tap into a new source of income, fundamentally altering the dynamics of the short-term rental market. Through its innovative model, Airbnb has ushered in a new era of travel, emphasizing local experiences and community engagement, thereby setting a new standard in the sharing economy.

Business Question

The dataset that the report will revolve around entails the listing activities and evaluation metrics of properties in New York City in 2019. Considering and using features from the dataset, how can Airbnb property owners in New York City optimize their listings in order to maximize their rental income while maintaining competitive pricing?

Motivation

This question is crucial for Airbnb owners who seek to understand the multitude of factors influencing the pricing of their properties. It underscores the importance of identifying and leveraging the key attributes that contribute to the rental income potential. By addressing this question, the project aims to provide actionable insights that enable property owners to make informed decisions regarding property management, strategic pricing, and targeted improvement. The ultimate goal is to enhance the attractiveness of their listing, align prices with the current market demand, and increase occupancy rates, thereby optimizing their returns on investment.

Data and Statistical Question

This dataset encompasses a wide array of attributes, including but not limited to, the distribution of listings across various neighborhoods and boroughs, room types, the number of bedrooms, reviews and review scores, and availability throughout the year. The dataset provides a holistic view of the Airbnb market dynamics in New York City, capturing insights into trends, traveler preferences, and the impact of tourism.

The core statistical question we aim to address is: *"How accurately can we predict the listing price of an Airbnb property in New York City based on features that are relevant and significant, and what confidence level can we attach to these predictions?"* This question is designed to uncover the relationships between property features and listing prices, facilitating the development of a predictive model through supervised learning techniques. The model's predictions, and the confidence in these predictions, are pivotal for making informed pricing decisions.

This question is instrumental in several ways. First, it enables property owners to understand which features significantly affect listing prices, allowing them to emphasize these in their offerings. Second, the predictive model can guide owners in setting prices that are competitive yet profitable, based on identified trends. Lastly, insights gained from the model analysis can highlight areas for improvement or investment in properties to increase their appeal and, consequently, their earning potential.

However, the statistical question does not summarize the entirety of the business question. It primarily focuses on price prediction and optimization from static features, potentially overlooking the dynamic aspects of the market, such as fluctuating demand due to seasonal changes or special events. Additionally, it may not fully consider the qualitative elements that influence a property's attractiveness, like the host's hospitality level, the quality of guest interactions, or the aesthetic appeal of the listing photos. These factors, though harder to quantify, play a crucial role in a listing's success and overall rental income.

To enhance the predictive capability of our model, a multifaceted approach to feature engineering is proposed. This involves the application of one-hot encoding to transform categorical variables such as neighborhood and room type into a format amenable to machine learning algorithms, thereby facilitating the nuanced interpretation of these attributes. Additionally, the extraction of temporal features is planned to encapsulate seasonal trends and demand fluctuations, reflecting the rental market's dynamic essence. Furthermore, the generation of interaction features, by merging variables like neighborhood with room type, is aimed at explaining the intricate relationships that influence listing prices, offering a comprehensive analytical perspective.

Through meticulous feature engineering and the judicious selection of a model architecture (considering regression models, tree-based models, or neural networks based on the dataset's characteristics), we aim to construct a predictive model that not only estimates Airbnb listing prices with precision but also offers actionable insights for property owners. This model will serve as a tool for strategic decision-making regarding property management, pricing strategies, and targeted improvements, ultimately guiding owners towards optimizing their rental income within the competitive landscape of New York City's Airbnb market.

Exploratory Data Analysis

Dataset Features

Feature	Description
id	The unique identifier for each listing. Not useful for modeling as it's just an identifier and will be excluded in analysis.
host_id	The unique identifier for each host. This might be useful for identifying power hosts or those with multiple listings, however, because we are interested in predicting the price, so we will be excluding this feature.
neighborhood_group	This indicates the broader the listing is in, which can affect price and availability.
neighbourhood	More specific location information which can be very important as prices can vary greatly even within the same city by neighborhood.
Latitude and longitude	Geographic coordinates. These are crucial for any geospatial analysis and can be used to calculate distances to points of interest.
room_type	This indicates whether the listing is an entire home, private room, or shared room, which is a major factor in pricing.
price	The price for renting the listing, which is our target variable.
minimum_nights	The minimum stay requirement, which could impact the listing's popularity and price.
Number of reviews	Reflects the popularity or quality of a listing.
reviews_per_month	Another indicator of listing popularity and frequency of occupancy.
calculated_host_listings_count	How many listings the host has, which could indicate professional hosts versus casual ones; we may consider removing this feature as it is not as relevant as other features to predict price.
availability_365	How many days in the year the listing is available, impacting revenue potential.
days_since_last_review	A derived feature indicating how recently the listing was reviewed, which could correlate with current listing quality or popularity.

Annotation

- We create the column of days_since_last_review to help capture how active or in-demand a listing is.
- We replaced NaN values in reviews_per_month feature with 0, where no reviews in a month mean the listing wasn't reviewed, rather than missing data.

```
1 airbnb_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   host_id                         48895 non-null  int64  
 2   neighbourhood_group             48895 non-null  object 
 3   neighbourhood                   48895 non-null  object 
 4   latitude                        48895 non-null  float64
 5   longitude                       48895 non-null  float64
 6   room_type                       48895 non-null  object 
 7   price                           48895 non-null  int64  
 8   minimum_nights                  48895 non-null  int64  
 9   number_of_reviews               48895 non-null  int64  
 10  reviews_per_month               48895 non-null  float64
 11  calculated_host_listings_count  48895 non-null  int64  
 12  availability_365                48895 non-null  int64  
 13  days_since_last_review          38843 non-null  float64
dtypes: float64(4), int64(7), object(3)
memory usage: 5.2+ MB
```

Considerations

- Replace null values using imputation
- Remove outliers
- For the categorical features, we will use one-hot encoding to deal with them; in the following graphs, we will assess the price distribution among different types
- Understand the correlation between numerical features and the target

Plots Interpretations

The price distribution of Airbnb listings in New York City, as revealed by the histogram, is notably right-skewed (*Appendix 1*). This skewness is characterized by a high concentration of listings at lower prices and a minority of listings with exceptionally high prices. The distribution points to the presence of outliers—listings that are priced significantly above the norm. These outliers could potentially skew predictive analyses and model outcomes if not addressed, given their disproportionate influence on the overall data trends. To ensure a more accurate and unbiased analysis, it would be prudent to remove or adjust these extreme values prior to conducting further exploratory data analysis or predictive modeling.

Room type emerges as a crucial determinant of pricing structure, and the box plot analysis matches with the intuitive expectation (*Appendix 2*). The plot illustrates that entire homes or apartments are typically priced higher than private or shared rooms, as larger spaces that offer more privacy and can host more guests tend to command higher prices. Consequently, room type should be a primary feature in predictive models for Airbnb pricing, as it encapsulates the level of accommodation and privacy, factors that are pivotal in influencing guests' willingness to pay.

The distribution of listings across different neighborhood groups is uneven, with neighborhoods like Brooklyn and Manhattan dominating the dataset, while Staten Island and the Bronx feature far fewer listings (*Appendix 3*). Although the bar plot does not directly link neighborhood groups to pricing, the significant disparity in listing counts across neighborhoods can be indicative of varying demand and average pricing, making *neighborhood groups* a potentially influential predictor of listing prices.

Analyzing the heatmap of correlation coefficients between numerical features and price provides further insights (*Appendix 4*). The *number of reviews* has a moderate negative correlation with price, suggesting that higher-priced listings may receive fewer reviews. In contrast, *reviews per month* have a slight positive correlation with price, implying that more expensive listings could be popular or highly rated, attracting more frequent reviews. The *days since last review* metric shows a moderate negative correlation with price, hinting at a possible trend where listings with more recent reviews tend to be less expensive, potentially reflecting active management or competitive pricing strategies.

The *longitude* feature has a moderate correlation with price, which might be attributable to factors such as the proximity to city centers, attractions, or other desirable locations. This relationship could also be capturing variations across neighborhoods that are not fully explained by neighborhood categories alone. Since New York has a vertically narrow landscape, evaluating *longitude* is expected to yield influential results for price optimization. Conversely, *latitude* shows minimal correlation with price, suggesting that longitudinal location within a neighborhood has a more pronounced impact on pricing than latitudinal differences.

Finally, the calculated host listings count has a weak positive correlation with price, which could indicate that hosts with multiple listings set higher prices, possibly due to the advantages of professional management. Availability over the year (*availability_365*) also presents a weak correlation with price. This aspect merits further exploration to uncover any potential seasonal pricing trends or to understand how constant availability might correlate with pricing strategies adopted by hosts.

In summary, *room type* and *neighbourhood group* are likely strong categorical predictors for price. Among numerical features, *number_of_reviews, reviews_per_month,* and *days_since_last_review* could be significant predictors and should be included in any predictive models. The presence of outliers in the price data should be handled carefully, as they can affect model training and the interpretation of results.

Method & Results

Feature Preprocessing and Feature Selection

When preparing our dataset, we generated a feature 'days_since_last_review' to capture the recency of review activity, assuming that it might provide information about the listings' popularity, a factor potentially influencing their prices. We also replaced NaN values in 'reviews_per_month' with zeros because the corresponding 'number_of_reviews' for these listings were all zero, suggesting these listings had not previously been reviewed. Since columns like 'name' and 'host_name' are generally unique, not predictive, and may have privacy concerns, we opted to drop them. We also dropped the 'neighborhood' column because its information overlaps with that in the 'neighbourhood_group', 'longitude', and 'latitude' columns and is not as relevant as these variables based on previous exploratory analysis.

For data preprocessing, we customized a MaxPlusOne imputer to fill in missing values in the 'days_since_last_review' column with the maximum value plus one from the dataset. All listings with missing values for this column have zero reviews, as we found in the dataset. The rationale behind this is that these listings haven't been reviewed before and are likely new listings. Thus, their popularity, if considered from the recency of reviews, is similar to those listings that were last reviewed a long time ago. The new listings are not yet popular, while the listings that weren't reviewed recently are no longer popular.

We then constructed a data preprocessing pipeline with different methods of transformations applied to specific data types within our dataset. Numeric features were normalized using StandardScaler, negating the disproportionate influence of larger-scale variables. Categorical variables were transformed via OneHotEncoder to translate their qualitative information into a quantitative format suitable for machine learning models. By integrating these preprocessing steps into a ColumnTransformer pipeline, we ensured a streamlined application for both training and testing sets that can later be seamlessly integrated with a variety of machine learning models.

Reflection on Selected Performance Metrics

Selecting R^2 as our primary performance metric was crucial for evaluating our regression models' effectiveness in predicting Airbnb listing prices (*Appendix 5*). It quantified the proportion of variance in listing prices explained by the models, offering an intuitive measure for comparing their predictive power. While R^2's interpretability facilitated clear communication of model performance, its limitations, such as not accounting for model complexity or the magnitude of prediction errors, highlighted the need for a comprehensive evaluation approach. Therefore, complementing R^2 with error metrics like MSE or MAE ensured a well-rounded assessment of each model's accuracy and practical relevance, guiding us to select the most effective model for our analysis.

Model Building and Hyperparameter Tuning

We first built a dummy regressor using the "median" strategy as our baseline model because the Airbnb listing prices have few large outliers, and the median is less sensitive to these extremes compared to the mean. We then tried out various models including Decision Tree Regressor, Ridge Regression, Lasso Regression, and XGBoost Regression. The Decision Tree Regressor was chosen for its high interpretability. It allows us to trace the decision-making path, which is crucial when stakeholders need to understand the pricing logic. Both Ridge and Lasso Regression were selected for their regularization abilities to prevent overfitting. Ridge Regression's L2 penalty keeps all features with reduced influence, which is suitable when many features slightly affect the price. Lasso Regression's L1 penalty reduces some coefficients to zero, performing feature selection and enhancing model simplicity. We also selected XGBoost Regression due to its predictive power, especially with diverse data and non-linear relationships. Hyperparameters for these models were optimized using RandomSearchCV scored on the R^2 metric to develop a model that explains price variability effectively.

Model Selection

In our comprehensive model selection and evaluation process, we assessed various regression models, including Decision Tree, Ridge, Lasso, and XGBoost, using *RandomizedSearchCV* for hyperparameter optimization and R^2 as the primary evaluation metric. The XGBoost Regressor emerged as the superior model, distinguishing itself with the highest R^2 score, which signifies its exceptional ability to capture the complex dynamics and non-linear relationships affecting Airbnb listing prices. This model's comparative analysis against a baseline Dummy Regressor, which showed negligible predictive power, further validated XGBoost's enhanced predictive capabilities. Moreover, the feature importance analysis from the XGBoost model provided crucial insights, identifying room type and location as key determinants of pricing, thereby offering actionable intelligence for optimizing listings. This approach not only validated the robustness of our methodology but also illuminated actionable paths for leveraging these insights in the Airbnb market, making the XGBoost Regressor the optimal choice for predicting Airbnb prices with accuracy and depth.

Results and Insights

Communication of Results

- Predictive Factors:
    1) Room Type: Our model underscores the significant impact of room type on listing prices. Specifically, entire homes or apartments tend to command higher prices than private or shared rooms. This finding highlights the premium guests place on privacy and the exclusivity of space during their stays.
    2) Location's Influence: Geographical location proves to be a key determinant of pricing, with properties closer to city centers or major attractions fetching higher prices. High-demand neighborhoods are particularly lucrative, reflecting the value guests place on convenience and the desirability of the area.
- Model Selection: The XGBoost model stood out as the premier predictive tool in our study, offering unparalleled accuracy and reliability in forecasting Airbnb prices. Its selection was based on an impressive R^2 score, which signifies a robust ability to explain variations in listing prices across different factors.
- Key Predictors: Beyond room type and location, our analysis identified specific amenities as significant price influencers. The model's feature importance analysis provided a clear indication of which aspects contribute most to pricing, presenting a clear direction for hosts looking to improve their offerings.

Communication of Insights

For Airbnb Hosts:

- Optimize Pricing: Implement dynamic pricing strategies informed by key predictors like room type and location. Adjusting prices based on these insights can enhance the listing's competitiveness and profitability.
- Enhance Listings: Focus on high-quality photography and detailed descriptions that highlight the listing's unique features and amenities. Tailoring the listing to align with the model's key predictors can significantly attract more guests.
- Improve Guest Experience: Prioritize creating a positive guest experience to boost reviews and ratings. Consider the model's emphasis on reviews as a direct factor influencing the listing's appeal and pricing potential.

For Airbnb Tenants:

- Smart Booking: Utilize the insights on price influencers to find listings that offer the best value for needs. Focus on preferred 'room_type' and 'neighbourhood_group' to ensure a stay that matches budget and preferences.
- Leverage Reviews: Make informed decisions by analyzing listings with a high number of positive reviews. This can serve as a marker for quality and satisfaction, guiding others towards the best possible stay.

For Market Analysts

Stakeholders in the real estate and hospitality sectors can draw on these findings to better understand the factors driving the sharing economy, particularly in the context of urban tourism and accommodation.

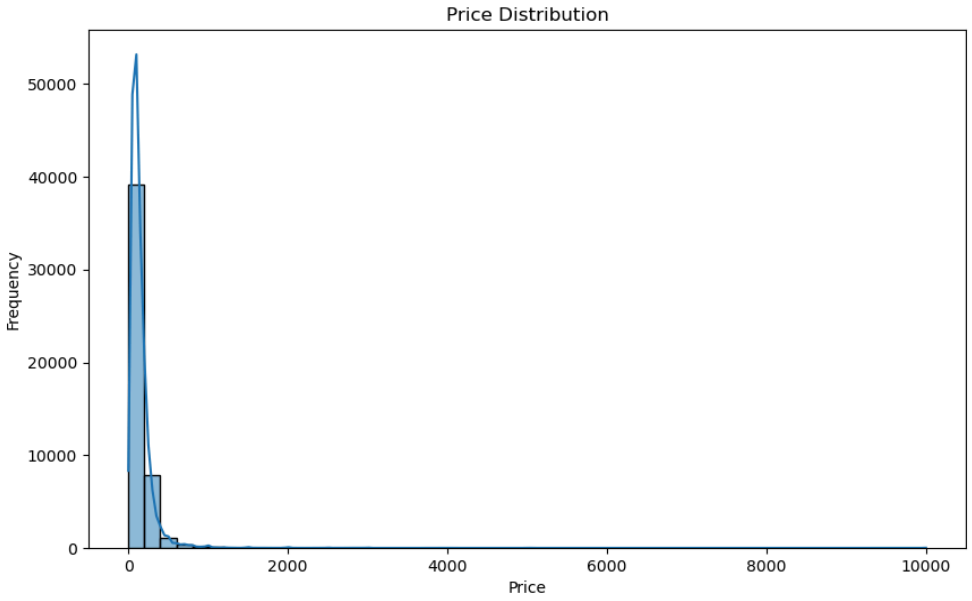
References

Building a custom Scikit-learn imputer. Stack Overflow. (2023, July 1). https://stackoverflow.com/questions/44575114/building-a-custom-scikit-learn-imputer

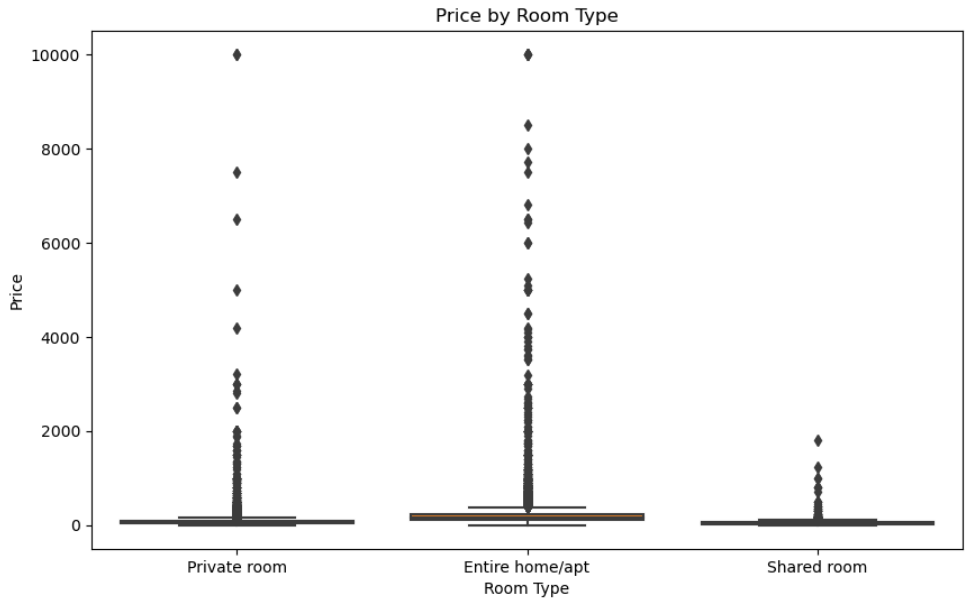
New York City airbnb open data. Kaggle. (2019, August 12). https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Appendices

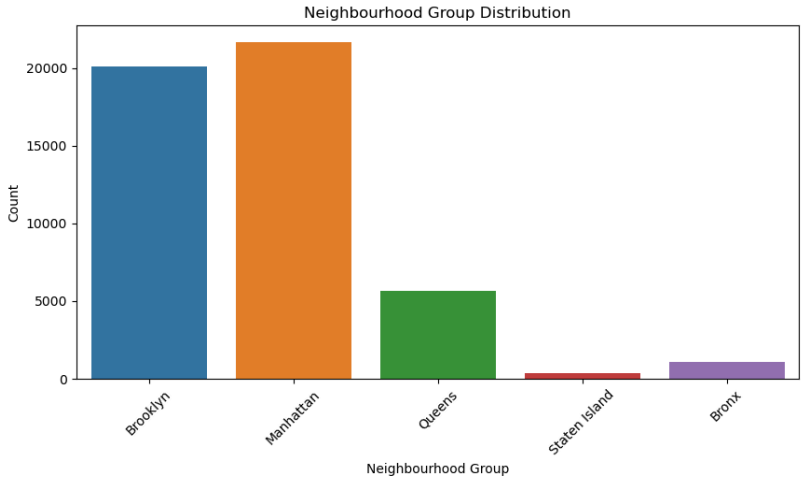
Appendix 1: Price Distribution



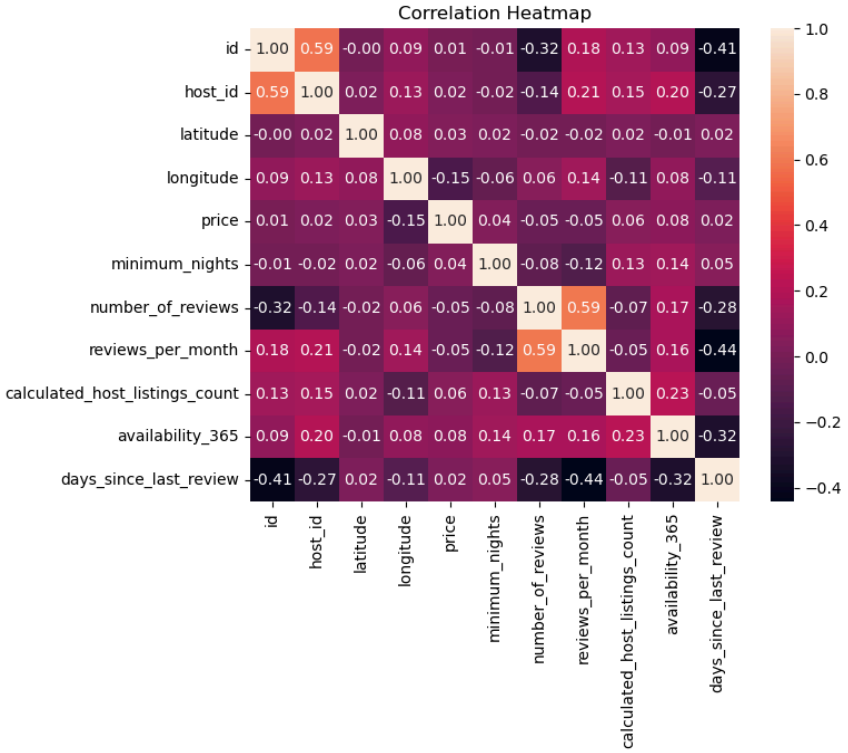
Appendix 2: Price by Room Type Boxplot



Appendix 3: Neighbourhood Group Distribution



Appendix 4: Correlation Heat Map



Appendix 5: Regression Models' Error Analyses

	Model	MSE	MAE	R^2
0	Linear Regression	47369.410165	71.315122	0.137790
1	Ridge Regression	47336.795801	71.245364	0.138383
2	Lasso Regression	47568.027269	71.264352	0.134175
3	Decision Trees	47936.956621	70.078889	0.127459
4	Random Forests	44275.805787	65.559934	0.194099
5	Gradient Boosting	45809.975337	66.416773	0.166174
6	XG Boosting	43874.633477	64.867069	0.201401