Springboard Data Science Capstone 2 - Predicting Short-term Listing Prices

Nizar Altawam

Table of Contents

1. Introduction:

1.1. Background:

Short-term rentals (STRs) are a type of accommodation that allows travelers to rent a property for a short period of time, usually less than a month. STRs have become increasingly popular in recent years, especially in urban areas like New York City (NYC), where they offer an alternative to hotels and other traditional lodging options. However, STRs also pose various challenges for the hosts, guests, and regulators, such as pricing, quality, safety, and legality.

1.2. Objective:

This is a project report for a data science capstone project that aims to build a machine learning prediction model for short-term listings. The motivation for this project is to provide consulting services for a start-up company that operates in the NYC market. This prediction model can help the company optimize its pricing strategy and increase its revenue and market share in the competitive STR industry.

1.3. Scope:

The scope of this project is to use the historical airbnb listing data from NYC with features including maximum capacity of listing, borough neighborhood information, property type, number of rooms and beds, minimum number of night stay, and price as the dependent variable to train the model

2. Data Collection and Wrangling:

2.1. Data Source

The data source for this project is InsideAirbnb.com, a website that collects and publishes historical Airbnb listing data. The data covers a 12-month period (ending December 4th, 2022). Three separate datasets were obtained:

- “listings.csv.gz” - Detailed Listings data
- “calendar.csv.gz” - Detailed Calendar Data
- “listings.csv” - Summary information and metrics for listings in New York City.

2.2. Data Wrangling

The data was imported into a pandas dataframe. 41,533 entries were observed. The detailed listings data included 75 features, and the summary information contained 18.

Using the summary dataset as a starting point, the null values were inspected. The Listing's with missing/incorrect information were either amended or dropped. Additional wrangling to be done in the EDA step after exploring relationships between features.

2.3. Feature Selection

Only the features relevant to the research question were retained in the dataframe.

- For location, these included: "neighbourhood", "neighbourhood_group", and the geocode location.
- The room_type feature which grouped listings into the following three categories (entire home/apt, private room, shared room). Hotel rooms were dropped.
- "Accommodates" is the maximum capacity of a listing.
- "Bedrooms" and "Beds" are numerical features.
- On the Airbnb web-site, the bathrooms field has evolved from a number to a textual description. For older scrapes, "bathrooms" is used. "Bathrooms_text" will need to be converted.
- Availability_365: The availability of the listing 365 days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
- Amenities is a json object of self-reported available amenities. This may be used in a future step.

The selected features and wrangled listings were exported into a new csv to be used in the Exploratory Data Analysis step, where they will be visualized and summarized. The distribution of price and its relationship with other features will be examined.

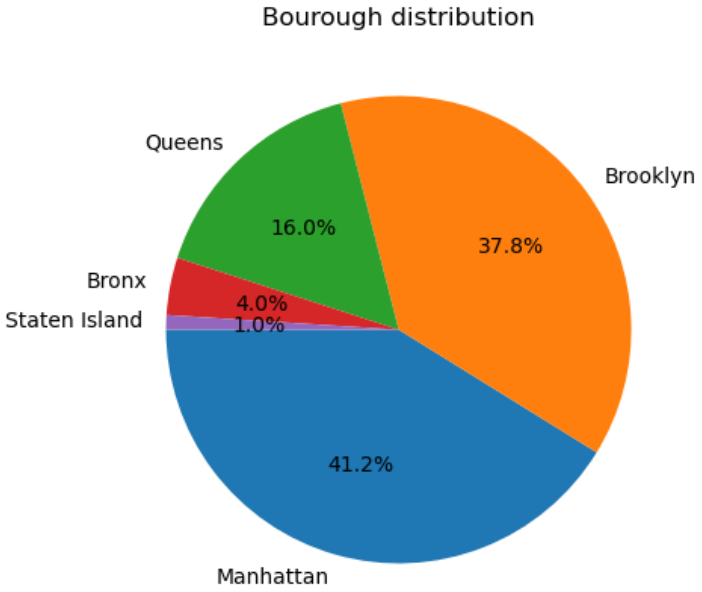
3. Exploratory Data Analysis:

3.1. Data Distribution Analysis:

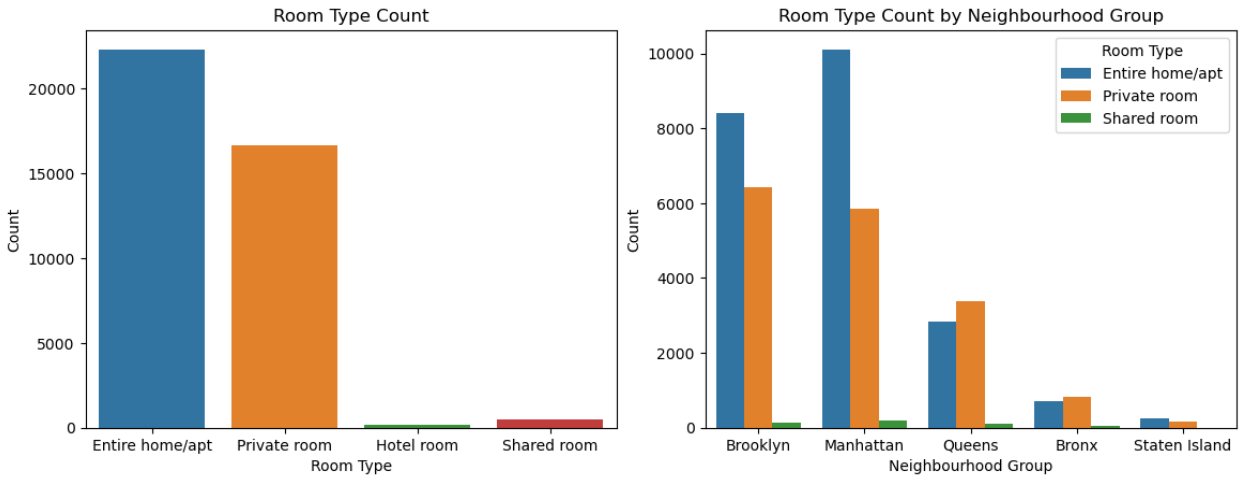
- Since the premise of this project focuses only on short-term listings, I will define those as listings with a `minimum_nights` to be less than 30. Upon visualization, I noticed there was a significant amount of data points (17,617) that bordered the predetermined definition of short-term listing.

Upon further research, it appears in New York City, short-term rental regulations apply to rentals of less than 30 days. These regulations are meant to protect housing for local residents. **For this project, I will amend the definition to include less than 31 days, and will note that the data indicates the market has shifted to longer-term stays**. This may have been to avoid these regulations or possibly due to the pandemic.

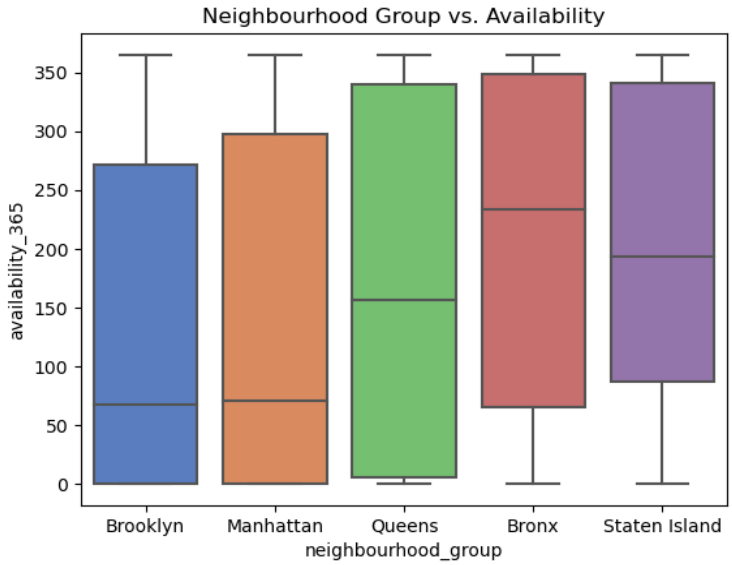
- We found that Brooklyn and Manhattan had the largest share of listings (37.8% and 41.2%, respectively), while Staten Island had the smallest (1%).



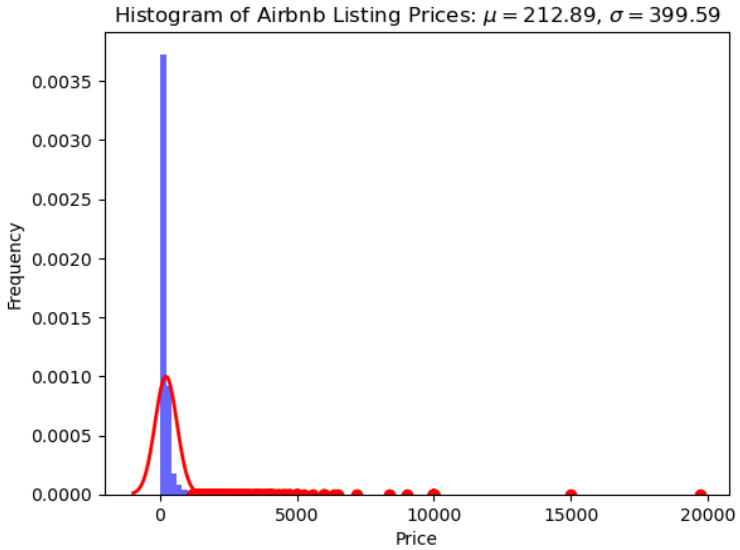
- The most common room types were entire home/apt and private rooms, but there were differences across boroughs. For example, Queens and the Bronx had a higher proportion of private rooms than entire apartments.



- We also examined the availability of listings and found that Brooklyn and Manhattan had lower availability rates than the other boroughs, suggesting higher demand.



- During price feature analysis, it was noted some listings were valued > $20,000. These were determined to be erroneously entered and were dropped from the dataset. After dropping these listings, there was still a significant amount of outliers as visualized below:



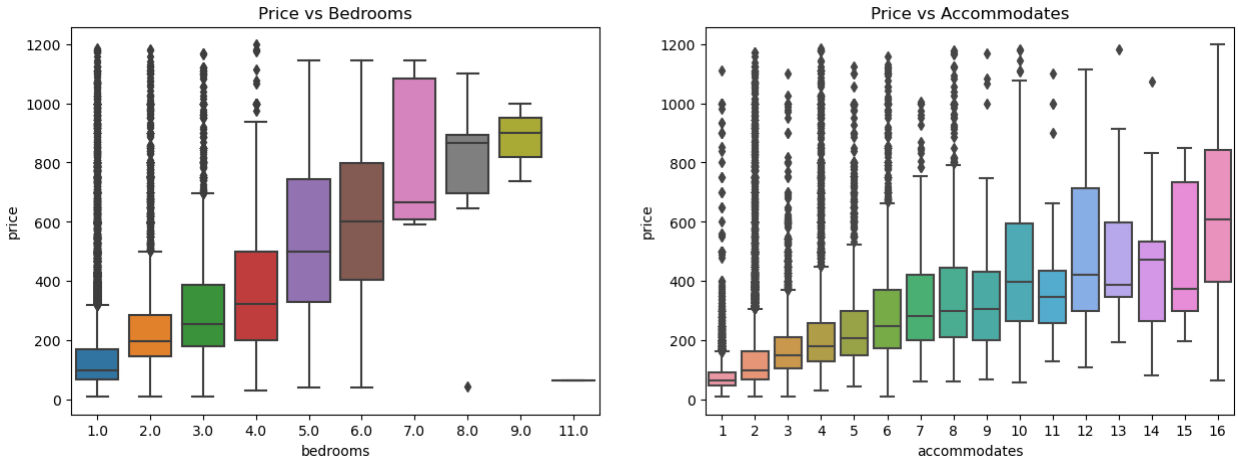
As a result, we inspected all listings above the 99.7 percentile of price. Without a way to accurately verify the price of these listings, the 598 rows were dropped.

The summary statistics for price after this was as follows:

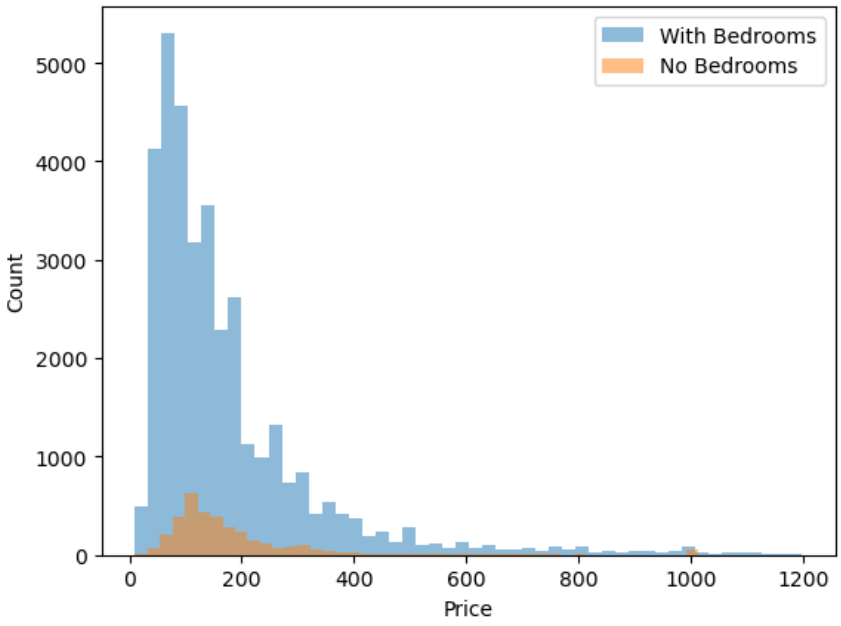
count:	mean	std	min	25%	50%	75%	max
38815.00	177.25	159.67	10.00	80.00	130.00	210.00	1198.00

3.2. Correlation Analysis:

We performed correlation analysis to examine the price distribution across the boroughs. Excluding the high priced listings, we found that Manhattan had the highest median price of $154 per night, compared to the other boroughs. We also found that more bedrooms and guests accommodated were typically associated with higher prices.



During this analysis, it was noted that there are 3555 null values for bedrooms, and 861 for beds. Upon closer examination, listings with no bedroom information were more likely lower priced rooms. Since we concluded there was some correlation between the price and number of bedrooms, then these listings were more likely to have a smaller number of bedrooms.



Using this logic, listings with a room_type of private room had bedrooms set to 1. Additionally, it appeared that listings that accommodate 1 person will most likely have 1 bedroom. These were adjusted as such. Finally for the remaining listings, the bedroom count was assigned the median value of bedrooms for that accommodates value.

Similarly, listings with a private or shared room had the beds set to 1 and the remainder were also imputed using the median for the respective accommodates value.

4. Data Pre-processing

4.1. Categorical Feature Encoding

Borough and neighborhood features were encoded using one-hot encoding. Room_type was encoded as an ordinal feature, assigning importance in decreasing order.

4.2. Data Splitting (Training and Testing)

The features were assigned to variable X and the price to y. The data was then split into training and testing datasets with a 70/30 ratio. The train and test sets had the following shapes: ((27170, 231), (11645, 231)).

4.3. Baseline Model - Dummy Model

A dummy model was created using the mean value of price as a constant prediction. This model served as a baseline for comparison with other models.

5. Modeling:

5.1. Models:

Two machine learning algorithms were applied to the data: Random Forest Regression and XGboost Regression. Both models were tuned using grid search and cross validation to find the optimal hyperparameters.

5.2. Feature Engineering

A new feature was also created using the latitude and longitude of the listings to calculate their distance from a popular landmark in NYC. This feature was added to the dataframe to test its impact on the prediction accuracy.

6. Model Performance Evaluation:

6.1. Evaluation Metrics

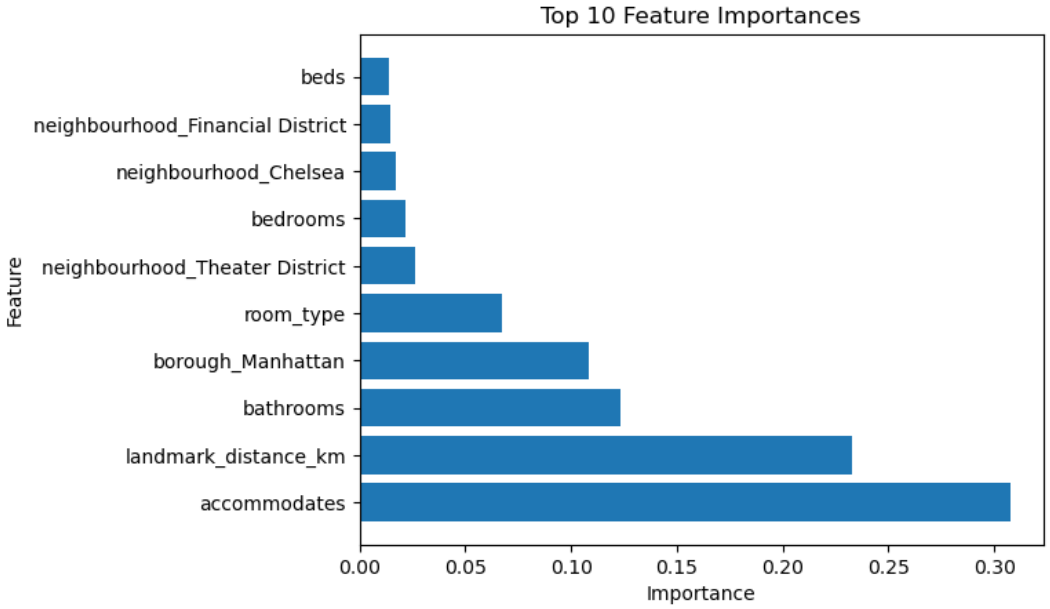
The performance of the models was evaluated using the root mean squared error (RMSE) and mean absolute error (MAE) metrics.

6.2. Results and Analysis

After adding the new feature, the XGboost model performed slightly better than the Random Forest (MAE: 63.45 vs 64.75). However, the RFR took half the amount of time to train and was less computationally intensive. As such I will use the RFR for final predictions.

6.3. Feature Importance Analysis

The Random Forest model was also used to conduct a feature importance analysis to identify which features had the most influence on price. The results showed that the number of guests accommodated, number of bathrooms, and distance from a popular landmark were the most important features in predicting price.



It's important to note that the importance score is a relative measure within the context of the specific model and dataset. It does not necessarily reflect the absolute importance of a feature or imply causality. Feature importance is a useful tool for understanding the relative importance of different features in a model but should be interpreted with caution and in the context of the specific problem domain.

7. Profitability Analysis:

The final part of the project was a profitability analysis that compared the predicted prices from the Random Forest model with a hypothetical pricing model used by the start-up company. The hypothetical model assumed that the company calculated prices by taking the average of listings aggregated by neighborhood and room_type. Here are the first 5 listings with the price and both predictions:

price	predicted_price	naive_price
144	140.5859	99.58113
150	138.2916	167.9686
109	167.8765	191.84
249	110.3573	106.2963
120	97.77564	98.6579

The generated revenue for each model was calculated by multiplying the occupancy rate by the price. To calculate the occupancy rate, using the “calendar.csv.gz” dataset I took the ratio of the number of nights the listing was not available to the total number of nights in a year. The results showed that the start-up company could potentially increase its revenue by adopting a more dynamic pricing strategy based on the Random Forest model. For the evaluated listings, over the course of 12 months, our model would generate $301,298,153.38 more in revenue than the hypothetical model. **This is a 1.4% increase in revenue.**

8. Assumptions and Limitations:

8.1. Data Limitations

- **Dataset Selection**: The project relies on historical Airbnb listing data obtained from the "insideairbnb" website. It's important to acknowledge that the dataset may not capture the complete population of listings or may have limitations in terms of data quality or representativeness.
- **Temporal Considerations**: The dataset spans a 12-month period, and the predictive model's performance may be influenced by the specific time range and potential variations in market dynamics or external factors not captured in the data.

8.2. Feature Limitation

- **Feature Selection**: The selected features, such as borough, neighborhood, room_type, minimum_nights, bedrooms, bathrooms, and beds, may not fully capture all relevant factors influencing the pricing of short-term listings. Other unobserved or uncollected features (e.g., property amenities, listing descriptions) could play a significant role but are not included in the analysis.
- **Feature Encoding**: The categorical features are encoded, which introduces assumptions about the relationship between categories. The effectiveness of the encoding scheme may be influenced by the specific dataset and could potentially limit the model's performance.

8.3. Model Limitations

- **Algorithm Selection**: The choice of using Random Forest Regression and XGBoost models may be based on assumptions and limitations associated with these specific algorithms. Other algorithms or ensemble methods might yield different results or better performance.
- **Model Generalization**: The model's performance may vary when applied to new, unseen data or different geographical regions outside of the NYC context. The model's predictive accuracy may be limited when encountering listings with unique characteristics or in markets with different dynamics.
- **Model Interpretability**: Random Forest Regression and XGBoost models may provide limited interpretability, making it challenging to explain the underlying factors driving the model's predictions.

8.4. Business Context:

- **Start-up Specifics**: The analysis is framed in the context of consulting for a start-up company. It's crucial to acknowledge that the project's findings and conclusions are specific to this particular start-up's situation and may not be directly applicable to other businesses or real-world scenarios.

- **Hypothetical Comparisons**: The profitability analysis involves comparing the model's predictions to a hypothetical model owned by the start-up. The assumptions and methodology used in the hypothetical model may introduce uncertainties and may not accurately reflect the actual performance of the start-up's existing model.

9. Future Steps:

9.1. Additional feature engineering

- Using `amenities` as an additional feature.
- Using different landmarks - possibly extracted from `name` feature using NLP.
- Analyzing the top reviewed listings and top hosts to gather insights on what makes a listing successful.

9.2. Applying models on other cities

- The models can be applied to other cities to see if the same features are important in predicting price.

9.3. Using different models:

- Try other models such as Support Vector Machines, Neural Networks, etc.

10. Conclusion:

In conclusion, the Random Forest model was able to predict the price of short-term listings with a reasonable degree of accuracy. The model can be used by the start-up company to optimize its pricing strategy and maximize its profitability. The model can also be applied to other cities to see if the same features are important in predicting prices.