Zivko Danicic

Final Report:
AirBnb Pricing Analysis & Prediction

Problem Statement

This Capstone Two project aims to develop a machine learning model that predicts the price of an Airbnb rental unit given the features of the property. The goal is to provide pricing strategies and informed decision-making in the Airbnb marketplace for both hosts and guests. Current and potential Airbnb hosts will have access to a reliable tool which estimates a fair market price for their rental properties, in order to accurately predict future projected revenues and subsequently their return on investment. Determining an appropriate price for an Airbnb rental unit can be challenging given the amount of various factors that can contribute to the price; such as the property type, location, amenities, seasonality etc. On the other hand, guests can utilize this model to understand the range in pricing they should be aiming for given the characteristics they're searching for in any given unit; ultimately guiding their purchasing decisions.

Data Wrangling

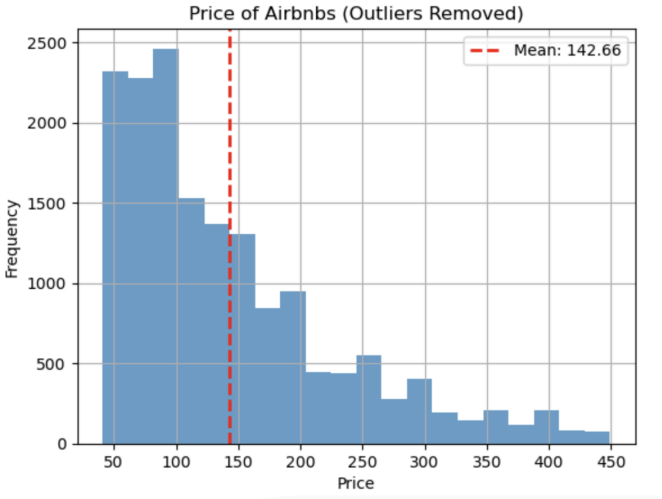
The raw dataset was taken from the official Airbnb website for the city of Toronto, which featured 75 columns and 17,997 unique rows. There was a breadth of data available to work with, however, it required ample cleaning and reduction in order to draw quality insights from it. To begin cleaning the dataset, first I took a look at what features contained missing values, and using some personal judgment, I dropped columns that either did not contain enough entries to garner strong insights and were unlikely to have an influence on the final predicted 'price' variable. Some columns such as 'bathrooms' needed to be converted to different data types to be made useful.

Some preliminary graphs and aggregations were used to get a sense for the distribution in some columns, this also helped develop a sense for the data and relation between columns. As a result, some columns with NaN values were able to be imputed in order to preserve the feature column and apply it to our model. Similarly, new columns were created to aggregate and clean previous features, an example of this was on 'host_acceptance_rate' in which NaN values were replaced with the mean and the previous column was dropped to avoid redundancy. Exploring the data further, an important column that contained the neighborhood's in which Airbnb rentals existed in Toronto, contained over 140 unique entries. To simplify this, a separate dataset that contained 6 boroughs across Toronto was uploaded, merged and matched to the corresponding neighborhood within a borough. This resulted in a smaller grouping of locations that would provide more effective analysis.

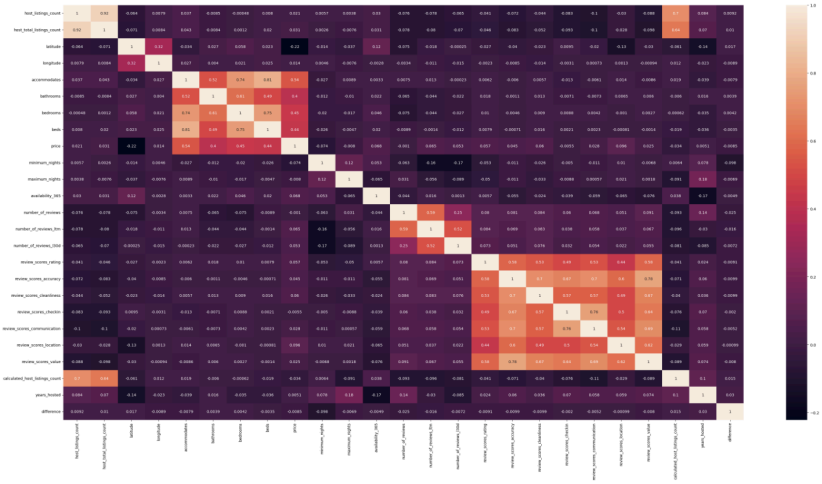
The goal of the initial wrangling stage was to gain a sense of the data, and clean the feature columns in order to prepare it for the EDA stage. More wrangling/cleaning would take place at later stages in the process as well.

Exploratory Data Analysis

In this step, I'm looking for correlations between the feature columns and the target variable, as well as correlation between the columns themselves. To begin the EDA process, we truncated the data based on extreme outliers in the price column, removing 5% of the data contained in the upper and lower bounds. Then I created a histogram of the price column to get a sense for the distribution and variation of the data, as well as the mean price across all units:

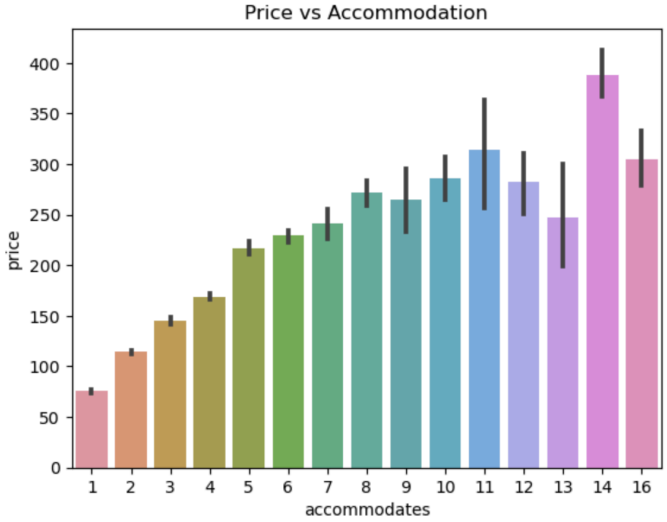


To give us a sense of how the features are correlated, and to quickly see which columns initially show strong correlation to price, I plotted a correlation heatmap:



(heatmap can more visibly be seen in Github)

The heatmap reveals some strong correlation to price with columns such as ‘accommodates’, ‘bathrooms’ and ‘bedrooms’, which fits the initial assumption that an Airbnb rental unit will be able to charge more given greater occupation. I explored the relation with the accommodation column further:



There is a clear positive relationship between the number of occupants and the price of a rental unit. Further, a pairplot along with the heatmap shown previously, were used to confirm that the number of occupants correlated with ‘beds’, ‘bedrooms’ and ‘bathrooms’ as they facilitate the ability to host more occupants. Other columns were tested in this process to search for potential influence on price. Interestingly, the box plot indicated that being a superhost did not have a significant impact on the price. Similarly, the borough a unit was located in had much less influence on price than initially assumed.

Modeling and Analysis

Given the dataset and the mission of predicting the price of listings, I chose to utilize Linear Regression, Ridge and Lasso Regression, and the Random Forest Regressor as my models to predict our target variable and extract important features.

The best performing model was the Random Forest with an R-squared value of 0.55. This means that approximately 55% of the variability in the price is explained by features included in the Random Forest model, which off the bat may suggest there is room for improvement, given that 44.8% of the variability in the price is unexplained. However, to better understand the context and interpretation of the model’s performance, we need to compare it to a baseline model to have a relative assessment. That is; we want to know if a stakeholder not utilizing any machine learning model would be better or worse off predicting the price.

In order to do this, I constructed another Random Forest model to establish a baseline, employing a max_depth of 2 as an approximation of the predictive capability comparable to making estimates without utilizing a model. The results are the following:

Model 1 - Best Performing Random Forest Model:

- R-squared: 0.5520419259035176
- Mean Absolute Error: 39.843021907933604

Model 2 - Random Forest Model (Max Depth 2):

- R-squared: 0.30458910596373046
- Mean Absolute Error: 53.06985544361114

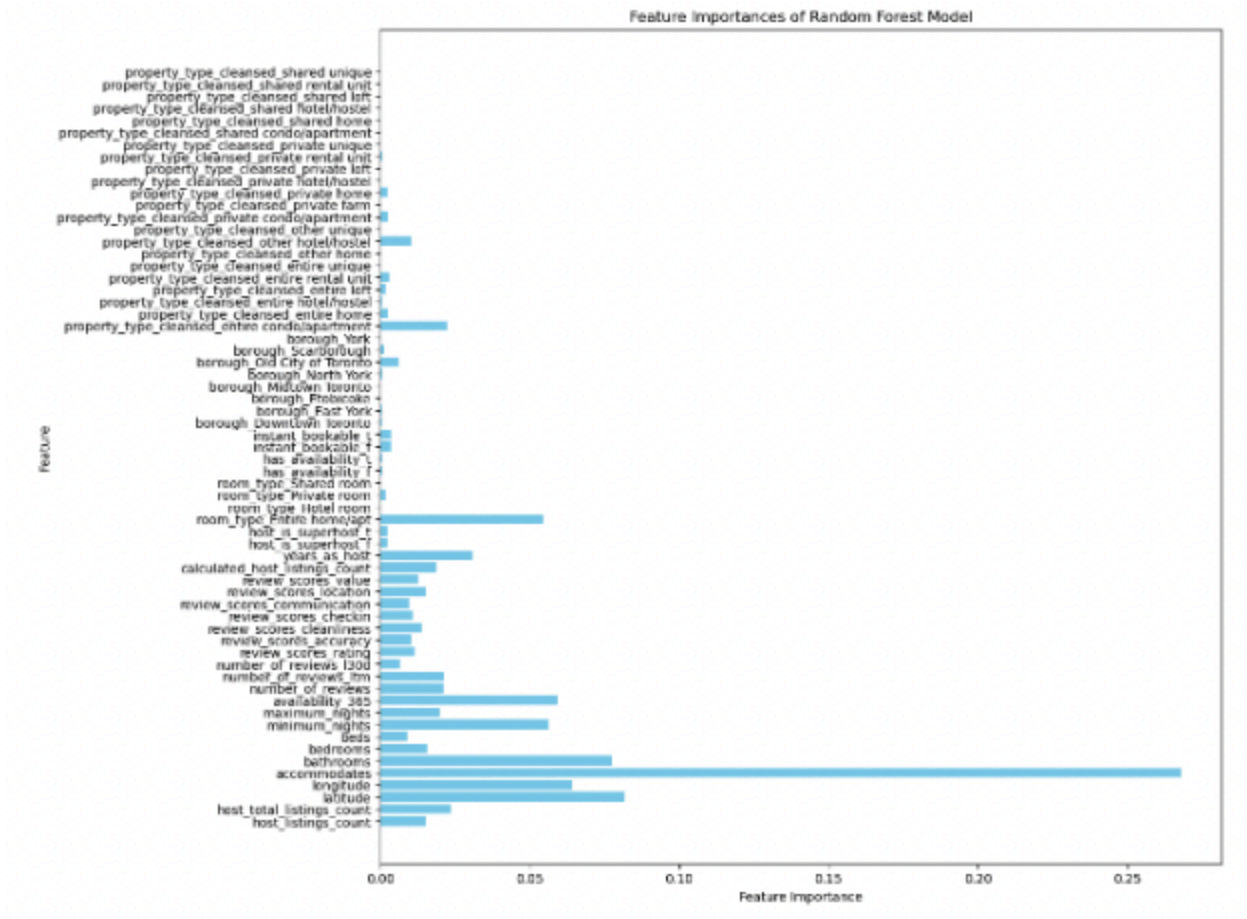
Model 1 outperforms the latter by 25% in explaining the variance in the dependent variable. Similarly, it also shows better accuracy in its predictions with a lower MAE of 39.84, a difference of 13.22. For context, stakeholders leveraging the model could optimize their pricing strategy more effectively compared to those neglecting the model. This discrepancy in application could translate into a $13.22 variation in unit price per night. Such a disparity holds substantial ramifications for return on investment (ROI) over the long term, potentially accumulating to thousands of dollars across multiple years.

Extracting feature importance from Model 1 provides insights into the areas where stakeholders and potential investors should allocate their resources to optimize property pricing and, consequently, enhance their returns.

Top 10 Feature Importances from Random Forest Model – Model 1:

	Feature	Importance
4	accommodates	0.268000
2	latitude	0.081711
5	bathrooms	0.077488
3	longitude	0.064219
10	availability_365	0.059364
8	minimum_nights	0.056243
25	room_type_Entire home/apt	0.054655
22	years_as_host	0.030805
1	host_total_listings_count	0.023604
41	property_type_cleansed_entire condo/apartment	0.022159

To get a better visual understanding of the Feature Importance:



Unsurprisingly, the model suggests that stakeholders should prioritize facilitating the amount of occupants a rental unit can accommodate over other potential features in order to maximize profits. This aligns with the intuitive understanding that more occupants can lead to increased revenue opportunities. However, the model underscores the importance for stakeholders to strategically utilize the space of their rental units. Rather than focusing on adding additional features that hosts might consider relevant, priority should be given to accommodating occupants effectively. As observed in our previous heatmap, accommodation exhibits strong correlations with the number of beds (0.81), bedrooms (0.74), and bathrooms (0.52). Based on these correlations, stakeholders should prioritize optimizing the number of beds, bedrooms, and bathrooms in their rental units to enhance the accommodation capacity. This strategic approach not only aligns with the model's emphasis on accommodating more occupants but also ensures that the rental units are tailored to meet the primary needs of potential renters.

While the model highlights latitude and longitude (indicating location) as relatively important features, their importance values are not overwhelmingly high. This could

imply that the Toronto rental market exhibits stability across the city, with no specific location exerting a dominant influence on rental prices. Availability throughout the year and minimum nights collectively contribute to around 10% of the influence on price. Although they exhibit similarities without directly overlapping, the insights from these results suggest that stakeholders should prioritize offering flexibility in availability and minimum nights to optimize pricing.