Machine Learning Applications for Airbnb Data

Dhruv Shah, Jenn Hong, Setu Shah, Sonya Dreyer

- **Clearly state the problem you are solving.**

The problem of pricing and review prediction on Airbnb is a complex challenge that involves finding a balance between maximizing revenue for hosts, providing fair and transparent pricing for guests, and accurately predicting the quality of the accommodations. The goal of our project is to develop a predictive model that can accurately estimate the price of Airbnb listings based on a set of relevant features, and another model to predict the likely review outcomes for guests. These models help hosts set competitive prices for their listings and help guests to make more informed decisions by taking into account predicted review quality. We aim to provide data-driven insights into setting and evaluating Airbnb listing prices and enhancing the review prediction system. This can lead to a more efficient and satisfying marketplace for everyone, offering a holistic approach to the Airbnb experience.

- **Explain what data sources you are going to use.**

Our dataset is called “Airbnb Listings & Reviews”, sourced from Maven Analytics, a well-known platform for high-quality datasets and data-related educational resources. Please find the dataset here.

- **Describe your dataset: how many rows, how many columns, what types of variables are included?**

Our dataset contains ~280,000 rows and 33 columns. The types of variables include numeric, categorical, date, and string. We will clean the data by transforming the columns into their relevant data types, treating missing values and outliers, and performing exploratory data visualizations.

- **Demonstrate that you can load the data in Python, for example by showing a couple of interesting figures motivating your project and showing summary statistics.**

Please find our visualizations and summary statistics in the appendix. Figure 1 plots the heatmap of correlations of the numeric variables in our data. We observe that price, bedrooms, and accommodations have positive correlations, which makes sense as we expect prices to increase with more guests and bedrooms. Additionally, we see that different types of ratings are correlated as well.

Figure 2 represents the trend of increase in mean price and mean accommodations as the number of bedrooms increases. We notice an interesting dip in price and accommodations for properties with 8 bedrooms and 21 bedrooms. On the other hand, Figures 3 and 4 show that despite variations in price (USD) across cities, the average ratings range from 9.3 to 9.7 (out of 10) for each review type. We will investigate this interesting relationship between reviews and price in our project.

- **What are the anticipated results? What type of analysis do we plan to do?**

In our study, we plan to develop two distinct predictive models aimed at understanding Airbnb pricing and listing review ratings.

Firstly, we wish to develop a model focused on pricing prediction, where we anticipate that the key features returned by our analysis would be on factors regarding location (city, neighborhoods), listing size (number of bedrooms, property type, and room type), listing ratings, host metrics (including host tenure, location and responsiveness), and provided amenities.

Our approach here will begin with rigorous preprocessing steps, including data cleaning, normalization of variables, hot encoding, and splitting into test and train sets. Then, we will perform feature selection through cross-validation to identify the most relevant features before we fit our models to the training sets (also with cross-validation). The models under consideration include Decision Trees (Random Forest), OLS Linear Regression, Multiple Linear Regression, Support Vector Regression, and Gradient Boosting Algorithms. Each model will be measured on metrics like R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which we will refine with hyperparameter tuning and regularization techniques like Ridge and Lasso Regression. After obtaining the final metrics for each model, we will select the best performing model, refit it on the entire training dataset and use that to predict on our hold-out test set.

For the classification prediction of listing review ratings, our analysis will also undergo similar preprocessing steps and feature selection processes. We anticipate the results will highlight different features, with more focus likely to be placed on things like superhost status, host tenure, response rate, acceptance rate, and host identity verification. The classification models we plan to explore here include Logistic Regression, Support Vector Machines, and Naive Bayes. As with the pricing prediction model, we will employ regularization techniques and hyperparameter tuning for enhanced performance. Evaluation metrics for this classification task will include Accuracy, Precision, Recall, Area Under the ROC Curve (AUC-ROC), and Confusion Matrix. Generally, we expect the workflow to be similar, just with different models and metrics being used for evaluation.

- **What are the potential implications of your results? How can they be used in practice? Why is the project worth undertaking?**

The results of the Airbnb Pricing and Review Prediction Project can have significant implications for various stakeholders and the hospitality industry. For hosts, accurate price and review predictions can lead to improved hosting strategies, revenue optimization, efficient resource allocation, and better guest experiences. As for guests, they benefit from finding accommodations that fit their budget and preferences, resulting in greater satisfaction, trust, and loyalty in the Airbnb platform. Airbnb can also use the insights for data-driven decision-making to set standards for pricing and rating transparency.

In practice, hosts can use the results of the project to maximize their earnings by adjusting prices to match supply and demand. They can also consider market variations to keep prices aligned with external conditions. In terms of reviews, hosts will understand which factors most influence reviews and can make adjustments to improve their chances of receiving positive feedback. Guests, on the other hand, can use the results of the project to make more informed decisions by selecting accommodations that align with their budget, preferences, and expectations. Airbnb can leverage the project's findings to develop features for both hosts and guests, allowing them to make data-driven decisions and maintain its competitive advantage in the market.

This project is worth undertaking because it addresses significant challenges in the Airbnb system and provides valuable solutions for hosts and guests. It improves the overall user experience by making pricing and reviews more transparent and efficient. Additionally, it supports Airbnb's growth by providing insights for platform improvement and data-driven decision-making. It is also important to mention that the project contributes to the discussion on pricing policies and review transparency, influencing the future of short-term rentals and how they are priced and managed globally. In summary, the project's potential implications are extensive and can have a significant positive impact on the Airbnb platform, its users, and the overall hospitality industry.

Appendix

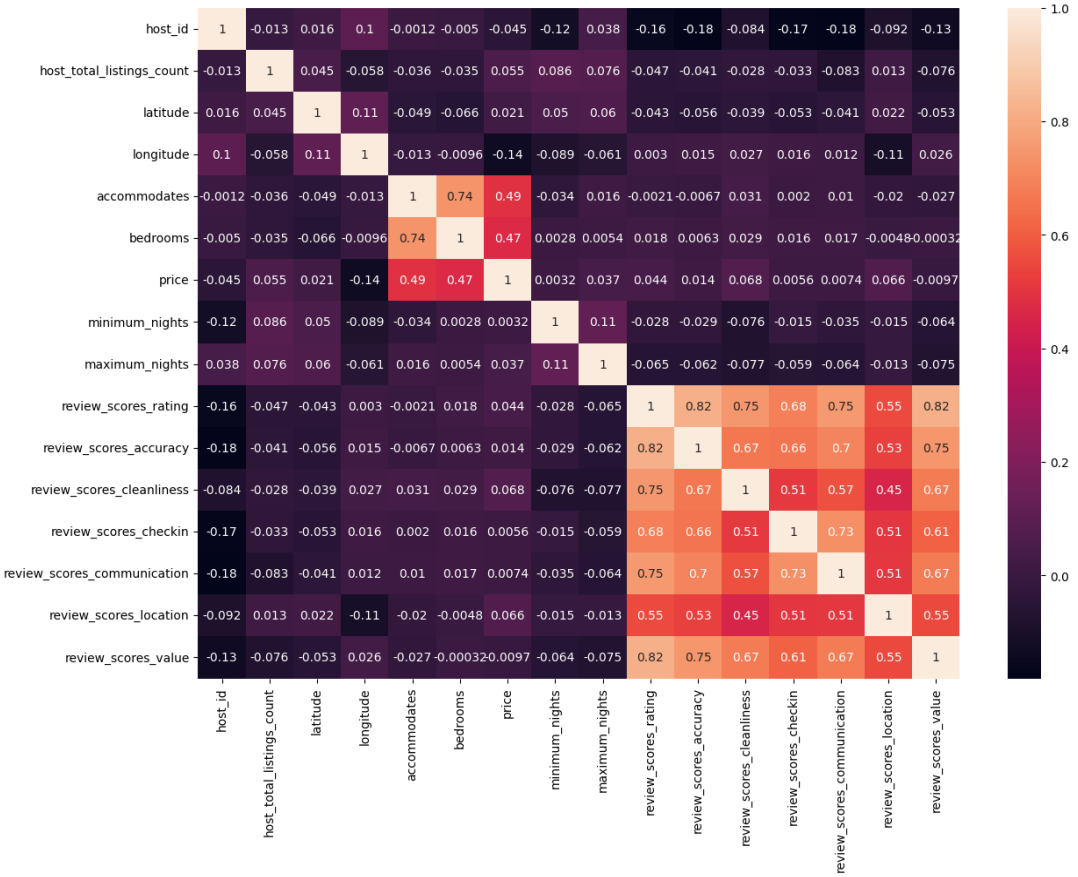

Fig. 1: Heatmap of Numerical Variables

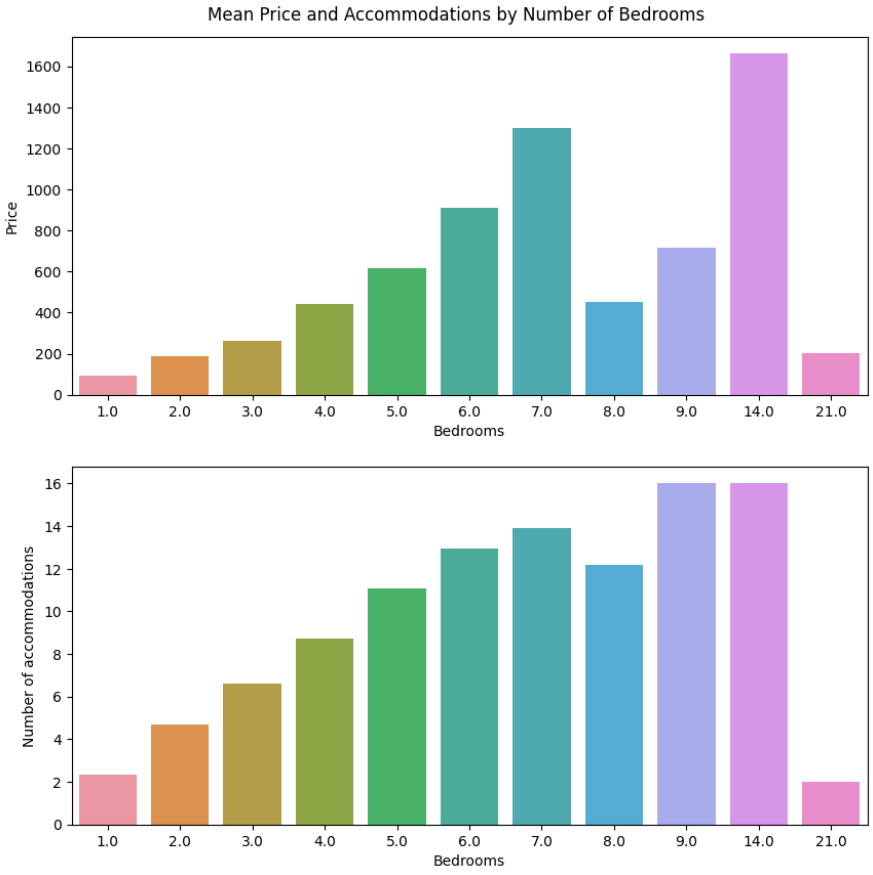

Fig. 2: Mean Price and Accommodations by Number of Bedrooms

Fig. 3: Mean Listing Price (USD) by City

	count	mean	std	min	25%	50%	75%	max
host_total_listings_count	279547.0	24.58	284.04	0.00	1.00	1.00	4.00	7.235000e+03
latitude	279712.0	18.76	32.56	-34.26	-22.96	40.71	41.91	4.890000e+01
longitude	279712.0	12.60	73.08	-99.34	-43.20	2.38	28.99	1.513400e+02
accommodates	279712.0	3.29	2.13	0.00	2.00	2.00	4.00	1.600000e+01
bedrooms	250277.0	1.52	1.15	1.00	1.00	1.00	2.00	5.000000e+01
price	279712.0	608.79	3441.83	0.00	75.00	150.00	474.00	6.252160e+05
minimum_nights	279712.0	8.05	31.52	1.00	1.00	2.00	5.00	9.999000e+03
maximum_nights	279712.0	27558.60	7282875.16	1.00	45.00	1125.00	1125.00	2.147484e+09
review_scores_rating	188307.0	9.34	1.01	2.00	9.10	9.60	10.00	1.000000e+01
review_scores_accuracy	187999.0	9.57	0.99	2.00	9.00	10.00	10.00	1.000000e+01
review_scores_cleanliness	188047.0	9.31	1.15	2.00	9.00	10.00	10.00	1.000000e+01
review_scores_checkin	187941.0	9.70	0.87	2.00	10.00	10.00	10.00	1.000000e+01
review_scores_communication	188025.0	9.70	0.89	2.00	10.00	10.00	10.00	1.000000e+01
review_scores_location	187937.0	9.63	0.83	2.00	9.00	10.00	10.00	1.000000e+01
review_scores_value	187927.0	9.34	1.04	2.00	9.00	10.00	10.00	1.000000e+01
usd_price	279712.0	111.38	424.27	0.00	34.59	64.85	110.24	1.270939e+05

Fig. 4: Summary Statistics