# DSCI 550 - Final Project Report

# Predicting Optimal Pricing for Airbnb Listings in Los Angeles City

**Tejas Rajurkar, Shivaani Kabilan, Qasim Riaz Siddiqui**

## 1. Problem Definition and Background

A surge in distance learning, remote work and a return towards traveling after the pandemic have led to an increase in consumers' demand for rental properties - especially those listed by Airbnb [1]. Given this surge in demand, and indeed supply, of rentals, property owners must choose optimal prices to rent out their properties to remain competitive in the market. Price determinants in the digital age of Airbnb have also been examined by researchers looking to analyze this growing industry [2]. Hence, it is important to solve this problem of predicting optimal prices for Airbnb listings. In a related work [3], wherein warehouse rental prices are estimated using relevant features such as warehouse size and location, machine learning models like Linear Regression, Regression Tree, Random Forest Regression and Gradient Boosting Regression Trees were used. Given this context, our team is focusing on answering the specific question of: "**Given certain attributes of a rental property in the City of Los Angeles, what is the most competitive price to list it on Airbnb?"**

## 2. Dataset

The raw dataset that we are using has been acquired from 'Inside Airbnb', a website offering web-scraped Airbnb listing details categorized by different cities and neighborhoods. Our dataset has detailed information on Airbnb listings in Los Angeles. Whilst there is data on 45,815 properties, we will narrow it down to 19,474 properties listed in the Los Angeles City to keep our analysis more specific. This dataset will help us to answer our aforementioned question, not only because it has information regarding different types of properties (including entire homes/apartments, private, and shared rooms) but also data on a plethora of attributes, continuous, categorical and even text - 75 columns. There is important data on the property host with information on their names, locations, ratings and response rates. Further, details about the properties themselves include its neighborhood, amenities provided, number of bedrooms, latitude and longitude and even a variety of ratings provided by previous tenants giving us ample details to run our analyses and find answers to our question.

## 3. Data Preprocessing

We began our analysis by working on data cleaning and preprocessing making sure our features and target variables are in accordance with what we require for our machine learning models. These steps have been highlighted below.

### 3.1 Transformation of Target Variable (Price)

Given that the aim of our project is to predict the optimal prices for the properties, we began by exploring the target variable *price*. After preprocessing this variable, we noted that the distribution in prices is heavily skewed with a majority of the property prices lying in the range of $20 - $1,000 per night. However, the data contains information on luxury apartments and mansions as well with prices that are much higher, going up to a maximum of $23,181 for a luxury villa. To deal with this highly skewed data, we transformed our target variable by taking a log-transform. This made the distribution of price much more symmetric as Figure 1 depicts below.

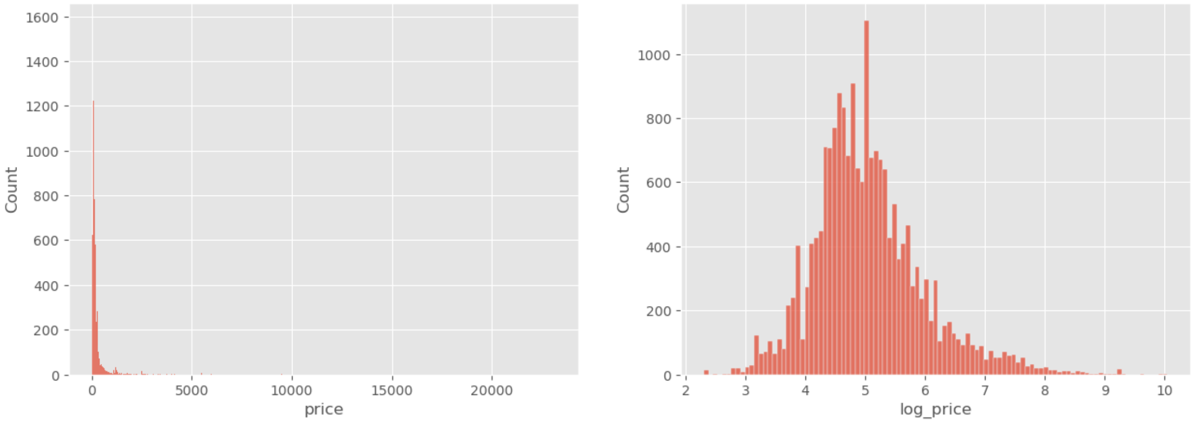

Figure 1: Transforming Price

### 3.2 Selecting Relevant Features

Next, we got rid of the features that only provided metadata information and did not contribute much to our predictive analyses, these features include *host_id*, *listing_url*, *picture_url*, *host_about*.

### 3.3 Handling Initial Missing Values

Initial exploration of our data showed us that certain variables contained too many missing values to be useful for us in our analyses. To deal with this issue, we simply got rid of the features for which more than 30% of the data points were missing. There were still other feature variables containing some missing values that were dealt with in other fashions. We mention this when discussing those features specifically.

### 3.4 Feature Transformation and Engineering

In this subsection, we discuss all the pertinent steps that were taken to transform the relevant columns in our dataset to be used for our models.

The amenities and verifications columns provided information on the property features and the type of verifications the property host had. To accommodate these variables in our regression models, we transformed them into numeric variables by considering just the total count of property amenities and host verifications provided.

Next, we looked at the boolean features in our data. These included columns such as *host_is_superhost*, *host_identity_verified* and *instant_bookable* which were handled by simply converting to 1s and 0s. Further, we also checked for any difference in means between the two categories for each of these variables to understand their importance.

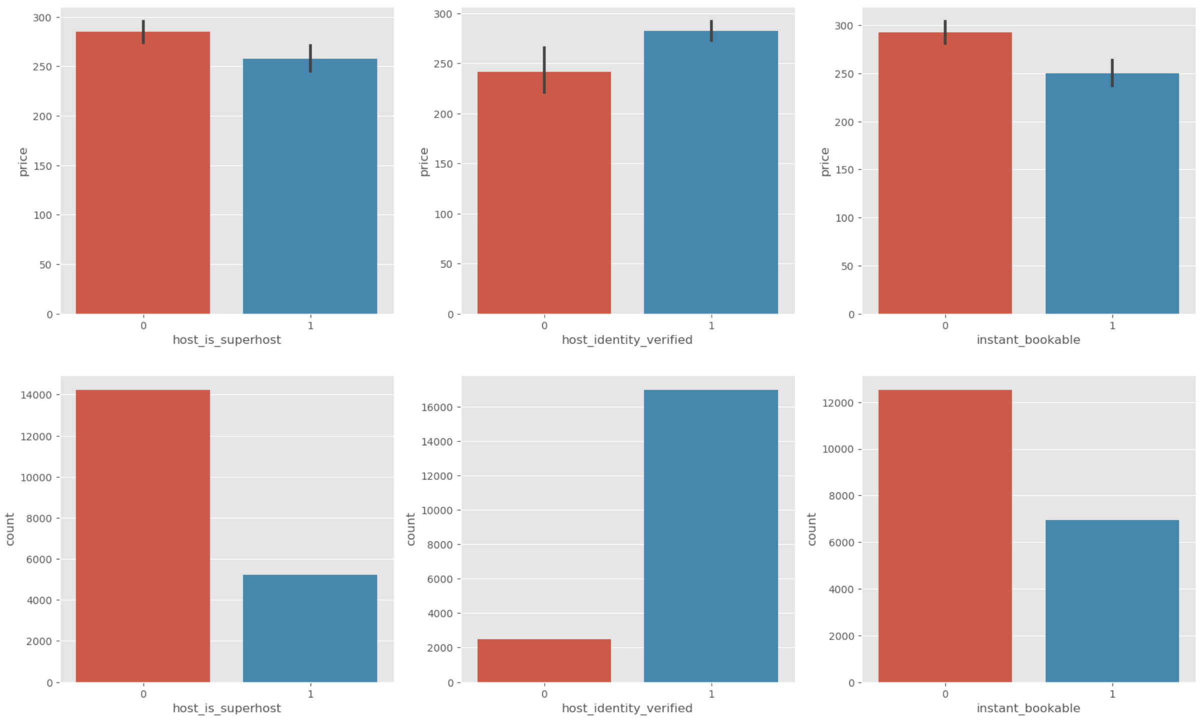

Figure 2: Boolean Variables

Through this, we could see some intuitive results such as the host having their identity verified raising the average property price or that the property being instantly bookable lowering the average price (assuming that a readily bookable property may not have high demand). However, we make any generalizations with a pinch of salt given the imbalance in the data points for the boolean categories in each of the features depicted in the second row in Figure 2 above.

Further, we looked into a couple of other important independent variables namely *room_type* and *neighbourhood_cleansed* which gave information on the exact property type on sale and the neighborhood where it was being sold. Firstly, we noted that the three different categories in the *room_type* variable (Entire Home/Apartment, Private Room, Shared Room) seemed to provide key insight into the price of the property as the illustration of Figure 3 below shows.

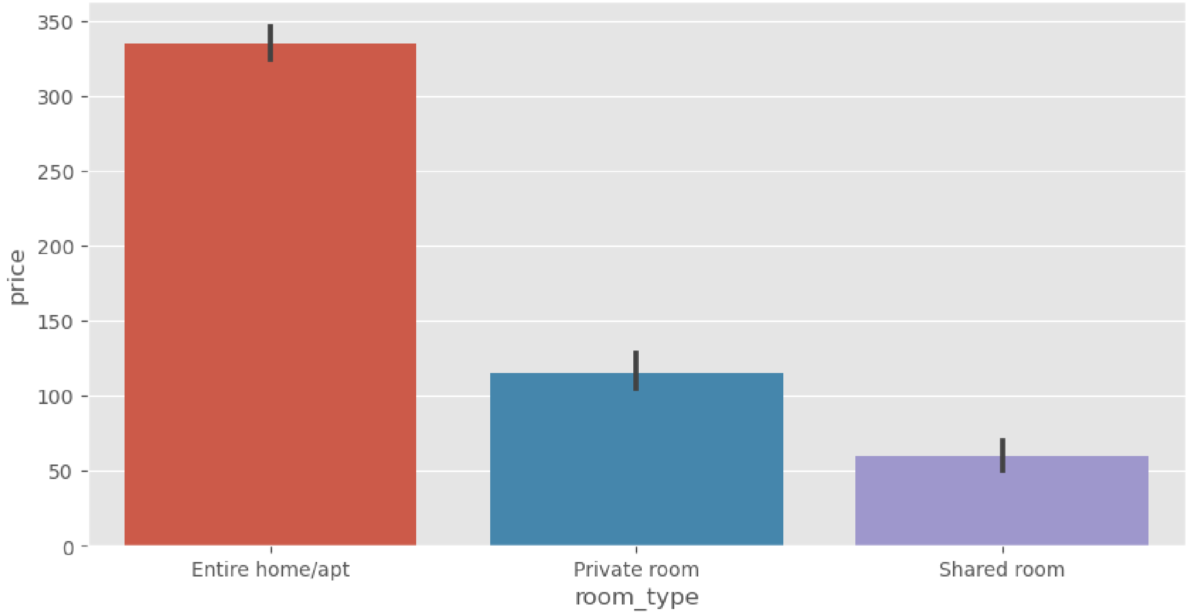

Figure 3: Room Types

Secondly, information on the location of the properties provided key insights as well.

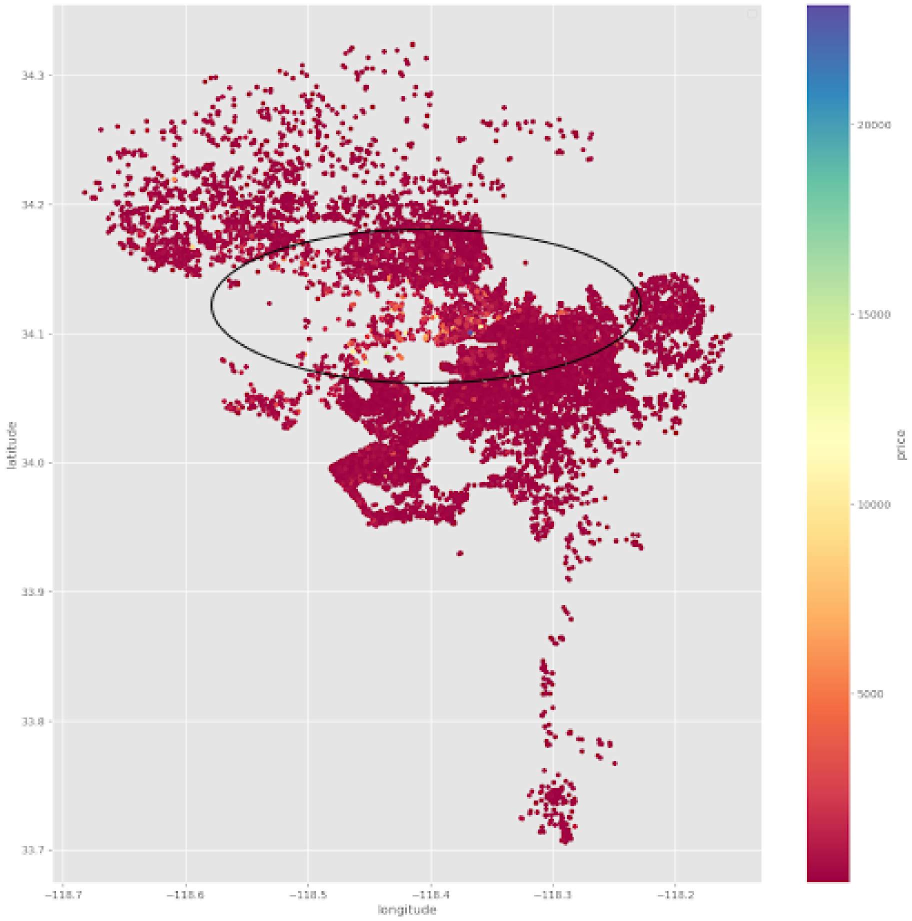

Figure 4: Prices with Neighborhoods

Figure 4 above shows that there is a horizontal strip in the region of latitude 34.1 where the property prices are relatively higher. This makes sense since that region corresponds to neighborhoods in the Beverly, Bel-Air, and Hollywood area that are generally considered more posh. Then, to use both of these variables for our regression, we make use of One-Hot Encoding by creating a total of 3 indicator variables for the *room_type* feature (one for each category) and 113 indicator variables corresponding to the 113 neighborhoods in the dataset.

Moreover, we also had data on user reviews on aspects such as cleanliness, checkin, communication and location. However, we noted that these features are highly intercorrelated. Therefore, we created a new feature *review_scores_average* that considered the mean of the other review ratings and also imputed the missing values using the median. We further preprocessed the texts in the columns *bedrooms* and *bathrooms* to determine their count in a particular property. To deal with missing values in these particular columns, we used another column called *accommodates* which has a proportional relationship with the count of bathrooms and bedrooms. Hence, we used group imputation based on the *accommodates* column.

Lastly, we used a couple of other features indicating the hosts' total listings, the latitude and longitude information of the property, and the number of days the user is available for in a year. Using these variables, we computed the correlation with the target variable and noted that the variables *accommodates, bedrooms, beds and number_of_bathrooms* were most strongly linearly correlated with price. However, eventually, we ended up using a total of 151 independent columns including 113 of the One-Hot Encoded indicator variables for the neighborhood feature.

## 4. Train-Test Split

Before running our different regression models, we split our data up into training and test sets. However, instead of a simple random split, we decided to stratify this split based on the neighborhood the property is in. This was done to avoid cases where, for example, our model does not train on properties where prices are generally lower and then ends up predicting for properties in expensive neighborhoods. A stratified, 80-20 split on the neighborhood group ensured that this case does not happen.

## 5. Methods

Given that our target variable is continuous, we made use of three main regression models to evaluate and compare. The three models we used were Multiple Linear Regression, Random Forest Regression and XGBoost Regression. Multiple Linear Regression was used initially since it is a more simple and interpretable model and to set a benchmark for comparisons with other more complex models. Then, the Random Forest Regressor was utilized as a more flexible model, without assuming any linear relationship between the predictor and predicted variables to evaluate whether results would improve given that other research has shown it to be more effective for more complex datasets [4]. Lastly, the XGBoost Regressor was made use of as another ensemble technique, given its efficacy with predicting continuous target variables, for comparison purposes. For all three regression models, we used the mean squared error as our cost function, which was iteratively minimized by each model to produce the best results possible.

## 6. Model Evaluation and Observation

### 6.1 Cross Validation

We computed the R2 score and Root Mean Squared Error (RMSE) as evaluation metrics to compare the results of each of these models. Here, we mention the fact that whilst we eventually used the log-transformed price as our target variable, we computed the RMSE by exponentiating our predicted results and comparing with the original prices. This was done to obtain a better interpretation for these values which is mentioned below. Further, for fairer comparison, we made use of 10-Fold Cross-Validation to compute the RMSE resulting from each model. The table of results with the model used, the R2 score, and RMSE of each model from 10-Fold Cross-Validation are provided below:

| **Model** | **R2 Score** | **RMSE** |
|---|---|---|
| Multiple Linear Regression | 0.679 | 475.79 |
| Random Forest Regressor | 0.761 | 433.87 |
| XGBoost Regressor | 0.754 | 442.78 |

Table 1: Model Results

From the results in Table 1 above, we see that, for instance, the RMSE of 475.79 for Linear Regression means that on average, our predicted model makes an error of $475.79 in determining the correct price of the airbnb listings. At first glance, this figure seems too high. However, we note that the values of the price variable in our data range from $10 to $23,181. Given the presence of extremely high values for price in our data, the RMSE makes more sense especially with the good R2 score that was achieved which, for this case, means that 67.9% of the variation in log price was explained by our Linear Regression model.

### 6.2 HyperParameter Fine-Tuning

On comparing the RMSE values for all three models, Random Forest Regressor performed the best with the lowest RMSE and highest R2 Scores. Therefore, we selected this model as our best model and applied Randomized Search to fine-tune some of the hyperparameters including the number of estimators and features used. The final model delivered an R2 score of 0.772 and RMSE of 428.88 which was a marginal but still an appreciable improvement over our original ensemble model. Hence, we re-trained this fine-tuned model on the entire training set and selected it as our final model.

## 7. Limitations and Conclusions

In conclusion, our observations showed that the fine-tuned Random Forest Regressor predicted the best prices for properties by resulting in the highest R2 score and lowest RMSE of 0.772 and 428.88 respectively. This meant that the model explained 77.2% of the variation in log price and made an average error of $428.88 - not that high given the extreme prices.

Here, we do acknowledge that our data contains only advertised property rates meaning may not be the exact prices the guests pay for. Further, the data is only a snapshot of the prices when the data was scraped and hence lacks any temporal information that may be used to perform seasonal analyses. Nonetheless, our investigation provided a good benchmark for property price predictions tasks and may be used to develop further research on.

## 8. References

[1] Ashley, Sara. “Airbnb's business is booming — and rates are rising.” *CNN*, 3 May 2022, https://www.cnn.com/2022/05/03/tech/airbnb-first-quarter-earnings/index.html. Accessed 4 December 2022.

[2] Wang, Dan, and Juan Luis Nicolau. “Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com.” *International Journal of Hospitality Management*, vol. 62, no. 2017, 2017, pp. 120-131, https://www.sciencedirect.com/science/article/pii/S0278431916305618.

[3] Ma, Yixuan, et al. “Estimating warehouse rental price using machine learning techniques.” *International Journal of Computers, Communications & Control 2018.*, vol.13, no.2, https://www.researchgate.net/publication/324512689_Estimating_Warehouse_Rental_Price_using_Machine_Learning_Techniques.

[4] Schonlau, Matthias, and Rosie Yuyan Zou. “The random forest algorithm for statistical learning.” *The Stata Journal*, vol. 20, no. 1, 2020, pp. 3-29, https://journals.sagepub.com/doi/10.1177/1536867X20909688.