Los Angeles Airbnb Analysis Capstone

Problem Statement

Airbnb has significantly changed the nature of the short-term property rental market. Not only it has allowed homeowners to earn extra revenue by listing their properties but also provided travelers with great and convenient places to stay. Running Airbnb requires lots of investment, it seems simple on the surface, but very involved and complicated in today's hospitality market. Encouraged by the idea of making money from my home when I am not living in it, in January 2022, I started traveling and working remotely and rented out my two-bedroom condo in Venice near Santa Monica Pier in the county of Los Angeles, often referred to by its initials L.A. So far, the revenue does not look promising due to the low occupancy rate. What has gone wrong? Now, that I have had experience being an Airbnb host for more than a year, I decided to use my data science skills to analyze the Airbnb data of LA to improve the revenue by 30% in 2023. The goal of the project is to answer the questions below:

- How are the listings distributed across the area of Los Angeles?
- How do the prices vary with respect to the neighborhood, property type, month, and day?
- Can machine learning be used to predict prices?

Data

The datasets were obtained from insideairbnb.com/get-the-data/ . I am interested in listings and calendar data sets for Los Angeles City.

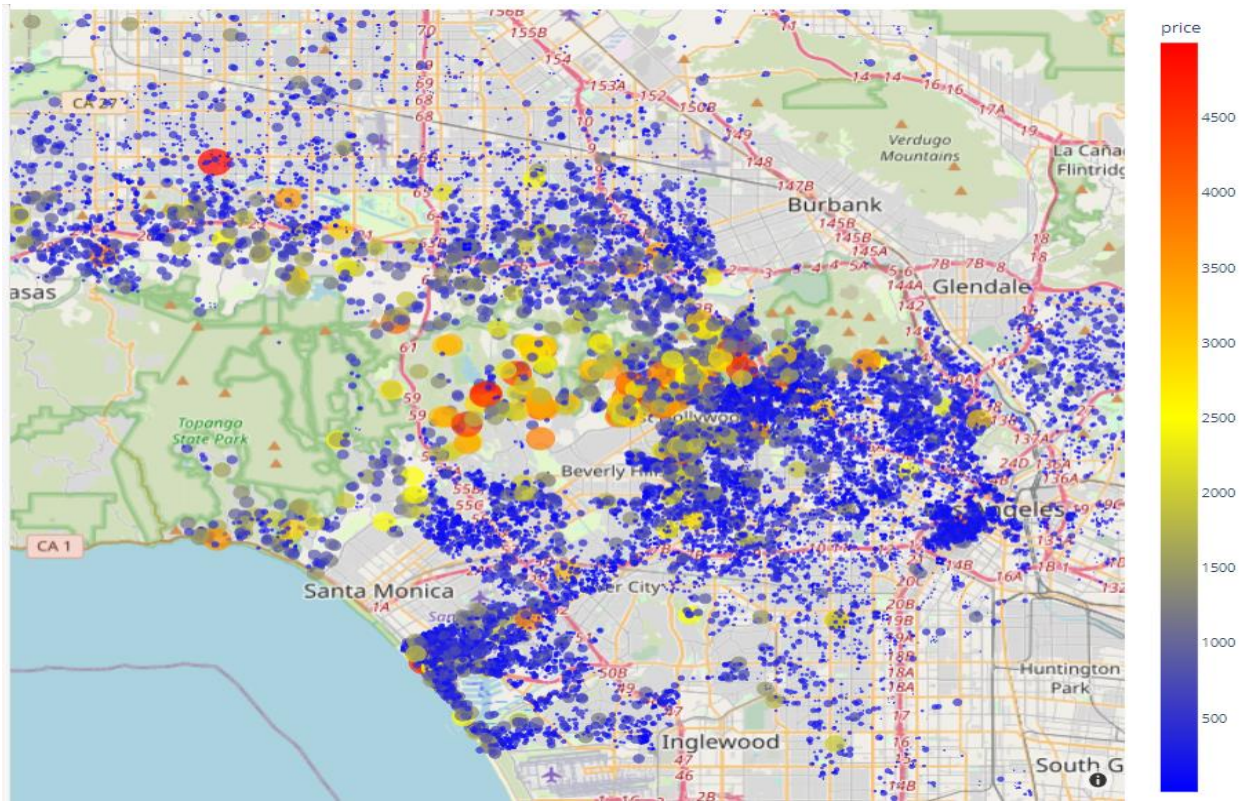
Data Wrangling

The listing dataset was 21,003 rows and 75 columns. After carefully reviewing the data, I decided to remove 41 columns due to no data or un-useful data to keep the columns at a manageable size that has good information for the solving problem. It does not make sense to keep listings with prices equal to zero or over $5,000 as they may skew the analysis, so I removed them. Also, I removed special characters in host-response-rate, host-acceptance-rate, and price and convert them from object to float. Split bathroom text into qty and type. By the end of data wrangling, the data shape is 20,993 rows and 34 columns. I did the same to the calendar dataset, converted the price to numeric, and split the date into month and day of the week

Exploratory Data Analysis (EDA)

Top expensive listings by city in LA

When it comes to finding a place to stay for any occasion, everything starts with a location so the very first thing I did was to create the beautiful color map below for all Airbnb listings in LA using scatter-map along with latitude, longitude, and prices.

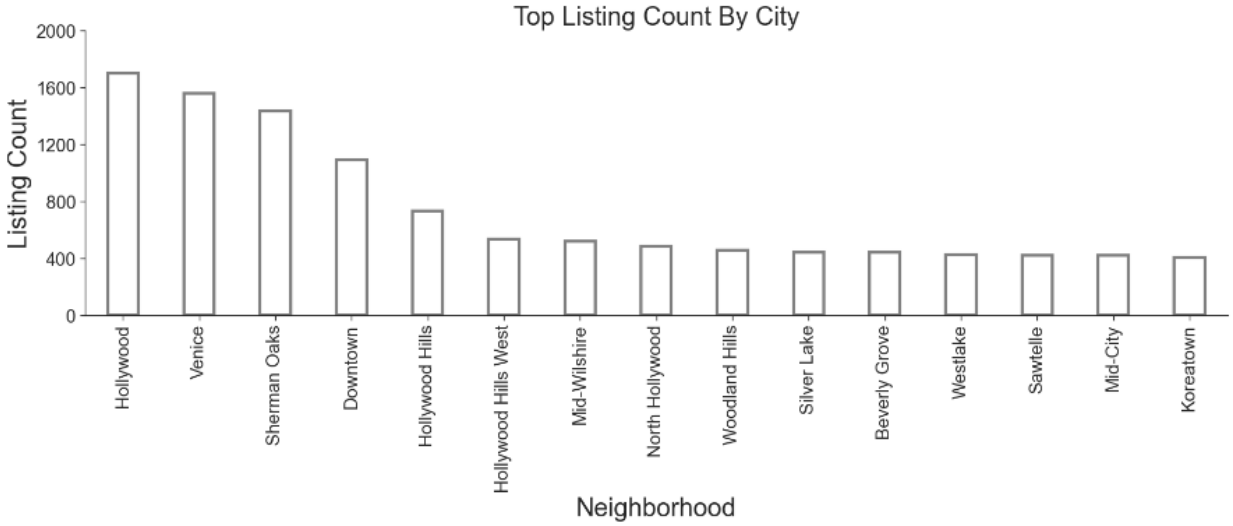


You may wonder why there are not many listings in the middle of the map. Because LA is both flat and hilly, there are not many properties on the hilly side, and if there are, they are very expensive like a couple of listings shown in big red, orange, and yellow circles with prices as high as five thousand dollars a day. That is insane, right? In the box plot below, you can find the listing price for the most expensive

listings in LA. Bel-Air, Beverly Crest, and Hollywood Hills West are the most prestigious upscale neighborhoods in the Western part of Los Angeles with very high standards of living. Who can afford to rent a property in these areas?

Top listing count by city in LA

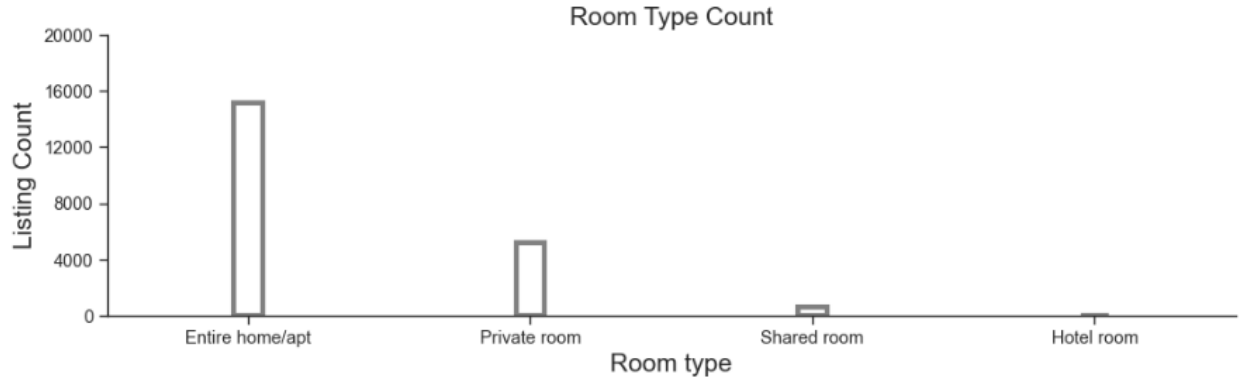
It is a challenge to choose where to stay in Los Angeles since it is one of the largest and most dynamic cities in the US and the main tourist attractions are scattered everywhere. Moreover, public transportation is not convenient here as it is in New York or San Francisco so almost everyone who lives in here uses a car for transportation. The plot below shows the top listing by cities with more than 300 listings. Let's dive into the top three cities and find out why they are popular with Airbnb.



Hollywood is the best area to stay in Los Angeles due to its location and security. It is the headquarters of the most famous film studio in the world, Universal Studios, and the Walk of Fame. The food is also good here. Venice is one of the most popular white sand beaches in Southern California with many things to do and see like street performers, colorful street art, trendy bars, and restaurants. Sherman Oaks is more affordable, has bigger properties, and is not too far from the main attractions

Room type

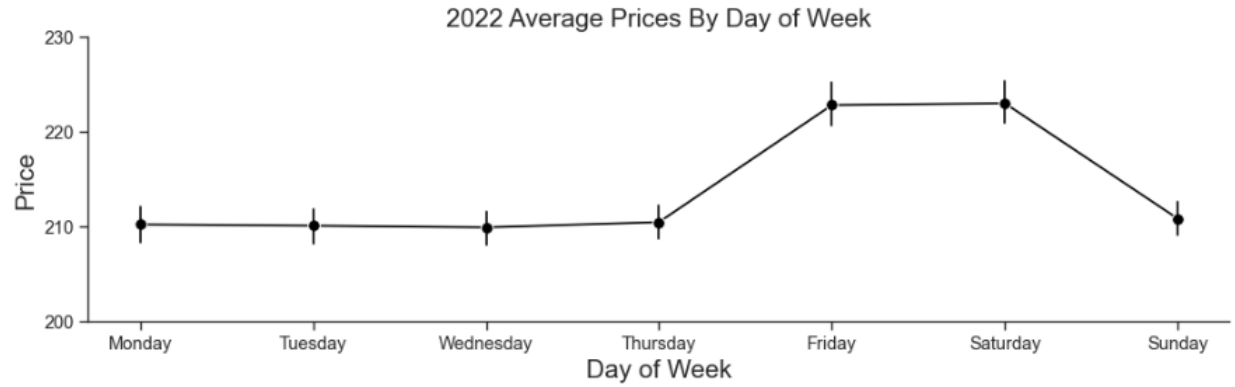
15,000 entire home/apt for rent? Do you see that number in the plot below? I was shocked by the number and could not stop thinking about how many of them belong to big investment corporates as more Airbnb listings mean higher rent and are bad for the housing market.

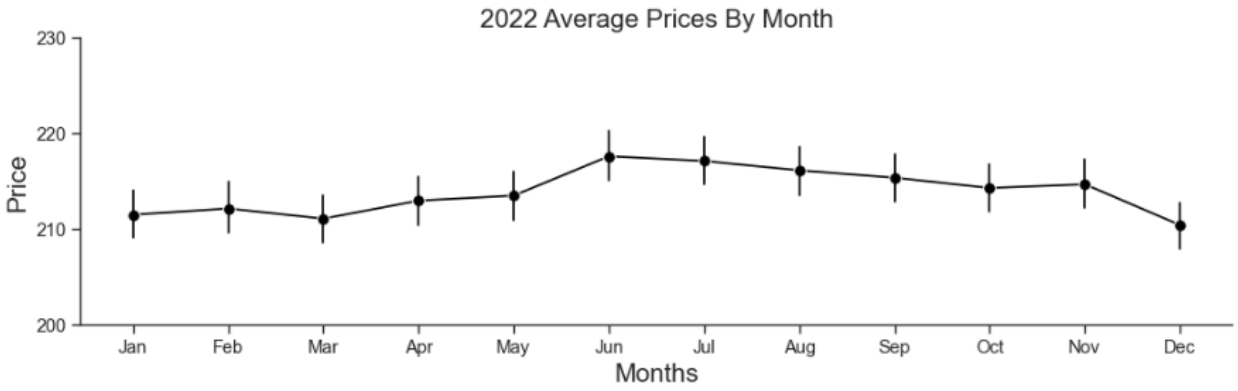


Anyway, the entire home/apt is the dominant room type which represents 70% of all types of rooms. Private room is the second most popular room type. In many expensive and high-demand cities such as New York and San Francisco, the research shows it may be possible to pay the entire rent on a two-bedroom apartment by filling one of the rooms for 21 days a month. Basically, it is rent-free for the hosts but keep in mind that lots of time commitment is involved.

I am surprised to see the hotel as one of the room types, even though there are not many listings, but it shows how popular Airbnb is that hotels even use it as an additional platform for advertising. Will drop all hotel listings for modeling but wondering how Airbnb impacts the hotel industry.

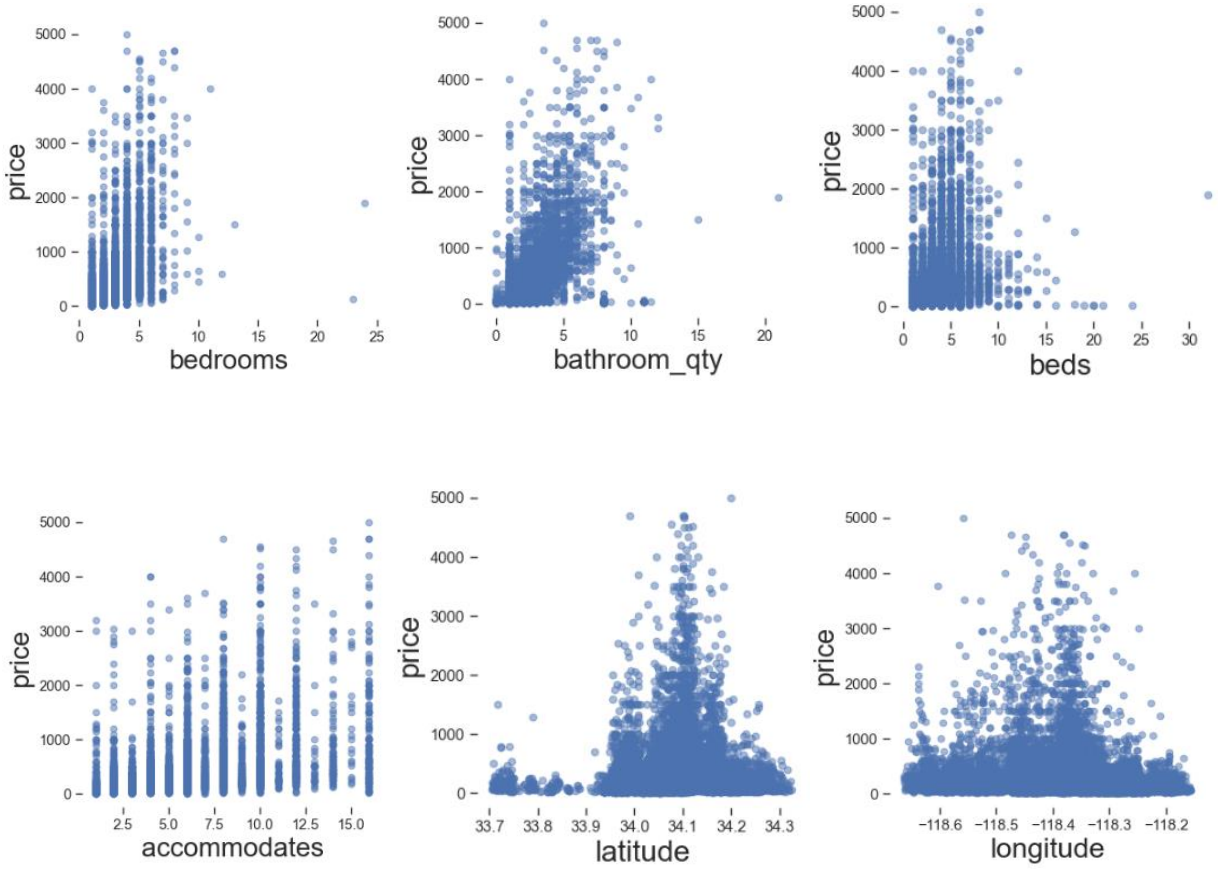
Should I be flexible with the price? Of course, I should.

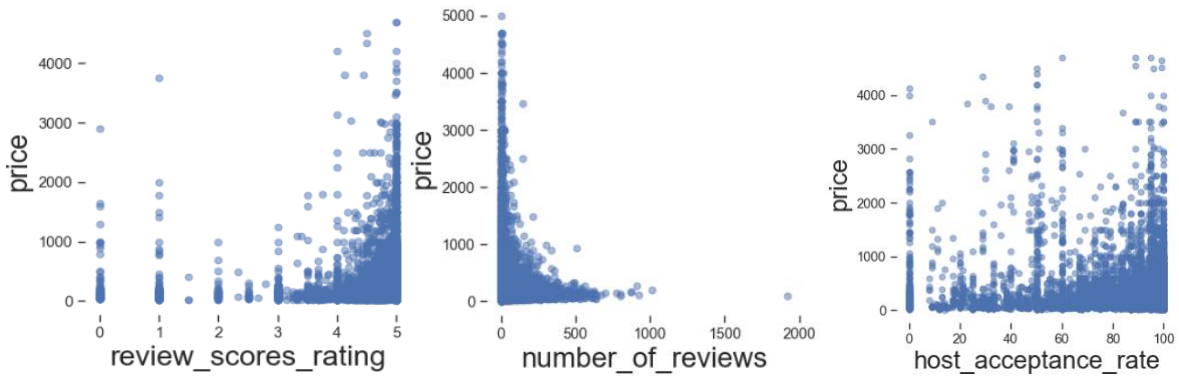




It intuitively makes sense to see the higher price for Friday and Saturday but why I don't see the price change much over the months. In the hotel industry, price is based on seasonal market demand. It is usually soft during school months then peaks in the summer months and the holiday season. LA is a sunshine city. It has two seasons, dry summer, and very mild winter. As a result, all tourist attractions are open year-round. I am wondering if that is the reason why the price doesn't fluctuate much from January to December. Despite the facts, I should set higher rates for the high season to maximize profit and lower prices for the low season to maximize occupancy.

Price Correlation





I used scatter plots to view the correlations between prices with all numerical features. Obviously, price correlate well with bedrooms, beds, bathroom-qty but not so much on accommodates. Even though the price does not show a linear correlation with latitude and longitude, but you know the location plays a major role in pricing. Moving on to the next step, I chose the features below for independent variables:

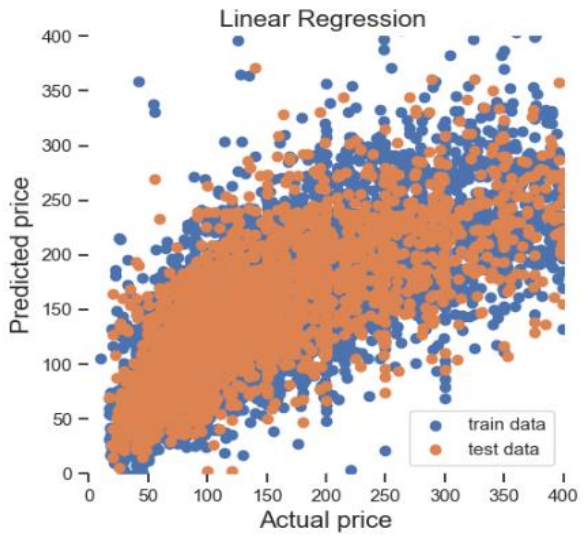
- latitude
- longitude
- neighborhood
- review-scores-rating
- room-type
- accommodates
- bedrooms
- beds
- bathroom-qty
- number-of-reviews

Feature Engineering

Recall during data wrangling, many missing values have been left untouched as we wanted to view the distribution from all listings. Now, with a good understanding of the data and their relationship. Let's transform the raw data into meaningful columns to prepare for modeling. Dropped rows with missing values in bedrooms and beds as they are two important features for pricing. Also, keep prices less than $400 as 99% of listings in this range plus I don't want the analysis skewed by the outliers. Next, I split the data into X and Y, got dummies for X, imputed missing data with median, and scaled all data using StandardScaler.

Modeling

Linear regression



I started the modeling with linear regression. As show in the plot above, there is a sign of a positive correlation between actual and predicted prices on both train and test datasets, however, r2 = 0.51 is not good. This means the model can only predict 51% of the relationship between price and independent variable price. I went back and adjusted the price range to $400 and removed listings with zero reviews and the r2 got improved from 0.51 to 0.54.

Lasso and Ridge

Lasso coefficients ranked from high to low

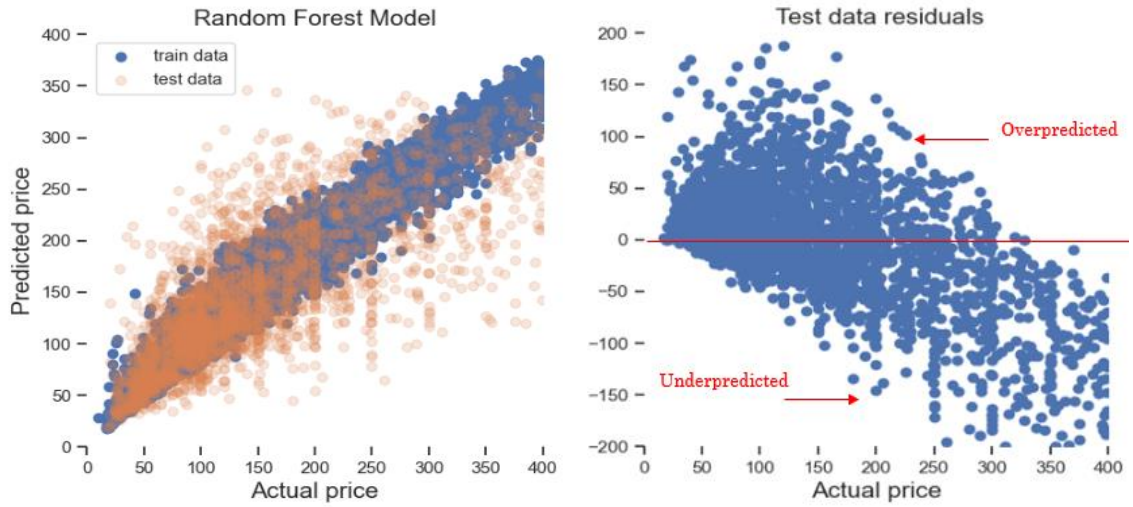
```
lasso=Lasso(alpha=0.01)
lasso_coef=lasso.fit(X_tr,y_train).coef_
order=np.argsort(np.abs(lasso.coef_))[::-1]
for i in order:
    coef_=lasso.coef_[i]
    if coef_>5:
        print(X_train.columns[i] + ", " + str(lasso.coef_[i]))
```

```
bedrooms, 27.847008624530567
room_type_Entire home/apt, 21.615449486313587
accommodates, 19.839178963669678
neighborhood_Downtown, 9.14686786441091
neighborhood_Venice, 7.621504832562207
neighborhood_Hollywood Hills, 7.3875255026857225
neighborhood_Silver Lake, 6.837357156239405
neighborhood_Hollywood Hills West, 5.915490538941012
review_scores_rating, 5.347658520298893
bathroom_qty, 5.054577608052501
```

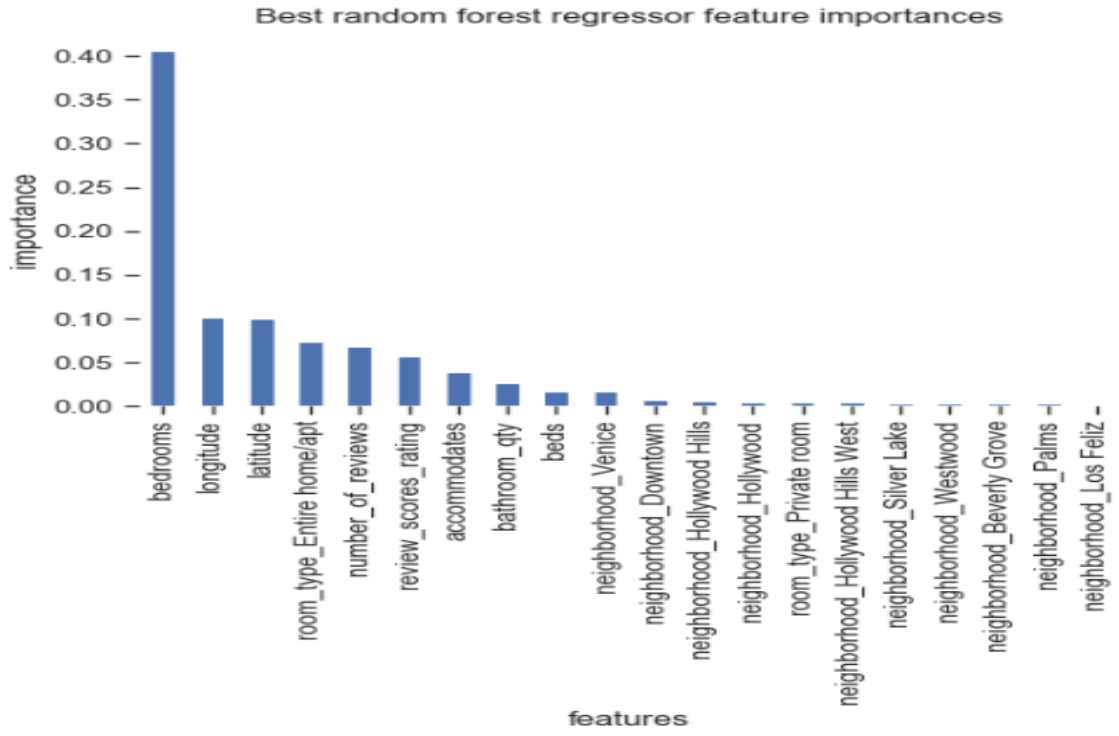
I then searched for the best alpha and applied them to Lasso and Ridge models and got similar results. The standard deviation of 5 runs seemed quite large. Lasso was able to pick up bedroom, room type,

and accommodates as the top three features. Downtown, Venice, and Hollywood were the top three neighborhoods.

Random forest regressor

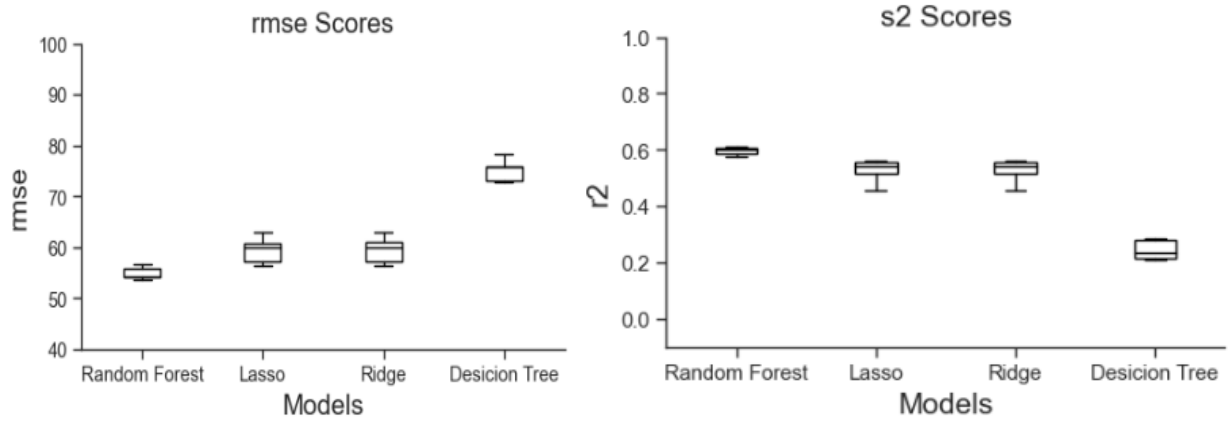


This is the fourth model. Look closely at the plot on the left, the training set has a very high positive correlation, but it is overfitted. To improve the performance, I did play with max-depth, min-samples-leaf, and min-sample-split but they did not help. Then I used grid search to go over different types of imputers and n-estimators values and observed better scores with an n-estimator equal to 1000 which of course will increase the execution time. The plot on the right above is the test data residuals. Half of the data is overpredicted and the other half is underpredicted especially for prices from $300 to $400.



The top three features from the model are bedrooms, longitude, and latitude. Even though, the neighborhoods are not in the top ten but in a way, latitude and longitude represent them. Glad to see the model can pick up room type, ratings, and reviews as important features as they are the features that we often pay attention to while searching for a rental on Airbnb

Cross-validation scores



Model conclusion

Which model has the best performance? I mainly used the rmse and r2 scores for comparison and observed that random forest is the best model which has the lowest rmse, highest r2, and smallest variation from cross-validation. Even though Random Forest is best, it is overfitted as r2 on the training data set is 0.94 while only 0.57 on the test dataset which means the model can only predict 57% of the unseen data. The rmse score equals 57 suggesting wide variation on the predicted price which could be averaging $57 off from the actual price. You can find the matrix scores for all the models in the table below.

	Linear regression	Lasso	Ridge	Random Forest
r2	0.51	0.54	0.54	0.57
mae	45.6	45.9	42.5	40.81
mse	3667	3700	3185	3244
rmse	60	60	62	56.9

Future Improvement

1/ Bedroom shows as the number one important feature for both linear and random forest models, I plan to hire an interior designer to convert the loft on the second floor to a small bedroom to increase the bedrooms count from two to three to target bigger groups as beach location is the ideal for parties and families get together

2/ Ideally set higher rates for the high season to maximize profit and lower prices for the low season to maximize occupancy so I will work with a property management company to re-evaluate my listing price.

3/ Good to know that my beach property in Venice is in the second most popular location in LA with the most visited by locals, and tourists. With the price set at $310 per day for two bedrooms house, the price seems reasonable when compared to properties nearby but the machine learning best model has been underpredicted for all prices from $300 to $400. With rmse equal to $57, my price can fluctuate from $250 to $367. I learned that It seems almost impossible to correctly predict the Airbnb price based solely on bedrooms, beds, room type, bathrooms, reviews, and locations which are no doubt important but from my experiences with Airbnb, the pictures of the properties are the feature that makes the final call. Two properties may have the same features, but one that is new and nicely decorated will, of course, cost way more than the older property. How can I add property pictures to my analysis to improve the prediction?