Exploration of LA's Airbnb Market

Zhiwen Cao

November 2022

1 Abstract

I wanted to predict the rental price given some features of the listing in LA area and investigate whether seasonality has an effect on the listing price. I used ANOVA and several machine learning models such as random forest and xgboost to achieve this goal. Using the quarterly data for the last 12 months, Sep 2021 to Sep 2022, from Inside Airbnb I found out in the last year seasonality has no effect on listing price in LA. Xgboost performs the best on the processed data, giving r2 value 0.68 and RMSE 398.1 on the testing data. Feature importance analysis from catboost model indicates number of bathroom, number of total listing counts, availability 365 and listing price level determined by neighborhood are prominent in determining listing price.

2 Background

Airbnb becomes more and more popular for short-term home and apartment rentals. One challenge that Airbnb hosts face is determining the optimal rent price. Although Airbnb provides the host with general guidance, there is no easy way to find the best price given the market is so dynamic. Airbnb pricing is important to get right, particularly in big cities like Los Angeles where there is lots of competition. No one will book if price is set too high, while potential income will be missed out if it is set too low. The aim of the project is to solve this problem by using machine learning to predict rental prices and find out the most influencing factors.

3 EDA

3.1 Cleaning

The data is scraped by a separate group named Inside Airbnb, containing all publicly available information about listings in LA. Variables such as 'listing_url', 'scrape_id' are not relevant and therefore are removed in the first place. The following list is all variables considered irrelevant to the project: 'host_id' 'listing_url', 'scrape_id', 'name', 'description', 'neighborhood_overview', 'picture_url', 'host_url', 'host_name', 'host_location', 'host_about', 'host_thumbnail_url', 'host_picture_url', 'host_neighbourhood','host_verifications', 'source','calendar_last_scraped', 'latitude','longitude','last_scraped',"neighbourhood_group_cleansed". Next, number of NAs in each remained columns is examined, shown in figure 1 below. All the columns with more than 30 percent value being NA are also dropped.

Because there are multiple columns related to listing counts, their correlation matrix is examined. 'host_listings_count' and 'calculated_host_listings_count' are highly correlated with 'host_total_listings_count', so they are dropped. Same analysis is performed on min max nights, availability and number of reviews.

All correlation plots are shown in figure 2. In the end, 'minimum_nights', 'maximum_nights', 'availability_30', 'availability_365' 'number_of_reviews_ltm', 'number_of_reviews_l30d' and 'number_of_reviews' are kept, while the rest are dropped.

'has_availability' and 'host_has_profile_pic' are categorical variables, both of which are extremely imbalanced. Because columns related to availability are already included and pictures are not the focus of study, both variables are dropped. I wanted to add an variable 'host_days_active' which is number of days between now and first day joining Airbnb, I had to remove rows with NA in 'host_since'.

id	0.000000
host_id	0.000000
host_since	0.000284
host_response_time	0.193823
host_response_rate	0.193823
host_acceptance_rate	0.173917
host_is_superhost	0.002139
host_listings_count	0.000284
host_total_listings_count	0.000284
host_has_profile_pic	0.000284
host_identity_verified	0.000284
neighbourhood	0.416086
neighbourhood_cleansed	0.000000
neighbourhood_group_cleansed	0.172193
property_type	0.000000
room_type	0.000000
accommodates	0.000000
bathrooms	1.000000
bathrooms_text	0.001550
bedrooms	0.082069
beds	0.017331
amenities	0.000000
price	0.000000
minimum_nights	0.000000
maximum_nights	0.000000
minimum_minimum_nights	0.000109
maximum_minimum_nights	0.000109
minimum_maximum_nights	0.000109
maximum_maximum_nights	0.000109
minimum_nights_avg_ntm	0.000109
maximum_nights_avg_ntm	0.000109
calendar_updated	1.000000
has_availability	0.000000
availability_30	0.000000
availability_60	0.000000
availability_90	0.000000
availability_365	0.000000
number_of_reviews	0.000000
number_of_reviews_ltm	0.000000
number_of_reviews_l30d	0.000000
first_review	0.230907
last_review	0.230907
review_scores_rating	0.230907
review_scores_accuracy	0.236778
review_scores_cleanliness	0.236757
review_scores_checkin	0.236909
review_scores_communication	0.236778
review_scores_location	0.236975
review_scores_value	0.236997
license	0.738164
instant_bookable	0.000000
calculated_host_listings_count	0.000000
calculated_host_listings_count_entire_homes	0.000000
calculated_host_listings_count_private_rooms	0.000000
calculated_host_listings_count_shared_rooms	0.000000
reviews_per_month	0.230907

Figure 1: proportion of NAs in each column

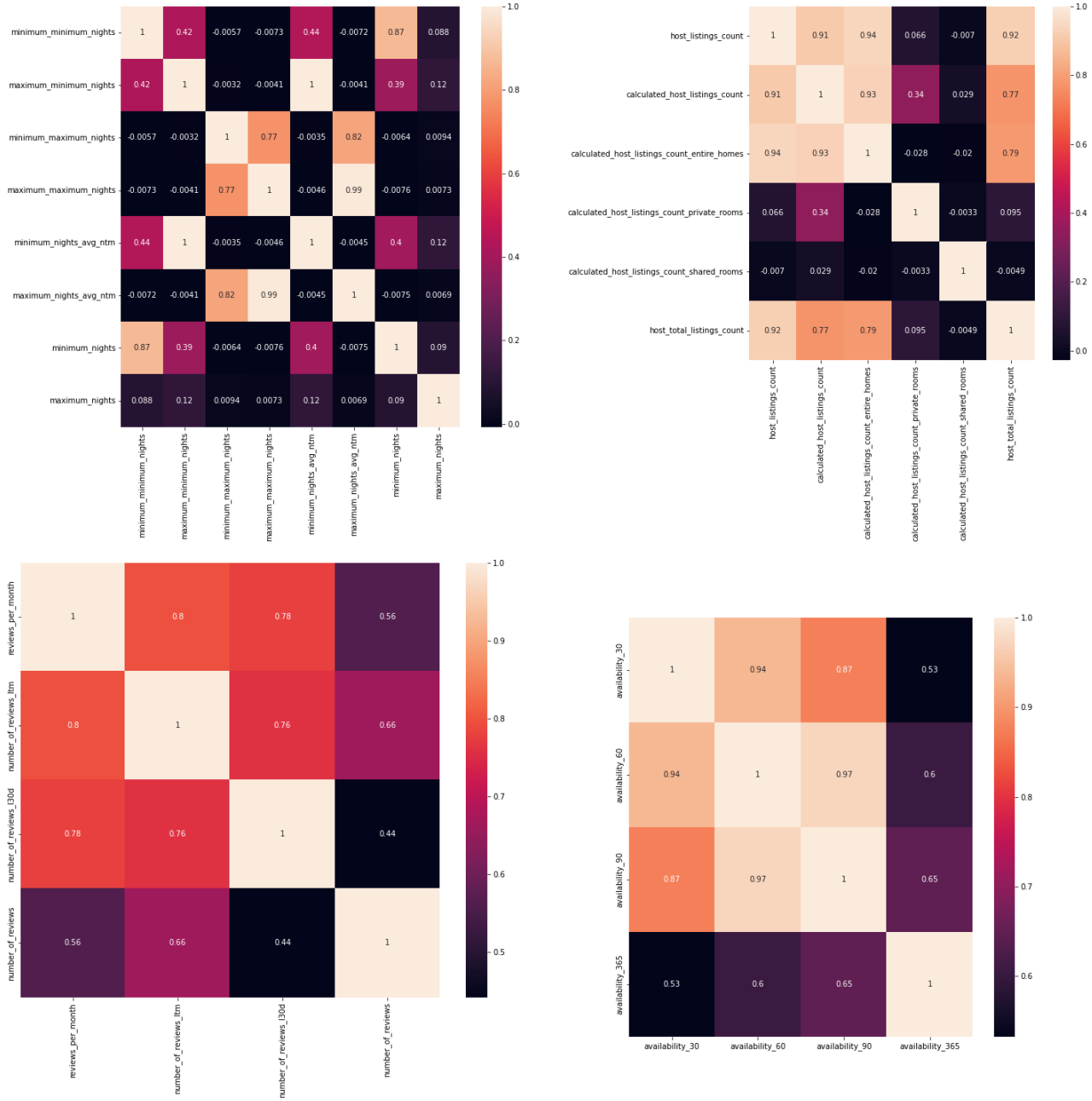

Figure 2: top left: min and max nights, top right: listings counts, bottom left: reviews, bottom right: availability

3.2 Augmentation and Transformation

As mentioned previously, 'host_days_active' is added. 20 percent of categorical variable 'host_response_time' is NA. I fill NA with the category 'unknown' instead of simply dropping rows because it may be important considering long response time will not satisfy urgent customer needs. Similarly, 20 percent of 'host_response_rate' is NA, and based on histogram shown in figure 3, it is grouped into several groups with NA being kept as its own category 'unknown'. 'host_acceptance_rate' is processed in the same way. Transformed variables are visualized in figure 4. 'bathrooms_text' is a string variable describing number and type of bathrooms. Number is extracted as a new column 'n_bathroom', and type is extracted as 'bath_type'. Bath type lack of either 'shared' or 'private' is given 'unspecified'. NA in 'bedrooms' and 'beds' are filled with corresponding median. 'amenities' is a list of additional features in the property. It is hard to determine which are more important than others or which combination has greater attractiveness. To simplify the process, the original 'amenities' is substituted with 'nun_amenities', indicating number of amenities provided in the listing.

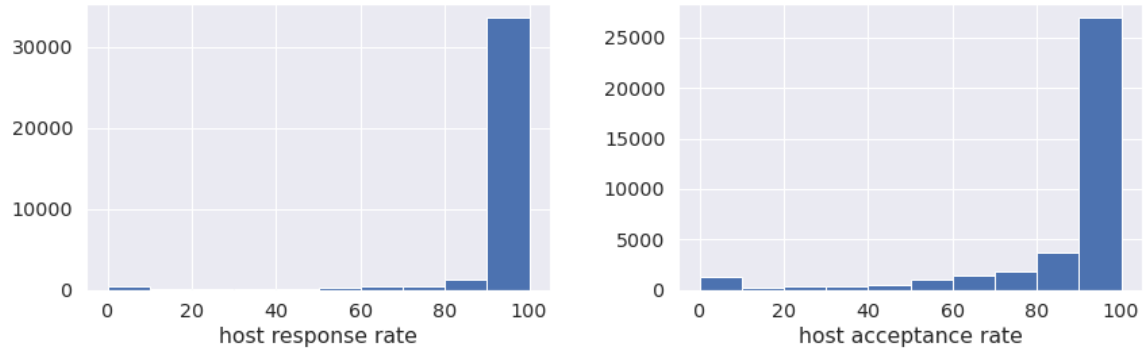

Figure 3: left: response rate, right: acceptance rate

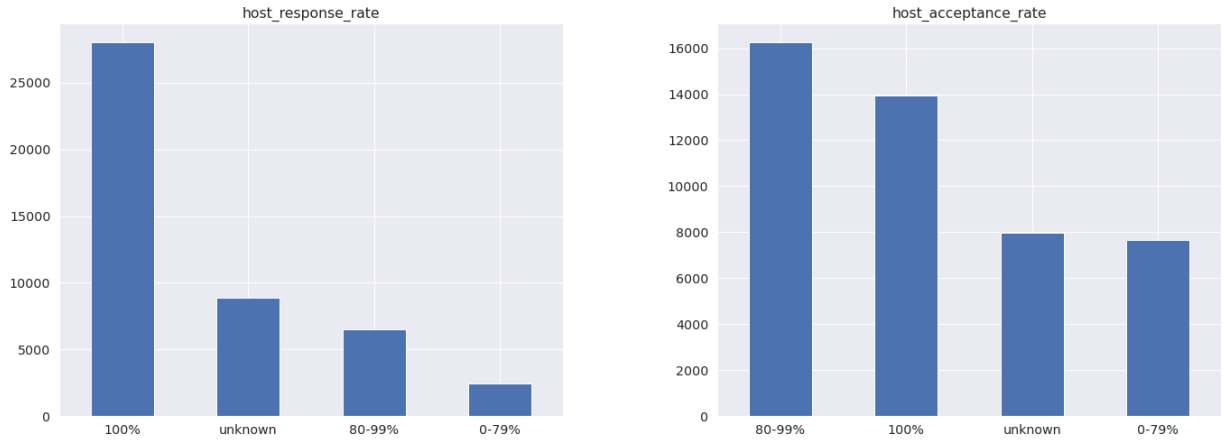

Figure 4: left: response rate, right: acceptance rate

'time_since_first_review' and 'time_since_last_review' are created in the same way as 'host_active_days', plotted in figure 5. Then they are transformed into categorical variables, plotted in figure 6.

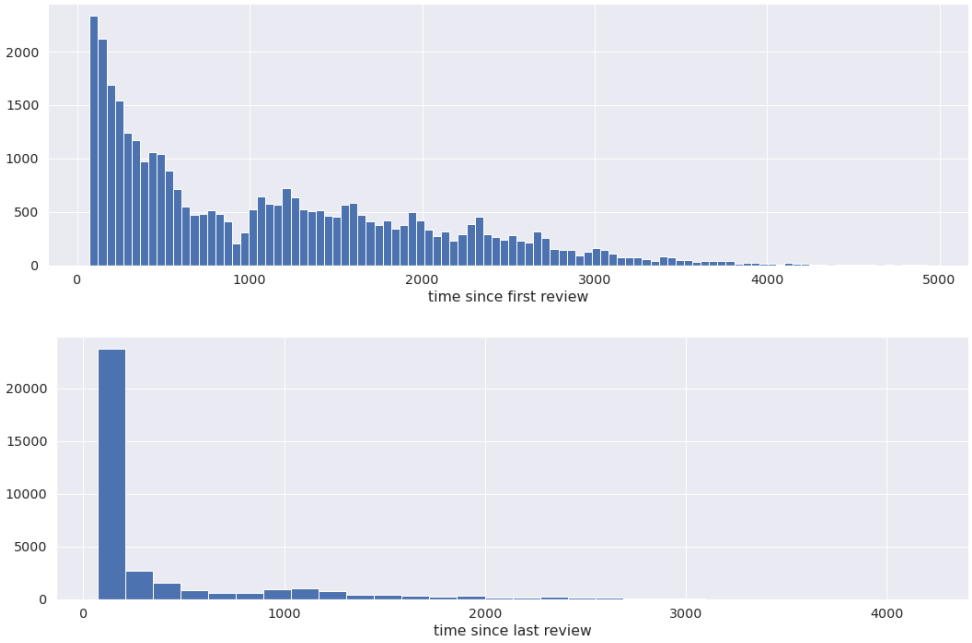

Figure 5: top: time since first review, bottom: time since last review

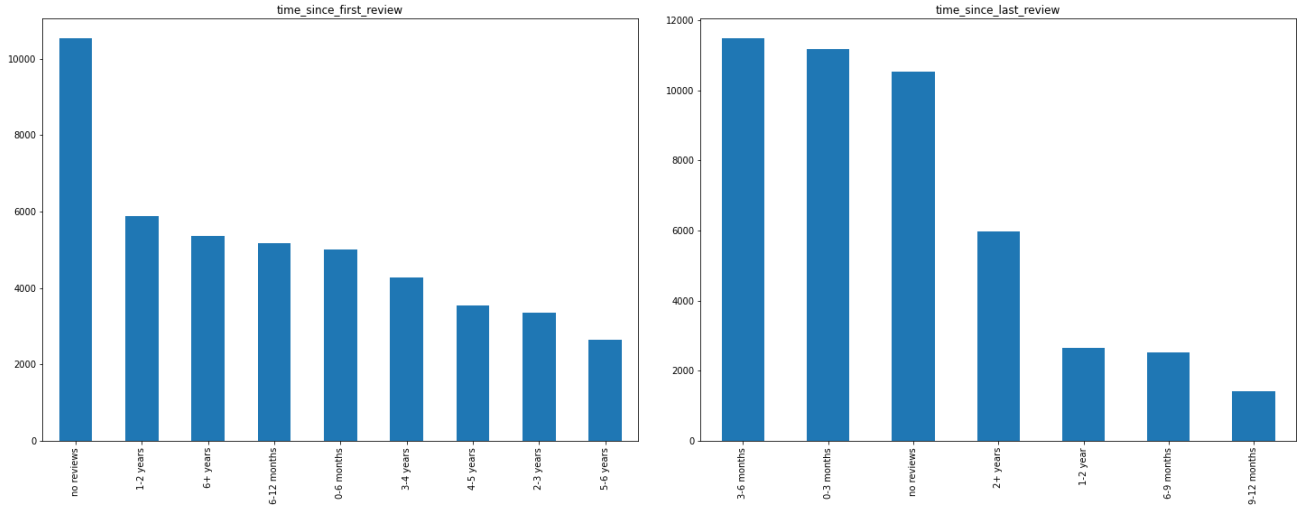

Figure 6: top: time since first review, bottom: time since last review

All reviews scores have more than 20 percent NAs. Because review scores are intuitively important in determining whether to book a specific listing, NA values have their own significance. They are transformed in the same way as 'host_response_rate' based on the histogram in figure 7. Transformed variables are shown in figure 8 and figure 9.

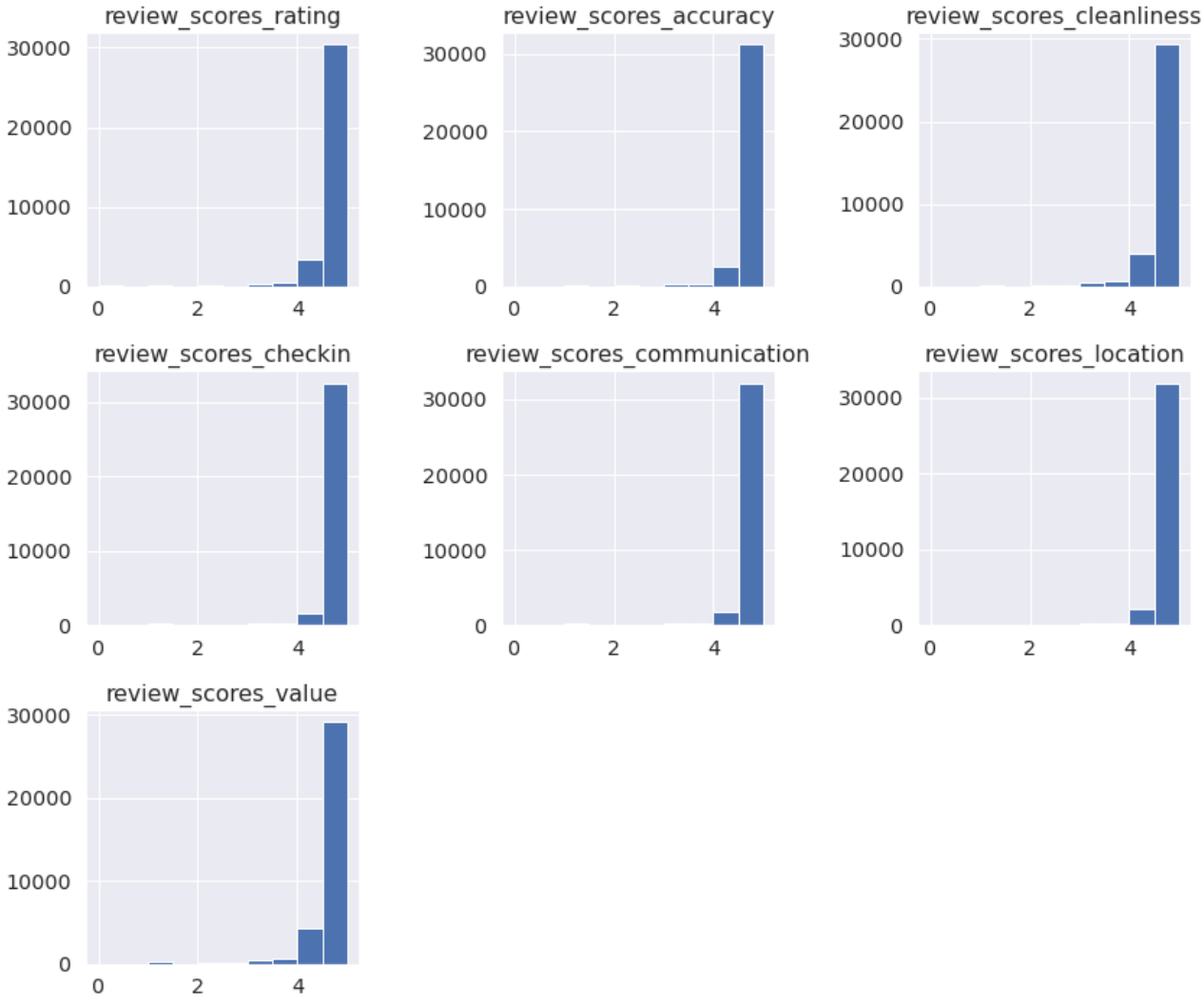

Figure 7: all review scores

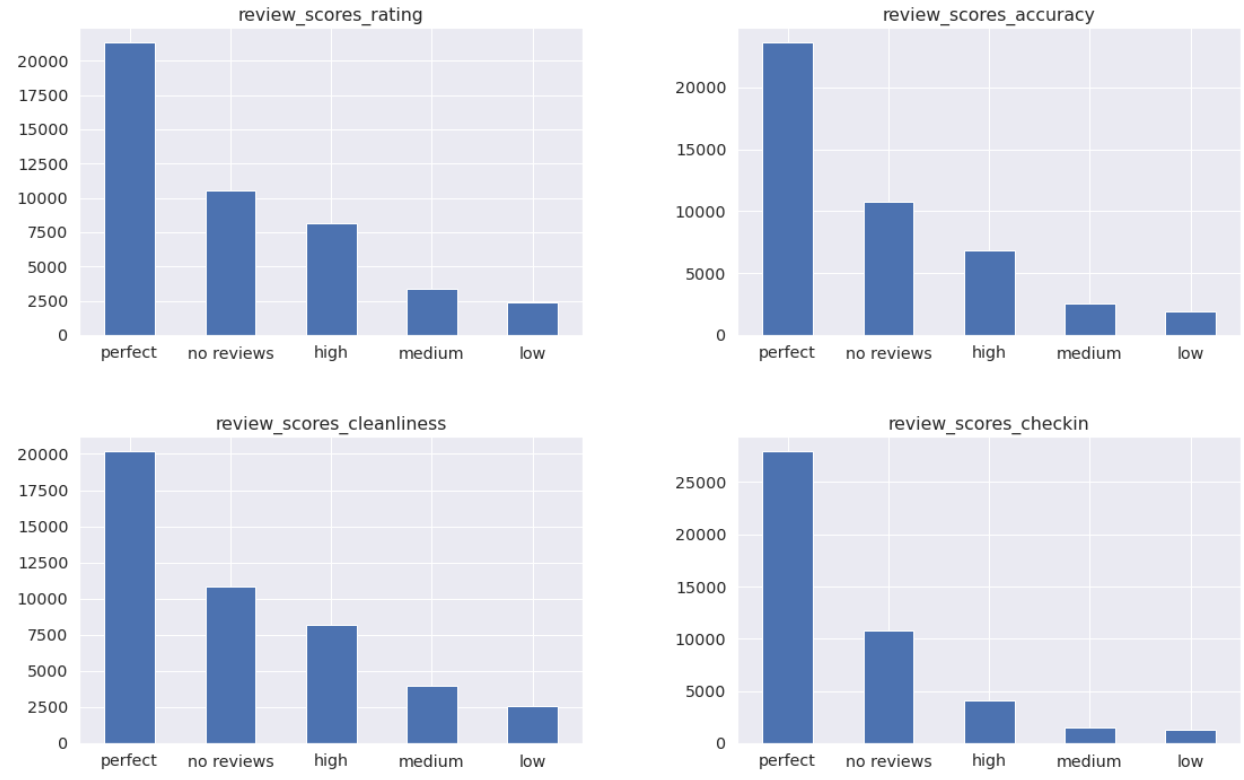

Figure 8: rating, accuracy, cleanliness, checkin

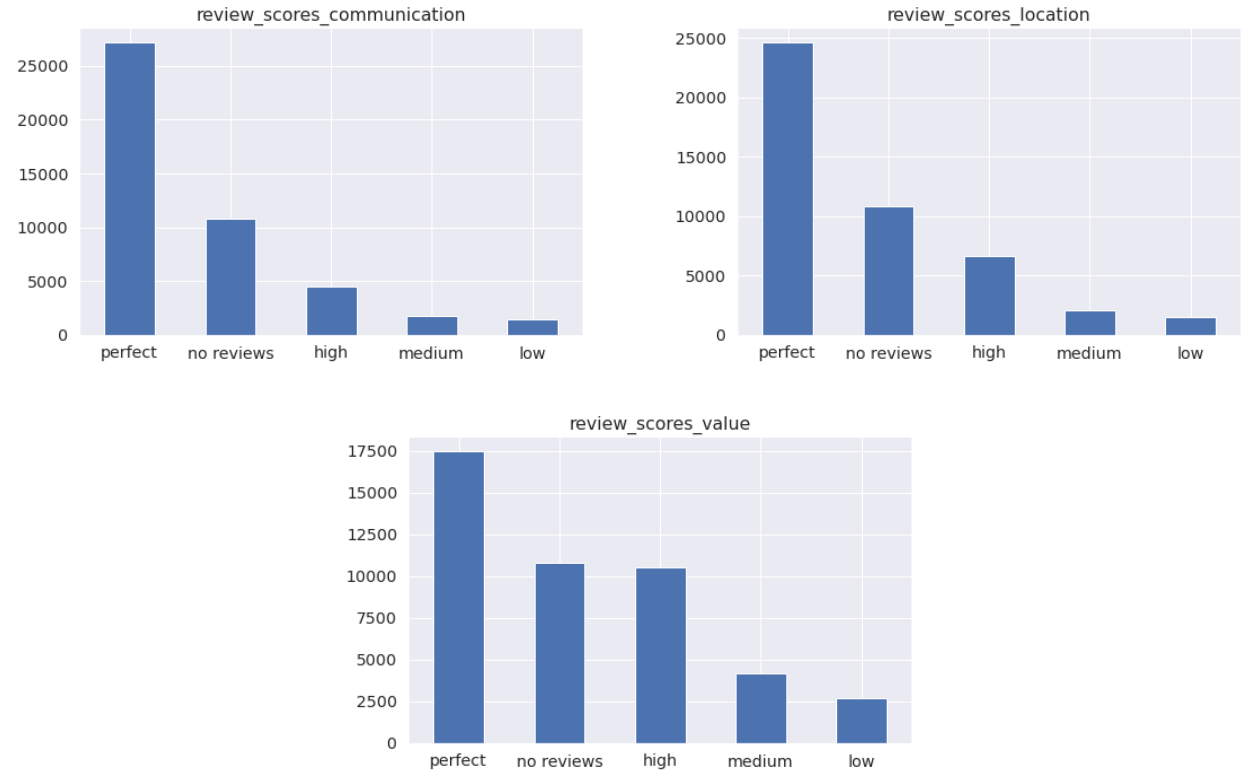

Figure 9: communication, location, value

'Property_type' has many categories, and those with less than 1000 elements are grouped together into 'other'. 'room_type' only has a few categories, so nothing is modified. 'above_average' is a new column indicating whether the price of a listing is above the average price of the neighbourhood it belongs to. By categorizing average price of listings in each neighbourhood into 5 levels based on histogram shown in figure 11, 'price level' indicates which level each listing belongs to.

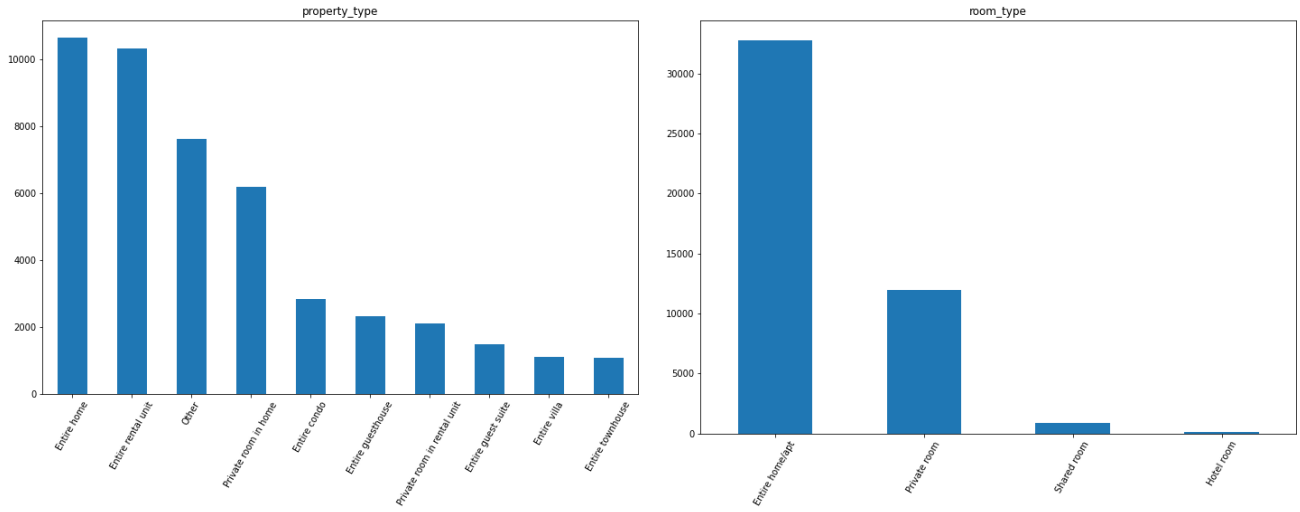

Figure 10: left:property type, right:room type

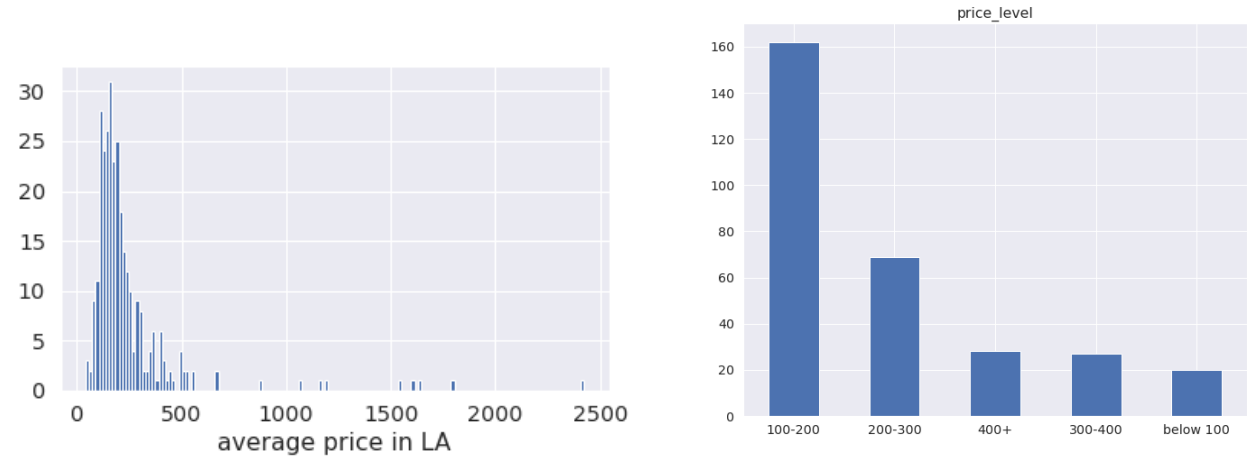

Figure 11: left:average price in LA, right:price level

4 Hypothesis Tests

I am interested in determining if seasonality affects the average price. Only listings appear in each quarter are selected to perform analysis. Average price of each quarter from sep 2021 to sep 2022 is denoted by μ_1, μ_2, μ_3, and μ_4 respectively. Use ANOVA to test:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1 : \mu_i \neq \mu_j$ for at least one pair (i, j)

The result is shown in figure 12, where p-value indicates I fail to reject H_0. However, price in different neighbourhoods varies a lot, so it is reasonable to block by neighbourhoods. P-value also indicates failure to reject H_0, shown in figure 13.

	sum_sq	df	F	PR(>F)
C(time)	2.652402e+06	3.0	1.526372	0.205349
Residual	5.519686e+10	95292.0	NaN	NaN

Figure 12: ANOVA result without blocking by neighbourhood

	sum_sq	df	F	PR(>F)
C(location)	6.239746e+09	263.0	46.052327	0.000000
C(time)	2.661817e+06	3.0	1.722254	0.159993
Residual	4.895711e+10	95029.0	NaN	NaN

Figure 13: ANOVA result with blocking by neighbourhood

5 Modeling

Final data is split into 70 percent training and 30 percent testing data. Standard scaler is performed on numeric variables, and one hot encoder is performed on categorical variables. Result are summarised in the following table. Feature importance of catboost model is also presented in figure 14

Model	RMSE	R2
Elasticnet	594.6	0.28
Random forest	419.7	0.64
Xgboost	398.1	0.68
Catboost	399.1	0.67

Figure 14: feature importance from catboost model

6 Future Work

Neighbourhood can be a really good indicator of listing price. In this project, straightforward categorization is performed for simplicity. In future work, one can incorporate more geographical information of each listing such as distance to major avenues, shopping center, grocery stores and so on in each neighborhood. Different combination of amenities may also affect a listing's attractiveness, so it should be analyzed more carefully. Using NLP to perform sentiment analysis on listing reviews and CNN to assess listing images is also a potential way to improve model performance.

7 Conclusion

With and without blocking by neighbourhood, seasonality has no effect on average price of listings. Xgboost model performs the best when predicting prices, giving r2 value 0.68 and RMSE 398.1. Although feature importance is not defined for booster 'dart' in xgboost, the second best model, catboost, still gives a relative good picture of what factors affect the price the most. In future investigation, geographical information should be exploited more thoroughly. NLP and CNN are additional tools to predict listing price.