Los Angeles

Airbnb Rentals Data Analysis

- Madhura Dighe, Shohini Ghosh, Surbhi Gupta, Manas Mishra, Pratik Pandey, Yash Srivastava (Team 9)

AGENDA

01 Kaggle Competition

02 Define Business Cases

03 Data Cleaning and Exploration

04 Explanatory Analysis

05 Predictive Analysis

KAGGLE Competition

1. Data Processing

- Cleaning of data.
- Conversion of variables.
- Handling of missing values (filled with zeroes, FALSE, median date, value, mean value).
- Mutation of columns.
- Dropping missing data or single-valued columns.
- Conversion of characters to factor.
- Processing and expansion of ‘amenities’ and ‘host_verifications’ columns.
- Removal of character columns containing NA values.

2. Final method used along with its parameters

Extreme Gradient Boosting:

xgboost(xtrain, label = ytrain, nrounds = 600, params = (eta = 0.07, colsample_bylevel=0.7],
subsample = 0.7, max_depth = 7)))

3. Variables included in final model

‘accommodates', 'availability_30', 'bathrooms', 'bedrooms', ‘'cancellation_policy', 'cleaning_fee', 'extra_people', 'guests_included', 'host_identity_verified', 'host_is_superhost', 'host_listings_count', 'host_response_rate', 'host_response_time', 'instant_bookable', 'is_location_exact', 'maximum_nights', 'minimum_nights', 'price', 'property_type', 'require_guest_phone_verification', 'require_guest_profile_picture', 'requires_license', 'review_scores_accuracy', 'review_scores_checkin', 'review_scores_cleanliness', 'review_scores_communication', 'review_scores_location', 'review_scores_rating', 'review_scores_value', 'room_type', 'host_days_active', 'email', 'phone', 'reviews', 'weibo', 'offline_government_id', 'google', 'facebook', 'selfie', 'work_email', 'balcony', 'bed_linen', 'breakfast', 'tv', 'coffee_machine', 'cooking_basics', 'white_goods', 'elevator', 'child_friendly', 'parking', 'outdoor_space', 'host_greeting', 'internet', 'long_term_stays', 'pets_allowed', 'private_entrance', 'secure', 'self_check_in'

Note: ‘Amenities’ and ‘host_verification’ columns have been exploded into subcategories for better analysis.

Performance (AUC) achieved: 94.018

Los Angeles Airbnb Market

24.7k

Total Properties

$1458

Monthly Income

1.3 million

Reviews Provided

Business Cases

New Acquisitions

Focus on new purchase using market characteristics such as events, attractions, locations etc. to increase bookings.

CASE 1

Pricing and Management

Study the impact of price/fees, professional management to increase bookings.

CASE 2

Upgrade and Renovations

Suggest improvements in existing rentals to increase bookings.

CASE 3

Data Cleaning and Exploration

To model the spatial relationship between Airbnb rental booking rate to various factors such as proximity to certain venues, attractions etc, data cleaning was performed to overcome limitations present in data.

- Columns with 70% data missing were eliminated such as is_business_ready.
- String cleaning for city and neighborhood columns.
- Breakdown of columns such as amenities, host verification to granular level.
- Null values were handled by substituting with mean/median/NA values.
- Some character columns were converted to number/factors.

Some of the important features this project considers are:

- Bedrooms, Bathrooms, Neighborhood, Price, Property Type, Host Properties

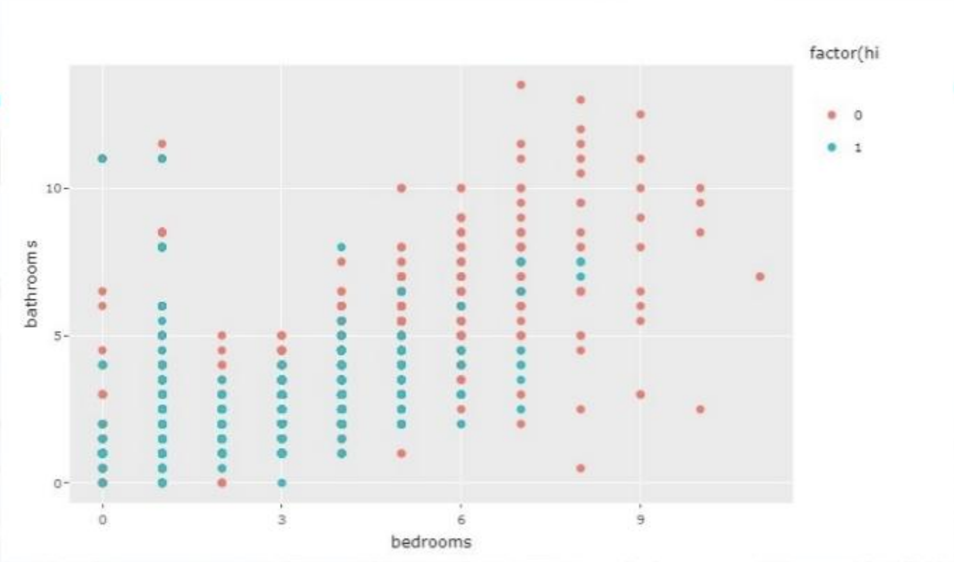
Explanatory Analysis

Analysis 1

Bedroom versus Bathroom Stats for high booking properties

Observations

- Properties with Bedroom & Bathroom count within range of 0-5 have higher bookings.
- As the count of both increases further chances of bookings drop.



Investment Suggestions

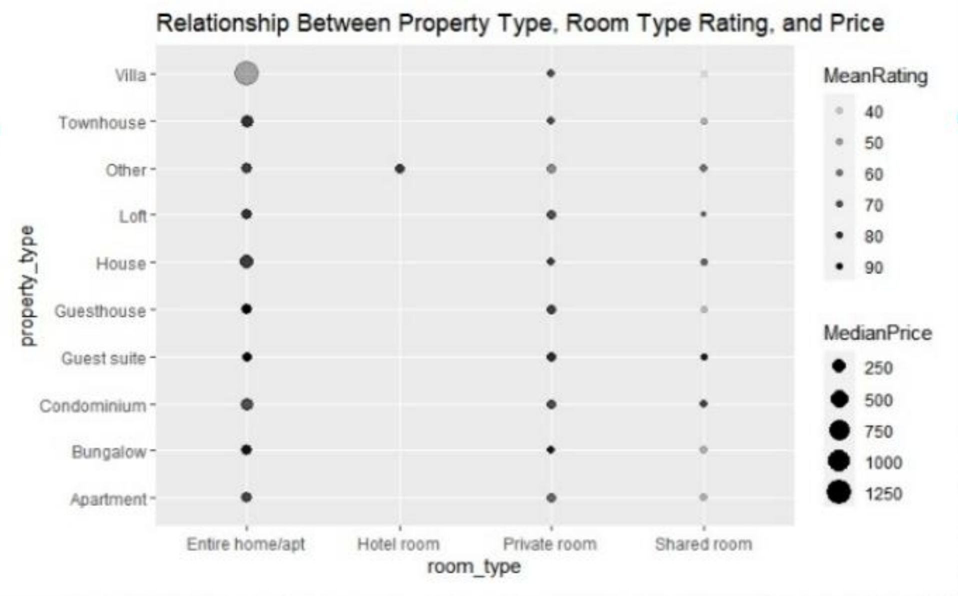
- Invest in rentals with with limited number of Bedrooms (not more than 5).
- Private bedroom properties seem to perform well with variable bathrooms, hence this can be safest option to invest.

Analysis 2

Property Type versus Room Type w.r.t. Price & ratings

Observations

- In entire home & private room category, townhouse, guest house, guest suit & bungalow are cheaper with good ratings.
- Villas being costlier are rated very low.
- Hotel rooms are also highly rated with low price & shared rooms perform poor w.r.t all.



Investment Suggestions

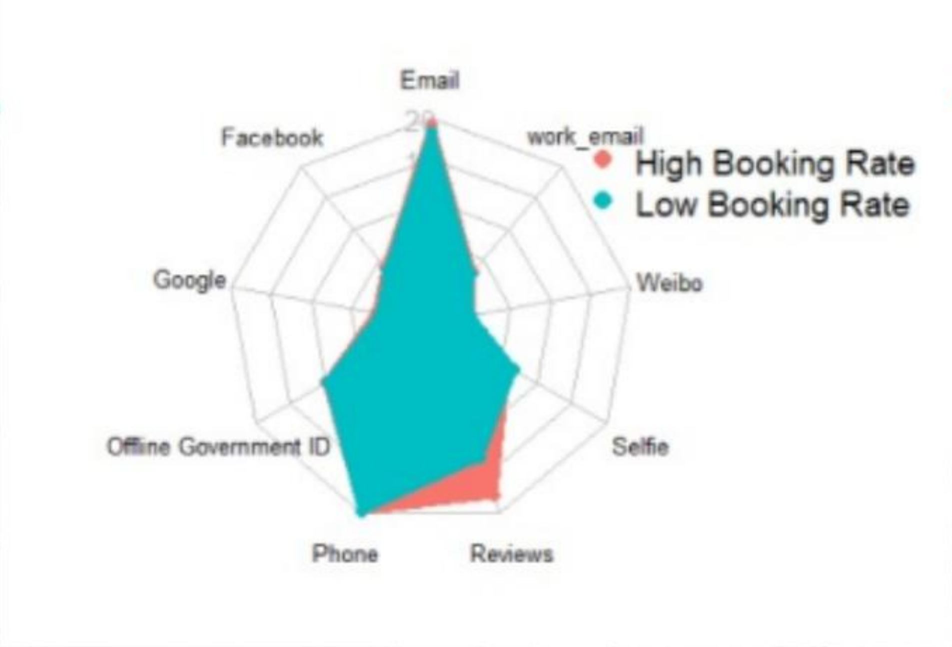
- Invest in cheap properties with higher ratings & get good business outcome at lower risk.
- Avoid shared rooms & properties mentioned above.
- Explore hotel rooms in other category as a investment option.

Analysis 3

Radar Plot : Host Verification versus Booking Rate

Observations

- Hosts for low booking and high booking properties are verified by all social media platforms.
- But reviews seems to be the distinguishing feature for higher bookings.



Investment Suggestions

- Try to get feedbacks & reviews for the host which is going to be investor.
- Highly reviewed hosts can attract more bookings thus beneficial for the business.

Analysis 4

Top 10 percentile of Neighbourhood versus Avg Price w.r.t. Booking Rate

Observations

- Avg price for Bel Air Beverly Crest, Beverly Hills is very high with no booking data.
- Places such as Downtown, Echo Park, Hollywood have lower average price with some bookings.

Investment Suggestions

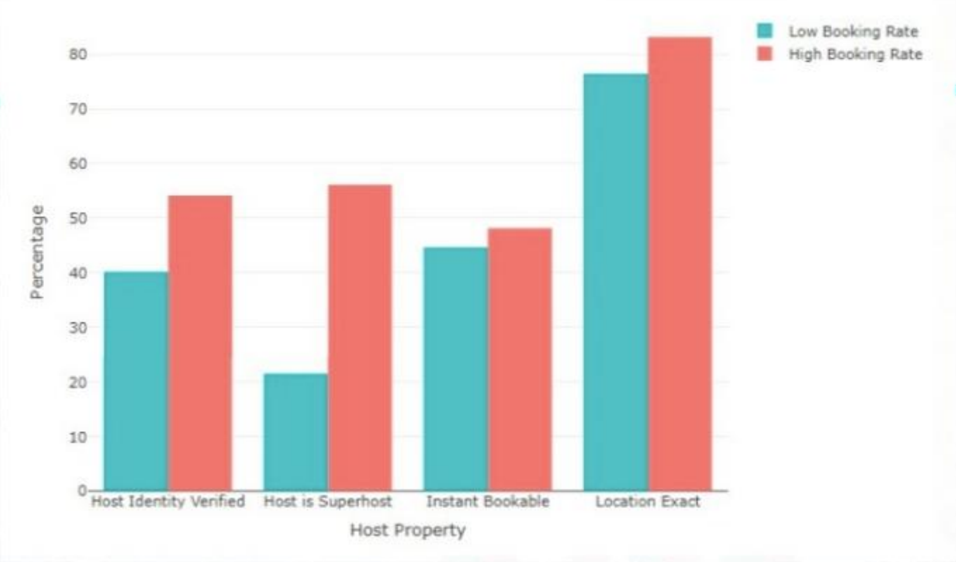
- Invest in properties such as Mar Vista, Silver Lake, Westchester as they have 100% booking rate.
- Avoid purchasing rentals in Westwood, BelAir, Beverly Hills.

Analysis 5

Host Property w.r.t. Booking Rate

Observations

- Properties where host is a superhost is the most statistically significant in terms of high booking rate.
- Overall, all the host properties have high booking rate.

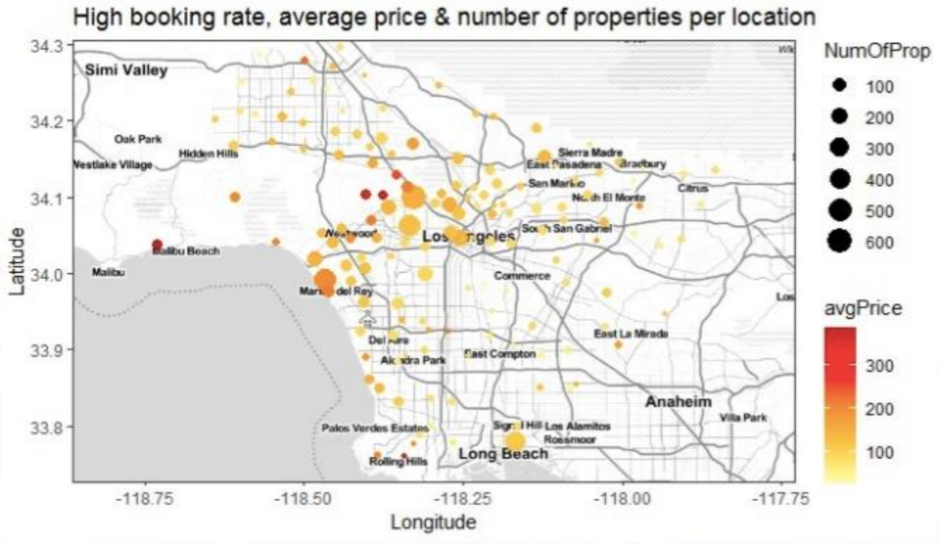


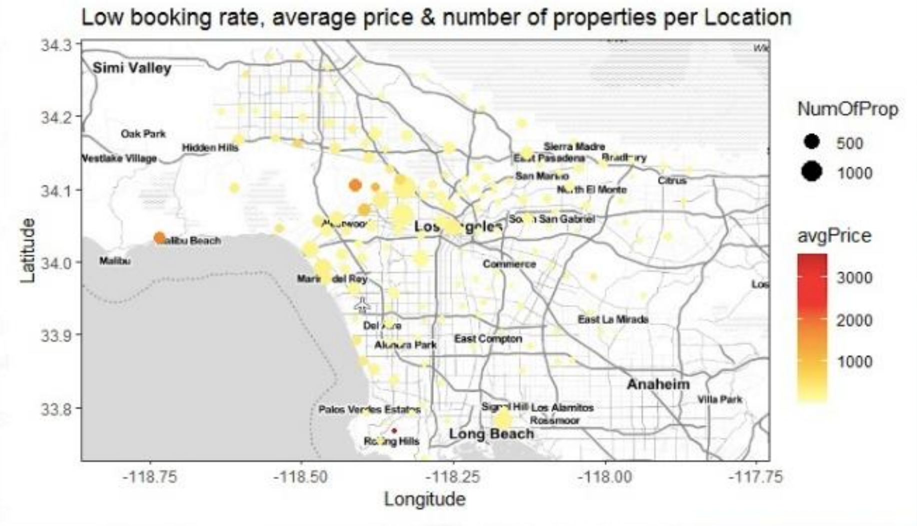
Investment Suggestions

- Invest in properties where the host is a superhost for higher booking rate.

Analysis 6

Neighborhood





While all neighbourhoods perform well on average, when we group data by neighbourhood, we see different picture for price and population of properties for high versus low booking rentals.

Analysis 6

Neighborhood: Observations and Investment Suggestions

Price range for properties with high bookings is comparatively lower than the low booking properties for similar locations.

Neighbourhoods with higher bookings for investment:

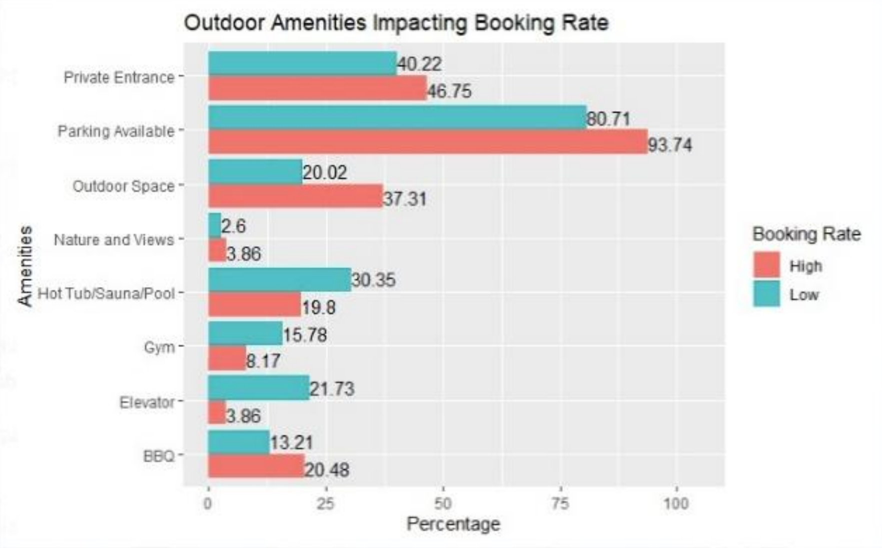
- Marina Del Rey (Nearby Marina beach)
- Hollywood (Nearby Hollywood walk of fame)
- Long Beach
- China Town
- Mid City
- Central LA
- South Los Angeles
- Westwood

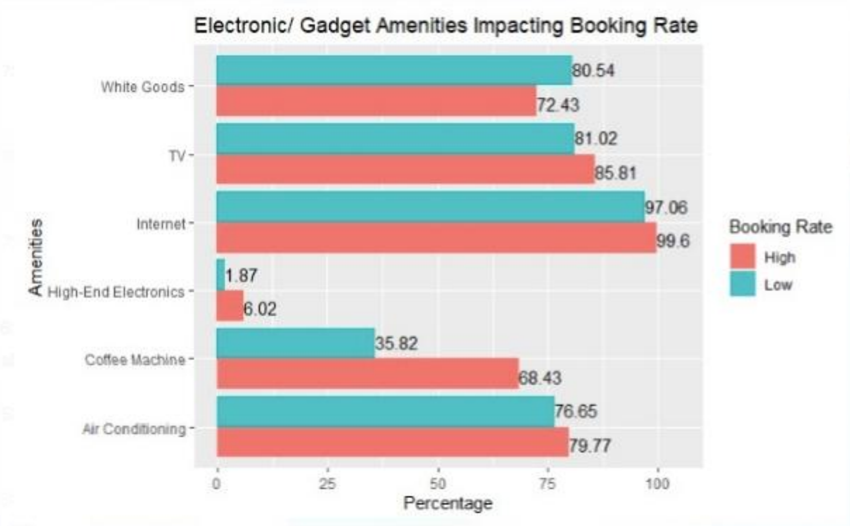
Investors should buy cheaper properties.

Customers are looking for cheap rentals.

Analysis 7

Amenities w.r.t. Booking Rate





Observations & Suggestions

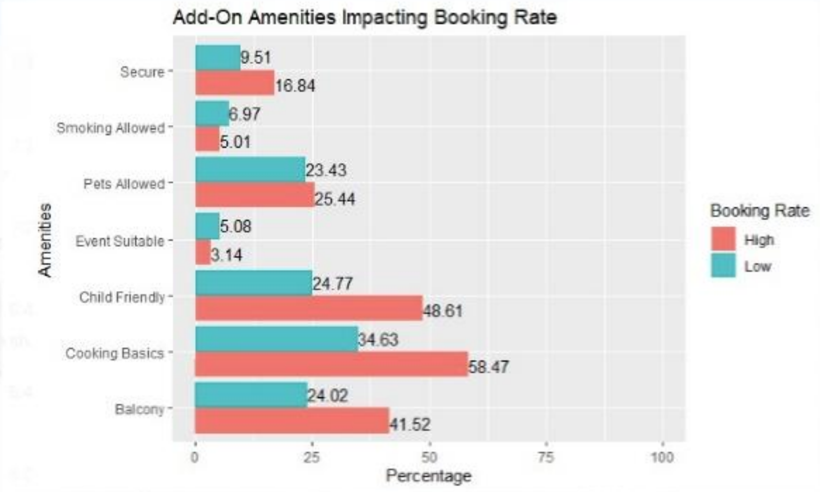
- Outdoor amenities don't significantly impact the pricing decision.

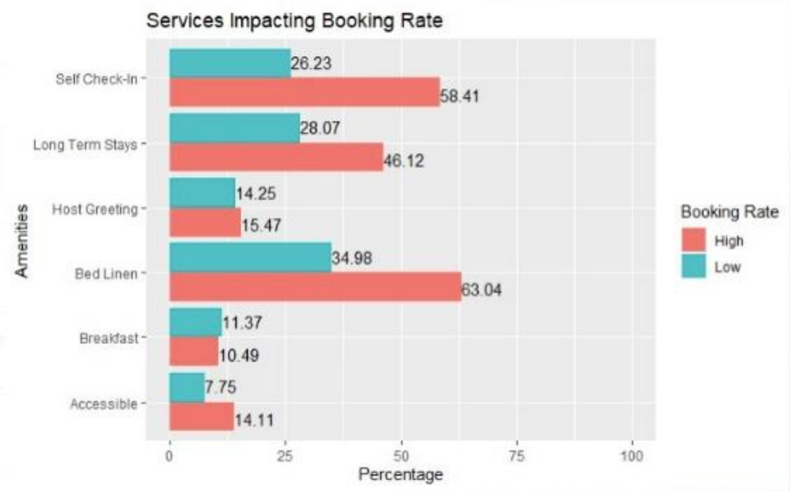
Observations & Suggestions

- Electronics and gadgets are basic amenities and they don't contribute heavily to the booking rates going up; although, coffee machine does drive up the costs.

Analysis 7

Amenities w.r.t. Booking Rate





Observations & Suggestions

- Add-on amenities like the property being child friendly and having cooking basics help increase the booking rate.

Observations & Suggestions

- Self Check-In is a major service that contributes to high booking rates.

Predictive Analysis

1. Data Processing

- Filtering out 'Los Angeles' data from the full data set.
- Cleaning of data.
- Conversion of variables.
- Handling of missing values (filled with zeroes, FALSE, median date, value, mean value).
- Mutation of columns.
- Dropping missing data or single-valued columns.
- Conversion of characters to factors.
- Handling 'city' column (managed case-sensitivity, string processing to avoid redundancy)
- Processing of 'property_type', 'host_verification' and 'amenities' column.
- Removal of character columns containing NA values.

2. Proposed final method used along with its parameters

fitLPM1 <-train (factor(high_booking_rate)~ review_scores_rating + cleaning_fee + host_is_superhost + check_in_24h + latitude+longitude + child_friendly+ internet + price + high_end_electronics + white_goods + self_check_in + room_type + bed_type + cancellation_policy + reviews + instant_bookable + secure + property_type + bedrooms+bathrooms + host_listings_count + availability_365, family = 'binomial', method= 'glm',data = dfTrain1, na.action = na.exclude)

3. Variables included in proposed final model

'review_scores_rating', ' cleaning_fee', 'host_is_superhost', 'check_in_24h', 'latitude', 'longitude', 'child_friendly', 'internet', 'price', 'high_end_electronics', 'white_goods', 'self_check_in', 'room_type', 'bed_type', 'cancellation_policy', 'reviews', 'instant_bookable', 'secure', 'property_type', 'bedrooms', 'bathrooms', 'host_listings_count', 'availability_365'

4. Parameters to test efficiency of proposed model

Accuracy: 0.7988
Sensitivity : 0.3980
Specificity : 0.9416
RMSE: 0.9854261 (0.6 cutoff)
MAE: 0.8849509 (0.6 cutoff)

Performance (AUC) achieved: 85.59

CONCLUSION

- Not all property types (rentals) would help investors in achieving high booking rates.
- Cheaper rentals have better bookings stats than costliers ones.
- Investors should target properties around prime locations.
- Upgradations should be done based on statistically significant features.

THANK YOU

Any Questions?