Data Mining and Predictive Analytics

Project Title: Airbnb LA Business Case Analysis

Market Assigned: Los Angeles

"We, the undersigned, certify that the report submitted is our original work; all authors participated in the work in a substantive way; all authors have seen and approved the report as submitted; the text, images, illustrations, and other items included in the manuscript do not carry any infringement/plagiarism issue upon any existing copyrighted materials."

Member ID	Member Name
Contact Member	Yash Srivastava
Team Member 2	Madhura Dighe
Team Member 3	Shohini Ghosh
Team Member 4	Surbhi Gupta
Team Member 5	Manas Mishra
Team Member 6	Pratik Pandey

Executive Summary

This report provides an analysis and evaluation of the current Los Angeles Airbnb market to help businesses make strategic decision regarding acquisition, update or renovation of Airbnb rentals. We have performed explanatory analysis to understand the impact of a market factor on the booking rate and used those significant factors to develop a predictive model that most accurately predicts the booking rate.

By exploring the Los Angeles Airbnb data, we were able to extract interesting and useful insights. For example, for risk neutral decisions, businesses should focus on locations such as Venice Beach and Marina Bay to maximize booking rate and not incur loss as these areas are major water recreation destination.

Research Questions

To assist the business make risk neutral strategic decision, we focused on answering the following questions:

Location, location, location:

1)In which location should the business invest in to get high booking rate?

Assessing the competition:

2) What are the most common property and room type in order to maximize booking rate?

Is it worth it:

3) What is the impact of factors such as price, host property type (superhost) on the booking rate?

Amenities the guests love:

4)What upgrade and renovations can lead to an increase in the booking rate?

Guests heart Airbnb:

5) How are the reviews affecting the booking rate?

Since Los Angeles already has a booming market for Airbnb, to have an edge in the competition, it is crucial to know from where to generate positive cash flow per month, which segment of market to target. We believe by answering the above research questions, businesses will be equipped to make informative decision.

Methodology

Load and Clean Data

The Los Angeles data set has a lot of entries and is messy. Our project is data cleaning intensive and we have manipulated the data in ways we found could help us with our analysis.

The larger model for the Kaggle competition gave us an idea about important features overall for Airbnb market and high booking rates. Based on our analysis for the larger model, we converted a few columns to numeric types as numeric entries made more sense. We handled missing values in a variety of ways, filling them with median or mean values in some cases and with simple zeroes in others. Mean imputations were done carefully keeping in mind the bias that could generate. We factorized a few variables to fetch their levels. We mutated a few columns, dropped columns with single values and NA values. We processed the 'amenities' and ' host_verification' columns to gain more insight from our data.

Once this was done, we analyzed our dependent variable, 'high_booking_rate', against a number of different features to determine the significance of these features on the booking rate.

Mutate Columns

```
trainLA <- trainLA %>% mutate(host_days_active = Sys.Date() - host_since)
testLA <- testLA %>% mutate(host_days_active = Sys.Date() - host_since)

trainLA$host_days_active <- as.numeric(trainLA$host_days_active)
testLA$host_days_active <- as.numeric(testLA$host_days_active)
```

Drop Missing-Data/Single-Valued Columns

```
drop <- c("monthly_price", "square_feet", "randomControl", "weekly_price",
"host_acceptance_rate", "state", "is_business_travel_ready",
"host_has_profile_pic", "amenities", "host_verifications")

trainLA <- trainLA[ , !(names(trainLA) %in% drop)]
testLA <- testLA[ , !(names(testLA) %in% drop)]
```

Process Amenities and Host Verifications

```
trainLAamenities <- read_csv("airbnbTrainLAamenities.csv")
testLAamenities <- read_csv("airbnbTestLAamenities.csv")

trainLA <- dplyr::inner_join(trainLA, trainLAamenities)
testLA <- dplyr::inner_join(testLA, testLAamenities)

trainLAhost_ver <- read_csv("airbnbTrainLAhost_verifications.csv")
testLAhost_ver <- read_csv("airbnbTestLAhost_verifications.csv")

trainLA <- dplyr::inner_join(trainLA, trainLAhost_ver, by = 'id')
testLA <- dplyr::inner_join(testLA, testLAhost_ver, by = 'id')
```

Remove Columns containing NA values

```
trainLAna <- trainLA[ , colSums(is.na(trainLA)) == 0]
testLAna <- testLA[ , colSums(is.na(testLA)) == 0]
```

Predictive Analysis

From our analysis, we short-listed a few variables that we thought were best suited to make predictions from. We realized that reviews, some important amenities like check in, friendliness in terms of the customer, rooms, the type of property, the reputation of the host, geographical importance and some other services like electronics and security impacted the booking rate positively.

The variables are listed as follows:

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

Based on the variables mentioned above, we ran models to test the performance metrics and found out that 'Gradient Boosting' performed the best in predicting booking rates for properties in LA. This was decided based on the metrics generated at multiple cutoff levels

giving us the best sensitivity, specificity and accuracy in addition to a high 'Area-under-the-curve (AUC)'.

```
modelBoost <- gbm(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, data = dfTrain1, distribution = "bernoulli", n.trees = 500,
shrinkage = 0.08, interaction.depth = 7, cv.folds = 5, n.cores = NULL,
verbose = FALSE)
```

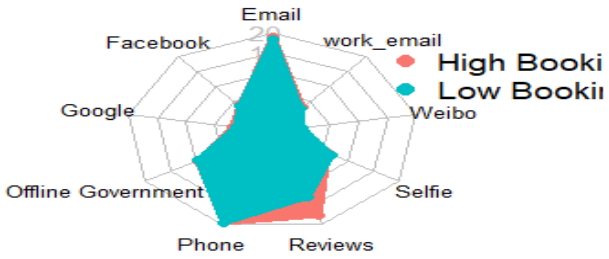
Results and Findings

Booking Rate Data Split

```
trainLAHBR <- trainLAna %>% filter(high_booking_rate == 1)
trainLALBR <- dplyr::setdiff(trainLAna, trainLAHBR)
```

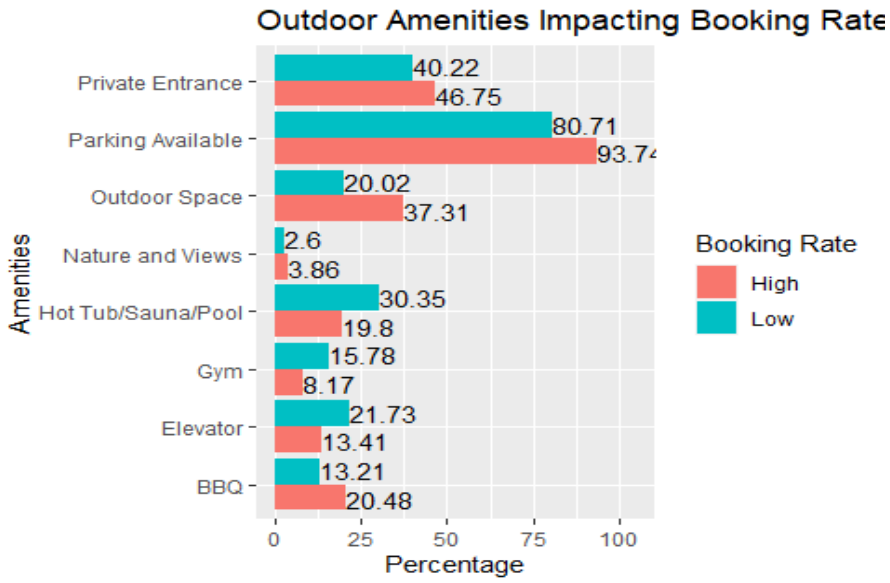
Host Verification Methods Analysis

In the radar chart below, we compare the different host verification methods for the high and low booking rate properties. We see that for most of the verification methods including phone and email, there isn't much difference between the two types of booking rates. But, in case of ***reviews*** as a host verification method, we observe that properties having higher booking rate tend to have been verified via reviews compared to low booked properties.

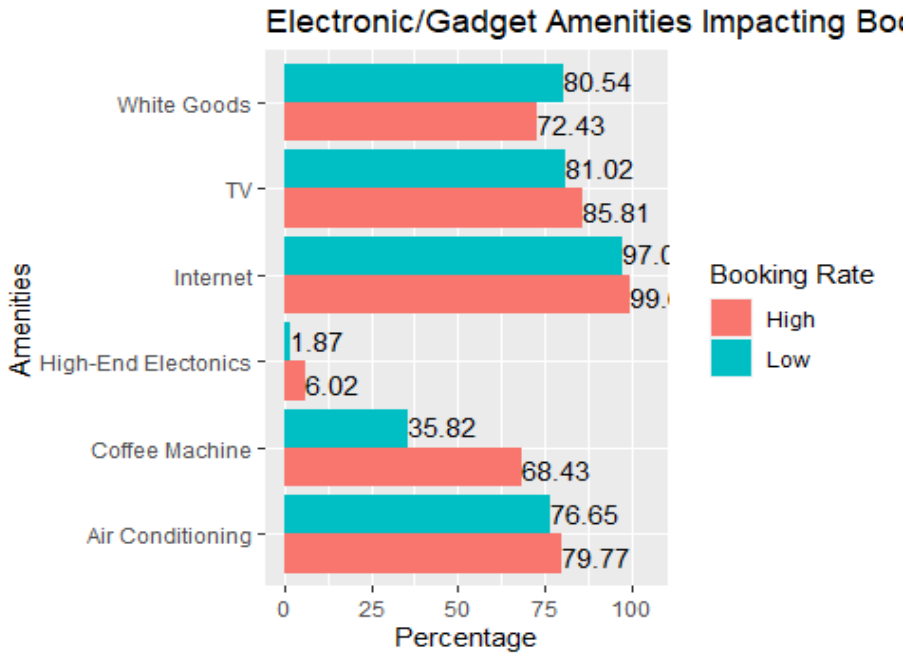


Amenities Analysis

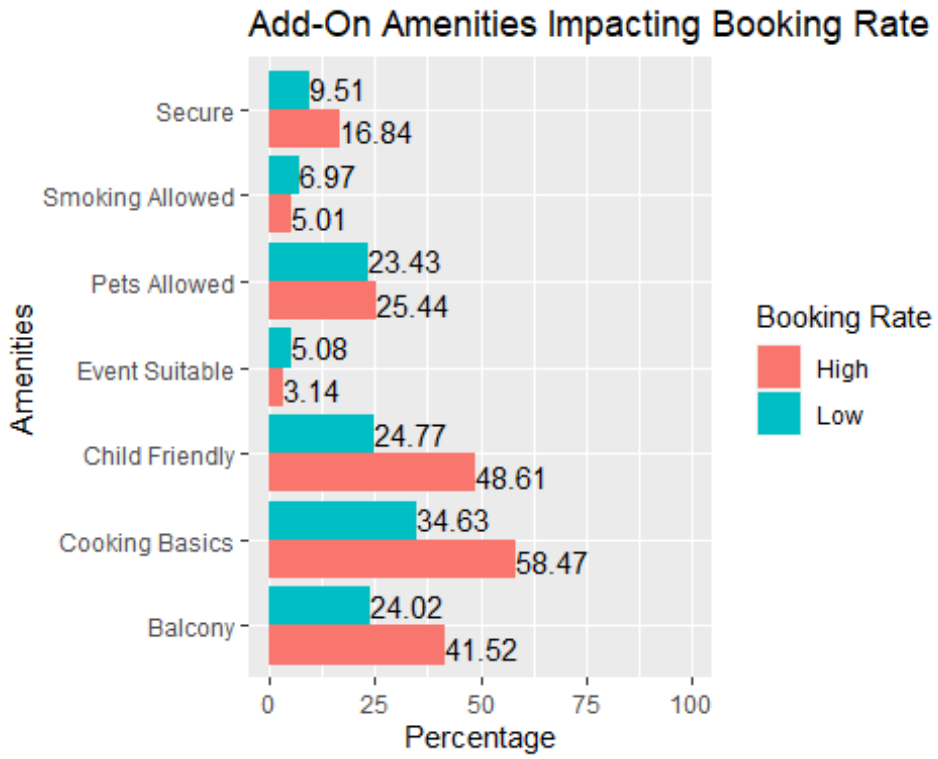
We can observe from the graph below that providing outdoor amenities has a positive impact on the booking rate but is not very significant. There are a few exceptions like gym, hot tub/sauna/pool, and elevator, which have a negative effect on booking rate. This might be due to the fact that LA is a coastal area and tourists prefer the beaches over indoor pools, a house over an apartment (eliminating the need for elevators) and relaxing over working out.



Electronics and gadgets like TV, internet and air conditioner are basic amenities in this modern era and thus are expected from the owner. Also, the owners keep up with these demands and thus these amenities do not contribute towards high booking rate. High-end electronics such as home entertainment systems are generally very expensive. Usually, customers are not looking for these amenities and they do not play a major role in helping them make booking decisions. Although, we observed that the customers really get upset if a coffee machine is not provided.

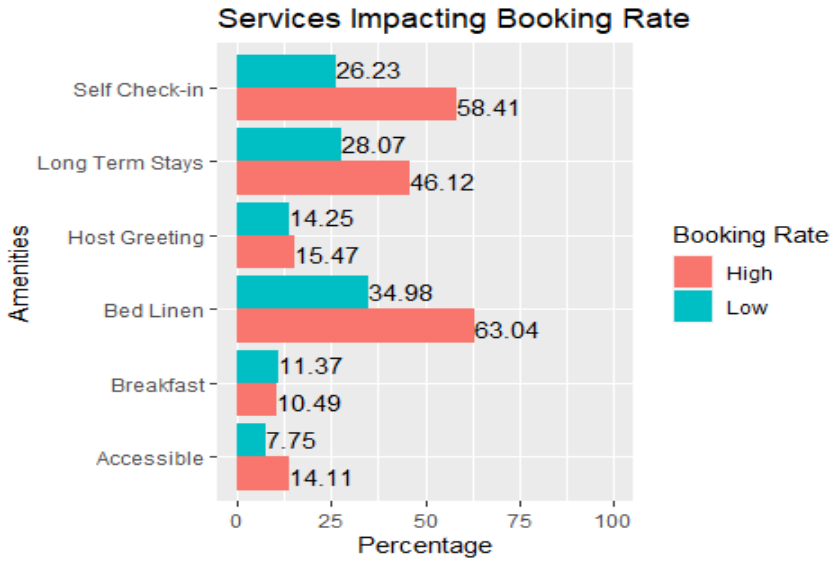


It is evident from the graph below that listings which are child friendly, have a balcony or provide cooking basics drive up the booking rates (almost double the rate of booking). This means investors should focus on investing in properties with these add-on amenities.



The graph below depicts the importance of services when it comes to making booking decisions.

As can be seen, providing the option of check-in to the customers boosts the chances of the property getting booked significantly as customers prefer a hassle-free experience. Also, Los Angeles being a major tourist attraction, people like spending time here. Hence, having long-term stays as a service widens the potential customer range. In all, each service increases the chances of achieving a high booking rate, the exception being a breakfast service which is not a major pull for customers.



Host Properties Percentage Ratio

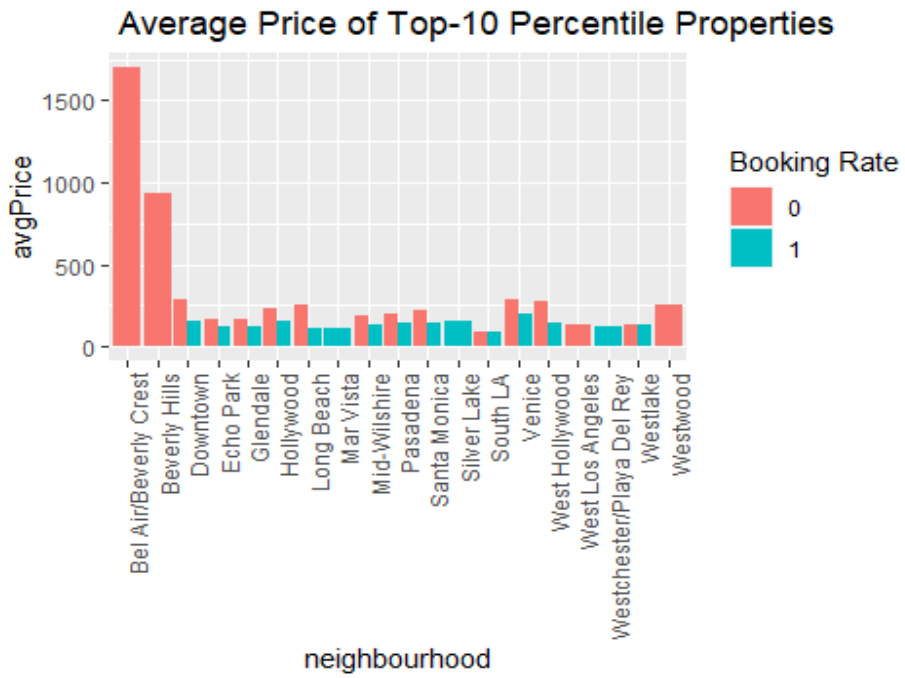
Some of the host properties that were analyzed in the below graph are:

- Host Identity Verified
- Host Is Superhost
- Instant Bookable
- Location Exact

Observing these properties for both high and low booking rate properties, we see a significant difference when the host is a superhost. Properties having the superhost tag have higher booking rates when compared to other properties. Whereas, the other host properties have a marginal edge in the case of high booking rate properties.

Pricing by Neighbourhood (High/Low Booking Rate)

This graph plots the top 10 percentile of the most populated neighbourhoods vs average price of the property with respect to booking rate. We observe that properites in Bell Air and Beverly Hills have a very high price and because of this, they might not have high booking rates. Whereas properties in Downtown, Echo Park, Venice and others that have considerably lower average prices tend to have high booking rates.

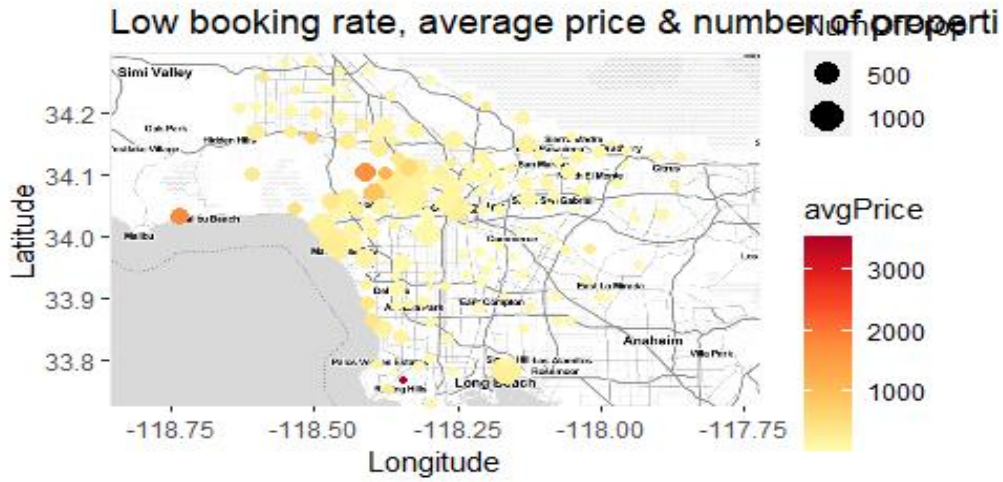


Map Visualization of High and Low Booking Rate Properties

The first map plot shows the distribution of properties in LA and their average prices with respect to high booking rates. We observe that the number of properties in and around Long-beach, Venice beach and Mid- Wildshire are significantly higher than other places. Also the average prices of these properties lie somewhere in the midrange. Properties in Malibu beach and Rolling Hills are few and average prices are pretty high.

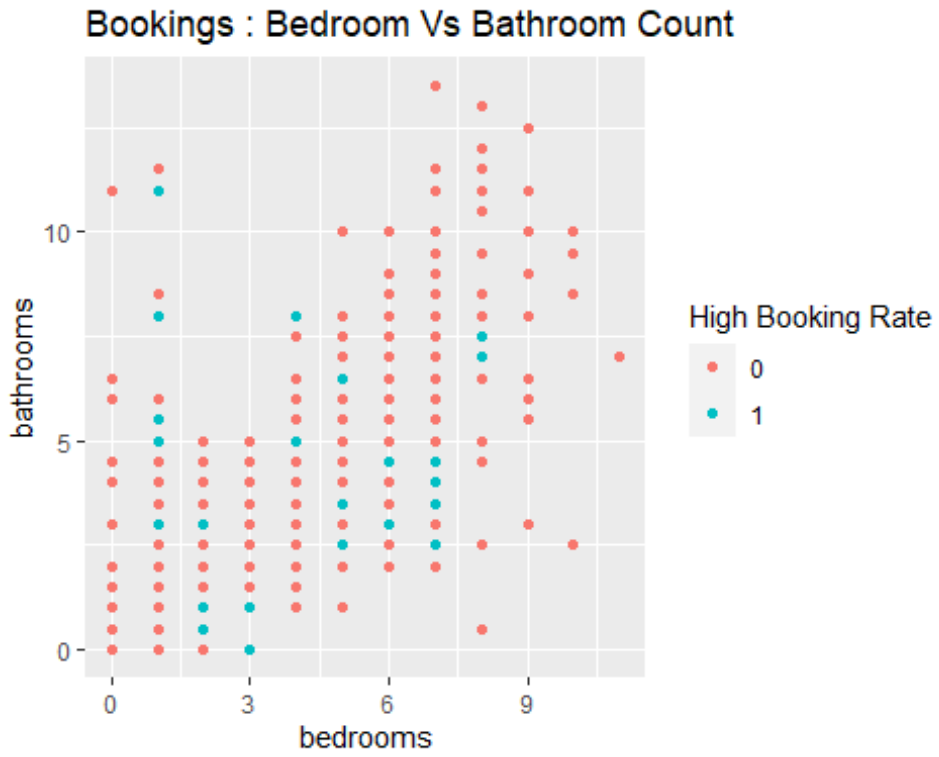
The second map plot shows the distribution of properties in LA and thier average prices with respect to low booking rates. In this map plot, we observe a similar trend as the first map, but average price range is this case is considerably higher than the properties with high booking rate. Thus, we can conclude that, properties with lower average prices tend to have a high booking rate.

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



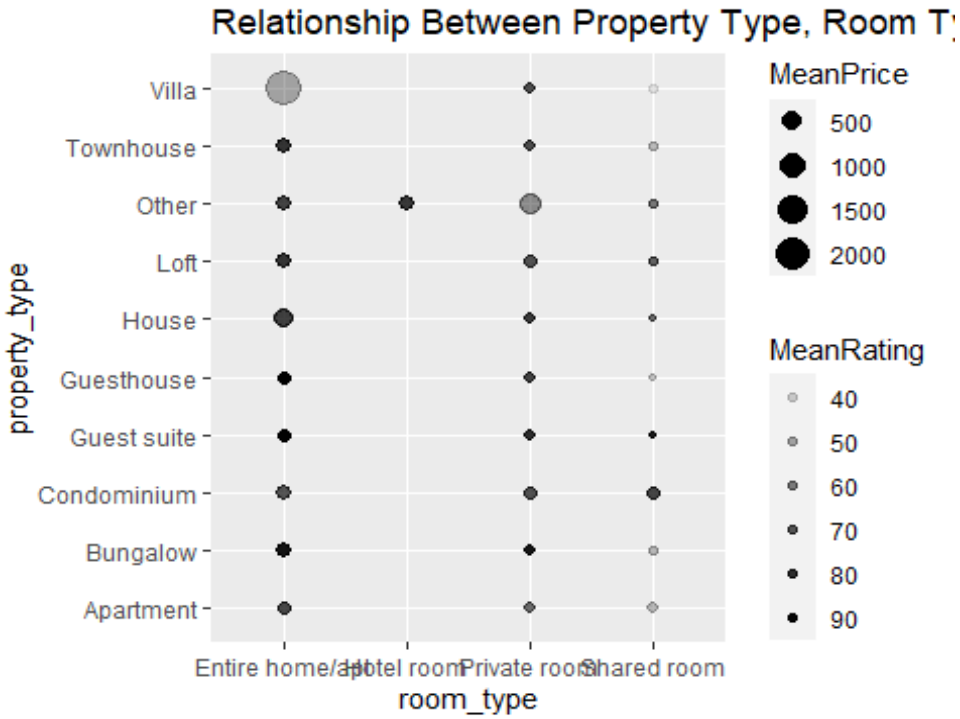
Bedroom and Bathroom Requirement Analysis on High/Low Booking Rate

The stats for bedroom vs bathroom do not follow a specific trend. From the plot, we can get an idea of combinations of bed vs bath arrangement for higher booking rate properties. We can definitely say that a higher number of bedrooms and bathrooms won't contribute to an increase in booking rate. We will check its actual significance later in the modeling.



Property and Room Type Analysis

This graph shows us which property type and room type can be better for investment by considering overall ratings vs price. For the entire room and private room category, guest house, guest suite, and bungalows are cheaper with good ratings. Villas are not only costlier but also rated low. Although hotel rooms are rated very highly, those are cheaper as compared to shared rooms. Thus, overall we can say that cheaper properties are highly rated. Hence, it's wise to invest in them. It's like an investment with lower risk and higher chances of benefits.

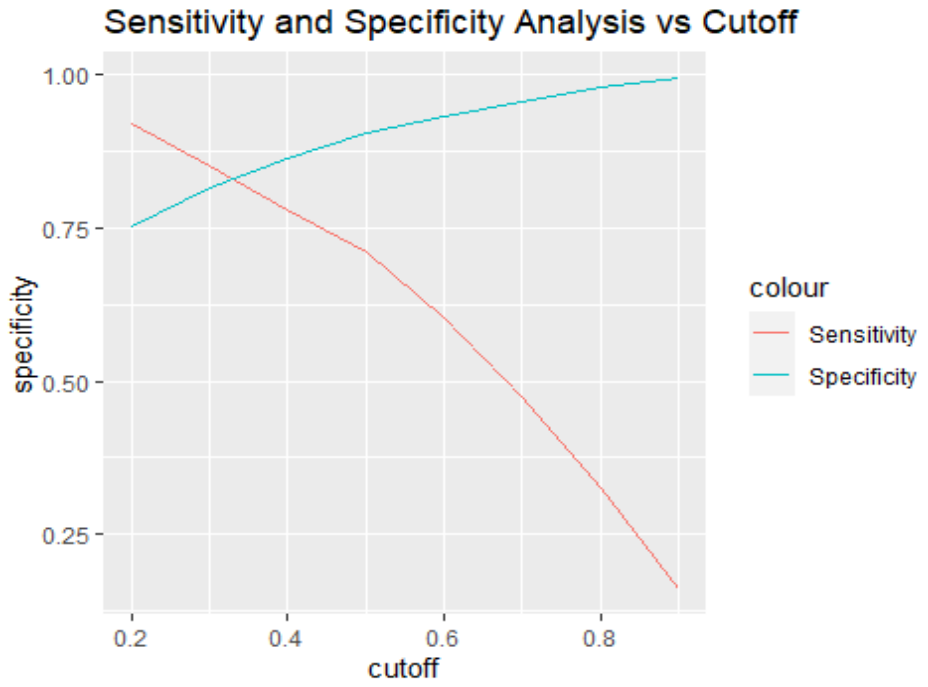


Cut-off, Specificity and Sensititvity Analysis for the Gradient Boost Model

- Accuracy tells us how correctly our model classifies the booking rate of a property in LA.
- Sensitivity tells us, out of all the properties classfied as high booking rates, how many does our model predict the same correctly.
- Specificity tells us, out of all the properties classfied as low booking rates, how many does our model predict the same correctly.
- Our goal here is to find the optimum threshold value that provides us with high Accuracy, Specifity and Sensitivity.

```
stats <- function(x, model) {

            results <- model %>%
            predict(dfTest1, type='response') %>%
              bind_cols(dfTest1, predicted=.) %>%
              mutate(predictedBookingRate =
as.factor(ifelse(predicted>x,1,0)))

            spec = specificity(results$predictedBookingRate,
as.factor(results$high_booking_rate))
            sens = sensitivity(results$predictedBookingRate,
as.factor(results$high_booking_rate))

            return(c(list(cutoff = x, specificity = spec, sensitivity =
sens), results))
}
```

```r
x <- c(0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)
data <- data.frame(cutoff = c(), specificity = c(), sensitivity = c())

for (val in x) {
  data <- rbind(data, stats(val, modelBoost)[1:3])
}

ggplot(data, aes(cutoff)) +
  geom_line(aes(y=specificity, color = 'Specificity')) +
  geom_line(aes(y=sensitivity, color = 'Sensitivity')) +
  ggtitle("Sensitivity and Specificity Analysis vs Cutoff") +
  scale_fill_discrete(breaks = c("Specificity", "Sensitivity"))
```

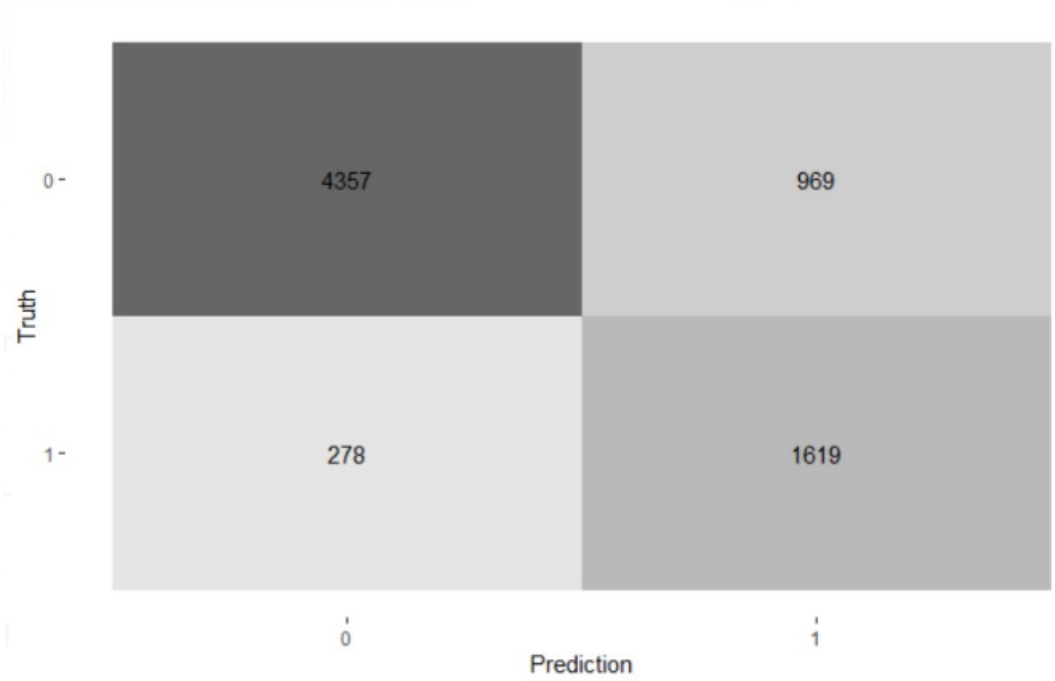


- Based on the graph above, we observe that the optimum threshold value is approximately 0.33.
- If the investor is risk seeking, we would aim for higher sensitivity and lower specifity which can be obtained by lowering the threshold value.
- If the investor is risk averse, we would aim for a lower sensitivity and higher specifity which can be obtained by increasing the threshold value.
- In our model, we are assuming that the investor is risk neutral, thus we have taken a cut-off of 0.33 which strikes the right balance between specificity and sensitivity.

Confusion Matrix and Statistics for best Cutoff

We observe that the model produces a high AUC value of 91.72%. We can conclude that the investor can rely on our predictions for high booking rate properties in LA region.

```
results <- modelBoost %>%
            predict(dfTest1, type='response') %>%
              bind_cols(dfTest1, predicted=.) %>%
              mutate(predictedBookingRate =
as.factor(ifelse(predicted>0.3,1,0)))

cm <- results%>%
  xtabs(~predictedBookingRate+high_booking_rate, .) %>%
  confusionMatrix(positive='1')


results%>%
  mutate(high_booking_rate= as.factor(high_booking_rate)) %>%
  conf_mat(truth='high_booking_rate', estimate='predictedBookingRate') %>%
  autoplot(type='heat_map')
```



Conclusion

The exploratory analysis above highlighted some interesting patterns and factors that can help elevate the booking rate:

- Investors should invest in properties that are around the prime locations such as Marina Bay, Silver Lake, Westchester.
- For upgrades, investors should invest in statistically significant factors such as Host is superhost, child friendliness, to achieve higher booking rate.
- Cheaper Airbnb rentals have higher booking rate than costlier rentals. Hence, investors should invest in the cheaper rentals for significant higher booking rate.

LA being one of the biggest business hubs, we wanted to gather insights from this segment of the market. However, we were limited in our research as the variable "is_business_ready" did not have well distributed data. We would like to further analyze this segment of the market and suggest solutions to cater to their needs.

References

Barzilay, O. (2017. April 4). 10 Things To Consider Before Buying An Airbnb Investment. Retrieved May 2, 2020, from https://www.forbes.com/sites/omribarzilay/2017/04/04/things-to-consider-before-buying-an-airbnb-investment/#2ce74d5c4869

Inside Airbnb: Los Angeles. Adding data to the debate. (n.d.). Retrieved May 2, 2020 from http://insideairbnb.com/los-angeles/

Shatford, S. (2019, June 27). Vacation Rental Investment Case Study | Best Places To Buy. Retrieved May2, 2020, from https://www.airdna.co/blog/short-term-real-estate-investment