How can I maximize my Airbnb list price in San Francisco?

Capstone Project by Jessie Huang
Mentor: Shmuel Naaman

Project Motivation

Purpose

Understand how San Francisco Airbnb hosts can maximize their listing price.

Goal

Identify the most important features to predict San Francisco listing prices.

Data Source

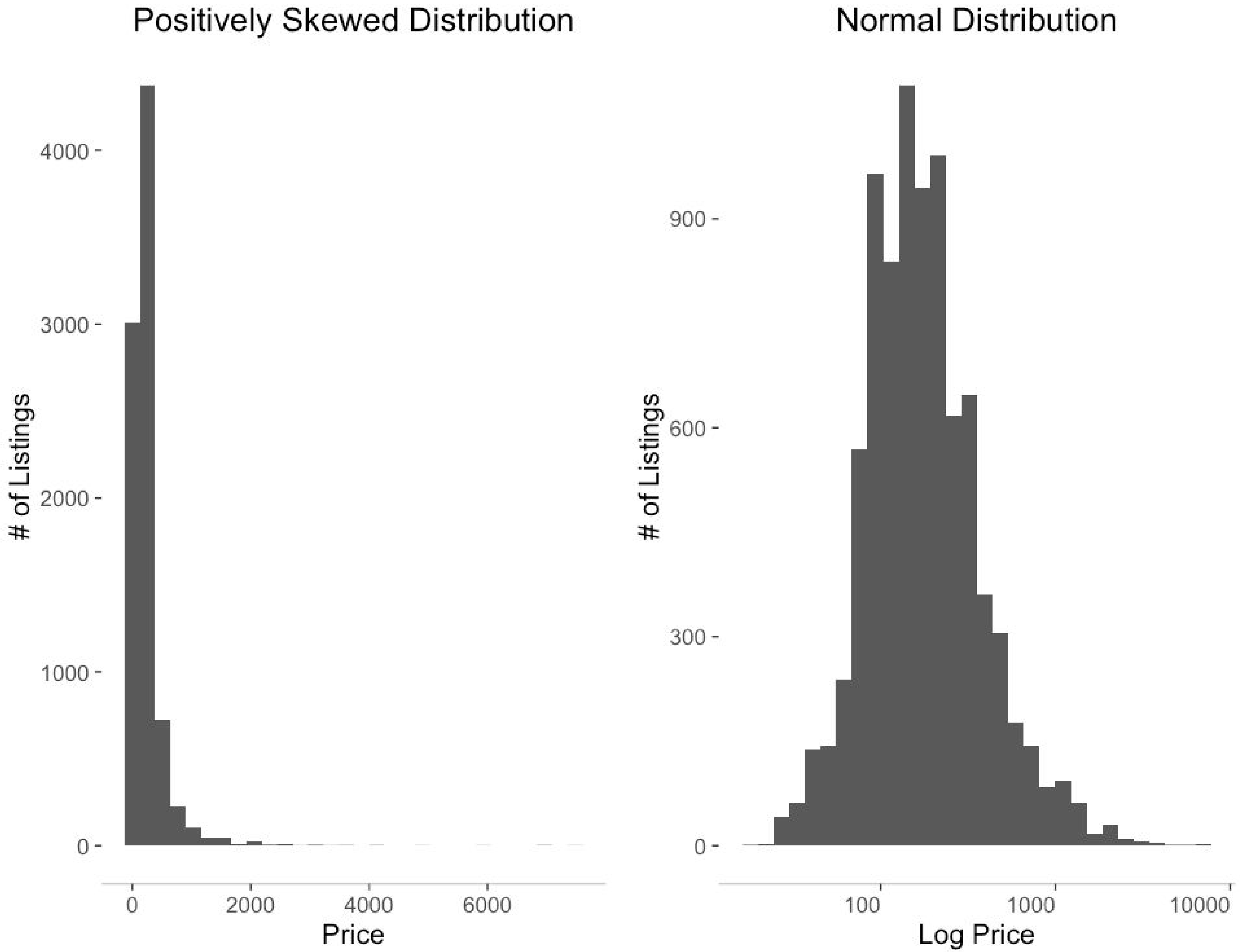
- The data came from Inside Airbnb, which is an independent and non-commercial website that provides tools and data to understand and explore how Airbnb is being used around the world.
- The raw dataset is 8,619 rows long and 96 columns wide
- Data limitations:
  - No booking data - cannot understand trends throughout the calendar year or most frequently booked listings
  - Dataset is small

Exploratory Analysis

Initial observations

- Total Listings: 8,588 (as of July 2, 2016)
- Price per night:
  - Average price: $245
  - Median price: $165
  - Price range: $19 to $7,500
- Accommodation size-related features influence the price.
- ~25% of the listings seem to have never been booked before.

The log price scale allows for a normal distribution.



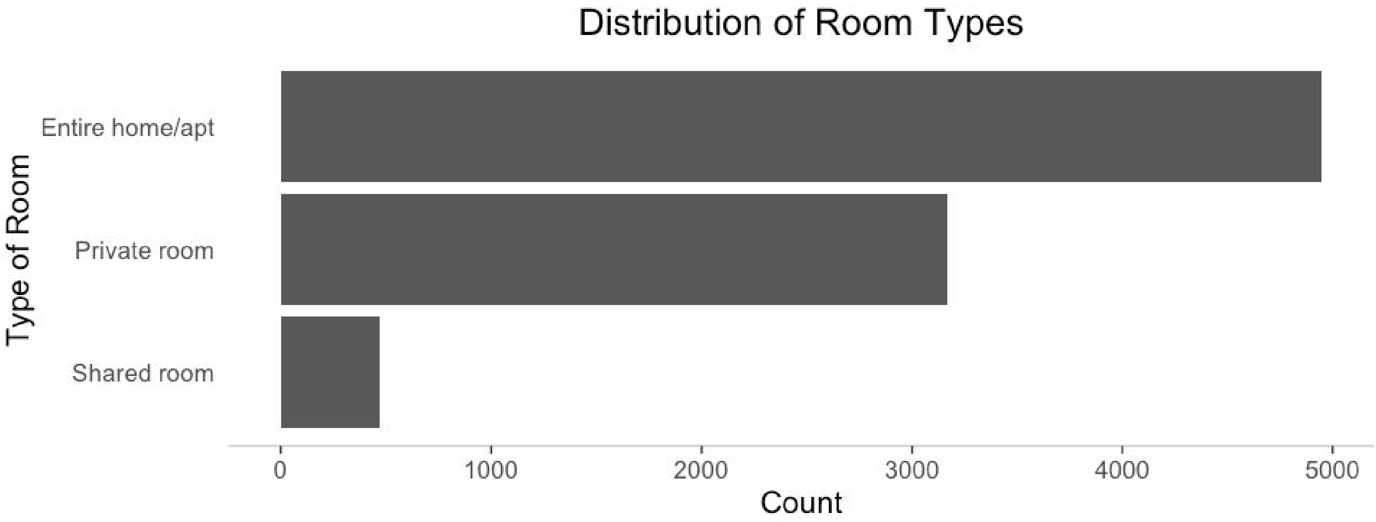
Statistical Analysis

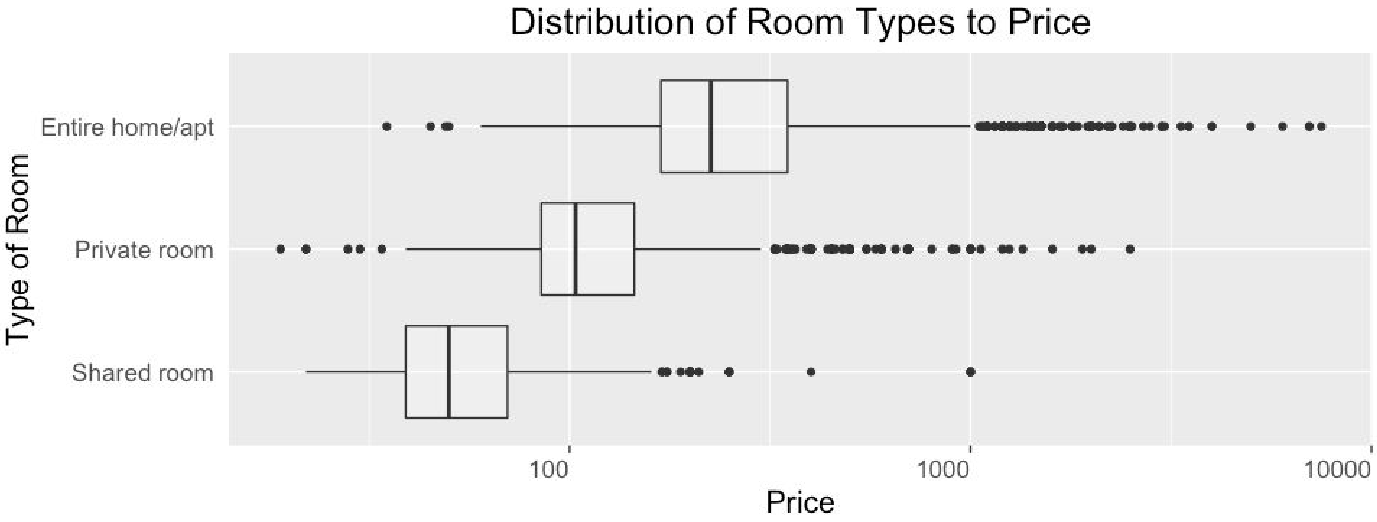
Analyzing data:

- Used the log of price feature for a normal distribution
- Performed linear model to test the categorical features
- Performed correlation test on numerical features
- Identified the features that are statistically significant different from the mean log price for predictive modeling, where $p < 0.05$

58% of listings are of an entire home.

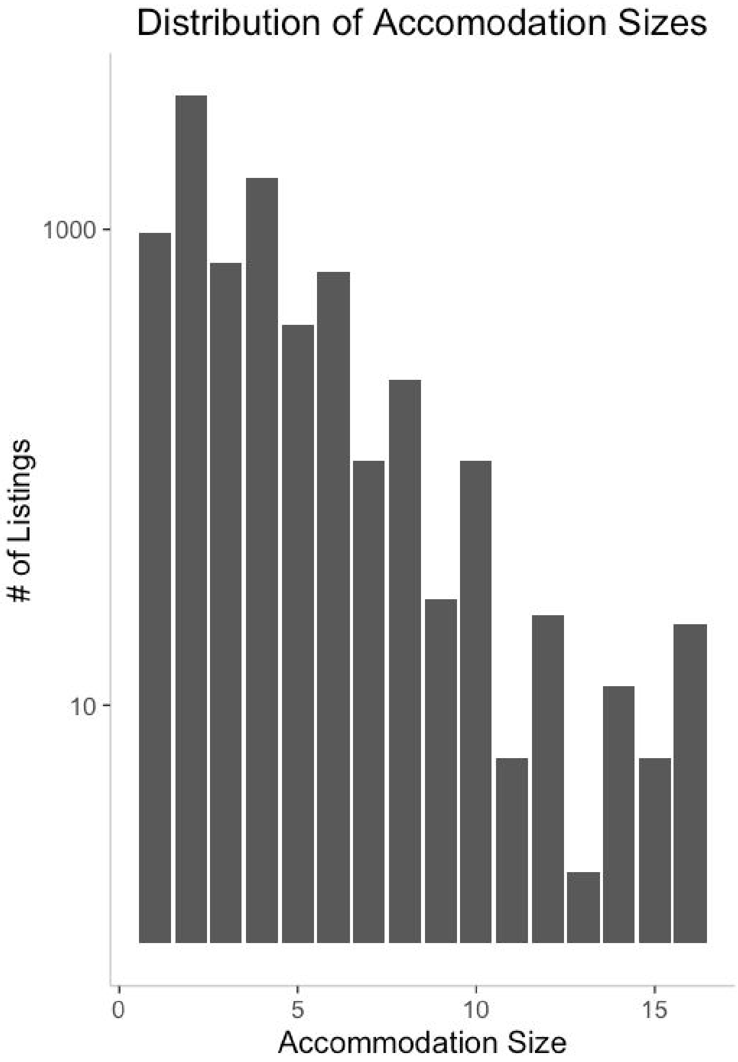
All room types have a statistically significant difference from the mean to the listing price with $p < 0.05$.

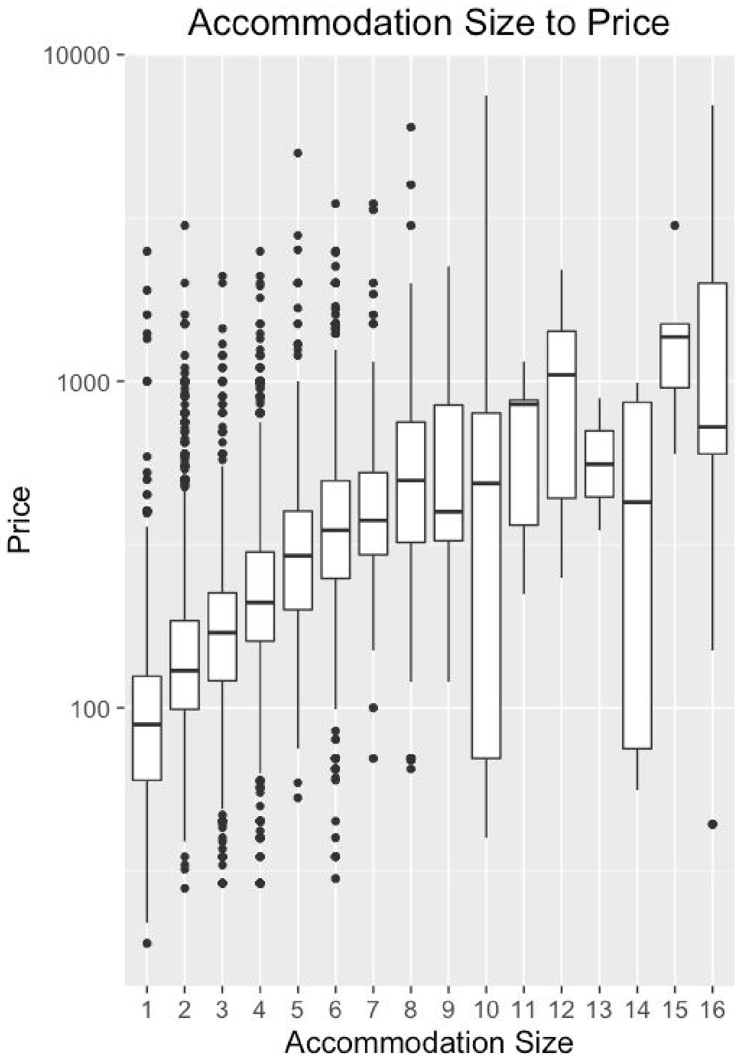




43% of listings accommodate 2 people.

However, **all accommodation sizes** have a statistically significant difference from the mean to influence the price, where $p < 0.05$.

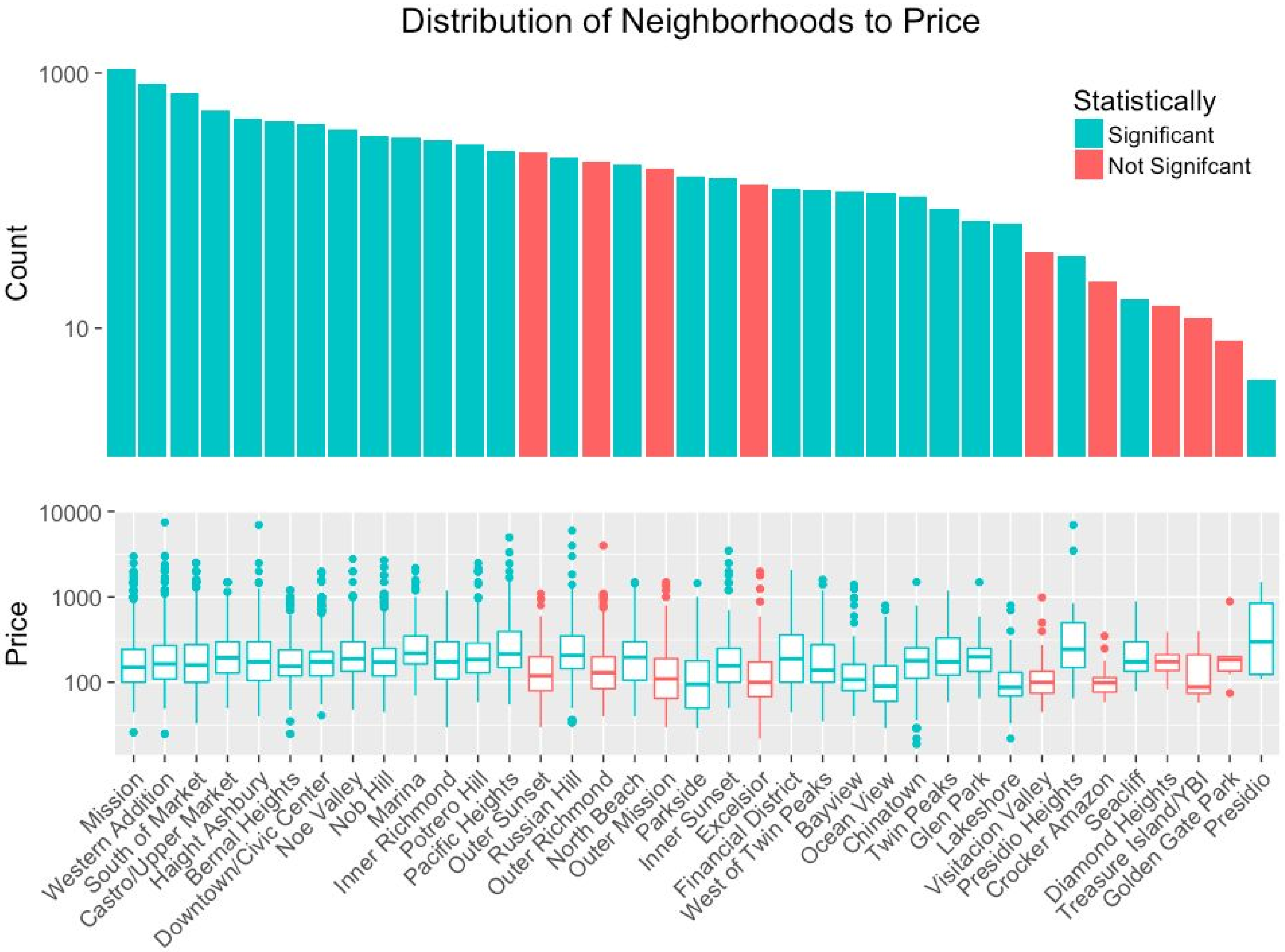




27% of listings have not received any review scores.

Only review score ratings of 95-100 have statistical significance where $p < 0.05$.

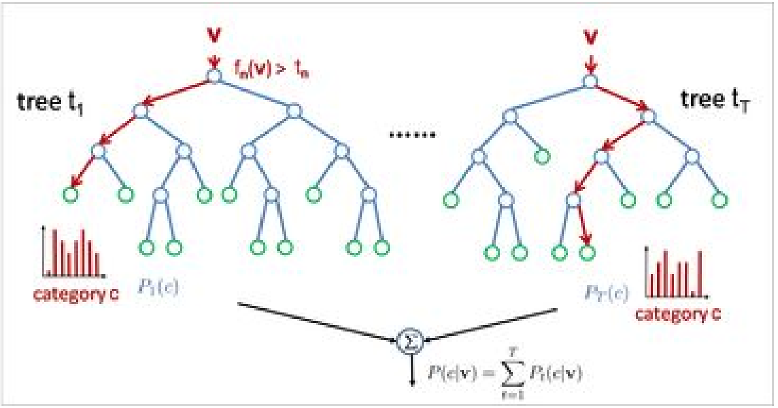
The price for 9 of 37 neighborhoods is not statistically different from the average (indicated in red). All other neighborhoods are significant with $p < 0.05$.



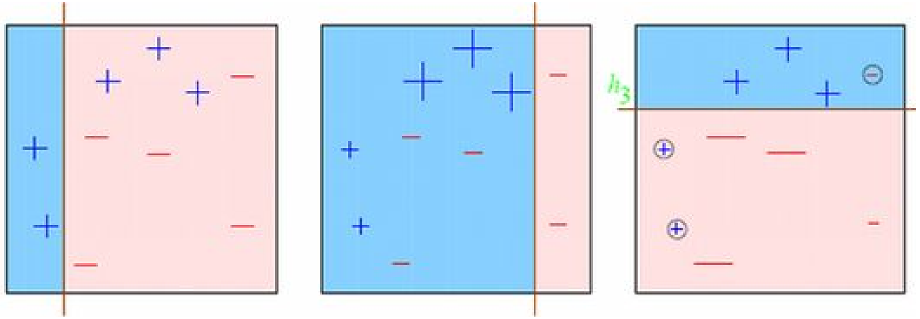
Predictive Modeling

Predictive Models:

Random Forest



Generalized Boosted Regression (GBM)

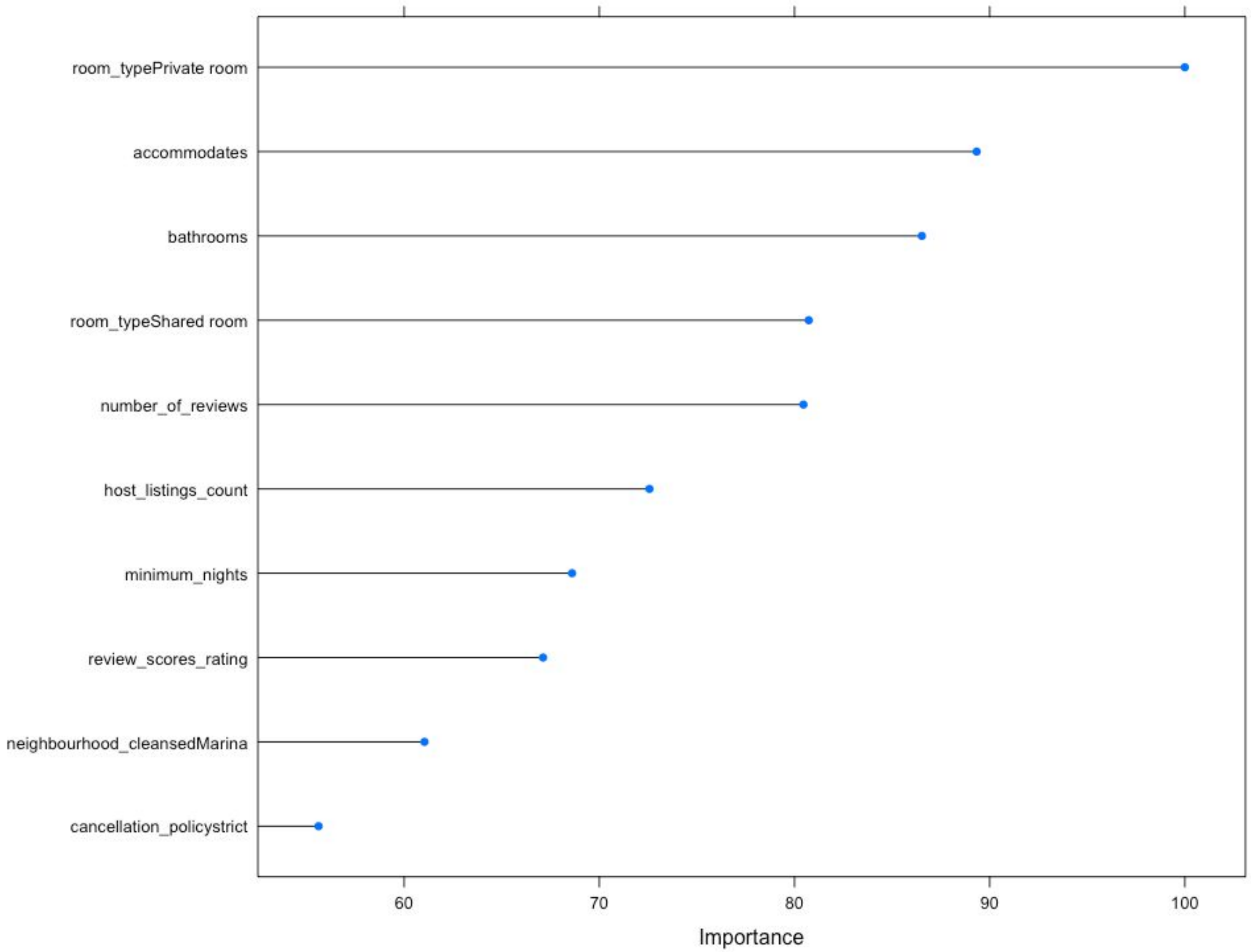


Features to be tested:

- neighbourhood_cleansed
- room_type
- accommodates
- host_listings_count
- minimum_nights
- is_dorm
- bathrooms
- beds
- bedrooms
- bed_type
- number_of_reviews
- reviews_per_month
- review_scores_rating
- cancellation_policy

Random Forest Results

R squared: 0.6275



Random Forest Grid Search

After tuning the model using grid search, we chose the model where ntree = 250 and r-squared = 0.6711.

```
##
## Call:
## summary.resamples(object = results)
##
## Models: 50, 100, 150, 200, 250
## Number of resamples: 10
##
## RMSE
##          Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## 50  0.4363  0.4550 0.4595 0.4604  0.4665 0.4893    0
## 100 0.4371  0.4532 0.4593 0.4594  0.4642 0.4893    0
## 150 0.4355  0.4521 0.4588 0.4590  0.4646 0.4904    0
## 200 0.4350  0.4506 0.4596 0.4586  0.4636 0.4902    0
## 250 0.4362  0.4495 0.4592 0.4583  0.4632 0.4897    0
##
## Rsquared
##          Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## 50  0.6053  0.6528 0.6583 0.6653  0.6886 0.7139    0
## 100 0.6100  0.6574 0.6636 0.6686  0.6886 0.7169    0
## 150 0.6098  0.6587 0.6656 0.6699  0.6912 0.7182    0
## 200 0.6099  0.6594 0.6664 0.6709  0.6937 0.7197    0
## 250 0.6104  0.6587 0.6667 0.6711  0.6944 0.7209    0
```

GBM Cross Validation

The GBM model is a much better predictor than the random forest model. The model we select is where ntree = 250, depth = 10 and r-squared = 0.6890.

```
Stochastic Gradient Boosting

6871 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 6184, 6183, 6184, 6185, 6182, 6185, ...
Resampling results across tuning parameters:
```

interaction.depth	n.trees	RMSE	Rsquared
2	50	0.5062636	0.5876091
2	100	0.4673742	0.6217289
2	150	0.4528664	0.6407134
2	200	0.4442050	0.6520843
2	250	0.4386009	0.6594164
5	50	0.4665818	0.6297245
5	100	0.4378124	0.6623628
5	150	0.4283085	0.6742325
5	200	0.4237667	0.6802034
5	250	0.4213890	0.6835016
7	50	0.4588906	0.6389040
7	100	0.4313829	0.6707679
7	150	0.4237993	0.6801691
7	200	0.4204499	0.6847993
7	250	0.4188586	0.6870795
10	50	0.4512608	0.6491306
10	100	0.4263867	0.6771808
10	150	0.4208796	0.6842425
10	200	0.4188025	0.6871583
10	250	0.4175592	0.6890170

Conclusion

- The most important features that influence the price is the number of people the listing can accommodate, the listing's reviews, and the location of the listing in San Francisco.
- However, we cannot conclude whether the statistically significant features help to increase the likelihood of a listing to being booked.
- Also mentioned earlier is that there is no booking data, so it would be interesting to collect this data in a creative way to understand booking trends.

Learnings

- Have a strategy to analyze your data so that data wrangling is a more efficient process.
- Make sure you understand the statistical foundations and concepts rather than trying to achieve a sophisticated end product. This also helps you know what is happening in R so you have an idea of how hard it needs to work to provide the results.
- Know the limitations of your dataset and understand how that could impact your results.