Predicting the Booking Rate for Airbnb San Diego Market

Market assigned to the team: **San Diego**

"We, the undersigned, certify that the report submitted is our own 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."

	Names of the signed team members
Contact member	Ramana Sriwidya
Team member 2	Shilpa Gupta
Team member 3	Laura Hayes
Team member 4	Yu (Zoey) Zhou
Team member 5	Kyle Reedy

Table of Contents

Background and Introduction

Airbnb[1] is a marketplace that offers customers lodging, homestays and different amenities such as experiences and tourism in the location they have booked the Airbnb. The company acts as a platform that brings the tourists and hosts together and collects a commission on each booking or transaction of experiences. Airbnb does not own any of the listed properties.

We have taken our data from Kaggle [2] to study the booking rates of Airbnb listings across the US market as a whole and then focused our research on the San Diego Market. The main objective of our project is to help Airbnb to predict and maximize their booking rates with the help of various variables we analyze. This research can help Airbnb hosts to maximize their bookings and hence increase their revenues as well as the revenues of Airbnb.

Executive Summary

This report focuses on the Airbnb data for the US market as a whole and more focussed research on the San Diego market data. We have analyzed the booking rates of Airbnb listings in the market and tried to predict the high booking rates of certain hosts. This research has resulted in 88% accuracy of our predictions. These predictions and analysis of various variables will help Airbnb and the hosts to help understand the customer preferences in the Airbnb market. This will in turn help them to increase their bookings and revenues.

This report will further describe the research methodology used for the various models tried in predicting the accurate booking rates, the variables which are statistically significant in the predictions and the impact those variables have on the predictions. Further, we will answer the questions included in the report and will present our findings and recommendations for the Airbnb company and their hosts to help them improve their listings and increase their revenues.

Main Focus and Questions

On starting the business case development for our San Diego market, below two questions came to our mind:

Question 1:

- How can we predict the probability of a high booking rate for all the listings of Airbnb?
- What is the accuracy of predicting such a probability?

Question 2:

- What all variables are significant for predicting the high booking rate probability?
- Which all variables are statistically significant?
- Among the statistically significant variables, which one has the highest impact on the booking rate?
- Can the booking rate be changed by changing these variables?

We analyzed the Airbnb data available from Kaggle and performed an extensive cleaning of the dataset. After obtaining the variables from the dataset and running several regressions, we found that there are a few variables that are statistically significant in predicting the booking rates of Airbnb listings. Few such variables are as below:

1. If a host is a superhost: If the host has gotten consistently good reviews over at least a year of hosting

[1] Airbnb
[2] Airbnb dataset on Kaggle

2. If host identity is verified
3. If a cleaning fee is included in the price
4. How strict is the cancellation policy
5. If guests are allowed and included
6. Count of host listings
7. Minimum nights allowed in the listing
8. Price of the particular listing
9. Review scores and
10. Room type (private or shared)

Among the statistically significant variables, we identified that being a superhost is particularly influential in changing the probability of listing. We obtained a coefficient of 1.26 for this variable which means being a superhost can increase the odds of a high booking rate by a factor of 3.5.

Hence, if a host has consistently got good reviews on his/her booking for a year, it can result in a high booking rate for that particular host and listing. This research is extremely useful and important for any company if they want to study customer preferences and how these preferences can be leveraged to increase the revenues for hosts and Airbnb and to provide customers with the experience they want.

Methodology

We used two methods to uncover the hidden pattern in the data. Random forest and logistic regression.

We first used the random forest model from the national Kaggle project for prediction. The random forest model is a form of machine learning method, and we chose this model because this model performs well with big data sets and our model achieved 88% AUC for validation, which implied the high performance of accuracy in prediction ability. With the random forest model, we have the ability to find properties with high booking rates in the future.

Before we ran the random forest model, we carefully cleaned and prepared our data. Here are the steps:

1. Remove all non categorical and no numeric columns, such as house description.
2. Remove dollar signs so that R can recognize all the number values.
3. Putting median or mean values to the houses that have missing values. For example, we used median for bedroom numbers and mean for square feet.
4. Transform the logical variables into factors so that we can use them as categorical variables, which means they only have certain dummy variables such as “bed_type”has “Couch”, “Futon”, “Pull-out Sofa”, and “ Real Bed”. Each variable has a value of either 0 or 1 corresponding to no or yes. Each dummy variable absorbs the variety in groups so that you can compare which variable is more statistically important. We transformed “market”, “bed_type”, “room_type”, “host_response_time”, and “cancellation policy”.
5. Transform the host_since value to host_days so that we can know how many days since the house joined the Airbnb listing. We used the current date minus host_since to get the host_days value.

6. Other variables we chose are numeric.

Random Forest Model Results

```
summary(fit_rf$fit)
```

```
##                 Length Class  Mode
```

```
## call               6   -none- call
## type               1   -none- character
## predicted        618   factor numeric
## err.rate        1500   -none- numeric
## confusion          6   -none- numeric
## votes           1236   matrix numeric
## oob.times        618   -none- numeric
## classes            2   -none- character
## importance        38   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            14   -none- list
## y                618   factor numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
```

Confusion Matrix

```
summary(conf_mat_assess, event_level='second')
```

```
## # A tibble: 13 x 3
##    .metric              .estimator .estimate
##    <chr>                <chr>          <dbl>
##  1 accuracy             binary         0.827
##  2 kap                  binary         0.470
##  3 sens                 binary         0.489
##  4 spec                 binary         0.933
##  5 ppv                  binary         0.697
##  6 npv                  binary         0.853
##  7 mcc                  binary         0.482
##  8 j_index              binary         0.422
##  9 bal_accuracy         binary         0.711
## 10 detection_prevalence binary         0.168
## 11 precision            binary         0.697
## 12 recall               binary         0.489
## 13 f_meas               binary         0.575
```

Even though the random forest model is good at prediction, we would still want to know more about the key variables that are statistically important. And the important variables would show us the strategy on how to improve a property's booking rate. Thus, we used a logistic regression model to find the statistically important variables. We used the same steps to clean the data for the logistic regression. After running the regression, we used AIC, AUC values to check the accuracy of the model. Then we looked at t-statistics and p-value which can show how significant the variables are.

We used 0.5 as cutoff for classification in prediction for the random forest model.

Log Model Results

```
summary(fit_log$fit)
```

```
##
## Call:
## stats::glm(formula = high_booking_rate ~ . - id, family = stats::binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9534  -0.5714  -0.1672   0.6226   2.6923
##
## Coefficients: (1 not defined because of singularities)
##                                              Estimate Std. Error z value
## (Intercept)                                 2.089e+02  6.338e+03   0.033
## accommodates                                8.121e-02  1.104e-01   0.736
## availability_365                            1.308e-03  1.129e-03   1.159
## bathrooms                                   4.364e-01  2.984e-01   1.462
## bed_typeCouch                              -3.204e+01  4.567e+03  -0.007
## bed_typeFuton                              -2.446e+00  4.615e+03  -0.001
## bed_typePull-out Sofa                      -3.387e+00  4.762e+03  -0.001
## bed_typeReal Bed                           -1.972e+01  3.956e+03  -0.005
## bedrooms                                   -5.385e-01  2.367e-01  -2.275
## beds                                        2.157e-01  1.602e-01   1.346
## cancellation_policymoderate                 9.603e-01  3.493e-01   2.749
## cancellation_policystrict_14_with_grace_period  1.048e+00  3.490e-01   3.002
## cancellation_policysuper_strict_30         -1.340e+01  1.519e+03  -0.009
## cancellation_policysuper_strict_60          1.693e+00  1.051e+00   1.611
## cleaning_fee                               -1.033e-02  3.061e-03  -3.373
## extra_people                               -2.709e-03  5.576e-03  -0.486
## guests_included                             1.551e-01  7.768e-02   1.996
## host_has_profile_pic                        1.503e+01  2.382e+03   0.006
## host_identity_verified                      5.613e-01  2.683e-01   2.092
## host_is_superhost                           1.477e+00  2.740e-01   5.389
## host_listings_count                        -1.997e-02  9.948e-03  -2.007
## host_response_timeN/A                       1.439e+01  1.700e+03   0.008
## host_response_timewithin a day              1.420e+01  1.700e+03   0.008
## host_response_timewithin a few hours        1.603e+01  1.700e+03   0.009
## host_response_timewithin an hour            1.625e+01  1.700e+03   0.010
## instant_bookable                            3.538e-01  2.574e-01   1.375
## is_location_exact                           6.768e-02  3.119e-01   0.217
## latitude                                    3.911e+00  2.469e+00   1.584
## longitude                                   2.834e+00  2.383e+00   1.190
## marketSan Diego                             1.359e+01  1.955e+03   0.007
## marketTijuana                              -4.873e+00  2.756e+03  -0.002
## maximum_nights                             -9.730e-05  2.364e-04  -0.412
## minimum_nights                             -4.755e-02  2.106e-02  -2.258
## price                                      -2.865e-03  1.358e-03  -2.109
## require_guest_phone_verification           -1.426e+00  1.329e+00  -1.073
## require_guest_profile_picture              -1.702e-01  1.502e+00  -0.113
## requires_license                                   NA         NA      NA
## review_scores_accuracy                      3.786e-01  4.060e-01   0.933
```

```
## review_scores_checkin                            3.549e-01  4.557e-01   0.779
## review_scores_cleanliness                       -4.537e-01  2.639e-01  -1.719
## review_scores_communication                     -6.164e-01  4.577e-01  -1.347
## review_scores_location                           1.623e-01  3.339e-01   0.486
## review_scores_rating                             7.078e-02  4.748e-02   1.491
## review_scores_value                             -5.020e-01  2.920e-01  -1.719
## room_typeHotel room                             -1.605e+01  2.143e+03  -0.007
## room_typePrivate room                           -9.476e-01  3.637e-01  -2.605
## room_typeShared room                            -1.065e+00  1.232e+00  -0.865
## security_deposit                                -9.367e-04  5.037e-04  -1.859
## square_feet                                     -4.620e-02  4.971e+00  -0.009
## host_days                                        3.474e-04  1.767e-04   1.966
##                                                Pr(>|z|)
## (Intercept)                                    0.973707
## accommodates                                   0.461830
## availability_365                               0.246613
## bathrooms                                      0.143678
## bed_typeCouch                                  0.994402
## bed_typeFuton                                  0.999577
## bed_typePull-out Sofa                          0.999433
## bed_typeReal Bed                               0.996022
## bedrooms                                       0.022877 *
## beds                                           0.178210
## cancellation_policymoderate                    0.005979 **
## cancellation_policystrict_14_with_grace_period 0.002680 **
## cancellation_policysuper_strict_30             0.992966
## cancellation_policysuper_strict_60             0.107094
## cleaning_fee                                   0.000742 ***
## extra_people                                   0.627057
## guests_included                                0.045900 *
## host_has_profile_pic                           0.994966
## host_identity_verified                         0.036421 *
## host_is_superhost                              7.08e-08 ***
## host_listings_count                            0.044715 *
## host_response_timeN/A                          0.993246
## host_response_timewithin a day                 0.993336
## host_response_timewithin a few hours           0.992476
## host_response_timewithin an hour               0.992375
## instant_bookable                               0.169228
## is_location_exact                              0.828217
## latitude                                       0.113107
## longitude                                      0.234231
## marketSan Diego                                0.994454
## marketTijuana                                  0.998589
## maximum_nights                                 0.680637
## minimum_nights                                 0.023970 *
## price                                          0.034922 *
## require_guest_phone_verification               0.283467
## require_guest_profile_picture                  0.909740
## requires_license                                     NA
## review_scores_accuracy                         0.351029
## review_scores_checkin                          0.436141
## review_scores_cleanliness                      0.085586 .
## review_scores_communication                    0.178081
```

```
## review_scores_location                          0.626828
## review_scores_rating                            0.136037
## review_scores_value                             0.085618 .
## room_typeHotel room                             0.994026
## room_typePrivate room                           0.009183 **
## room_typeShared room                            0.387249
## security_deposit                                0.062958 .
## square_feet                                     0.992584
## host_days                                       0.049268 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 767.47  on 617  degrees of freedom
## Residual deviance: 464.00  on 569  degrees of freedom
## AIC: 562
##
## Number of Fisher Scoring iterations: 16
```

Results and Findings

Our main goal with this study was to predict the model that explains the best whether an Airbnb property will observe a high booking rate. Since we were assigned to San Diego, we focused our efforts on this specific Airbnb market. We accomplished our goal by running our data through the random forest model, first by using 1500 trees, then by using a less varied random forest model. Our last random forest model ended up being our strongest predictor with 88% accuracy. The random tree model was our best option for supporting our objective and revealing unknowns about the San Diego market because they are able to correct for mistakes that models like decision trees make in overfitting to the training set. Additionally, they generate reasonable predictions across a wide range of data. Since we were faced with a huge data set, we wanted to use a model that would accommodate this.

Overall, our models did very well in supporting our objectives and revealing unknowns about the specific market. Our AUC model predicted 90% accuracy, indicating a greater ability to make predictions using the model. As referenced earlier, among the statistically significant variables, we found that being a superhost is particularly influential in changing the probability of a listing being booked. We obtained a coefficient of 1.26 for this variable which means being a superhost can increase the odds of a high booking rate by a factor of 3.5.

Log Model AUC

```
roc_auc(results_to_assess, truth = high_booking_rate, predictedProb, event_level = 'second')
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.855
```

In order to be considered a superhost, the host must meet the following requirements[3].

[3] Airbnb Superhost Description

- Complete at least three reservations that total at least 100 nights
- Maintain a 90% response rate or higher
- Maintain a 1% percent cancellation rate (1 cancellation per 100 reservations) or lower, with exceptions made for those that fall under our Extenuating Circumstances policy
- Maintained a 4.8 overall rating (this rating looks at the past 365 days of reviews, based on the date the guest left a review, not the date the guest checked out)
- Be a host for at least 12 months

Therefore, after deep analysis, our recommendation to Airbnb's investors would be to take our data models and have a team of software developers and engineers develop a new feature to the Airbnb host dashboard. This newly developed feature would use our random forest model to identify whether the host's listings are predicted to have a high booking rate. As predicted by our model, four of these significant variables would be being a superhost, having a cleaning fee that is included in the price, having a cancelation policy with a 14 day grace period and offering private rooms. We also propose some general advice for the property owners to increase the days prior to the booking date that the property is available to rent, to respond to potential customers as soon as possible and have longer descriptions of the host property. The Airbnb host relations team can also use this information in the host guideline packet which is a comprehensive guide to impress the guests.

While we believe our findings are comprehensive given the data and tools we had access to, there were some limitations to our analysis. In the future, there should also be a qualitative component to the research in terms of checking with super hosts on what has worked for them and noting if this lines up with our findings. Additionally, our data may be skewed towards specific areas of San Diego depending on the spread of the property listings. However, overall, our research has provided a greater starting point for Airbnb investors to make improvements to their models for determining properties that have high booking rates.