MGSC 661 Final Project – Predicting Airbnb Ratings

Presented to: Prof. Juan Camilo Serpa

Submitted by: Arunima Agrawal (261001915), Smitul Soni (261001914)

1. Introduction

An important success factor for a business is understanding customer experience and satisfaction for maintaining goodwill and longevity of the business. For our project, we analyzed the customer satisfaction ratings for Airbnb, one of the most popular lodging service providers. Airbnb is a global online marketplace for renting homestays/ villas/ private rooms.

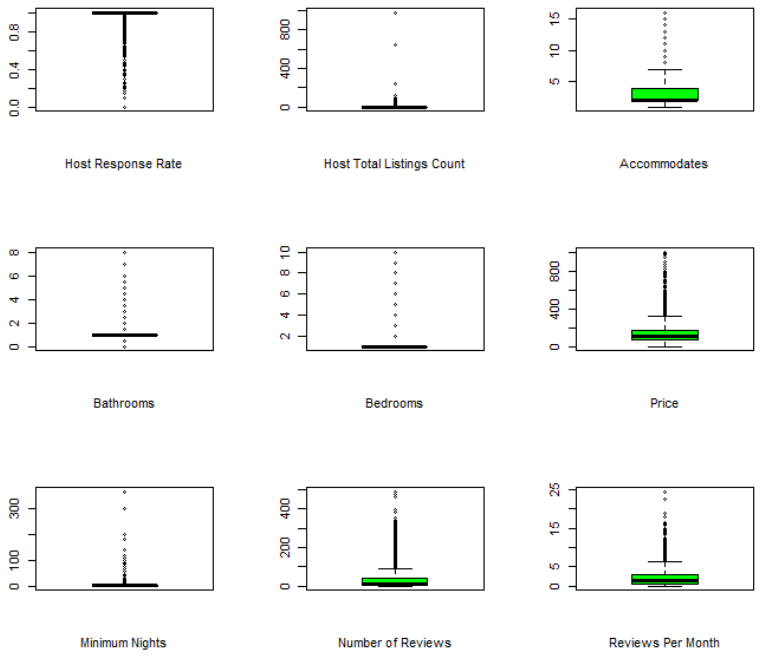
Airbnb maintains the user review scores for all its listings to ascertain customer satisfaction metrics. The aim of this project was to understand which factors affect the user ratings the most, and how Airbnb hosts can make modifications to their properties and services to boost customer satisfaction.

We had access to the Airbnb dataset on Kaggle containing around 18,000 listings in and around the city of New York, United States. The goal was to build a predictive model to predict the ratings for a subset of listings based on their feature definitions. The dataset consisted of 27 different features that could be used for a detailed analysis. The target variable was customer ratings, which was the final rating for the listing based on several criteria, including accuracy, cleanliness, communication, etc., given by each user.

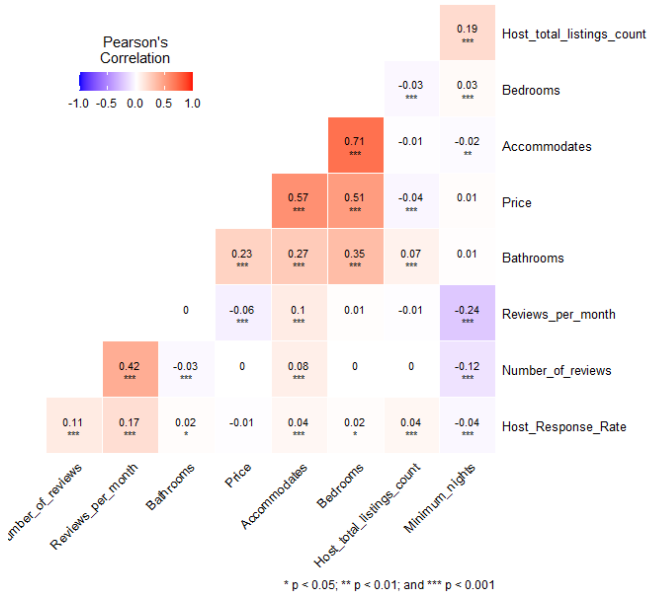
To build the best model for relevant prediction, it was necessary to find the importance of individual predictors and to choose optimal number of predictors in the model to avoid overfitting or underfitting the training data. Multiple tree-based machine learning models were tested to get reliable results and maximum accuracy in finding the rating of any Airbnb listing in New York.

2. Data Description

The first step towards the solution was exploratory data analysis and data transformation. We analyzed the predictors to understand their distribution in the dataset. For all the numeric variables, we plotted boxplots to visualize the distribution and detect outliers *(refer image below)*. We noted outliers for almost all the numeric variables and dropped them from the dataset to reduce the variability and avoid overfitting. *(for table describing the outliers removed for each variable, please refer* ***Appendix 1****).*

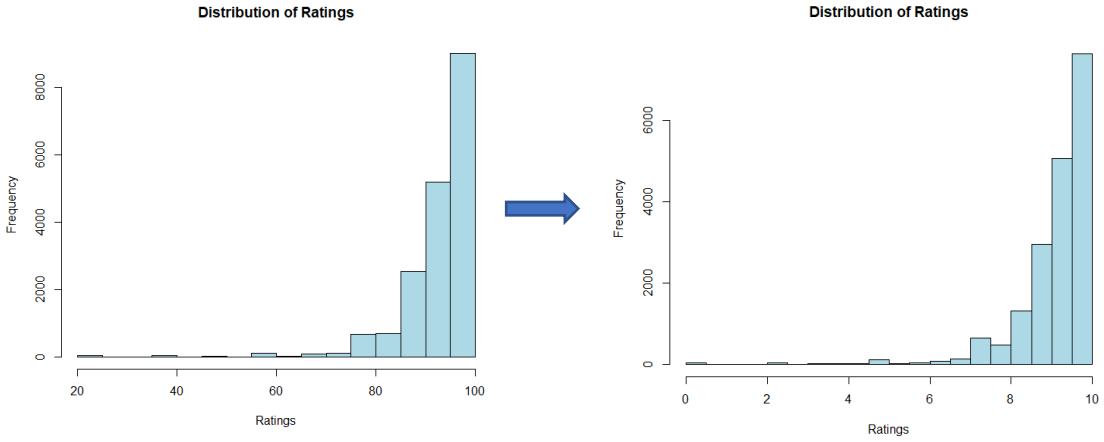


The correlation between numerical variables was identified by plotting the Pearson's correlation matrix.



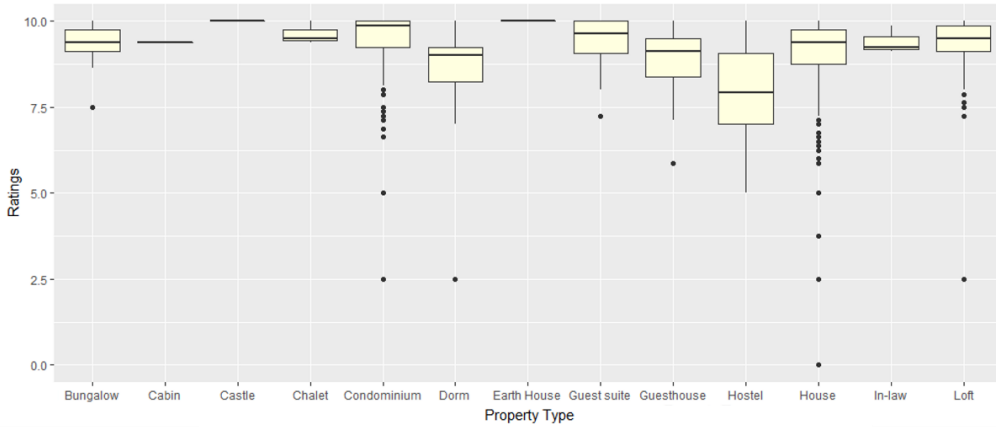
It was noted that the variable pair {accommodates, bedrooms} had a correlation of 0.71 and therefore bedrooms variable was removed from further analysis to prevent collinearity.

Initially, the target variable, 'Ratings' had the score in range of [0, 100]. We scaled down this variable to a range of [0, 10] to reduce the variability and to reduce the processing time for the model in predicting the target variable. The distribution of the Ratings field was plotted as histogram plots, and we observed that the ratings are left skewed.



In the initial dataset, the amenities provided by the listings were present in a consolidated fashion into one column named 'Amenities'. Multiple amenities were separated by semicolon. E.g., "Cable TV; Internet; Wireless Internet; Kitchen". From a total of 117 unique amenities, we identified 32 amenities that had a comparatively higher frequency and were more likely to be important. Further, we factored them to create 32 new columns for individual amenities.

We also roughly analyzed the spread of ratings for the different categories in categorical variables by boxplot visualizations. This included ratings analysis by city, property type, room type and host type. For instance, ratings for chalets or bungalows were in the higher range than the ratings of hostels, whose spread was comparatively larger (*for descriptions about the spread of other categorical variables, please refer* ***Appendix 2****).*

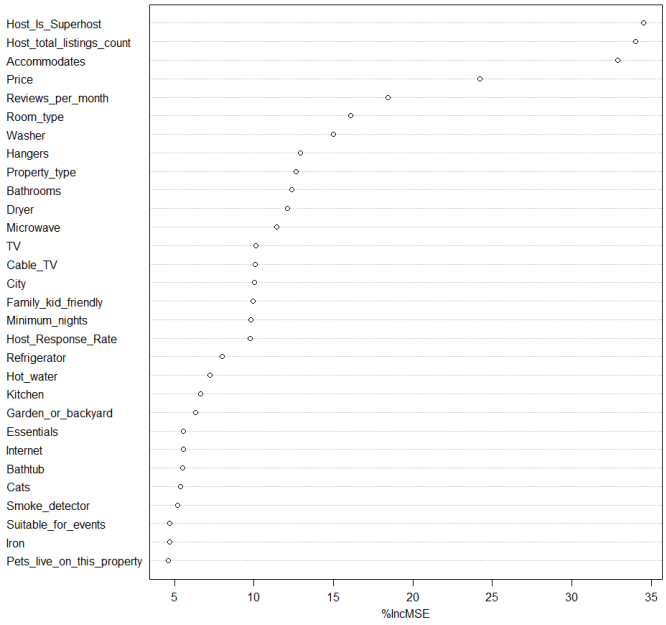


3. Model Selection & Methodology

Post exploratory analysis, we performed the following steps to create the optimal predictive model:

1. Finding the best predictors

Before selecting models for the prediction, it was necessary to identify the best predictors that could be included in our model. To identify the important variables out of the total 44 predictors, we ran random forest algorithm with all predictors in the dataset with the target variable Ratings and plotted the variable importance graph to check the variables with highest importance based on their 'Percent Increase in MSE'. As a result, it was identified that the host type is the most important predictor with 34.5% increase in MSE, suggesting that if we remove host type from our model, the MSE would increase by 34.5% of current MSE while all other predictors remain the same. This metric was further used as an ordered vector in individual models to find the optimal number of predictors according to the percent increase in MSE value (*for detailed table for % increase in MSE values, please refer* ***Appendix 3***).



'Percent Increase MSE' for each predictor indicating the predictor importance achieved by random forest

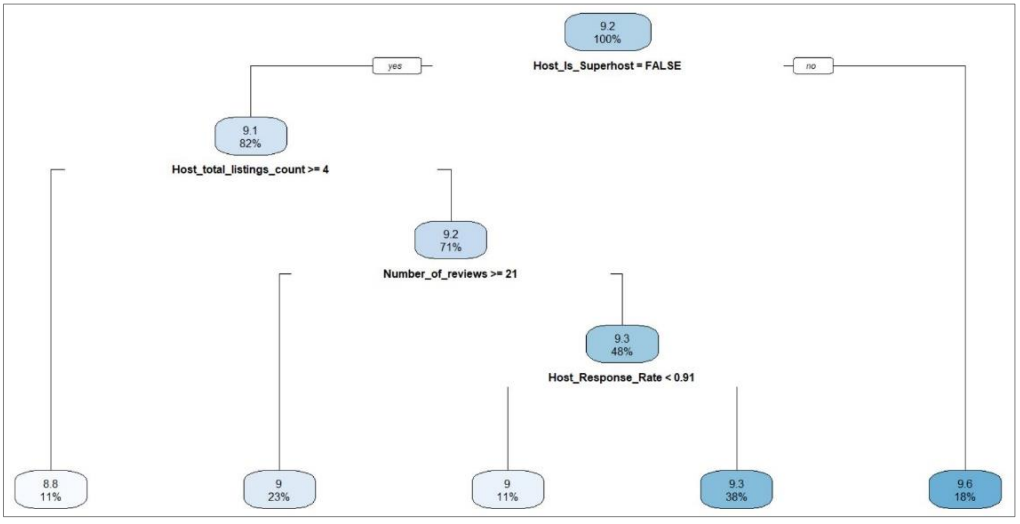
2. Prediction through multiple models

Since our target variable is a continuous value, we used tree-based regression models and not classification models. Hence, we performed regression analysis using Regression Tree, Random Forest, and Boosting models to predict the ratings.

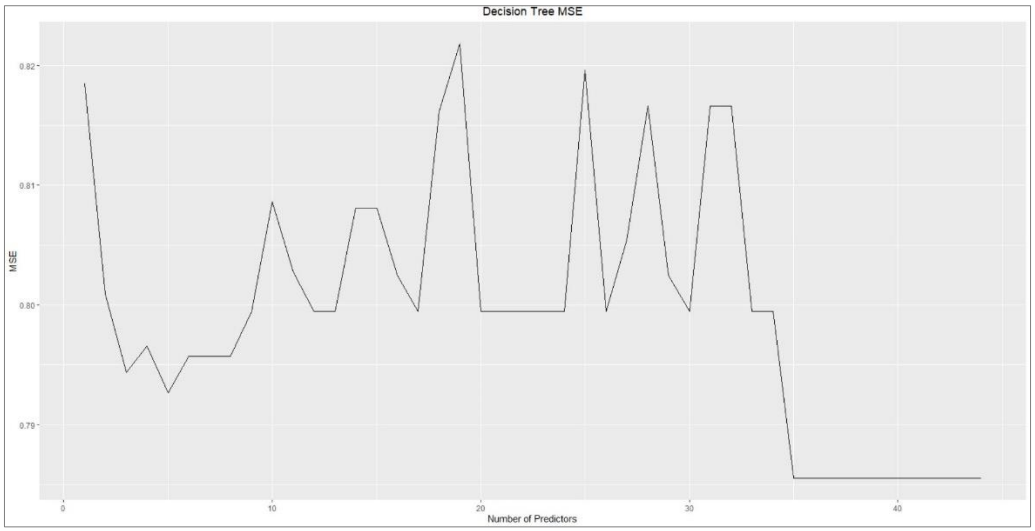
We split our complete dataset into train and test, with 70% for training the model and 30% for testing it. The optimal number of predictors were chosen dynamically for all the models. We ordered predictors based on their feature importance calculated by random forest. The first iteration picked the most important predictor, followed by the addition of the next important predictor in each iteration. Hence, for all the models, the MSE was calculated at each step, and the predictor combination with the lowest MSE was chosen as the best combination for that model. For selecting the final model, various factors like MSE score, compute time and potential associated costs were taken into account.

i. Regression Tree Model

The regression tree was built by splitting the data into optimal number of branches using the complexity parameter (cp) on predictors. Each iteration added the next most important predictor according to the random forest feature selection algorithm. The optimal cp value for the model was calculated and chosen individually at each iteration. Therefore, for each set of predictors, the regression tree was built using the optimal cp, and MSE was calculated. The optimal model with the lowest MSE of 0.7855 was obtained with the set of the first 35 most important predictors. However, the resulting plotted tree only consisted of 4 interaction nodes, considering only 4 predictors: host type, host total listings count, number of reviews, and host response rate. We inspected that the 'number of reviews' predictor was rated the 35th most important predictor by random forest, due to which the lowest MSE was obtained at 35th iteration. Hence, we note that the 'number of reviews' predictor is deemed more important by the regression model.



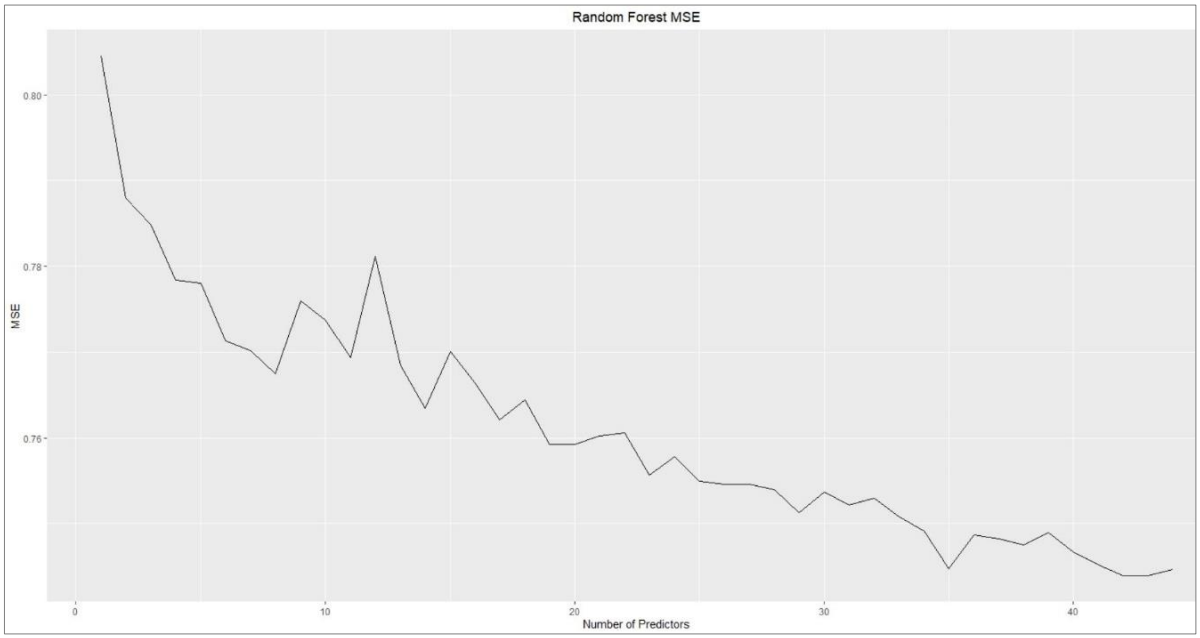
Regression tree obtained with the lowest MSE



MSE obtained for Regression Tree model at each iteration with different number of predictors

ii. Random Forest Model

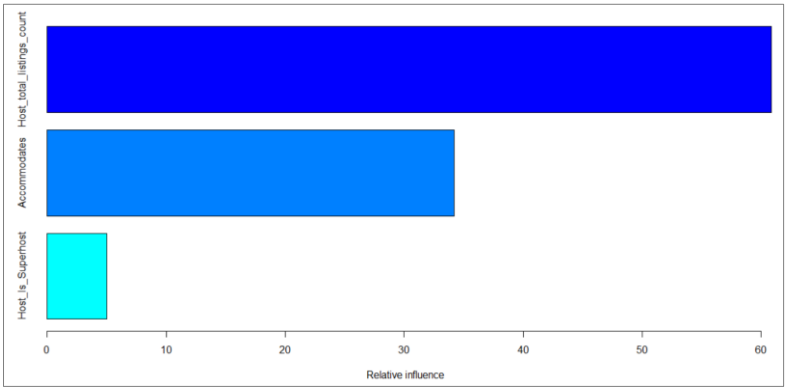
The second model under consideration was random forest. Like the regression tree, random forest model was also run p number of times, adding one predictor at a time in the order of their importance calculated by random forest feature selection. We specified 500 trees for the model, since the number of rows in the dataset were large (~18,000 rows). Due to this, the code took several hours to run the 44 loops (7-8 hours). The lowest MSE of 0.7439 was obtained at 42nd iteration. Therefore, this model deemed 42 out of the total 44 predictors to be important for analyzing the ratings. Out of all the 3 models used for the analysis, random forest took the maximum amount of time for running, and hence it was deemed to be the most expensive model computationally.



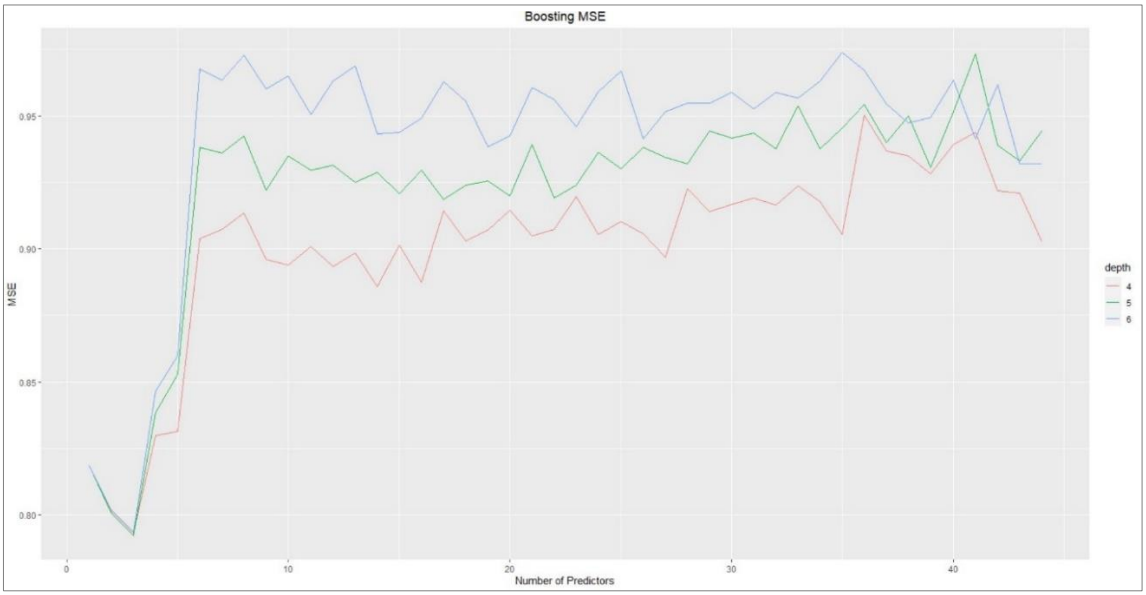
MSE obtained for Random Forest model at each iteration with different number of predictors

iii. Boosting

The third model used for our analysis was boosting using the gaussian distribution, since we are performing regression analysis. We considered 10,000 trees for both training and testing the boosting model. Since the number of rows are large, we considered multiple interaction depth values for the analysis. We analyzed the model for {4,5,6} set of interaction depth values, and ran nested loops to analyze the predictor combinations by varying interaction depths. For instance, for interaction depth 4, there were 44 iterations, adding predictors in the order of random forest feature importance. These 44 iterations were run for interaction depths 4, 5, and 6. The lowest MSE of 0.7921 was obtained at interaction depth 5 at the 3rd iteration with 3 predictors: host type, accommodation capacity, host total listings count. The model had 132 iterations in total (that is, 44 predictors * 3 interaction depths), but took lesser time (4-5 hours) as compared to the random forest model. The optimal boosting model expresses 'host total listings count' as the most influential predictor with 60.7% influence on the prediction, followed by 'accommodated' with 30.4% and lastly the 'host type' with 4.9% influence.



Relative influence of the three predictors obtained at interaction depth 5, 3rd iteration



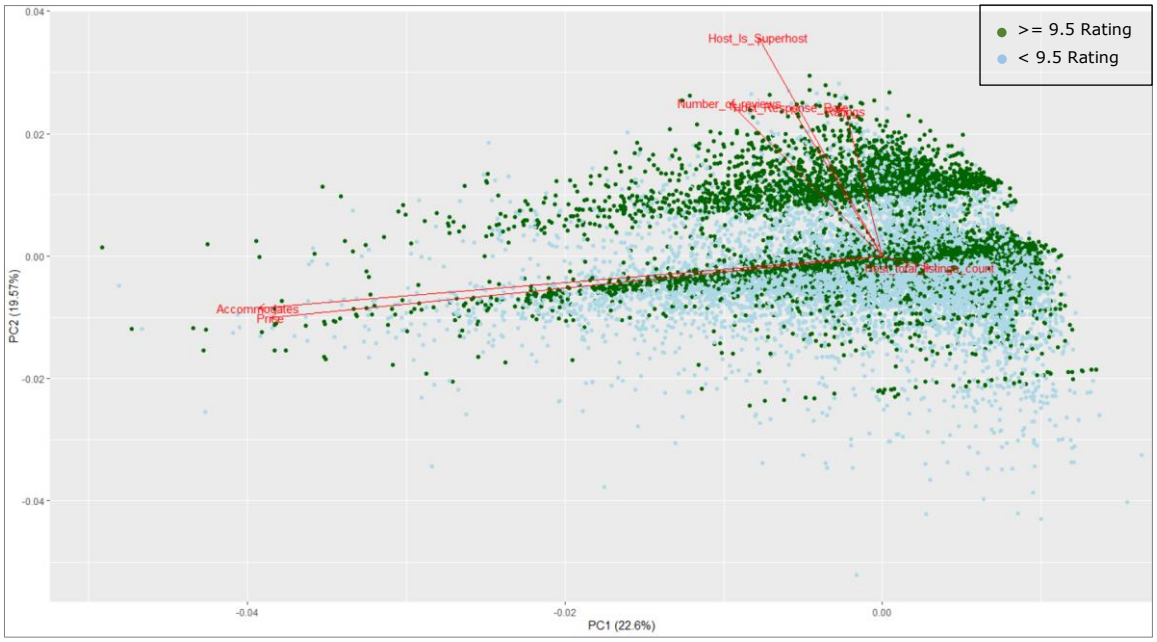
MSE obtained for Boosting model at each iteration with different number of predictors for interaction depths 4,5,6

3. Principal Component Analysis and Clustering

We did not use Principal Component Analysis for building models even though the number of predictors was high. This is because apart from obtaining the final predictions, we were also aiming to evaluate each predictor's relationship with the target variable. This would have been difficult to achieve with PCA. However, after building the model, we performed the PCA analysis over the 6 predictors that were identified to be important by the three models.

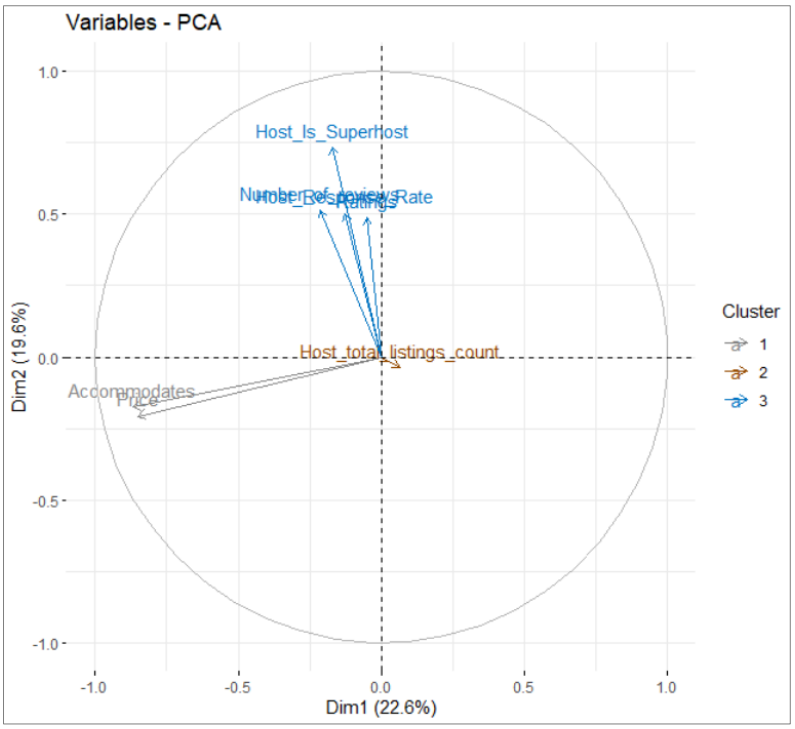
Noted for PC1 and PC2 that host type, number of reviews, and host response rate were directly affecting ratings. Host total listings count was affecting the ratings negatively. Although, price and accommodation capacity predictors were relevant in evaluating the ratings, they had no direct positive or negative relationship with ratings. We highlighted the highly rated property data points in green and observed that the density of properties with high ratings is high near the significant predictors.

We also analyzed PCA for all numeric predictors in data *(please refer to **Appendix 4** for details)*.



Principal Component Analysis for 6 important predictors in predicting the Ratings

We also attempted to run clustering analysis on the dataset, but the high number of rows and variability in the dataset resulted in clustering results that could not lead to relevant results. However, we applied the K-Means clustering algorithm on the PCA results, which resulted in dividing the predictors based on their relationship with the target variable ratings.



K-Means clustering on PCA analysis results

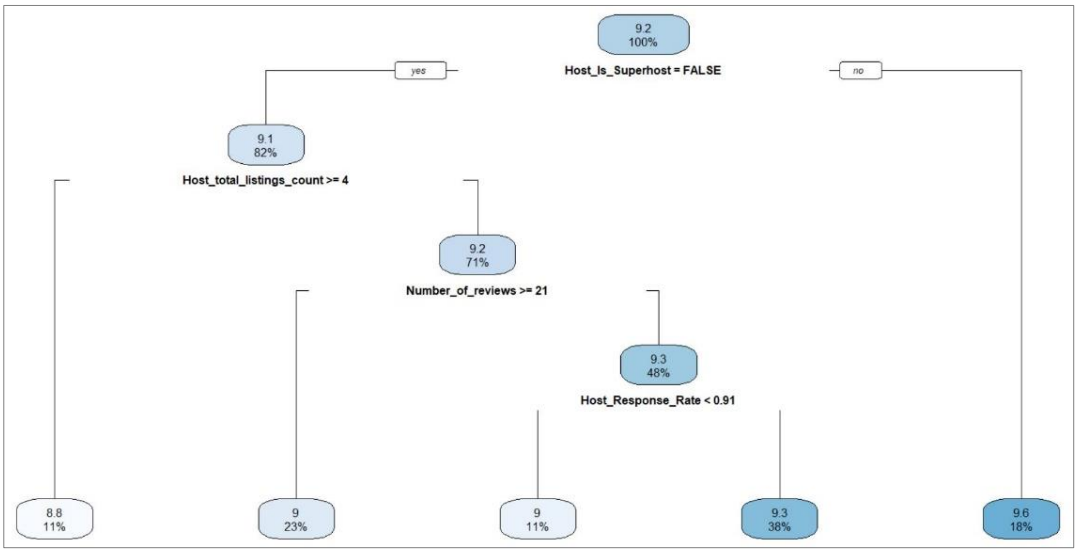
We noted that the cluster 3 member predictors, that is, host type, host response rate, and number of reviews, affect ratings positively. Likewise, the cluster 2 member predictor, that is, host total listings count, affect the ratings negatively. The member predictors in cluster 1, that is accommodation capacity and price, are significant in predicting the ratings, but do not have a direct positive or negative relation with ratings.

4. Results

Evaluating all the optimal regression models, we identified that the random forest model predicted the best ratings with the lowest MSE of 0.7439. In comparison to random forest, the regression tree and boosting models predicted the ratings with MSEs of 0.7855 and 0.7921 respectively. Although random forest model had the lowest MSE value, the number of predictors being used for prediction was 42, while regression tree model used 4 important predictors to get to the MSE of 0.7855.

The regression tree model yielded good results with just the 4 most important predictors required for analysis. Below are the advantages of finalizing this model:

i. Simplistic: We noted that its MSE score is almost as low as the MSE score obtained by random forest algorithm, and in fact better than the boosting technique. The model requires 4 predictors which makes the model and results easily interpretable by the non-technical Airbnb professionals and Airbnb hosts. The graphical representation is user-friendly and easy to read.



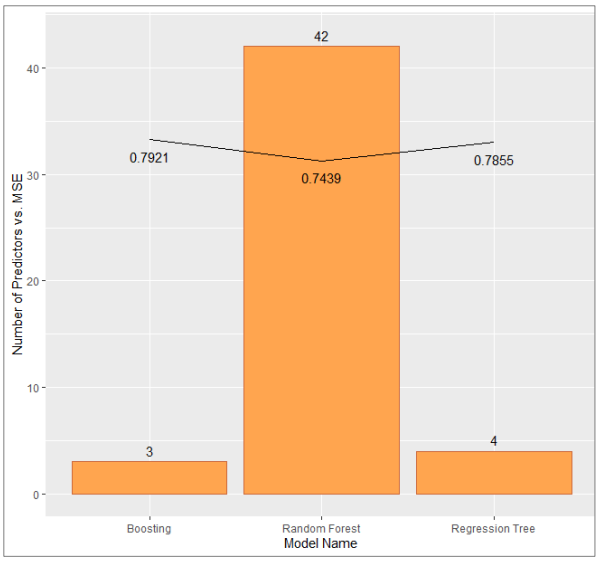
Interpreting the above graph, we note that:

- The 'host type' is the most important predictor influencing Airbnb ratings. If the host is a super host, the prediction for mean ratings is as high as 9.6.
- Among listings where the host is not a super host, the most important predictor becomes the 'host total listings count'. If the listings count by a single host is 4 or more, it reduces the ratings to the lowest mean of 8.8.
- The number of reviews becomes the next important predictor when the host listings count is less than 4. According to the model, the mean prediction for

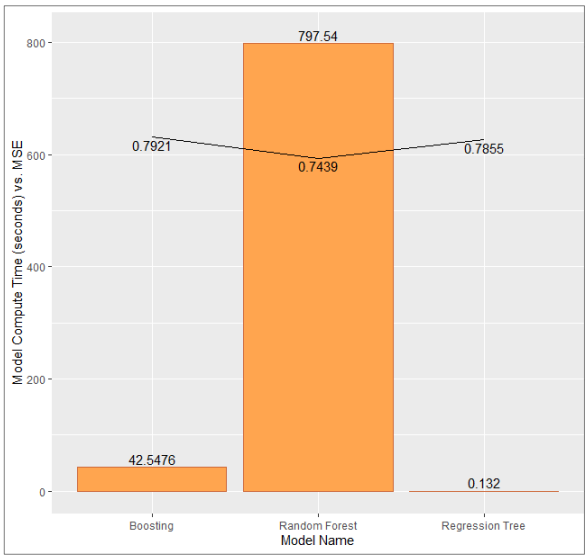
listings with 21 or more reviews, with listings count more than 4 and the host not being a super host is 9.

- Lastly, the fourth predictor is the 'host response rate'. Listings where the hosts tend to have a response rate of more than 0.91 end up with the mean ratings of 9.3. 38% of the observations from our dataset lie within this prediction.

ii. Least expensive computationally: While considering the computing power and time required for making predictions over the test dataset, the regression tree model is the clear winner, because it produced results approximately 400 to 6000x faster than its competition models. The MSE value was similar to that of random forest while saving the processing cost for predicting the ratings.



Number of Predictors (bars) vs. MSE (line) for each model



Model Compute Time seconds (bars) vs. MSE (line) for each model

We plotted the number of predictors vs. MSE values for each of the models and concluded that the regression tree model follows a simplistic approach, resulting in a similar MSE as other models while only taking 4 predictors into account. Likewise, we also plotted the model compute time in seconds for each model and concluded that the regression tree model takes considerably lower amount of compute time compared to boosting and random forest models while producing similar MSE values.

5. Predictions and Conclusions

From the best prediction model chosen, we can conclude that the rating of a listing majorly depends on factors that the attributes related to the host are more important than the property and the provided services. From the study, 3 out of the 4 most important predictors are directly related to the host's characteristics.

These insights would be helpful for Airbnb in guiding their registered hosts to provide better services. Airbnb can also utilize this study in redefining their criteria for approving the hosts' properties and recommend the hosts to make quick fixes to enhance overall customer satisfaction and experience.

- The hosts should respond to guests' requests and messages promptly, since low response rates can significantly result in lower ratings.
- Hosts should have enough resources if they are managing multiple properties. According to the model results, the number of hosts listings affected the ratings negatively.
- Hosts should focus on obtaining high number of reviews for their services. This is a cyclical process, since the better the service, higher would be the footfall, and the number of reviews.
- Other factors like the accommodation capacity, price, and the number of bathrooms, also affect the ratings significantly. There is no direct relation (like higher the price, lower the ratings) because high-priced bungalows are likely to get better ratings. But the hosts should consider these attributes carefully while adding property listings to Airbnb.

- Availability of amenities including microwave, dryer, hot water, cable TV, garden, and wireless internet are also important factors for achieving good ratings.

The model is effective in predicting Airbnb ratings, and it also comes with the benefit of being a relatively simpler model. The hosts can be enabled to simply plug in the information about their property and get quick results about the potential customer ratings. For example, in the final model, if the host is a super host, has no other listings added on Airbnb, has a count of 12 reviews on her property, and has a response rate of 1, the model would output a 9.6 rating for the property. The model can be quickly upgraded using the research and model development performed in this study, thus enabling addition of more complex factors as needed. We also performed additional research on the Airbnb business model *(please refer **Appendix 5** for details).*

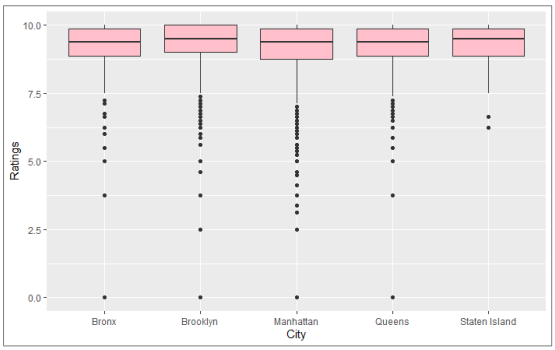
6. Appendix

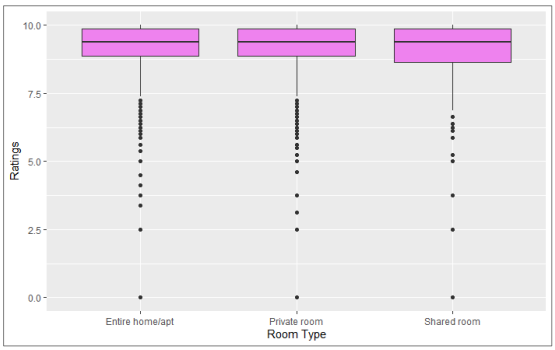
Appendix 1: Number of outliers in numerical variables

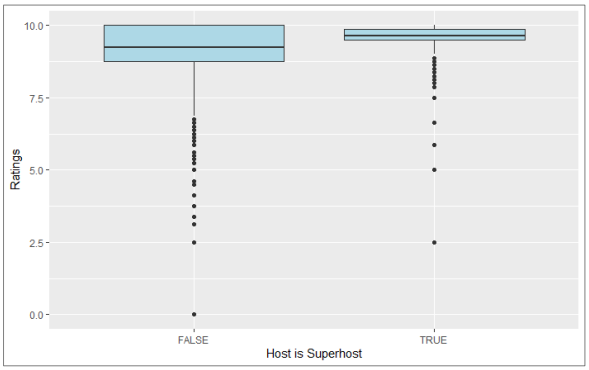
Variable	Threshold	No. of outliers removed
Host total listings count	count > 100	32 observations
Accommodates	value > 10	137 observations
Bathrooms	value > 4	20 observations
Bedrooms	value > 5	24 observations
Price	value > 900	12 observations
Minimum nights	value > 30	49 observations
Number of reviews	count > 270	41 observations
Reviews per month	count > 15	10 observations

Appendix 2: Spread of categorical variables

We note that spread of the ratings of cities and room types is similar for all categories. On average, the average ratings for Bronx and Manhattan are slightly lower than the other cities. For room type, the average rating for shared room is slightly lower than the average ratings of entire place or private room. For host type, we observed that the super hosts are generally highly rated and have a very thin spread. For non super hosts, the spread of ratings is considerably higher, and their overall average ratings are lower than ratings of the super hosts.



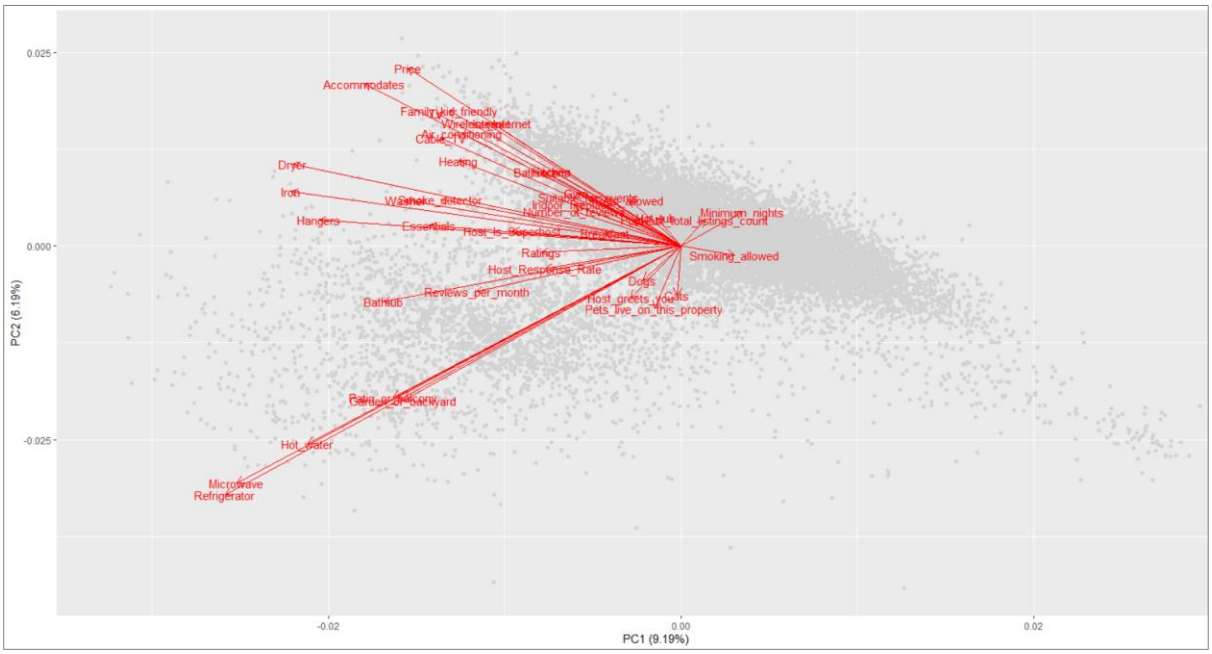




Appendix 3: Variable Importance by % increase in MSE values

predictor	PercentIncMSE	IncNodePurity
Host_Is_Superhost	34.510217	549.9297448
Host_total_listings_count	34.0268883	626.4724798
Accommodates	32.8825235	647.6273267
Price	24.194082	1506.894202
Reviews_per_month	18.413789	1622.256288
Room_type	16.0744869	274.0231313
Washer	15.0015657	177.0142688
Hangers	12.9211108	196.2871477
Property_type	12.6576675	588.0720705
Bathrooms	12.3742041	210.7706454
Dryer	12.0996029	161.6562443
Microwave	11.4420952	85.11624668
TV	10.1387807	205.3110919
Cable_TV	10.0735607	159.5041993
City	10.0297792	472.5983193
Family_kid_friendly	9.9444788	195.8549004
Minimum_nights	9.7919074	672.5672693
Host_Response_Rate	9.7590539	668.881312
Refrigerator	7.9784658	86.14904819
Hot_water	7.226719	105.3453725
Kitchen	6.6182755	119.4655065
Garden_or_backyard	6.3283099	56.40585968
Essentials	5.5708188	149.2624643
Internet	5.5403123	66.23004662
Bathtub	5.5198845	90.66124937
Cats	5.3944836	67.52666332
Smoke_detector	5.1970745	182.4468594
Suitable_for_events	4.6974032	126.8620548
Iron	4.6897761	186.8213893
Pets_live_on_this_property	4.6148187	89.67132137
Gym	4.5439941	102.8695559
Wireless_Internet	4.5130352	94.29709344
Pets_allowed	3.9627716	165.454192
Indoor_fireplace	3.2553989	81.51054501
Patio_or_balcony	3.1620468	51.77031434
Breakfast	3.1302456	145.7801308
Dogs	2.7930005	50.25032599
Number_of_reviews	1.6671864	1538.539126
Smoking_allowed	1.6437388	110.907405
Air_conditioning	0.9561362	153.4324257
Hot_tub	-1.6297316	89.72668247
Pool	-2.7286334	44.67574345
Host_greets_you	-3.5868428	97.1617712
Heating	-4.0304345	186.271463

Appendix 4: PCA Results after adding ratings with all the available numerical predictors



Along with the PCA analysis performed with the 6 important predictors, we also analyzed PCA for all the numeric variables. By plotting the PCA plot, it was identified that since the number of predictors were significantly high, the principal components 1 and 2 could explain very less variability of 9% and 6% respectively. It was established that this PCA was not completely reliable as it is not representative of all the numeric predictors. However, these results are in-line with our interpretations from the model, that is, important predictors like host response rate, reviews per month, minimum nights count, availability of bathtub, are directly affecting the property ratings.

Appendix 5: Additional Research and Motivation for Future Analysis

As we were researching about the Airbnb business model, we also read through the research work done by Prof. Maxime Cohen for analyzing the effects of Airbnb rentals on the area residential development. We brainstormed about the potential extensions to our model for integrating the effects of high vs. low Airbnb ratings on the area development. Reading through the research paper, we noted that the difference-in-difference method implemented for his study could also be helpful for the capstone project for one of us.