UNIVERSITY OF MARYLAND

ROBERT H. SMITH SCHOOL OF BUSINESS

BUDT 758T Final Project Report Spring 2023

Section 1: Team member names and contributions

Name	Contributions
Akhil Aenugu	Data Cleaning, Linear Regression, Feature engineering, Fitting Curve, Sentiment Analysis, Error debug, Business Understanding, Report development
Yu Hsiang Cheng	Data cleaning, Feature Engineering for text and non-text variables, Searching for External data, Tuning Linear Regression, Logistic Regression, Tree ,Ridge/Lasso and Xgboost Model, Plotting fitting curve for tree and Ridge/Lasso Models. Finalizing the final report
Sai Navadeep Reddy Ramireddygari	Feature Engineering, Ridge and Lasso for logistic model, Relationship graphs, Hyperparameter tuning, Learning Curve and inference, Report preparation
Praharsh Deep Singh	Data cleaning, Feature Engineering for the text variables by using word embeddings, and k-means clustering, Tuning hyperparameters for Xgboost, Optimizing the winning model (Xgboost Model) , Searching for External data, Making learning curve and fitting curve for Xgboost.
Udit Singh	Data cleaning, Feature Engineering for text and non-text variables, Researched external datasets, Turning random forest models. Analysis insights form different features and generate plots. Evaluation of model performance using cross validation.

Section 2: Business Understanding

Airbnb dataset is a humongous amount of information available on properties in the market, for rental purposes. The data consists of information on hosts, locality, amenities and many more aspects of the property. Study and analysis of the data provides insights, performance and future predictions about the properties in the market. Adopting machine learning algorithms like Boosting can drive business decisions to increase revenue and reach in the airbnb market.

By adopting Boosting technique, we predict whether an airbnb property will bag a 100% perfect rating score or not. In addition to supporting the airbnb platform to make significant value-increasing decisions, the analysis covers a wide range of clients such as property owners, real estate management firms and the hospitality industry to improvise and expand their business and services. The study will support the legal authorities to understand trends in the rental sector of the real-estate market.

The report built out of data mining and predictive analytics can be a base for upcoming new entrants in the business such as airbnb. The firm can start building their research and expand analysis on the report in order to cater the business requirements of their unique objective.

Tourism is one of the major revenue generating industries. Laying out a detailed analysis on the given airbnb housing will set to the rise of new businesses in the locality and increase popularity over due course of time. For instance, locations which see high inflow of renters can open up a window for restaurant chains set up in the locality.

The analysis adds value to actors in the Airbnb system in multiple dimensions:

1. Services and amenities: Amenities is one of the major supporting factors in increasing reviews and churning more customers. It has shown a lot of advantage in deriving a perfect rating score. This would enable firms to analyze on kind for services to provide in each type of property based upon requirement.
2. Investment and Revenue: The model gives understanding on properties historical performance and recommends on the factors that gave a good model. These features such as location and different property categories help our decision making process in choosing the right property at popular locations based upon different kinds of customers' preferences.
3. Demand Assessment: When we estimate that properties are getting good rating scores, it shows that the type of property, location and other several factors are leading to a good model. The team of airbnb can reach out to more properties to attend to the demand in future.

4. Loss potential properties: In addition to making a perfect rating score to give positive results in future, the analysis paves way to rectify the factors leading to non-perfect rating scores. Addressing the issues in factors might lead to perfect scores for the property.
5. Marketing Campaigns: To increase customer retention and more churn, release property discounts on highly preferred properties to targeted customers.

Section 3: Data Understanding and Data Preparation

1) Table of variables used:

ID	Feature Name	Brief Description	File Name	R Code Line Numbers
1	price	Original feature from dataset	Group23_Log	39-40
2	cleaning_fee	Original feature from dataset	Group23_Log	58-59
3	bedrooms	Original feature from dataset	Group23_Log	44-45
4	beds	Original feature from dataset	Group23_Log	47-48
5	host_total_listings_count	Original feature from dataset	Group23_Log	50-51
6	price_per_person	price/accommodates	Group23_Log	54-55
7	has_cleaning_fee	has_cleaning_fee is YES if there is a cleaning fee, and NO otherwise	Group23_Log	58-59
8	bed_category	bed_category is "bed" if the bed_type is Real Bed and "other" otherwise	Group23_Log	63-64
9	property_category	property_category has the following values: apartment:Apartment, Serviced apartment, Loft. hotel:Bed & Breakfast, Boutique hotel, Hostel. condo:Townhouse, Condominium. house:Bungalow, House. other:otherwise	Group23_Log	77-89
10	ppp_ind	ppp_ind is 1 if the price_per_person is greater than the median for the property_category, and 0 otherwise	Group23_Log	95-95
11	bathrooms	Original feature from dataset	Group23_Log	106-107
12	host_is_super	Original feature from dataset	Group23_Log	110-111

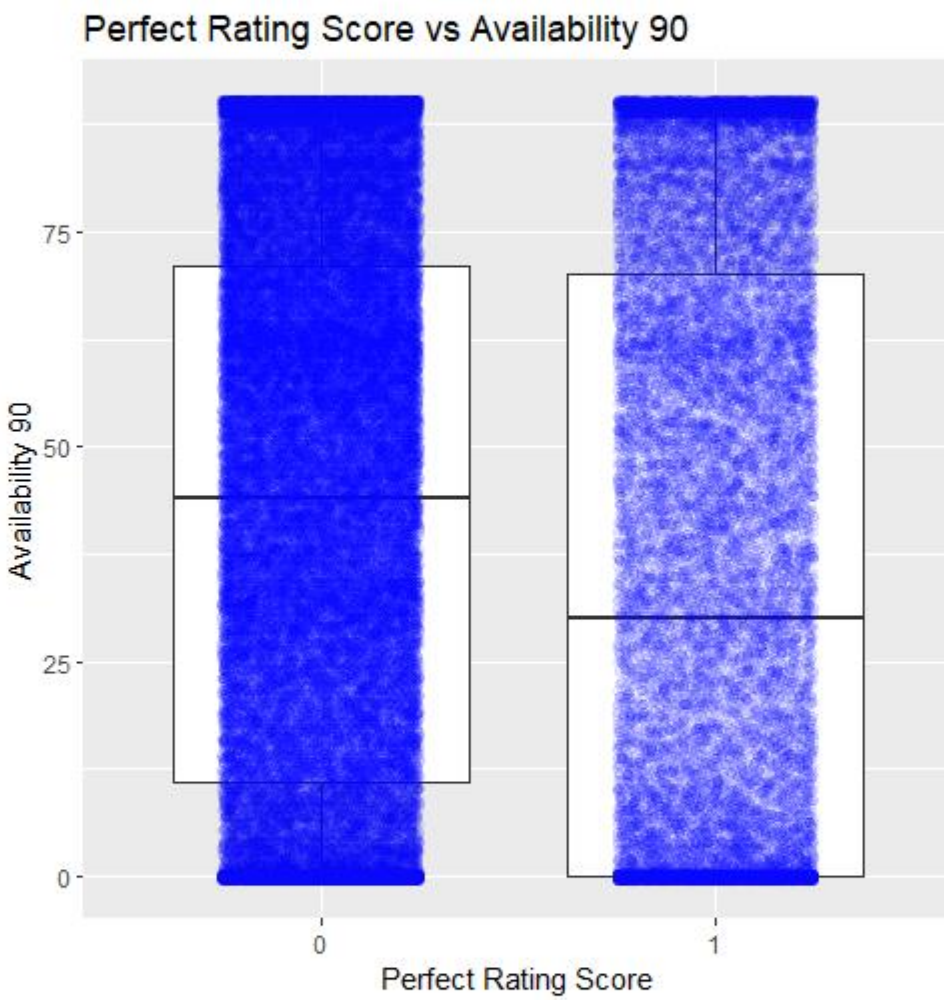
	host			
13	charges_for_extra	YES if extra_people > 0 and "NO" if extra_people is 0 or NA	Group23_Log	115-120
14	host_acceptance	ALL if host_acceptance_rate = 100%, "SOME" if host_acceptance_rate < 100%, and "MISSING"	Group23_Log	124-127
15	has_min_nights	YES" if minimum_nights > 1, and "NO" otherwise	Group23_Log	138-141
16	market	Original feature from dataset	Group23_Log	145-149
17	list_till_today	system date-host_since	Group23_Log	153-157
18	last_review_till_today	system date-first_review	Group23_Log	153-157
19	has_detector	if amenities contains the key word detector	Group23_Log	159-225
20	has_Internet	if amenities contains the keyword Internet	Group23_Log	159-225
21	has_TV	if amenities contains the keyword TV	Group23_Log	159-225
22	has_friendly	if amenities contains the keyword friendly	Group23_Log	159-225
23	has_Heating	if amenities contains the keyword Heating	Group23_Log	159-225
24	has_Kitchen	if amenities contains the keyword Kitchen	Group23_Log	159-225
25	has_Essentials	if amenities contains the keyword Essentials	Group23_Log	159-225
26	has_Smoke	if amenities contains the keyword Smoke	Group23_Log	159-225
27	has_Air	if amenities contains the key word Air	Group23_Log	159-225
28	has_Shampoo	if amenities contains the keyword Shampoo	Group23_Log	159-225
29	has_conditioning	if amenities contains the keyword conditioning	Group23_Log	159-225
30	has_Hangers	if amenities contains the keyword Hangers	Group23_Log	159-225
31	has_Carbon	if amenities contains the keyword Carbon	Group23_Log	159-225
32	has_Dryer	if amenities contains the keyword Dryer	Group23_Log	159-225
33	has_Washer	if amenities contains the keyword Washer	Group23_Log	159-225
34	has_Hair	if amenities contains the keyword Hair	Group23_Log	159-225
35	has_monoxide	if amenities contains the keyword monoxide	Group23_Log	159-225
36	has_Laptop	if amenities contains the keyword Laptop	Group23_Log	159-225
37	has_Iron	if amenities contains the keyword Iron	Group23_Log	159-225

38	has_dryer	if amenities contains the keyword Dryer	Group23_Log	159-225
39	rule_sent_score	Sentiment analysis for Rule column P for positive and N for negative	Group23_Log	229-278
40	host_response_time	Original feature from dataset	Group23_Log	282-283
41	instant_bookable	Original feature from dataset	Group23_Log	286
42	population	Population of the city - external dataset - source - [US Zip Codes Points- United States of America — Data hub (opendatasoft.com)]	Group23_xgboost	51-59
43	density	Population density - external dataset - source - [US Zip Codes Points- United States of America — Data hub (opendatasoft.com)]	Group23_xgboost	51-59
44	popular_destination	If the city is in the list of most popular destinations in the US amongst international travelers - source - https://www.bts.gov/archive/publications/state_transportation_statistics/summary/table_04_19	Group23_xgboost	60-65
45	popular_destination_yahoo	If the city is in the list of most popular destinations in the US amongst international travelers - source - https://finance.yahoo.com/news/30-most-visited-cities-american-114422642.html	Group23_xgboost	60-65
46	zipcode	Original feature from dataset	Group23_xgboost	43-49,172
47	monthly_price	If the original dataset has a value for monthly price - factor - binary	Group23_xgboost	178-188
48	weekly_price	If the original dataset has a value for weekly price - factor - binary	Group23_xgboost	178-188
49	security_deposit	Original feature from dataset	Group23_xgboost	178-188
50	guests_included	Original feature from dataset	Group23_xgboost	372
51	is_business_t	Original feature from dataset	Group23_xgboost	178-188

	ravel_ready			
52	minimum_nights	Original feature from dataset	Group23_xgboost	129-143
53	room_type	Original feature from dataset	Group23_xgboost	87-115
54	license	If the license field has a value or not - 0/1	Group23_xgboost	178-188
55	is_location_exact	Original feature from dataset	Group23_xgboost	
56	house_rules	Original feature from dataset	Group23_xgboost	155-162
57	neighborhood_overview	Original feature from dataset	Group23_xgboost	155-162
58	income_families	Mean income of families in the city - external dataset - source - S1901: INCOME IN THE PAST 12 MONTHS ... - Census Bureau Table	Group23_xgboost	166-171
59	income_non_families	Mean income of non family households in the city - external dataset - source - S1901: INCOME IN THE PAST 12 MONTHS ... - Census Bureau Table	Group23_xgboost	166-171

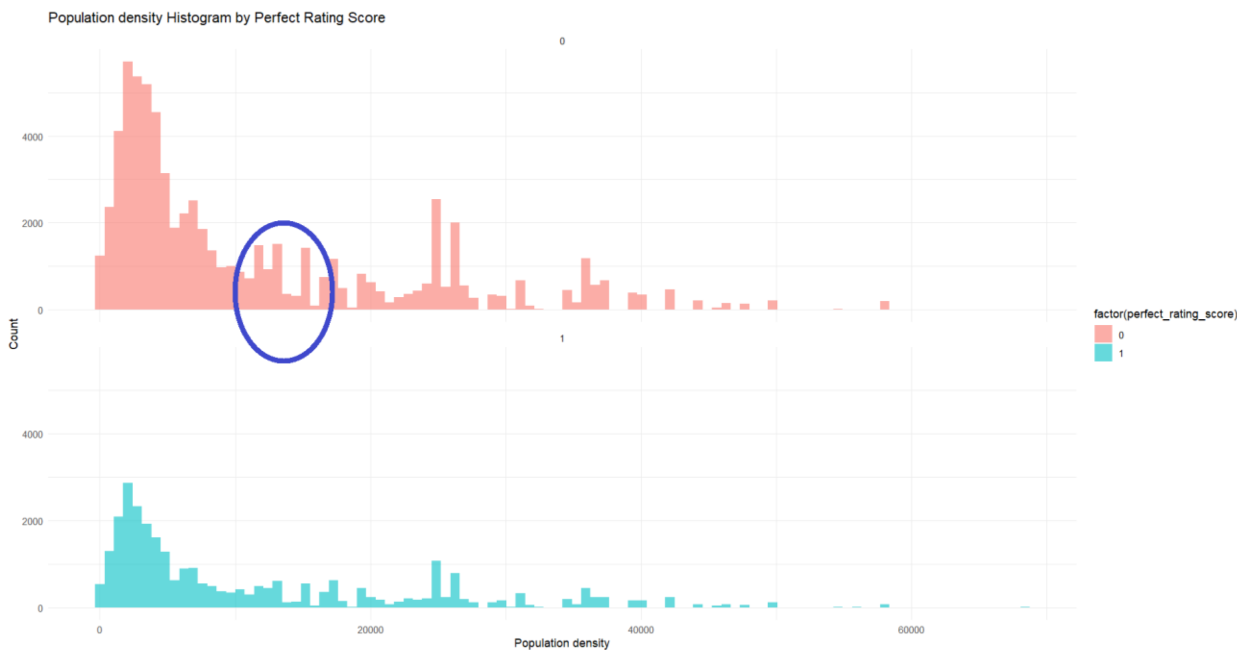
2) Graphs or tables demonstrating useful or interesting insights regarding features in the dataset.

1. Availability_90
   The boxplot between availability and target variable shows that:
   a)properties with a perfect rating score are more likely to have a low value for availability_90
   b) while properties with perfect rating score value are distributed evenly throughout the range, properties with a perfect rating score are more likely to have the availability value of 0 or 90.



2. Price:
   Scatterplot between price and target variable shows that higher priced airbnbs between 3000$ and 7500$ are more likely to have a perfect rating score.

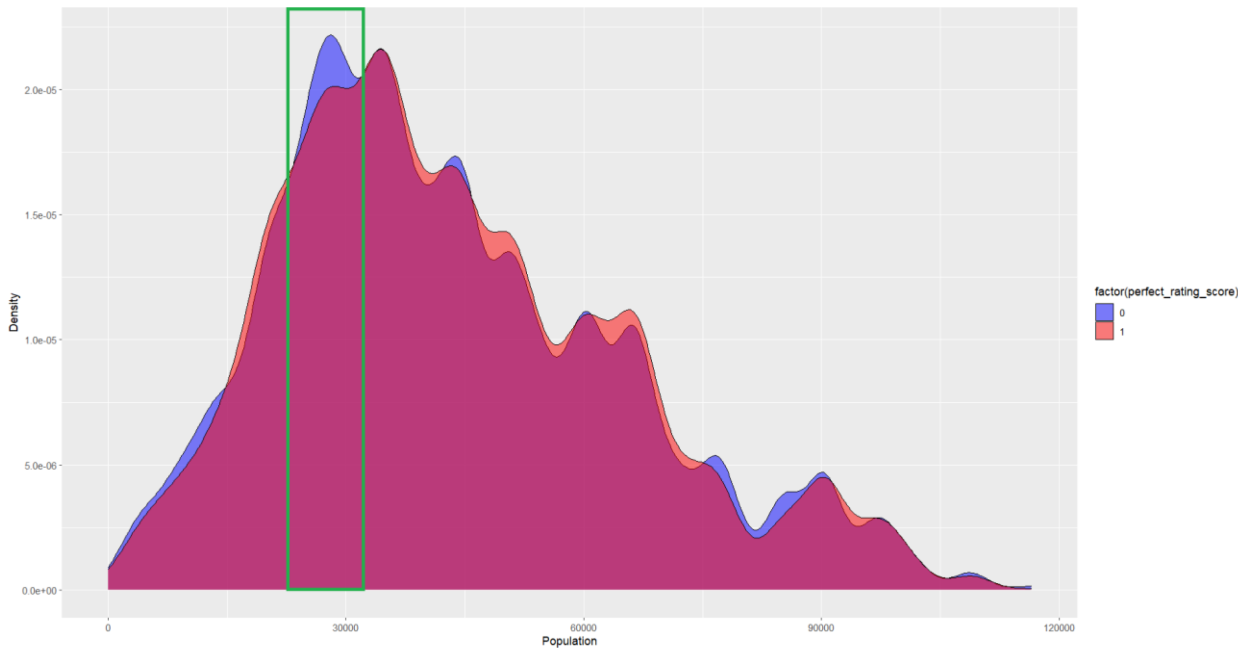
3. Population density:
   A slight peak in the count of properties that do not have a perfect rating score for the cities with population density between 2000 and 3000 when compared to the properties with a perfect rating score in the same range suggests that there might be higher chances of properties not having a perfect rating score for these cities (a peak in both categories suggests that there are a large number of cities with that particular density, but a peak in only one suggests there might be a relationship between that population density group and zero perfect rating score).



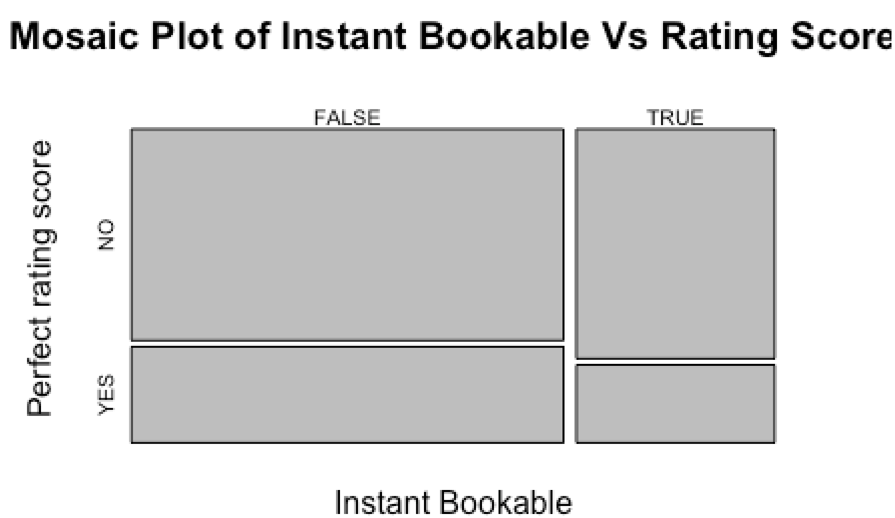
4. Zip Code :
   The table below shows that 40% of the zipcodes have a mode perfect rating score of , which means that majority of the properties in these areas have a perfect rating score, these zipcodes can be grouped together and regression analysis can be performed to gain further insight.

	mode perfect_rating_score = 1	mode perfect_rating_score = 0
Count of zipcodes	28159	71822

5. Population :
   The density plots of population show that it has no major effect on the perfect rating score but the enclosed region suggests that properties in cities with their population in this region tend to have more number of properties without a perfect rating score hence it is worth exploring with a regression model analysis.

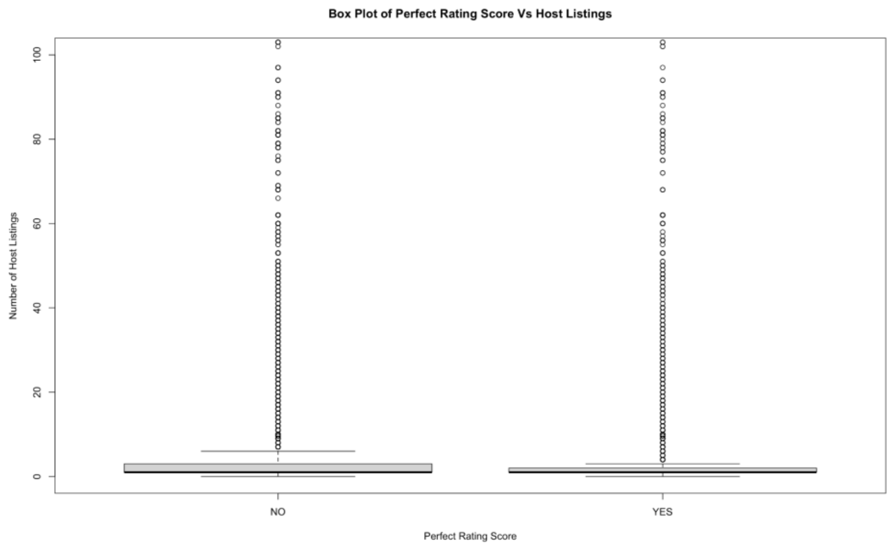


6. Instant Bookable



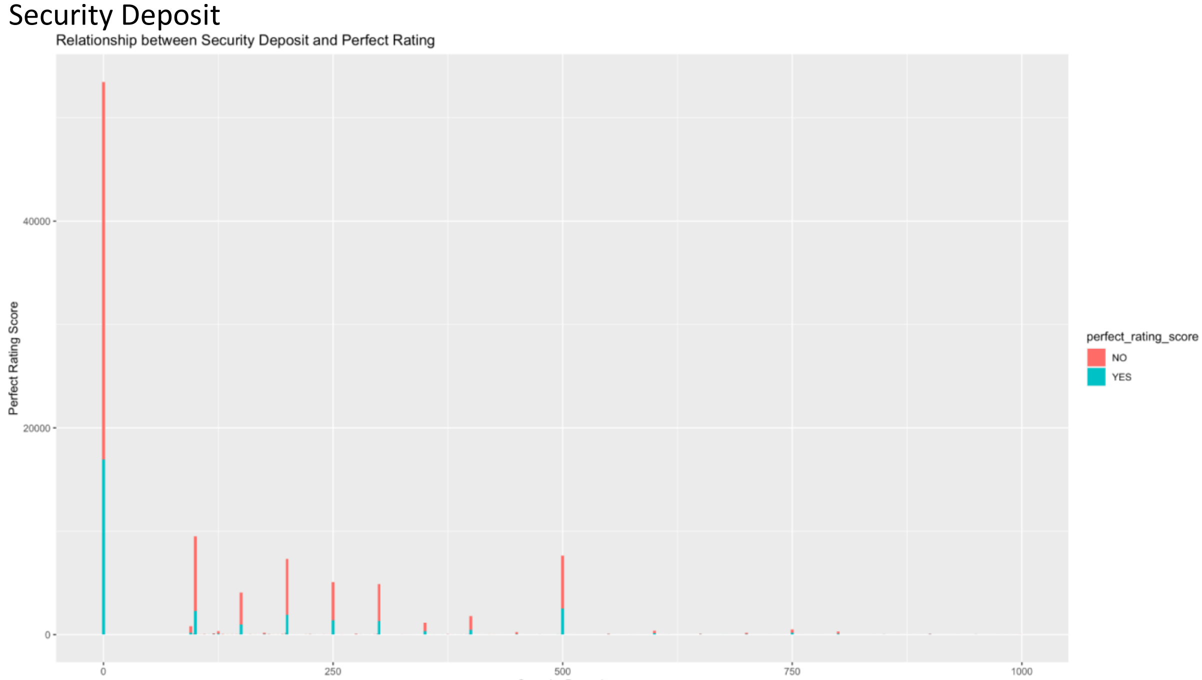
The variables have sufficient relation for study. Perfect rating predictions are the majority in FPs. In majority cases, non-instant bookable leads to non-perfect rating score.

7. Host total listings



The relation shows that having less number of listings helped in predicting the perfect rating score more accurately. Having a large number of listings shows as outliers.

8. Security Deposit



We can infer that having a decent amount of security deposit is better than having a high amount. Also, having a security deposit tends to make property non-preferred compared to having security.

9. Accommodates

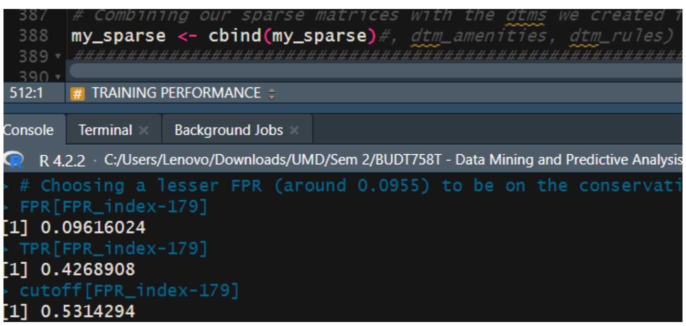
Accommodates show a mixed relationship to Perfect Rating score. But, there are more properties having 'NO' near 2-3 accommodates feature, evident with the wide strings of violin. This relationship is quite debatable as the data points are heavily located around 0-5 accommodates for both categories.

10. Availability 30

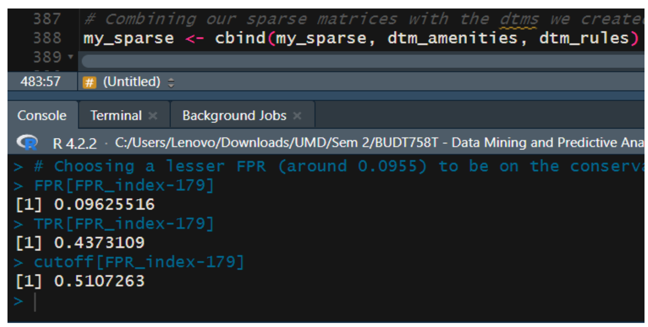
Properties that don't have a perfect rating score are more likely to be in Availability 30. Availability 30 is a factor that is not much important to rate a property to be perfect.

3) Preparing text columns (additional comments)

- While creating features from text columns like interaction, access, and description, we initially constructed a vocabulary comprising these words. Subsequently, we pruned the vocabulary and generated a document-term matrix (DTM) and a term-co-occurrence matrix (TCM). The TCM was utilized to obtain word embeddings for each word, resulting in 50 embeddings per word. To obtain weighted final embeddings for each observation, we computed a weighted sum across the DTM. These weighted embeddings were then used as input for the k-means clustering algorithm, enabling us to determine the centers associated with each observation. Our hypothesis suggests that these centers capture the sentiment linked to each observation.
- We also used text cleaning for the text columns of amenities, and house_rules and made DTM in a similar way. Our TPR increased from 0.4268908 to 0.4373109. Hence we concluded that these features are helpful in prediction. Here are the screenshots for reference:



TPR value without text features



TPR value with text features

Section 4: Evaluation and Modeling

1) **Top performing Winning Model**

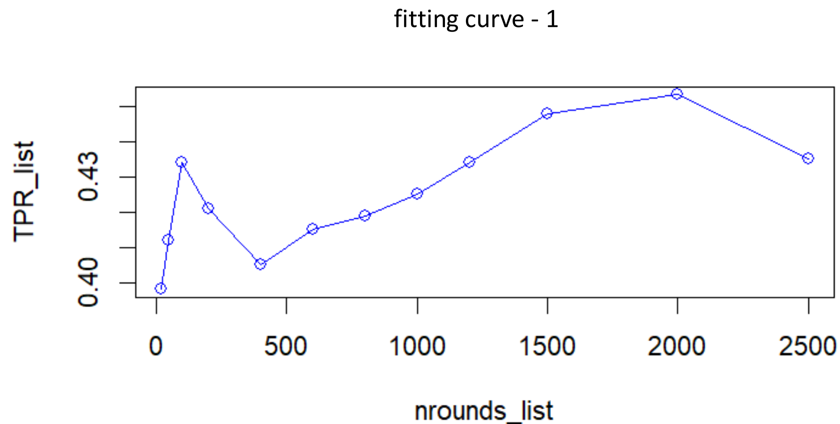
The xgboost() was our winning model. Our estimated Generalization Performance was of TPR = 0.453333 subject to FPR - 0.09549575, and Estimated Training Performance: TPR - 1 subject to FPR - 0.09999596. This was the highest TPR that we achieved after trying 6 different kinds of models. We used a simple train/validation split: 30% of data was reserved to validate the training results. We used complete 112186 observations of training_data to split into training and validation sets.

We performed the predictions in the R-Code line numbers: 538 - 537. The file name is: “Group23_xgboost”.

Hyperparameters that we tried include:

max.depth = (8, 10, 15, 20, 25, 30)
eta = (0.01, 0.05, 0.1, 0.5, 1.0)
nrounds <- (100, 500, 1000, 1500, 2000, 2500)

The best values of TPR subject to a FPR of 0.1 that we got for the hyperparameters are max.depth = 15, eta = 0.05, and nrounds = 200. Here's a fitting curve of nrounds as the hyperparameter in the xgboost()

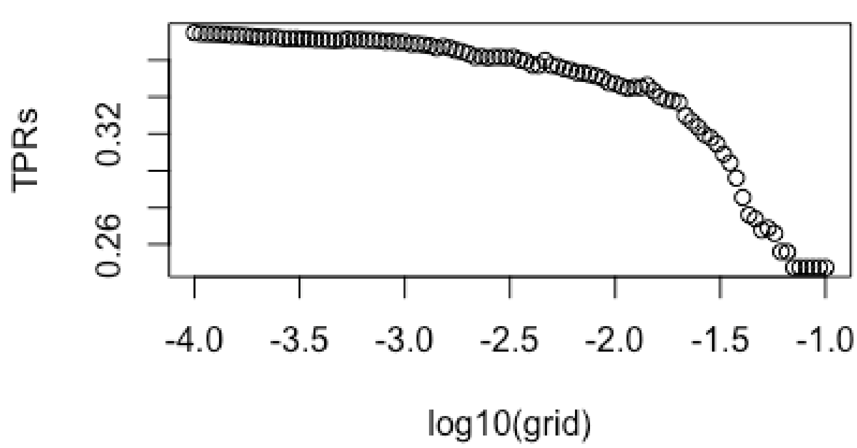


2) Modeling:

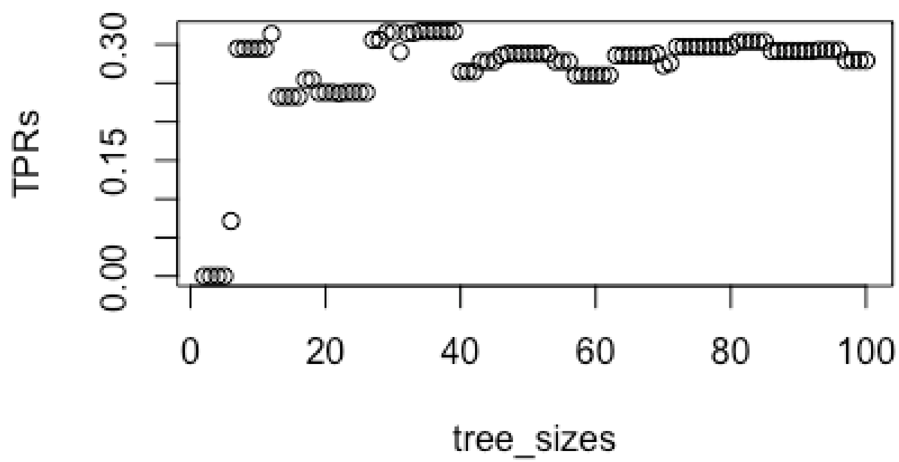
Model 1	Logistic Regression
R Function / Library	glm() function
Training & Generalization Performance	Estimated TPR with the FPR closest to 0.1 on Training Set: TPR: 0.3829798 FPR: 0.1000101 Estimated TPR with the FPR closest to 0.1 on Validation Set: TPR: 0.3723812 FPR: 0.1000189
Methodology	10-fold cross validation was used to evaluate the model performance and we used a simple train/validation split by using 30% for validation and 70% for training.
Best-performing features	accommodates,bedrooms,cancellation_policy,has_cleaning_fee,host_total_listings_count,price,ppp_ind,property_category,bathrooms,charges_for_extra,host_acceptance,has_min_nights,market,host_is_superhost,room_type,list_till_today,last_review_till_today,rule_sent_score,host_response_time,instant_bookable,has_Internet,has_TV,has_friendly ,has_Heating ,has_Kitchen ,has_Essentials,has_Smoke ,has_Air ,has_Shampoo

	,has_conditioning ,has_Hangers ,has_Carbon , has_Dryer ,has_Washer ,has_Hair ,has_monoxide,has_Laptop ,has_Iron,has_dryer
R Code Lines	File name: Group23_Log Lines: #352 - #399
Hyperparameter & Fitting curve	Not Applicable

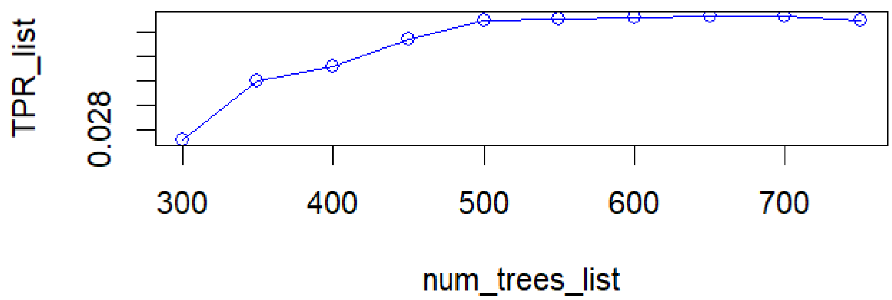
Model 2	Linear Regression
R Function / Library	lm() function
Training & Generalization Performance	Estimated TPR with the FPR closest to 0.1 on Training Set: TPR: 0.3535706 FPR: 0.1000189 Estimated TPR with the FPR closest to 0.1 on Validation Set: TPR: 0.359965 FPR: 0.1000101
Methodology	We used a simple train/validation split by using 30% for validation and 70% for training
Best-performing features	accommodates,bedrooms,cancellation_policy,has_cleaning_fee,host_total_listings_count,price,ppp_ind,property_category,bathrooms,charges_for_extra,host_acceptance,has_min_nights,market,host_is_superhost,room_type,list_till_today,last_review_till_today,rule_sent_score,host_response_time,instant_bookable,has_Internet,has_TV,has_friendly ,has_Heating ,has_Kitchen ,has_Essentials,has_Smoke ,has_Air ,has_Shampoo ,has_conditioning ,has_Hangers ,has_Carbon , has_Dryer ,has_Washer ,has_Hair ,has_monoxide,has_Laptop ,has_Iron,has_dryer
R Code Lines	File name:Group23_Linear Lines:#351 - #398
Hyperparameter & Fitting curve	Not Applicable

Model 3	Ridge and Lasso
R Function / Library	Library: library(glmnet) R functions: glmnet()
Training & Generalization Performance	Estimated TPR with the FPR closest to 0.1 on Training Set: TPR: 0.3858856 FPR: 0.1000041 Estimated TPR with the FPR closest to 0.1 on Validation Set: TPR: 0.373483 FPR: 0.09999525
Methodology	We used a simple train/validation split by using 30% for validation and 70% for training
Best-performing features	accommodates,bedrooms,cancellation_policy,has_cleaning_fee,host_total_listings_count,price,ppp_ind,property_category,bathrooms,charges_for_extra,host_acceptance,has_min_nights,market,host_is_superhost,room_type,list_till_today,last_review_till_today,rule_sent_score,host_response_time,instant_bookable,has_Internet,has_TV,has_friendly ,has_Heating ,has_Kitchen ,has_Essentials,has_Smoke ,has_Air ,has_Shampoo ,has_conditioning ,has_Hangers ,has_Carbon , has_Dryer ,has_Washer ,has_Hair ,has_monoxide,has_Laptop ,has_Iron,has_dryer
R Code Lines	File name: Group23_ridge_lasso Lines: #351-474
Hyperparameter	Lambda: A list of grids for different lambda were tested: grid <- 10^seq(-1,-4,length=100)
Fitting curve	

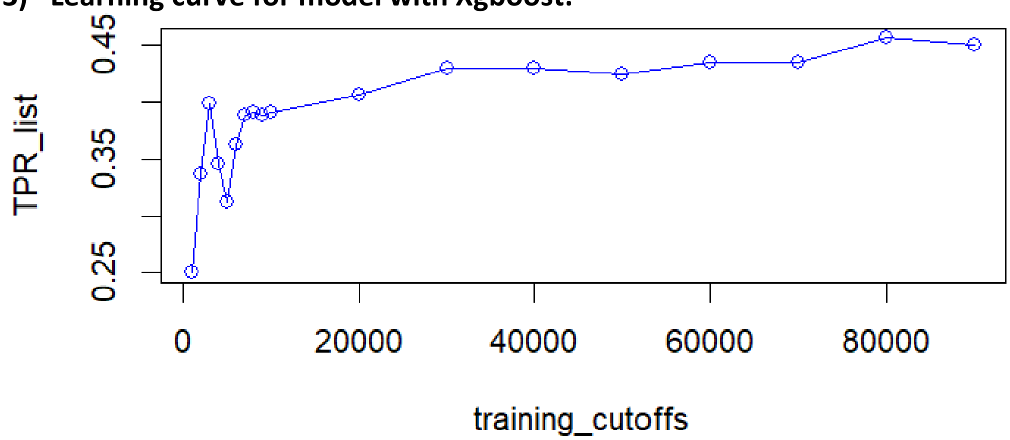
Model 4	Trees
R Function / Library	Library: library(tree) R functions: tree.control() tree() prune.tree()
Training & Generalization Performance	Estimated TPR with the FPR closest to 0.1 on Training Set: TPR: 0.5244957 FPR: 0.1047387 Estimated TPR with the FPR closest to 0.1 on Validation Set: TPR: 0.2925408 FPR: 0.1060252
Methodology	We used a simple train/validation split by using 30% for validation and 70% for training
Best-performing features	accommodates,bedrooms,cancellation_policy,has_cleaning_fee,host_total_listings_count,price,ppp_ind,property_category,bathrooms,charges_for_extra,host_acceptance,has_min_nights,market,host_is_superhost,room_type,list_till_today,last_review_till_today,rule_sent_score,host_response_time,instant_bookable,has_Internet,has_TV,has_friendly ,has_Heating ,has_Kitchen ,has_Essentials,has_Smoke ,has_Air ,has_Shampoo ,has_conditioning ,has_Hangers ,has_Carbon , has_Dryer ,has_Washer ,has_Hair ,has_monoxide,has_Laptop ,has_Iron,has_dryer
R Code Lines	File name: Group23_Tree Lines: #354 - #425
Hyperparameters	We tuned different sizes of trees by using prune.tree(full_tree, best = 100) Tree size = (10, 20, 30, 40, 50, 100) TPR: (0.3216783, 0.3240093, 0.3379953, 0.2645688, 0.2505828, 0.2925408) FPR: (0.1163008, 0.1195703, 0.1331154, 0.1060252, 0.09341429, 0.1060252)

Fitting curve	

Model 5	Random Forest
R Function / Library	ranger() function, and the library(ranger)
Training & Generalization Performance	Estimated TPR with the FPR closest to 0.1 on Training Set: TPR: 0.062449 FPR: 0.09999596 Estimated TPR with the FPR closest to 0.1 on Validation Set: TPR: 0.03697479 FPR: 0.09193602
Methodology	We used a simple train/validation split by using 30% for validation and 70% for training.
Best-performing features	accommodates,bedrooms,cancellation_policy,has_cleaning_fee,host_total_listings_count,price,ppp_ind,property_category,bathrooms,charges_for_extra,host_acceptance,has_min_nights,market,host_is_superhost,room_type,list_till_today,last_review_till_today,cluster_interaction,cluster_access,cluster_descrip,population,availability_30,availability_60, availability_90, zip code,monthly_price, weekly_price, security_deposit, guests_included,instant_bookable,is_business_travel_ready,minimum_nights,room_type,is_location_exact,license,income_families,income_non_families,density,popular_destination, popular_destination_yahoo
R Code Lines	File name: Group23_ranger Lines: #490-511
Hyperparameter	mtry values that we tried 15, 20, 22, 25, 28, 30 num.trees values that we tried 300, 400, 500, 600, 700

	We got best performance for mtry = 22, and num.trees = 500
Fitting Curve	

3) Learning curve for model with Xgboost:



By observing the learning curve, we can deduce that the TPR shows significant improvement with a smaller amount of data. This implies that the model performs well even with limited training data, as there is minimal change in TPR with an increase in the amount of data. We utilized this insight while evaluating hyperparameters for the xgboost model. Instead of training the model (for tuning hyperparameters) on the entire dataset, which would be time-consuming, we used a sampled dataset consisting of 20,000 observations. This approach enabled us to efficiently optimize our model.

The learning curve graph indicates a tendency towards high bias, as the increase in the training set has minimal impact on TPR. Conversely, high variance occurs when the model overfits and becomes overly sensitive to the training data.

Section 5: Reflection/takeaways:

1) What did your group do well?

Our group did well on the below parts.

1. Internal Competitions and Performance: Our group actively participated in internal competitions and consistently submitted models with the best True Positive Rate (TPR) and lowest False Positive Rate (FPR). This demonstrates our team's dedication and ability to optimize our model's performance.
2. External data acquisition: Our group successfully sourced additional external data to enhance our analysis. By incorporating information such as population/density by zip code, income by zip code, and the popularity of specific locations among tourists, we expanded the scope of our analysis and potentially gained valuable insights.

These achievements highlight our commitment to improve performance, utilizing diverse data sources, and leveraging external information to enhance your project's analysis and results.

3. Extensive work on the winning model - the winning model was improved consistently by tuning hyperparameters and exploring different combinations of features and the TPR increased every time as a result.

2) What were the main challenges?

The main challenges were:

1. Job allocation: Assigning tasks to different team members proved to be a challenging aspect. Distributing work efficiently and ensuring that everyone is effectively contributing requires careful planning and coordination. Also, due to the sequential nature of the work, some tasks could not be divided parallelly.
2. Obtaining external data: Finding suitable external datasets that met our project requirements posed a significant challenge. The process of searching for relevant datasets involved extensive research and evaluation to ensure they aligned with our needs. Additionally, before merging different datasets, we had to invest effort in cleaning and preprocessing them to ensure data consistency and reliability.

Overcoming these challenges required effective communication, coordination, and dedicated efforts from the team to address task allocation and data acquisition obstacles.

3) What would your group have done differently if you could start the project over again?

1. We will enhance our internal communication to avoid duplicating tasks. While we currently rely on internal competitions to submit our results, we have observed instances where multiple team members independently worked on the same task. For instance, after meeting with the professor, we received suggestions on cleaning the training and test datasets. Ideally, we could have assigned one person to handle this task. However, we ended up with three different sets of code accomplishing the same function.

2. To improve our efficiency, we aim to explore different models earlier in the project. Although we only started using the xgboost model in our last three submissions, incorporating it from the first submission onwards will significantly enhance our performance compared to our final results. By diversifying our model selection early on, we can explore the potential of different algorithms and optimize our overall performance.
3. We could have improved the contributions of feature engineering by performing feature analysis at an earlier stage to utilize the insights gained from it during regression analysis.

4) What would you do if you had another few months to work on the project?

1. We will explore additional features by incorporating more textual data from the original dataset and seek out external data sources.
2. We plan to experiment with different machine learning models, including Support Vector Machines (SVM).
3. To streamline the process and optimize our results, we aim to develop a machine learning system that automates parameter tuning. This system will enable us to systematically try and test various parameter combinations, ultimately generating the highest True Positive Rate (TPR) while minimizing the False Positive Rate (FPR) based on different parameter settings.

5) What advice do you have for a group starting this project next year?

1. Have fun while exploring and experimenting with various concepts taught in class and prioritize this project. Despite having assignments, mid-term, and final exams, this project provides the best opportunity to gain practical experience in implementing the knowledge acquired in class.
2. Make regular visits to the professor. From our group's perspective, every time we meet the professor, we gain a lot of valuable information. Office hours serve as a convenient platform to resolve any problems we encounter.
3. Plan your work according to the grading rubric, start early and proceed according to the steps mentioned in the rubric.