Airbnb: Classifying Whether a Host is a SuperHost

Mrunmayi Bhide
School of Computing
Dublin City University
mrunmayi.bhide2@mail.dcu.ie
Student Number: 22263299

Omkar Gokhale
School of Computing
Dublin City University
omkar.gokhale2@mail.dcu.ie
Student Number:22266982

Sayali Randive
School of Computing
Dublin City University
sayali.randive2@mail.dcu.ie
Student Number:21265052

Abstract: With the ever increasing number of travelers across the globe, there is a need for an efficient and cost effective property rental system which provides best service at a reasonable cost. One such service is Airbnb, which offers both short-term and long-term stays to its customers at competitive prices and with a variety of accommodation alternatives. Trust factor between a provider and consumer is important for a successful business and Airbnb fosters this trust by displaying feedback from past clients, verifying hosts, and awarding the hosts with a super host badge [4]. In this paper we focus on building a prediction model to identify whether a host is a Super Host by analyzing various factors which have influence on the super host status. In this study we have analyzed the Airbnb Listings in 31 US cities which consists of detailed information about all the listings. We applied feature selection techniques and compared various machine learning models to classify the superhost status and Random Forest classifier proved to be more efficient in comparison to logistic regression, Decision Tree and XGBoost.

Keywords- Airbnb, Superhost, Listings

I. INTRODUCTION

Airbnb stands for Air bed and breakfast [2] and it has become the most popular service over the past few years. There are 6.6 million active listings and 4 million plus hosts on Airbnb worldwide as of 2022. These listings were from more than 220 plus different nations and regions [3]. It is a service which is driven by a peer-to-peer business model allowing the house/property owners to rent out their place to the travelers [1]. The hosts are able to enter their property details over the platform as a listing along with all necessary details such as the nightly price, description of the property, the amenities etc. The property is not limited to an entire room or a house but also includes shared rooms, hotels and apartments which are rented at a reasonable price. This is different from the traditional services provided by the hotels and vacation resorts where they own a place and rent it to people on a nightly basis. Airbnb is just a mediator between the host and the customers who receives a commission for every single booking made through their platform [2].

There are multiple factors which influence the decision of customers while selecting a listing such as the amenities, location, neighborhood, property reviews. In addition to these aspects, information about the host, such as super host status, may have an impact on this choice. According to Airbnb, a superhost is a host who offers top-notch amenities and provides a remarkable lodging experience [5].

Airbnb performs quarterly assessment for identifying the superhosts who satisfy the below mentioned requirements defined by Airbnb. They consider the data over the past 12 months for identification [6].

- Hosts having greater than 90% response rate
- Has at least 10 stays in the last year.
- Hosts with less than 1% cancellation rate
- Average rating received by the host is 4.8.

However, there might be other parameters which might act as the contributing factors in identification of a superhost. The main aim of this research is to identify all the features influencing the super host status and develop an optimal classification model for determining whether a host is a super host or not. Hosts with a superhost badge on their profile are recognised to be the hosts providing excellent accommodation facilities and are known to be more experienced than others. As a result, the superhosts will have advantages such as an increase in the number of bookings and visibility, both of which will enhance their revenue. Superhosts receive added benefits from Airbnb as well [7].
The next section gives an overview of the previous works which have been done in this area. Most of them deal with price prediction for the listings.

II. RELATED WORKS

There has been some research done in this area related to the Airbnb listings data, and most of it has been in the travel and hospitality industries [1]. There has been noteworthy research in the creation of machine learning based price prediction models, however a limited amount of research has been done in relation to the prediction of superhost.

The authors [2] discuss various machine learning models such as linear regression, logistic regression and random forest to predict the price of a listing. The dataset used contains 305,564 rows and 108 columns of data for various cities in the United States. It also includes the information about the past reviews as well as all the further bookings. The features such minimum_nights, maximum_nights, availability_365, number_of_reviews, review_scores_rating, review_scores_accuracy and accommodates were used as input to Random forest to predict the price of a listing and it proved to outperform the other models with an accuracy of 87%. This paper used a limited number of features to train the model and use of other features such as summary attributes and other features such as is_superhost which can be related to price prediction. Additionally, this paper does not provide the exact source of the dataset, thus the models cannot be reproduced.

The study [1] examines the impact of neighborhood and transport links on Airbnb postings using 13097 listings dataset from 40 US cities web scraped from the Airbnb website. They have used the multilevel mixed-effect linear regression models to determine the relationship between the two outcome variables price and listings intensity (which was the no. of listings per 10000 housing units) and the other variables such as Population density, Location and transportation for understanding the relationship between Transit frequency, Room characteristics, Super-host listings. The difference between both models was in the usage of independent features such as star rating for pricing model and reviews for intensity model. Advantage of use of Multilevel mixed-effects models is that it can handle multicollinearity and reduce the bias in the results. Also, the authors employed Importance Analysis (IPA) to determine the significance of various factors. The results indicate that the listing price decreased, the further it was from the city center, as did the number of listings. A higher listing price and more hosts in the neighborhood appeared to be related to efficient transport service and presence of local eateries and bars in the neighborhood.

There has been some research done on the effectiveness of the superhost status on the listings. In the study [4], the authors discuss the impact of the introduction of the Super Host program. Airbnb assesses its business model and implements various strategies in order to improve their service, attract more customers and in turn increase their profit. Throughout the years, Airbnb has made a lot of improvements, including modifications to its rating and review systems and the addition of the Super Host program. According to [4], Airbnb began recognizing exceptional hosts who offered superior service and awarding them a Super Host badge as part of the Super Host program. This was awarded based on certain factors decided by Airbnb such as the number of cancellations reviews, their response rate and number of stays. There has been some work done in the analysis of these four factors affecting the superhost status such as the Gunter's [9] examination of the relative importance of these four characteristics revealed that in San Francisco and the Bay Area, having a strong rating is a crucial factor. An analysis of the listings in four european cities was conducted in order to understand the impact on price of listings which showed that superhost's received around 4-9% of additional profit than the ordinary hosts.

This study [4] only discusses the impact of the superhost program and its effect on the price listings for 4 european cities only. So this cannot be generalized.

We discussed the research which has been conducted over the Airbnb listings data and as we can see the study related to superhost only deals with the four categories defined by Airbnb but they don't consider the impact other parameters have on the superhost status. In this paper we are focusing on analyzing the parameters affecting the superhost status and we intend to implement a model which identifies whether a host is also a super host.

III. METHODOLOGY

Data Cleaning:

We have used the Kaggle Airbnb Listings[1] data for 31 cities in the United States which includes all listings information from 6th June 2022 to 6th June 2023. This dataset includes information about the listings such as the amenities, property/room type,

[1] https://www.kaggle.com/datasets/konradb/inside-airbnb-usa

price, neighborhood, location, availability, past reviews of the listing (see Table 1), as well as host details such as host acceptance rate, super host status, license etc (see Table 2). The dataset has a total of *261,508* rows and *75* columns. The Fig. 1 below highlights the states for which we have the listings data available in our dataset.

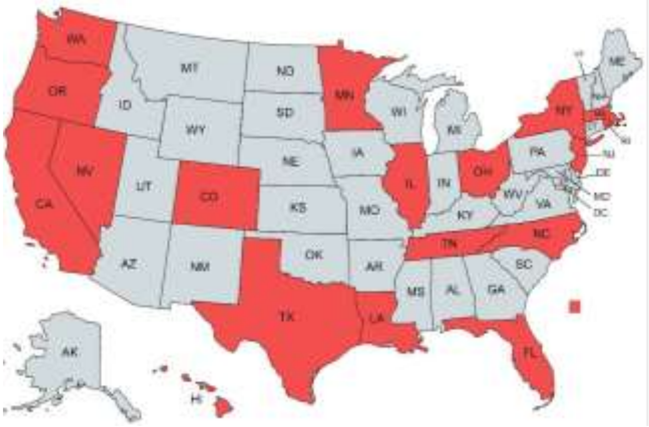


Fig. 1. State Wise Listings of Airbnb (USA)

We have followed the below mentioned steps as part of the initial Data Cleaning process.

- We dropped the columns which were not relevant for our analysis such as *Listing URL, Host URL, Last Scraped Date, Scrape ID.* We have also dropped the *Host Name* column in order to avoid data privacy issues. Additionally columns with greater than 50% of NA values were dropped such as the *bathrooms* and *calendar_updated* column.
- Some of the columns such as *neighbourhood_group_cleansed*, *neighbourhood_cleansed* and *license* had mixed data types, so we have converted all such columns to string datatype for uniformity.
- For the text features like *Listing Name and Description*, *host_location and host_about* we performed basic operations such as lower casing and removing everything except the characters.
- The columns with boolean values such as host_is_superhost, host_has_profile_pic which had 't' and 'f' string values were mapped to respective boolean values (1 and 0).
- For some of the listings multiple entries were present with the same Listing IDs, so we filtered and removed the multiple entries by sorting based on the Last Scrapped date to retain the most recent entry of that listing.

<u>Missing Data</u>:

- The values such as ***name, description, host name, host about*** are entered by hosts, and their missingness does not depend on any other variables as well so we can consider it as a case of MCAR. This can be considered as a case of entry error either by the Host or by Airbnb. Since this is a text column and imputing it with any other value is not possible we have replaced these missing values with the text as 'Not Entered'.
- Missing values for Boolean fields such as host_has_profile_pic, host_identity_verified, has_availability were replaced with false.
- The ***host_acceptance_rate*** **and** ***host_response_rate*** were related to the variables such as *host_response_time_score* as well as few other variables due to which we considered this as a case of MAR and the values were imputed using Lasso Regression. Additionally, these columns had greater than 30,000 missing values and as per [8], Lasso Regression imputation performs better than the other methods for high missing values within high dimensional datasets. Hence, we have used Lasso Regression for performing the imputation.

Data Preprocessing:

- Stop words were removed followed by lemmatization on the text columns such as *description, name, neighborhood_overview, neighbourhood_cleansed, property_type, amenities and neighbourhood_group_cleansed, host_location, host_about, host_neighbourhood*

Feature Engineering/Selection:

<u>Newly Created Features</u>:

- Using the date columns such as *host_since, first_review* and *last_review* we have created new features (ex. ***host_months***, ***host_years***) by extracting the number of months and years for easy processing.
- The *host_response_time* feature was used to calculate the ***host_response_time_score*** as below. This mapping is based on the time taken to respond, thus "within an hour" has the highest score.

Original Values	Assigned Scores
within an hour	4
within a few hours	3
within a day	2
a few days or more	1

- For every listing the *host_verifications* field had values such as *['email', 'phone', 'work_email']*. We calculated the ***host_verifications_score*** field based on the number of verifications the host has. For the above example, the score will be *3*.
- We have created dummy variables in order to encode the room_type (values such as 'Entire home/apt', 'Private room' etc) categorical variable.
- Another boolean variable ***bathroom_shared*** was created from the *bathrooms_text* which had the information whether a bathroom is shared or not (For instance: "1 shared bath").

After completion of all the data cleaning and preprocessing steps along with the addition of all the new features, we had a dataset with *256,887 rows* and *86* columns. In order to identify the most important features to be considered in our analysis we initially performed the classification using a base model.

The ***Logistic Regression Model*** was used to undertake our initial analysis of the data for the purpose of choosing significant features. With the exception of the text columns (which provide textual details about the host and the listings), we used all the other data in this model. The model's outcomes are shown in Table 3 below. Even though the accuracy fell short of expectations, the training and validation outcomes were in sync.

We looked at feature reduction as one approach to feature engineering. One method of achieving this involved organizing the host's and listings' textual information into two separate columns. This reduced the 11 textual columns to 2 columns namely listings_text and host_text respectively as mentioned below.

listing_text *= (name + description + neighborhood_overview + neighborhood + neighbourhood_cleansed + neighbourhood_group_cleansed + property_type + amenities)*

host_text *= (host_location + host_about + host_neighbourhood)*

We then used these two columns to derive useful numeric features from this text using Bag Of Words and further transforming the same to calculate the TF-IDF of the words. This helped us to analyze if textual content such as amenities and description of the property had any significant impact on the likeliness of the host being a super host. The most significant amenities for the listings where the host is a super host are displayed in the word cloud below.

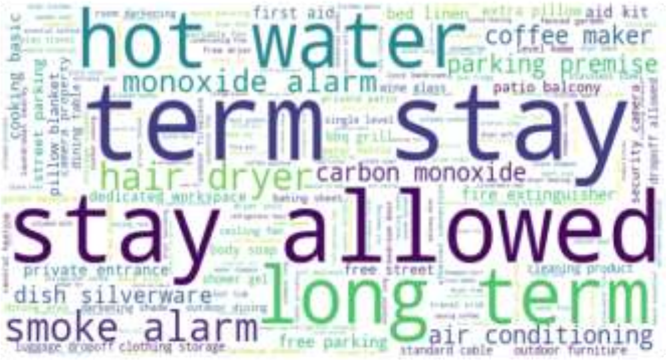


Fig 2. Word Cloud for Amenities

Any method of component analysis proved to be ineffective in improving the model through the feature selection due to the large dimensionality of the data as well as the inconsistent data types. We chose a Random Forest feature selection process instead, which reduced the features from 56 to 32. The majority of the features presented had strong correlations with one another. Interestingly, the listings' latitude and longitude were a contributing factor. The accuracy of the base model significantly increased once these characteristics were included, and the discrepancy between the training and validation scores was further narrowed. In the upcoming subsection we have implemented and evaluated various models.

Models Used and Evaluation:

Running the basic model on the top 32 characteristics that were retrieved in the previous part served as the basis for our initial analysis. The model's accuracy improved to 73%, and the difference between the training and validation sets further minimized. Additionally, we computed the AUC (Area Under the Curve) and the ROC (Receiver Operating Characteristic Curve). The ROC and AUC for the fundamental Logistic Regression Model are shown in the following graph.

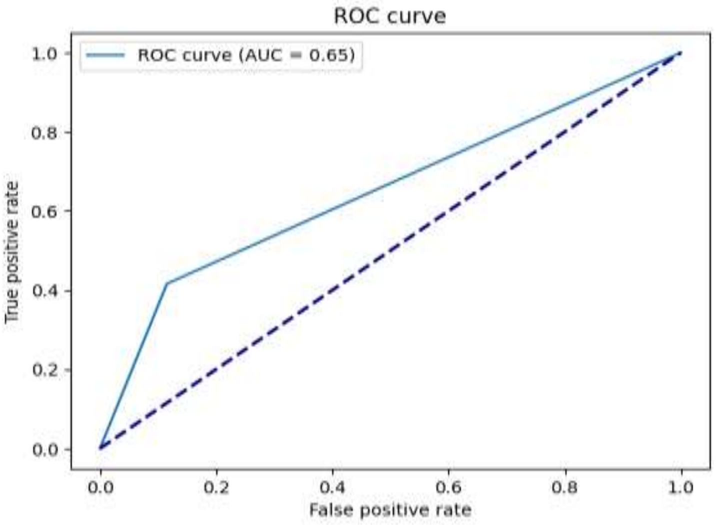


Fig 3: ROC Curve - AUC

We also expanded our features to include the textual information. We did basic cleaning, preprocessing, and feature reduction and lemmatizing the data while taking into account the POS Tagging of the tokens. This kept the text's true relevance to the data. By fitting the Bag of Words model to the listings_text and host_text columns, we were able to get the numerical characteristics. Additionally, we used the TFIDF Transformer to convert the fitted data. The columns that were obtained numerically were then combined with this data. We also standardized the 0–1 range of numerical columns in this phase. The data became more consistent as a result. The accuracy and the F1-score showed considerable improvement after incorporating the text columns in the model. The results and metrics have been depicted in Table 3.

We then used Tree Classifiers to assess our data. We utilized the Decision Tree Classifier with all the features at first, and its performance was about equivalent to the basic model. However, Decision Tree Classifier surpassed the outcomes of the basic model when the same model was applied with the chosen features. To further analyze, the tree classifiers we performed the same with Random Forest Classifiers and the outcomes were much better. The values of the metrics that were recorded when capturing the results are displayed in Table 3.

As of right now, the tree classifiers outperformed the rest of the other models we implemented, including SVM. There was one additional method, though, whose efficiency was comparable to that of the Random Forest Classifier. Additionally, XGBoost (Gradient Boosting) showed metrics that were quite similar to the top model. Additionally, XGBoost only returned 10 columns when we first used it to extract features from the 56 columns. Using only the 10 features, XGBoost performed admirably when we applied the same to the model. With 32 characteristics, it did not, however, outperform Random Forest. Table 3 shows the tabular analysis of the models and the corresponding metrics. In the section below we have further discussed the research gaps and critiqued the models implemented.

IV. CONCLUSION

Over the years there has been a significant amount of research done in the area of Airbnb listings price prediction as mentioned before. In this paper we have addressed a previously unexplored aspect of the superhost status by examining additional factors impacting superhost status in addition to the four factors that Airbnb has explicitly mentioned– cancellation rate, response rate, rating, and number of stays. We have used various machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, XGBoost for classifying whether a host is also a Superhost. Due to the high dimensionality of data we also used various feature selection techniques as mentioned in the previous section to retrieve the most important features. It was found that the models performed better on the 32 selected features as compared to the model trained on all features in the dataset. Including the textual features in addition to the numeric features, helped in improving the accuracy of all models. Random Forest Classifier outperformed the other models and achieved F1 scores of 95.23% and 89.53% for training and validation data respectively as seen in Table 3. We can infer from the analysis that factors like host acceptance rate, listing count, review criteria like review score, number of reviews, as well as whether the host has been verified, whether the host has a license, and whether the listing is instantly bookable all affect a host's superhost status. Thus, it becomes necessary for a host to consider these parameters, improve and further maintain their hospitality standards in order to attract the maximum number of customers and increase their profit.

V. REFERENCES

[1] J. Jiao, S. Bai, “An empirical analysis of Airbnb listings in forty American cities”, Cities,Volume 99, 102618, ISSN 0264-2751, 2020, https://doi.org/10.1016/j.cities.2020.102618.

[2] J. Dhillon et al., “Analysis of Airbnb Prices using Machine Learning Techniques,” 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), NV, USA, 2021, pp. 0297-0303, doi: 10.1109/CCWC51732.2021.9376144.

[3] “Fast facts”, Airbnb Newsroom, Dec. 2022, Available: https://news.airbnb.com/about-us/.

[4] E. Ert, A. Fleischer, "The evolution of trust in Airbnb: A case of home rental", Annals of Tourism Research, Volume 75, Pages 279-287, ISSN 0160-7383, 2019 https://doi.org/10.1016/j.annals.2019.01.004

[5] “About Superhost”, Help Centre, Available: https://www.airbnb.ie/help/article/828?locale=en&_set_bev_on_new_domain=1680960031_ZmNmMzE2NTA0ODE3.

[6] “How to become a Superhost”, Help Centre, Available: https://www.airbnb.ie/help/article/829.

[7] “Why strive for Superhost status”, Resource Centre, May 2021, Available: https://www.airbnb.ie/resources/hosting-homes/a/why-strive-for-superhost-status-50.

[8] M. Takada, H. Fujisawa and T. Nishikawa, “HMLasso: Lasso with High Missing Rate”, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), 2019 https://www.ijcai.org/proceedings/2019/0491.pdf.

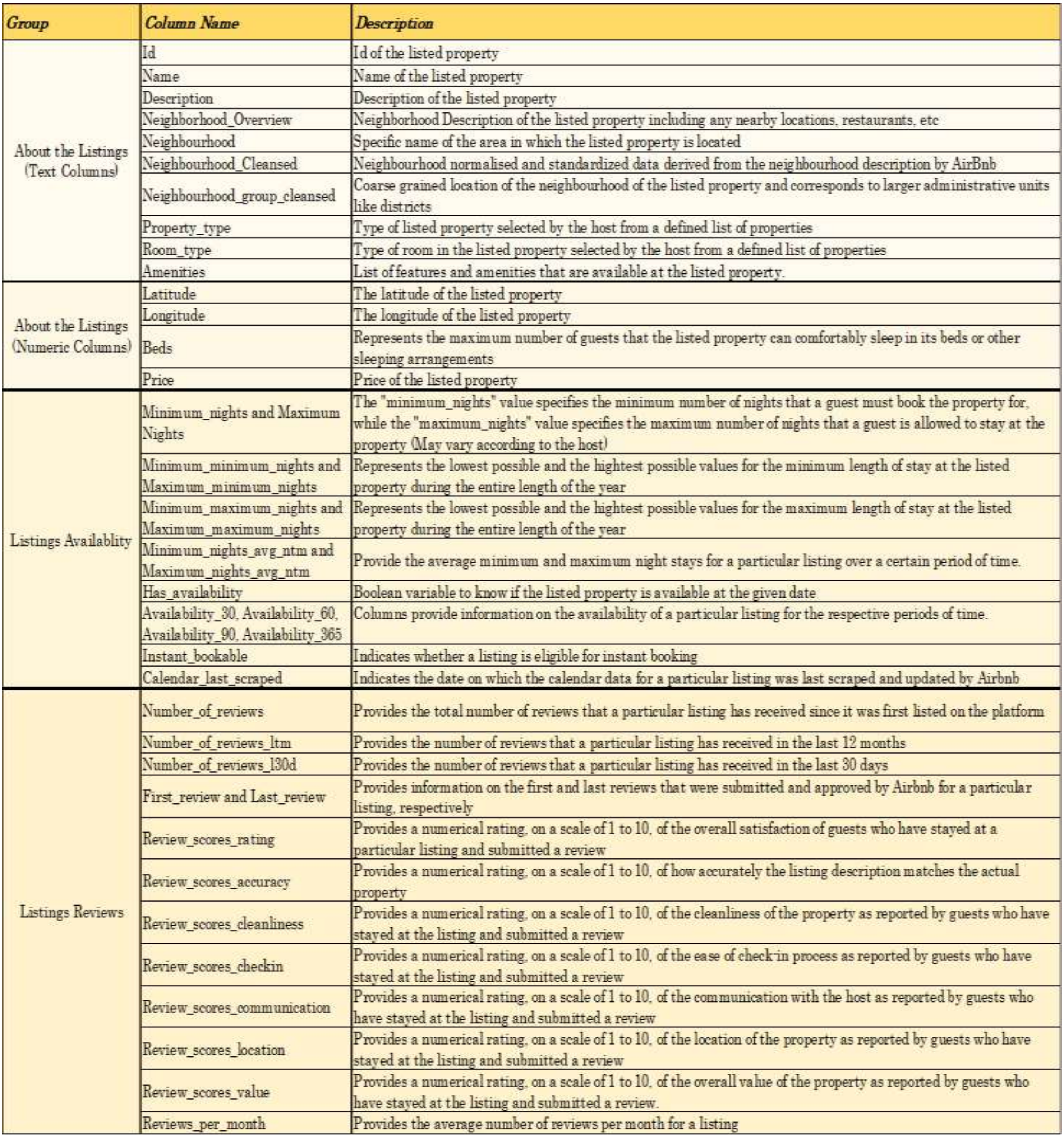
[9] U. Gunter, “What makes an Airbnb host a superhost? Empirical evidence from San Francisco and the Bay Area”, Tourism Management, Volume 66, Pages 26-37, ISSN 0261-5177, 2018 https://doi.org/10.1016/j.tourman.2017.11.003.

[10] Y. Liu, "Airbnb Pricing Based on Statistical Machine Learning Models," 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML), Stanford, CA, USA, 2021, pp. 175-185, doi: 10.1109/CONF-SPML54095.2021.00042.

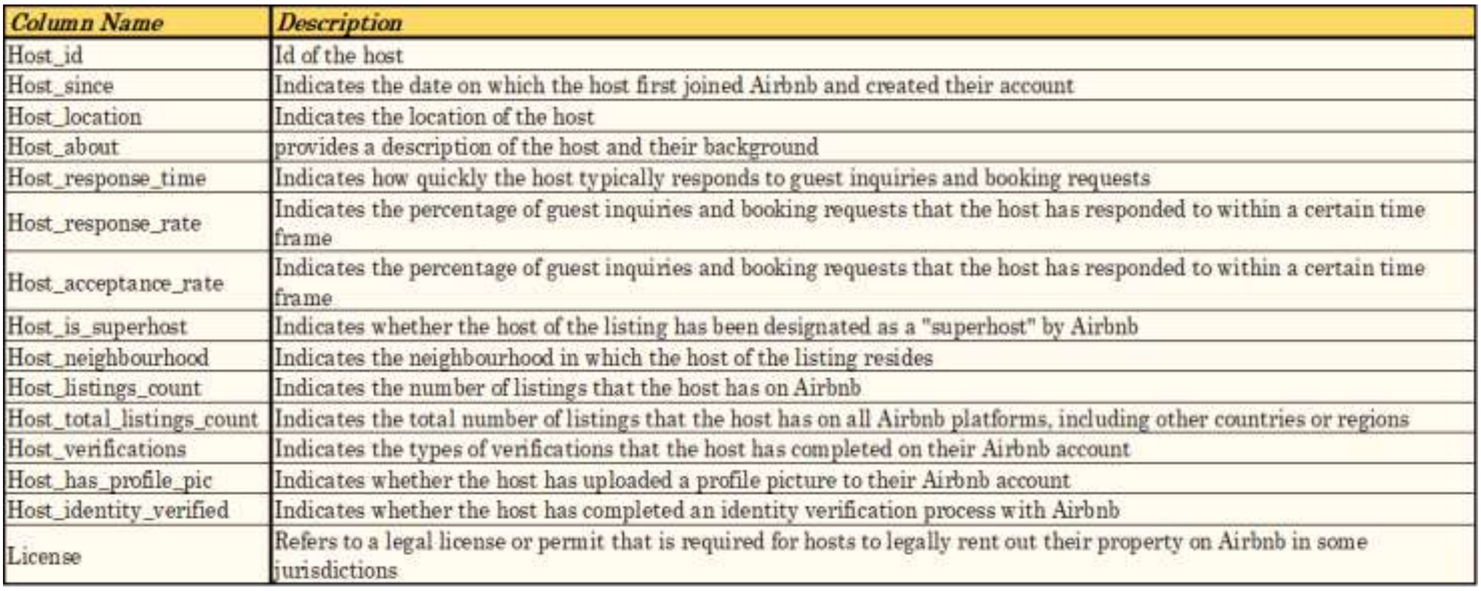
[11] M. Roelofsen, C. Minca, The Superhost. Biopolitics, home and community in the Airbnb dream-world of global hospitality, Geoforum, Volume 91, Pages 170-181, ISSN 0016-7185, 2018, https://doi.org/10.1016/j.geoforum.2018.02.021.

[12] A. Farmaki, P. Christou, A. Saveriades, “A Lefebvrian analysis of Airbnb space”, Annals of Tourism Research, Volume 80, 102806, ISSN 0160-7383, 2020, https://doi.org/10.1016/j.annals.2019.102806

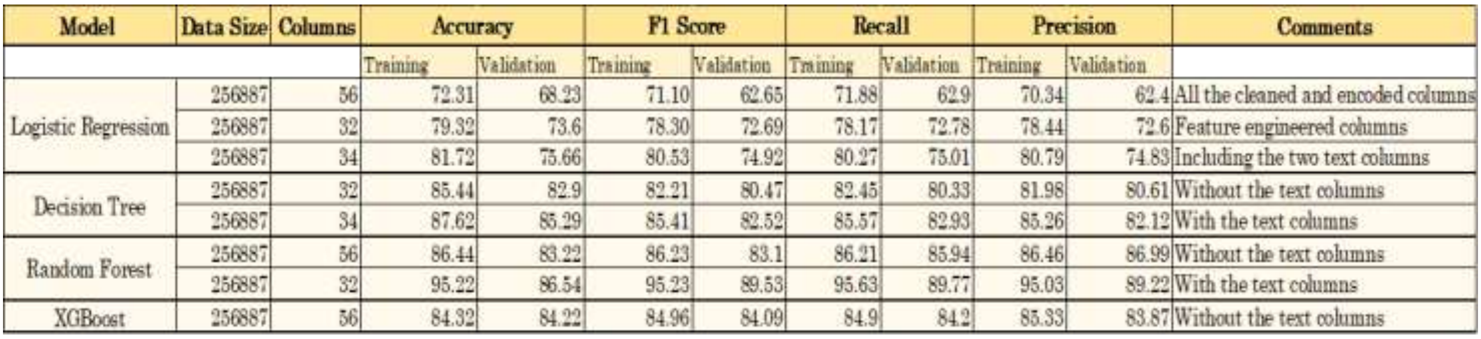
APPENDIX



Group	Column Name	Description
About the Listings (Text Columns)	Id	Id of the listed property
	Name	Name of the listed property
	Description	Description of the listed property
	Neighborhood_Overview	Neighborhood Description of the listed property including any nearby locations, restaurants, etc
	Neighbourhood	Specific name of the area in which the listed property is located
	Neighbourhood_Cleansed	Neighbourhood normalised and standardized data derived from the neighbourhood description by AirBnb
	Neighbourhood_group_cleansed	Coarse grained location of the neighbourhood of the listed property and corresponds to larger administrative units like districts
	Property_type	Type of listed property selected by the host from a defined list of properties
	Room_type	Type of room in the listed property selected by the host from a defined list of properties
	Amenities	List of features and amenities that are available at the listed property.
About the Listings (Numeric Columns)	Latitude	The latitude of the listed property
	Longitude	The longitude of the listed property
	Beds	Represents the maximum number of guests that the listed property can comfortably sleep in its beds or other sleeping arrangements
	Price	Price of the listed property
Listings Availablity	Minimum_nights and Maximum Nights	The "minimum_nights" value specifies the minimum number of nights that a guest must book the property for, while the "maximum_nights" value specifies the maximum number of nights that a guest is allowed to stay at the property (May vary according to the host)
	Minimum_minimum_nights and Maximum_minimum_nights	Represents the lowest possible and the hightest possible values for the minimum length of stay at the listed property during the entire length of the year
	Minimum_maximum_nights and Maximum_maximum_nights	Represents the lowest possible and the hightest possible values for the maximum length of stay at the listed property during the entire length of the year
	Minimum_nights_avg_ntm and Maximum_nights_avg_ntm	Provide the average minimum and maximum night stays for a particular listing over a certain period of time.
	Has_availability	Boolean variable to know if the listed property is available at the given date
	Availability_30, Availability_60, Availability_90, Availability_365	Columns provide information on the availability of a particular listing for the respective periods of time.
	Instant_bookable	Indicates whether a listing is eligible for instant booking
	Calendar_last_scraped	Indicates the date on which the calendar data for a particular listing was last scraped and updated by Airbnb
Listings Reviews	Number_of_reviews	Provides the total number of reviews that a particular listing has received since it was first listed on the platform
	Number_of_reviews_ltm	Provides the number of reviews that a particular listing has received in the last 12 months
	Number_of_reviews_l30d	Provides the number of reviews that a particular listing has received in the last 30 days
	First_review and Last_review	Provides information on the first and last reviews that were submitted and approved by Airbnb for a particular listing, respectively
	Review_scores_rating	Provides a numerical rating, on a scale of 1 to 10, of the overall satisfaction of guests who have stayed at a particular listing and submitted a review
	Review_scores_accuracy	Provides a numerical rating, on a scale of 1 to 10, of how accurately the listing description matches the actual property
	Review_scores_cleanliness	Provides a numerical rating, on a scale of 1 to 10, of the cleanliness of the property as reported by guests who have stayed at the listing and submitted a review
	Review_scores_checkin	Provides a numerical rating, on a scale of 1 to 10, of the ease of check-in process as reported by guests who have stayed at the listing and submitted a review
	Review_scores_communication	Provides a numerical rating, on a scale of 1 to 10, of the communication with the host as reported by guests who have stayed at the listing and submitted a review
	Review_scores_location	Provides a numerical rating, on a scale of 1 to 10, of the location of the property as reported by guests who have stayed at the listing and submitted a review
	Review_scores_value	Provides a numerical rating, on a scale of 1 to 10, of the overall value of the property as reported by guests who have stayed at the listing and submitted a review.
	Reviews_per_month	Provides the average number of reviews per month for a listing



Column Name	*Description*
Host_id	Id of the host
Host_since	Indicates the date on which the host first joined Airbnb and created their account
Host_location	Indicates the location of the host
Host_about	provides a description of the host and their background
Host_response_time	Indicates how quickly the host typically responds to guest inquiries and booking requests
Host_response_rate	Indicates the percentage of guest inquiries and booking requests that the host has responded to within a certain time frame
Host_acceptance_rate	Indicates the percentage of guest inquiries and booking requests that the host has responded to within a certain time frame
Host_is_superhost	Indicates whether the host of the listing has been designated as a "superhost" by Airbnb
Host_neighbourhood	Indicates the neighbourhood in which the host of the listing resides
Host_listings_count	Indicates the number of listings that the host has on Airbnb
Host_total_listings_count	Indicates the total number of listings that the host has on all Airbnb platforms, including other countries or regions
Host_verifications	Indicates the types of verifications that the host has completed on their Airbnb account
Host_has_profile_pic	Indicates whether the host has uploaded a profile picture to their Airbnb account
Host_identity_verified	Indicates whether the host has completed an identity verification process with Airbnb
License	Refers to a legal license or permit that is required for hosts to legally rent out their property on Airbnb in some jurisdictions



Model	Data Size	Columns	Accuracy		F1 Score		Recall		Precision		Comments
			Training	Validation	Training	Validation	Training	Validation	Training	Validation	
Logistic Regression	256887	56	72.31	68.23	71.10	62.65	71.88	62.9	70.34	62.4	All the cleaned and encoded columns
	256887	32	79.32	73.6	78.30	72.69	78.17	72.78	78.44	72.6	Feature engineered columns
	256887	34	81.72	75.66	80.53	74.92	80.27	75.01	80.79	74.83	Including the two text columns
Decision Tree	256887	32	85.44	82.9	82.21	80.47	82.45	80.33	81.98	80.61	Without the text columns
	256887	34	87.62	85.29	85.41	82.52	85.57	82.93	85.26	82.12	With the text columns
Random Forest	256887	56	86.44	83.22	86.23	83.1	86.21	85.94	86.46	86.99	Without the text columns
	256887	32	95.22	86.54	95.23	89.53	95.63	89.77	95.03	89.22	With the text columns
XGBoost	256887	56	84.32	84.22	84.96	84.09	84.9	84.2	85.33	83.87	Without the text columns