PREDICTING AIRBNB RATINGS: KEY DRIVERS OF PERFECT RATINGS USING NLP AND FEATURE ENGINEERING

1. ABSTRACT

In this study, we explore an extensive predictive analysis of Airbnb properties to determine factors that drive a perfect rating score. The focus of the analysis is to maximize the True Positive Rate (TPR) while keeping the False Positive Rate (FPR) below 10%, using a combination of machine learning and natural language processing (NLP) techniques. A novel approach involving feature engineering with word embeddings and k-means clustering was utilized to represent textual features effectively. By constructing a document-term matrix (DTM) and deriving word embeddings, we obtained meaningful clusters that were further used as features in our models. In addition, we performed feature engineering based on exploratory data analysis (EDA), which helped in identifying and transforming significant features from the dataset to improve model accuracy. Several machine learning algorithms, including logistic regression, random forests, and boosting methods, were implemented. Among them, the XGBoost model demonstrated the best performance, achieving a TPR of 45.33% while maintaining an FPR of 9.55%. This model, combined with hyperparameter tuning, provided significant insights into the factors contributing to high rental property ratings, offering value for business decision-making in the rental property market.

2. INTRODUCTION

The growth of short-term rental platforms has transformed the global tourism and hospitality industries. Airbnb, in particular, has seen rapid adoption, providing travelers with a wide variety of lodging options while allowing property owners to monetize their spaces. However, the success of Airbnb listings is not uniform across properties, and achieving a perfect rating has become an important benchmark for hosts seeking to attract more guests and increase their revenue. Understanding the factors that influence these ratings is crucial for both hosts and the platform itself in improving service quality and meeting customer expectations.

The dataset used in this study was sourced from Inside Airbnb, which provides publicly available data on Airbnb listings from cities worldwide[1]. It has information on 69 features for every listing, details of which are mentioned in the Data Dictionary. Prior research has shown that Airbnb has disrupted the traditional hospitality industry significantly, affecting hotel occupancy and pricing strategies (Dogru, Mody, & Suess, 2019[2]; Guttentag, 2015[3]). Understanding the factors that lead to success for Airbnb hosts, such as achieving high ratings, is vital for competing in this evolving market.

Factors influencing listing performance have been analyzed from various angles, including host attributes, location, pricing, and textual content. Xie and Mao (2017)[4] found that host characteristics, such as responsiveness, significantly influence listing performance. Similarly, Müller (2020)[5] demonstrated that location and neighborhood characteristics are critical determinants of listing success. Our study builds on these findings by including host responsiveness, number of listings managed, and neighborhood information to enhance the predictive model.

Wang and Nicolau (2017)[6] studied price determinants for Airbnb listings, noting that affordability is a major factor in attracting guests. Our analysis incorporated features such as price per person and cleaning fees to assess the impact of pricing on ratings. Additionally, Zhang, Law, and Li (2020)[7] highlighted the role of machine learning in improving predictive accuracy for hospitality research, which motivated the use of ensemble models like XGBoost in our study.

The textual content of Airbnb listings also plays an essential role in listing success. Studies by Guttentag (2015)[3] and Xie and Mao (2017)[4] have emphasized the importance of unique experiences and host engagement. To capture the sentiment and content richness of listings, we applied natural language processing (NLP) techniques, generating features from guest interactions, access descriptions, and listing descriptions. This comprehensive approach, integrating insights from multiple aspects of Airbnb performance, aimed to provide a holistic predictive model for Airbnb success.

3. DATA & METHODOLOGY

Exploratory Data Analysis (EDA)

To gain an initial understanding of the dataset, an exploratory data analysis (EDA) was conducted. This involved investigating the distribution of key variables, identifying correlations, and dealing with missing values. Visualizations such as scatter plots, histograms, and correlation matrices were employed to highlight important trends and patterns. Additionally, missing values were handled by imputation strategies such as replacing missing numerical values with their means and categorical variables with their modes. The conclusions drawn from our EDA are:

1. **Availability_90 (Fig. 1):** The boxplot between availability and the target variable shows that properties with a perfect rating score of 1 are more likely to have a low value for Availability_90. Additionally, the distribution of properties with a perfect rating score of 1 is focused on the extremities of Availability_90, while properties with a perfect rating score of 0 are more evenly distributed.
2. **Price (Fig. 2):** The scatterplot between price and the target variable indicates that higher-priced Airbnbs, particularly those priced between $3,000 and $7,500, are more likely to have a perfect rating score of 1.
3. **Population density (Fig. 3):** There is a slight peak in the count of properties without a perfect rating score for cities with population density between 2,000 and 3,000. This suggests that there may be higher chances of properties not having a perfect rating score in cities with this population density range.
4. **Zip Code (Fig. 4):** Analysis shows that 40% of the zip codes have a perfect rating score of 0. These zip codes can be grouped together, and regression analysis can be performed to gain further insights.
5. **Population (Fig. 5):** Density plots of population show no major effect on the perfect rating score, but the enclosed region suggests that properties in cities with a certain population range tend to have more properties without a perfect rating score. This finding warrants further exploration using regression analysis.
6. **Instant Bookable (Fig. 6):** The data shows that non-instant bookable properties are more likely to have a non-perfect rating score, suggesting a relationship worth studying further.

7. **Host total listings (Fig. 7):** The analysis indicates that having fewer listings helps in predicting a perfect rating score more accurately. Hosts with a large number of listings are often outliers, which may affect guest satisfaction.
8. **Security Deposit (Fig. 8):** The findings suggest that having a moderate security deposit is better than having a high amount. Additionally, properties without a security deposit tend to be preferred over those with one.
9. **Accommodates (Fig. 9):** The relationship between the accommodates feature and the perfect rating score is mixed. However, properties accommodating 2-3 people are more likely to have a non-perfect rating score, as evidenced by the wider strings of the violin plot. This relationship is debatable, as data points are densely located around 0-5 accommodates for both categories.
10. **Availability_30 (Fig. 10):** 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.

Feature Engineering

Feature engineering played a crucial role in enhancing the predictive power of our models. Based on the insights gained from EDA, several new features were derived to improve model accuracy. For instance, the price per person feature was created to account for the affordability of listings, while clustering techniques were applied to group similar properties based on amenities and textual descriptions. Text data, such as descriptions and interactions, were processed using natural language processing (NLP) techniques, including tokenization, word embeddings, and k-means clustering. These engineered features provided richer information to the model, enabling it to better capture the nuances that contribute to high ratings. Fig. 11 gives a flowchart of the 8 steps to obtain the clusters which are used as features.

Based on the insights gained from EDA, several new features were derived to improve model accuracy:

1. **Price per Person:** This feature was created by dividing the nightly price by the number of people the listing can accommodate. The rationale behind this was to capture the affordability of the listing, which is an important factor for many guests. Listings with a reasonable price per person are likely to receive better reviews, which could contribute to achieving a perfect rating.
2. **Bed Category:** This feature categorizes bed types into either "bed" (for real beds) or "other". The quality of sleeping arrangements is a significant determinant of guest satisfaction. By categorizing bed types, we aimed to capture the effect of different sleeping options on the overall rating.
3. **Property Category:** Property types were grouped into categories such as "apartment," "hotel," "condo," "house," and "other" to simplify the analysis. Different property types can have varying levels of appeal to guests, and categorizing them helped in understanding their influence on ratings.
4. **Price per Person Indicator (PPP Indicator):** This binary feature indicates whether the price per person is above the median for the property category. This indicator helps in understanding whether premium pricing affects guest satisfaction and the likelihood of achieving a perfect rating.
5. **Charges for Extra People:** This binary feature indicates whether there is an additional charge for extra guests. It was included to capture whether additional costs affect the guest experience and satisfaction.

6. **Host Acceptance and Response:** These features categorize the host's acceptance rate and response rate as "ALL," "SOME," or "MISSING." These features help in assessing the impact of host behavior on guest satisfaction. A responsive and accommodating host is likely to receive better ratings.
7. **Minimum Nights Requirement (Has Minimum Nights):** This feature indicates whether the listing has a minimum night requirement greater than one. It was engineered to understand how minimum stay requirements impact guest satisfaction and their likelihood of leaving a perfect rating.
8. **Market Consolidation:** Markets with fewer than 300 listings were grouped under "OTHER" to reduce the number of categories and improve the robustness of the model. This feature helps in understanding how the geographical location of a listing affects its rating.
9. **Textual Features (Interaction, Access, Description):** Textual data such as guest interactions, access descriptions, and listing descriptions were processed using natural language processing (NLP) techniques. These text columns were tokenized, embedded using word embeddings, and clustered using k-means to generate new features representing the sentiment and key topics discussed in these texts. The purpose of these features was to capture the effect of descriptive information and sentiment on guest satisfaction.
10. **Listing Duration and Last Review Duration:** The number of days since the listing was created (listing duration) and the number of days since the last review (last review duration) were calculated. These features help in understanding the effect of the listing's age and recent activity on guest satisfaction.

Methodology

The prediction of perfect ratings was approached using machine learning models, with a focus on maximizing TPR while minimizing FPR. Various models were implemented, including logistic regression, random forests, and XGBoost. The XGBoost model, in particular, was tuned extensively to optimize its performance. Hyperparameter tuning was carried out to find the best combination of parameters, such as maximum depth, learning rate (eta), and the number of boosting rounds, to achieve the highest TPR while keeping FPR below 10%.

The model training process involved several steps:

1. **Data Preparation**: The features were transformed into a format suitable for machine learning models, including encoding categorical variables and normalizing numerical features.
2. **Text Processing**: Textual data from various columns, such as descriptions and amenities, were transformed using word embeddings and then clustered to create additional features for the model.
3. **Hyperparameter Tuning**: A validation set was used to tune hyperparameters, optimizing the model to achieve a balance between TPR and FPR. This step was crucial to ensure that the model provided precise recommendations without overestimating the potential of suboptimal listings.

The final model was evaluated on a separate test set to ensure its generalizability. The results indicated that the XGBoost model was the most effective, achieving a TPR of 45.33% while keeping the FPR below 10%. This outcome demonstrated the ability of the model to accurately identify listings with the potential for a perfect rating, providing valuable insights for Airbnb hosts.

4. RESULTS

The results of our analysis indicate that different machine-learning models had varying levels of success in predicting perfect ratings. We evaluated seven models: Logistic Regression, Linear Regression, Ridge & Lasso Regression, Trees, Random Forest, XGBoost without clusters, and XGBoost with clusters, the details of which are in **Fig. 12**. The clusters refer to features added using NLP on the three text columns (interaction, access, and description). The performance of each model was assessed based on the True Positive Rate (TPR) and False Positive Rate (FPR).

The XGBoost model with clusters outperformed all other models, achieving the highest TPR while maintaining a relatively low FPR. This demonstrates the effectiveness of using advanced NLP techniques to generate features from textual data, which helped the model capture more nuanced relationships. The XGBoost without clusters also performed well, but the additional features derived from clustering the textual columns led to a noticeable improvement in TPR.

Random Forest also showed competitive results, with a TPR of 39.70% and an FPR of 9.19%. However, it did not reach the same level of predictive power as the XGBoost models. Logistic Regression and Ridge & Lasso Regression performed moderately well, with TPRs of 37.24% and 37.35%, respectively, and similar FPRs. Linear Regression had slightly lower performance compared to these models, while Trees had the lowest TPR of 29.25%, accompanied by a higher FPR of 10.60%.

The inclusion of NLP-derived clusters in the XGBoost model significantly enhanced its ability to predict perfect ratings, highlighting the value of incorporating textual data in prediction models. The clusters captured semantic information from the text, which provided additional context that improved the model's decision-making capabilities.

Overall, the XGBoost model with clusters demonstrated the highest performance, emphasizing the importance of feature engineering, particularly with textual data, in achieving accurate predictions. The results suggest that using ensemble methods combined with advanced NLP techniques can effectively enhance the predictive power of models in the short-term rental domain.

5. CONCLUSION

The XGBoost model with clusters outperformed all other models, achieving the highest TPR while maintaining a relatively low FPR. This demonstrates the effectiveness of using advanced NLP techniques to generate features from textual data, which helped the model capture more nuanced relationships. The additional features derived from clustering the textual columns led to a noticeable improvement in TPR, highlighting the value of feature engineering in predictive modeling.

Feature engineering played a pivotal role in increasing the True Positive Rate (TPR) across different models. The NLP-derived features, specifically the clusters representing interaction, access, and description text, provided richer contextual information that significantly improved model accuracy. By capturing the sentiment and important topics discussed in these text columns, the model was better able to predict which listings were likely to achieve a perfect rating. Overall, the combination of ensemble modeling, hyperparameter tuning, and feature engineering—particularly from textual data—was key to the success of this analysis, offering valuable insights for Airbnb hosts looking to improve their properties.

6. REFERENCES

[1]. https://insideairbnb.com/get-the-data/

[2]. Dogru, T., Mody, M., & Suess, C. (2019). The Airbnb paradox: Positive employment effects in the hospitality industry. Tourism Management.

[3]. Guttentag, D. (2015). Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. Current Issues in Tourism.

[4]. Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy-based accommodation rental: A study of listings from 33 cities on Airbnb.com. International Journal of Hospitality Management.

[5]. Xie, K. L., & Mao, Z. (2017). The Impacts of Quality and Quantity Attributes of Airbnb Hosts on Listing Performance. International Journal of Contemporary Hospitality Management.

[6]. Zhang, Z., Law, R., & Li, A. (2020). Machine learning in hospitality and tourism: A systematic review. International Journal of Contemporary Hospitality Management.

[7]. Müller, A. (2020). Location and neighborhood characteristics in predicting Airbnb performance. Journal of Real Estate Research.

7. FIGURES

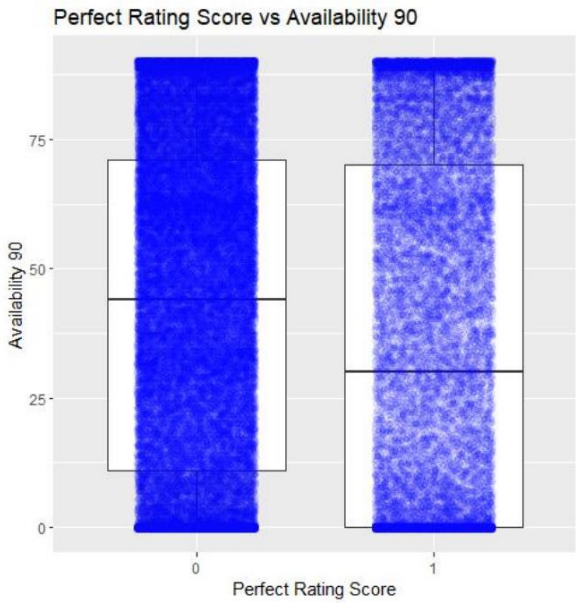

Fig. 1: Boxplot between Perfect Rating Score vs. Availability_90

Fig. 2: Scatterplot between Price and Perfect Rating Score

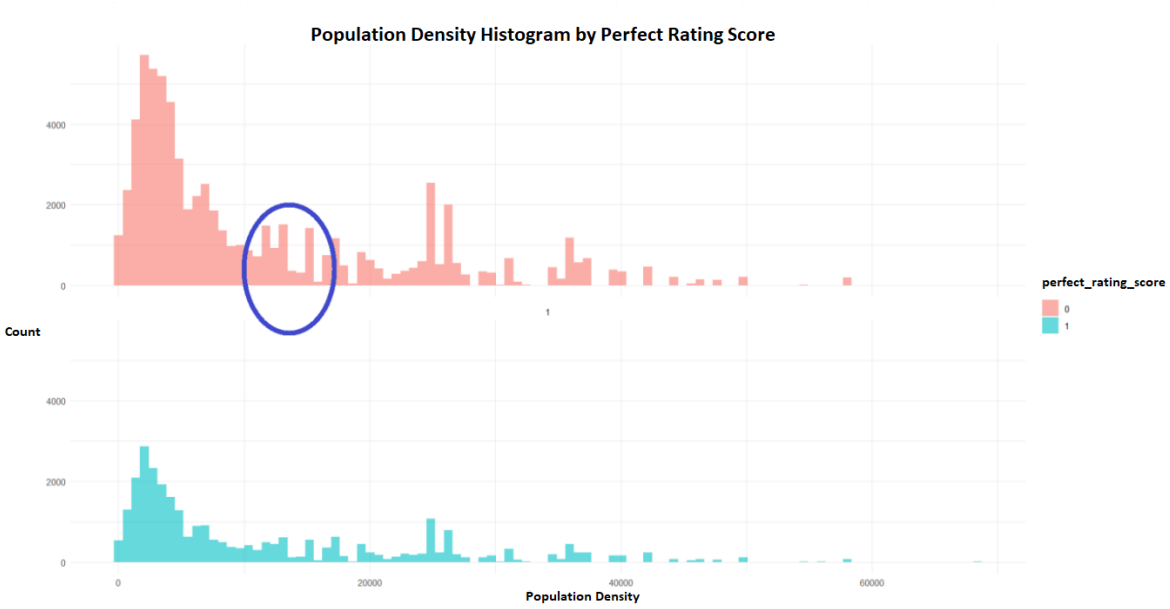

Fig. 3: Population Density Histogram

	perfect_rating_score = 1	perfect_rating_score = 0
Count of zipcodes	28159	71822

Fig. 4: Zipcodes for each value of Perfect Rating Score

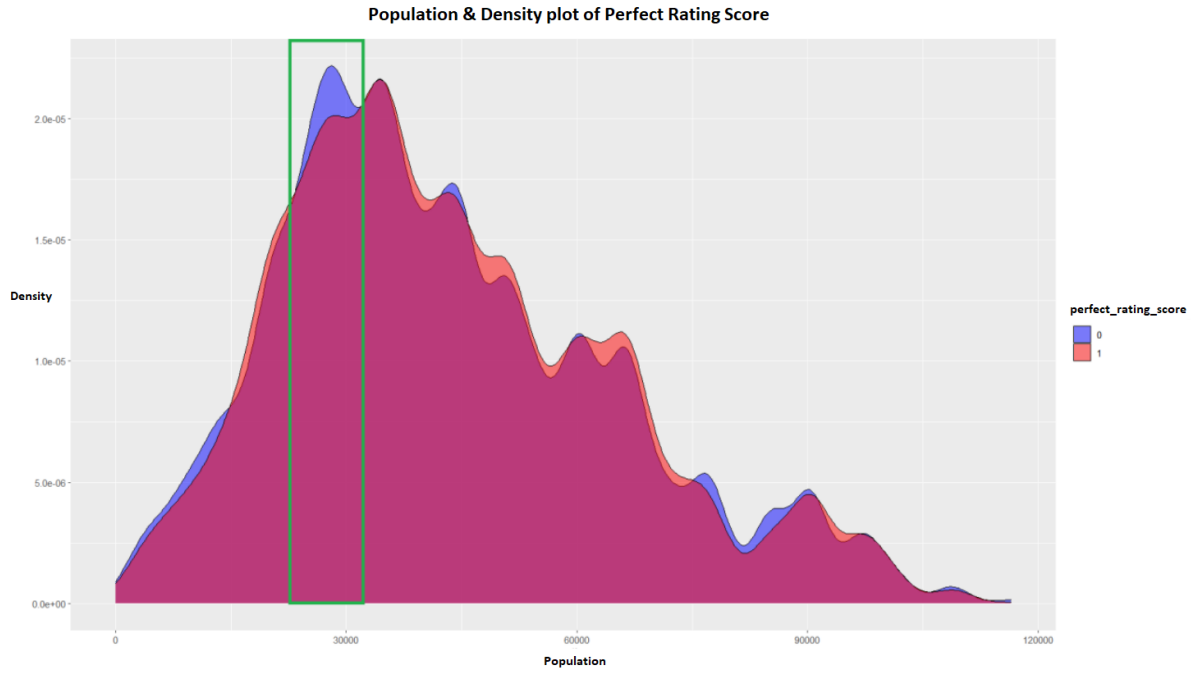

Fig. 5: Population Density plot for Perfect Rating Score

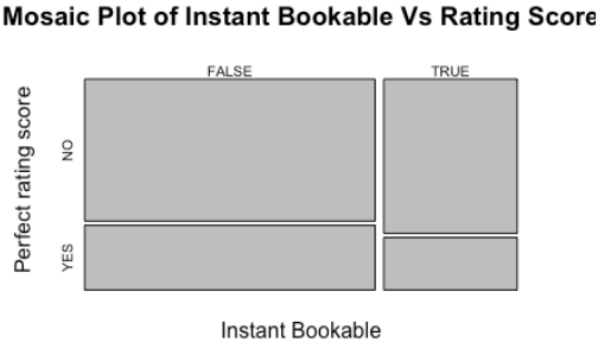

Fig. 6: Mosaic Plot of Instant Bookable vs. Perfect Rating Score

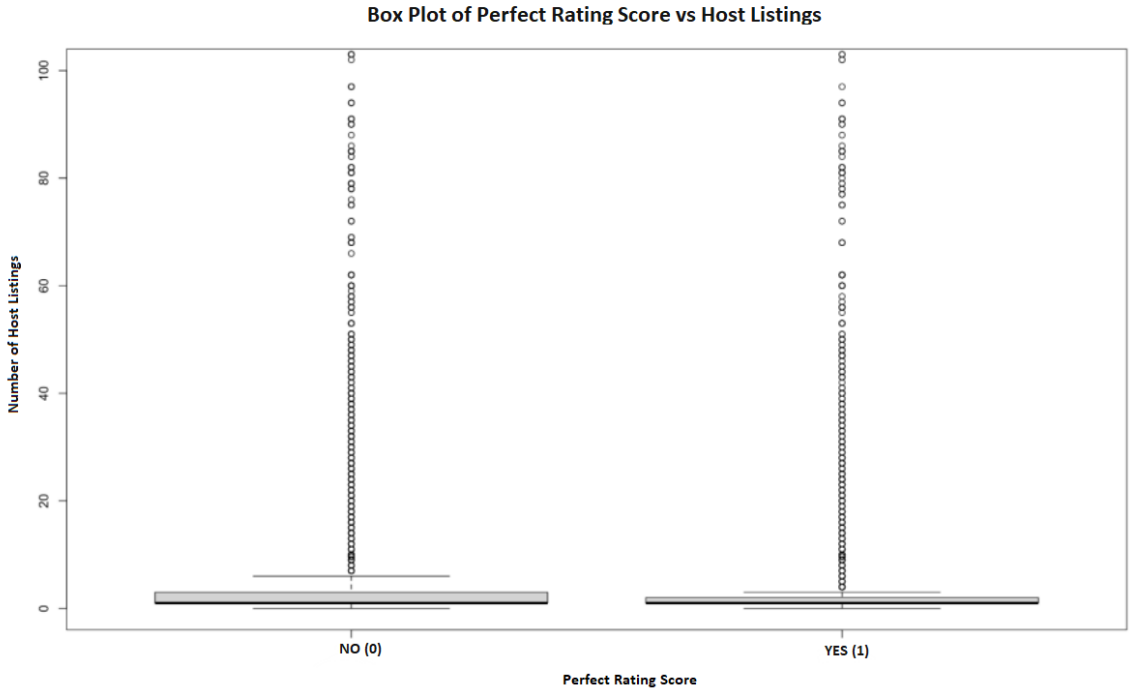

Fig. 7: Box Plot of Perfect Rating Score vs Host Listings

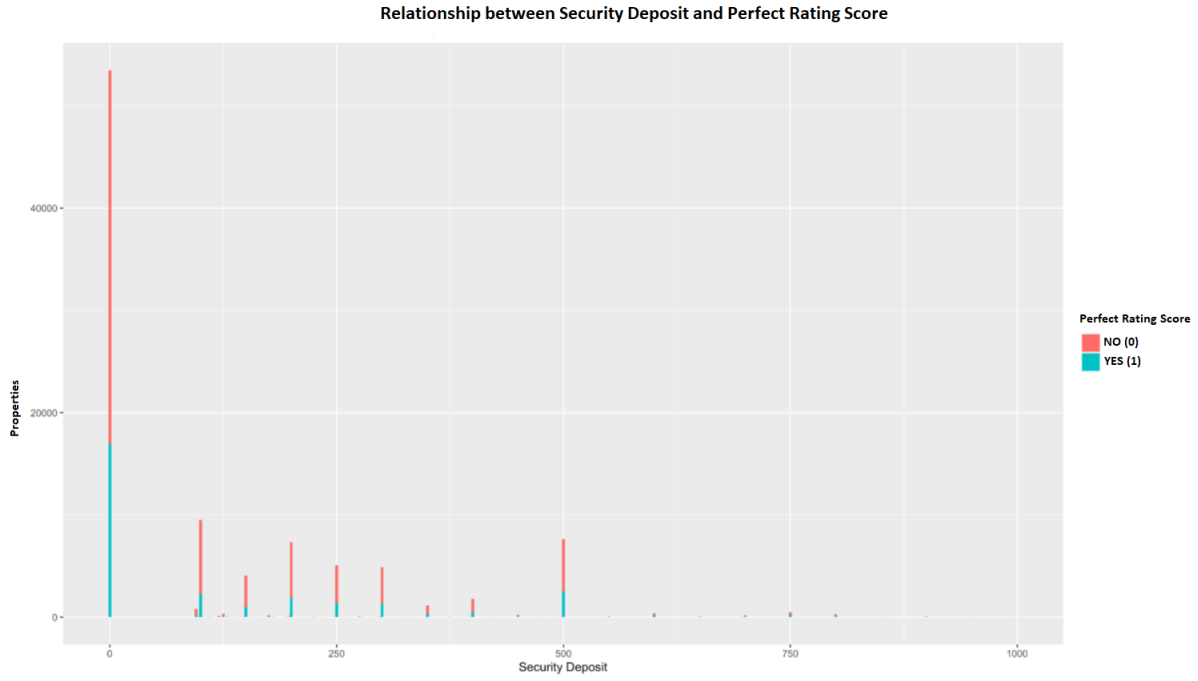

Fig. 8: Column Chart for Security Deposit and Number of Properties

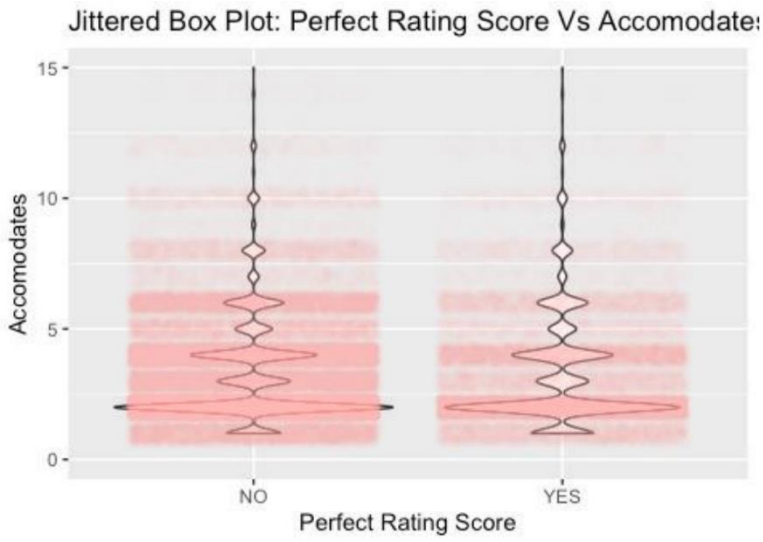

Fig. 9: Jittered Box Plot of Perfect Rating Score vs. Accomodates

Fig. 10: Jittered Box Plot of Perfect Rating Score and Accomodates

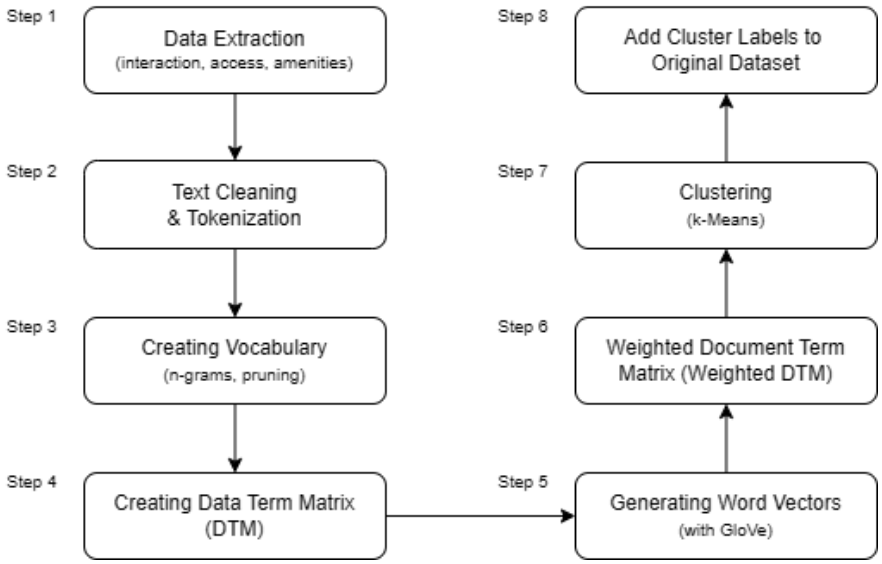

Fig. 11: Flowchart explaining feature engineering behind the text columns: 'interaction', 'access', 'amenities'

Model	True Positive Rate (TPR)	False Positive Rate (FPR)
Logistic Regression	37.24%	10.00%
Linear Regression	35.36%	10.00%
Ridge & Lasso	37.35%	9.99%
Trees	29.25%	10.60%
Random Forest	39.70%	9.19%
XGBOOST (w/o clusters)	41.52%	9.75%
XGBOOST (with clusters)	45.33%	9.55%

Fig. 12: Results of TPR and FPR for different machine learning models

8. DATA DICTIONARY

S.No.	Feature Name	Description	Feature Type
1	access	free text field describing what portion of the residence guests will be able to use	text
2	accommodates	how many guests can stay in the listing	numerical
3	amenities	list of amenties available in the listing	categorical list
4	availability_30	how many days out of the next 30 the listing is available for	numerical
5	availability_365	how many days out of the next 365 the listing is available for	numerical
6	availability_60	how many days out of the next 60 the listing is available for	numerical
7	availability_90	how many days out of the next 90 the listing is available for	numerical
8	bathrooms	number of bathrooms in the listing	numerical
9	bed_type	description of the bed	categorical
10	bedrooms	number of bedrooms in the listing	numerical
11	beds	number of beds in the listing	numerical
12	cancellation_policy	description of how strict the cancellation policy is	categorical
13	city	the actual city that the listing is in	categorical
14	city_name	the broader metro area the listing is in (for instance, city=Takoma Park, city_name = Washington DC)	categorical
15	cleaning_fee	how much, if any, is the cleaning fee the host charges	numerical
16	country	country	categorical
17	country_code	abbreviated country name	categorical
18	description	free text field written by the host describing the listing	text
19	experiences_offered	whether there are "experiences" offered with the listing (t) or not (f)	categorical
20	extra_people	additional charge for extra people in the rental	numerical
21	first_review	when the first review for the listing was written	date
22	guests_included	how many guests are included in the price of the rental	numerical

23	host_about	free text field written by the host describing themselves	text
24	host_acceptance_rate	percent of stay requests the host accepts	numerical
25	host_has_profile_pic	if the host has a visible profile picture	categorical
26	host_identity_verified	whether the host's identity has been verified using Airbnb's process	categorical
27	host_is_superhost	whether the host is a "superhost"	categorical
28	host_listings_count	how many total listings the host has	numerical
29	host_location	city, state, and country where the host is located	categorical
30	host_name	the host's first name	text
31	host_neighbourhood	more fine-grained location for the host	categorical
32	host_response_rate	percent of stay requests the host responds to	numerical
33	host_response_time	how long it takes the host to respond to requests	categorical
34	host_since	date the host joined Airbnb	date
35	host_total_listings_count	how many total listings the host has ever had	numerical
36	host_verifications	ways the host has verified their identity	categorical list
37	house_rules	free text field describing the rules in the residence	text
38	instant_bookable	whether you can instantly book the airbnb (t) or not (f)	categorical
39	interaction	free text field describing how potential clients will interact with the host	text
40	is_business_travel_ready	whether the listing is available for business travel (t) or not (f)	categorical
41	is_location_exact	whether the listing reports the exact location (t) or not (f) (usually for privacy purposes)	categorical
42	jurisdiction_names	the legal jurisdiction that the listing falls under	categorical
43	latitude	the number of degrees west of the prime meridien	numerical
44	license	whether the host has a hotelier license (t) or not (f)	categorical

45	longitude	the number of degrees north of the equator	numerical
46	market	Airbnb's definition of the market that the listing competes in	categorical
47	maximum_nights	maximum nights you can book the listing for	numerical
48	minimum_nights	minimum nights you can book the listing for	numerical
49	monthly_price	price to rent the listing for a month	numerical
50	name	short free text field describing the listing	text
51	neighborhood_overview	free text field written by the host describing their neighborhood	text
52	neighbourhood	fine-grained neighborhood name of the listing	categorical
53	notes	free text field with additional notes on the listing	text
54	price	price to rent the listing for one night	numerical
55	property_type	description of the type of dwelling the listing is in	categorical
56	require_guest_phone_verification	whether the host requires a phone number to verify the guest's ID (t) or not (f)	categorical
57	require_guest_profile_picture	whether the host requires the guest's profile picture (t) or not (f)	categorical
58	requires_license	whether the listing is in a jurisdiction that requires the host to have a license (t) or not (f)	categorical
59	room_type	description of the type of accomodation of the listing	categorical
60	security_deposit	the amount of security deposit required to rent the listing	numerical
61	smart_location	another description of the location of the listing	categorical
62	space	free text field describing the space in the listing	text
63	square_feet	how many square feet the listing is	numerical
64	state	the state the listing is in	categorical
65	street	the street address of the listing	text
66	summary	free text field summarizing the description of the listing	text

67	transit	free text field describing nearby transit options for the listing	text
68	weekly_price	price to rent the listing for a week	numerical
69	zipcode	zipcode of the listing	categorical