# Exploring the Key Drivers of Airbnb Review Scores: A Machine Learning Approach with Sentiment Integration

By Sarika Gaind

## ABSTRACT

This study aims to investigate the key drivers of Airbnb guest review scores using a machine learning approach integrated with sentiment analysis. The research tries to identify and quantify factors affecting guest satisfaction by integrating structured data, such as property attributes and host features, with unstructured guest comments. In total, several predictive models have been performed, including Random Forest, Decision Trees, K-Nearest Neighbors, and Linear Regression. Each of these required huge pre-processing, feature engineering, and sentiment analysis using VADER. Among all models performed, the Random Forest emerged as the most reliable with the lowest mean squared error and highest R-squared, thus best to explain complex relationships among variables. The results of sentiment analysis showed that positive guest sentiment was strongly associated with higher review scores, indicating the importance of emotional tone in reviews. This study combines numerical and textual data to provide actionable insights for Airbnb hosts in their quest to improve guest experiences and optimize review outcomes, with key recommendations focused on enhancing communication, cleanliness, and strategies for improving guest satisfaction.

## INTRODUCTION

Guest review scores in the listings on Airbnb work as a critical indicator of how desirable a property is and increase or decrease its booking potential. These scores reflect not just the quality of the listing, but also act as an indicator of a host's performance or success in the competitive short-term rental market. Recognizing the pivotal role these ratings play, this project focuses on identifying and quantifying the factors driving guest satisfaction, using advanced predictive modeling techniques to assess their impact on review scores. This study deploys machine learning methods, further than previous research that usually tends to be descriptive in analysis, in order to unlock the predictive capabilities of both structured data, such as host attributes and property features, and unstructured data like textual reviews, for a more comprehensive understanding of what drives guest satisfaction.

A methodical approach has been made for modeling the satisfaction of the guests using various predictive models, such as Decision Trees, Random Forest, K-Nearest Neighbors (KNN), and Linear Regression. The individual models were subjected to comprehensive cross-validation in order to assess their accuracy and robustness. The Random Forest model came out to be the most reliable predictor, with the lowest mean cross-validation mean squared error and the highest R-squared value on test data, hence really effective in capturing complex interactions among features. To enrich the analysis even more, this study incorporated sentiment analysis using VADER to extract emotional tone from guest reviews, adding a qualitative layer to the quantitative predictions. Merging this with structured features, such as host responsiveness, property location, and amenities, it can point out which of those factors are determining review outcomes.

The work accomplished in this project includes extensive preprocessing for the cleaning and normalizing of the dataset, extraction of features that are rich in predictors, and some exploratory data analysis done for finding preliminary patterns or relationships. Sentiment analysis played a vital role in decoding the subjective impressions of guests, bridging the gap between textual feedback and numerical ratings. Key insights from top positive and negative reviews, together with average sentiment scores, shed light on how aspects like communication, cleanliness, and the meeting of guest expectations align with overall satisfaction. The predictive framework was then applied to a newly introduced dataset to test its applicability, and the results provided actionable recommendations for hosts. These include enhanced guest experiences through improved communication, amenity personalization, and improvement prioritization with sentiment-driven feedback.

In all, the project delivers a full-rounded, evidence-based toolkit for Airbnb hosts, merging numerical predictors with qualitative insights. From analyzing structured to unstructured data, it puts forward a multidimensional look at guest satisfaction to let hosts proactively adapt their offerings and optimize review outcomes. The dual-focused approach not only advances knowledge about the drivers of review scores but also serves as a practical guide for better performance in listings on the competitive platform of Airbnb.

## LITERATURE REVIEW

### *Predicting sentiment and rating of tourist reviews using machine learning*[1]

In the work by Puh and Babac, machine learning models such as Naive Bayes, SVM, and Random Forest were used to address sentiment analysis and rate prediction of traveler evaluations. Their study concentrated on using factors including host traits, property characteristics, and visitor input to predict sentiment polarity and rating scores. The goal of our research, which predicts the "review_scores_ratings" for Dallas Airbnb houses, is similar in that it uses machine learning models like Linear Regression, KNN, Decision Tress, and Random Forest to identify the main elements influencing review ratings. Although the use of machine learning for prediction is emphasized in both papers, our effort goes one step further by incorporating sentiment analysis using VADER, which provides a more thorough level of analysis on the textual evaluations than Puhand Babac did.

Furthermore, our study attempts to predict evaluations more comprehensively by combining sentient and structured data, such as host traits, property parameters, and geography, whereas Puh and Babac primarily concentrated on the sentiment categorization portion. With a more complex prediction model, this all-encompassing method is a step toward closing the knowledge gap about how qualitative and quantitative factors work together to influence ratings. Furthermore, Puh and Babac primarily concentrated on the sentiment and rating prediction accuracy without conducting such a thorough analysis, but our project stresses the use of model evaluation metrics as MAE and R^2 scores, which are essential for evaluating model performance. Through data-driven insights, this comparison analysis will help us better understand how applicable the model is in the real world and help Airbnb hosts perform better.

***What do Airbnb users care about? An analysis of online review comments by Mingming Chenga, Xin Ji [2]***

Cheng and Jin's study, "What do Airbnb users care about? An analysis of online review comments," provides key insights into guest satisfaction on Airbnb, highlighting factors like location, amenities, and host interactions as primary determinants, while price plays a surprisingly minor role. Their research employs text mining and sentiment analysis to analyze unstructured review data, offering a descriptive understanding of guest preferences by identifying broad themes from online comments. While the study provides a valuable foundation for understanding qualitative aspects of guest experiences, it lacks a quantitative approach to determine how these factors impact overall review scores.

The proposed research project seeks to advance Cheng and Jin's work by adopting a predictive modeling approach, focusing on how specific factors—such as host responsiveness and property amenities—affect review scores for Dallas Airbnb properties. By utilizing a structured dataset and conducting a more detailed analysis, this project aims to quantify the impact of individual variables on guest satisfaction. In doing so, it not only builds on the themes identified in Cheng and Jin's work but also provides more actionable insights for hosts to improve guest experiences and enhance their ratings.

***Airbnb listings' performance: determinants and predictive models [3]***

With an emphasis on criteria like property attributes, host traits, and location, the article "Airbnb listings' performance: determinants and predictive models" offers a thorough examination of the variables that affect Airbnb listing performance. Key predictors of visitor satisfaction and performance indicators are identified through empirical analysis in this study. It reveals that the success of an Airbnb listing is significantly influenced by characteristics such as property type, room quality, host responsiveness, and area attractiveness. To provide hosts and platform administrators with useful information for improving listing quality and guest happiness, the study also investigates how predictive modeling approaches, such as regression analysis and machine learning models, might be used to forecast performance outcomes.

The essay also emphasizes how useful sophisticated computer models are for comprehending intricate relationships between different factors, which enables more precise listing performance forecasts. The paper highlights the potential of data-driven approaches in the short-term rental market by analyzing the advantages and disadvantages of various models, offering insightful information to practitioners in the business as well as to scholars. To provide a thorough knowledge of what propels success in the Airbnb ecosystem, this work emphasizes the significance of combining diverse data sources and predictive analytics. It also offers a framework that can be tailored to different markets and environments.

***Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis [4]***

Using machine learning (ML) and sentiment analysis, Alharbi (2023) created a sustainable price prediction model for Airbnb listings, highlighting the significance of predictive pricing tactics for Airbnb hosts. In order to predict listing prices, the study used a dataset from Barcelona and a variety of machine learning (ML) algorithms, such as Lasso regression, Ridge regression, Bayesian regression, and support vector regression (SVR). The study found that while amenities had little effect on pricing, features like the number of bedrooms, maximum accommodation capacity, and review sentiment polarity had a significant impact.

Model performance was improved by integrating qualitative review data into the predictive framework through the use of sentiment analysis via TextBlob. With a remarkable R square score of 99%, Ridge and Lasso regression beat other models, proving their effectiveness in managing complicated datasets. The study gave Airbnb hosts practical advice on how to strategically price their listings while preserving sustainability and competitiveness by identifying important price determinants

The study also emphasized the wider applications of machine learning in tackling issues in peer-to-peer marketplaces and the sharing economy. In contrast to conventional pricing models, sentiment analysis's incorporation provided a fresh viewpoint by

***Rating prediction of peer-to-peer accommodation through attributes and topics from customer review* [5]**

The article by Subroto and Christianis (2021) delves into rating prediction mechanisms for peer-to-peer (P2P) accommodations, specifically focusing on Airbnb in Indonesia. The authors analyze customer reviews and listing attributes to understand factors influencing customer satisfaction and ratings. Employing a dataset of over 55,000 reviews, the study applies machine learning models such as Classification and Regression Tree (CART), Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO) Logistic Regression, Artificial Neural Network (ANN), and Multi-Layer Perceptron (MLP). The findings indicate that customer review attributes significantly impact rating predictions, with ANN achieving the highest accuracy (84.79%) when attributes were considered. Furthermore, topic-based features yielded relatively lower accuracy, with RF achieving 60.09%, demonstrating the value of attributes over topics in predictive performance.

This study adds to the existing research, on peer-to-peer accommodations by introducing a defined approach to examining customer reviews and forecasting satisfaction levels. The research emphasizes the importance of customer focused information in guiding decisions for platforms such as Airbnb. By employing machine learning for analysis within the sharing economy sector this study showcases its ability to improve customer satisfaction and operational effectiveness. This research closes the divide between customer feedback and numerical rating assessments by providing guidance on how businesses can utilize data analysis, for making informed decisions. The increasing significance of intelligence, in customer actions and improving service provision within the fiercely competitive P, TO P accommodation sector is also underscored.

***Machine Learning: Algorithms, Real-World Applications and Research Directions*** *[6]*

It covers all aspects of ML algorithms, from simple to complex, applied in real-life problems and also gives a view of future directions. In the context of the current project, this paper puts great emphasis on structured and unstructured data integration, which will drive actionable insights-a principle followed in this project, where the numerical attributes are combined with sentiment analysis. The review has also highlighted the importance of ensemble methods, such as Random Forests, in handling feature interactions effectively, thus justifying the inclusion of Random Forest in this study for the prediction of review scores on Airbnb. Sarker's discussion of nonlinear methods and their applications in personalized recommendations resonates with the project's goal of understanding nuanced guest satisfaction factors through models like Decision Trees and sentiment-driven insights.

Besides, Sarker mentions the increasing role of sentiment analysis in making sense of user behavior across domains and thus strengthens the utility of VADER and TextBlob for the derivation of qualitative insights from comments left by the guests. The paper further emphasizes the use of strong evaluation metrics like R-squared and MAE, which are used to evaluate model performance in this project comprehensively. Further, the work of Sarker on real-world challenges regarding data imbalance and model interpretability underlines the need to address the skewed ratings of the Airbnb dataset and to choose models that predict accurately yet provide actionable recommendations for the hosts. The current study is grounded on these very fundamental principles and illustrates how machine learning frameworks can be tuned for guest satisfaction optimization within the sharing economy.

## METHODOLOGY

### Sentiment Analysis of Comments'

To that end, the guest comments had sentiment analysis conducted using VADER, or Valence Aware Dictionary and sentiment Reasoner, an effective tool designed to quickly assess sentiment in short texts like reviews. VADER assigns a compound sentiment score to each comment, indicating, based on the score, the intensity level from negative to positive. The preliminary sentiment analysis showed that the majority of comments fell under the category "Positive," which agrees with the high average "review_scores_rating" and suggests generally positive guest experiences.

Further text preprocessing was done to make the sentiment analysis more qualitative and relevant. First, lemmatization was performed on the comments. It is a process that reduces words to their base or root form; hence, different forms of a word, such as "running" and "ran," are reduced to "run." Lemmatization improves the analysis in terms of standardizing the words, which helps in capturing the true sentiment without being misled by different word forms.

Besides, the frequency filtering technique was used to refine the processed text. It consisted of removing the top 5% most frequent and the bottom 5% least frequent terms in the comments. The point of this step is to remove the noise from the very common words that carry little meaning, such as "good," "nice," "awesome," and rare outliers that are unlikely to provide any useful insights. We took away these extreme terms so that the sentiment analysis focused on more relevant words that may be more representative of the true sentiment of the guest experience.

Together, lemmatization and frequency filtering smoothed out the text, so the sentiment analysis would not get biased by either irrelevant or redundant terms. This gives a more refined approach to the guests' perceptions of the properties and hence increases the relationship between the sentiment scores and the actual review ratings.

The VADER sentiment analysis tool determines the sentiment of text by scoring individual words and phrases based on their emotional intensity. It calculates negative, neutral, and positive scores, which represent the proportion of the text that corresponds to each of those sentiments, and combines them into a compound score ranging from -1 to +1. These are labeled as Positive, Neutral, or Negative based on the compound score. A compound score above 0.05 means positive. The VADER system also takes into consideration some contextual modifiers, such as punctuation, capitalization, and intensity amplifiers. For instance, "Cute modern stay" or "Very clean, well decorated" were considered Positive because their compound score is very high with positive phrasings like "great", "recommend", and "comfortable." This also illustrates the point of the entire review where, once positive language has dominated a text, its overall classification leans toward positivity, with accurate breakdowns into negative, neutral, and positive proportions.

**Input and target variables**

In our dataset, we have multiple columns. Out of them review_scores_rating which represents average overall rating score from guest reviews is considered as the target variable. And from the rest of the columns, we found out that 10 columns are relevant to target variables reducing multicollinearity and provided description below.

| Column Names | Description |
|---|---|
| accommodates | The maximum capacity of the listings. |
| bedrooms | The number of bedrooms in the listing. |
| beds | The number of beds in the listing |
| comments | Guest comments or reviews for the property |
| review_scores_accuracy | Rating for accuracy of the listing description |
| Number_of_reviews | Number of reviews for the property |
| Host_total_listings_count | The total Number of listing the host has. |
| review_scores_communication | Rating for communication with the host |
| price | Daily price in local currency. |
| minimum_nights | minimum number of night stay for the listing |

| | |
|---|---|
| maxmum_nights | maximum number of night stay for the listing |
| availability_365 | The availability of the listing 365 days in the future as determined by the calendar |
| Compound_sentiment | Cumulative sentiment score score given to a comment |

TABLE 1

**Data Preprocessing**

The preprocessing workflow was an essential part in setting up the dataset for feature selection and training the model. This consists of the following key stages of the workflow:

i. Text Preprocessing for Sentiment Analysis: The cleaning and normalizing of the comments column was done with a custom function. It handled non-string values, tokenized the text, put it in lowercase, removed non-alphanumeric tokens and stopwords, and lemmatized the remaining tokens down to their base forms. Then, the processed tokens were joined into one cleaned text string.
ii. Removing Extreme Terms: After tokenization and filtering out stop words, the comments were further processed to remove the top 5% most frequent and the bottom 5% least frequent terms to remove noise from overly common words and rare outliers and enhance the focus on meaningful terms.
iii. Feature Engineering:The feature matrix included key numerical and categorical variables such as property characteristics and review-related metrics. The target variable was the review score rating. Any missing values in the predictor variables were imputed to ensure complete data for model training.
iv. Feature Selection:To enhance model efficiency and relevance, the top 5 features most correlated with the target variable were selected. This step helped reduce the dimensionality of the dataset and focus on the most important predictors.

v. Normalization:The selected features were scaled to a consistent range, ensuring that all features contributed equally to the model by addressing differences in their scale.

vi. Data Splitting:The dataset was split 70% for training and 30% for testing. The splitting was done to enable the validation of the models, and the results will be reproducible.

**Models Implemented**

i. Linear Regression: It is a statistical model which is used to find the relationship between independent and dependent variables by fitting a linear equation. It provides coefficients for the independent variables to minimize the error in predicting target variable. It helps us in identifying which variables has linear relationship with our target variables 'review_score_rating'. It provides meaningful insights and makes it easier to identify which variable have an influence on target variables but at the same time it cannot capture complex and non-linear relationships.

ii. K-Nearest Neighbors: It is a non-parametric and instance-based algorithm.It predicts the target variable by identifying k nearest neighbors in the feature space. The distance between the listings is calculated using Euclidean distance. It can capture non-linear patterns in the data. For example, listing of same room types and amenities may have similar review scores because they have non-linear relationship between them. The only disadvantage is that it is too expensive in terms of large datasets and can be sensitive for noisy data.

iii. Decision tree: Decision trees are supervised learning models where it splits the data into subsets based on their feature values and hence creating a tree structure in which each nodes contains a decision rule. The model makes predictions by traversing from root node to leaf node where each path represents a decision for predicting review_score_rating. Decision trees can be useful in identifying non-linear relationships and complex relationships between variables. They reveals hierarchical dependencies between attributes like certain host attributes combined with property features impact review scores.

But the disadvantage in decision tree is they can prone to overfitting which may require tuning and hyper optimization.

iv. Random forest: It is an ensemble learning algorithm that is combination of multiple decision trees and it aggregates their predictions. Each tree in the forest is trained using a subset of features, the final decision is the average(regression) or majority vote(classification) of all the trees. Random forest is highly effective for prediction tasks with complex datasets because it captures nonlinear relationships and reduces overfitting by averaging multiple decision trees. It can be very useful for our research question because it can handle interaction between host attributes, property features which helps in identifying strong influencers of review scores.

**Evaluation Metrics**

To find out the best model of all the models, metrics like Cross-Validation MSE (Mean Squared Error):, Test MSE and R-squared are used.

i. Cross-Validation MSE is the average of the squared differences of the predicted and actual review scores calculated over cross-validation. It means the data is divided into several subsets, and on different subsets, the model is trained and validated. This gives insight into how well the model generalizes across different parts of the data. A lower Cross-Validation MSE means the model generalizes better to unseen data and thus provides more stable and reliable predictions.

ii. Test MSE: Test MSE is computed over the test set, which is the portion of data that one retains for evaluation after training the model. It gives an indication of the model's ability to generalize on new data. The lower the test MSE, the closer will be the model's predictions to the real values of the test set, reflecting better accuracy and a high predictive power.

iii. R-squared: R-squared is a statistical metric that reflects the proportion of variance in review scores explained by the model. It gives an indication of the goodness of fit for the model: the higher the R-squared, the better it is, since it captures most of the variation in the review scores. An R-squared value closer to 1 indicates that the model does a very good job of explaining the variability in the data, while a lower value suggests that the model is less effective at fitting the data.

**Relationship Study of Target and Compound Sentiment Score**

The compound score is the overall positivity/negativity of guest comments, derived using VADER sentiment analysis. Its relationship to the target variable, review_scores_rating, was examined as a way of determining how the expressed sentiments by guests align with their numeric ratings.

## DATA ANALYSIS PLAN

### Overview of Exploratory Data Analysis (EDA)

The EDA analyzes the dataset's primary numerical and categorical properties to determine how they are distributed, relate to one another, and affect the target variable, "review_scores_rating." To find any patterns, trends, or possible abnormalities in the data, this will entail creating descriptive statistics and visualizations. The EDA seeks to identify elements that influence property ratings and to learn more about how satisfied customers are as shown by the reviews.

### Descriptive Statistics Of Target Variable

An overview of the rating distribution is given by the first descriptive statistics for "review_scores_rating." The majority of properties often receive high ratings, ranging from 4.32 to 5.00, with a mean rating of 4.84 and a comparatively low standard deviation of 0.13, suggesting generally positive reviews. The rating distribution is probably skewed toward the upper end, according to these figures. which would limit the model's variability for prediction purposes but might also suggest great overall satisfaction.

```
count    4500.000000
mean        4.802367
std         0.187560
min         3.250000
25%         4.750000
50%         4.850000
75%         4.920000
max         5.000000
Name: review_scores_rating, dtype: float64
```

FIGURE 1

**To enhance the analysis, the following visual aids were employed**

**Histogram of "review_scores_rating"** [Figure 1]**::** This histogram helps us comprehend the general distribution of ratings across properties and validates the key found in the descriptive statistics. It is easier to interpret the consistency and quality of the evaluations when one is aware of this distribution.

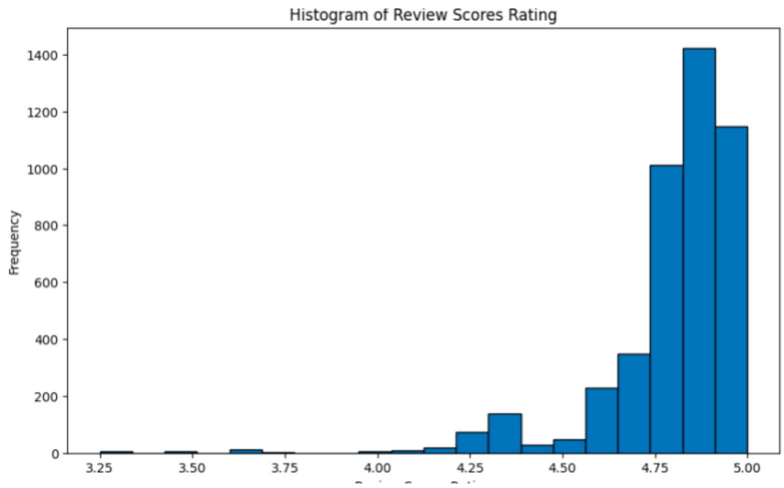


FIGURE 2

**Numerical variable correlation heatmap**[Figure 2]**:** This type of heatmap shows how characteristics relate to one another and to the target variable. This method made it easier to see any multicollinearity problems and comprehend how various features affect "review_scroes_ratings." Stronger correlations between the features and the target variable are then taken into account for additional analysis and model improvement.

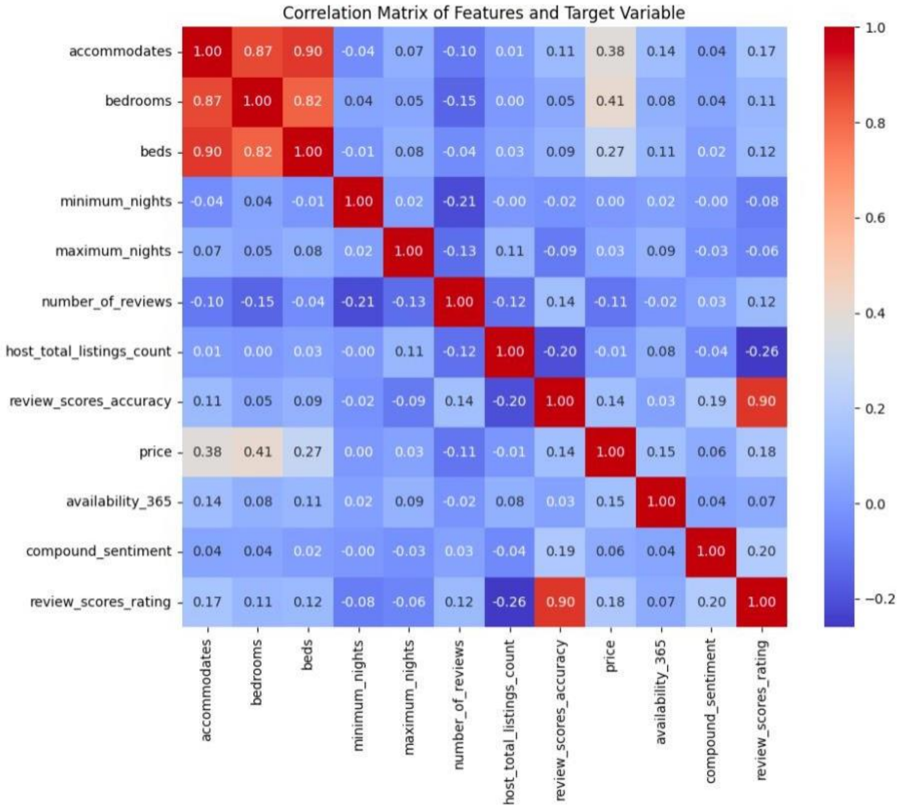


FIGURE 3

**Word Cloud for Comment Sentiment Analysis** [Figure 3]**::** To visualize the most commonly worst and best used terms in customer evaluations, a word cloud is created in the "comments" column. This graphic will provide information on popular subjects or features of the properties that visitors find interesting. By further classifying input into positive or negative feelings using the "comments" column, sentiment analysis will enhance the quantitative analysis by providing a qualitative knowledge of consumer feedback.


FIGURE 4

**RESULTS**

In fact, the most accurate model in predicting review scores of Dallas Airbnb is the Random Forest model, which had a mean cross-validation MSE of 0.000296 and a test MSE of 0.000945, with an R-squared value of 0.973484. This shows that it is better at modeling complex relationships between predictor variables such as host characteristics, property features, and guest sentiments. The performance was also very good from the Decision Tree model, which, in some respects, outperforms with a mean CV MSE of 0.000518, but slightly higher in test MSE of 0.001004 and a high R-squared of 0.971808, placing Random Forest at the best in predictions. Although the KNN model was equally efficient, with a mean CV MSE of 0.002497 and R-squared of 0.968633, it was not that strong in handling nonlinear relationships. Linear Regression had the poorest performance, probably due to its linear assumptions, with a mean CV MSE of 0.006675 and R-squared of 0.822106.

| | Model | Cross-Validation MSE | Test MSE | R-squared |
|---|---|---|---|---|
| 0 | KNN | 0.002497 | 0.001117 | 0.968633 |
| 1 | Decision Tree | 0.000518 | 0.001004 | 0.971808 |
| 2 | Random Forest | 0.000296 | 0.000945 | 0.973484 |
| 3 | Linear Regression | 0.006675 | 0.006337 | 0.822106 |

TABLE 2

Sentiment analysis, through VADER, was crucial in understanding how guest sentiment was related to review scores. Positive sentiment strongly correlated with higher scores, as shown in the top five reviews with compound sentiment values greater than 0.99, reflecting highly satisfied guests praising hosts and accommodations. On the other hand, the lowest five reviews had compound values less than -0.90 and reflected significant dissatisfaction due to poor communication, cleanliness, or unmet expectations. Overall, the average sentiment value was 0.7587, indicating positive feedback and reinforcing that maintaining guest satisfaction is important. This integration of features such as property attributes with sentiment data allowed for nuanced and actionable insights that equipped hosts with strategies to enhance their offerings. The best performing Random Forest model was then used to predict scores for a newly introduced dataset, showing its robust generalization to future data.

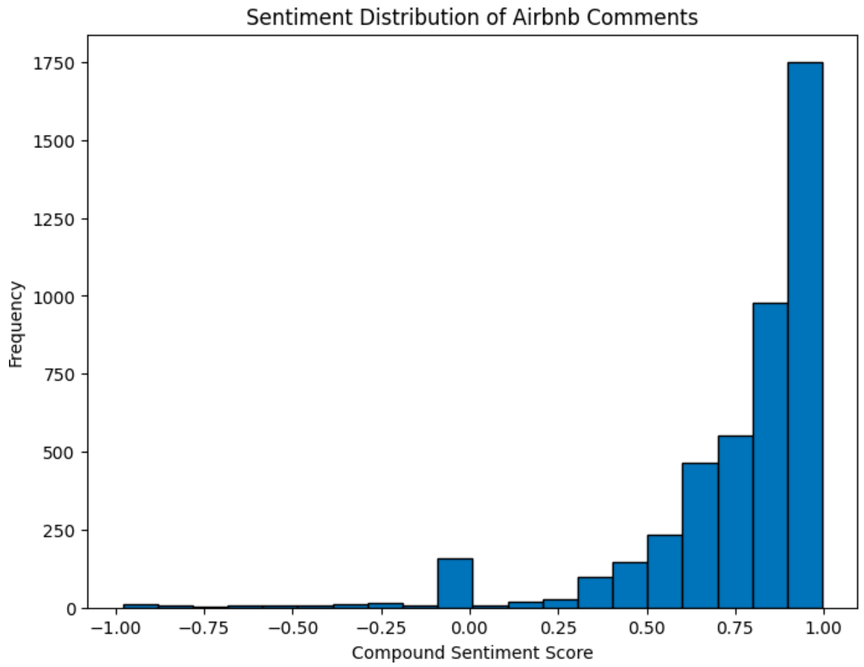


FIGURE 5

The correlation coefficient is 0.2033 for review scores versus compound sentiment score, which describes a weak positive relationship. In other words, as the sentiment score increases-that is, derived from guest reviews-the review score tends to increase a little, but not strongly.

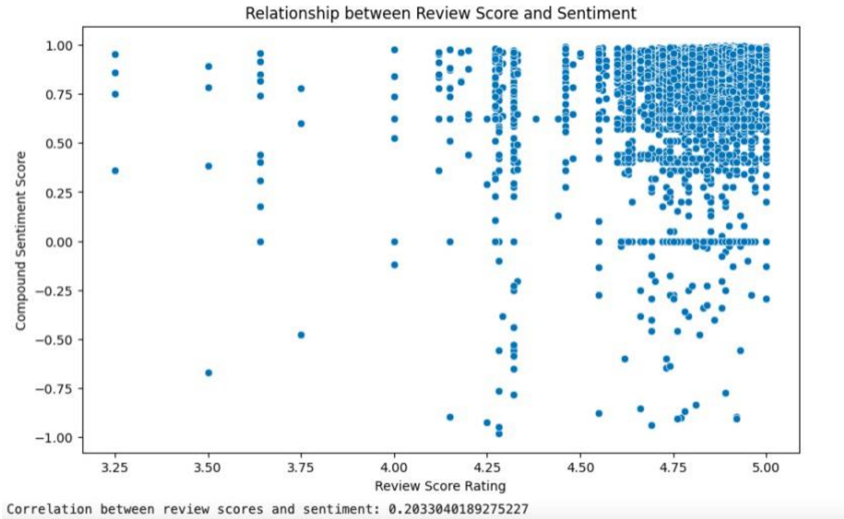


FIGURE 6

Positive Correlation: The fact that the correlation is positive suggests that guest sentiment, as measured by tools like VADER, does align somewhat with review scores. Positive comments are generally reflected in higher ratings, and negative comments in lower ones.

Weak Magnitude: The weak magnitude of the correlation, being close to zero, suggests that sentiment is not a dominating factor in review scores. Other factors such as property features, responsiveness of hosts, or location may be more influential.

Possible Reasons for Poor Correlation:

Subjectivity in Scores: Guests might give high scores despite some critical feedback in their reviews, since ratings are influenced by cultural norms or personal biases.

Linguistic Nuances: VADER and similar sentiment analysis tools, although very helpful, can fail to pick up subtleties of language or sarcasm that could lead to less precise sentiment scores.

Other Influences: Structured features such as price, amenities, or even host interactions might mask the impact of textual sentiment on review scores.

The Random Forest model was used to predict review scores for a new dataset introduced. The Random Forest model was trained and tuned before being tested on the new dataset, showing great predictive performance with high accuracy. By effectively combining structured data with sentiment analysis from unstructured text, the model provides actionable insights for hosts, allowing them to predict guest satisfaction and enhance their offerings to obtain higher review scores.

## CONCLUSION

This study is a comprehensive investigation of the drivers of the review scores left by guests on Airbnb by integrating structured data, like property attributes and host characteristics, with unstructured text data represented by comments left by guests. Property features related to accommodation capacity, responsiveness of the host, and communication ratings have been identified as key significant factors. The analysis revealed that the relationship between `review_scores_rating` and the compound sentiment score is a weak positive correlation of 0.203. Although this is not a strong predictor in and of itself, it suggests that the emotional tone of the feedback being positive is coupled with a higher rating, further emphasizing the qualitative impressions. Among all machine learning models explored, Random Forest was the best capable of capturing the interaction among features nonlinearly; hence, it had the best R-squared and minimum mean squared error.

These insights give actionable wins for Airbnb hosts: the need for clear communication, maintaining high standards of cleanliness, and acting on specific guest feedback. This may result in increased satisfaction and, therefore, improved ratings. For Airbnb as a platform, this could be even further improved by implementing tools that automatically analyze guest reviews for personalized suggestions on how hosts can improve.

Despite these strengths, there were a number of limitations to this study. First, because of computational resources, the advanced deep learning models like RoBERTa could not be employed, which would, in turn, have enabled deeper insights into the sentiment and nuances of the language used by guests in reviews. The analysis used VADER, a rule-based approach for sentiment analysis, which is effective yet may lack contextual subtlety when compared to transformer-based language models. Second, the dataset is robust but comes from a single geographical region, namely Dallas; therefore, there may be limits to generalizing results across cities or countries due to a mix of guests' expectations and cultural influence. This might be done in future research with higher computational power to incorporate models like RoBERTa or BERT for richer insights from the textual data. Secondly, the dataset can be extended to include listings in several regions to enable comparative analysis that can help in the validation of the generalizability of the findings.

Furthermore, it would be interesting to extend the current predictive framework by considering other types of unstructured data, such as images of properties or audio reviews. Finally, temporal trend analysis-such as how satisfaction might vary seasonally with guests, or how reviews themselves change over time-provides a dynamic perspective that complements the understanding of factors influencing ratings.

## REFERENCES


[1] Puh, K., & Babac, M. B. (2022). Predicting sentiment and rating of tourist reviews using machine learning. Journal of Hospitality and Tourism Insights, 6(3), 1188–1204. https://doi.org/10.1108/jhti-02-2022-0078

[2] Cheng, M., & Jin, X. (2017). What do Airbnb users care about? An analysis of online review comments.TourismManagement,60,99–111. https://doi.org/10.1016/j.tourman.2017.01.019

[3] Fakhrutdinova, E., Wendt, N., & Lyu, J. (2024). Airbnb listings' performance: determinants and predictive models. European Journal of Tourism Research, 32(2), 2142-518. https://www.ejtr.vumk.eu/index.php/about/article/view/2142/518

[4] Alharbi, Z. H. (2023). A sustainable price prediction model for Airbnb listings using machine learning and sentiment analysis. Sustainability, 15(17), 13159. https://doi.org/10.3390/su151713159

[5] Subroto, A., & Christianis, M. (2021). Rating prediction of peer-to-peer accommodation through attributes and topics from customer review. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-020-00395-6

[6] Sarker, I. H. (2021). Machine learning: algorithms, Real-World applications and research directions. SN Computer Science, 2(3). https://doi.org/10.1007/s42979-021-00592-x

[7] Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

[8] Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in Machine learning Algorithms. Lecture Notes in Networks and Systems, 15–25. https://doi.org/10.1007/978-3-031-35314-7_2

[9] Grefenstette, G. (1999). Tokenization. In Text, speech and language technology (pp. 117–133). https://doi.org/10.1007/978-94-015-9273-4_9