**Predictive Modeling and Analysis of Factors Affecting Airbnb Ratings in Oakland**

Team members: Siqing Xu, Amber Zhu, Chang Su, Jiatong Liu, Siyuan Ma

# **Introduction**

## **Motivation**

Nowadays, a sharing economy like Airbnb has reshaped the hospitality industry, which creates a platform where hosts and travelers can connect. However, with countless listings, understanding what drives guest satisfaction and high ratings is essential for the properties to stand out. The ratings and reviews not only influence the visibility of listings but also guide guests to make informed decisions.

We seek valuable insights that create a fair and more user-friendly Airbnb community. Our analysis focuses on identifying the key factors that influence ratings for Airbnb listings in Oakland. By analyzing the data, we aim to help hosts improve their listings and provide guest recommendations for better experiences by addressing aspects that affect the reviews and ratings. Also, the findings can also assist Airbnb platform in enhancing the recommendation algorithm tailored to guest preferences.

## **Database**

We obtain the [Oakland Airbnb dataset](https://insideairbnb.com/get-the-data/) from the official Airbnb open data website. Listing is our primary file. it consists of 2410 Oakland listings, and each contains core information about house amenities such as room type, bedroom count, bathroom count, price, and descriptions; host information such as host response rate, acceptance rate, total number of listings, and superhost identification; Also, the review subscores for accuracy, cleanliness, communication, location and so on. Review file contains 123088 entries, each containing the listing id, comment date comment, reviewer, and written comments.

# **Data Processing**

To prepare the data for analysis, we first select relevant columns intuitively and remove columns with excessive missing data (containing more than 5 missing values). Then we identify the columns with numerical and object data types to fill in missing values with np.nan and NA respectively. Next, we standardize data types for consistency. We handle rate percentages into decimals, map boolean (‘t’ to 1 and ‘f’ to 0), and remove dollar signs for the price. In addition, we perform a sentiment analysis on the guest reviews to obtain an average sentiment score (-1 for most negative and 1 for most positive) for each listing. Finally, we merge the sentiment score and the listing file to obtain a comprehensive dataset that includes important features that can be used to identify listings with high ratings. By randomly splitting the data into 70% training and 30% testing, we aim to build robust models and deliver meaningful insights for all stakeholders.

# **Natural Language Processing**

## **NLP Analysis Method**

The NLP analysis began with thorough text preprocessing, including HTML removal, tokenization, and stopword elimination. We then applied TF-IDF vectorization to capture the nuanced importance of words across listings. This technique weighs words based on both their frequency in individual listings and their uniqueness across the entire dataset, providing deeper insights than simple word counting.

## **Key Findings in NLP**

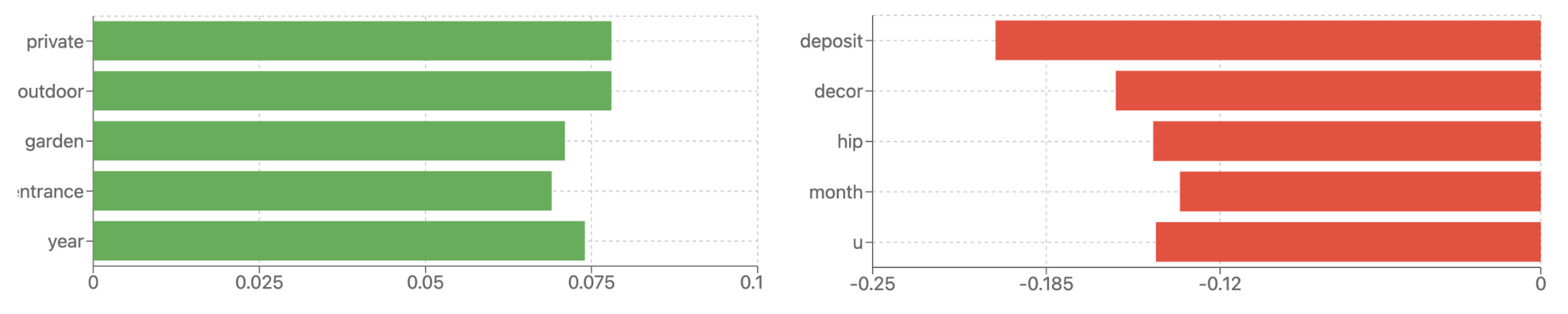


Figure 1. Positive impact words (the left) and Negative impact words (the right)

* **Positive Impact Words** (Figure 1, the left): Words emphasizing *permanent features* and *outdoor spaces*, such as “private,” “outdoor,” and “garden,” showed strong positive correlations with high ratings.
* **Negative Impact Words** (Figure 1, the right): Words related to *financial arrangements* (e.g., “deposit”) and generic terms negatively impacted guest satisfaction.

# **Models**

## **Linear Regression Model**

The linear regression model was constructed using structured features such as price, number of bedrooms, host response rate, and review scores across different dimensions. The dataset was preprocessed to remove multicollinearity by calculating the Variance Inflation Factor (VIF) and selecting significant predictors using p-values.

Table 1. The performance of the Linear Regression model

| Model | R-Squared | Adjusted R-Squared |
| --- | --- | --- |
| Linear Regression | 0.794 | 0.791 |

The high R-squared value indicates that the model explains 79.4% of the variance in review scores, demonstrating a strong fit. The important features (table 1) are Host Response Rate, Communication Ratings, Bedrooms, Location Ratings, Price Sensitivity, Check-in Experience.

## **Logistics Regression Model**

To classify Airbnb listings as "useful" based on a binary target variable (1 if the review score rating is >= 4.89, 0 otherwise) and to evaluate the significant predictors for achieving high ratings. The logistic regression model used the same features as the linear regression model, focusing on structured data combined with review subscores to capture the nonlinear relationship between features and high ratings.

Table 2. The performance of the Logistics Regression model

| Model | Pseudo R-Squared | Accuracy | TPR | FPR |
| --- | --- | --- | --- | --- |
| Logistics Regression | 0.3871 | 0.869 | 0.894 | 0.158 |

While the logistic regression model performs well on classification metrics, the relatively low pseudo-R-squared value suggests the potential benefit of using more sophisticated models like Random Forest or Gradient Boosting for nonlinear relationships.

## **CART Models**

A decision tree is a non-parametric supervised learning algorithm that recursively splits data based on feature values, forming a tree structure where nodes represent features, branch decisions, and leaves predictions.

Table 3. The performance of the CART (Decision Tree Classifier) model

| Model | Accuracy | TPR | FPR | Precision |
| --- | --- | --- | --- | --- |
| 1 Baseline DTC NLP | 0.574790 | 0.561290 | 0.410526 | 0.597938 |
| 2 DTC NLP CV | 0.586555 | 0.445161 | 0.259649 | 0.650943 |
| 3 DTC CV | 0.658824 | 0.600000 | 0.277193 | 0.701887 |
| 4 DTC Subscores CV | 0.840336 | 0.809677 | 0.126316 | 0.874564 |

\* random state being 42; ‘NLP’ Methods are introduced and the words are used as predictors; Cross Validation (CV) and tuning on parameters are introduced; Result was found based on ratings on dimension-specific review scores (SUBSCORES).

* **Baseline DTC NLP** (table 3): The model is designed using room description and rating text extracted by Natural Language Processing (NLP) as features. The model tries to capture the detailed text features as much as possible. The results show the limitation of relying on text features alone.
* **DTC NLP CV** (table 3): The model is designed based on a baseline model with fine-tuning applied, and it significantly increased the difficulty of computation given a slight improvement.
* **DTC CV** (table 3): The model introduces structured features such as the number of rooms, the number of beds, the number of bathrooms, the price, and the geographic location, sentiment scores while excluding words as natural language features. Feature importance analysis shows that Number of Reviews, Average Sentiment Score, Host Total Listings Count, etc. are important factors.
* **DTC Subscores CV** (table 3): The model combines structured features with detailed ratings of listings (e.g., cleanliness, location, quality of communication with the landlord, etc.) in the advancement of model 3. The significant factors indicating that Review Scores Accuracy, Review Scores Cleanness, Review Scores Communication are dominant factors.

## **Random Forest and Boosting Models**

Random Forest and Boosting are applied to enhance score prediction, with Random Forest reducing overfitting and Boosting optimizing errors, tested on DTC model variants.

Table 4. The performance of the Random Forest and Boosting model

| Model | | Accuracy | TPR | FPR | Precision | ROC AUC |
| --- | --- | --- | --- | --- | --- | --- |
| 1 DTC | Random Forest | 0.7076 | 0.7258 | 0.3123 | 0.7166 | 0.7794 |
| Gradient Boosting | 0.6891 | 0.7065 | 0.3298 | 0.6997 | 0.7769 |
| 2 DTC Subscores | Random Forest | 0.8639 | 0.8516 | 0.1228 | 0.8829 | 0.9482 |
| Gradient Boosting | 0.8706 | 0.8548 | 0.1123 | 0.8923 | 0.9448 |
| 3 Logistics Regression | Random Forest | 0.8756 | 0.8677 | 0.1158 | 0.8907 | 0.9462 |
| Gradient Boosting | 0.8639 | 0.8452 | 0.1158 | 0.8881 | 0.9411 |

With the same features included, Model 2 and Model 3 give similar results and measurements as Decision tree and logistic regression.

# **Final Model and Key Factors**

We chose ‘Random Forest DTC Subscores’ as the final model. There are two main reasons:

* **Best Performance**: Random Forest has the best performance with the highest accuracy, TPR, AUC, and lowest FPR. The introduction of Subscores significantly improved the accuracy and reliability of the model.
* **Robustness**: Random Forest is more robust in processing multi-dimensional features. Even though the Gradient Boosting model has equivalent performance, it’s more sophisticated in processing detailed features.

We counted the Top 20 feature importance in the model which was shown by the graph (Figure 2). The blue bars represent the Random Forest. The key factors influencing high ratings are review\_scores\_accuracy, review\_scores\_value, review\_scores\_cleaniness, review\_scores\_communications, review\_scores\_checkin and review\_scores\_location.

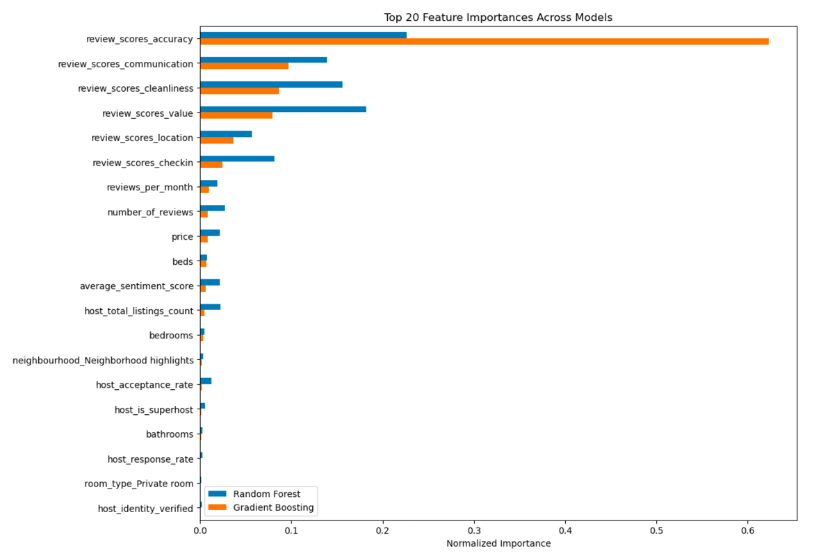


Figure 2. Top 20 feature importances across models

# **Recommendations**

## **For Hosts to Improve their Listings**

1. **Improve Information Accuracy**:
   * Ensure listing descriptions are precise and up-to-date regarding amenities, location, and property details.
   * Use high-quality photos and highlight unique features to reduce guest uncertainty.
2. **Enhance Value for Money**:
   * Perform competitive pricing analysis to align with market expectations.
   * Upgrade infrastructure: offer high-quality mattresses, modern bathrooms, high-speed Wi-Fi, and essential hygiene products.
3. **Focus on Cleanliness**:
   * Implement rigorous cleaning protocols and communicate cleanliness standards to guests.
   * Regularly deep clean rooms and replace bedding after every stay.
4. **Provide Exceptional Service**:
   * Respond promptly to inquiries (within 24 hours).
   * Offer flexible check-in options using electronic locks or self-check-in systems.
   * Share check-in guides, local recommendations, and emergency contacts.
5. **Encourage Positive Reviews**:
   * Motivate guests to leave reviews through small gestures like thank-you notes or follow-up messages.

## **For Airbnb to Improve its Platform**

1. **Enhance Recommendation Algorithms**:
   * Prioritize listings with high scores in accuracy, cleanliness, communication, and value for money.
   * Use model insights to personalize recommendations based on guest preferences.
2. **Improve Platform Standards**:
   * Implement stricter host verification processes.
   * Offer Airbnb-recommended cleaning services to maintain hygiene standards.
   * Conduct periodic random inspections and display results to enhance transparency.
3. **Optimize the Feedback System**:
   * Use NLP algorithms to analyze guest reviews and provide automated suggestions for improvement to hosts.
   * Simplify the feedback process for guests, focusing on key features like cleanliness, communication, and check-in experience.

# **Appendix**

Video:

<https://youtu.be/YaG2Hd6bAHk>

Slide: (named INDENG242 project slide.pdf):

<https://drive.google.com/drive/u/0/folders/1vj_ul8aXlt_KfRvaFpI-qdT5a8IvuWCj>

Code:

PDF：

<https://drive.google.com/file/d/1T-SM_6pdLWujRQYq7fworbyO9lV0eCG6/view?usp=sharing>

Google colab:

<https://drive.google.com/file/d/1_MLcfwqiKNaUqshfpfSoWmZwEwHNWS4m/view?usp=sharing>