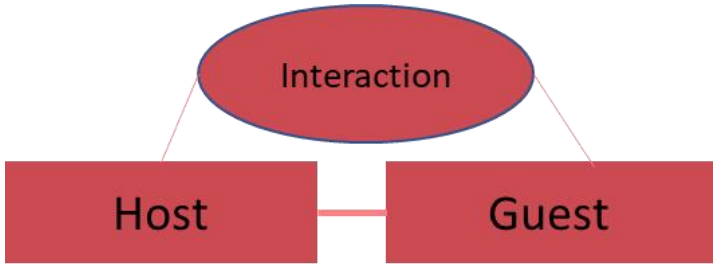
Uncovering Key Factors in Airbnb Guest Satisfaction: A Data-Driven Analysis of Ratings

Yunzhen Wu

Background

Platform introduction



- A structure of host-customer interaction
- Emphasize "physical environment" and “human interactions”
- Airbnb's review and rating system

- Competition among hosts within the platform grows.
- Rent short term lodging
- Short-term rentals has reshaped the way travelers choose accommodation.

Problem Statement

Research Question:

- What are the most important **factors** Airbnb users care about when choosing short-term rentals, and how do these **factors** affect their ratings?

Problem Statement:

- Guests struggle to find suitable listings
- Hosts need insights to improve ratings and bookings
- Misalignment leads to negative reviews and revenue loss

Literature Review

Big Data And Data Science Methods For Management Research (George et al., 2016)

- Big data analytics helps analyze Airbnb user behavior beyond traditional surveys. By examining reviews and ratings, key factors like cleanliness, location, and host responsiveness can be identified, optimizing listing quality and service.

Identifying Salient Attributes of Peer-to-Peer Accommodation Experience (Tussyadiah & Zach, 2016)

- Airbnb guest satisfaction is driven by **service**, **facilities**, **location**, **feeling welcome**, and **comfort**. Hosts' responsiveness, cleanliness, proximity to attractions, and a home-like ambiance significantly impact ratings.

What Drives Purchase Intention on Airbnb? (Chen & Chang, 2018)

- High ratings, numerous reviews, **detailed descriptions**, and **high-quality images/videos** enhance trust and booking likelihood. Hosts should optimize listing content to attract more guests.

Exploratory Data Analysis - Data Overview

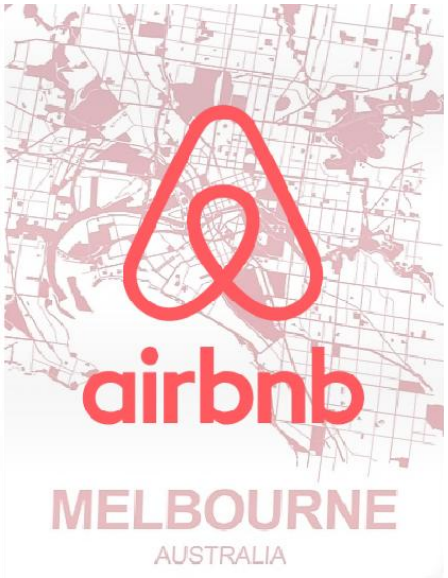
Dataset: Melbourne Airbnb listings (collected Dec 7, 2018) from Kaggle

Size: 22,895 rows, 84 columns

Data Includes: Host characteristics, pricing, availability, guest ratings

Preliminary Data Cleaning:

- Removed irrelevant columns (e.g., ID, listing URL, host name)
- Consolidated review scores into a single variable
- Verified data integrity (no duplicates)



Exploratory Data Analysis - Missing Value Analysis

Columns with high missing values:

- **License:** 99.9% missing
- **Monthly price:** 91.74% missing
- **Weekly price:** 88.98% missing

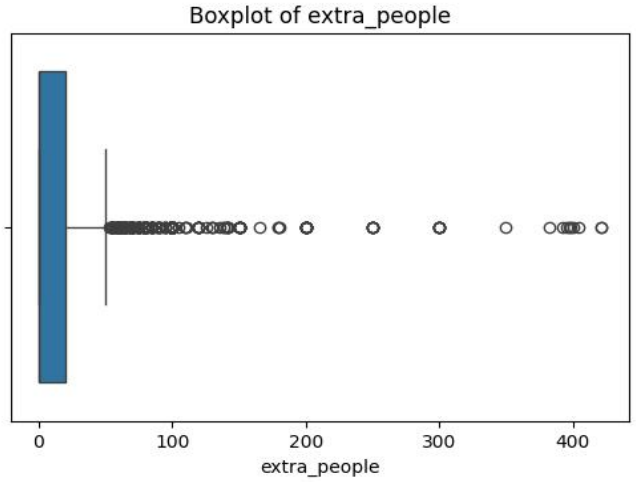
Review Scores: ~25% missing (some listings have no guest ratings)

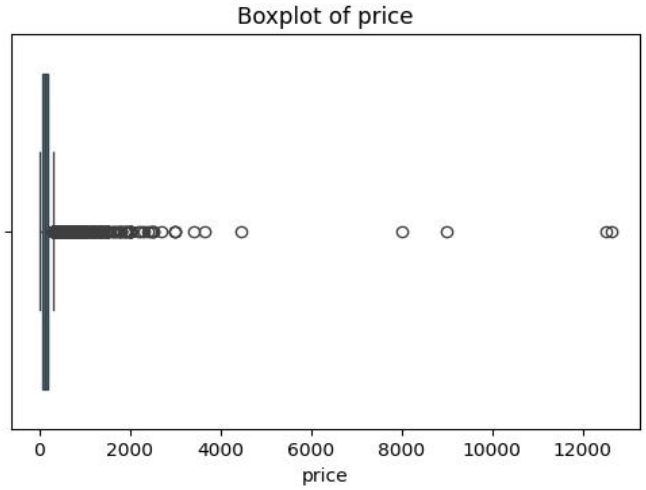
Missing values would be handled through column removal or imputation in the following part.

Exploratory Data Analysis - Outlier Detection

Method: Used Interquartile Range (IQR) method

- **Price Outliers:**
  - Listings priced **above $10,000 per night** (potential ultra-luxury or errors)
- **Extra People Fee Outliers:**
  - Some listings charge **$400+ per extra guest**, exceeding typical rates

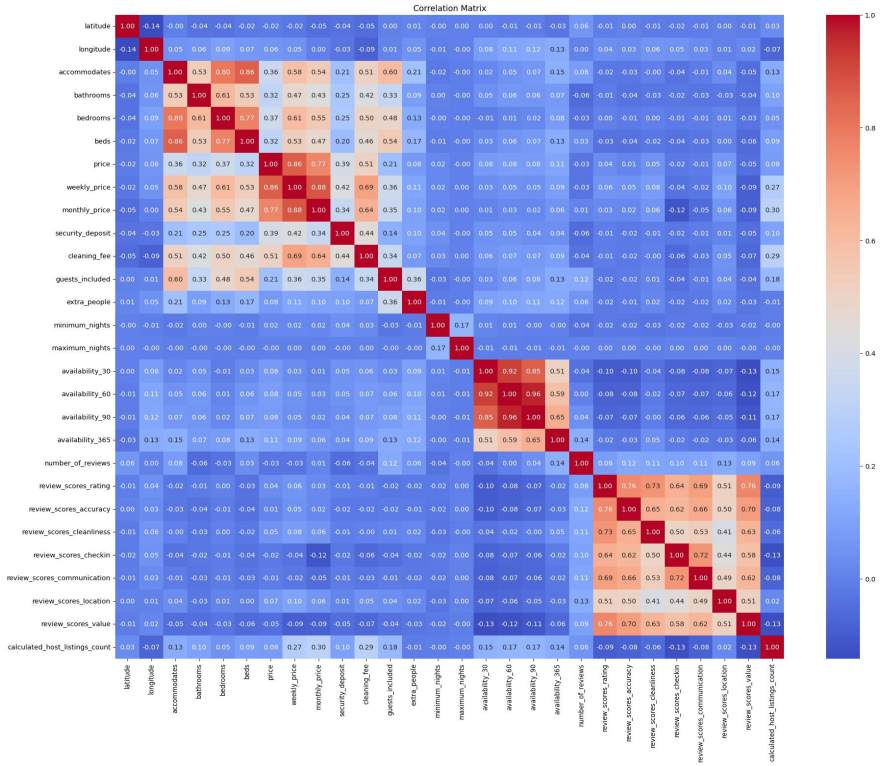




Exploratory Data Analysis - Correlation Analysis

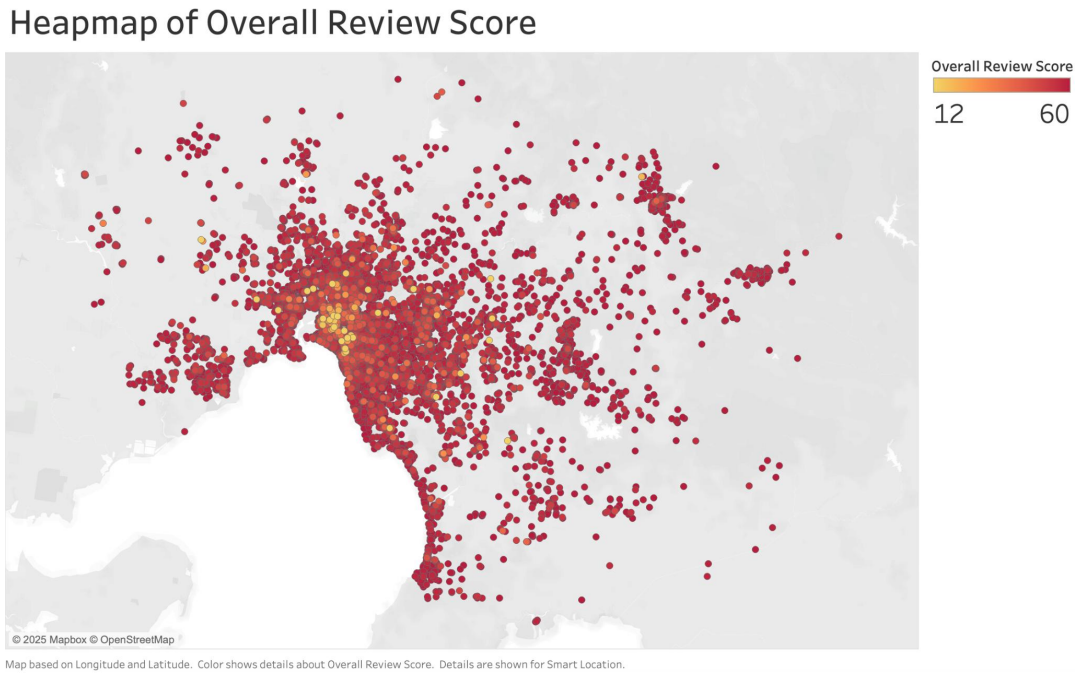
Strong correlations found between:

- **Accommodates, bedrooms, bathrooms, beds** (0.53 - 0.86) → Larger properties host more guests
- **Price, weekly_price, monthly_price** → Strong correlation in pricing structure
- **Availability (30, 60, 90 days)** → Highly correlated (0.85 - 0.96), suggesting redundancy



Exploratory Data Analysis -Spatial Distribution of Review Scores

- Heatmap analysis of Airbnb listings using **Tableau**.
- **Low-rated listings** cluster in the city center.
- **High-rated listings** are more spread in suburbs.
- Possible reasons: noise, congestion vs. better space and amenities.



Data preparation - Handling Missing Values

Dropping Entire Columns:

- Columns with high missing rates removed (e.g., *'license', 'monthly_price'*).

Dropping Rows with Missing Review Scores:

- Rows with missing review scores removed (listings without review scores were deleted.)

Imputing Missing Values:

- Text columns filled with empty strings (*'summary', 'description*').
- Other text columns filled with 0 (*'transit', 'neighborhood_overview')*.

Data preparation - Handling Outliers

To ensure data consistency, extreme values in numerical variables were clipped to remain within valid ranges.

Clipping Price:

- To avoid situations where prices are set at free or close to no rentals, a minimum clip set at 10 was placed in order to get rid of prices that are set lower than 10.

Clipping Accommodates:

- Minimum guest capacity set to 1 to prevent invalid values.

Clipping Other Numerical Variables: *"bathrooms", "bedrooms", "beds", "security_deposit", "cleaning_fee", "extra_people","number_of_reviews"*

- These columns were clipped to 0, restricting the negative values that were placed into the columns.

Data Preparation
- Removing Highly Correlated Variables

High correlation leads to redundancy and collinearity issues.

Variables removed:

- 'availability_30', 'availability_60', 'availability_365' (highly correlated with each other).
- 'beds' (correlated with 'bedrooms').
- 'weekly_price' and 'monthly_price' (correlated with 'price').

Feature Engineering

Creating New Features:

- Created ***distance_to_city_center*** using the Haversine formula between each property and Melbourne's city center (-37.8136, 144.9631), replacing latitude, longitude, and smart_location.
- Created ***overall_review_score*** by aggregating 6 review ratings.
- Two additional numerical features: ***summary_length*** and ***description_length*** to measure listing detail richness.
- Transformed *host_since* into ***host_duration_days*** by calculating the total days since the host's registration, capturing the effect of host experience on guest reviews.

Feature Engineering

Binary Encoding for Text Variables:

- Converted descriptive text fields into binary variables (0/1).
- Encoded columns: *' transit', 'notes', 'neighborhood_overview', 'house_rules'*.
- Ensures machine learning models can process categorical descriptions efficiently.

One-Hot Encoding for Categorical Variables

- Used drop_first=True to prevent multicollinearity in models.

Standardization of Non-Target Variables

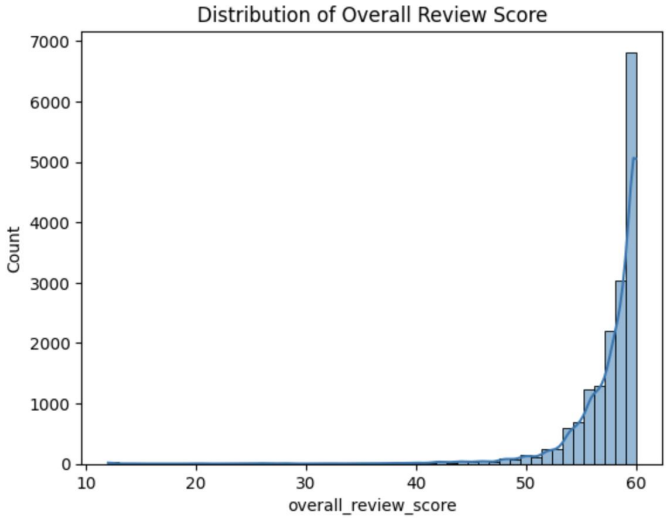
- Features were Z-score normalized using **StandardScaler.**
- Ensures features have similar scales for better model performance.
- Prevents models from being biased towards large-magnitude features.

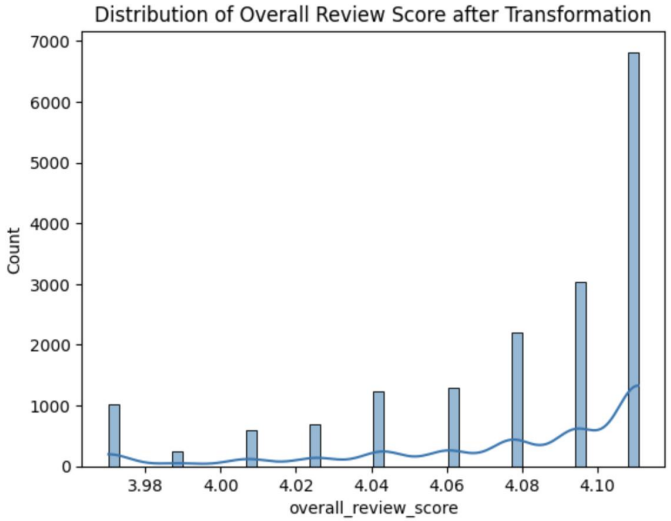
airbnb

Feature Engineering

Target Variable Transformation

- **Outlier Mitigation:** Applied **winsorization** at the 5th and 95th percentiles to cap extreme values of overall_review_score, preventing skewed ratings from distorting the model.
- **Distribution Normalization:** Performed a log transformation to further reduce right skewness, ensuring the target variable aligns better with linear model assumptions.





Methodology and Deployment

Tooling and Environment: Google Colab A100 GPU

- **Linear Regression - Baseline**
- **ElasticNet Regression**
- **XGBoost**
- **Random Forest**

Methodology and Deployment

- **Hypertuning**

  ElasticNet:

  Best Alpha: 0.000297; Best L1 Ratio: 0.7

  XGBoost:

  'subsample': 0.7, 'n_estimators': 800, 'min_child_weight': 5, 'max_depth': 4,

  'learning_rate': 0.01, 'gamma': 0, 'colsample_bytree': 0.9

  Random Forest:

  'max_depth': 30, 'min_samples_leaf': 10, 'min_samples_split': 2, 'n_estimators': 300

- **Cross Validation of 5 fold**

Methodology and Deployment

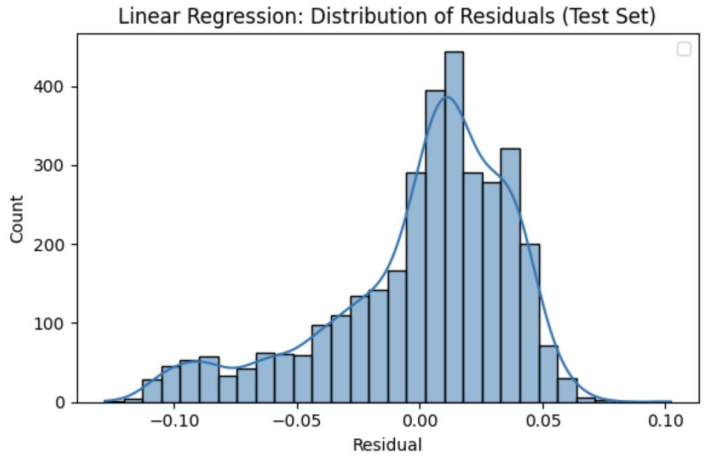
Evaluation:

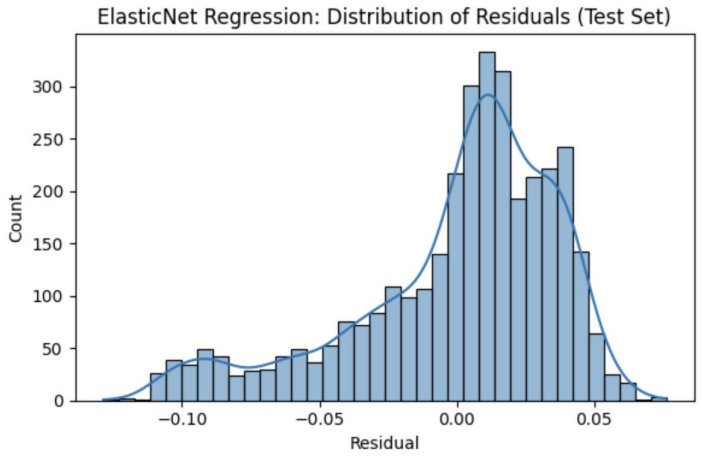
- R2
- MSE
- True vs Predicted values
- Residual Analysis

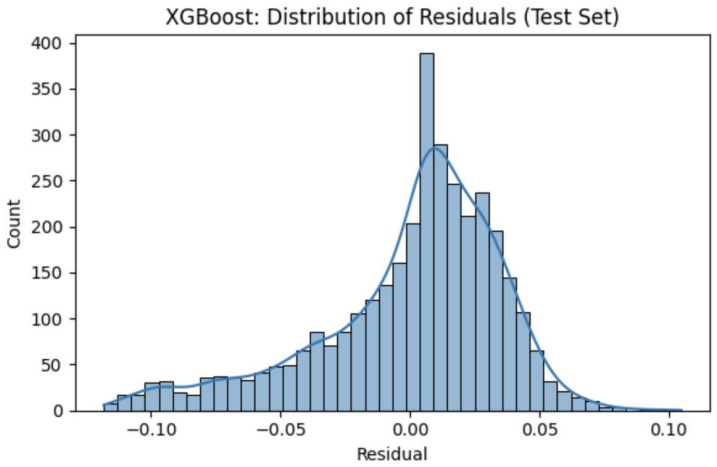
Factor Analysis:

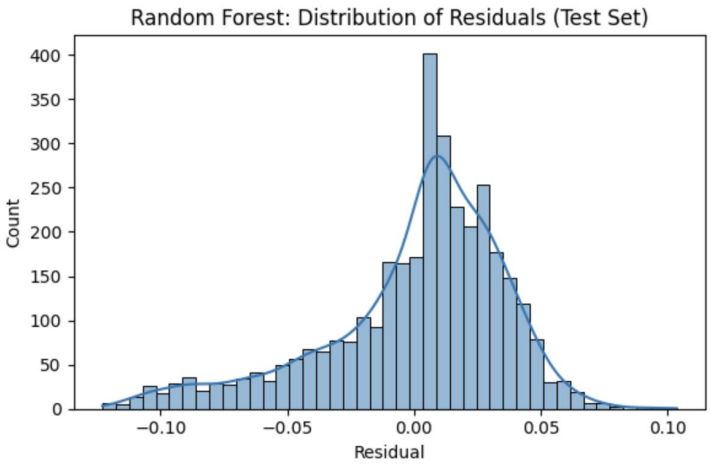
- Feature Importance

Results - Residual Analysis



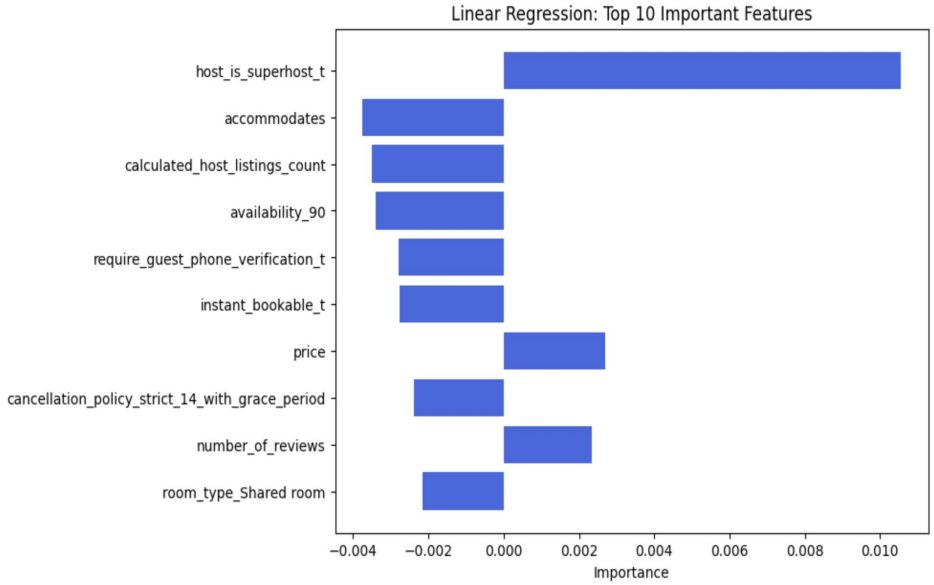


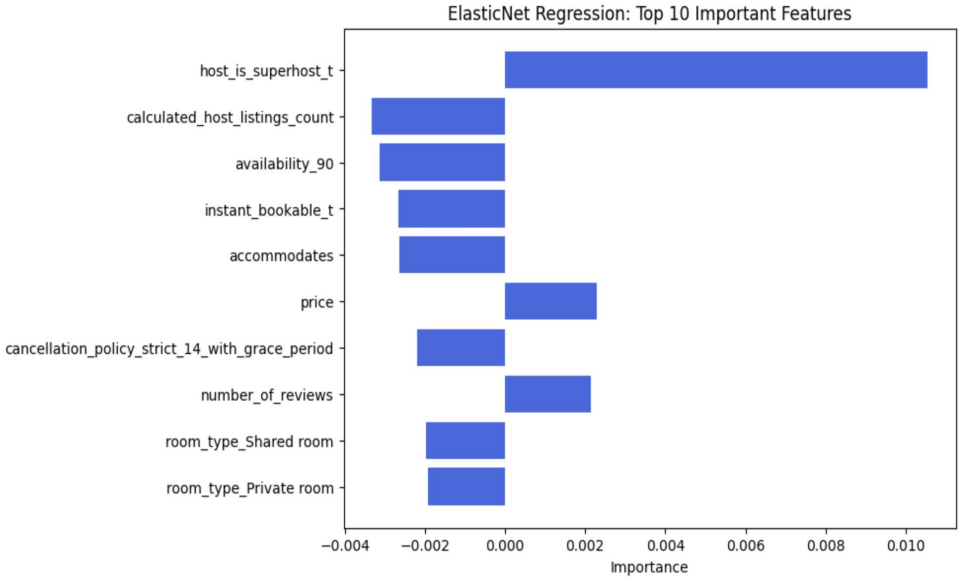


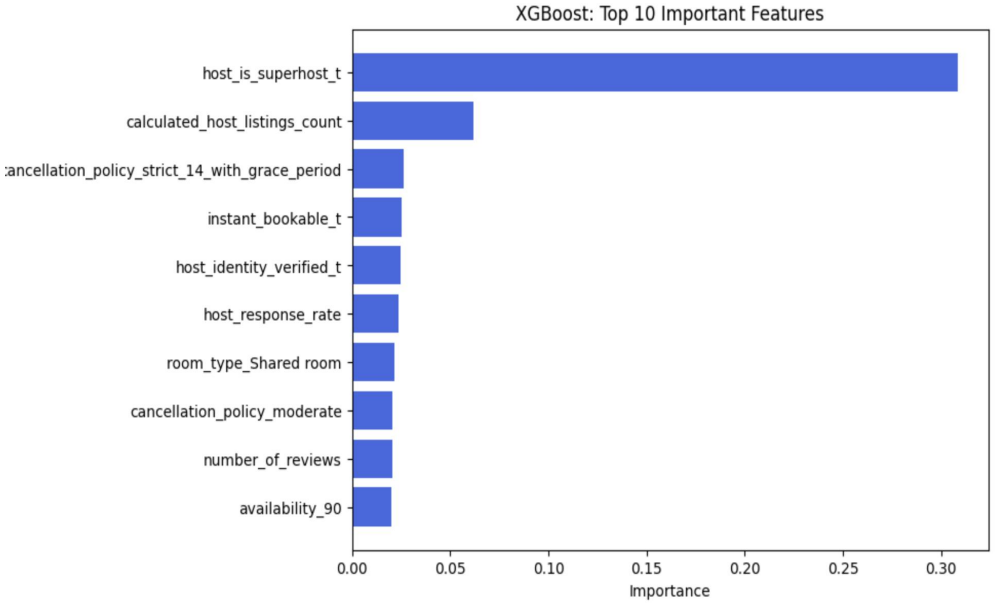


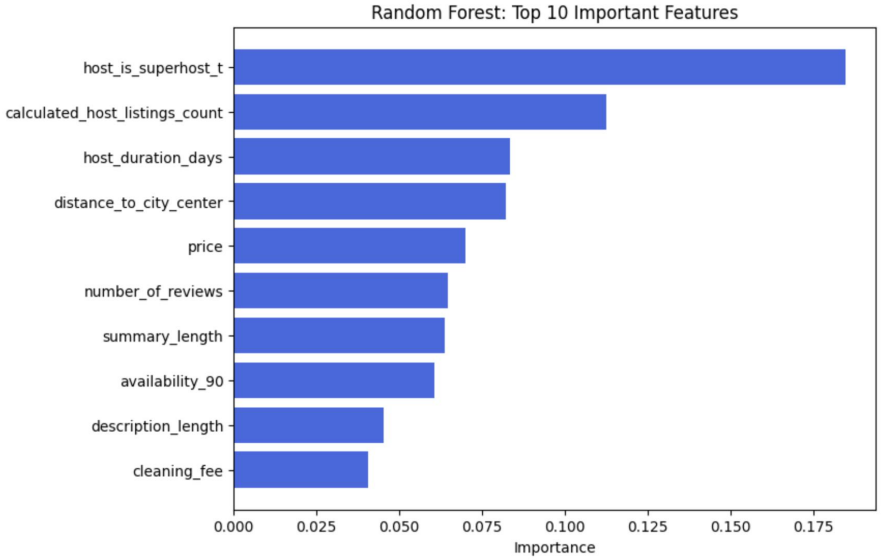
- All models' performances are not bad, with residuals concentrated around 0.
- All models exhibit slight left skew, indicating a tendency to slightly over-predict lower ratings or under-predict higher ratings.
- **Tree-based models** perform better, reducing prediction errors.

Findings - Feature Importance



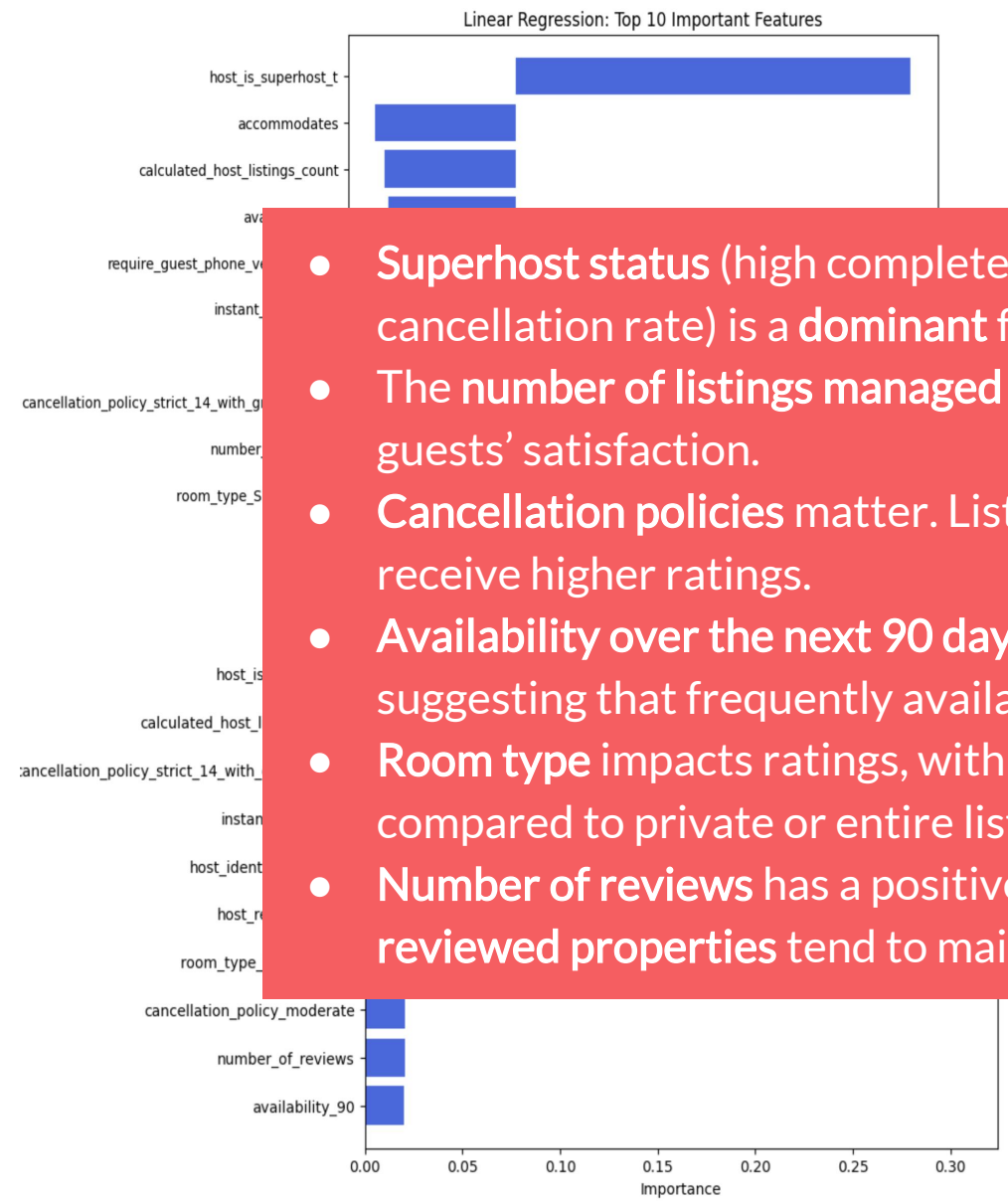


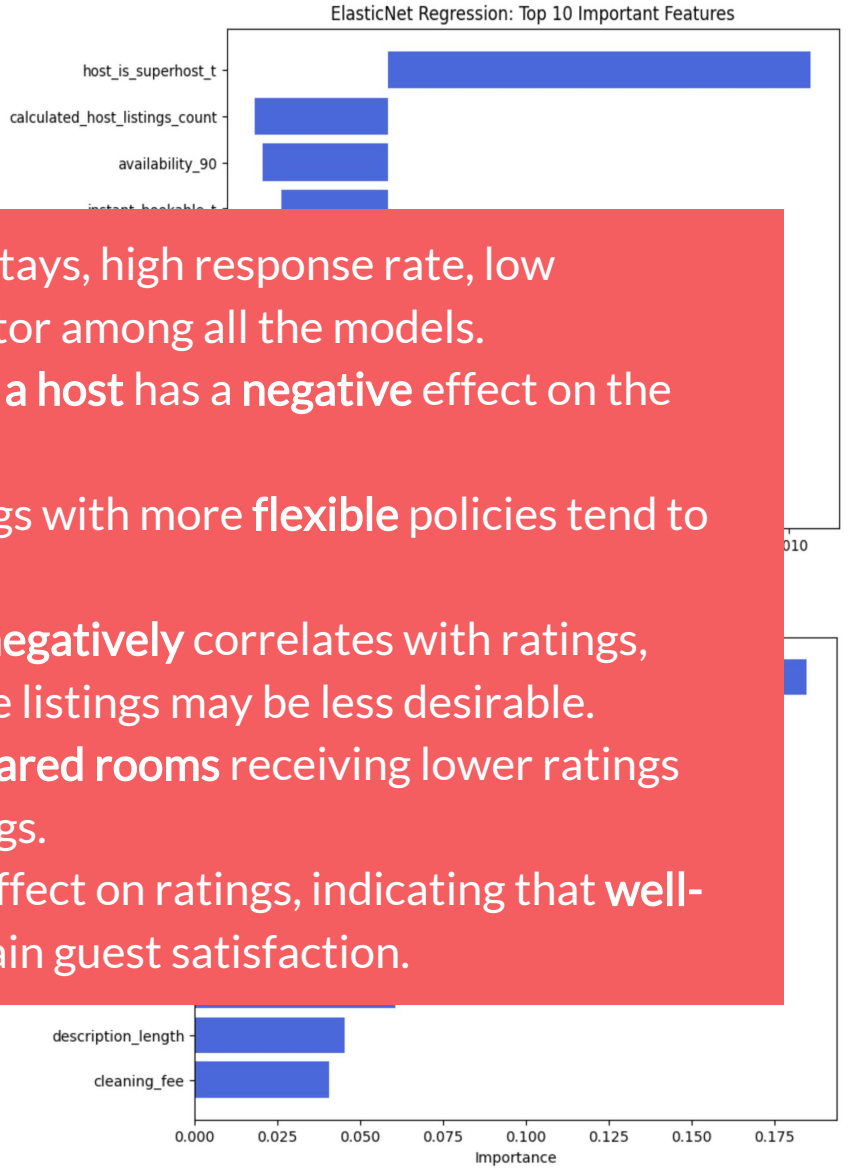




airbnb

Findings - Feature Importance





- **Superhost status** (high completed stays, high response rate, low cancellation rate) is a **dominant** factor among all the models.
- The **number of listings managed by a host** has a **negative** effect on the guests' satisfaction.
- **Cancellation policies** matter. Listings with more **flexible** policies tend to receive higher ratings.
- **Availability over the next 90 days negatively** correlates with ratings, suggesting that frequently available listings may be less desirable.
- **Room type** impacts ratings, with **shared rooms** receiving lower ratings compared to private or entire listings.
- **Number of reviews** has a positive effect on ratings, indicating that **well-reviewed properties** tend to maintain guest satisfaction.

airbnb

Conclusion

Model Performance Summary:

- **XGBoost** and **Random Forest** achieved the best predictive performance, outperforming ElasticNet and Linear Regression.

Key Factors from Feature Importance:

- Superhost status, number of listings managed by a host, availability_90, room type, number of reviews, and cancellation policies.

Recommendations for Airbnb Hosts:

- Aim for **Superhost status** by ensuring high-quality service, and quick response times.
- **Limit the number of managed listings** to ensure consistency in service.
- **Avoid listing shared rooms**, which are rated lower than private or entire spaces.

Limitations

- **Feature Limitations:** The dataset lacks **guest demographics**, booking purpose, and stay duration.
- **External Factors** Excluded: The impact of local events and neighborhood safety was not incorporated.

Future Directions

- Try **other techniques**, such as Sentiment Analysis & NLP, Deeper Learning Models (MLPs, transformers), clustering (K-Means, DBSCAN).
- **Find External Datasets**, like crime rates, tourism statistics, weather conditions, and real estate trends to assess their impact on guest ratings.

References

Chen, Chih-Chien, and Yu-Hsiang Chang. 2018. "What Drives Purchase Intention on Airbnb? Perspectives of Consumer Reviews, Information Quality, and Media Richness." *Telematics and Informatics* 35 (5): 1512–23. https://doi.org/10.1016/j.tele.2018.03.019.

George, Gerard, Ernst C. Osinga, Dovev Lavie, and Brent A. Scott. 2016. "Big Data and Data Science Methods for Management Research." *Academy of Management Journal* 59 (5): 1493–1507. https://doi.org/10.5465/amj.2016.4005.

Tussyadiah, Iis P., and Florian Zach. 2016. "Identifying Salient Attributes of Peer-to-Peer Accommodation Experience." *Journal of Travel & Tourism Marketing* 34 (5): 1–17. https://doi.org/10.1080/10548408.2016.1209153.