AirBnb - Broward County, Florida

Application of Artificial Intelligence (AI) Models
SCH-MGMT 661

PROJECT PRESNTATION

Aishwarya Anandan
Vipin Kumar Karthikeyan
Varshini Kuppusami
Puneeth Rao Gunuganti

UMassAmherst | Isenberg School of Management

Predicting Revenue & Optimizing Guest Experience in Broward County

Why Broward? 17,000+ listings. High-stakes tourist market (Ft. Lauderdale) where competition is fierce.

Problem 1: **Price Prediction** (Revenue): Hosts struggle to value their properties correctly. We aim to predict fair market price using advanced signals (amenities, location). (Listings Dataset)

Problem 2: **Sentiment Analysis** (Quality): Pricing is only half the battle. We apply NLP to guest reviews to understand what drives positive sentiment and high ratings. (Reviews Dataset)

Data Cleaning & Formatting

1. **Type Conversion:** Standardized non-numeric data by stripping symbols from currency (`$150` to `150.0`) and percentages (97% to `0.97`), and mapping boolean strings to `True/False(host_is_superhost', 'host_identity_verified')`
2. **Target Processing:** Cleaned the target variable **`price`** by removing zero/missing values and applying a **Log Transformation** to normalize the skewed distribution.
3. **Outlier Management:** Capped extreme price outliers using the **1st and 99th percentiles** to prevent ultra-luxury listings from distorting the model.

Imputation Strategy

1. **Structural Imputation**: Filled missing `bedrooms`, `beds`, and `accommodates` using **Grouped Medians** (grouped by Room Type and Neighborhood) to preserve local market variance.
2. **Consistency Checks (Beds/Accomodates):** Reconciled data logic, such as imputing missing `beds` counts based on `accommodates` capacity where applicable.
3. **Host Response Imputation:** Imputed categorical policies like `host_response_time` using the mode within specific host cohorts (e.g., Superhosts vs. non-Superhosts).

Feature Engineering – Property & Host Signals

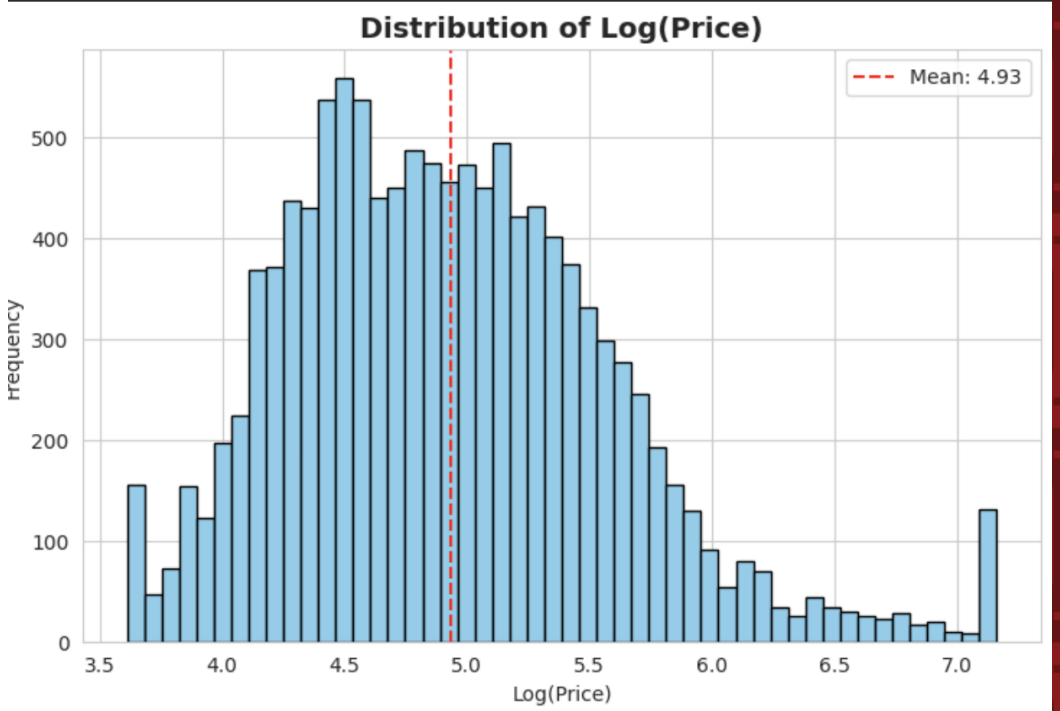
1. **Amenities Mining:** Parsed complex JSON amenity lists to create binary flags (0/1) for high-value drivers like `pool`, `waterfront`, `kitchen`, and `air_conditioning, washer.`
2. **Host Tenure:** Engineered a `host_tenure_days` feature (calculating days since `host_since`) to quantify host experience.
3. **New Host Flag:** Created an `is_new_host` binary indicator to explicitly flag inexperienced hosts who might price unpredictably. **(90 Days)**

Feature Engineering – Text & Market Taxonomy

1. **Text Scrubbing:** Cleaned raw text fields (names/descriptions/amenities) by removing HTML tags, unicode characters, and repeated punctuation. (Unify "wifi" and "wi-fi", etc)
2. **Keyword Extraction:** Created binary flags for high-signal keywords like `beach`, `ocean`, `las_olas`, and `newly_renovated` found in descriptions.
3. **Cardinality Reduction:** Collapsed ultra-rare `property_type` and `room_type` categories (those appearing <1%) into an "Other" category to prevent model overfitting.

Exploratory Data Analysis

Dataset Overview & Price Distribution

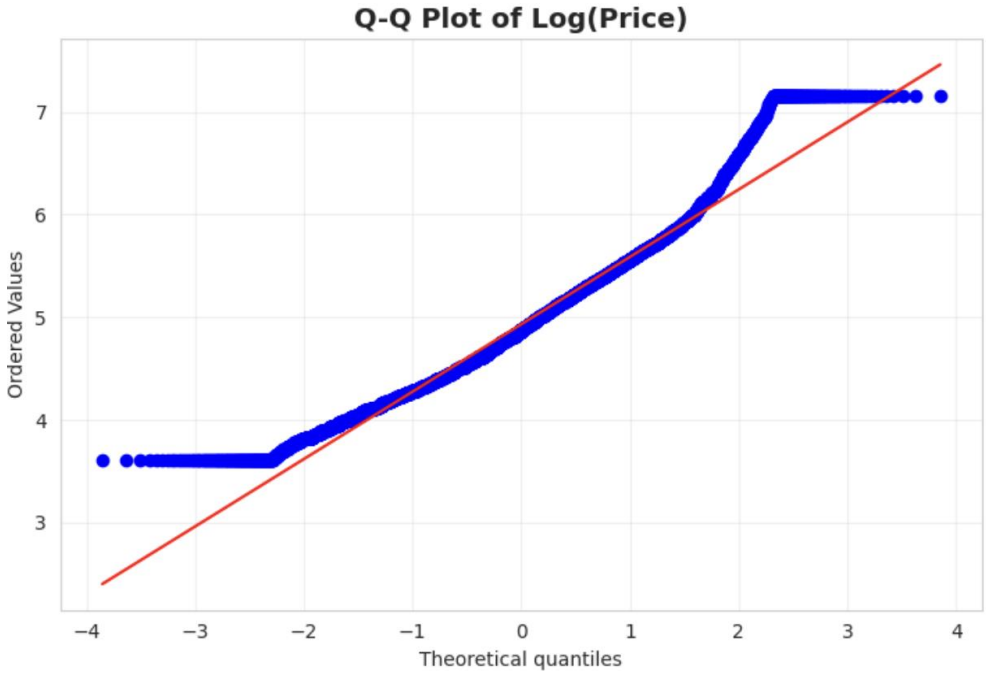


Key Statistics

- Mean Price: **$179**
- Median Price: **$130**
- Range: **$36 - $1,289**

Why Log Transformation?

1. Reduces right skew in price data
2. Improves model performance
3. Enables linear regression assumptions



UMassAmherst | Isenberg School of Management

Exploratory Data Analysis

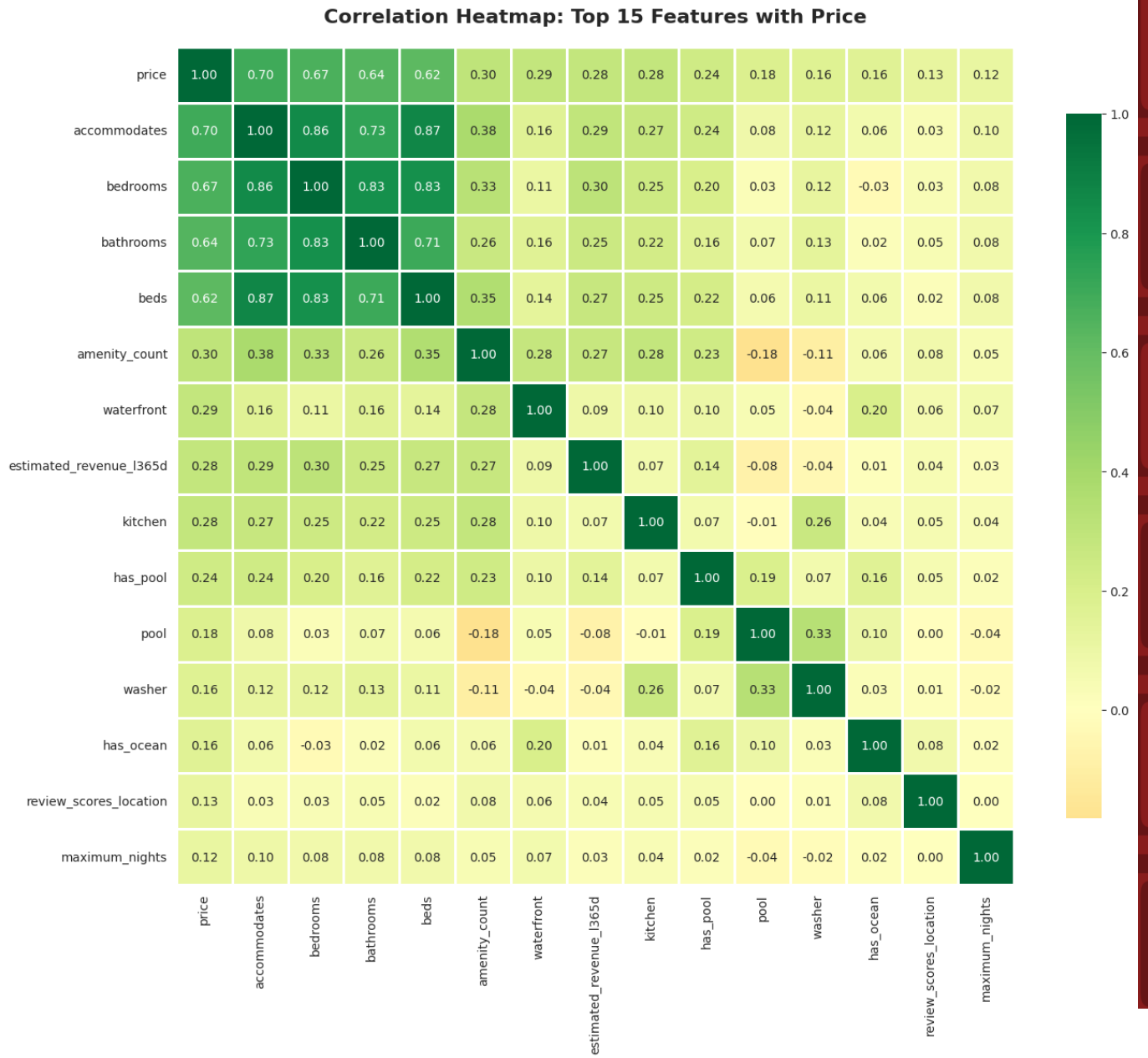
Key Price Drivers - Feature Correlation

Property Characteristics

- Accommodates: **+0.70**
- Bedrooms: **+0.67**
- Bathrooms: **+0.64**
- Beds: **+0.62**

Amenities

- Waterfront: **+0.29** (71.8% premium)
- Pool: **+0.18** (34.5% premium)



Exploratory Data Analysis

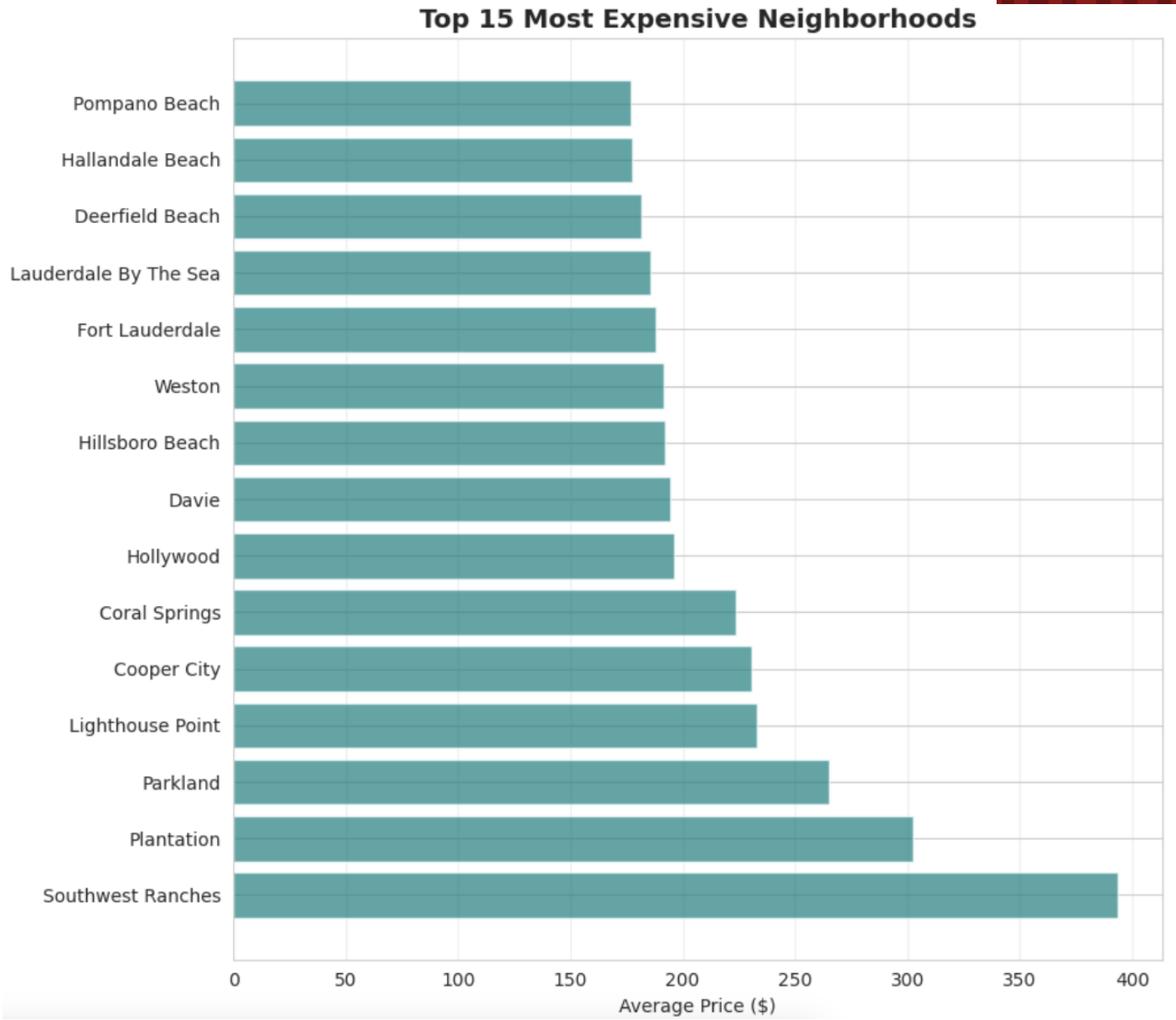
Geographic & Categorical Insights

Geographic Patterns

- **Most Expensive:** SW Ranches ($393), Plantation ($303)
- **Most Affordable:** Pembroke Park ($75), Lauderdale Lakes ($76)
- **Highest Volume:** Hollywood (25.8%), Ft. Lauderdale (23.1%)

Room Type Distribution

- Entire home/apt: 83.6% (Avg: $194)
- Private room: 15.5% (Avg: $100)



Exploratory Data Analysis

Review Patterns & Host Quality

Review Analysis

- Average Rating: **4.75 / 5.00**
- Median Reviews: **18 per listing**
- Price Correlation: **-0.11 (weak)**

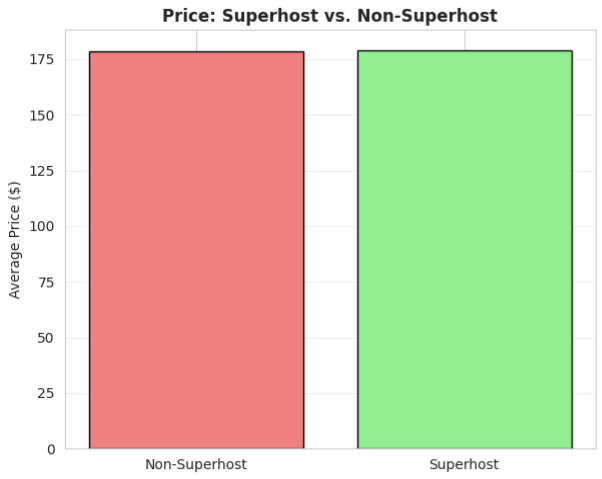
Superhost Impact

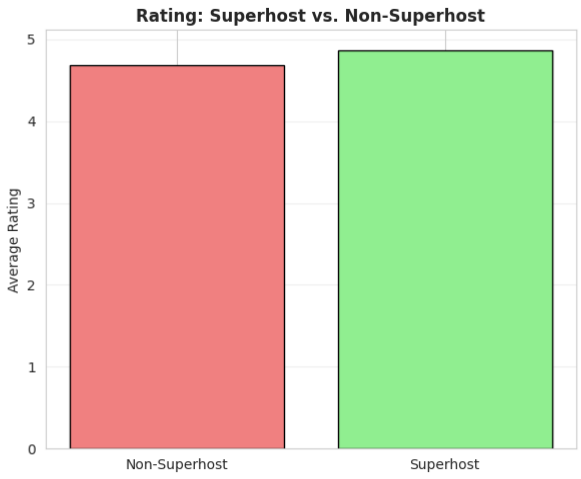
- Price difference: **Minimal**
- Rating advantage: **+18%**
- Review volume: **+119%**

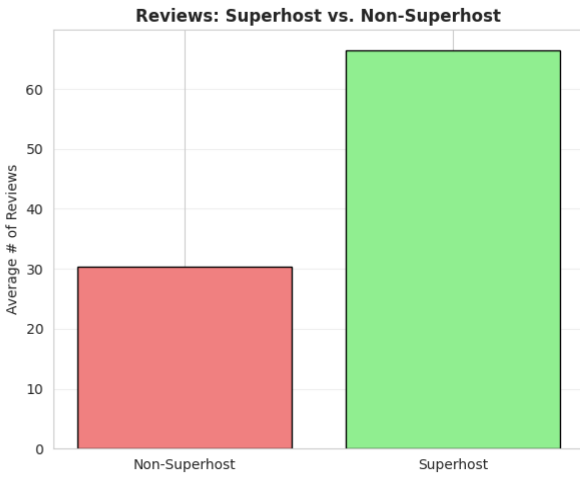
Availability Patterns

Bimodal distribution reveals two strategies:

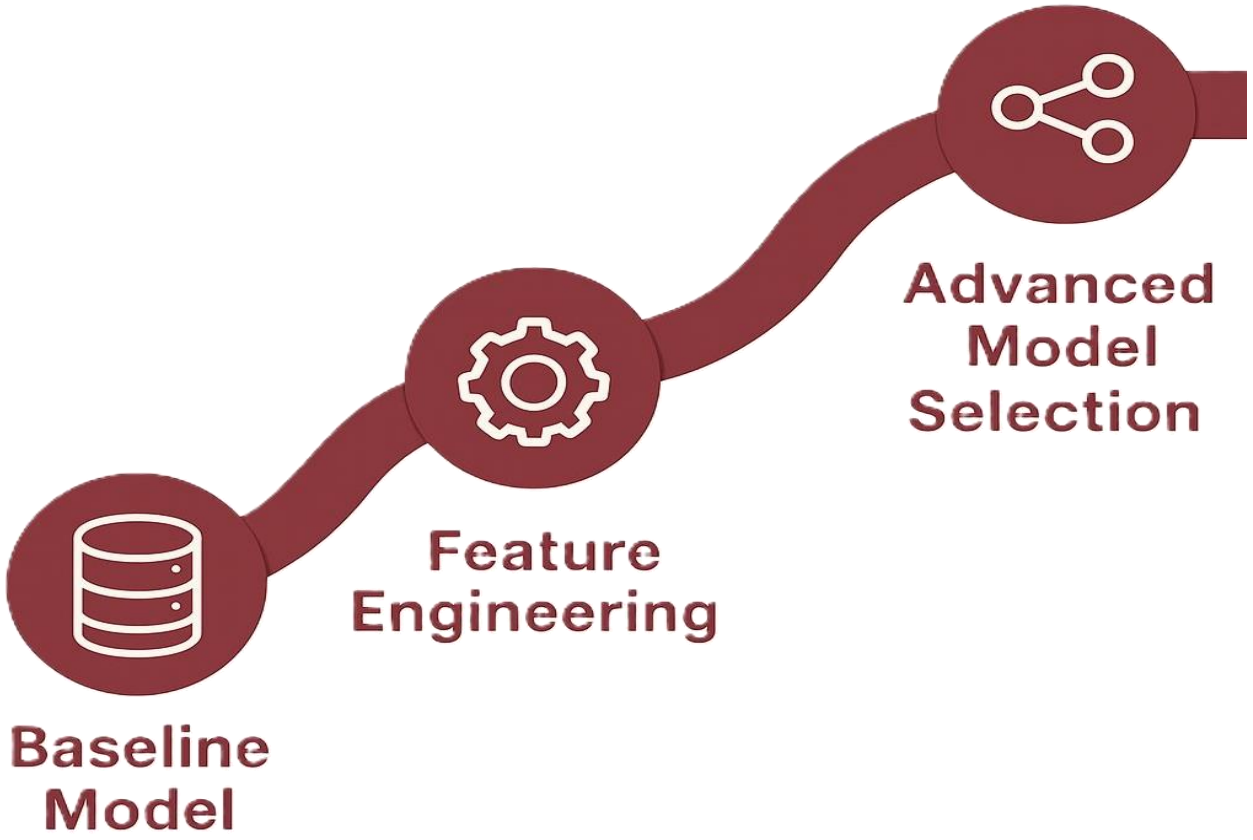
- High availability: Struggling inventory
- Low availability: High demand/personal use







How We Approached Price Prediction



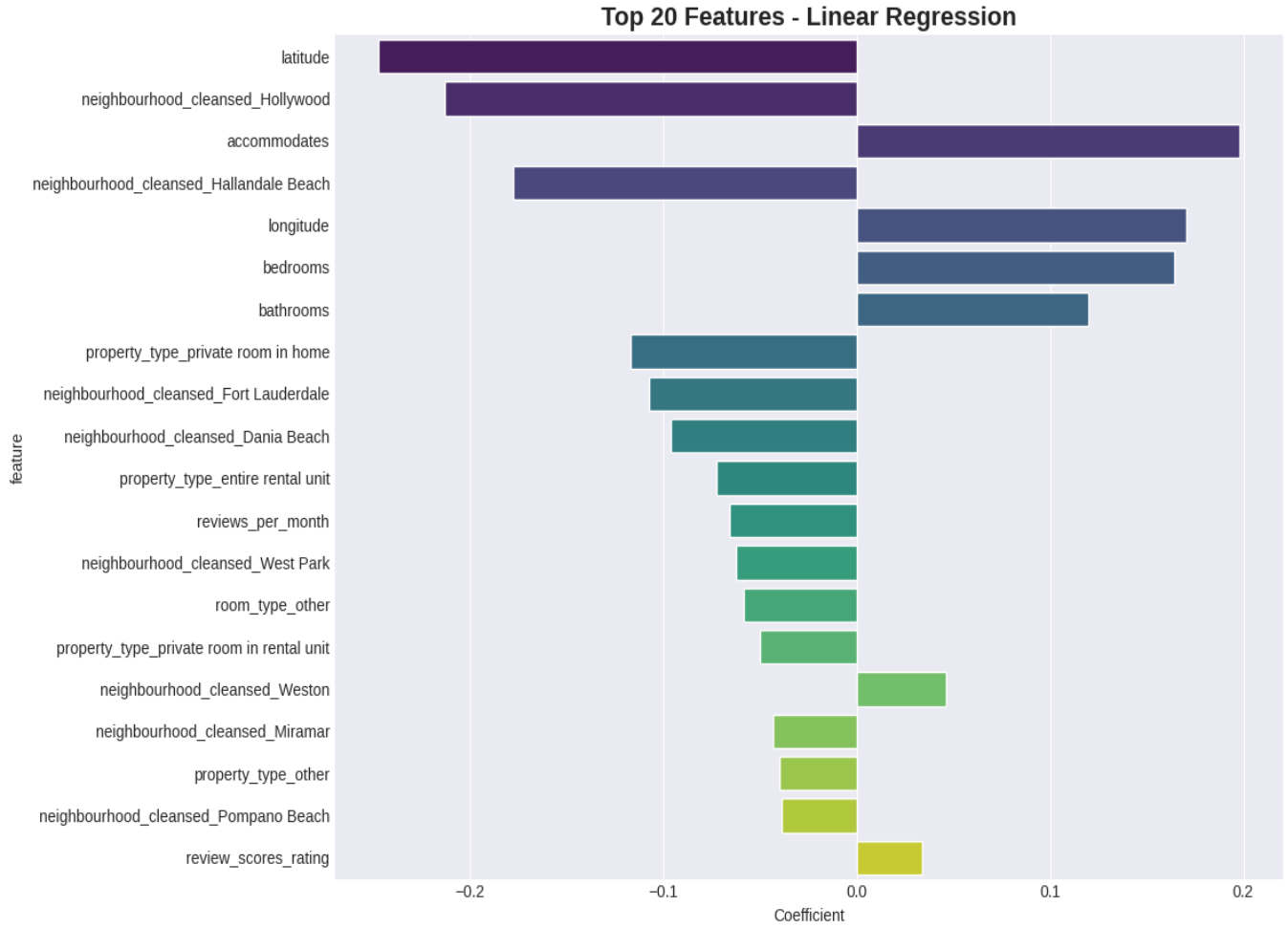
This allowed us to validate early assumptions, strengthen the signal in our data, and ultimately choose a model that balances accuracy, stability, and business interpretability.

Price Prediction

Baseline Model performance

Linear Regression

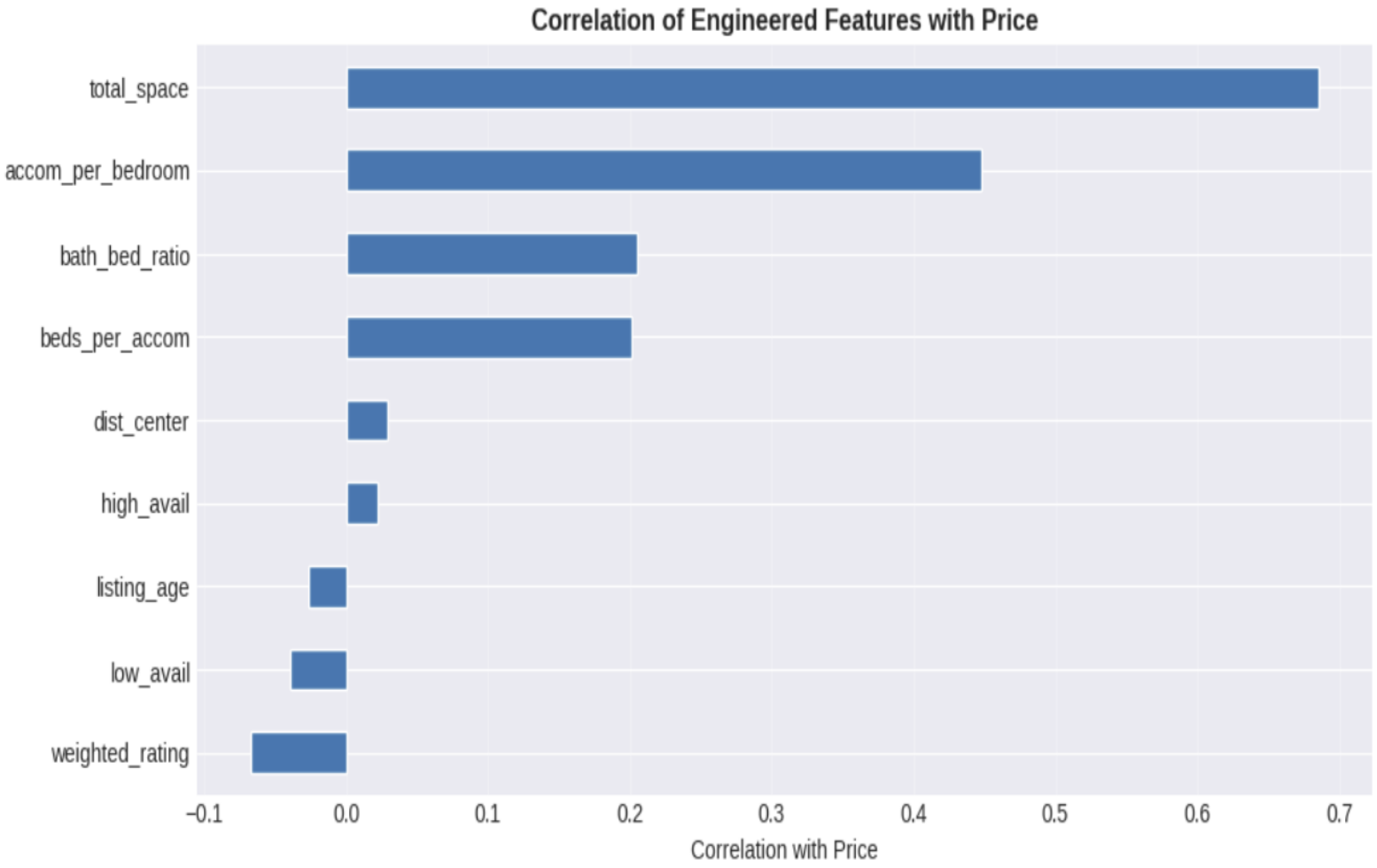
```
============================================================
Linear Regression - Training Set
============================================================
MAE:  $0.30
MSE:  $0.16
RMSE: $0.40
R²:   0.6392
============================================================
============================================================
Linear Regression - Test Set
============================================================
MAE:  $0.31
MSE:  $0.18
RMSE: $0.43
R²:   0.5960
============================================================
```



What the Model Learned:

The Power of Engineered Features

Feature Engineering

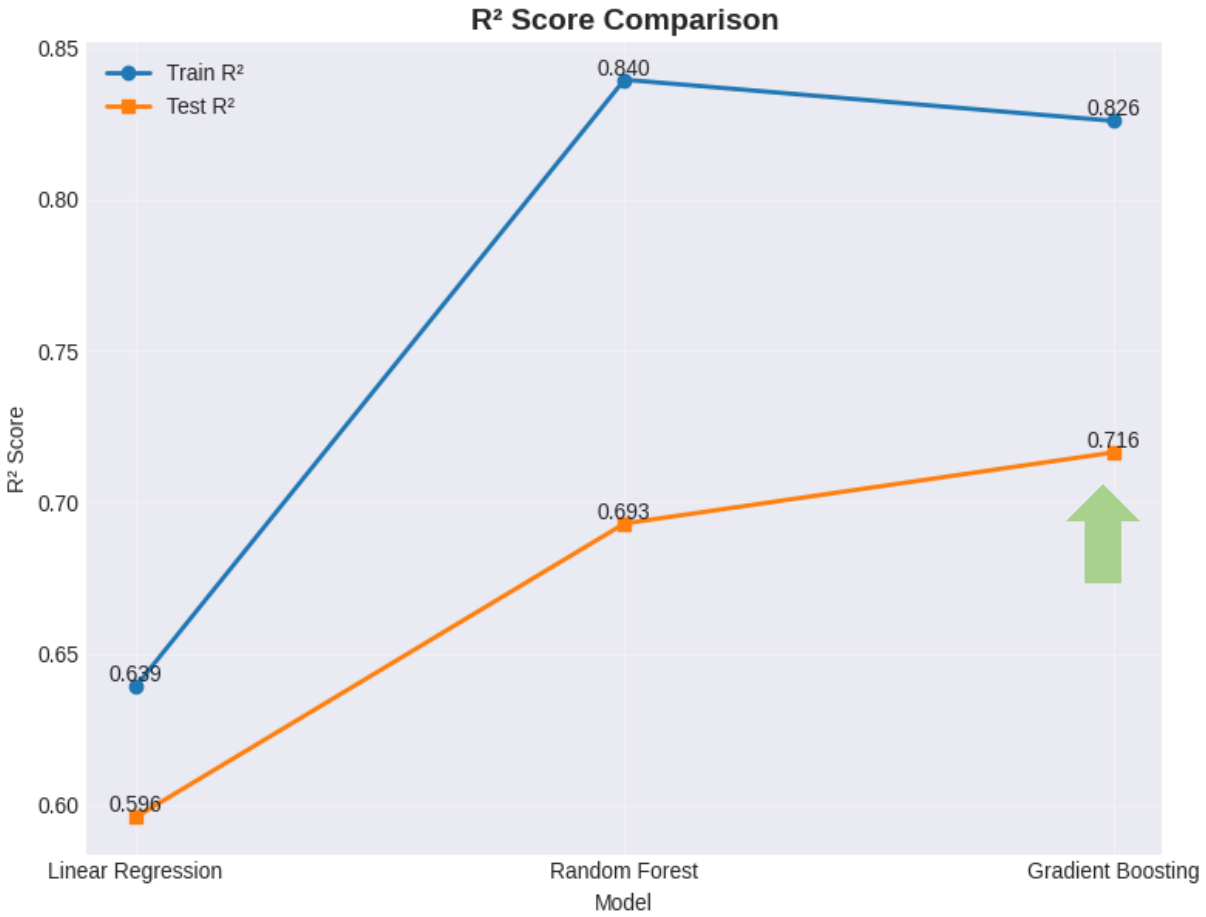


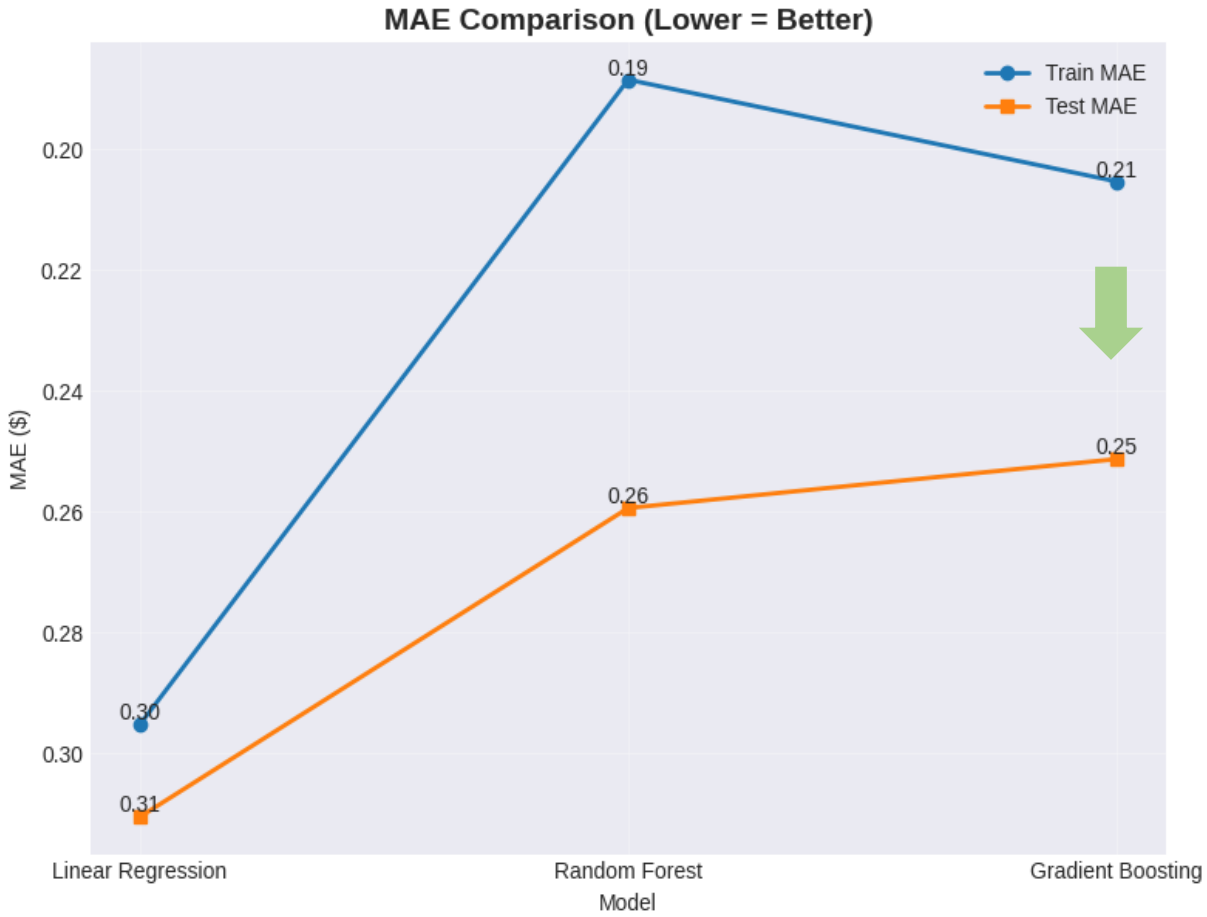
These engineered features reshaped the model's understanding of value. Instead of looking at bedrooms alone, we capture *efficiency*, *engagement*, and *location influence*.

The result: a smarter, more context-aware price model.

Advanced Model

Random Forest & GB





Sentiment Analysis for Airbnb Reviews

Reviews Modelling

- **Data Collection & Cleaning:** Collected Airbnb reviews, removed duplicates, and handled missing values.
- **Preprocessing:** Performed language detection, translation, tokenization, and stop-word removal.
- **Sentiment Analysis (VADER):** Labeled reviews as Positive, Negative, Neutral, or Unknown.
- **Sentiment Analysis (Hugging Face):** Used a transformer model to classify sentiments into Positive, Negative, and Unknown.
- **Agreement Rate Calculation:** Compared outputs to measure consistency between VADER and Hugging Face.
- **Visualization:** Built bar charts, pie charts, and a confusion matrix to display sentiment distribution.
- **Insights & Comparison:** Noted key differences (e.g., Hugging Face identifies more negative reviews; VADER assigns more neutral labels)

Sentiment Analysis for Airbnb Reviews

Original, Translated, And Cleaned Comments (Top 10 Reviews)

- **Language Detection:** Used the langdetect library to automatically identify the language of each review.
- **Translation:** Applied the **Google Translator (googletrans library, version 4.0.0-rc1)** to convert non-English reviews into English.
- **Standardization:** Stored translated text in a new column (english_comments) so all reviews are in a consistent language.
- **Text Cleaning:** Applied your custom preprocessing function to remove noise, stopwords, and lemmatize words, producing cleaned_comments ready for modeling.
- **Final Output:** Displayed a table with the original, translated, and cleaned comments for the first 10 entries

	comments	english_comments	cleaned_comments
108825	Great place for a night!	Great place for a night!	great place night
453910	This airbnb is perfectly/centrally located, be...	This airbnb is perfectly/centrally located, be...	airbnb perfectlycentrally located beach pier l...
63326	This place was perfect for our family! We wer...	This place was perfect for our family! We wer...	place perfect family able stay near beach quai...
421587	Nice house. Not a bad area. Seems to be lackin...	Nice house. Not a bad area. Seems to be lackin...	nice house bad area seems lacking bit quality ...
490146	Everything was excellent!	Everything was excellent!	everything excellent
298505	Great place to stay if you want to be close to...	Great place to stay if you want to be close to...	great place stay want close wilton manner rest...
170530	Great and privat spot for a few nights. Bed wa...	Great and privat spot for a few nights. Bed wa...	great privat spot night bed really good
587894	Excelente lugar con una gran vista! Los mueble...	Excellent place with a great view!The furnitur...	excellent place great viewthe furniture descri...
209557	Great host, great location, great vacation!	Great host, great location, great vacation!	great host great location great vacation
618667	El apartamento esta muy lindo, en un condomini...	The apartment is very nice, in a safe condomin...	apartment nice safe condominium easy accessthe...

Sentiment Analysis for Airbnb Reviews

2. Text Vectorization: Representing Words as Numeric Features

Bag-of-Words (BoW):

- Represents text as word frequency counts and ignoring grammar and word order.
- Simple and effective, but treats all words equally without considering importance.

TF-IDF (Term Frequency–Inverse Document Frequency):

- Weighs words by how often they appear in a document versus across all documents.
- Highlights distinctive terms, reducing the impact of common words like “the”, “and” etc.

Vectorization Output	
BoW	(5000, 3491)
TF-IDF	(5000, 3491)

3. Topic Modelling Techniques

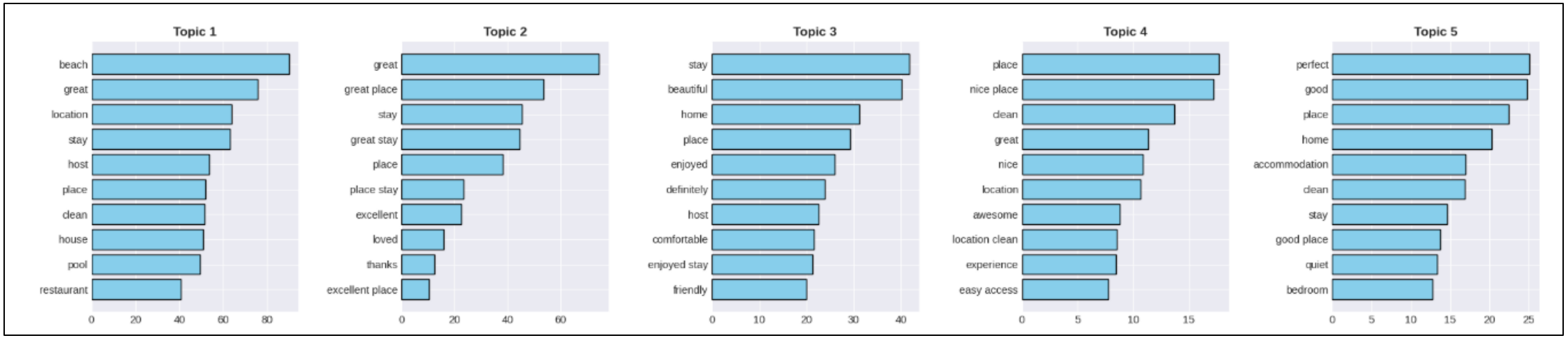
•**LDA (Latent Dirichlet Allocation):** Used to uncover hidden topics in reviews by modeling word distributions.

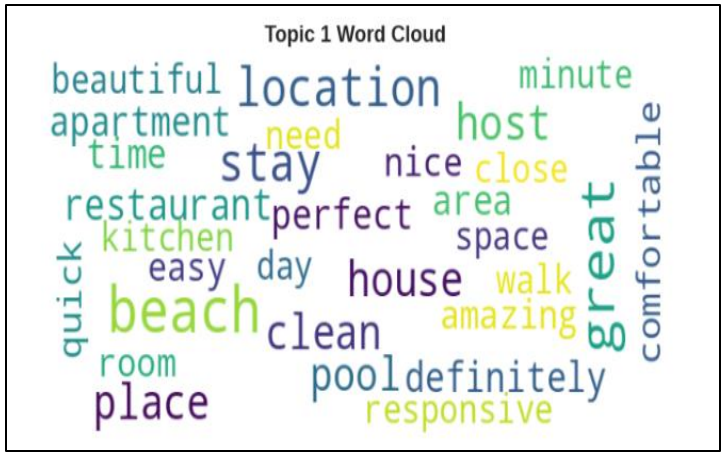
•**NMF (Non-negative Matrix Factorization):** Applied to extract interpretable topics by decomposing text into additive parts.

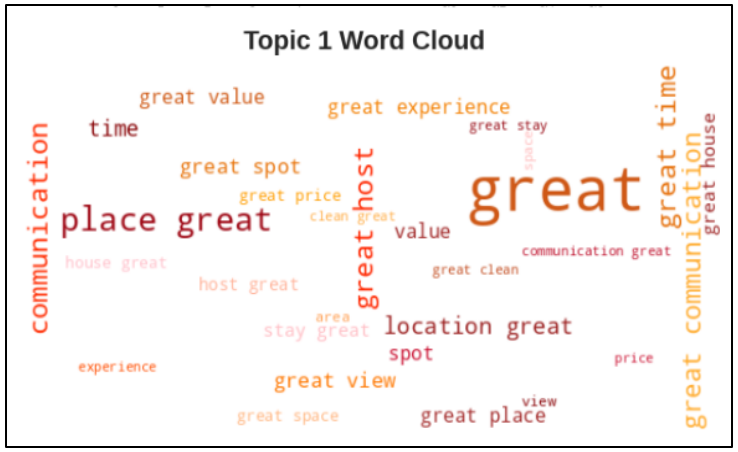
•**LSA (Latent Semantic Analysis):** Used to capture semantic relationships between words and documents through dimensionality reduction.

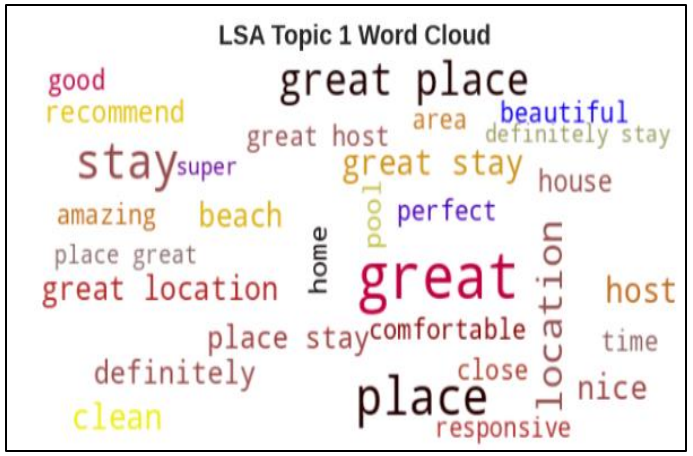
Sentiment Analysis for Airbnb Reviews

TOP WORDS & WORD CLOUD





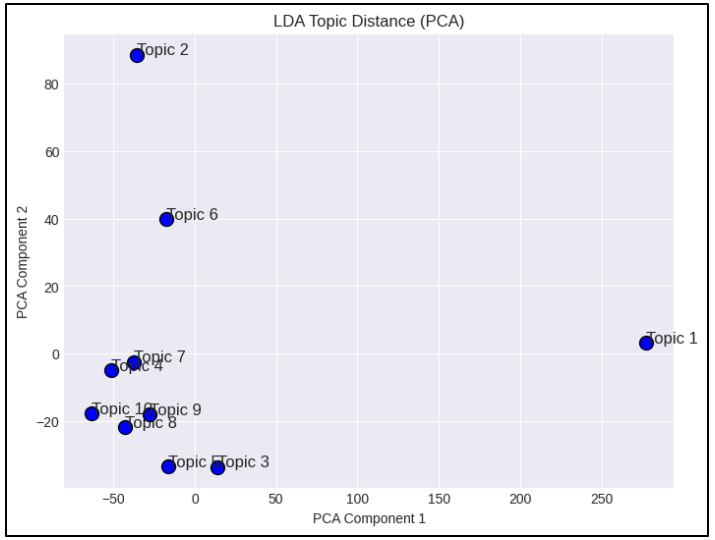


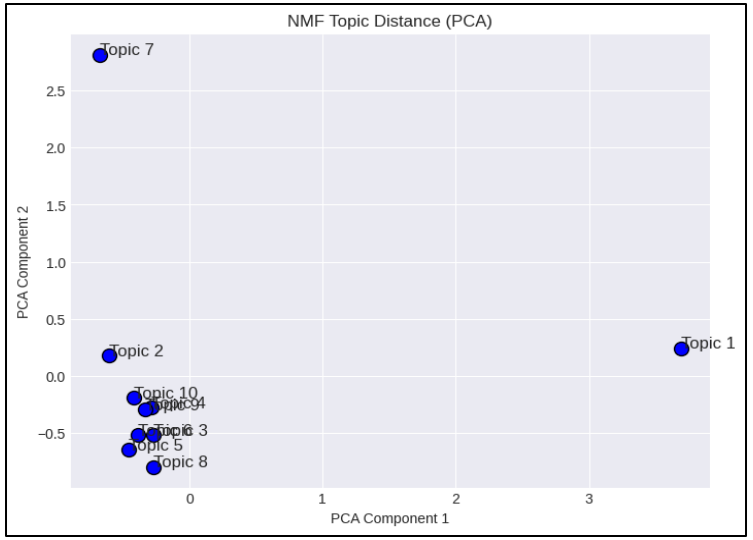


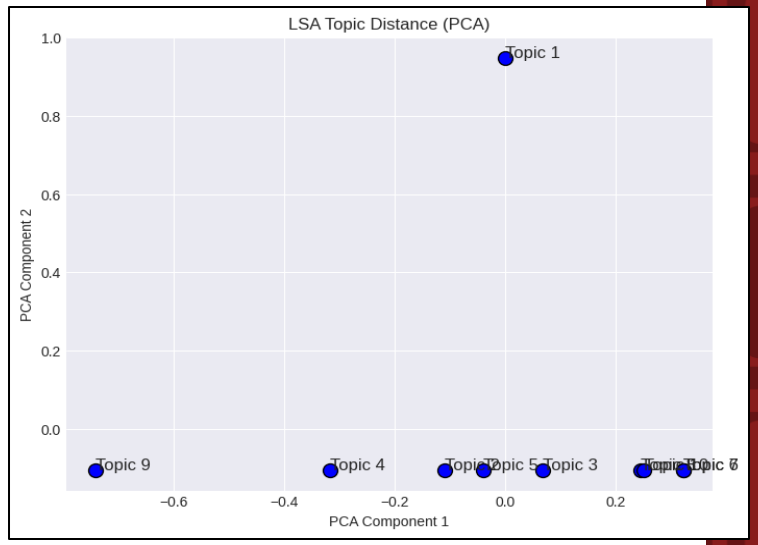
Sentiment Analysis for Airbnb Reviews

TOPIC DISTANCE

- **LDA** shows the strongest topic separation, producing highly distinct and well-differentiated topics across the PCA space.
- **NMF** produces moderately separated topics, with a couple of clearly distinct themes like Topic 1, Topic 2 and Topic 7 while the rest form a compact but interpretable cluster.
- **LSA** topics cluster tightly, indicating strong similarity and less thematic distinction. Only topic 1 , topic 9 and topic 4 are meaningfully separated.



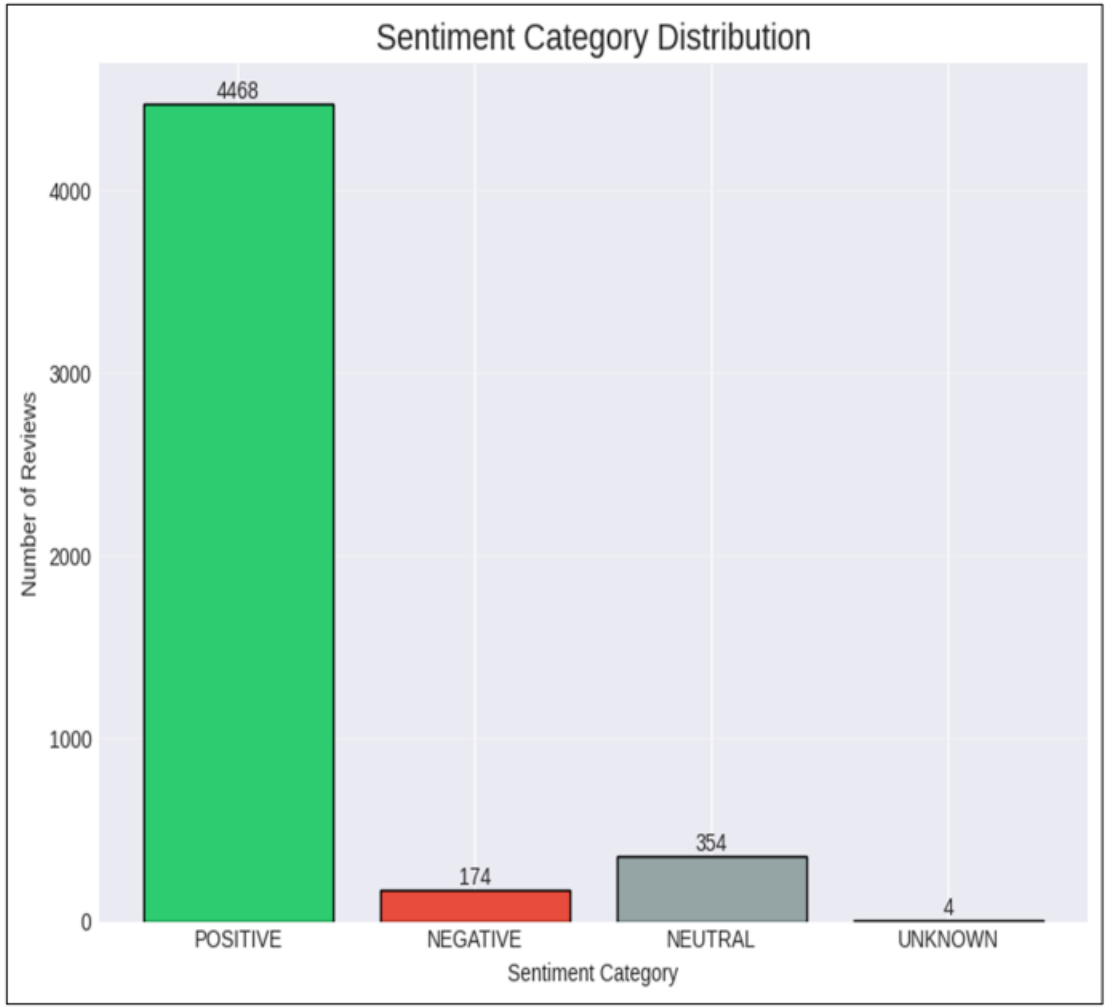




Sentiment Analysis for Airbnb Reviews

Sentiment Analysis with VADER (Valence Aware Dictionary and sEntiment Reasoner)

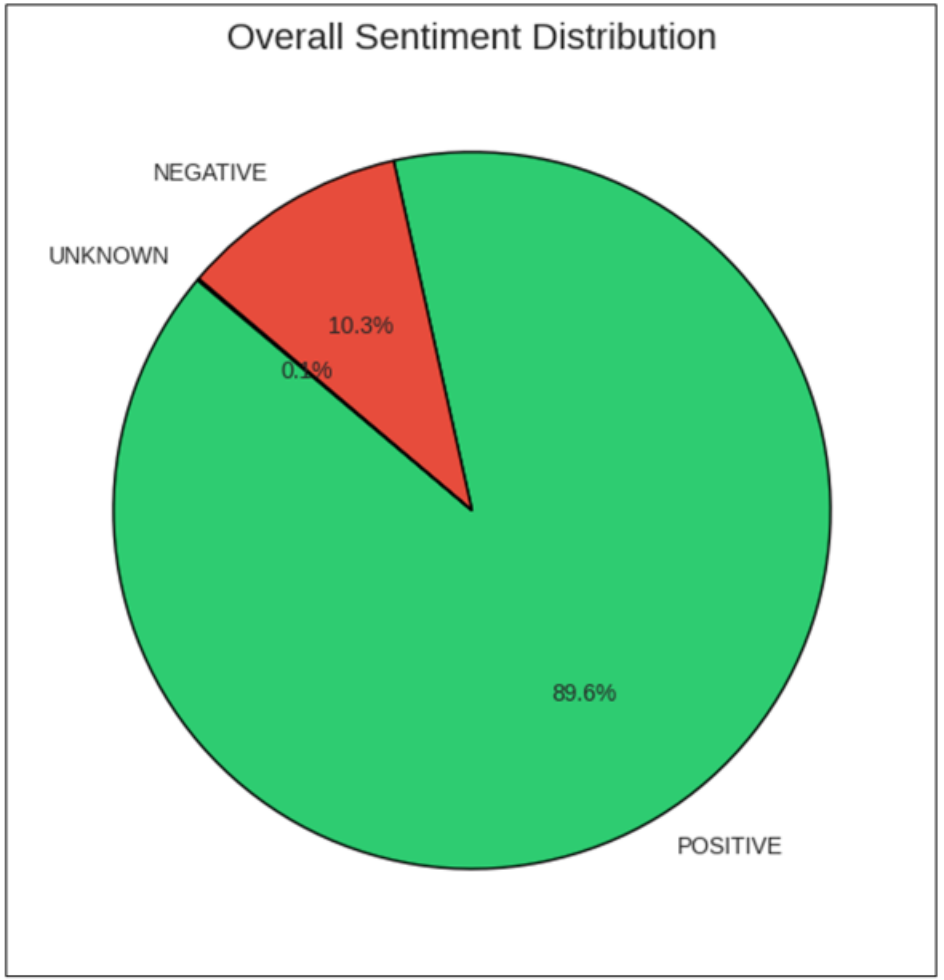
- **Positive reviews**: Out of 5,000 sampled Airbnb reviews, 4,468 (≈89%) are labeled as Positive, indicating strong guest satisfaction.
- **Neutral sentiment**: Only 354 reviews (≈7%) are marked as Neutral, suggesting VADER tends to favor polarized sentiment or that guests rarely leave neutral/mixed feedback.
- **Negative reviews**: Just 174 reviews (≈3.5%) are classified as Negative, which may reflect either genuinely good experiences or polite phrasing that VADER doesn't flag as harsh.
- **Minimal data issues**: Only 4 reviews fall under Unknown, showing that the dataset is clean and well-processed with very few unusable entries.



Sentiment Analysis for Airbnb Reviews

Sentiment Analysis using Hugging Face Transformers

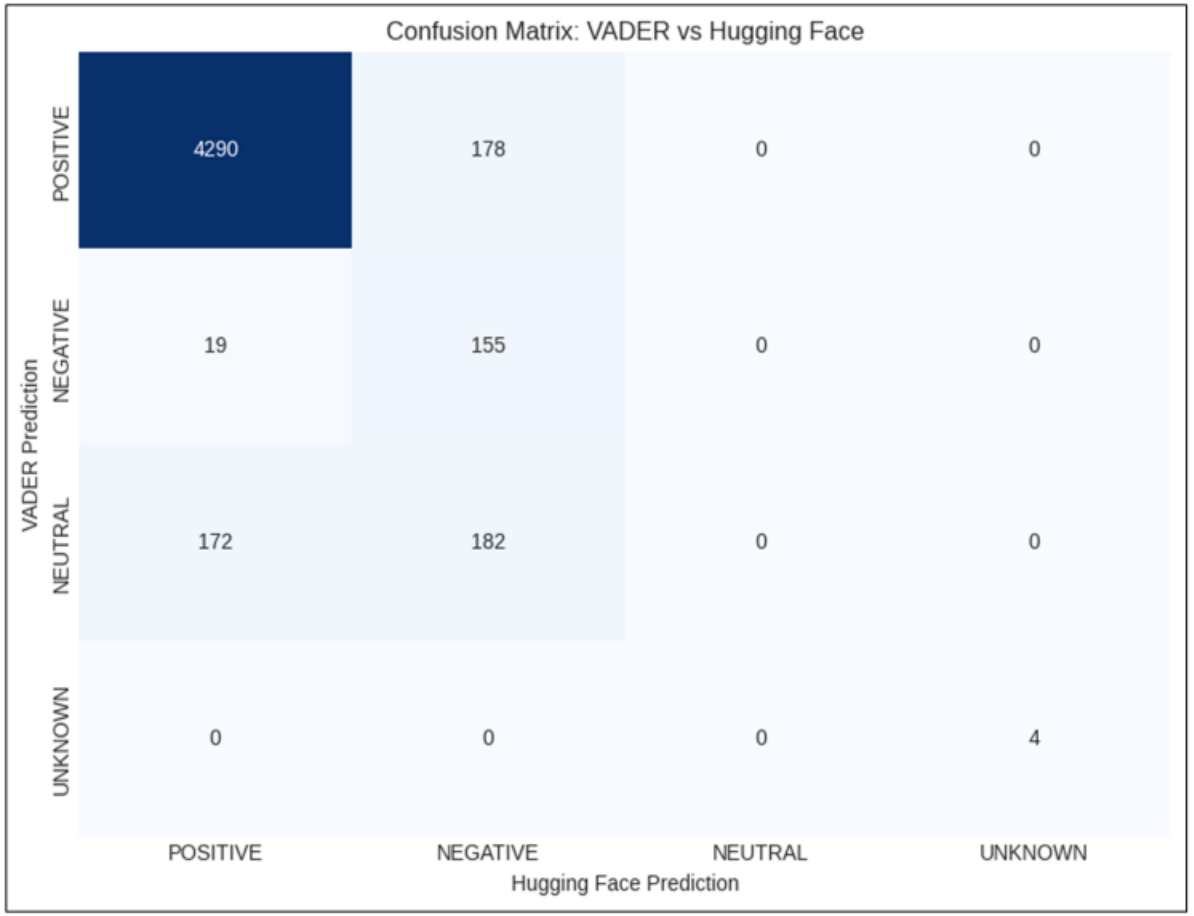
- **Positive reviews**: 89.6% show strong satisfaction across the dataset which means majority of the reviews are positive.
- **Negative sentiment is higher than VADER's output**: 10.3%, which suggests that Hugging Face is more sensitive to critical tone.
- **No neutral category detected**: The model forces polarity, unlike VADER which allows neutrality.
- **Minimal unknowns**: Only 0.1%, indicating robust language handling and clean preprocessing



Sentiment Analysis for Airbnb Reviews

Confusion Matrix: VADER vs Hugging Face

- **Strong agreement on positive reviews**: 4,290 reviews were labeled Positive by both models, showing high alignment on positive sentiment.
- **Disagreement on neutral cases**: VADER marked 354 reviews as Neutral, but Hugging Face assigned Positive to 172 and Negative to 182,which suggests that Hugging Face forces polarity where VADER stays cautious.
- **Minor mismatches in negative sentiment**: 19 reviews labeled Negative by VADER were seen as Positive by Hugging Face, indicating some optimism bias in the transformer model.
- **Perfect match on unknowns**: All 4 Unknown cases were consistently labeled by both models, confirming clean handling of edge cases.



Business Implications

1. Smarter, Market-Aligned Pricing
AI-powered price prediction helps hosts set fair and competitive rates by factoring in amenities, location, and property size, which in turn maximizes occupancy and overall revenue.

2. Actionable Guest Insights
Sentiment analysis of guest reviews identifies what drives satisfaction or complaints, allowing hosts and Airbnb to improve service quality and boost ratings.

3. Strategic Market Positioning
Geographic and feature-based insights reveal which neighborhoods and property traits yield the highest returns, guiding investment and marketing strategies.

4. Revenue Optimization driven by Data
Integrating predictive models like Random Forest and Gradient Boosting creates a reliable, scalable system that continuously refines pricing and performance decisions hence, ensuring long-term profitability and competitive advantage.

Thank you for your attention

Any Questions?

Any Comments?

UMassAmherst | Isenberg School of Management