Understanding Customer Satisfaction & Pricing in Airbnb Listings (Austin, TX)

Combining Structured and Unstructured Analysis

Akshat Gadgil, Sienna Amorese, Vernon Walker, Parker Jones, Brody Kasprzak
MSBX 5420 - Group 11

Problem & Importance

PROBLEM

Airbnb market is highly competitive

Success depends on:

- Price
- Property features
- Customer experience

GAP

- Most analyses focus only on structured data
- Guest perception and experience are ignored

CORE QUESTION

What drives Airbnb success — the property itself, or how guests experience it?

Data Overview

10,500
Listings

79
Variables

Austin, TX
Market

Inside Airbnb Dataset

Structured Data

- Price, location, amenities
- Property type, host attributes

Unstructured Data

- Description text
- Natural language, open-ended responses

Project Objective

We Aim To

- Use customer reviews to understand drivers of customer satisfaction
- Identify drivers of pricing
- Compare perception (text) vs. reality (features) and find out which reviews (words) correlate the most with the most satisfied customers

Approach

Combine two analytical methods:

Word2Vec

Text analysis of reviews

ML Regression

Pricing prediction & drivers

Feature Structure

Target Variables

Satisfaction

Review scores

Price

Listing price

Features Used

Structured

- Location
- Amenities
- Property type

Text

- customer reviews
- Word2Vec embeddings

Overall Approach

Two-Part Framework

Part 1

Text Analysis (Word2Vec)

Understand perception & experience from guest review language

VS

Part 2

Machine Learning (Supervised)

Understand pricing drivers from structured listing data

Goal: Compare "What people feel" vs "What determines price"

Section 1

Word2Vec Analysis

Understanding Guest Perception Through Review Language

Word2Vec Method

1

Tokenization

Convert descriptions into individual words

2

Model Training

Train Word2Vec on description corpus

3

Vectorization

Words become numeric vectors based on context

Evaluation Methods

- Centroid Similarity
- Pearson Correlation
- Cosine Similarity

Word2Vec: Key Finding 1

Host Relationship Matters Most

High-rated reviews frequently have host names included in the description. Listers are most likely more satisfied when the host was reachable and helpful.

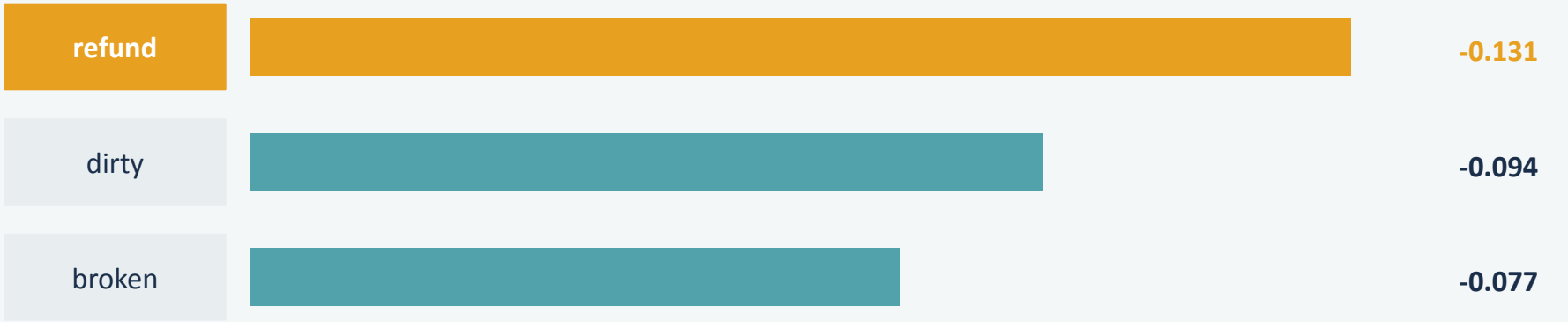
Insight

Most ratings are more influenced by relationships, not only actual properties

Word2Vec: Key Finding 2

Strongest Negative Signal

The word "refund" had a correlation of -0.131, which implies a stronger negative signal than "dirty" or "broken"



Insight: Listers should consider doing refunds. Especially if property has a lot of demand

Word2Vec: Key Finding 3

Cleanliness

- "Dirty" has a strong negative impact
- "Spotless" shows only a weak positive effect

Cleanliness is less important,
but being unclean is a major negative

Emotional Language

- "incredible", "felt like home"
- These expressions drive top ratings

Invoking an emotional connection
is the real key for better reviews

Insight: Meeting baseline expectations (like cleanliness) is not enough — emotional experiences are what earn top reviews

Word2Vec: LDA Analysis

```
— Topic word profiles —
  Topic 0: clean, comfortable, neighborhood, location, definitely, recommend, perfect, responsive, quiet, needed
  Topic 1: group, pool, perfect, location, recommend, space, responsive, amazing, beautiful, beds
  Topic 2: beautiful, loved, amazing, perfect, definitely, recommend, wonderful, space, comfortable, hosts
  Topic 3: clean, location, easy, definitely, comfortable, close, recommend, check, responsive, super
  Topic 4: location, perfect, clean, space, comfortable, close, restaurants, recommend, neighborhood, easy
  Topic 5: location, apartment, clean, restaurants, easy, walking, walk, perfect, downtown, check

— Mean review score by topic —
  Topic 2 | score=4.929 +/-0.142 | n=842 | [beautiful, loved, amazing, perfect, definitely, recommend]
  Topic 0 | score=4.928 +/-0.130 | n=2,644 | [clean, comfortable, neighborhood, location, definitely, recommend]
  Topic 4 | score=4.917 +/-0.112 | n=1,009 | [location, perfect, clean, space, comfortable, close]
  Topic 1 | score=4.876 +/-0.212 | n=1,216 | [group, pool, perfect, location, recommend, space]
  Topic 5 | score=4.831 +/-0.228 | n=1,009 | [location, apartment, clean, restaurants, easy, walking]
  Topic 3 | score=4.652 +/-0.492 | n=2,189 | [clean, location, easy, definitely, comfortable, close]
```

Emotional Language

- Topic 2 scores highest — listings whose reviews use emotional, descriptive language (beautiful, loved, wonderful) outperform everyone else

This directly reinforces the Word2Vec finding that superlatives drive top scores

Overall Analysis: LDA reinforces how different words create different emotions and those emotions can be ranked based on how likeable they are

Word2Vec: Conclusion

01

Experience Over Features

How guest feel matters. Making a guest feel at home is an important thing to gain positive reviews

02

Emotional Response Drives Ratings

Top scores are earned through emotional connection, not just basic satisfaction; these are expected, not a selling point

03

Language Patterns Are Consistent

Positive and negative language signals are reliable and predictable. The more positive the review language, the higher your price can be

Section 2

Machine Learning (Pricing)

Predicting Listing Price from Features

Modeling Approach

1

Data Cleaning

Handle missing values, outliers, and standardizing inputs

2

Feature Engineering

Selected top features from 79 variables

3

Encoding

host_is_superhost, host_response_time

4

Assembly

Build feature vector

Target Variable: Listing Price

Price	accommodates	bedrooms	bathrooms	host_is_superhost	room_type
97.0	3.0	1.0	1.0	1.0	Entire home/apt
160.0	2.0	1.0	1.0	1.0	Entire home/apt
38.0	2.0	1.0	1.0	0.0	Entire home/apt

Models & Evaluation

Model Performance Comparison

Model	RMSE	MAE	R^2
Linear Regression	1715.00	403.64	0.743
Gradient Boosted Trees	**1672.52**	**178.26**	**0.756**
Random Forest	1724.33	194.72	0.740

Key Insight

- Tree-based models outperform linear regression
- This indicates nonlinear relationships between features and price
- Gradient Boosted Trees achieved the best overall performance (R^2 = 0.756, MAE = 178.26)

Pricing Insights

Location

Proximity to attractions, neighborhood

Property Type

Entire home vs. private room

Amenities

Premium features boost perceived value

Key Insight

Pricing is driven by tangible, physical features primarily

Connection

Price Quartile	Category	Price Range	Avg Review Score	Sample Size (n)
Q1	Budget	$8 – $85	4.7724	2,228
Q2	Mid-Low	$86 – $130	4.842	2,247
Q3	Mid-High	$131 – $215	4.8733	2,214
Q4	Premium	$216 – $50,000	4.8769	2,212

Key Insight

- Ratings improve only a little when price is increased
- Price is not a significant contributor to ratings (Seen by very low correlation)

Combined Insights

Customer Satisfaction

Driven By

Experience & Perception

Emotional connection, host relationships, problem resolution

Listing Price

Driven By

Property Features & Location

Amenities, property type, neighborhood

To get the best possible listing, optimize BOTH dimensions for Airbnb success: Improve physical features where can & deliver emotional experiences to earn top reviews