Airbnb Project presentation

QTM 347
Cheryl Chung, Frank Feng,
Shawn Chen, Peter Zhao

Challenges

- Rising Airbnb prices, especially in NYC.
- Travelers face affordability challenges, particularly younger generations.
- Hesitation among users due to escalating costs.
- Motivation: Simplify booking for price-sensitive users.

Why is this problem interesting

- Supports accessible and equitable travel opportunities.
- Simplifies decision-making for travelers.
- Strategic pricing insights for hosts.
- Informs Airbnb policy recommendations in urban markets.

Introduction

Problem Statement and Motivation

★★★★★

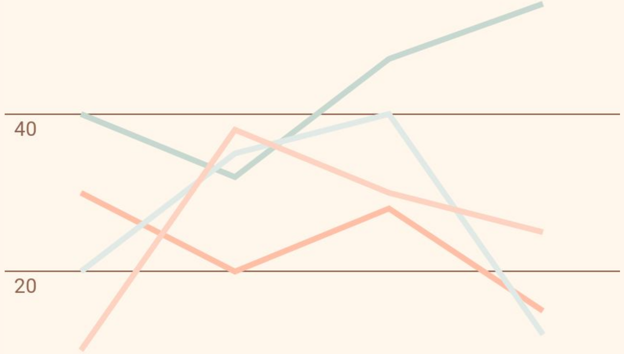
Approach

To tackle this problem:

1. **Identify key features impacting price** using statistical and machine learning models.
2. **Segment the market** into distinct categories based on price ranges and listing characteristics (budget, mid-range, and luxury accommodations
3. Use **Random Forest & Gradient Boosting** for feature importance analysis.
4. Apply **clustering** algorithms for segmentation.

Dataset Overview

2019 New York City Airbnb Open Data dataset
(48,000 rows x 16 columns)



This dataset is sourced from publicly available information from the Airbnb site. It includes various attributes that provide a snapshot of the Airbnb market and its price dynamic in NYC.

Key Variables:

Price: The target variable, indicating the listing price/ night.

Room Type: Type of accommodation (e.g., entire home/apt, private room, shared room)

Availability_365: Num of days the listing is available/year, indicating its overall accessibility to potential guests.

Number of Reviews: Total number of reviews, serving as an indicator of popularity and customer feedback.

Neighbourhood_group: 3 geographic subdivisions within NYC, including Williamsburg, Bedford-Stuyvesant, and Harlem

Set up

★★★★★

1

Data cleaning

`name`: 16 missing values.
`host_name`: 21 missing values.
`last_review` and `reviews_per_month`: 10,052 missing values each.

We dropped these rows with missing values

2

One-hot encoding

One-hot encoding for non-ordinal categorical variables:

`room_type` (2 categories)
`neighbourhood_group` (3 categories)

3

Price Segmentation

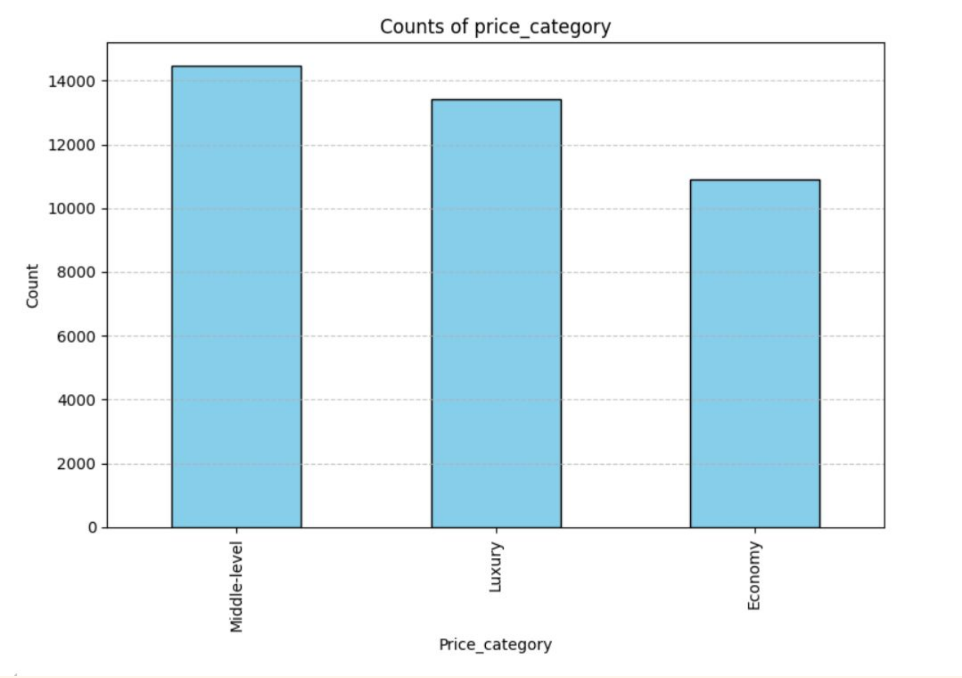
National Economy hotel: $70/night
Middle-income travelers: $140/night on average.

Segment `price` into 3 levels:
Economy: <$70
Middle-level; $71-$140
Luxury: >$140

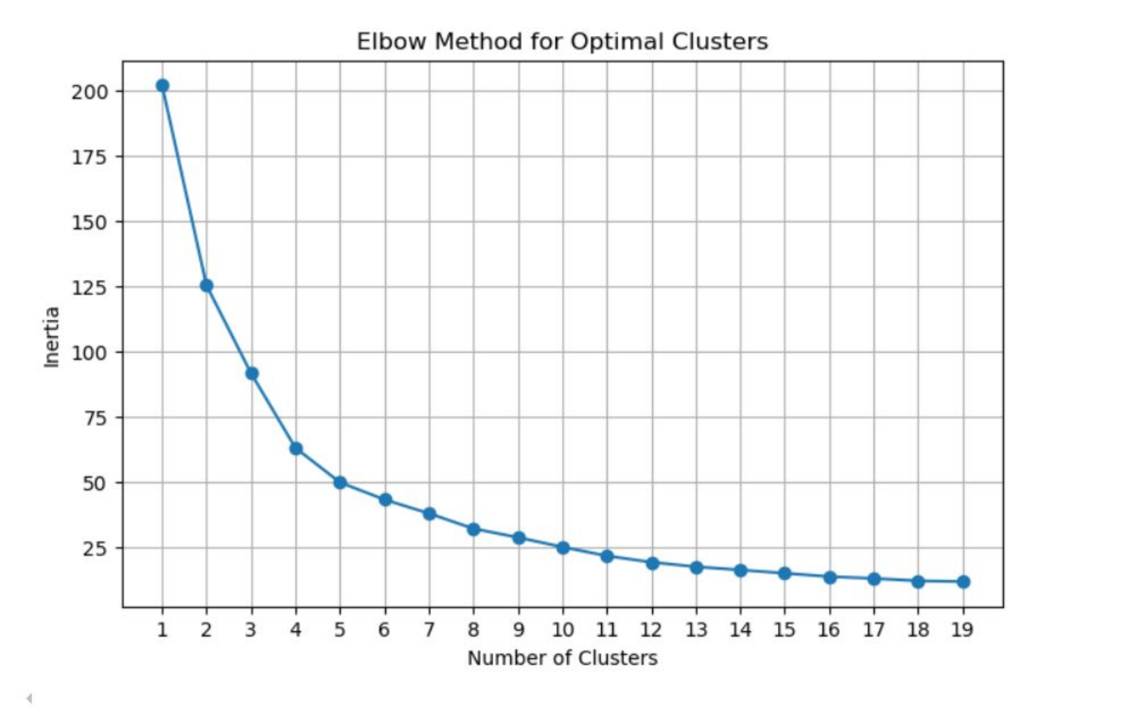
Data Description

- 38812 Data entries after cleaning

	μ	σ
Price	118	65
Review / month	1.12	1.19
Availability	114.9	129.5

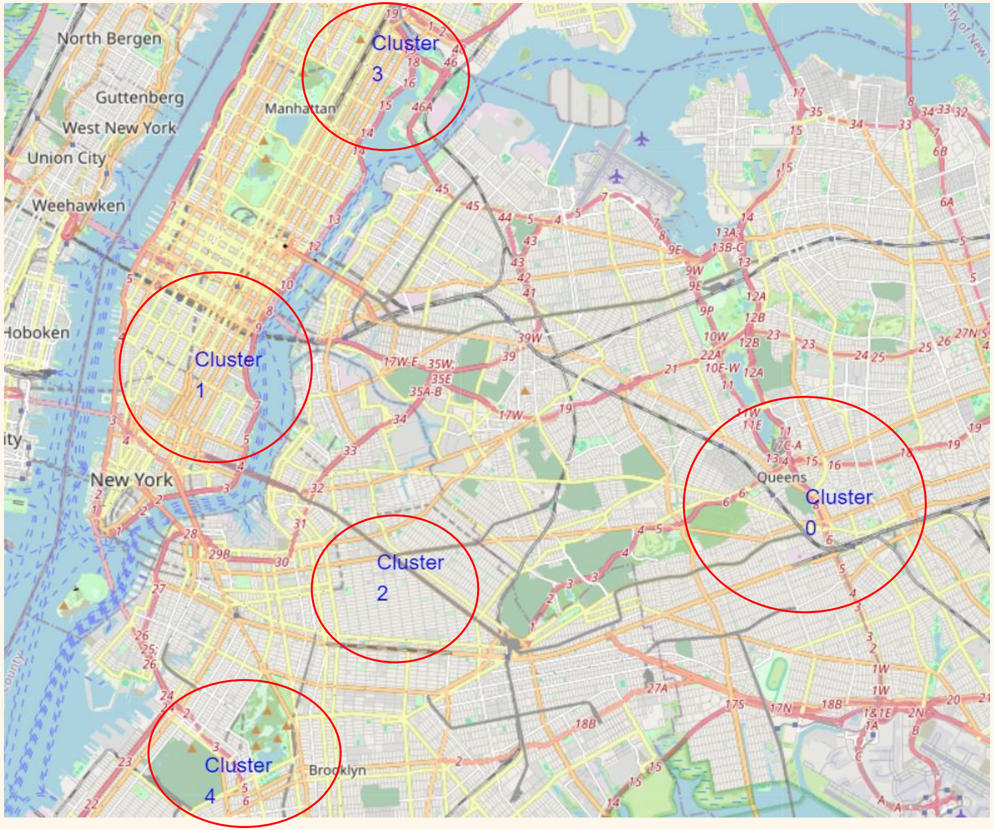


Results: Location Cluster

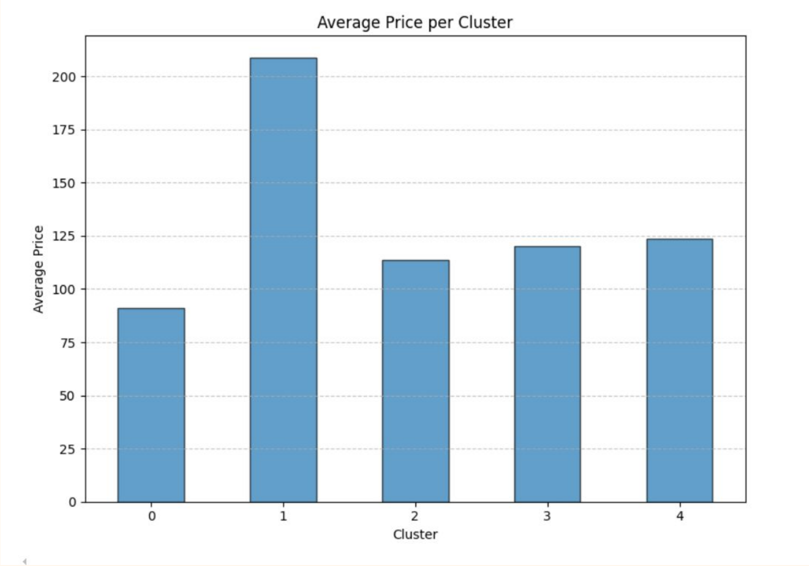


- Use **K-Means** to cluster latitude and longitude information into clusters
- Based on Elbow graph, the optimal number of cluster is decided (balancing how well fit the data and overfitting)
- Lower inertia—better defined clusters; High inertia—poorly defined clusters
- Optimal cluster number is chosen when the slope slows the decreasing rate

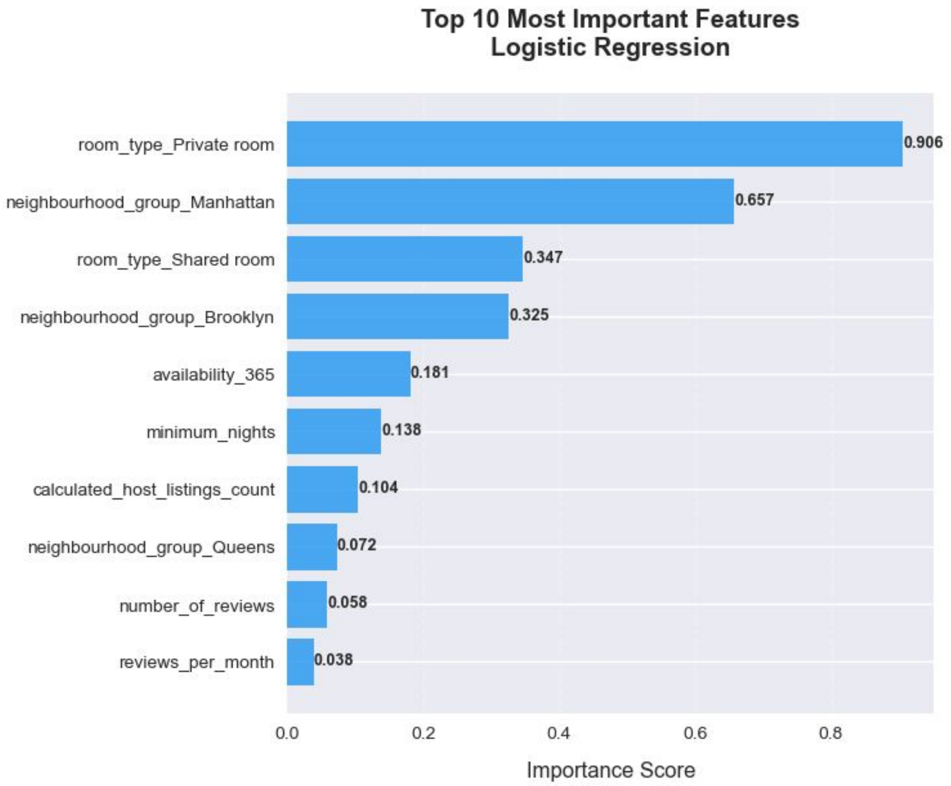
Results: Location Cluster

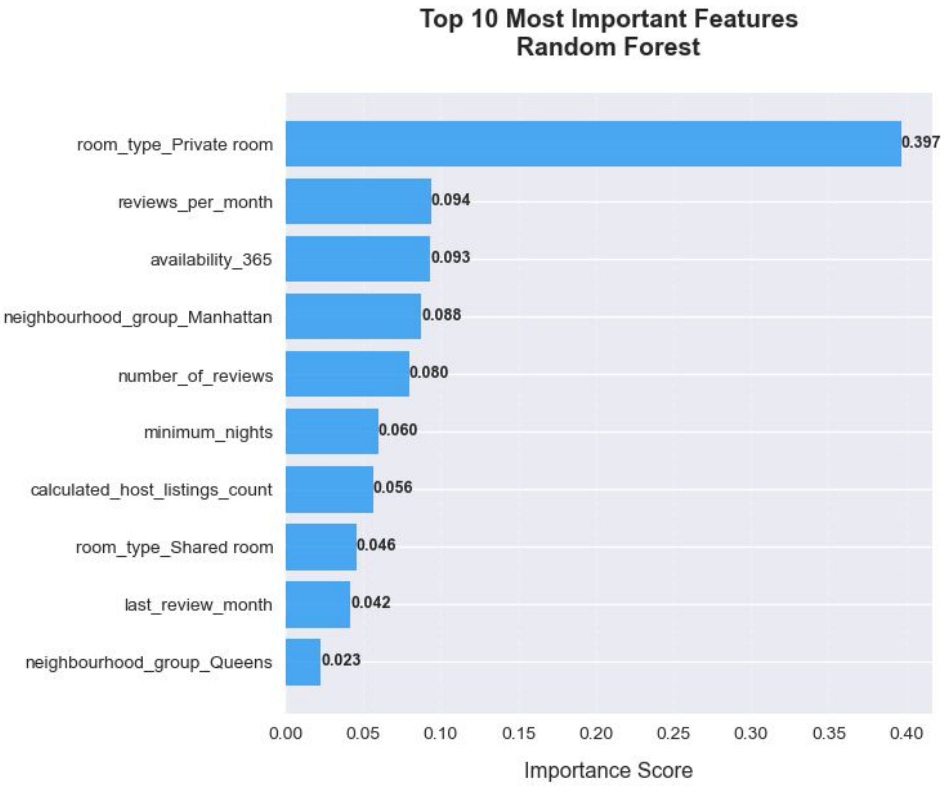


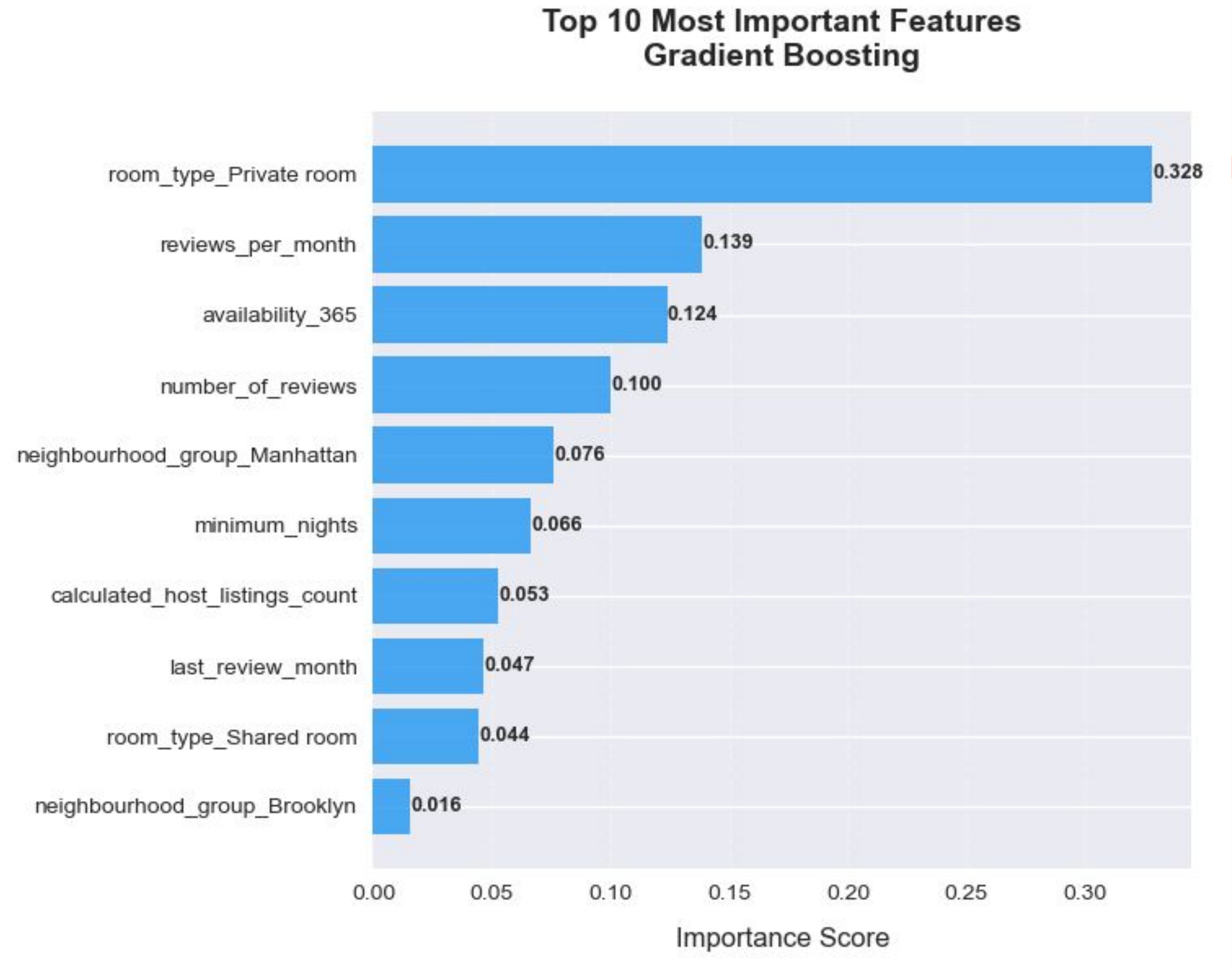
Cluster 0: Queens
Cluster 1: Lower Manhattan
Cluster 2: Bushwick
Cluster 3: Upper Manhattan
Cluster 4: Brooklyn



Results: feature importance

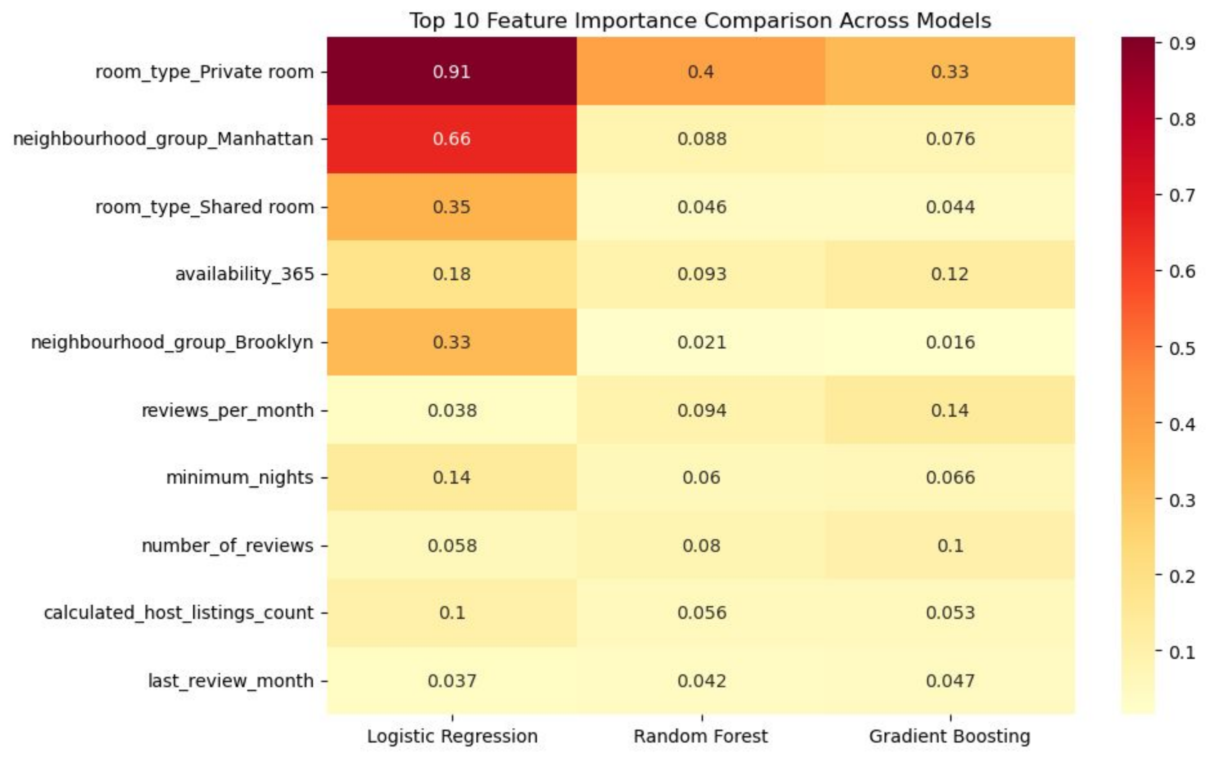






Top 10 Most Important Features
Gradient Boosting
room_type_Private room
reviews_per_month
availability_365
number_of_reviews
neighbourhood_group_Manhattan
minimum_nights
calculated_host_listings_count
last_review_month
room_type_Shared room
neighbourhood_group_Brooklyn
0.328
0.139
0.124
0.100
0.076
0.066
0.053
0.047
0.044
0.016
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance Score

Results: feature importance



Key Patterns:

- **Logistic Regression** assigns higher importance to categorical features (room type, neighborhood)
- **Random Forest** tends to distribute importance more evenly across features
- **Gradient Boosting** often falls between the two, suggesting it captures both linear and non-linear relationships

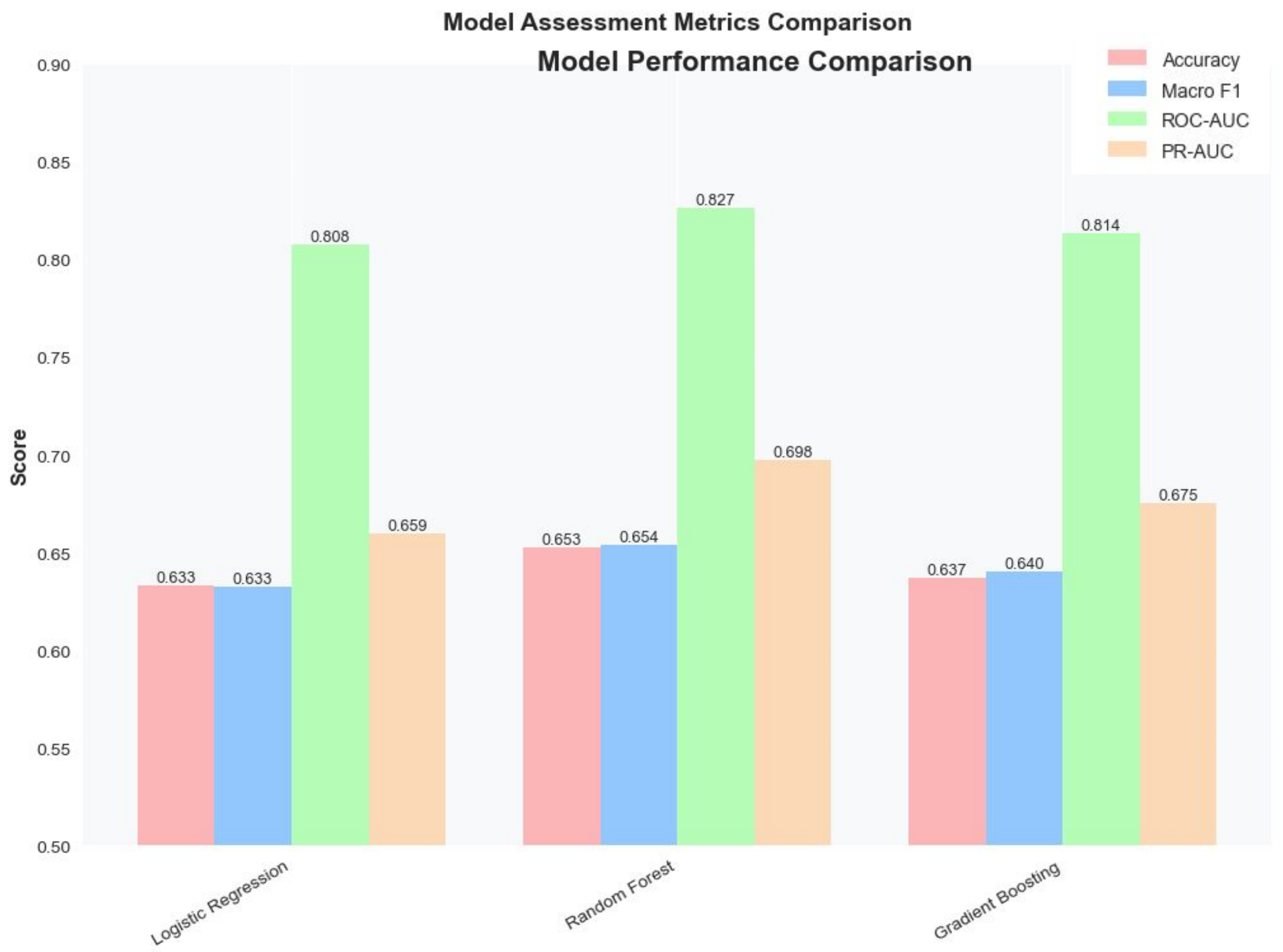
Assessment Metrics

Accuracy: The percentage of all predictions that were correct (e.g., 63.3% means the model correctly classified 63.3% of all listings into their proper price categories).
Macro F1 Score: The balanced average of precision and recall across all price categories, where a higher score (like 0.633) indicates the model is good at both finding actual listings in each category and avoiding false classifications.
ROC-AUC Score: Measures the model's ability to distinguish between price categories, where 0.808 indicates good discriminative ability (0.5 = random guessing, 1.0 = perfect separation).

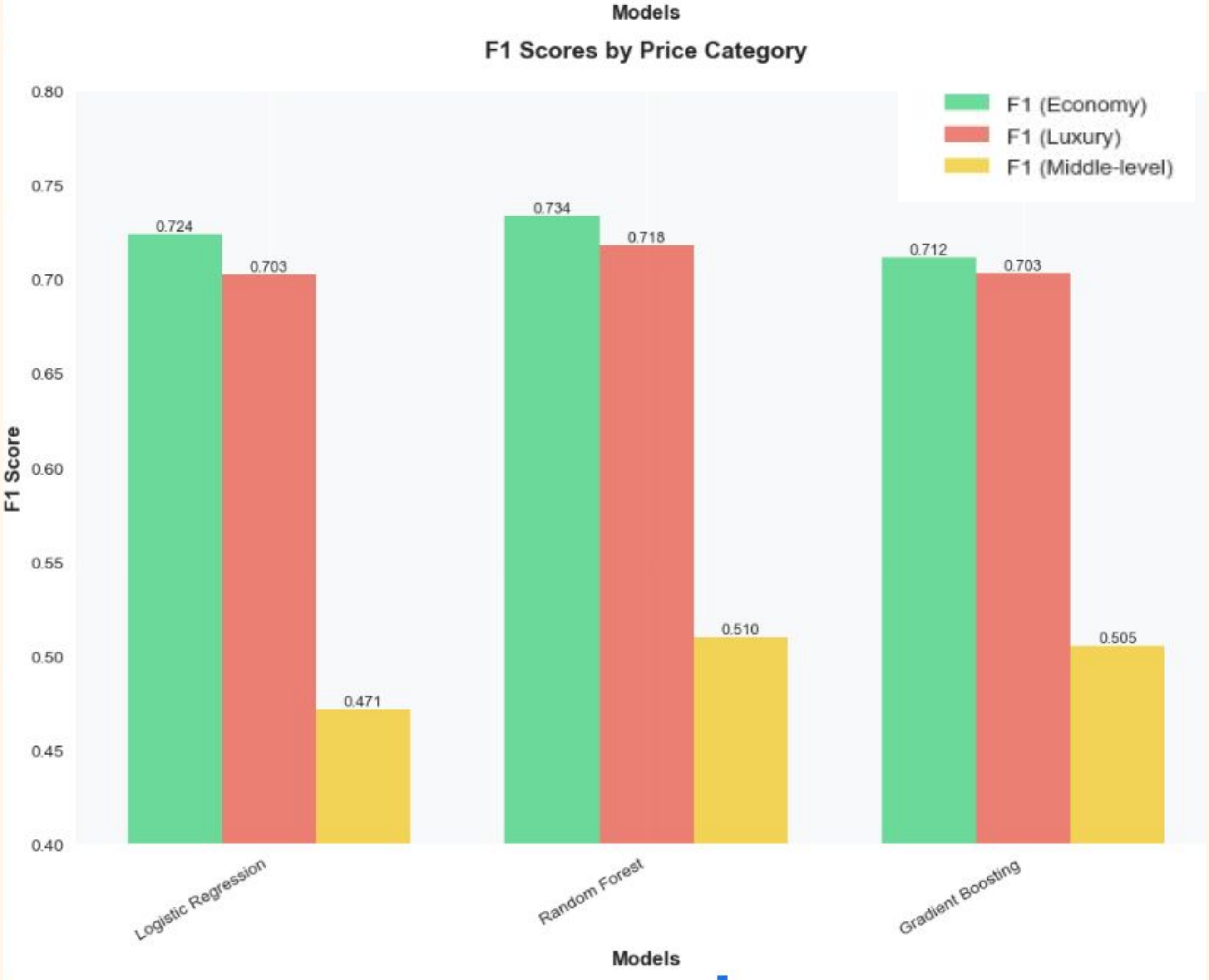
Per-Class F1 Scores:

- **Economy F1 (0.724):** Strong at identifying and classifying budget-priced listings
- **Luxury F1 (0.703):** Strong at identifying and classifying high-end properties
- **Middle-level F1 (0.471):** Suggests the model struggles most with correctly identifying and classifying mid-range properties, possibly due to overlap with other categories

Assessment part I - general scores



Assessment (part 2)



Discussion and Conclusion

Which variable impacts the price the most?

- **Room type** (shared or private) is the most important variable for price
- **Lower Manhattan** is the most expensive neighborhood
- **Romdom forest** has the best combined perdictions

Limitations:

1. lacks detailed customer information
2. Lack of customer reviews context

Future Directions:

- Can segment customers and analyze the impact of pricing strategies on specific groups.
- Can analyze how customer feedback context impacts pricing decisions or influences booking behavior

Thank you!

Any questions?