

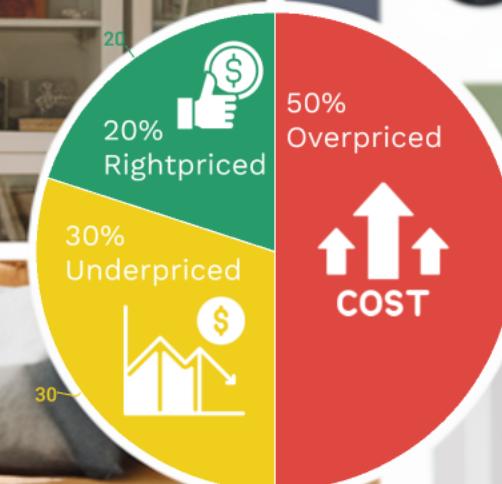


Airbnb Price Prediction

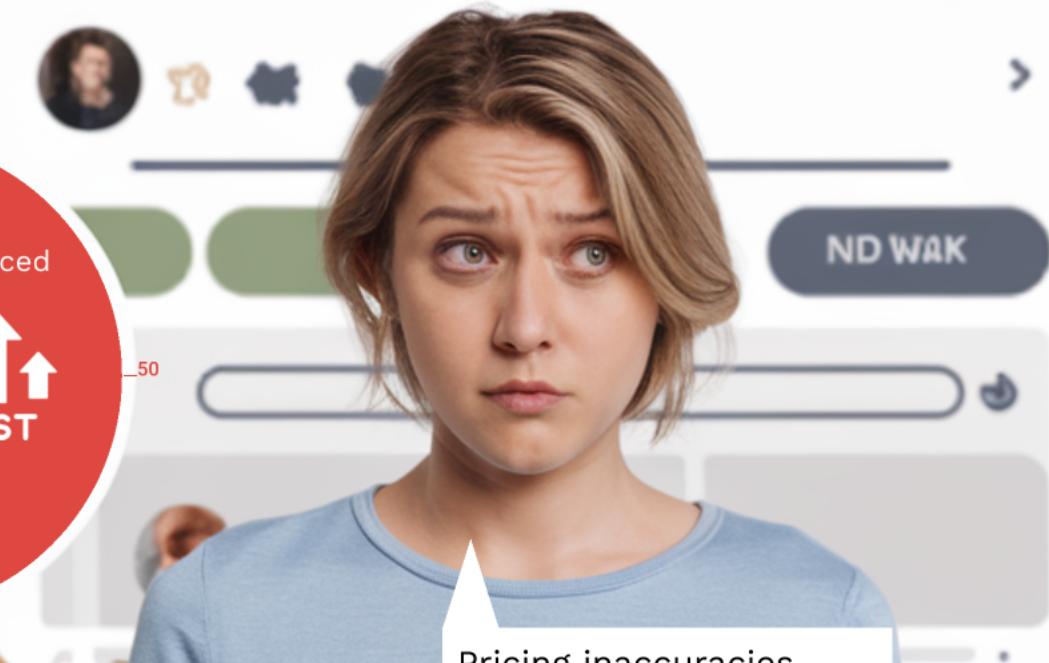
Optimizing Pricing through Supervised Machine Learning

\$12 billion in revenue lost annually

Revenue Down 60%



Overpriced



Pricing inaccuracies
harm trust, satisfaction,
and revenue

Pricing Chaos in Airbnb Listings

Why This Matters

Hosts



Maximize **Revenue** and minimize **Uncertainty**

Guests



Gain **transparency** and **affordability**

Airbnb



Build **trust** and strengthen **market competitiveness**

 74% of Airbnb hosts are individuals relying on income consistency

Fair pricing increases bookings by **15%** on average

Data Description

74,111 Airbnb listings

6 major cities



New York



Boston



Chicago



Los Angeles



San Francisco



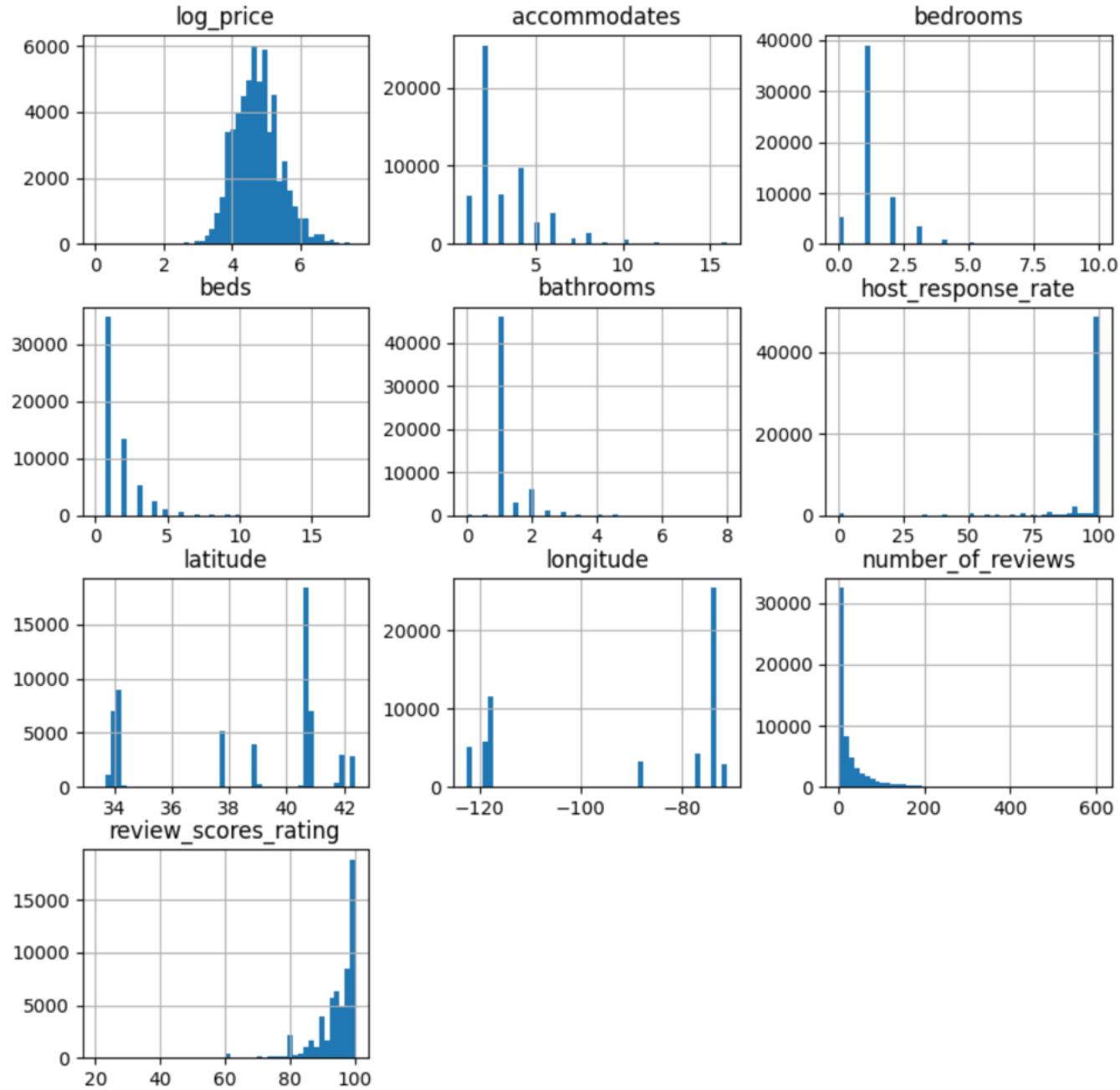
DC

29 features

Key features include:

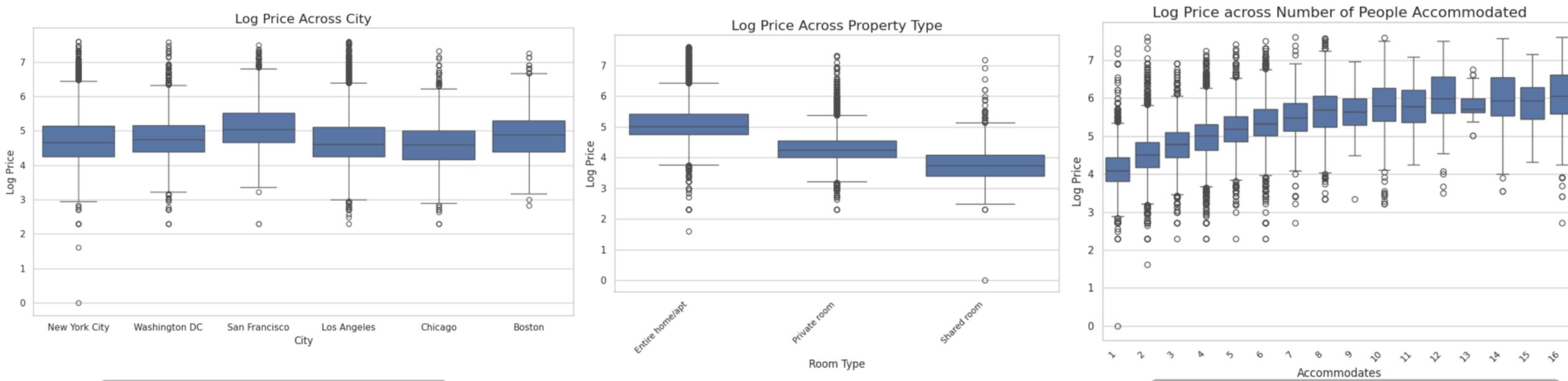
Room type: Entire home, private room, shared space.
City: SF, NYC, LA, Boston, Chicago, DC.
Accommodates, latitude, longitude.

Exploratory Data Analysis



Log-transformation normalizes price distribution, reducing skew





Median prices in NYC and SF are ~2x higher than Chicago or Boston

Entire homes priced ~50% higher than private rooms

Linear trend shows ~20% price increase per additional guest capacity





Data Preprocessing for Airbnb Price Prediction



1. Pipeline Introduction

A robust preprocessing pipeline was built to address data inconsistencies and improve feature representation for the Airbnb price prediction model.



2. Boolean Feature Transformation

Columns such as instant_bookable and cleaning_fee were transformed by replacing 't'/'f' with 1/0. A custom transformer was created for future scalability.



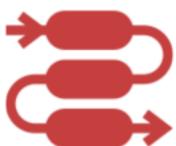
3. Numerical Features Standardization

StandardScaler was applied to numerical columns like accommodates and bedrooms to ensure consistency in scale, preventing large values from skewing the model.



4. Categorical Features Encoding

TargetEncoder was used for columns like property_type to manage feature space efficiently, avoiding memory overhead while maintaining relationships with the target variable.



5. ColumnTransformer Integration

Combined various preprocessing methods into a single ColumnTransformer to streamline the pipeline, enhancing modularity and extensibility.



6. Pipeline Application

The pipeline was fitted on training data and transformed both training and test sets, ensuring consistent data shapes and performance evaluation.

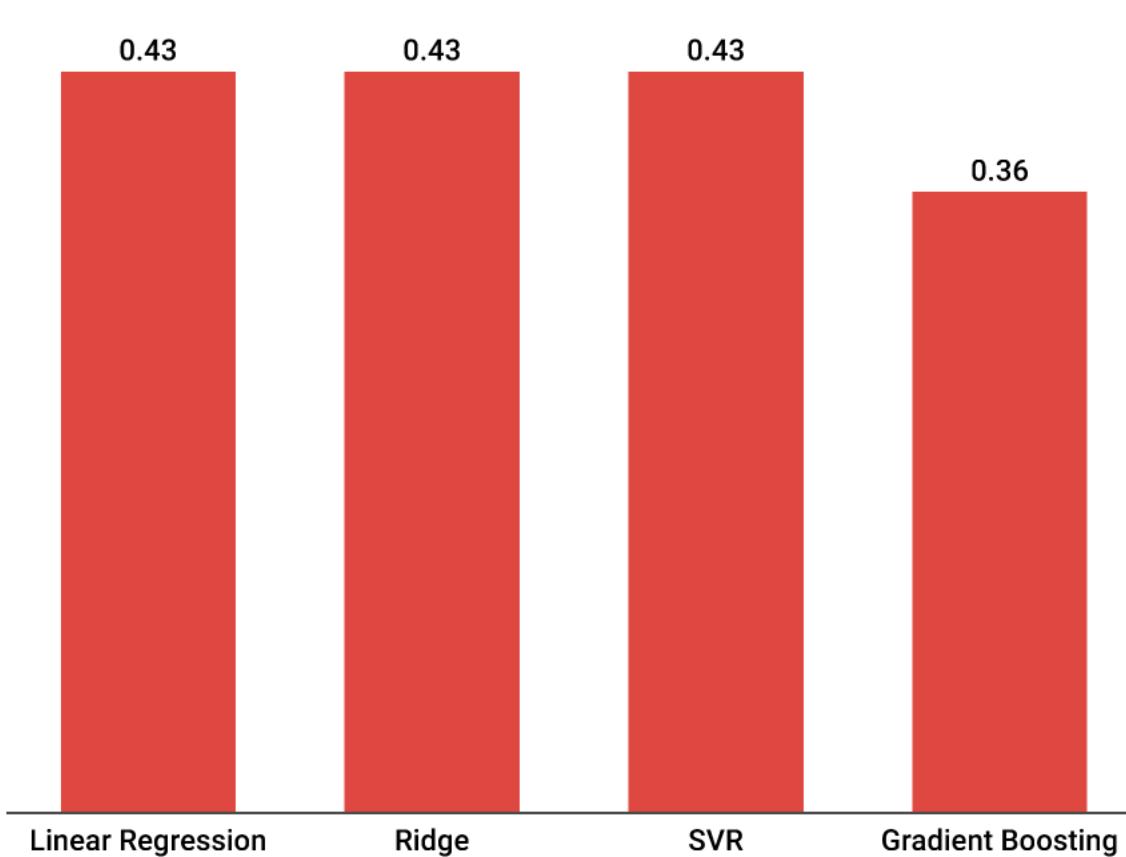


7. Key Insights

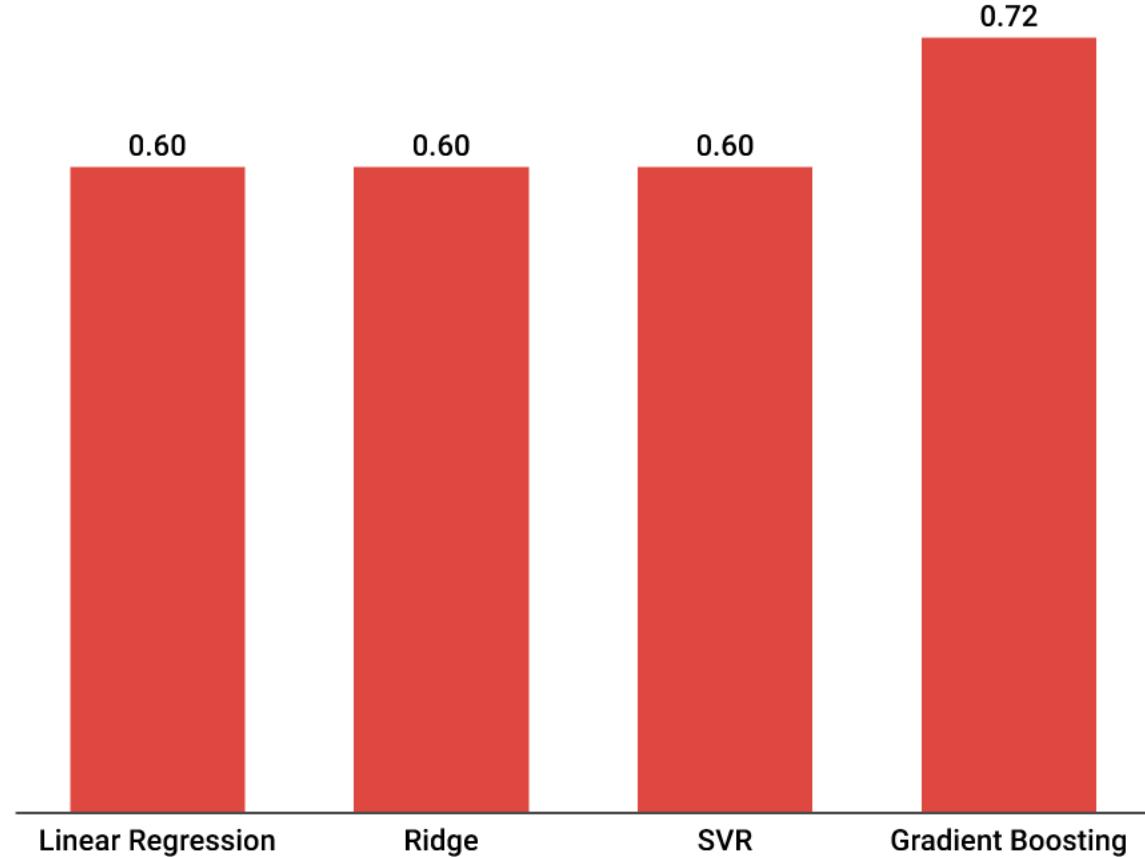
Avoiding OneHotEncoder minimized memory usage, while proper encoding and scaling contributed to improved model efficiency and convergence speed.



Machine Learning Model Evaluation



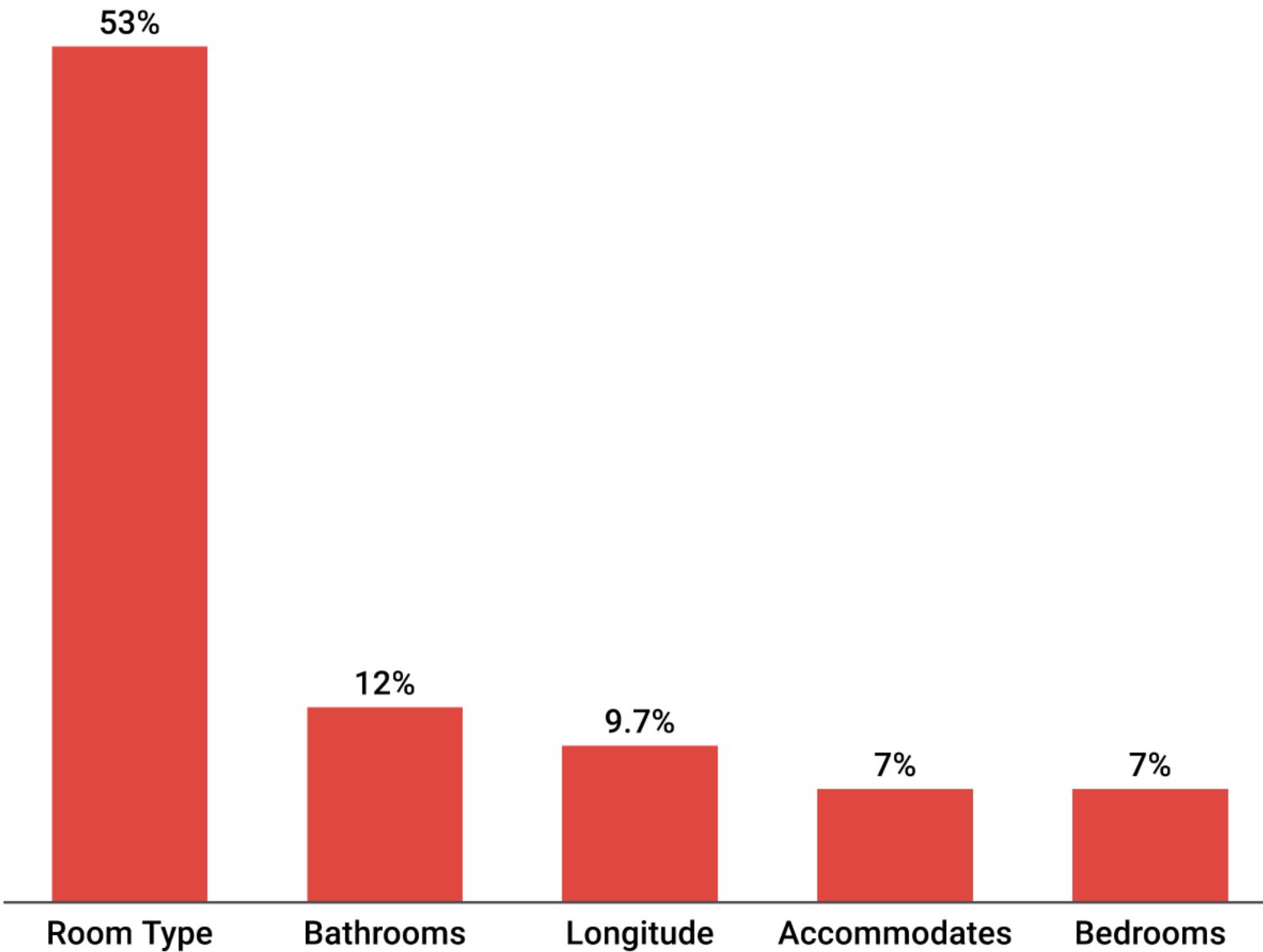
Comparison of RMSE Scores



Comparison of R² Scores



Feature Importance - Gradient Boosting Insights



Model Fine-Tuning

Hyperparameter Optimization with Randomized Search

01	learning_rate	0.1
02	max_depth	5
03	max_features	None
04	min_samples_leaf	1
05	min_samples_split	10
06	n_estimators	150
07	sub_sample	0.8



Model Evaluation & Final Results

Comprehensive Overview of Predictive Model Performance



Model Evaluation Metrics

Root Mean Square Error (RMSE) on training data is **0.33**, indicating a **strong fit**.

The R² score of 0.76 highlights the model explains a significant portion of variance in the data.



Best Model

The Gradient Boosting Regressor was identified as the **best-performing model** based on evaluation metrics, demonstrating its robustness in capturing complex patterns in the data.



Test Predictions

Predictions made by the model show a **close alignment** with actual log prices, reinforcing the reliability of the model in real-world scenarios. RMSE on test data is **0.34**, indicating a strong fit.



Insight

By fine-tuning model parameters, we achieved **high accuracy** in predicting listing prices, successfully reducing prediction error by over **8.5%**. This reflects the model's adaptability and effectiveness.

Challenges

Data Quality:

Addressing **missing values**, **iNcoNsisTENT** forMATS, and **outliers** required significant effort.



Processing Time:

Parallel processing resolved **delays** caused by sequential cross-validation, improving computation time significantly



Scalability:

Ensuring computational **efficiency** for larger datasets remains a critical challenge



Recommendations



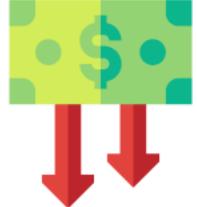
Empower Hosts with Data-Driven Insights:

Provide tools to help hosts analyze pricing trends and compare their listings against competitors. Offer suggestions to optimize property descriptions and amenities for better appeal.



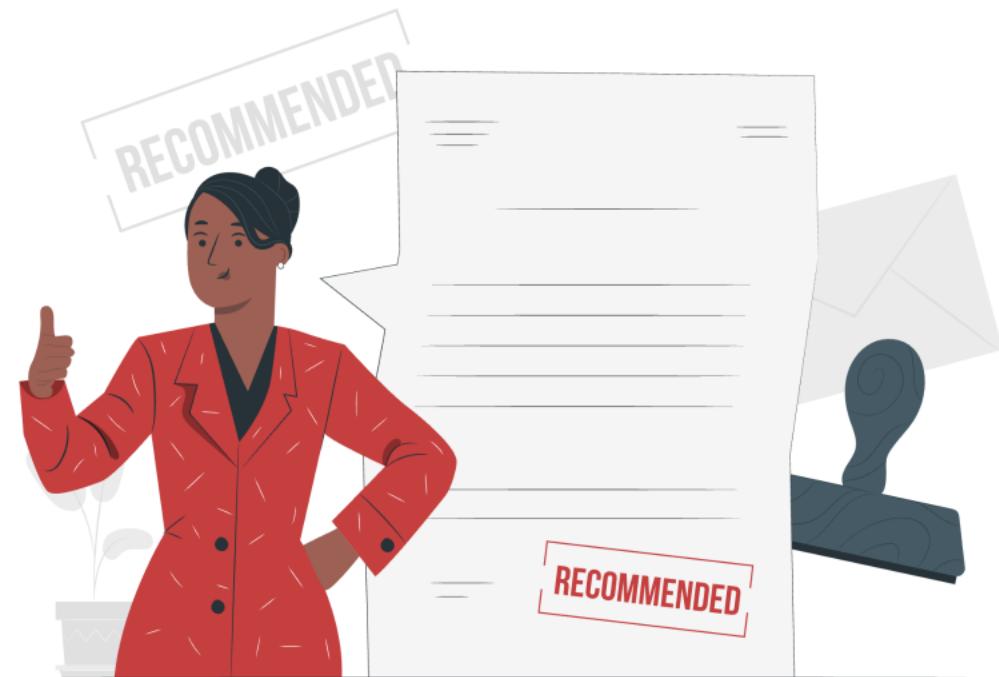
Leverage Geographical Insights:

Use clustering and ranking to recommend property optimizations and adjust pricing strategies by location.



Implement Advanced Pricing Strategies:

Introduce machine learning-driven pricing tools for dynamic adjustments based on: Seasonality, demand, and competitor analysis.



Target Niche Segments:

Promote premium and unique listings through targeted marketing campaigns to attract high-value customers.



Future Steps

Enhanced Property Features:

Develop advanced ranking systems using zip code-based clustering and price normalization to identify growth opportunities.



Sentiment Analysis:

Examine customer reviews to extract sentiments and their impact on pricing and occupancy.

Expanding Dataset Scope:

Incorporate new variables like local events, weather conditions, and transportation access.



Demand Prediction and Seasonality:

Model seasonal trends using time-series data and optimize peak-period pricing dynamically.

