The University of Texas at Austin
McCombs School of Business

100

Airbnb Price Prediction

Exploring patterns in airbnb dataset to understand the implications of different factors that affect the individual airbnb price

Anudeep Kumar, Chu Nie, Aishwarya Sarkar, Yingjia Shang
MIS 381N Final Project, Summer 2022

Traveling after summer classes?

Airbnb has revolutionized the travel industry with simple & convenient places to stay.

Hosts — How to list properties to generate additional income?

Travelers — What features should pay attention to find the optimally priced property?

Agenda

Project Goals

Exploratory Data Analysis

Data Preprocessing

Modeling & Conclusion

Project Goals
Exploratory Data Analysis
Data Preprocessing
Modeling & Conclusion

Data Description

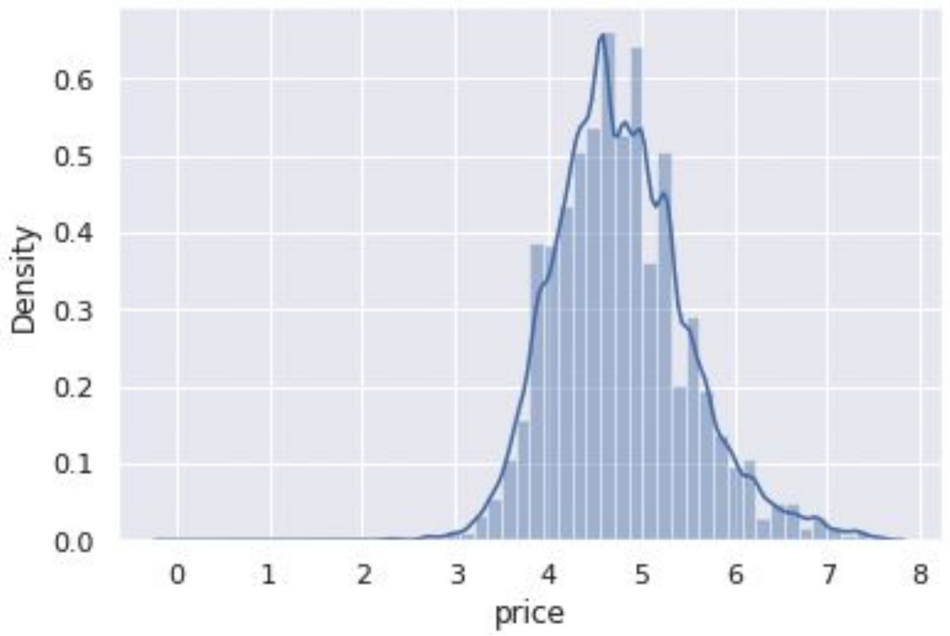
Source: Kaggle (Dataset Here)

Number of Features: **29**

- Number of numerical features: **9**
- Number of categorical features: **18**
- Number of date features: **2**

Target: log_price

The dataset consists of **74,111** records



Features

Rating/Review Feature
first_review
last_review
number_of_reviews
review_score_rating

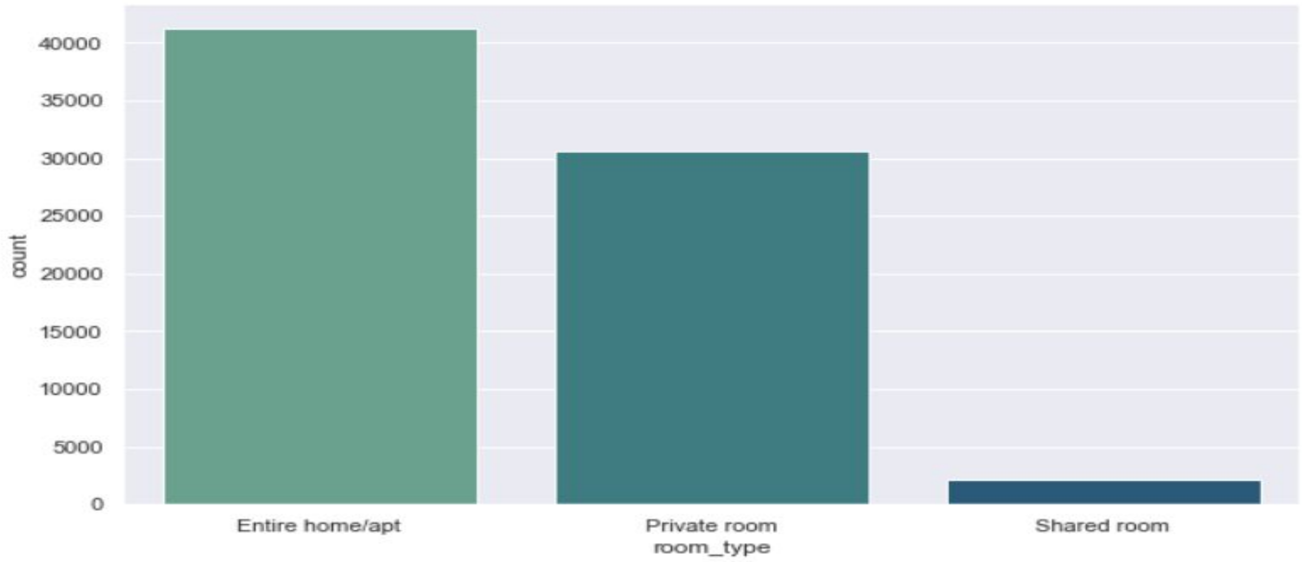
Host-related Feature
cancellation_policy
host_has_profile_pic
host_identity_verified
host_response_rate
host_since

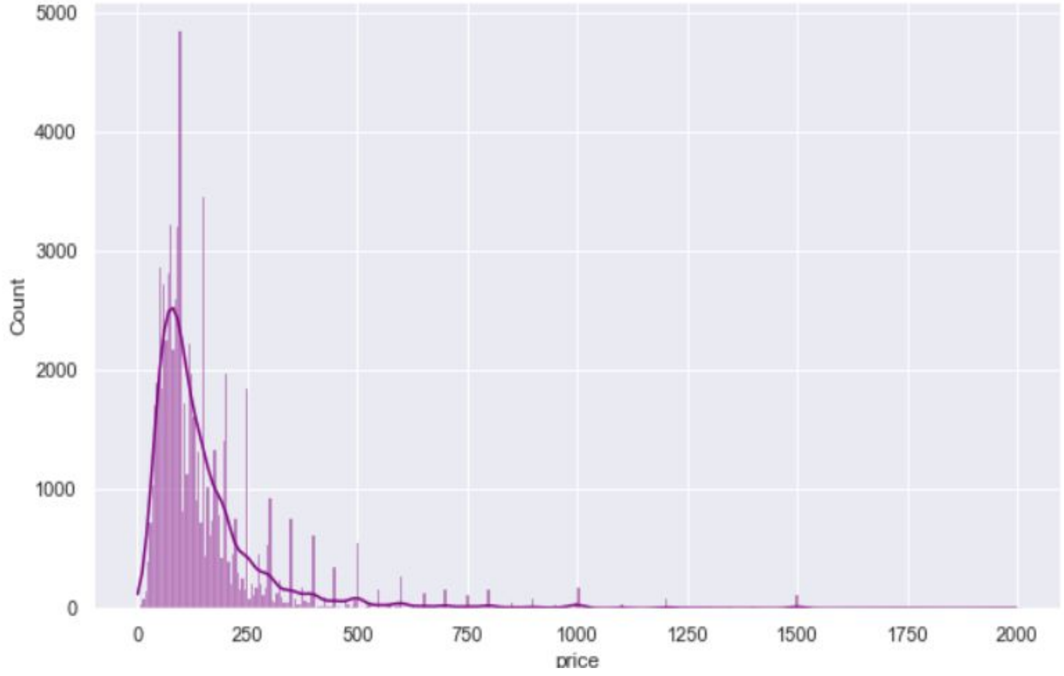
Location Feature
city
description
latitude
longitude
neighbourhood
zipcode

Property Feature
id
log_price
property_type
room_type
amenities
accommodates
bathrooms
bed_type
cleaning_fee
instant_bookable
bed
bedrooms
thumbnail_url
name

Exploratory Data Analysis

- Only 3% of the listings are for shared rooms
- 97.2% have real beds
- 73.9% of the listings are in NYC and LA
- Host response rate has a mean of 94.3%
- ~30% have a flexible cancellation policy
- Review scores > 85

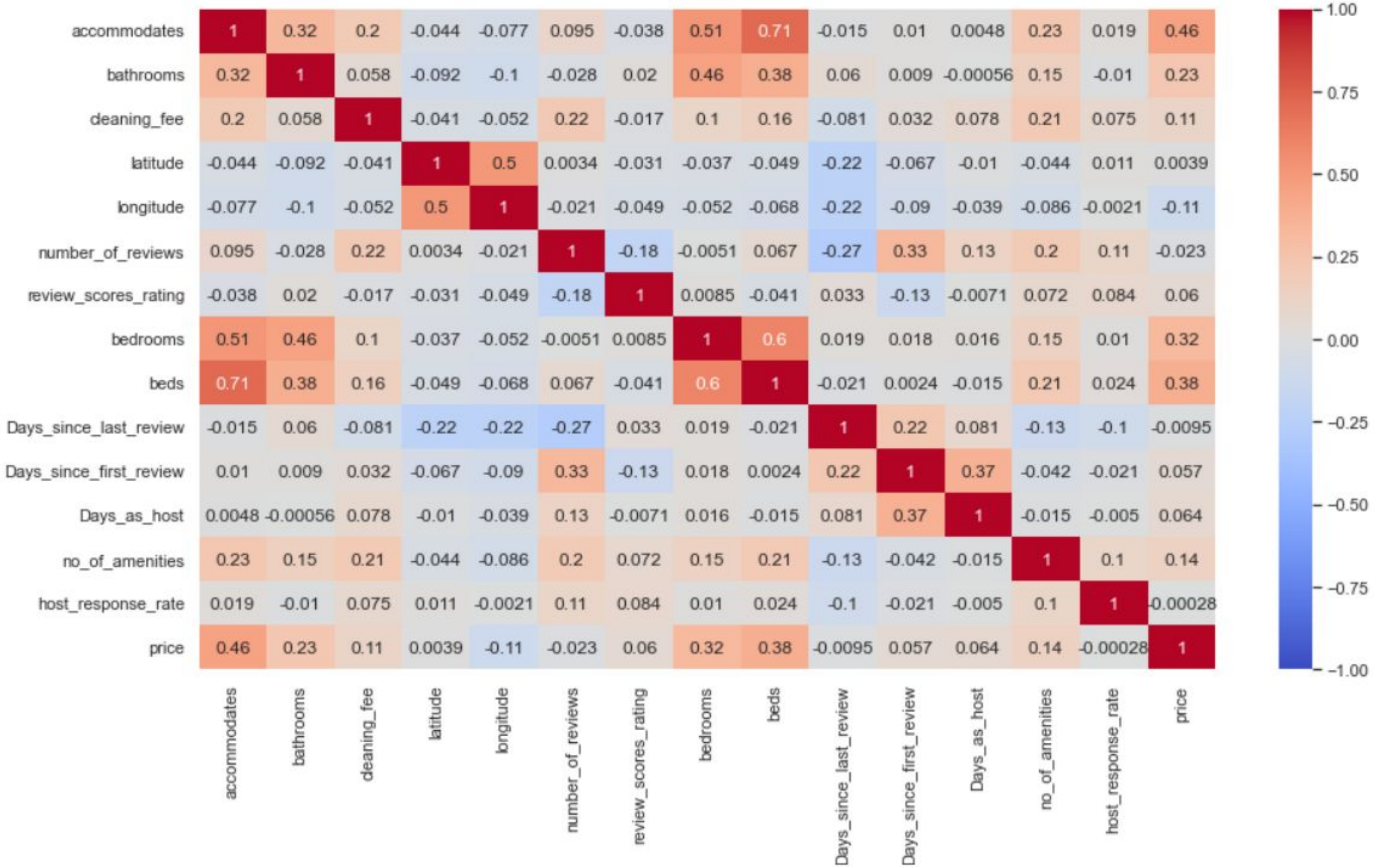




Exploratory Data Analysis

- More expensive airbnb's also have a stricter cancellation policy
- Number of bedrooms and bathrooms have a significant impact on price
- Host features seem to have a negligible effect on price
- San Francisco has the highest avg number of ratings per airbnb, followed by Chicago
- SF has the highest average median price of airbnb's
- NYC has the highest number of expensive neighborhoods followed by LA

Correlation between variables



- Number of people a room accommodates and bathrooms is correlated to the number of bedrooms
- Cleaning fee has a positive correlation with number of amenities offered
- Number of reviews is highly correlated with number of days elapsed since first review
- Number of bedrooms has a correlation with price

Data Preprocessing

Drop Columns

Dropped columns with **no predicting power** or **duplicate usage**

Drop Null

Drop records with **null values**

Transform Variables

Encoding **categorical variables** and **dates**

Split Train/Test

Split the data into **80% training data** and **20% testing data**

Linear Regression

Results Summary	
RMSE	0.4096
MAPE	0.0672
R^2	0.63

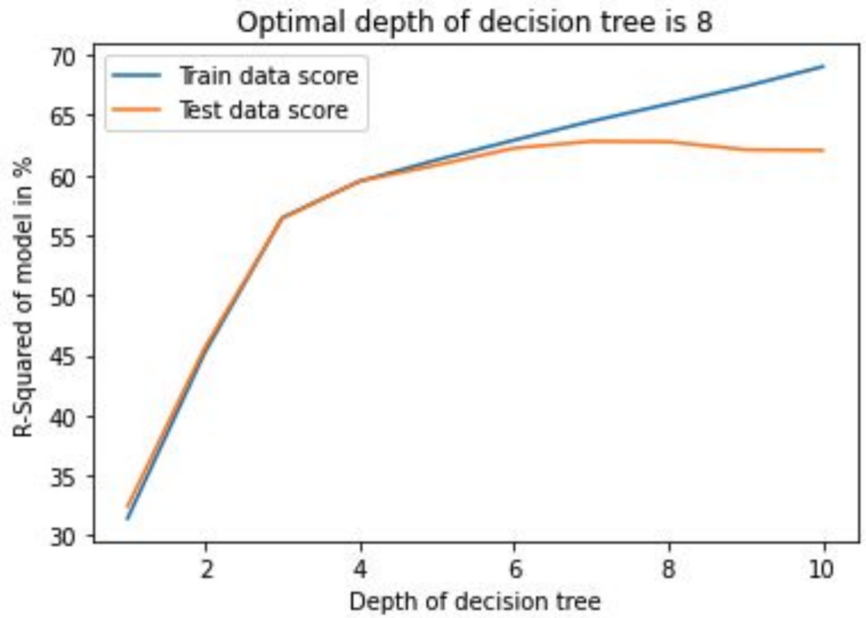
- We used all features to fit linear regression model
- Features with p-value > 0.05:
  - number_of_reviews
  - property_type: boutique hotel, bungalow, cabin, cave, chalet, earth house, island, serviced apartment, treehouse, villa, yurt, other
  - cancellation_policy_moderate
  - cleaning_fee

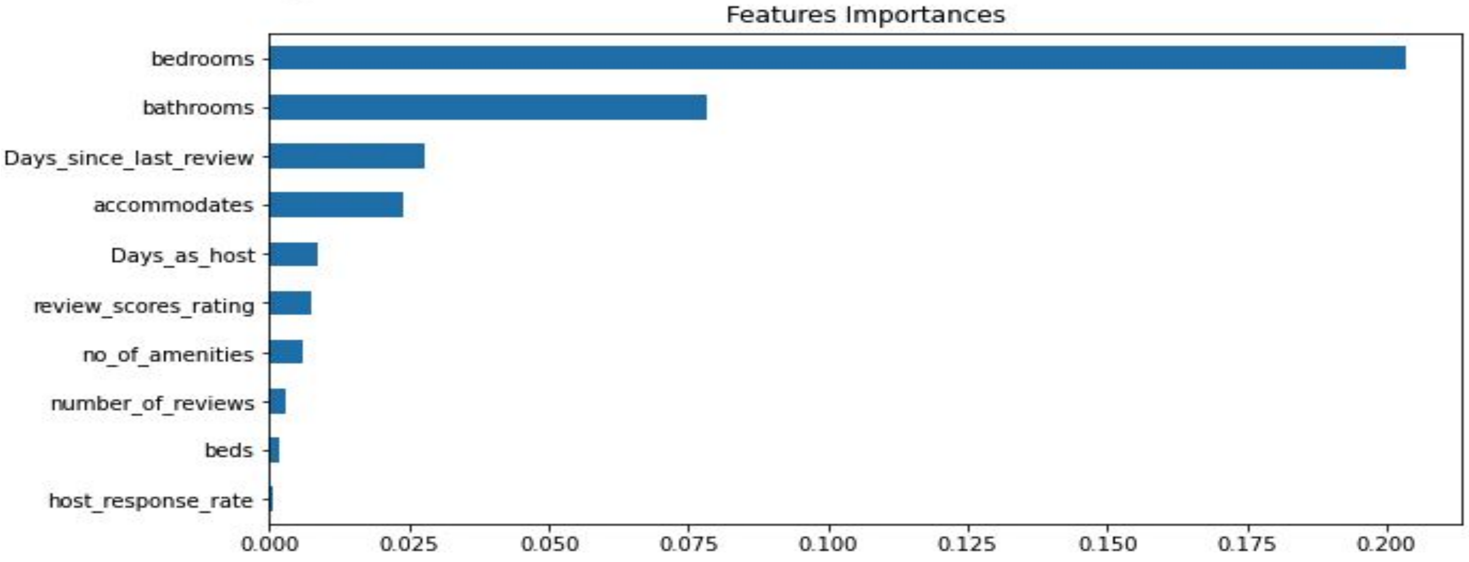
	coef	std err	t	P>ltl
accommodates	0.0793	0.002	44.139	0.000
bathrooms	0.1476	0.004	34.697	0.000
host_response_rate	0.0008	0.000	5.425	0.000
number_of_reviews	-9.315e-05	4.81e-05	-1.935	0.053
review_scores_rating	0.0110	0.000	42.735	0.000
bedrooms	0.1376	0.004	37.491	0.000
beds	-0.0409	0.003	-14.734	0.000
Days_since_last_review	0.0005	1.23e-05	43.528	0.000
Days_as_host	5.048e-05	3.24e-06	15.600	0.000
no_of_amenities	0.0066	0.000	21.808	0.000
property_type_Bed & Breakfast	0.0911	0.023	3.889	0.000
property_type_Boat	0.2359	0.063	3.754	0.000
property_type_Boutique hotel	0.1288	0.069	1.855	0.064
property_type_Bungalow	-0.0347	0.026	-1.360	0.174
property_type_Cabin	-0.1244	0.055	-2.262	0.024
property_type_Camper/RV	-0.2338	0.052	-4.467	0.000
property_type_Castle	0.3486	0.117	2.988	0.003
property_type_Cave	0.2649	0.297	0.891	0.373
property_type_Chalet	0.1015	0.188	0.540	0.590
property_type_Condominium	0.0950	0.011	9.006	0.000
property_type_Dorm	-0.4148	0.042	-9.780	0.000
property_type_Earth House	0.0834	0.243	0.344	0.731
property_type_Guest suite	-0.1229	0.042	-2.896	0.004
property_type_Guesthouse	-0.0645	0.021	-3.001	0.003
property_type_Hostel	-0.5097	0.058	-8.791	0.000
property_type_House	-0.0595	0.005	-11.396	0.000
property_type_Hut	-0.3741	0.159	-2.353	0.019
property_type_In-law	-0.2195	0.053	-4.119	0.000
property_type_Island	0.8016	0.420	1.907	0.057
property_type_Loft	0.1478	0.015	10.169	0.000
property_type_Other	0.0318	0.021	1.478	0.139
property_type_Serviced apartment	0.1899	0.109	1.749	0.080
property_type_Tent	-0.2393	0.113	-2.125	0.034
property_type_Timeshare	0.4897	0.077	6.368	0.000
property_type_Tipi	0.6421	0.243	2.643	0.008
property_type_Townhouse	-0.0327	0.013	-2.597	0.009
property_type_Train	0.6677	0.297	2.246	0.025
property_type_Treehouse	0.3765	0.210	1.790	0.073
property_type_Vacation home	0.3694	0.172	2.152	0.031
property_type_Villa	0.0476	0.038	1.252	0.211
property_type_Yurt	0.1708	0.172	0.995	0.320
room_type_Private room	-0.5956	0.005	-123.024	0.000
room_type_Shared room	-1.0389	0.013	-78.424	0.000
bed_type_Couch	0.5274	0.044	11.871	0.000
bed_type_Futon	0.4860	0.031	15.687	0.000
bed_type_Pull-out Sofa	0.5555	0.032	17.270	0.000
bed_type_Real Bed	0.5870	0.025	23.815	0.000
cancellation_policy_moderate	0.0076	0.006	1.296	0.195
cancellation_policy_strict	0.0411	0.006	7.465	0.000
cancellation_policy_super_strict_30	0.2096	0.048	4.339	0.000
cancellation_policy_super_strict_60	0.7239	0.133	5.427	0.000
cleaning_fee_t	-0.0017	0.005	-0.322	0.748
city_Chicago	-0.3475	0.012	-29.947	0.000
city_DC	-0.1398	0.011	-12.371	0.000
city_LA	-0.1792	0.010	-18.750	0.000
city_NYC	0.0453	0.009	4.992	0.000
city_SF	0.3028	0.011	27.905	0.000
host_has_profile_pic_t	1.5330	0.035	43.542	0.000
host_identity_verified_t	-0.0232	0.005	-4.980	0.000
instant_bookable_t	-0.0108	0.004	-2.451	0.014

Decision Tree Regressor

Results Summary	
RMSE	0.1702
MAPE	0.0674
R^2	0.63

- Max depth of tree considered: 8
- Top 5 most important features: bathrooms, bedrooms, number of days since last review, number of people that can be accommodated and number of days as host



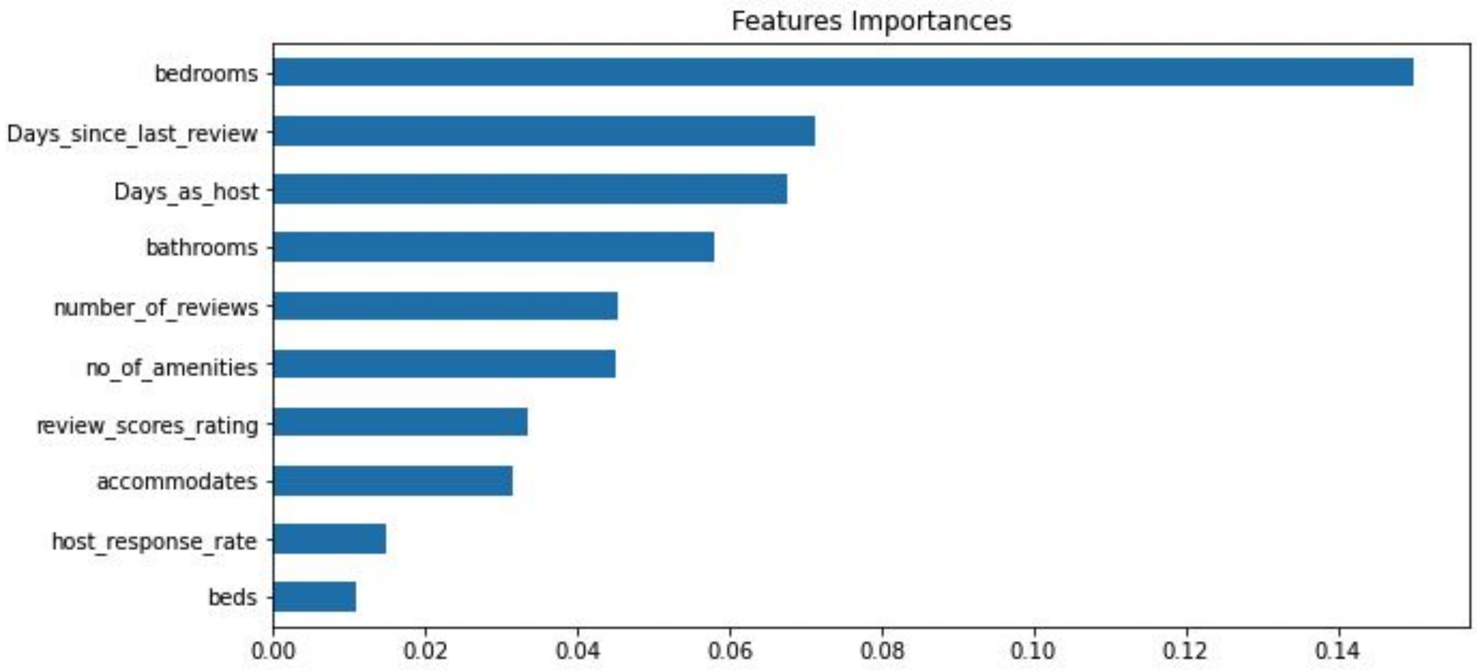


Random Forest

Results Summary	
RMSE	0.3880
MAPE	0.063
R^2	0.67

Values of Parameters selected after tuning:

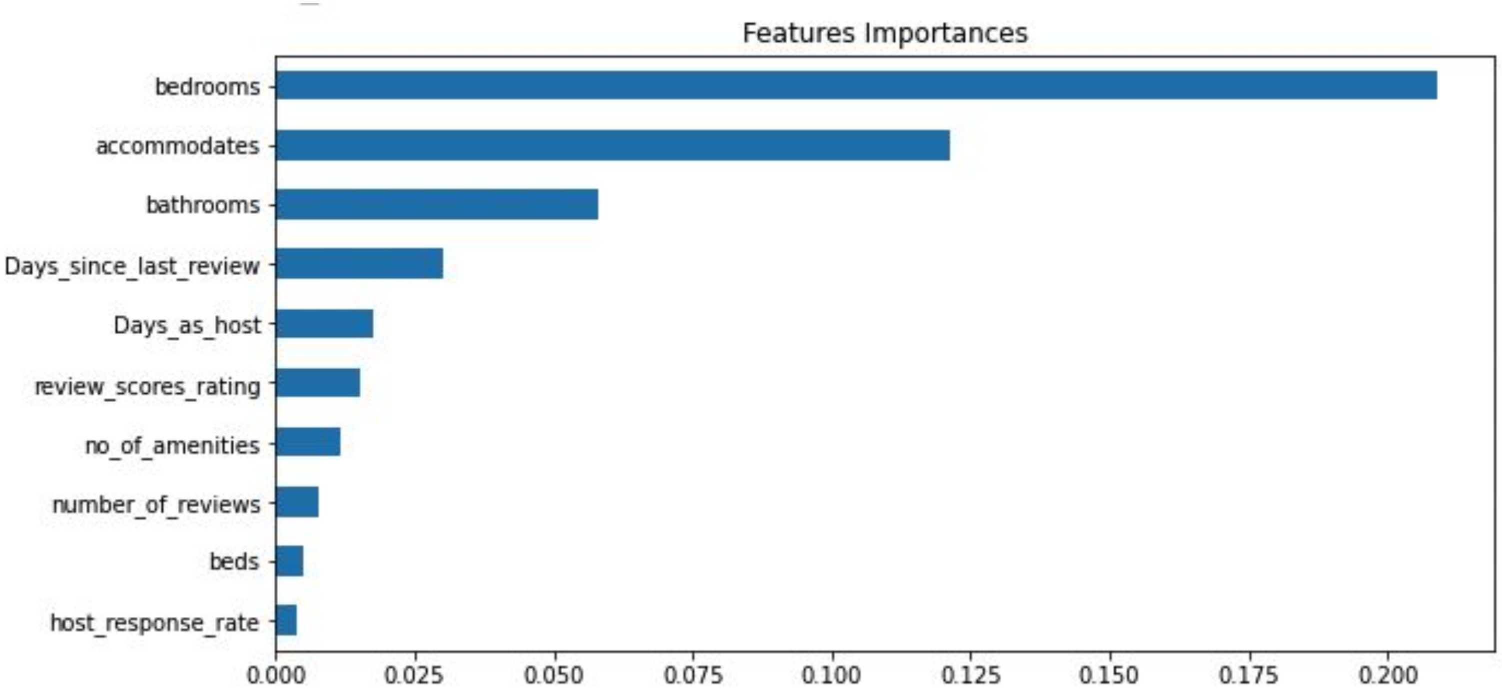
- n_estimators = 300
- max_depth=80
- random_state = 42



Gradient Boosting

Results Summary	
RMSE	0.3844
MAPE	0.0631
R^2	0.68

- Parameters tuned:
  - n_estimators = 1000
  - max_features = ‘auto’



Model Performance & Output Comparison

	Linear Regression	Decision Tree	Random Forest	Gradient Boosting
RMSE	0.4096	0.1702	0.3880	0.3844
MAPE	6.72%	6.73%	6.31%	6.31%
R^2	0.63	0.63	0.67	0.68

Insights & Conclusion

- **Gradient Boosting** yielded the best R-squared result, followed by Random Forest, Decision Tree, and OLS Linear Regression.
- Overall, **bedrooms, bathrooms, days_since_last_review and days_as_host** are the top 4 features with the highest importances.

Thanks!

Any Questions?