Predicting Airbnb prices in US

Mahshid AMIR

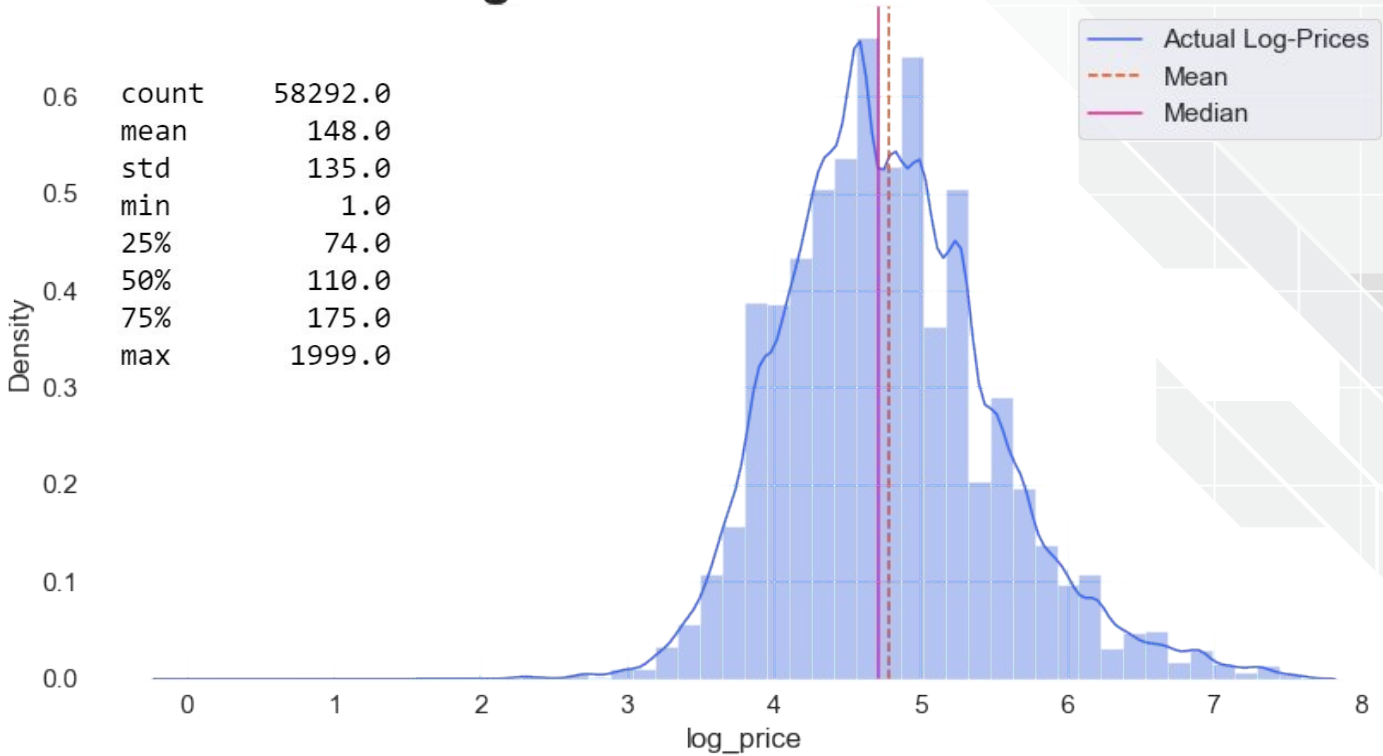
Predict AirBnB Price in major US cities

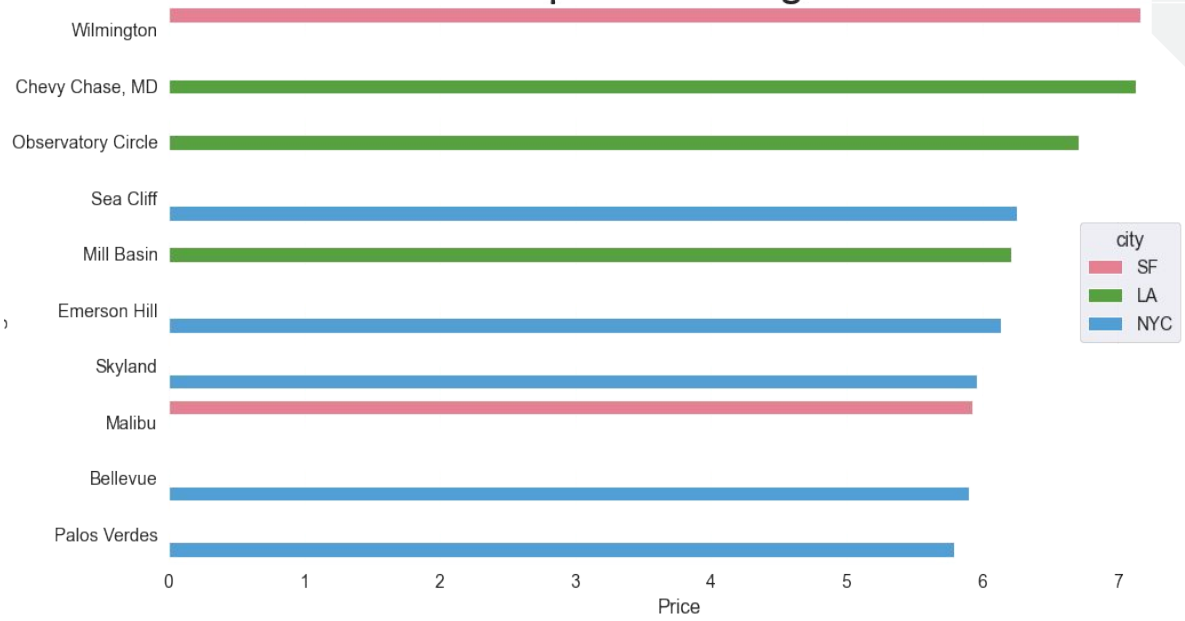
- Location
- Property type
- Amenities
- Host profile
- Reviews
- etc

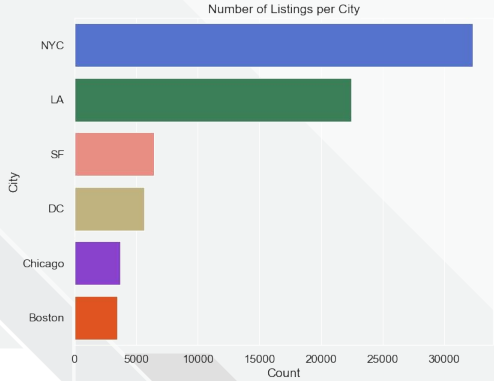
Exploring the Data

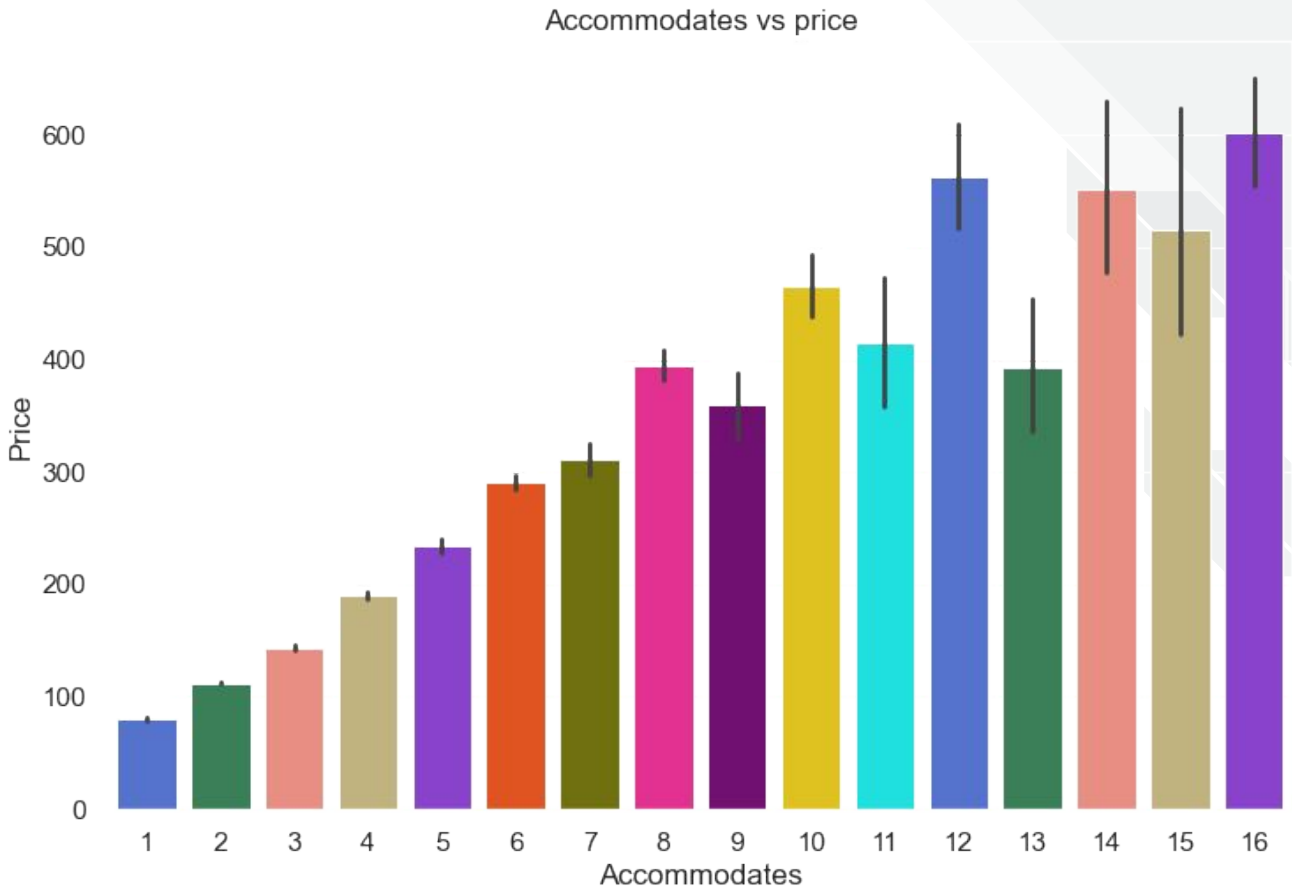
Log-Price Distribution Plot

count	58292.0
mean	148.0
std	135.0
min	1.0
25%	74.0
50%	110.0
75%	175.0
max	1999.0

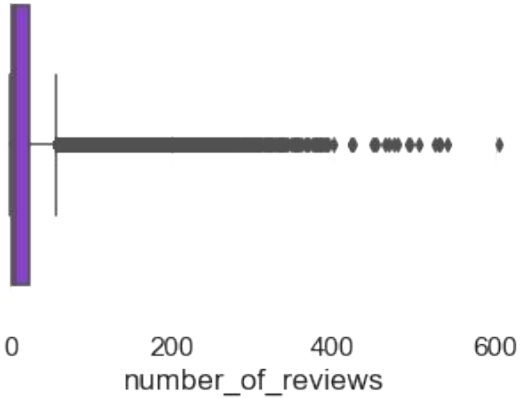


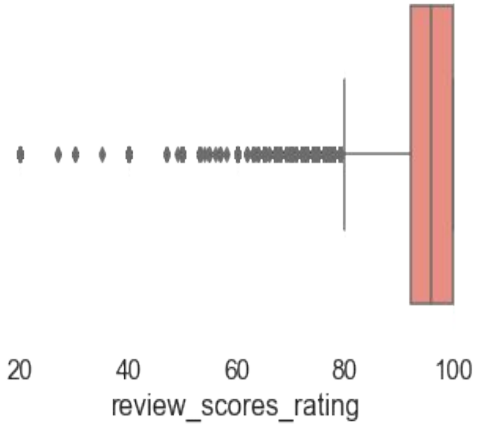

10 most expensive neighborhoods.
Wilmington
Chevy Chase, MD
Observatory Circle
Sea Cliff
Mill Basin
Emerson Hill
Skyland
Malibu
Bellevue
Palos Verdes
0
1
2
3
4
5
6
7
Price
city
SF
LA
NYC


Number of Listings per City
NYC
LA
SF
DC
Chicago
Boston
City
0
5000
10000
15000
20000
25000
30000
Count


Accommodates vs price
Price
0
100
200
300
400
500
600
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Accommodates

Data Cleaning

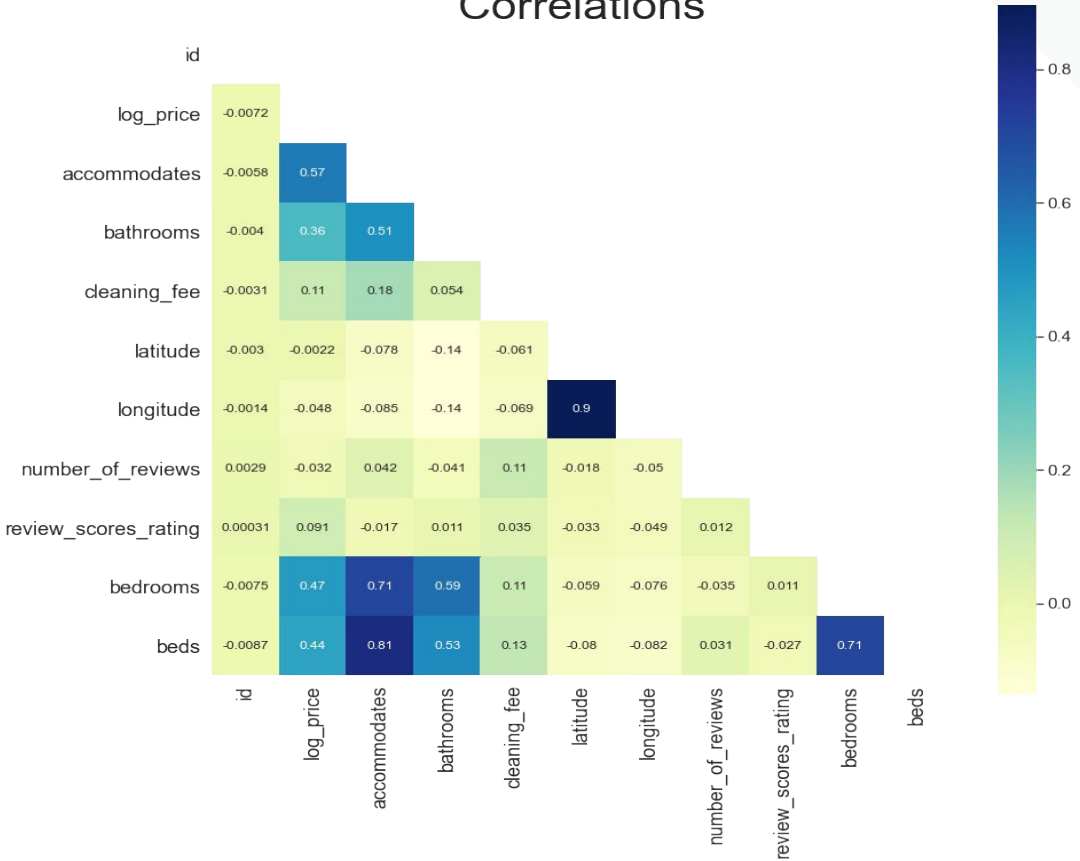




- Missing Values
- Outliers
- Incorrect values
- Transformation

	column_name	missing_count
15	host_response_rate	18299
24	review_scores_rating	16722
12	first_review	15864
18	last_review	15827
25	thumbnail_url	8216
22	neighbourhood	6872
26	zipcode	966
6	bathrooms	200
13	host_has_profile_pic	188
14	host_identity_verified	188
16	host_since	188
28	beds	131
27	bedrooms	91

Correlations



Data Preprocessing

Dummies, Features Transformations

Dummies Encoders

Room_type
Property_type
Cancellation_policy
City
Neighbourhoods
Amenities

Scalars

Number of reviews
Review scores
Host since rank
Accommodate
Distance

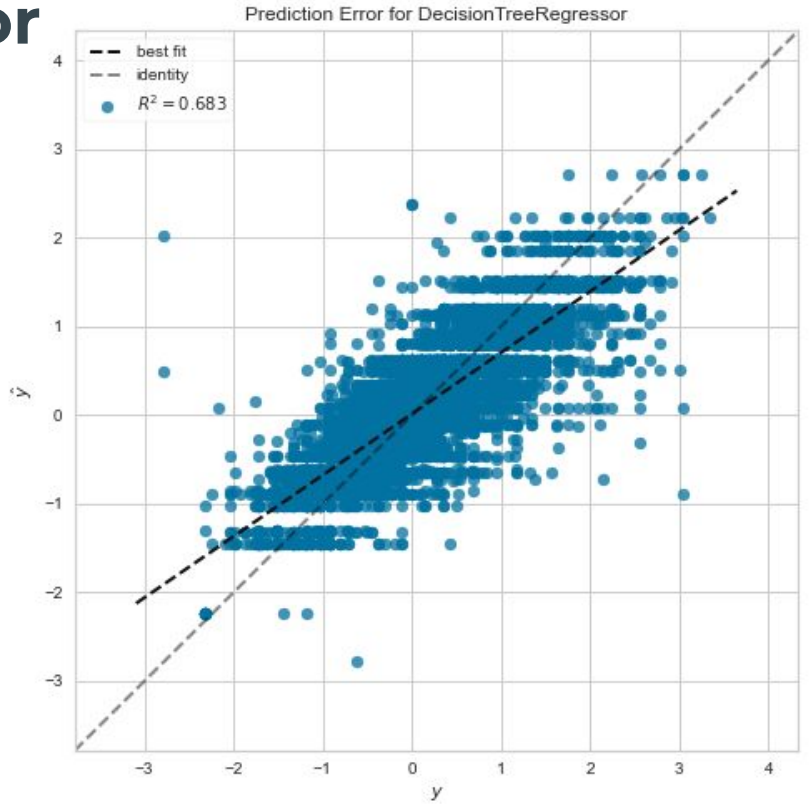
Features selection

	Features
1	accommodates
2	bathrooms
3	cleaning_fee
4	host_has_profile_pic
5	host_identity_verified
6	host_response_rate
7	instant_bookable
8	number_of_reviews
9	review_scores_rating
10	bedrooms
11	dist_to_citycenter
12	host_since_rank
13	amenities_frequency
14	neighbourhood_encoder
15	property_type_Condominium
16	property_type_House
17	property_type_Loft
18	property_type_OTHER_prop_type
19	property_type_Townhouse
20	room_type_Private room
21	room_type_Shared room
22	bed_type_Couch
23	bed_type_Futon
24	bed_type_Pull-out Sofa
25	bed_type_Real Bed
26	cancellation_policy_moderate
27	cancellation_policy_strict
28	cancellation_policy_super_strict_30
29	cancellation_policy_super_strict_60
30	city_Chicago
31	city_DC
32	city_LA
33	city_NYC
34	city_SF

Modeling

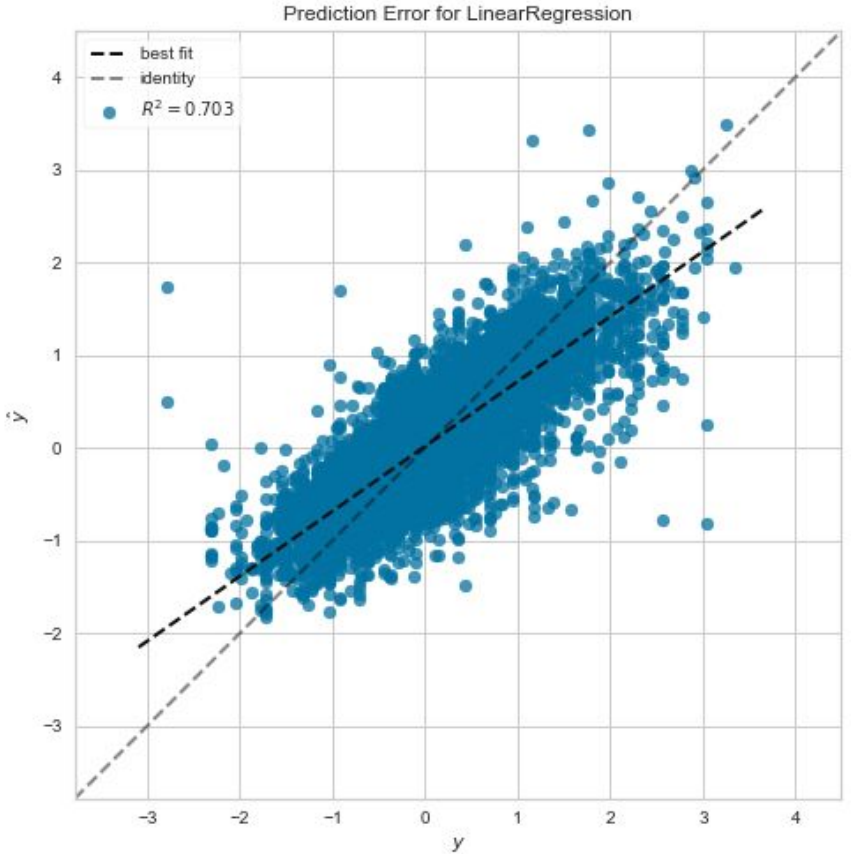
Decision Tree Regressor

R2 = 0.683



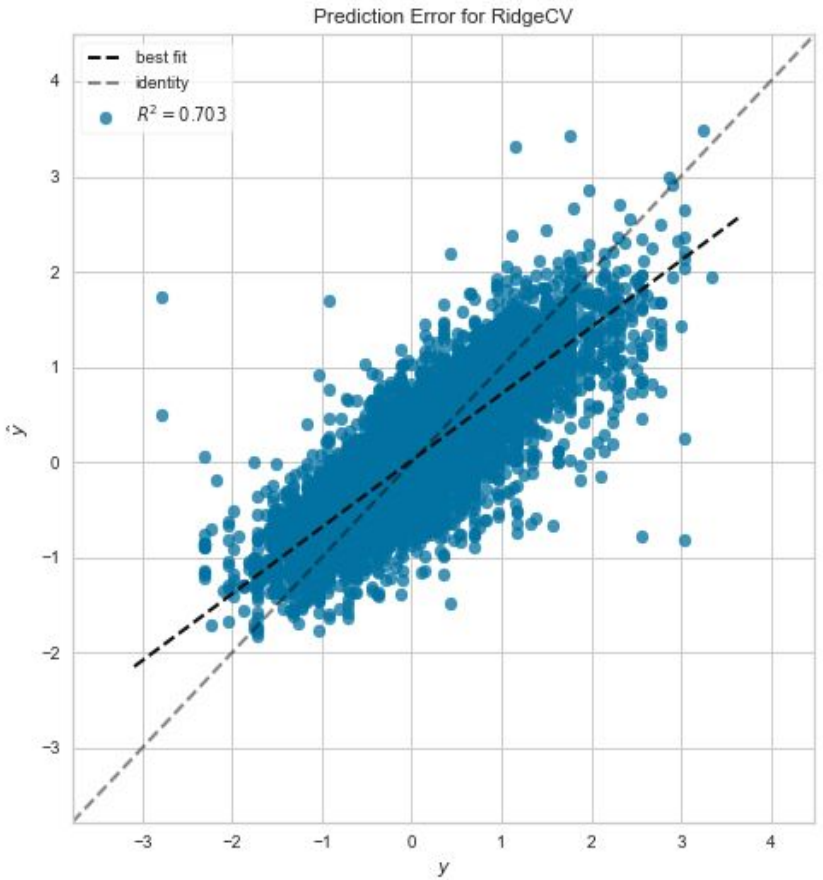
Linear Regression

R2 = 0.703



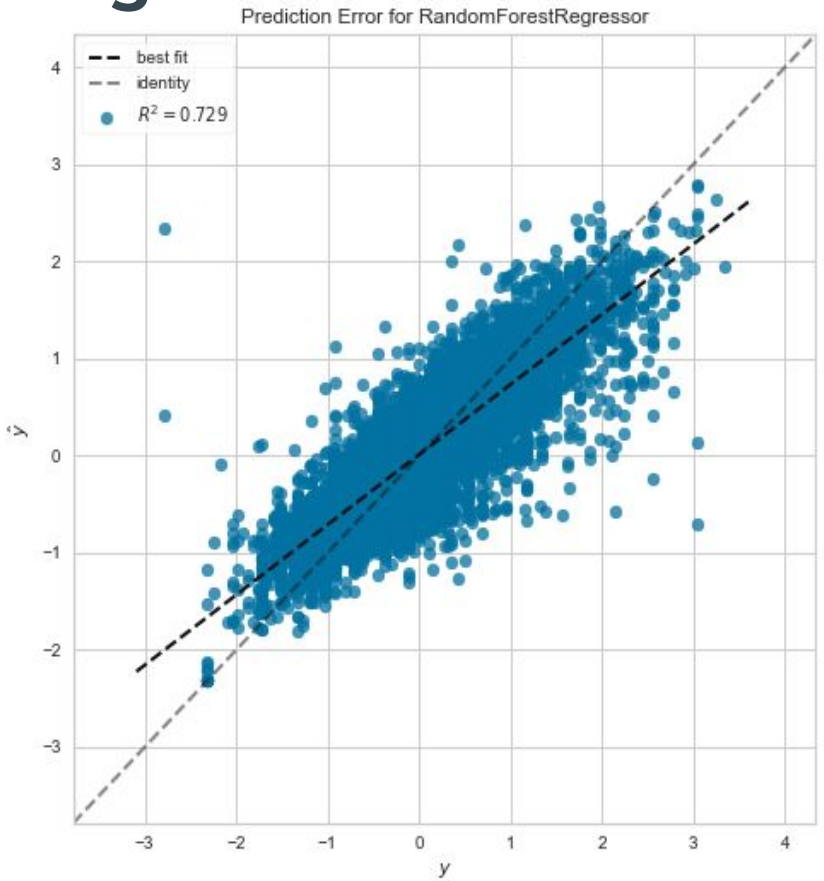
RidgeCV

R2 = 0.703



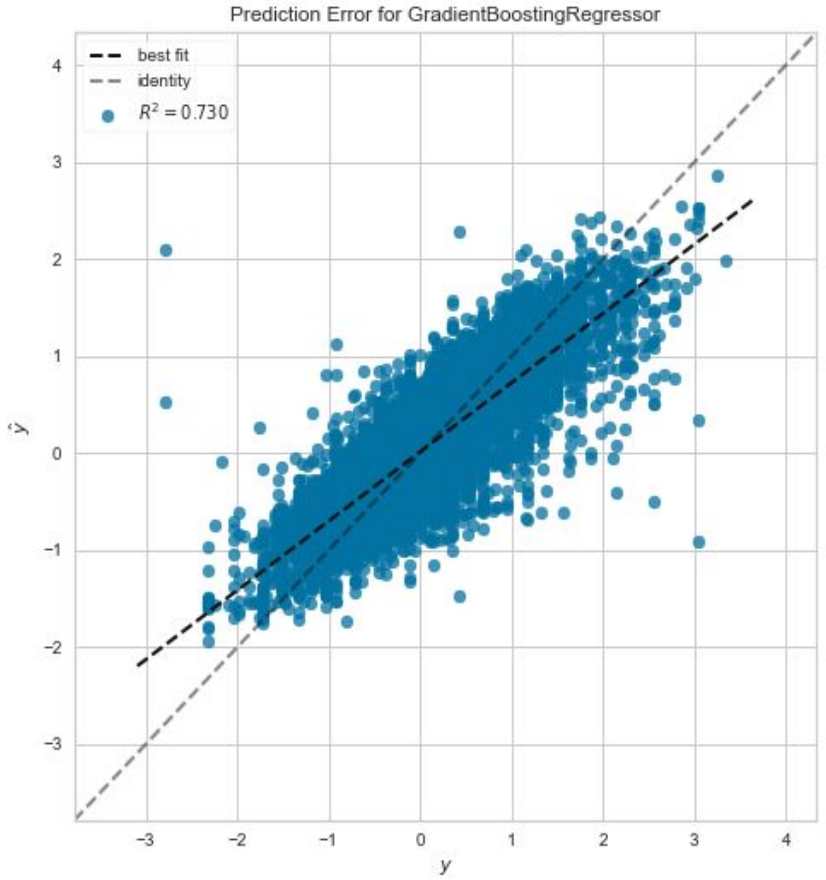
Linear Random Forest Regressor

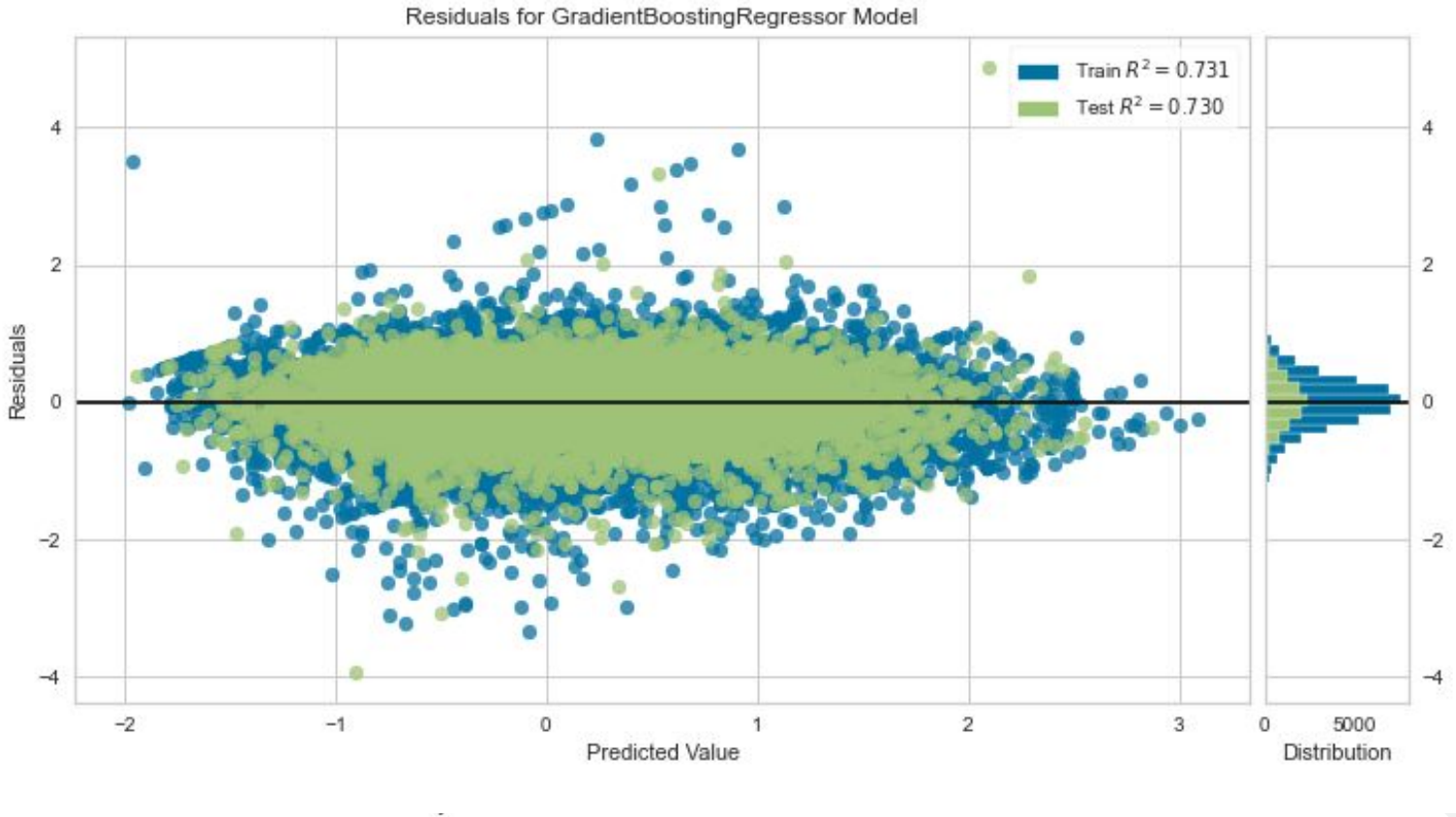
R2 = 0.729

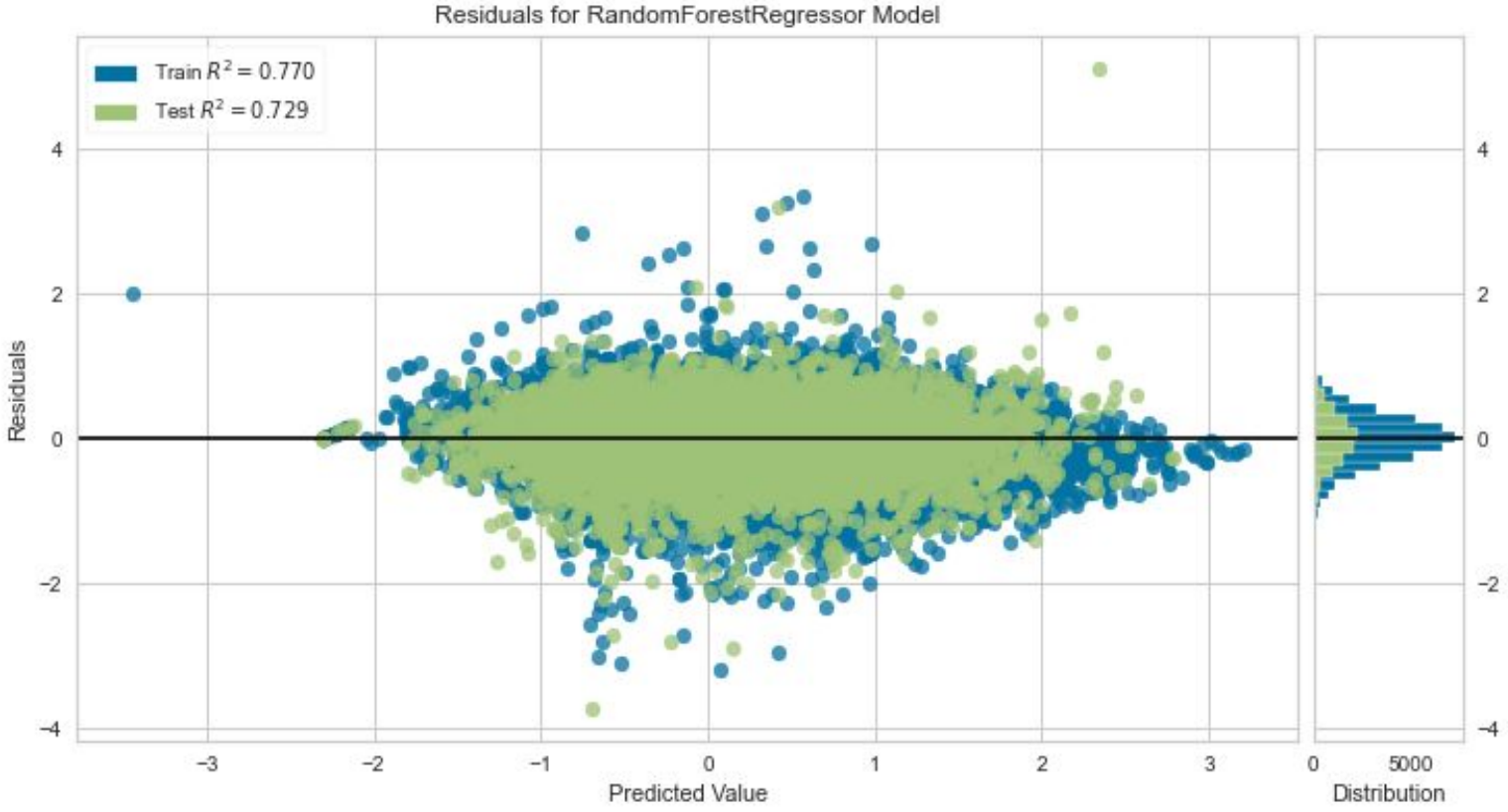


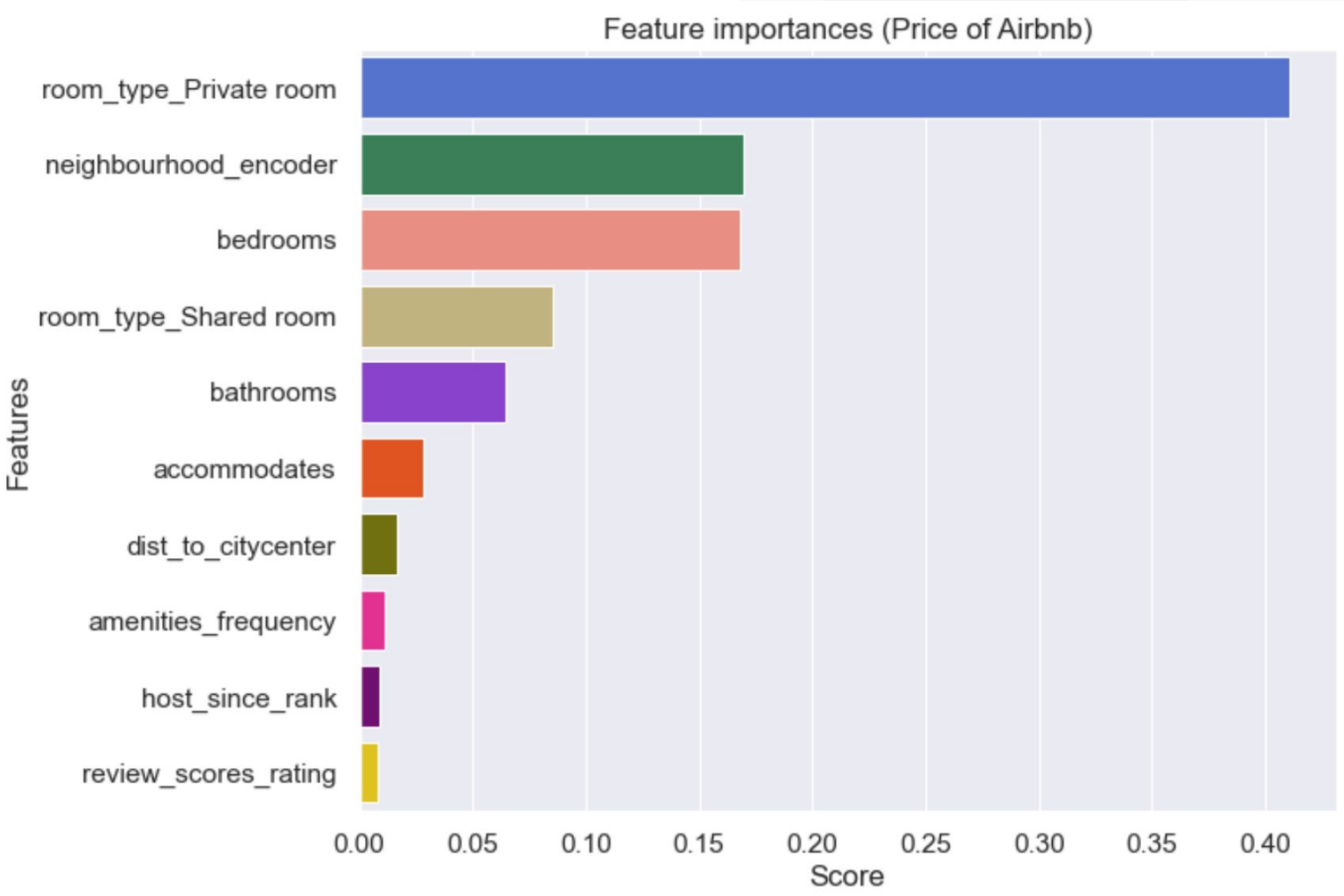
GBoost Regressor

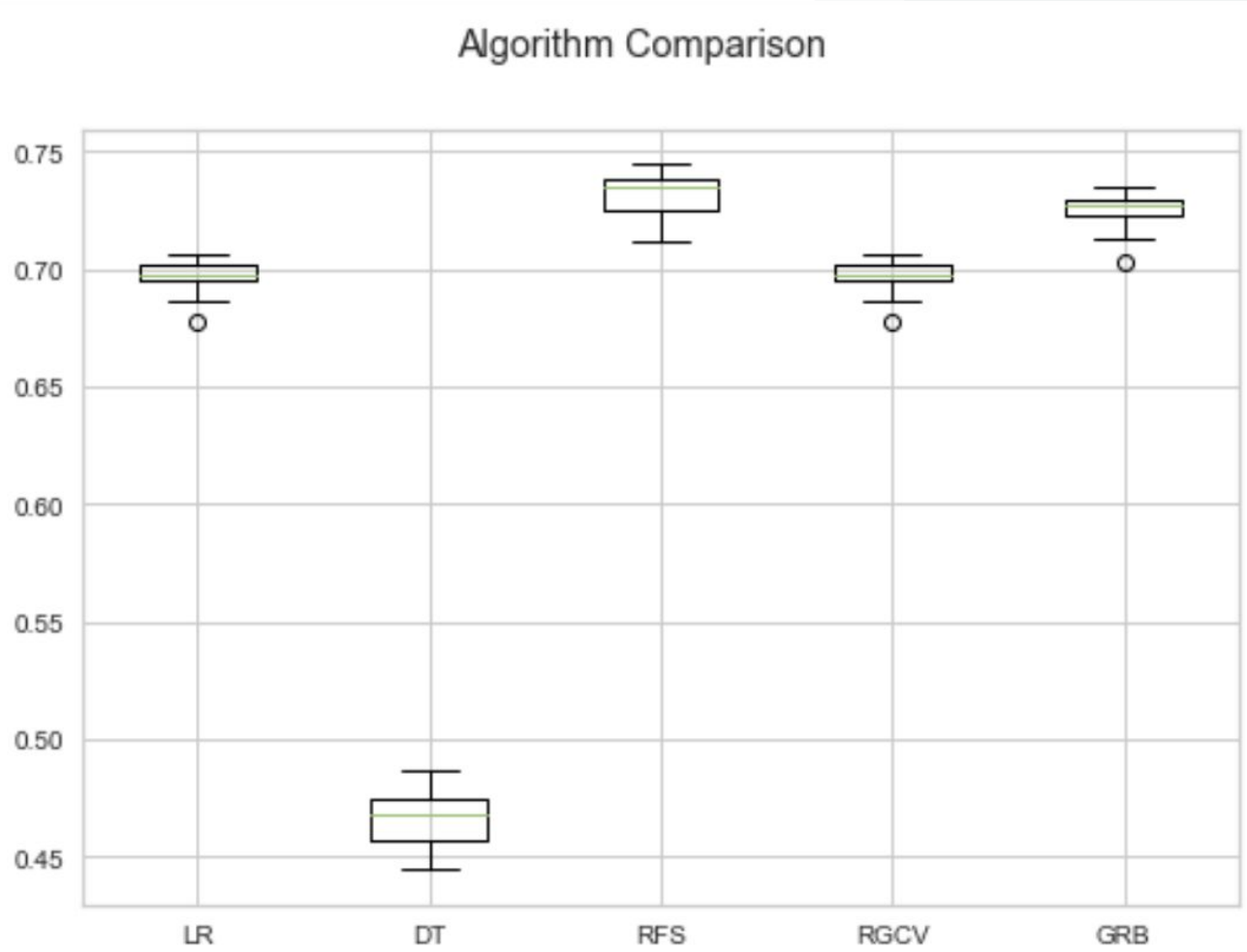
R2 = 0.730




Residuals for GradientBoostingRegressor Model
Train $R^2 = 0.731$
Test $R^2 = 0.730$
Residuals
Predicted Value
Distribution
4
2
0
−2
−4
−2
−1
0
1
2
3
0
5000


Residuals for RandomForestRegressor Model
Train $R^2 = 0.770$
Test $R^2 = 0.729$
Residuals
Predicted Value
Distribution
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
0
5000


Feature importances (Price of Airbnb)
room_type_Private room
neighbourhood_encoder
bedrooms
room_type_Shared room
bathrooms
accommodates
dist_to_citycenter
amenities_frequency
host_since_rank
review_scores_rating
Features
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Score


Algorithm Comparison
0.75
0.70
0.65
0.60
0.55
0.50
0.45
LR
DT
RFS
RGCV
GRB

Thank you!

Any questions?