Aizhan Borubaeva
Prediction with Machine Learning
Assignment 2
Airbnb prediction models
Technical report

The goal of the project is to help a company set an efficient price for their new apartments that just entered the market. The mentioned company operates small and mid-size apartments hosting 2-6 guests.

I used Airbnb data for Paris with a scraping method (Inside Airbnb. "City Name Dataset." Inside Airbnb, http://insideairbnb.com/ Paris). I chose the data for December 2022 for Task 1 and for September 2023 for Task 2.

I dropped missing values for price and the following columns: 'id', 'listing_url', 'scrape_id', 'last_scraped', 'source', 'name',
'description', 'neighborhood_overview', 'picture_url', 'host_id', 'bedrooms',
'host_url', 'host_name', 'host_since', 'host_location', 'host_about',
'host_response_time', 'host_response_rate', 'host_acceptance_rate',
'host_thumbnail_url', 'host_picture_url','host_neighbourhood',
'host_total_listings_count', 'host_verifications','host_has_profile_pic',
'host_identity_verified', 'neighbourhood', 'neighbourhood_group_cleansed',
'latitude', 'longitude','bathrooms', 'bathrooms_text', 'amenities',
'minimum_minimum_nights','maximum_minimum_nights', 'minimum_maximum_nights',
'maximum_maximum_nights', 'minimum_nights_avg_ntm','maximum_nights_avg_ntm',
'calendar_updated', 'availability_30', 'availability_60', 'availability_90',
'availability_365', 'first_review', 'last_review',
'calendar_last_scraped', 'number_of_reviews_l30d', 'license'
I dropped that columns because they have data that was not useful, a lot of missing values, data which were difficult to interpret or data very similar to other ones (dublicated).

Then I kept only ['room_type'] == 'Entire home/apt'] according to assignment insrtuctions.Because I don't need this column later, I dropped it.

Keep only data for accomodates from 2 till 6 according to assignment instructions

Keep 'property_type' == 'Entire rental unit' or 'Entire condo') or 'Entire loft') or 'Entire serviced apartment'

Additionally, I created dummies for categorical variables. Because there were some extreme values for the price (see Table 1), I kept only the price up to USD 1.500.

count	43771.000000
mean	157.960625
std	581.055713
min	9.000000
25%	77.000000
50%	110.000000
75%	167.000000

max 99140.000000

Created values for missing review scores rating and beds

Change all data to numeric variables to avoid data errors later

dropped missing variables to avoid future errors in models (Random Forest, Boosting)

Used random seed and split data to training and test datasets

Created feols models to choose the best for final OLS and calculated RMSE and R-squared:

M1: RMSE: 105.913 Adj. R2: 0.178 Adj. R2 Within: 0.178
M2: RMSE: 101.132 Adj. R2: 0.25 Adj. R2 Within: 0.25
M3: RMSE: 100.157 Adj. R2: 0.264 Adj. R2 Within: 0.264

Create OLS Model, Randon Forest and Gradient Boosting, calculated BIC manually:

	RMSE	R-squared	BIC
Linear Regression	96.725764	0.336366	62889.784113
Random Forest	85.549885	0.480862	63077.612877
Gradient Boosting	84.416858	0.494521	62889.784113

Scraped data for September
url = "http://data.insideairbnb.com/france/ile-de-france/paris/2023-09-04/data/listings.csv.gz"
repeat all steps with data cleaning

Check the same LS Model, Randon Forest and Gradient Boosting models. calculated RMSE, R squared and BIC:

	RMSE	R-squared	BIC
Linear Regression	109.877101	0.375464	79583.496555
Random Forest	103.403791	0.446884	79950.968573
Gradient Boosting	101.211479	0.470089	79583.496555

Calculated SHAP values for a set of samples, build a graph

Code:
https://github.com/Aborubaeva/Prediction-with-Machine-Learning-for-Economists-Course