Aizhan Borubaeva
Prediction with Machine Learning
Assignment 2
Airbnb prediction models

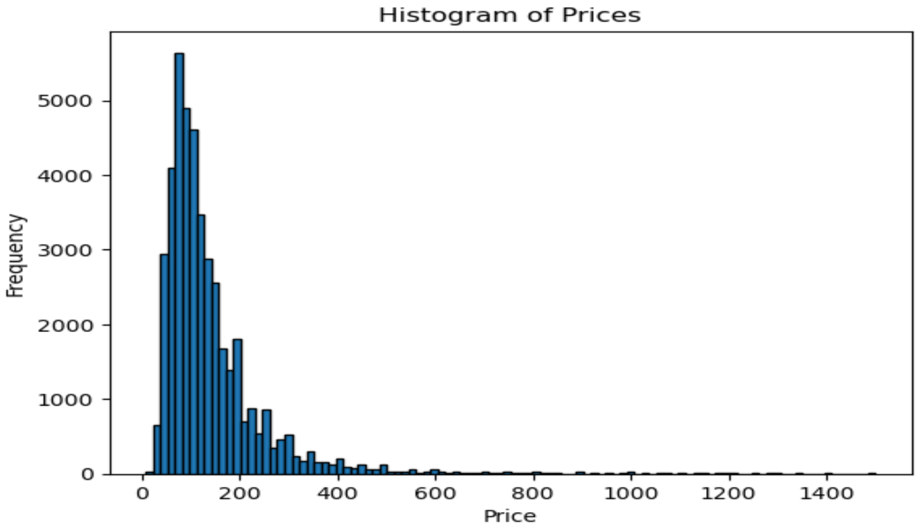
Summary report

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 have chosen Paris for this project as one of the most popular cities for tourism due to its many well-known tourist locations, cultural life, museums, and architecture. The analysis is based on Airbnb data.[1] I chose the data for December 2022 for Task 1 and for September 2023 for Task 2.

First, I started with the data for December 2022. Due to the goal of the project, I have kept only apartments hosting 2-6 guests located in entire rental units, entire condos, entire lofts, and entire serviced apartments. I deleted several variables that could be bad potential predictors or had a lot of missing values. 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. Histogram 1 below shows the distribution of price after this manipulation. The histogram is skewed to the right which shows that there are still some extremely high values for the price. I kept them due to their density. The given analysis needs to keep them because they can show which qualities of an apartment can predict a higher price. Further data were split for training and test datasets, to test the prediction of the models.

Histogram 1. Price



I have chosen three models for the analysis, including OLS, Random Forest, and Gradient Boosting. To create the best OLS model I used several fixed-effect OLS models with different potential predictors (Table 2):

[1] Inside Airbnb. "City Name Dataset." Inside Airbnb, http://insideairbnb.com/Paris/.

Table 2. OLS models

Models	Dependent variable	Predictors
Model_1	Price	host_listings_count, accommodates, beds
Model_2	Price	host_listings_count, accommodates, beds, minimum_nights, maximum_nights, number_of_reviews, number_of_reviews_ltm, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value
Model_3	Price	host_listings_count, accommodates, beds, minimum_nights, maximum_nights, number_of_reviews, number_of_reviews_ltm, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, calculated_host_listings_count, calculated_host_listings_count_entire_homes, calculated_host_listings_count_private_rooms, calculated_host_listings_count_shared_rooms, reviews_per_month, host_is_superhost_t, has_availability_t, instant_bookable_t
Model_4	Price	host_listings_count, accommodates, beds, minimum_nights, maximum_nights, number_of_reviews, number_of_reviews_ltm, review_scores_rating, review_scores_accuracy, eview_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, calculated_host_listings_count, calculated_host_listings_count_entire_homes, calculated_host_listings_count_private_rooms, calculated_host_listings_count_shared_rooms, reviews_per_month, flag_review_scores_rating, host_is_superhost_t, neighbourhood_cleansed_Bourse, neighbourhood_cleansed_Buttes-Chaumont, neighbourhood_cleansed_Buttes-Montmartre, neighbourhood_cleansed_Entrepôt, neighbourhood_cleansed_Gobelins, neighbourhood_cleansed_Hôtel-de-Ville, neighbourhood_cleansed_Louvre, neighbourhood_cleansed_Luxembourg, neighbourhood_cleansed_Ménilmontant, neighbourhood_cleansed_Observatoire, neighbourhood_cleansed_Opéra, neighbourhood_cleansed_Palais-Bourbon, neighbourhood_cleansed_Panthéon, neighbourhood_cleansed_Passy, neighbourhood_cleansed_Popincourt, neighbourhood_cleansed_Reuilly, neighbourhood_cleansed_Temple, neighbourhood_cleansed_Vaugirard, neighbourhood_cleansed_Élysée, property_type_Entire loft, property_type_Entire rental unit, property_type_Entire serviced apartment, has_availability_t, instant_bookable_t

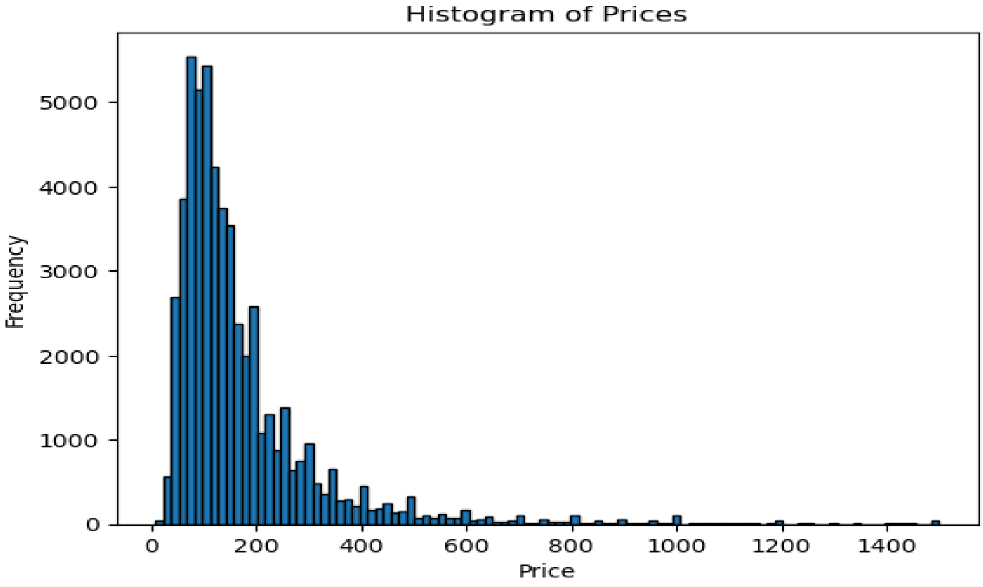
According to the results of these regressions, it can be concluded that the third model is the best because of the lowest RMSE -100.16 and highest R2 - 0.264 (Table 3). However, further, I added the OLS model including all indicators and it demonstrated even better results with RMSE - 96.73 and R2 - 0.34. Therefore, I decided to use the last fourth OLS model for comparison with Random Forest and

Gradient Boosting. I chose Random Forest and Gradient Boosting due to the possibility of manual rather than machine model selection in this case, and because they allow you to process many scenarios simultaneously through automatic algorithms.

Based on Table 4 we can evaluate the predictive quality of these models in the given case. The Gradient Boosting model is the best for prediction in this case because of the lowest RMSE – 84.42 and highest R-squared - 0.49. In both cases, Random Forest is better than OLS due to higher R-squared and lower RMSE. According to BIC, both OLS and Gradient Boosting are the best models because of the equal BIC (62,889.78) which is lower than in the case of Random Forest.

For September data I did the same data cleaning manipulations as in the December case, so as filtering room type, and number of guests, dropping unnecessary variables, creating dummy variables for categorical ones and deleting extreme values for price (above USD 1,500). You can find the descriptive statistics of the price before the filtering in Table 5. Histogram 2 below shows the distribution of price after this manipulation. The histogram is skewed to the right. Further data were split for training and test datasets, to test the prediction of the models.

Histogram 1. Price

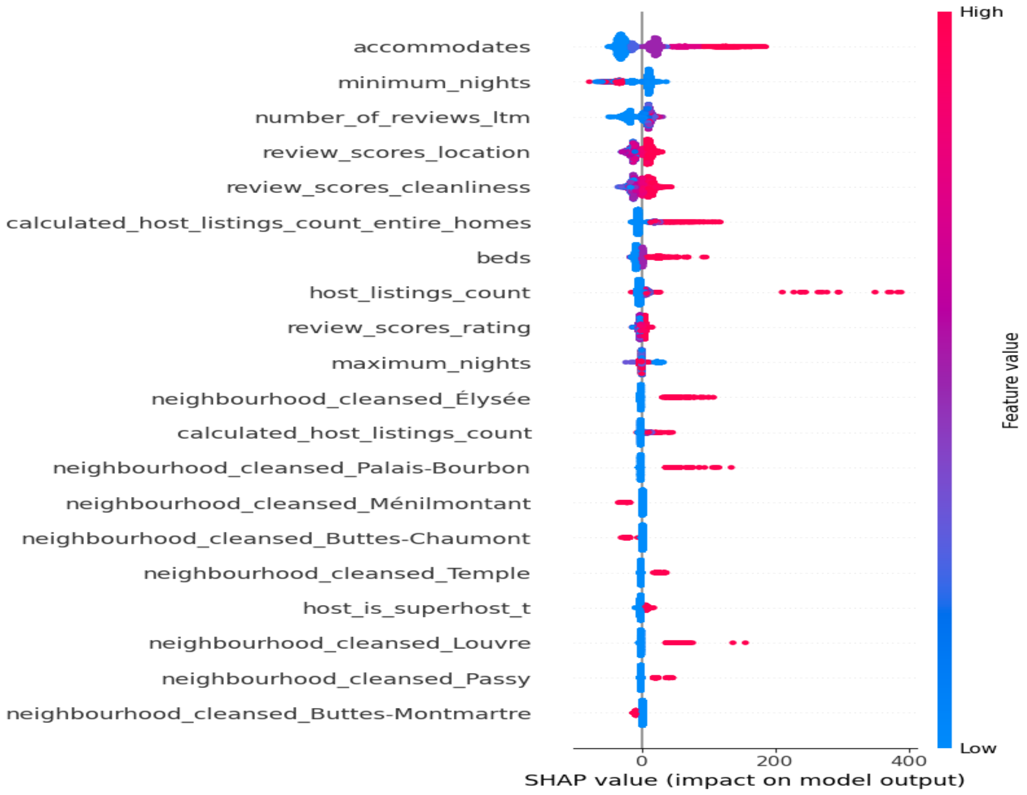


I used the same OLS, Random Forest, and Gradient Boosting as I used in Task 1 for December. The results in Table 7 show that the Gradient Boosting model is still the best for prediction because of the lowest RMSE – 101.21 and highest R-squared - 0.47. In both cases, Random Forest is better than OLS due to higher R-squared and lower RMSE. According to BIC, both OLS and Gradient Boosting are the best models because of the equal BIC (79,583.50) which is lower than in the case of Random Forest.

However, the prediction quality of all models for September is worse than for December, which can be demonstrated in the combined Table 8. For example, the RMSE of the best gradient-boosting model in September is much higher than in December. We can see the same tendency for other indicators. It can happen because of the differences in the market in September and December, for example, people in December can travel more because they have Christmas holidays, or such unlisted indicators as good Christmas entertainment reconsider the model with the time to be in line with all market changes.

Using a SHAP model we can understand which predictors affect our price the most. In Figure 1 we can see that the number of guests, minimum nights to stay and the number of reviews during the last month have the highest effect on the price, while the neighbourhoods are less important.

Figure 1. SHAP



Based on the analysis above we can conclude that the Gradient Boosting is the best model for the given case which can help to predict prices more accurately. However, we should remember that this algorithm takes more time than Random Forest or OLS, so we should evaluate the size of dataset and potential cost of this additional time. The model should be reconsidered with the time to adapt to market trends, inflation, and seasonality. The better model can be created using SHAP method to drop less-important variables and creating better dataset, which will lead to better model and therefore better predictions. We can also consider additional models to find the best.

Appendix

Table 1. Descriptive statistics for price – December 2022

count	Mean	std	min	25%	50%	75%	max
43,771	158	581	9	77	110	167	99,140

Table 3. Fixed-effect, OLS models – December 2022

Models	Model_1	Model_2	Model_3	Model_4
RMSE	105.91	101.13	100.16	96.73
R-squared	0.18	0.25	0.26	0.34

Table 4. Final models – December 2022

Models	OLS	Random Forest	Gradient Boosting
RMSE	96.73	85.55	84.42
R-squared	0.34	0.48	0.49
BIC	62,889.78	63,077.61	62,889.78

Table 5. Descriptive statistics for price – September 2023

count	mean	std	min	25%	50%	75%	max
55,036	190	357	9	87	129	200	30,000

Table 7. Final models – September 2023

Models	OLS	Random Forest	Gradient Boosting
RMSE	109.88	103.40	101.21
R-squared	0.38	0.45	0.47
BIC	79,583.50	79,950.97	79,583.50

Table 8. Combined table – December 2022 vs September 2023

Models	OLS		Random Forest		Gradient Boosting	
Date	December	September	December	September	December	September
RMSE	96.73	109.88	85.55	103.40	84.42	101.21
R-sqr.	0.34	0.38	0.48	0.45	0.49	0.47
BIC	62,889.78	79,583.50	63,077.61	79,950.97	62,889.78	79,583.50

Sources:

Database: Inside Airbnb. "City Name Dataset." Inside Airbnb, http://insideairbnb.com/Paris/.

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