Price Predictions for AirBnB Listings in Buenos Aires:

Using Data as of September 22, 2023 to Build Prediction Models for Price on Listings

Introduction

Data from AirBnB listings in Buenos Aires was taken from insideairbnb.com for listings in September 22, 2023. It contains data for all listings on AirBnB's website as of the specified date including many different features and amenities listed and also lists prices in the local currency, this in case the Argentine Peso. A dictionary of all the features and explanations of each within the original data[1].According to the company's request, the data was filtered for only showing listings that list a total number of guests accommodated between 2-6 because the company is looking to set prices of their new apartments that aren't on the market and want to compare them to small and mid-size listings. The code for the entire process can be viewed on github[2].

Data Decisions

The data pre-filtering for guests accommodated between 2-6 was n=29,346. Post filtering it was n=27,340 observations. The dataset itself was dirty and required significant cleaning with several decisions being made as to how. They are as follows:

- Dropping columns that contained specific information about the host's residency beyond whether they lived in the city or not as these columns were mostly null values
- Filtering the dataset further to account for only entire homes/apartments (classified together within the data) and shared rooms which meant excluding hotels (n=56)
- Imputing the mean values of columns that were null/missing from the data with their respective means per number of guests accommodated (meaning imputing the mean for a listing that accommodates 4 people with the mean review scores for existing reviews amongst other listings that accommodates 4 people). For all values that were imputed flag variables were created. This was performed for the following columns:
    - Review scores
    - Bedrooms
    - Host Acceptance and Response Rates
- Determining which listings within the data qualified as new listings
    - Listings that had a number of reviews all time were classified as a new listing for this review
- Reclassifying the bathrooms_text column in the data as a numeric column adding a dummy variable to continue to capture whether or not the listing has a shared bathroom or not
- Quantifying the following amenities to be considered as "luxury" amenities:
    - Pools, Pool Tables, Hot Tubs, Saunas, Dishwashers, Grills, Rooftops, Premium Cable, Gyms
- The following amenities were noted as important with dummy variables:
    - Patios, Toasters, Microwaves, Bidets, Skyline views, Pets being permissible
- Additional dummy variables were created for comparison for higher review scores, minimum stays for a listing over 3 days, whether the host was local, and accounting for certain more expensive neighborhoods (Palermo, Recoleta, Puerto Madero)
- Additional null values in the assortment of review scores available that couldn't be rectified or determined as to why they may have been null (because they may have been a new listing, etc.) were dropped (n=55)

After data cleaning and feature engineering descriptive statistics were run on price. It showed a mean price of 23,371.014 with an extremely large standard deviation of 320,394.898. The decision to perform IQR on price to exclude extreme values was made by determining that any value greater or less than 1.5 times the value of either the 25% or 75% quartiles can be considered as extreme values to be dropped. Descriptive statistics were run post-IQR and provided a much more reasonable standard deviation of 6,333.464 with a total of n=2155 observations dropped. The comparison of descriptive tables can be seen below:

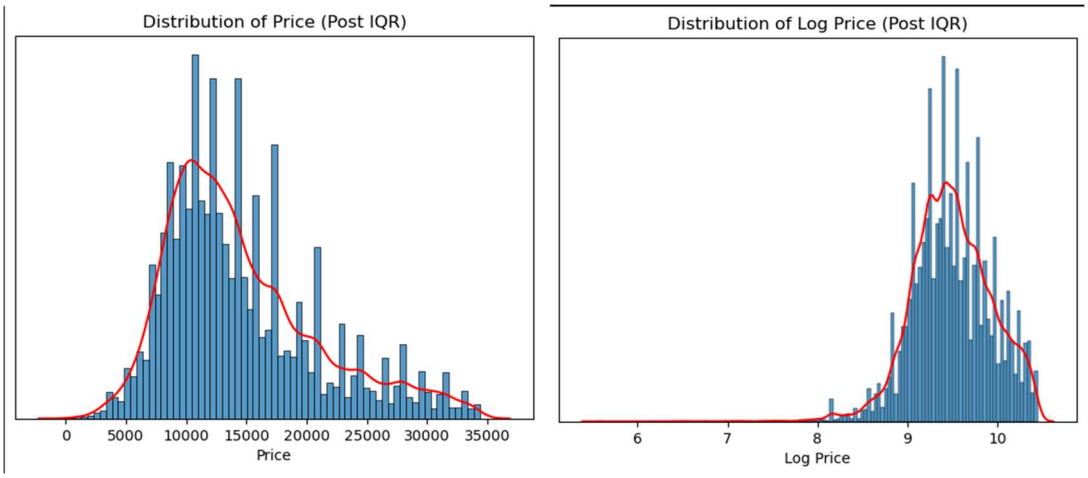
[1] https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=1322284596

[2] https://github.com/nxfern/DA3_Assignment_2/blob/main/Assignment_2_NF.ipynb

```
count      26970.000
mean       23371.014
std       320394.898
min          260.000
25%        10301.000
50%        13753.000
75%        19951.000
max     35002562.000
Name: price, dtype: float64
```

```
count    24815.000
mean     14546.011
std       6333.464
min        260.000
25%      10064.000
50%      13046.000
75%      17501.000
max      34352.000
Name: price, dtype: float64
```

A comparison of the distributions of price post IQR was also made to determine whether to transform price into it's log values for better predictions however it was determined that the log transformation did not make an improvement and that the level distribution was close enough to normal, therefore no transformation was made. Graph comparisons are below:



Another plot was created to check for any potential non-linearity in the data however it did not show any particular non-linearity compared to a basic OLS plot model when examining price according to number of guests a listing accommodates.

Models

Four models were created to compare them to each other and provide the best predictions. A work set and a holdout set were created for training the models at an 80/20% split along with setting up cross validation with 5 folds. All models used 'accommodates' as the main explanatory variable. RMSE values were taken from both performances on the work set and the holdout set for comparison.

Simple Linear OLS

An OLS model was created by hand first. The model accounts for bathrooms, bedrooms, beds, amenities, neighborhood the listing is in, whether the host is a superhost, amongst other things.

LASSO

A LASSO model was created using lambda turning. All of the parameters from the hand-made OLS model prior along with additional explanatory variables and all dummy variables were used with the LASSO algorithm to allow for LASSO to determine which features were important or not.

Random Forest (RF)

A Random Forest model was created using all the values from the LASSO model without the dummy variables, interaction terms, and the flag variable for new listings.

Gradient Boosting Machine (GBM)

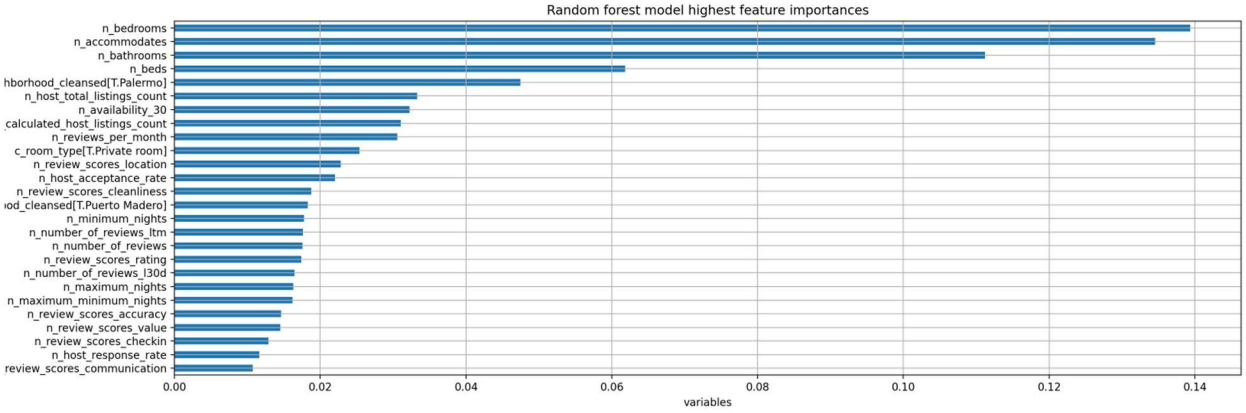
A GBM model was created using the exact same variables from the RF model. The GBM model used was a naïve tuning parameter to cutdown on the required computational power.

Results and Conclusion

The RMSEs from all models are shown in the table below:

	model	Workset RMSE	Holdout RMSE
0	OLS	5230.963	5054.841
1	LASSO	5227.052	4820.723
2	RF	5306.906	4236.691
3	GBM	4743.295	1867.580

From these results the GBM model has by far the lowest RMSE on both sets. The hand-made OLS model performs the worst, which implies that there is a fair amount of variance that is not being accounted for properly in the model. The LASSO model performs better on the work set than the Random Forest model but performs worse on the holdout set while also taking longer to process than any of the other models chosen. Feature importance within the RF model can be seen below:



Interestingly, bedrooms, accommodates, and bathrooms are the most important. Next, the number of beds and the listing location in Palermo. Beyond that most of the remaining features are similar in importance.

The holdout set can be construed as the best attempt to simulate live data, and the model that gives the best prediction RMSE can be the best predictive model. The GBM model performs the best on the work set and by far the best on the holdout set. It appears that gradient boosting for this dataset is far and away the superior model for running predictions with live data. The RMSE values being as high as they are fine given the mean price value being 14,546.01. Additionally, the GBM model takes little time to reproduce without tradeoff.

Using the GBM model you can then view the predicted prices and determine prices for new listings depending on the neighborhood the listing is in and whether or not the listing is small or not. See the following table below:

		RMSE	Avg Pred Price	rmse_norm
c_neighborhood_cleansed	d_small_size			
Belgrano	0	1643.550	17929.570	0.090
	1	1978.320	13355.760	0.150
Palermo	0	2062.820	19685.090	0.100
	1	2099.870	14753.640	0.140
Puerto Madero	0	1063.090	27514.460	0.040
	1	1372.610	23506.110	0.060
Recoleta	0	1848.760	18994.600	0.100
	1	2013.460	13773.660	0.150
Retiro	0	1638.360	17419.170	0.090
	1	1551.580	13447.450	0.120
San Nicolas	0	1654.520	16338.580	0.100
	1	1680.760	11784.750	0.140

In the table a 1 indicates that the apartment is small for each respective neighborhood with the mean RMSE value for each, the average predicted price by the GBM model, and rmse_norm being the normalized RMSE values. Based on viewing the rmse_norm values you can see what a competitive price may be for a new listing with the lower the normalized RMSE score the better the model does in predicting the price.