Analytical Report - DA 03 Homework 2

Submitted by: Gabriella Zsiros, Hanna Asipovich

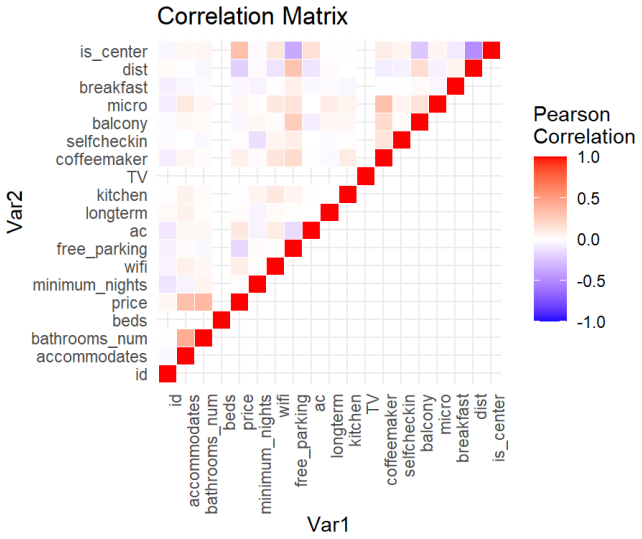
1. Introduction

The purpose of this assignment is to a build price prediction model for a small business running Airbnb apartments in Rome. As the new entrants on the market, they need to find out the best model for pricing their mid-range apartments for 2-6 people. All codes related to this assignment are available on Github:

https://github.com/gabizsiros/DA3_Prediction/tree/main/A2

2. Data Analysis and Transformation

We worked with data collected by Inside Airbnb: Get the Data , which is based on the information scraped from Airbnb webpage in December 2022. The database provided us with historic information on the accommodation rented out in Rome, its main features, as well as price listings for the immediate future. After narrowing down our inquiry to apartments with capacity to accommodate 2-6 people, we had a dataset of about 14,000 observations. With further data analysis, we dropped off the group of "luxury accommodation" which was in the top 5% of price listings. We added additional information on "distance to the center" based on the geo location of the accommodation, so that this factor can also be taken in consideration in future pricing.

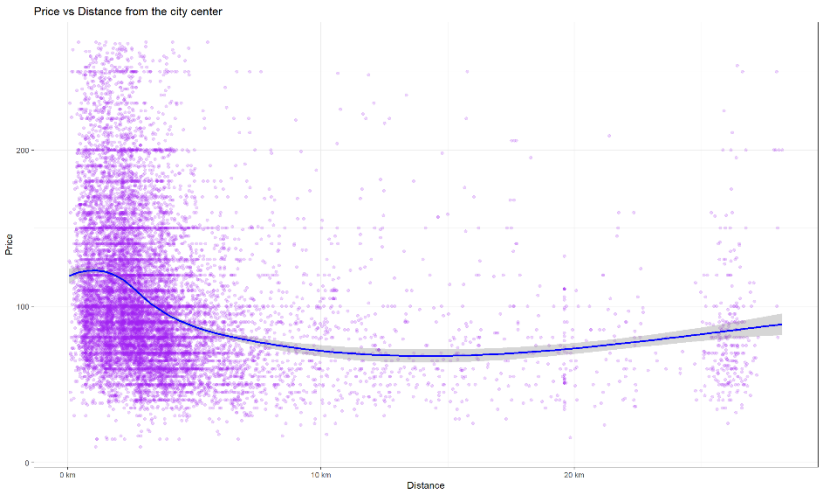


Looking at correlation between various features of the accommodation (see Correlation Matrix), we can see that certain accommodation features relate to final pricing. Hence, we decided to extract following features for building our prediction models: distance to the center, number of guests ("accommodates"), and the neighborhood, where the apartment is located, and, finally, availability of wifi, air conditioning and free parking.

3. Building prediction models

We took these three models: ***OLS*** for simplicity, ***LASSO*** to introduce penalty for overfitting, and ***Random Forest (RF)*** to enhance our model's performance with a machine learning algorithm. For cross-validation, we created a holdout set as a random 15% of our cleaned data table, resulting in approximately 2,000 observations for the holdout.

OLS regression was run according to 5 models of increasing complexity. For example, initially we built a simple regression model on distance to center, which intuitively should represent correlation between price change and the apartment location in relation to the center. Majority of our dataset is close to the center. We can also see it clearly in the density plot, with a bit of an uptick in the accommodation on the outskirts, most likely located close to a transportation hub, possibly the airport. We later added additional features and used OLS method in determining our preferred model out of the five. After building and cross-validating models, Model 5 looked the best positioned based on the lower average RMSE and highest R2. Accordingly, model 5 is taken further to be compared between LASSO and Random Forest to find the best predicting model.



LASSO: As our second model, we selected LASSO as it penalizes over-fitting. The model iterates through various values for lambda, with 0.05 steps between 0 and 1. The best lambda selected by the algorithm was 0.2, and this what we considered when comparing with the rest of our models.

RF: We later use a bagging algorithm Random Forest to further fine-tune our model. RF tends to perform better than OLS with non-linear patterns in the data, as well as showed us better results on RMSE than LASSO. While OLS and LASSO could handle the neighborhood variables a factors, we decided to focus on the first district as a binary decision point when making the regression trees. Here we take our model with the most variables, which does not contain the neighborhoods as factors, and RF's best result has 5 predictors when selecting model 4 formula (with 15 variables).

4. Choosing an optimal model

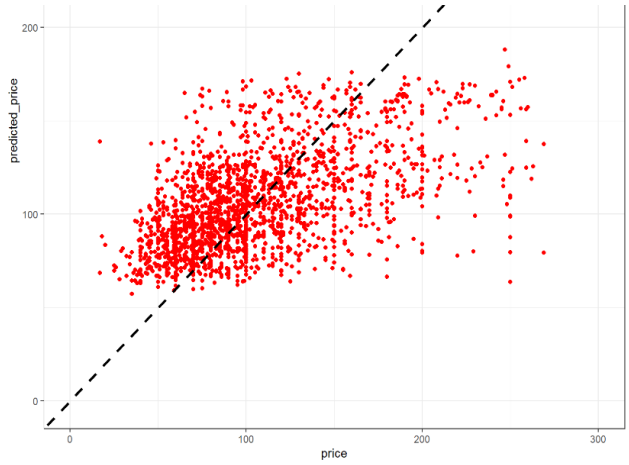
We compare below all three models to decide on the best one for our purposes of price setting for a mid-range apartment in Rome. We see that RMSE for the best OLS and LASSO are very close, similarly to their MAE and R2. However, when we compare either of these models to our best performing RF it outperforms with a lower RMSE=39.24, and the highest R2=0.31

Model Summary

Model	Variables	RMSE	MAE	R2
OLS	16	39.92	30.29	0.28
LASSO	16	39.91	30.29	0.28
Random Forest	5	39.24	29.77	0.31

We test the performance of RF on the holdout dataset. Variable importance shows that distance and location have a high importance, as well as the number of bathrooms and accommodation capacity, which indicate the size of the apartment, followed by the number of minimum nights required variable. Our RMSE on the training sample and the hold-out set are very similar, 39.26 being the RMSE of the Random Forest when tested on the hold-out set. (See the scatterplot graph on predicted price under RF model).

5. Applying predictive model to the immediate future listings: extra task



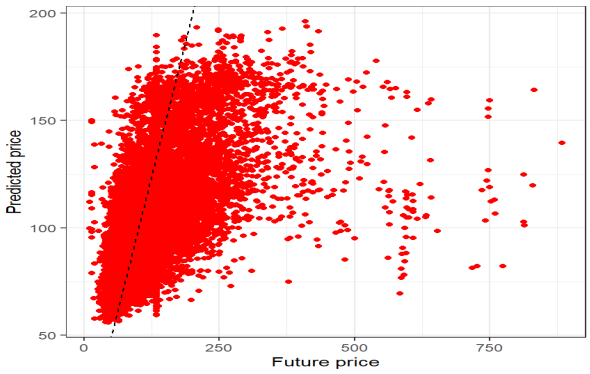
Our analyzed dataset comes from a specific date in 13-12-2022 with data going back to one year. What is also available however, is a booking data of future dates, where we can test if our prediction models can predict the prices for the future. Having the calendar data for the immediate future, we have a unique position to test our models' performance in real life. We applied the predictive model to the future dataset after some minimal data manipulation. When summarizing the data, we can see that the RMSE is a lot higher.

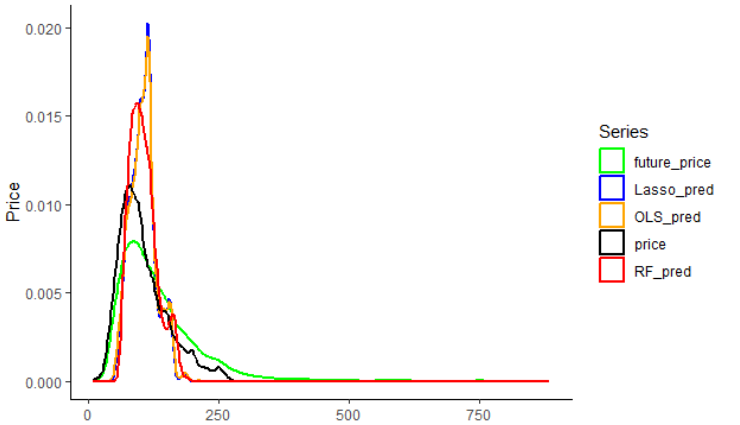
Model comparison with future data

Model	Past.Data.RMSE	Future.Data.RMSE
OLS	39.92	74.81
LASSO	39.91	74.90
Random Forest	39.24	72.72

The order of the model fit did not change; however, random forest performs better than the previous two, with Lasso having slightly better results than OLS.

What skews our data are the listings with high prices, which are not frequent but take on extreme values. We can argue that while for the accuracy of future prediction it could have been better to keep our outliers in the very beginning, but when it comes to the business case, and on how to position the new apartments on the market, being mid-sized apartment without any special feature to our knowledge, it is best to keep our 95% threshold. The skewness also means that the future listings will also have outliers that might not be relevant to our business case.





When comparing the prediction models, we can see that all of our prediction models overestimate the distribution and the frequency around the mean value, but Random forest captures both the current and future distribution visibly better and follows the curve of both the current and future data distribution.