Assignment 2: Business Report

Alima Dzhanybaeva

Introduction

In this project, I tried to find the best price prediction model for mid-size apartments in Los Angeles. The results for 5 models (OLS, LASSO, CART, Random Forest, GBM) showed the lowest RMSE of 73.69 for *Random Forest*, suggesting the best predictive performance.

Cleaning data

The original dataset for Los Angeles includes 40438 observations and 75 variables, and was scraped from Airbnb on December 7th 2022.

Firstly, the variables that won't be helpful for the prediction of prices and the ones that had a lot of missing values were dropped from the dataset.

As the goal of the project is to predict prices for mid-size apartments the dataset was filtered for the property types ("Entire condo","Entire loft", "Entire serviced apartment", "Entire home/apt", "Entire rental unit") and number of accommodates between 2 and 6. Additionally, the 'price' variable was filtered for the values between 0 and 800 USD.

After additional transformation of existing variables, addition of new variables, and imputation of the values for the missing values we ended up 11890 observations and 59 variables.

Models

First of all we need to decide which variables to include to the models. Therefore, I created 8 models with different combinations of variables and their functional forms to later run the regressions, look at the results for different measurements and decide which one to use. In the table below you can get familiar with the variables that were added to each model:

Model	Predictors
M1	# of Accommodates
M2	M1 + Property Type + Number of Beds + Number of Days Since the First Review
M3	M2 + # of bathrooms + Neighbourhood group + Reviews per Month + Review Score Rating + # of Reviews
M4	M3 + # of Accommodates Squared + Squared and Cubic Terms # of Days Since the First Review
M5	M4 + (Property Type * Number of Accommodates) + (Property Type * Neighbourhood Group)
M6	M5 + (Property Type * several Amenities Dummies)
M7	M6 + all Dummy Vars
M8	M7 + (# of Accommodates * Neighbourhood group) + (Property Type * all Dummy Vars)

The results for all eight models are presnted in the table below:

Model	N predictors	R-squared	BIC	Training RMSE	Test RMSE
(1)	1	0.1376757	111464.4	84.66793	84.65799
(2)	6	0.1843692	110980.7	82.33904	82.38185
(3)	12	0.2597859	110112.7	78.43036	78.57805
(4)	15	0.2633928	110093.8	78.23495	78.42164
(5)	24	0.2653805	110150.5	78.11666	78.43121
(6)	56	0.3023260	109952.8	76.08310	76.83552
(7)	67	0.3183255	109832.9	75.19032	76.08590
(8)	102	0.3257105	110049.9	74.70709	76.40225

Root Mean Squared Error (RMSE) and Bayesian Information Criterion (BIC) are a statistical model selection criterion used for comparing different models. The lower the RMSE and BIC values, the better the fit of the model to the data. Even though **Training RMSE** is the lowest for the *Model 8*, the difference between *Model 7* is negligible. Moreover, **BIC** and **Test RMSE** are the lowest for *Model 7*.

As for other models (**CART**, **Random Forest**, and **GBM**), following variables are used: number of accommodates, property type, number of beds, number of days since the first review, factor variables for the number of bathrooms, neighbourhood group, factor variable for number of reviews, rating, number of reviews per month, different polynomials, and all dummy variables.

The results for all 5 models are presented in the table below:

	CV RMSE	Holdout RMSE	CV R-squared
OLS	76.06700	75.94949	0.3046565
LASSO	75.89403	75.91083	0.3076597
CART	79.19440	79.92913	0.2467668
Random forest	73.68993	73.42125	0.3502718
GBM	74.17428	74.29148	0.3397116

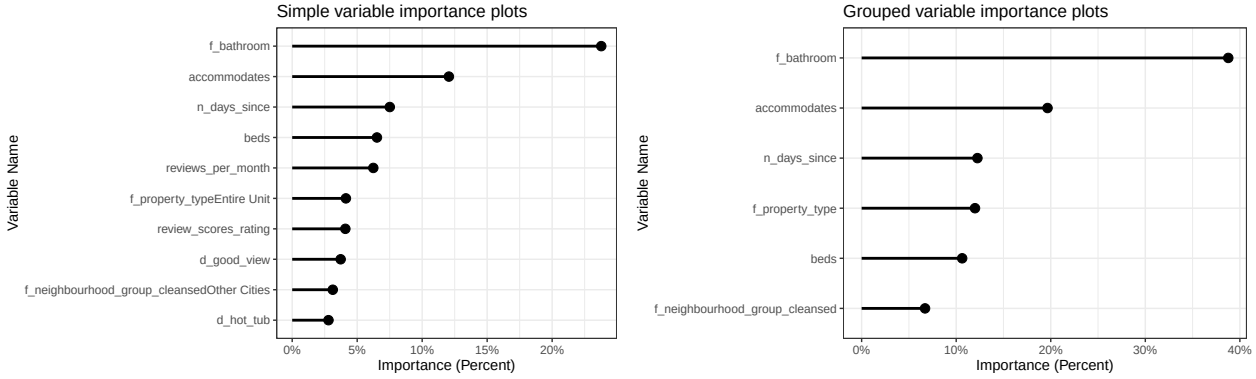
Random Forest showed the best performance with the lowest values for **CV RMSE**, **Holdout RMSE** and the highest value for **CV R-squared**.

Diagnostics

As **Random Forest** is a "black box" method it is quite hard to understand the inner workings of the model and the relationship between the input variables and the predictions, as it is represented by a combination of many decision trees and not by mathematical formula like in linear regressions.

Nevertheless, we can try to uncover the obtained patterns with the variable importance plot, partial dependence plot, and plot for actual vs predicted prices.

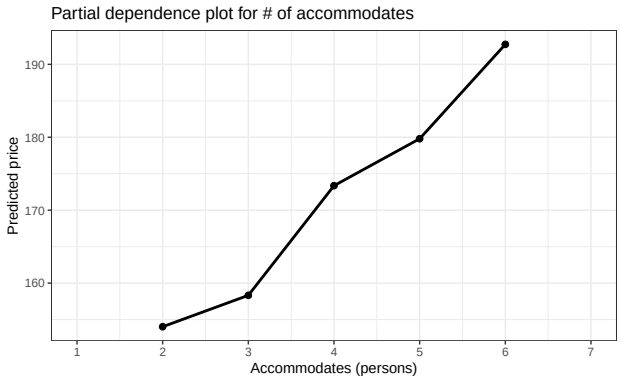
Importance of the variables



It is noticeable that 'f_bathroom' and 'accommodates' are the two most important variables, as they together account for around 30% in the 'Single variable importance plot' and 60% in the 'Grouped variable importance plot'.

Another important variables are: 'n_days_since', 'f_property_type', 'beds', 'f_neighbourhood_group_cleansed', and 'reviews_per_month'.

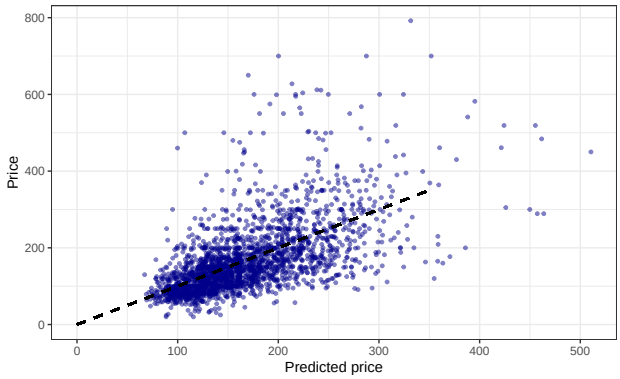
Partial dependence plot



As we can see from the partial dependence plot for the number of accommodates, as the number of guests that can check in into apartment rises the predicted price also increases.

Actual VS Predicted prices

The figure below implies that the model is performing better for cheaper apartments.



Conclusion

This project tried to find the model that will predict prices for mid-size apartments in Los Angeles. The obtained values **CV RMSE**, **Holdout RMSE** for all 5 models (OLS, LASSO, CART, Random Forest, GBM) showed that **Random Forest** has the best performance.

Using variable importance plots it was found out that 'f_bathrooms' and 'accommodates' have the biggest influence on the prices for apartments. The high importance of these two variables is not surprising, as their higher values suggest the bigger apartment size.

Additionally, other diagnostics showed that the model performs better for cheaper and smaller apartments. This is may be due to the fact that the dataset doesn't include variables that actually capture the features that make apartments higher in value.