DA3 - Assignment 2 - Summary review

Oszkar Egervari

2/10/2022

Introduction

In this report I build models, that predict the prices of Airbnb rentals in San Diego, USA. The first two models are linear regressions estimated by OLS and LASSO, I compare them with predictions from a single regression tree with CART, and a Random Forest model. The aim of this report is to find the model with the best performance based on RMSE.

Data

The data used for the report is the San Diego data from http://insideairbnb.com/ (https://raw.githubusercontent.com/oegervari/da2_a2/main/airbnb_sandiego_prepped.csv). The data I've linked is already cleaned and prepared for the report. The number of observations is 10298 for all of our key variables.

This report is aimed to predict rental units that accommodate 2-6 people per night. After the necessary filtering, we are left with 3839 observations. The distribution of price can be found in the appendix.

Before building the models, we split the data into train and holdout sets. The models are built using the train set, while the holdout set is reserved for the final test. The dimensions of the two sets can be found in the appendix.

The last step before the modelbuilding is determining the variables. It's done in four groups, first is the basic variables, next the review related variables, the third is amenities, which are all binary variables, the last is creating interactions for LASSO. Out of these groups are created three predictors. The variables and the predictors can be found in the appendix.

Building Random Forest models

With the 5 fold cross-validation is set for the train set, two random forest models are built. The simpler one uses the first set of predictors (containing only the basic variables), the more complex one uses the second set of predictors (containing the review related and amenity variables on top of the basic ones). The RMSE of the two models can be seen below, the whole results table can be found in the appendix.

```
##    user  system elapsed
##  17.527   0.481   7.242
```

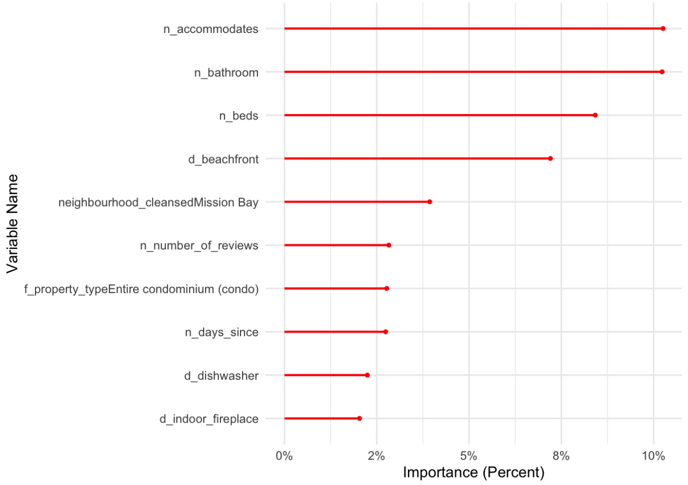
```
##    user  system elapsed
##  19.297   0.486   7.920
```

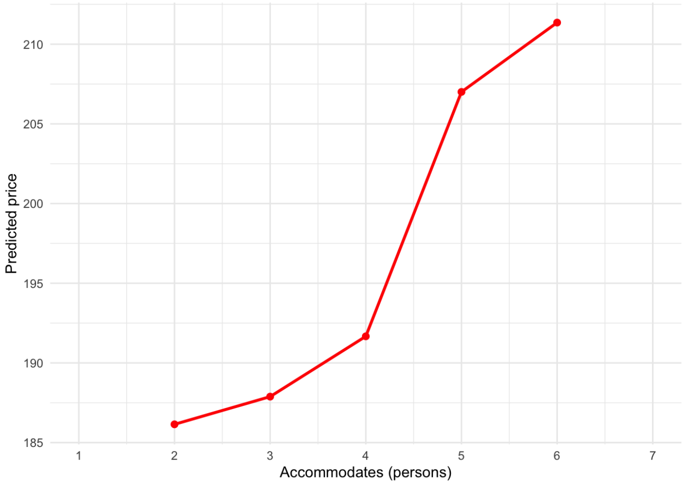
```
##
## Call:
## summary.resamples(object = results)
##
## Models: model_1, model_2
## Number of resamples: 5
##
## MAE
##               Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## model_1 60.03020 65.13785 66.39429 65.90689 68.91608 69.05601    0
## model_2 58.35068 62.35886 64.82257 63.59424 65.45043 66.98865    0
##
## RMSE
##               Min.   1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## model_1 87.71929 100.96670 102.2987 114.6135 140.5729 141.5097    0
## model_2 85.70122  97.41799 101.4013 112.2192 137.4660 139.1094    0
##
## Rsquared
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## model_1 0.1825382 0.2307887 0.3172644 0.3204699 0.4093993 0.4623588    0
## model_2 0.2581820 0.2630383 0.3993865 0.3762865 0.4692204 0.4916055    0
```

Based on the RMSE, the more complex model provides better prediction.

Model Diagnostics

The variable importance plot shows which variables reduce the sample MSE the most. Our plot displays the top 10 variables. Our other plot, the partial dependence plot shows the direction and shape of association of the number of accommodates with price.





As the last step of the model diagnostics, we check the fit of the models on different subsets of the holdout data. The subsets inform us on the external validity of our prediction across groups represented by the subsamples. We compare the RMSE values relative to the corresponding mean predicted y. We use three subsamples, first we devide the data to small (apartments accommodating 3 or less people) and large apartments. With the second subsample we compare the performance of our model in six neighbourhoods. Lastly we do the same with two types of properties, entire rental units and entire condominiums.

The RMSE/predicted price show similar predicting performance in case of apartment size and type, with slightly better performance in case of smaller apartments and rental unit compared to condos. However in case of the neighbourhoods we see more significant differences, which means that the prices are harder to predict there. Most likely because of the uneven number of observations.

Comparing models

Lastly let’s compare the random forest models with the linear regression models and the regression tree (CART). The evaluation on the holdout set can be seen below (Results of other models and CART tree can be found in the appendix).

```
##    user  system elapsed 
##   3.471   0.206   4.065
```

```
##    user  system elapsed 
##  49.500  19.091  83.664
```

```
##    user  system elapsed 
##   5.342   0.353   6.054
```

	Holdout RMSE
	<dbl>
OLS	105.02613
LASSO (model w/ interactions)	104.42299
CART	102.93622
Random forest 1: smaller model	100.58695
Random forest 2: extended model	98.45799

5 rows

Conclusion

Evaluating the models on the holdout set, we see that the best performer based on the RMSE is the more complex random forest model, followed by the simpler random forest model. Meanwhile the worst performer is the OLS model, which could suggest, that nonlinear functional forms and interactions could be important for the prediction. On the other hand the LASSO with interactions performed very similarly to the OLS, but it might be because the interactions weren’t the right ones.

Appendix

Datasummary for Price

	Mean	Median	Min	Max	P25	P75	N
price	193.80	155.00	45.00	2300.00	119.00	227.00	3839

Summary table of the RF models

```
##            Length Class      Mode     
## values     15     data.frame list     
## call        2     -none-     call     
## statistics  3     -none-     list     
## models      5     -none-     character
## metrics     3     -none-     character
## methods     5     -none-     character
```

Dimensions of the train and holdout sets

```
## [1] 2688   93
```

```
## [1] 1151   93
```

Variables, interactions and predictors

```
## [1] "n_accommodates"         "n_beds"                 "n_days_since"          
## [4] "f_property_type"        "f_number_of_reviews "   "n_bathroom"            
## [7] "neighbourhood_cleansed"
```

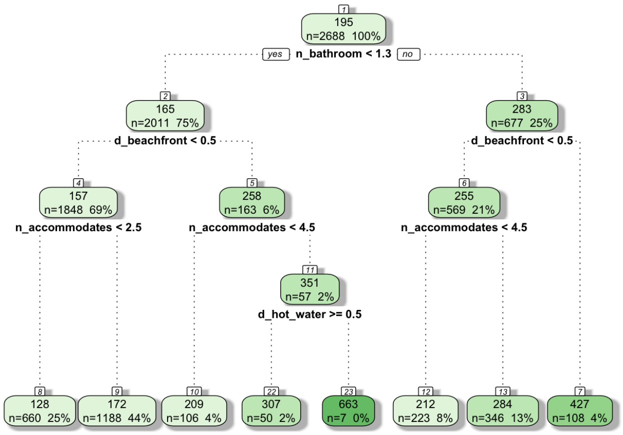
```
## [1] "n_number_of_reviews"         "flag_n_number_of_reviews"   
## [3] "n_review_scores_rating"      "flag_review_scores_rating"  
## [5] "n_review_scores_cleanliness"
```

```
##  [1] "d_host_is_superhost"                 "d_instant_bookable"                 
##  [3] "d_air_conditioning"                  "d_backyard"                         
##  [5] "d_barbecue_utensils"                 "d_bathtub"                          
##  [7] "d_bbq_grill"                         "d_beach_essentials"                 
##  [9] "d_beachfront"                        "d_bed_linens"                       
## [11] "d_board_games"                       "d_breakfast"                        
## [13] "d_ceiling_fan"                       "d_central_heating"                  
## [15] "d_childrenu2019s_books_and_toys"     "d_cleaning_before_checkout"         
## [17] "d_coffee_maker"                      "d_cooking_basics"                   
## [19] "d_dedicated_workspace"               "d_dining_table"                     
## [21] "d_dishwasher"                        "d_elevator"                         
## [23] "d_ethernet_connection"               "d_extra_pillows_and_blankets"       
## [25] "d_fire_extinguisher"                 "d_first_aid_kit"                    
## [27] "d_free_parking_on_premises"          "d_free_street_parking"              
## [29] "d_freezer"                           "d_gym"                              
## [31] "d_heating"                           "d_host_greets_you"                  
## [33] "d_hot_tub"                           "d_hot_water"                        
## [35] "d_indoor_fireplace"                  "d_iron"                             
## [37] "d_laundromat_nearby"                 "d_luggage_dropoff_allowed"          
## [39] "d_microwave"                         "d_mini_fridge"                      
## [41] "d_oven"                              "d_pool"                             
## [43] "d_private_fenced_garden_or_backyard" "d_refrigerator"                     
## [45] "d_roomdarkening_shades"              "d_security_cameras_on_property"     
## [47] "d_shampoo"                           "d_smart_lock"                       
## [49] "d_stove"                             "d_tv"                               
## [51] "d_washer"                            "d_carbon_monoxide_alarm"            
## [53] "d_cable_tv"                          "d_hair_dryer"                       
## [55] "d_long_term_stays_allowed"           "d_dryer"                            
## [57] "d_patio_or_balcony"
## [1] "n_accommodates*n_review_scores_rating"
## [2] "d_pool*n_accommodates"
## [3] "f_property_type*d_beachfront"
## [4] "d_air_conditioning*n_review_scores_rating"
## [5] "d_beachfront*n_review_scores_rating"
## [6] "d_host_is_superhost*n_review_scores_rating"
```

```
## [1] "f_property_type*neighbourhood_cleansed"
## [2] "d_pool*neighbourhood_cleansed"
## [3] "n_accommodates*neighbourhood_cleansed"
```

Comparing models

```
## CART
##
## 2688 samples
##   69 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 2151, 2150, 2149, 2151, 2151
## Resampling results across tuning parameters:
##
##   cp           RMSE      Rsquared    MAE
##   0.006782342  118.2423  0.24498299  67.48069
##   0.007013581  118.1966  0.24590957  67.62682
##   0.007246904  118.0308  0.24804342  67.49184
##   0.010804084  116.9911  0.25313269  67.21560
##   0.014369247  117.2980  0.24645213  67.86415
##   0.015508626  117.8386  0.23851888  68.41347
##   0.016621120  117.8386  0.23851888  68.41347
##   0.031543004  119.9409  0.21096219  71.01047
##   0.054151816  122.7673  0.17071037  73.32764
##   0.143237608  129.9601  0.09862751  80.31255
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.01080408.
```



```
##
## Call:
## summary.resamples(object = .)
##
## Models: OLS, LASSO (model w/ interactions), CART, Random forest 1: smaller model,
Random forest 2: extended model
## Number of resamples: 5
##
## MAE
##                                     Min.  1st Qu.   Median     Mean  3rd Qu.
## OLS                             58.54697 62.75276 64.49504 63.99642 65.66015
## LASSO (model w/ interactions)   55.84669 58.87975 61.39092 61.28218 63.09898
## CART                            61.49785 65.65516 68.28424 67.21560 69.31555
## Random forest 1: smaller model  60.03020 65.13785 66.39429 65.90689 68.91608
## Random forest 2: extended model 58.35068 62.35886 64.82257 63.59424 65.45043
##                                     Max. NA's
## OLS                             68.52716    0
## LASSO (model w/ interactions)   67.19457    0
## CART                            71.32519    0
## Random forest 1: smaller model  69.05601    0
## Random forest 2: extended model 66.98865    0
##
## RMSE
##                                     Min.   1st Qu.    Median     Mean  3rd Qu.
## OLS                             85.03466  92.61374  99.27335 109.4833 134.5037
## LASSO (model w/ interactions)   83.93068  89.89516  98.73961 108.4257 133.6870
## CART                            93.41792 101.89889 104.44393 116.9911 142.2543
## Random forest 1: smaller model  87.71929 100.96670 102.29871 114.6135 140.5729
## Random forest 2: extended model 85.70122  97.41799 101.40133 112.2192 137.4660
##                                     Max. NA's
## OLS                             135.9911    0
## LASSO (model w/ interactions)   135.8759    0
## CART                            142.9407    0
## Random forest 1: smaller model  141.5097    0
## Random forest 2: extended model 139.1094    0
##
## Rsquared
##                                      Min.   1st Qu.    Median      Mean
## OLS                             0.2476822 0.2555340 0.3347972 0.3469574
## LASSO (model w/ interactions)   0.2489213 0.2652267 0.3408111 0.3589322
## CART                            0.1673777 0.1688349 0.2518634 0.2531327
## Random forest 1: smaller model  0.1825382 0.2307887 0.3172644 0.3204699
## Random forest 2: extended model 0.2581820 0.2630383 0.3993865 0.3762865
##                                   3rd Qu.      Max. NA's
## OLS                             0.4193787 0.4773947    0
## LASSO (model w/ interactions)   0.4320167 0.5076854    0
## CART                            0.3113693 0.3662182    0
## Random forest 1: smaller model  0.4093993 0.4623588    0
## Random forest 2: extended model 0.4692204 0.4916055    0
```

	CV RMSE <dbl>
OLS	109.4833
LASSO (model w/ interactions)	108.4257

	CV RMSE <dbl>
CART	116.9911
Random forest 1: smaller model	114.6135
Random forest 2: extended model	112.2192

5 rows