Data Analysis 3 : Assignment II : Technichal Report

Ghazal Ayobi

Introduction

The aim of this report is to provide detailed description for the price prediction model. The main goal of this project is to help the a company to set price for their new apartments which are not yet in the market.To build a price prediction model for a company which is operating small and mid-size apartments hosting from two to six guests in the New York, the data is taken from Inside Airbnb which can be found here. As a result of data cleaning, munging, and analysis, five price prediction models, OLS, Lasso, Cart, Random Forest and GBM, Gradient Boosting Machine, are created. As a result GBM showed the best prediction result with 65.38 USD RMSE. The major predictor features are the number of accommodates, number of bathrooms, the number of beds, neighborhood, and amenities such as availability of washer, gym, elevator are the most important. Other characteristics such as days since the first review is also an important predictor. Eventually the final goal of the project is to finalize a better prediction model measured in relative RMSE values.

Cleaning and Preprationg Data

The data for the project is taken from the Inside Airbnb website. The original data set has a single data table which contains 38186 observations and 74 columns. Some of the important columns are as following: ID, price, number of accommodates, property type, room type, amenities and other host and rental unit characteristics. The data refer to the one night rental prices between January 6 to January 9, 2022. The target variable is price per night per person in US dollars. The original data set contains significant amount of information which requires data cleaning because it is easier to work with Tidy data tables. Initially, I dropped columns such as, host URL, host picture URL, house rules, notes, host location and others.Because these columns are not target of this project. Moreover, To transform the original data to tidy data table, the major part of transformation consists of processing *amenities* column to binary variables. After the preliminary transformation of amenities to binaries, the total number of columns resulted to 3177, from which 3134 are different types of amenities. The code for this transformation is as following. further codes can be found in the 1_cleaning_preparing section.

```
#define levels and dummies
levs <- levels(factor(unlist(df$amenities)))
df<-cbind(df,as.data.frame(do.call(rbind, lapply(lapply(df$amenities, factor, levs), table))))
```

As a result of the above code, meaningful grouping is required to combine similar amenities to one binary variable group. For example, the data contained information about different types of TVs. Such as HDTV, 18 inches HDTV, 32 inches HDTV and many others. To narrow variable grouping, only binaries between 1% to 99% values are kept in the data set. After the grouping of similar amenities the total number of variables decreased to 133 containing 90 amenities. The codes is as following

```
# function to merge columns with the same key word in them
for (i in column_names) {
  xdf <- amts %>% select(matches(i))

  amts$new_col <- ifelse(rowSums(xdf)>0, 1, 0)

  names(amts)[names(amts) == "new_col"] <- paste0("have_", i)

  amts <- amts %>% select(-colnames(xdf))
```

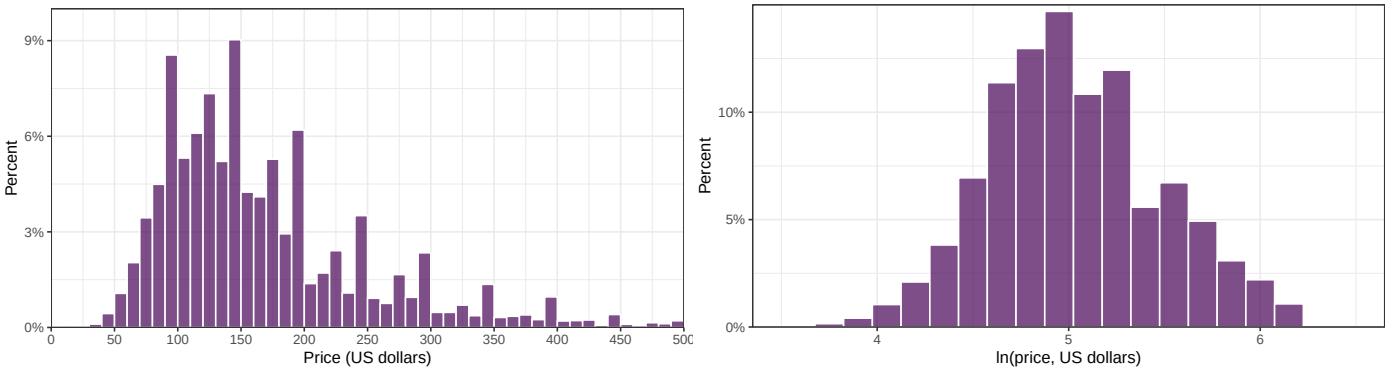
```
}

# keep only columns where the percentage of 1s is at least 1% and at most 99%
selected <- sapply(names(amts), function(x){
  ratio <- sum(amts[[x]])/nrow(amts)*100
  if (between(ratio, 1, 99)) {
    return(TRUE)
  } else { return(FALSE) }
})

# taking only selected or grouped columns
amenities <- amts[,selected]
```

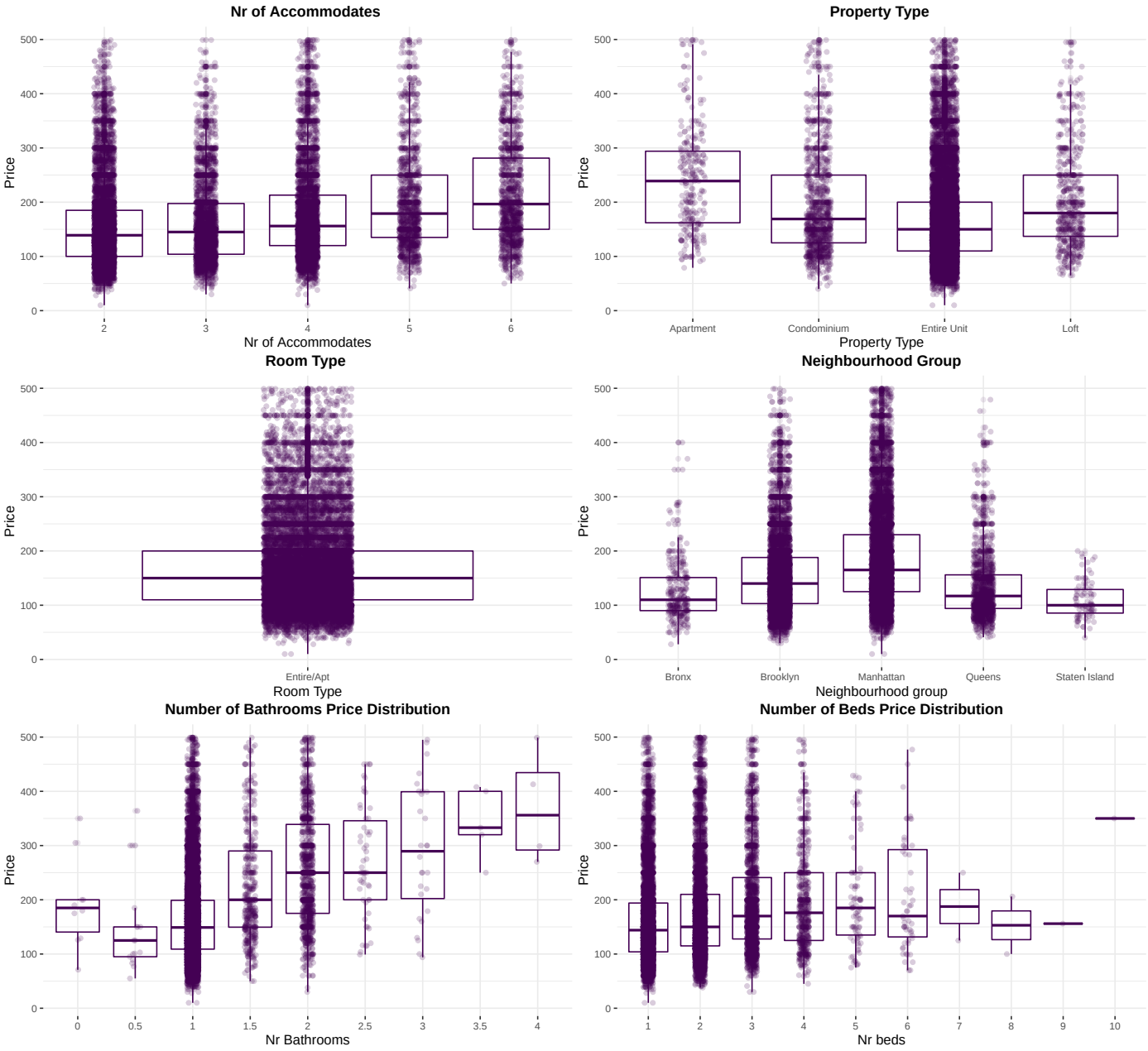
For detailed description of data cleaning please find the codes here. It is worth mentioning that having right structured data, proper ID variable are the essential while working with this big data sets. Moreover, after the preliminary stage of cleaning, data preparation comes next. Preparing data for analysis consists of the following steps: identifying types of the variables, removing duplicates, and addressing missing values. The selected data set, Airbnb New York, consists of numerical values such as price, factor variables such as property type and neighborhood, and binary variables such as all amenities and host characteristics. Other data types are first review and the date scraped. I identify the data types such as price as numeric value and factor variables are created for the predictors such as, property types and neighborhood types. For this process, I used data summaries which can be found in the appendix section.

Filters The main goal of the project is to predict prices of apartments which accommodates between 2 to 6 guests, thus the data was filtered accordingly. The key variable, price, contained extreme value and missing observations. To narrow down the project goal the target variable, price per night included extreme values above 900 USD per night which contained of less than 1% of the observations, thus, price was filtered to less than 500 USD and dropping the observation where price is missing. Moreover, As the goal of project is to build price prediction model, it is crucial to check price and log of price distribution. I created of price and log of price distributions are as following:



Price distribution shows Airbnb apartment prices is skewed with a long right tail and the log price is close to normally distributed. In this project, log of price is not considered, prediction is carried out with price for all of the models.

In addition as this project's main goal is to build price prediction for small and mid-size apartments, thus, the following categories for the property types is filtered for the analysis which are as following categories: entire home or apartment, entire serviced apartment, entire condominium, entire loft, entire rental unit. Based on the dictionary definition, a loft, condominium and a rental unit refer to different types of apartments. The below figures show the number of guests, property type, room type, neighborhood groups, number of bathrooms and number of beds along with their mean prices. It is shown that there is one room type across all data set.



Factoring I created factor variables such as, property type, neighborhood groups based on the above plots.

New Variables the second step in the process is creating new meaningful variables such as number of days since the first review, which is subtracting date of first review from the date when the data was scraped and log, square and cubic functional forms of number of days since first review were created.

Grouping Numeric Variables some of the numeric variables such as number of bathrooms. This variable contained information such as half bathroom. I created another variable as a factor of bathrooms with four cuts of 0, 1, 2, and 10. This means to add all the variables in the mentioned groups. Another example is number of reviews for 0, 1-51 and 51 above.

After the key filters and grouping variables there were 15 variables with missing values. Missing values were addressed as following: first assumption is there is at least one bathroom in each apartment, second assumption is if the number of guests are less than 4 then impute 1, otherwise imputing 2. Missing number of beds were replace with half of number of accommodates, assuming there are double beds in the apartments. It is assumed that the minimum number of nights is one and minimum number of reviews is also 1. Flags were created to indicate the missing values in the each predictor.

As a result of data cleaning, and preparation the total number of observations are 16271 with 136 columns.The cleaned data is saved to this repository.

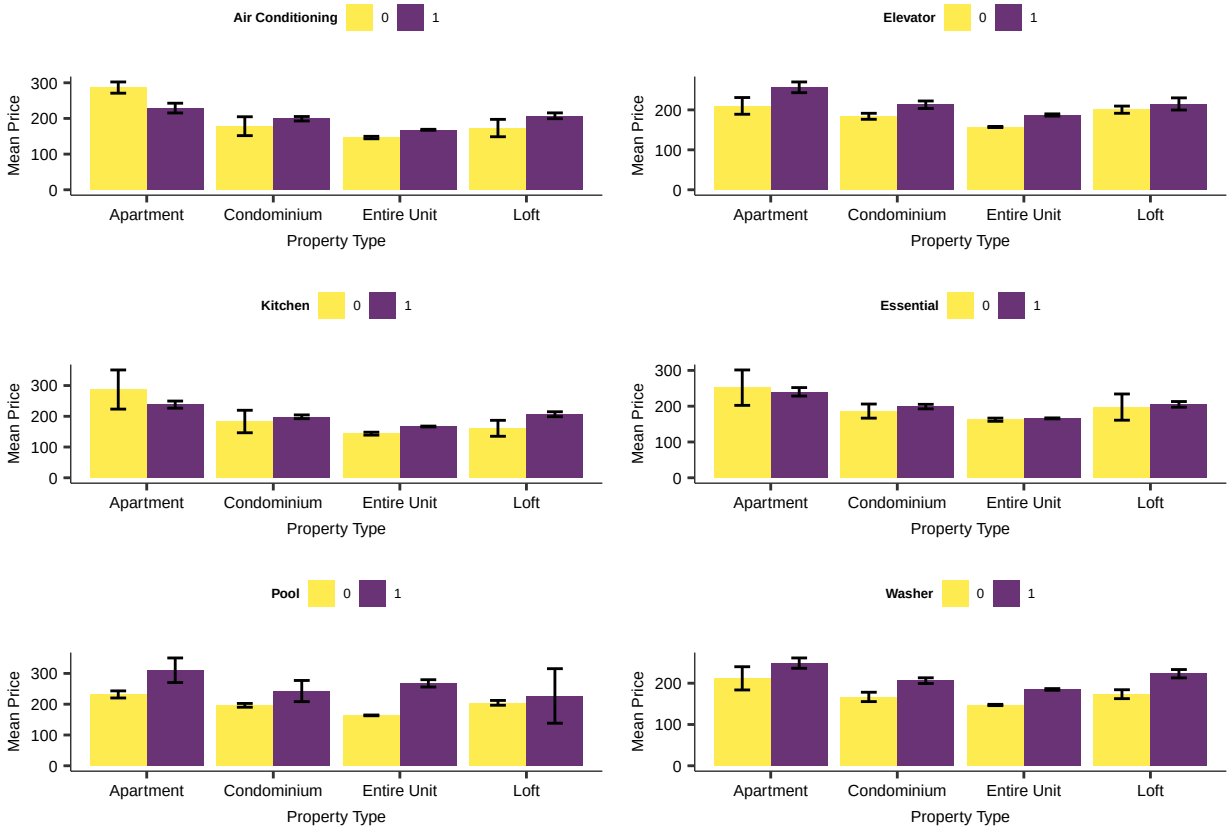
Data Analysis and Feature Engineering

The next most crucial step in this project is data analysis which is conducted in the following steps. First and far most importantly I defined and grouped variables as following.

Feature engineering includes what type of predictor variables to include, and deciding about functional forms of predictors and possible interactions. The data is grouped as following:

- **Basic variables** which consists of the main predictors such as: number of accommodates, property types, number of beds, number of days since the first review and flag variable of number of days since the first review to indicate missing. As the focus of project is on small and mid-size apartment thus, property types are as following: Entire home or apartment, serviced apartment, Condominium, and entire rental unit.
- **Basic addition** this includes key factorized variables, such as, neighborhoods groups, and host response time.
- **Review Variables** consists of the crucial guests reviews predictors such as total number of reviews, number of reviews per month and host review score rating and reviews flags which shows missing variables.
- **Polynomial level** consists of squared terms of guests and squared and cubic terms for days since the first review.
- **Amenities dummies** which consisted of the binary values for all of amenities.

The next step in the process is finding the right interactions. The below plots were used to see prices changes across each interaction.



Based of the above plot I created three categories of interactions which are as following:

- **X1** property type times number of accommodates and property type times host response time (categorical variables)
- **X2** property type time air conditioning, elevator, dryer washer, wife, kitchen and breakfast dummy variables
- **X3** property type time neighbouhood groups, and all amenities

Table 1: New York Airbnb apartment price prediction Models

Model	Predictors
M1	Nr of accommodates
M2	M1 + number of beds + number of days since first review + property type
M3	M2 + Nr bathrooms + Neighbourhood group + host reponse rate + reviews per month + average review score rating + number of reviews
M4	M3 + squared termof guests + squared and cubic terms of number of days since first review
M5	M4 + property type and number of guests interaction + property type and host response time interaction
M6	M5 + property type interaction with adummies as air conditioning, elavator, dryer, washer, wifi, kitchen and breakfast
M7	M6 + all other amenities
M8	M7 + all other amenities, Neighbourhoods interacted with property type

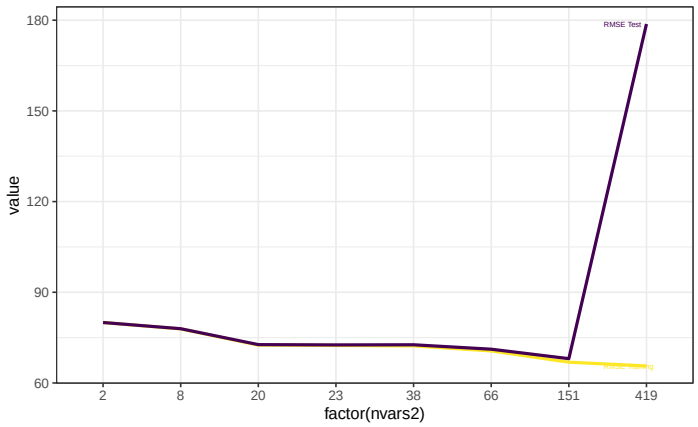
Table 2: Models Evaluation

Model	N predictors	R-squared	BIC	Training RMSE	Test RMSE
(1)	1	0.0532355	151049.2	79.99399	79.99901
(2)	7	0.1023267	150413.0	77.88634	77.95243
(3)	19	0.2214444	148673.5	72.52479	72.68633
(4)	22	0.2238793	148661.2	72.40917	72.59340
(5)	37	0.2265689	148758.1	72.26510	72.63687
(6)	65	0.2612903	148425.5	70.60399	71.17225
(7)	150	0.3358859	147845.1	66.88213	68.04437
(8)	418	0.3564556	149974.6	65.61656	178.74911

Modeling

Regressions The best model gives the best prediction in the live data. Before turning to the modeling part of the project, it is worth mentioning that in order to avoid over fitting, the original data is split into two random parts by 20% to 80% ratio. Holdout set contains the 20% and the rest is work data set. In addition the k-fold cross-validation is a good way to find a model which gives the best prediction for the original data. For the purpose of this project 5-fold cross validation is used. This means splitting the data into five random samples and calculating and deciding based on the average of 5 CV RMSE result. Eight basic OLS regression models from simplest to the most complex one were used to find a better model to use for further analysis. Below are the list of models,

5-fold cross-validation RMSE suggests that Model 7 regression has a better performance and it has the lowest RMSE value. The table is as following.



Models

It is important to run and evaluate different models for a given data set. In order to predict apartment prices, the following models and algorithms were used. Naming them as following:

- **OLS** and **LASSO** using model 7 based CV-RMSE result
- **CART**, **Random Forest** , **GBM** using basic level variables, basic additions, review variables and amenities as dummy variables.

Setting LASSO tuning parameter Before runing the models is sat the lasso tuning parameter, it servers as a weight for the penalty term versus OLS fit. As a result it derives the strength of the variables selection. Lamda, value for tuning parameter is set between 0.05 and 1.

OLS Model:

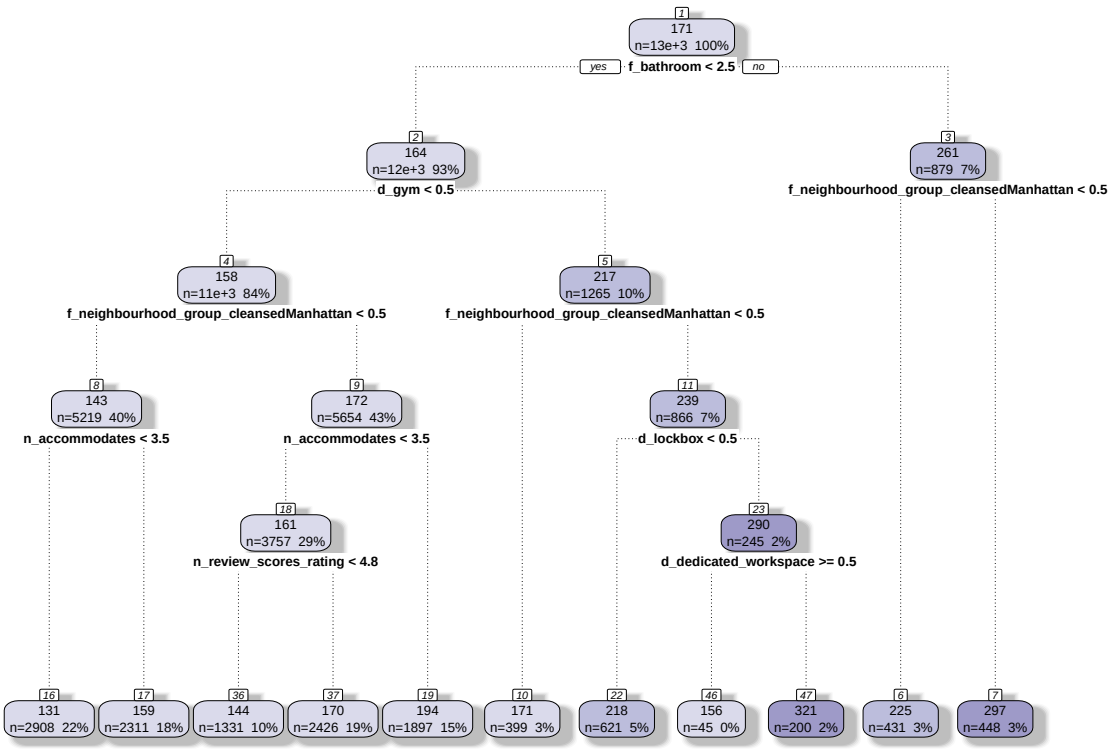
OLS model is the fastest among all models which has CR RMSE of 67.99 and R-squared of 31.67% and with 107 predictors.For this purpose I model 7 which has basic variables, basic additions, reviews, polynomials and a dew interactions

LASSO:

LASSO is the most widely used shrinkage method. It is an algorithm that fits a model by shrinking coefficients, some them to zero by adding a penalty term. After running 5-fold cross validation for the selecting the optimal value for lambda. In the end LASSO picked 149 predictors out 150 with the CR-RMSE of 67.97247. For this method, I used the same Model 7 with few interactions but it does not have as manay interactions as Model 8.

CART

Regression trees are called CART. It is a building and growing a tree. This algorithm has no formula, the goal is to arrive at a set of bins of predictors. This algorithm splits the bins into smaller bins. For CART algorithm no functional or interactions were given. The variables were basic variables, basic additions, reviews and dummies with cp values of 0.005054542. CART selected 104 predictors. CART result is as following



	CV RMSE	Holdout RMSE	CV Rsquared
OLS	67.99696	68.21073	0.3167363
LASSO	67.97247	68.12913	0.3171112
CART	71.88612	71.42794	0.2360636
Random forest	67.06593	66.07839	0.3644942
GBM	65.38265	64.39938	0.3681212

Random Forest

Random forest is an example of ensemble method which uses the results of many predictive models and combine those results to generate a final prediction. I used basic variables, basic additions, reviews and dummies for this model the minimum nod size of 50. The CR RMSE is 67.06593.

GBM

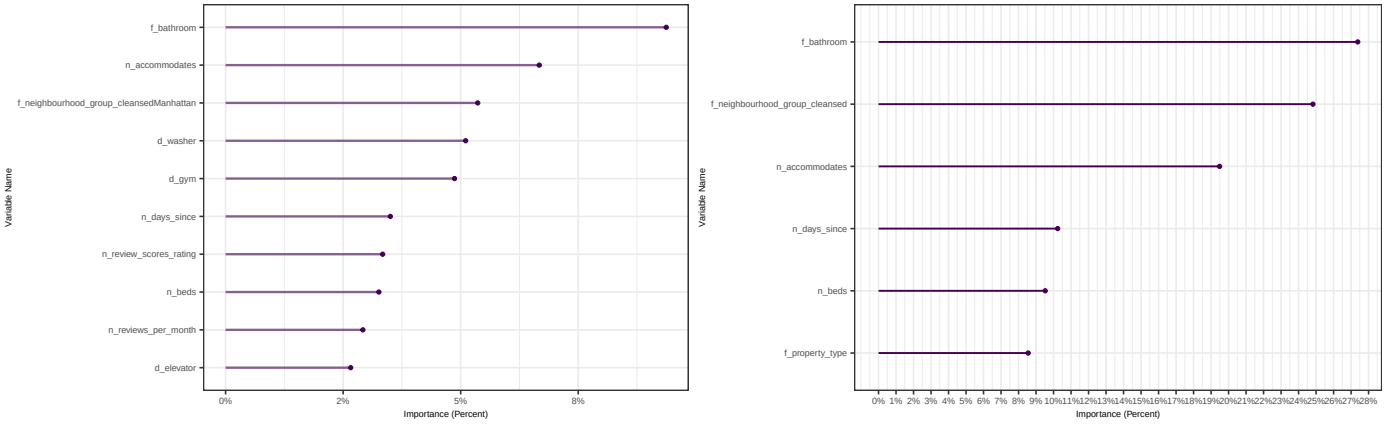
The last model for this case study is GBM. It gives the best result across all models in terms of relative RMSE. GBM basic tuning model is referred as black box because there is no more regression trees. The complexity of tree in this GBM model is 5 and 10 which RMSE of 65.75680, and 65.38265 respectively. The Rsquared is 36.8%.

Result Based on the below table of models, it is can be seen that GBM model has the best performance. 5-fold cross validation RMSE for the data is 65.38 USD RMSE which is 1.683 USD RMSE less than the second best model which is Random Forest. Moreover, The 5-fold cross validation RMSE for the GBM model using Holdout set is 64.38 USD RMSE indicating a better performance than other models illustrating a better performance from second best model of random forest by 1.671 USD RMSE. GBM model tend to be robust, thus, the selected model for this project is GBM BASIC TUNING model.

Diagnostics

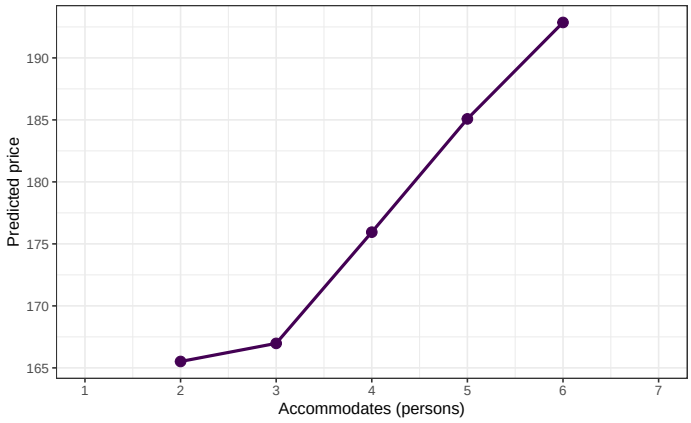
The best selected model, GBM, is an ensemble method which is a black box model, because it does not reveal the pattern of association that drive prediction. However, diagnostic tools can be used to uncover information about the patterns of association which drive prediction. Some of them are as following:

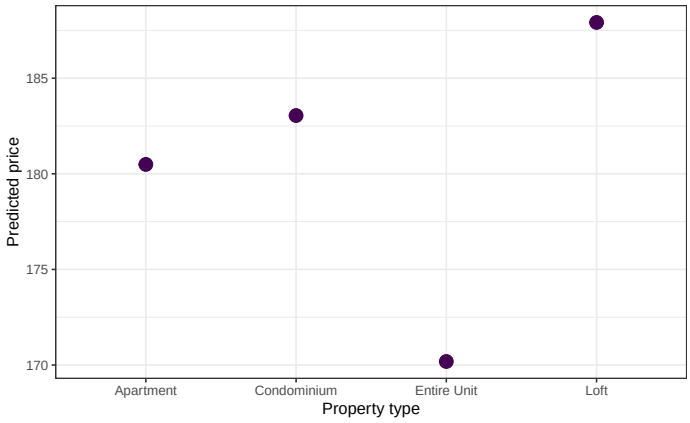
Variable Importance plot it shows the average importance of fit when we use an x variable or group of x variables. Variable importance plot for Top 10 important variables shows that number of bathrooms, number of accommodates, Manhattan neighborhood are the most important along with amenities such as washer, and gym. The grouped variable importance shows that bathrooms, neighborhoods, number of accommodates are the most important variables.



Partial Dependence Plot it shows how average y differs for different values of x conditional on all other predictor variables. Partial dependence plot is based on predictors for the holdout set. Partial dependence plot for number of accommodates and price shows that price increases as the number of accommodates.

Var.1	RMSE	Mean.price	RMSE.price
Apartment size	NA	NA	NA
large apt	71.74585	185.9207	0.3858949
small apt	62.09615	154.7917	0.4011595
Type	NA	NA	NA
Apartment	73.55792	236.1250	0.3115211
Condominium	86.55398	201.3830	0.4297979
Entire Unit	64.35807	162.9873	0.3948654
Loft	66.96040	186.8140	0.3584336
Neighbourhood	NA	NA	NA
Bronx	54.40533	128.2857	0.4240950
Brooklyn	60.19156	152.8951	0.3936788
Manhattan	71.42782	184.2427	0.3876833
Queens	57.87346	131.7437	0.4392883
Staten Island	56.82891	103.4500	0.5493370
All	66.07839	167.0940	0.3954563





Performance Across Subsamples

Examining the fit in the various sub samples can inform us about external validity of the prediction.

Actual vs Predicted Price another post prediction diagnostics is comparing the predicted prices versus the actual prices. The figure below shows that prediction does a better job for lower than higher prices.

Compraring results with case study:

In the case study of Gabors book for Data Analysis one of the main predictors is different categories of room type, however in this case study of New York, there was only one type of room.

The purpose of the case study was to build a model to predict property rental prices in london. In the case study as a result of running different models GBM won the contest in terms of best fit. It just marginally beating random forest. Based on the case study result Random Forest worked well and faster and importantly random forest is relatively easy to implement. In this case study to build a prediction model for New York prices, GBM outperformed however models, same as the case study in the book Random Forest was the second best for New York.

In the case of London Airbnb prices the top most important variables are number of accommodates, room type, neighbourhood and beds. In the case study of New York the most important variables are: number of bathrooms and accomodates, and other amenities such as washer, gym and elvator.

Conclusion

The goal of this report was to find a better model to predict Airbnb prices in New York for a small to mid-size apartments. Five models were illustrated to compare and contrast across models performance. GBM resulted to be the best model by 65.38 USD RMSE, besides this GBM model also shows better performance in holdout set with 64.38 USD RMSE. The second best model was basic Random Forest which has highlights meaningful characteristics about the nature of Airbnb apartments in New York. Key price drivers based on post prediction diagnostics are the number of bathroom, number of accommodates, and availability of amenities such as washer and gym. partial dependence plot also illustrated that the model better predicts Manhattan neighborhood.

Notes:

The codes for the regressions are taken from seminars of Data Analysis 3, Gabors book of Data Analysis for Business, Economics and Policy. Moreover, I would like to source that amenities transformation are taken from this GitHub repository. However, I further added small changes to the codes.

Appendix

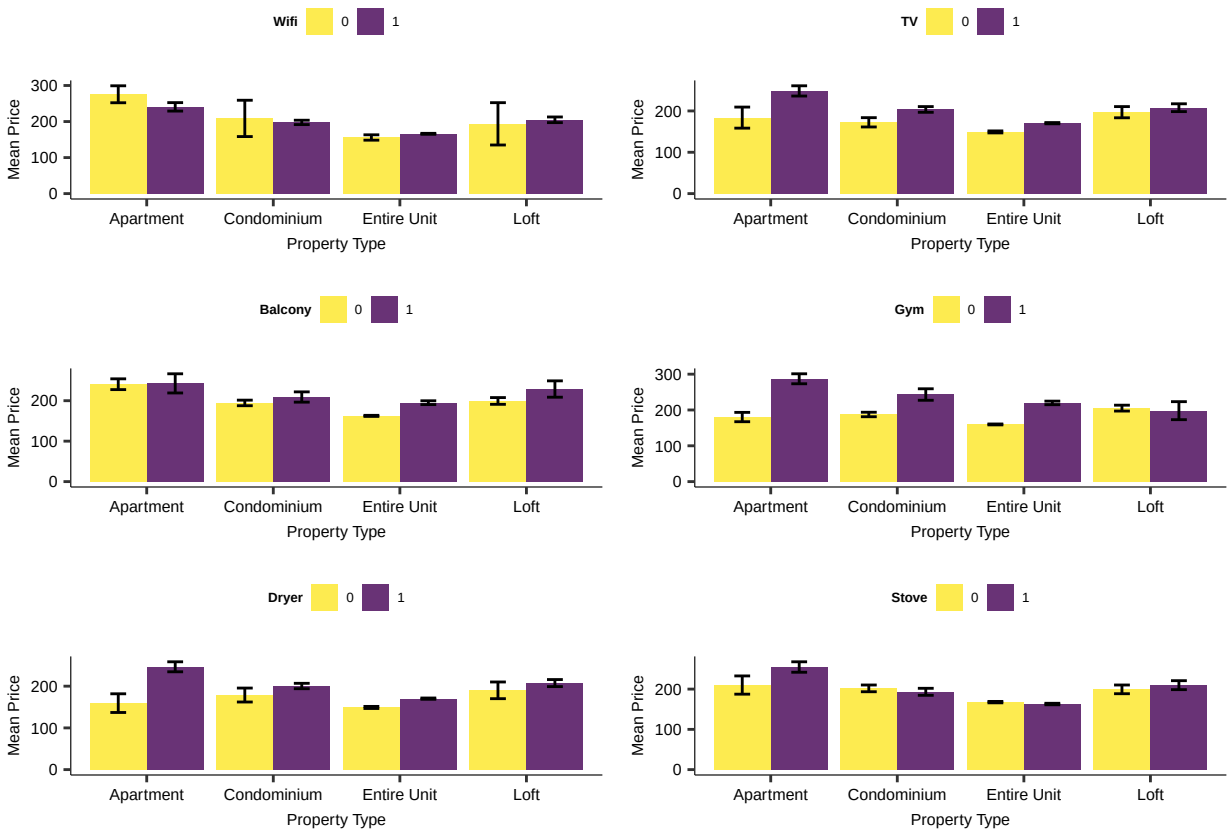


Table 3: Accommodates Summary

	Accommodates	N	min	max	Percent	mean
price	2	6863	10.00	499.00	42.18	155.00
	3	2795	30.00	499.00	17.18	160.95
	4	4285	10.00	499.00	26.34	178.80
	5	1140	41.00	499.00	7.01	201.32
	6	1188	50.00	499.00	7.30	217.74

Table 4: Beds Summary

	Beds	N	min	max	Percent	mean
price	1	8263	10.00	499.00	50.78	159.13
	2	5489	38.00	499.00	33.73	175.11
	3	1844	30.00	499.00	11.33	193.89
	4	529	45.00	495.00	3.25	197.90
	5	85	75.00	429.00	0.52	202.06
	6	55	70.00	477.00	0.34	206.13
	7	2	125.00	250.00	0.01	187.50
	8	2	100.00	206.00	0.01	153.00
	9	1	156.00	156.00	0.01	156.00
	10	1	350.00	350.00	0.01	350.00

Table 5: Bedrooms Summary

	Bedrooms	N	min	max	Percent	mean
price	1	11546	10.00	499.00	70.96	158.11
	2	3977	30.00	499.00	24.44	194.58
	3	696	51.00	499.00	4.28	225.53
	4	49	90.00	495.00	0.30	220.14
	5	3	125.00	350.00	0.02	275.00

Table 6: Room Type Summary

		N	min	max	Percent	mean
price	Entire/Apt	16271	10.00	499.00	100.00	170.12

Table 7: Property Type Summary

	Property Type	N	min	max	Percent	mean
price	Apartment	265	79.00	499.00	1.63	241.17
	Condominium	991	40.00	499.00	6.09	197.90
	Entire Unit	14483	10.00	499.00	89.01	165.65
	Loft	532	65.00	495.00	3.27	204.67

Table 8: Host Response Time Summary

	Room Type	N	min	max	Percent	mean
price	a few days or more	594	48.00	480.00	3.65	182.38
	N/A	6263	10.00	499.00	38.49	160.46
	within a day	1661	30.00	495.00	10.21	167.82
	within a few hours	2244	39.00	499.00	13.79	174.52
	within an hour	5509	30.00	499.00	33.86	178.67

Table 9: Minimum nights Summary

	N	min	max	mean
Min Nights	16271	1.00	1250.00	22.48

Table 10: Neighbourhood Group Summary

Neighbourhood Group	N	min	max	mean
Bronx	289			
Brooklyn	5948			
Manhattan	8678			
Queens	1265			
Staten Island	91			