Airbnb Price Prediction & Analysis

Team 1

Agenda

1. Business Problem/ Introduce Dataset
2. Data Cleaning
3. Show Descriptive Analyses in R
4. Model & Main Results & Challenges
5. Conclusion

Imagine you are an House owner...

Project Mission

Use the property and host information to predict a reasonable price.

From Airbnb price prediction dataset

Dataset

- **Kaggle - Airbnb price prediction (Feb, 2018)**
- **29 columns with Mixture of data types : String/Categorical/Numerice**
- **74k rows**

Detail | Compact | Column — 29 of 29 columns

id	log_price	property_t...	room_type	amenities	accommo...	bathrooms
6901257	5.010635294096256	Apartment	Entire home/apt	{"Wireless Internet","Air conditioning",Kitchen,Heating,"Family/kid friendly",Essentials,"Hair dryer...	3	1.0
6304928	5.1298987149230735	Apartment	Entire home/apt	{"Wireless Internet","Air conditioning",Kitchen,Heating,"Family/kid friendly",Washer,Dryer,"Smoke de...	7	1.0
7919400	4.976733742420574	Apartment	Entire home/apt	{TV,"Cable TV","Wireless Internet","Air conditioning",Kitchen,Breakfast,"Buzzer/wireless intercom",H...	5	1.0
13418779	6.620073206530356	House	Entire home/apt	{TV,"Cable TV",Internet,"Wireless	4	1.0

Data Cleaning

Delete: **ID, description, zip code, neighborhood,**

Deal with missing Data: numerical variables -> impute with mean

categorical ->drop

Create Dummies using one hot encoding

Data Cleaning

Convert to time length: **host_since**

Clean categorical variable by regex: **Amenities**

amenities
{"Wireless Internet","Air conditioning",Kitchen,Heating,"Family/kid friendly",Essentials,"Hair dryer",Iron,"translation missing: en.h
{"Wireless Internet","Air conditioning",Kitchen,Heating,"Family/kid friendly",Washer,Dryer,"Smoke detector","Fire extinguisher",E
{TV,"Cable TV","Wireless Internet","Air conditioning",Kitchen,Breakfast,"Buzzer/wireless intercom",Heating,"Family/kid friendly",
{TV,"Cable TV",Internet,"Wireless Internet",Kitchen,"Indoor fireplace","Buzzer/wireless intercom",Heating,Washer,Dryer,"Smoke d

amenities_Air conditioning	amenities_Bath towel	amenities_Bathtub	amenities_Coffee maker
1	0	0	0
1	0	0	0
1	0	0	0

Final: 117 column and over 73k rows

Create Train and Test Dataset with ratio of 30-70

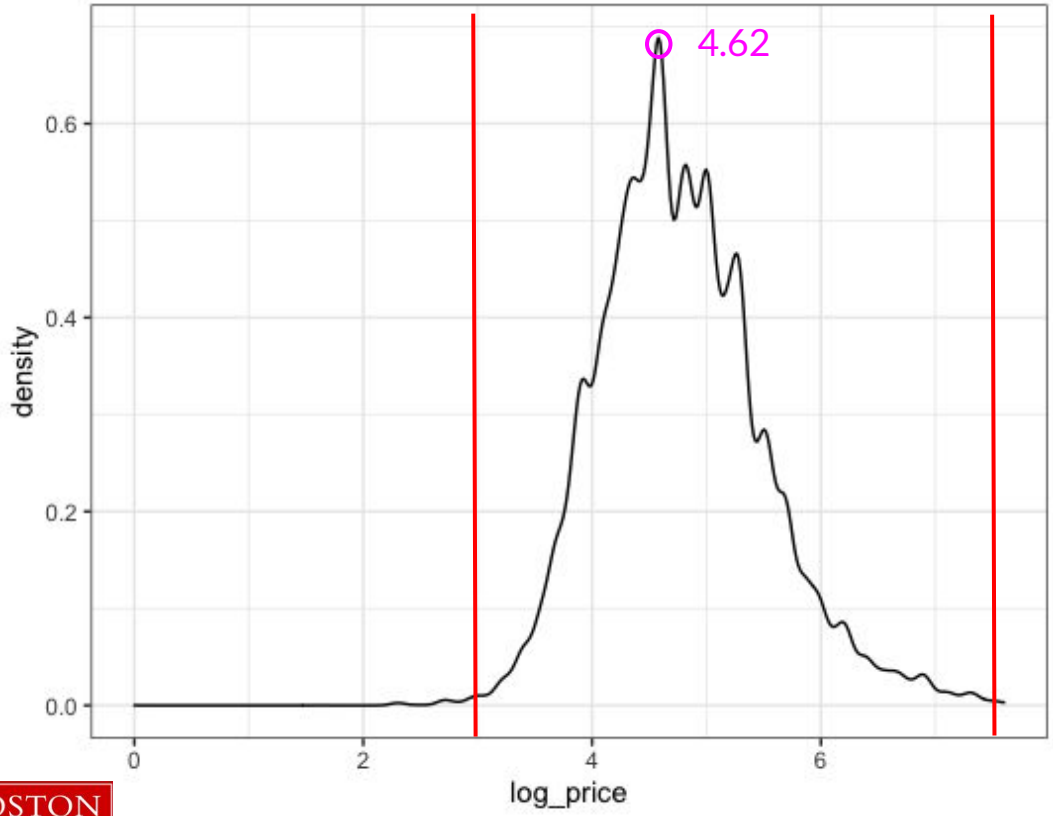
Descriptive Analyses

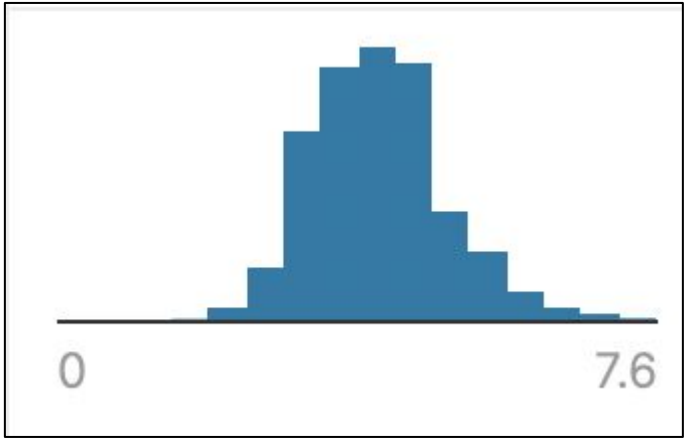
Three Dimensions:

1. Single variable plot- distribution of the predicted variable
2. Two-variable plot - rough relationship plot
3. Three- variable plot- with a hue

Descriptive Analyses

Single Variable - the Predicted Variable

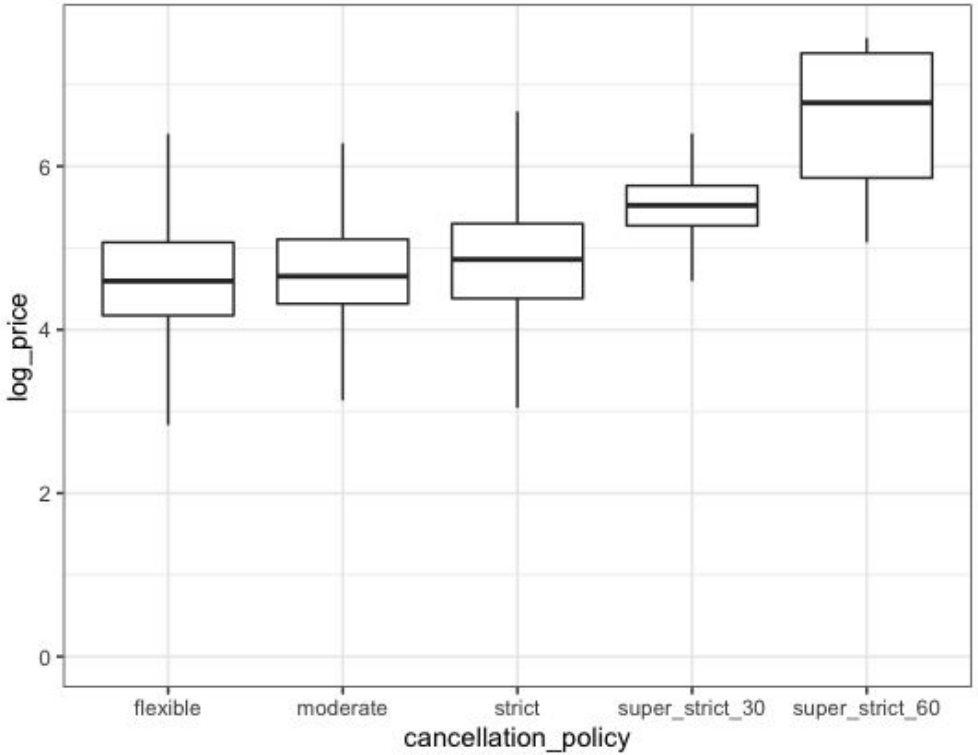




- Central range: [3, 7]
  = Price range: [20.08, 1998.19]
- Highest density: 4.62 (about 70%)
  = Price: 101.49

Descriptive Analyses

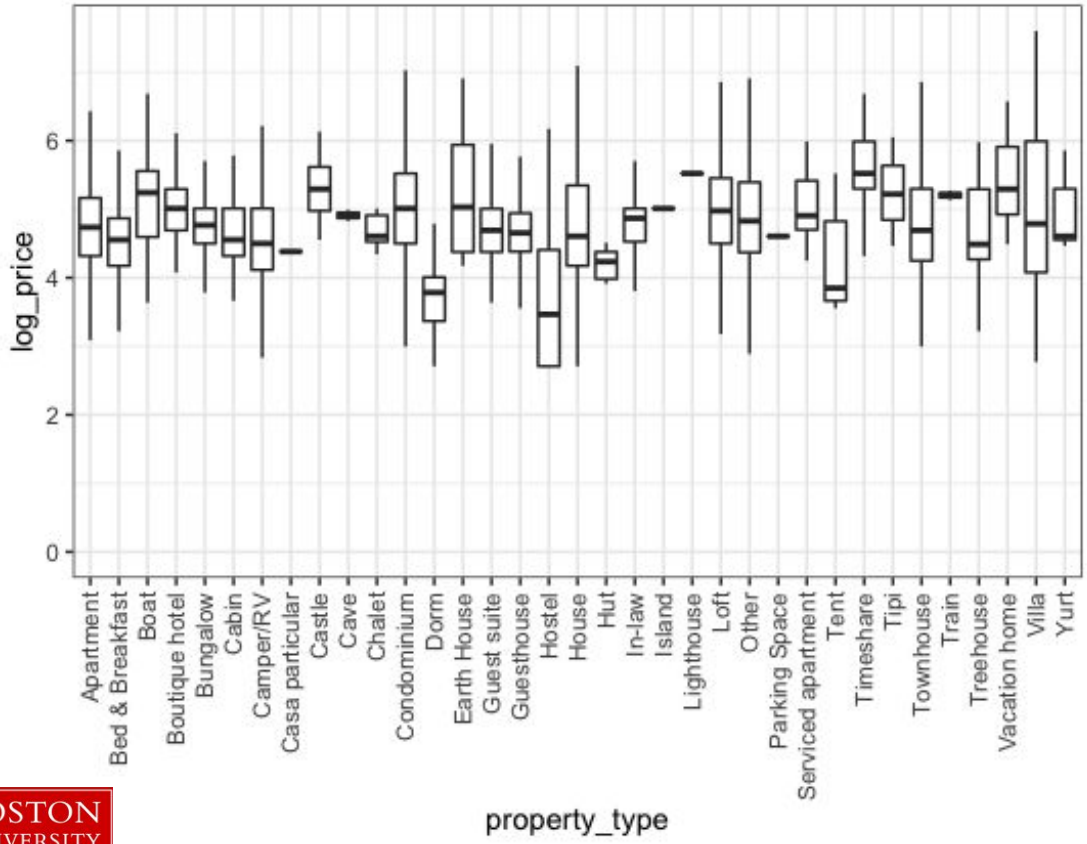
Two-variable plot - the predicted variable with one predictor (simple)



- **Assumption**: positive relationship: stricter policy - higher price
- **Result**: positive relationship
- Median
- Range between the 25th and 75th percentile
- Outliers are omitted in this case

Descriptive Analyses

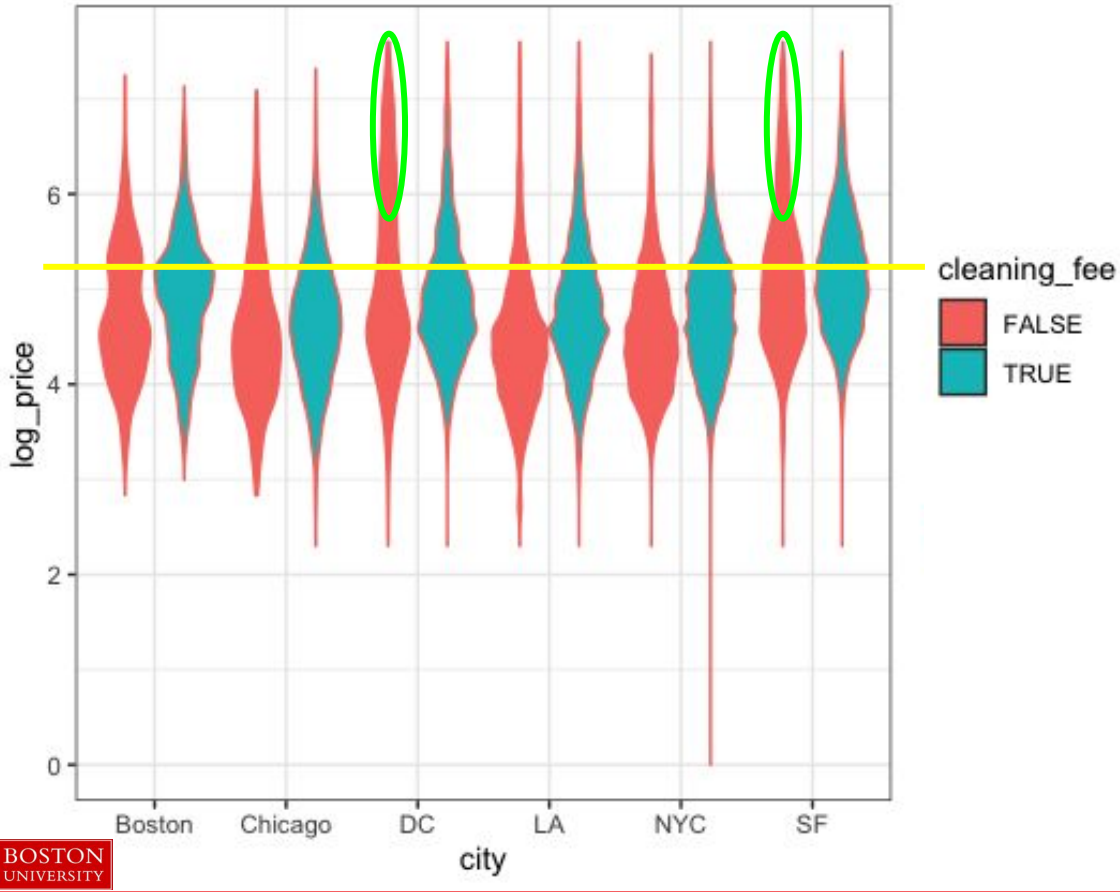
Two-variable plot - the predicted variable with one predictor (complicated)



- Log price and property types
- Highest median : Tent and lighthouse
- Assumed reason: High cost
- Interesting finding: Cave and island

Descriptive Analyses

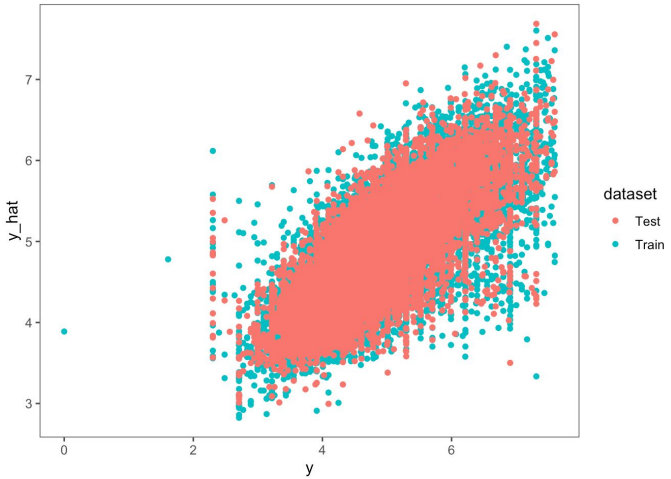
Three-variable plot - the predicted variable with two predictors (hue added)



- Mean and median are not enough
- The distribution of log price with more variables
- Cleaning fee: higher log price
- Similar ranges but different distributions
- Fatter tails for DC and SF

Main Results

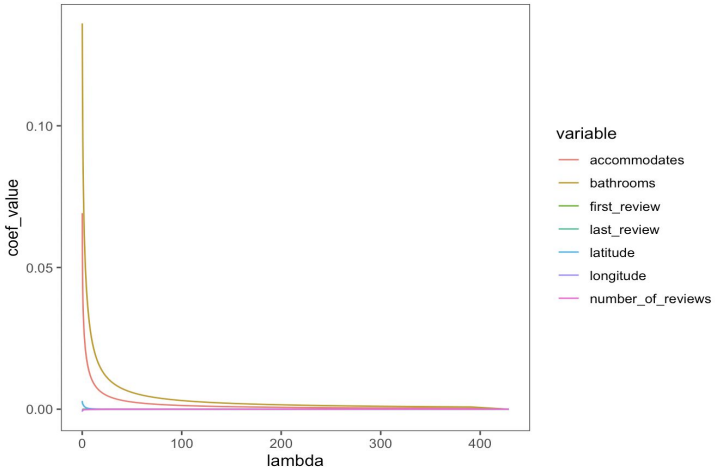
Linear Regression



MSE Train: 0.2072

MSE Test: 0.2109

Ridge Regression



MSE Train: 0.2173

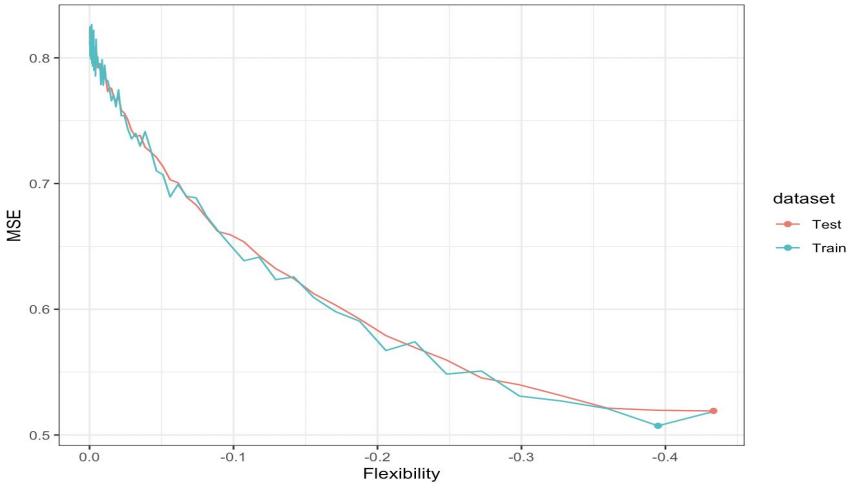
MSE Test: 0.2170

BOSTON UNIVERSITY

airbnb

Main Results II

Lasso



MSE Train: 0.3948

MSE Test: 0.4333

Challenge

```
## amenities_Air conditioning          6.298421e-02
## amenities_Bath towel                1.282988e-02
## amenities_Bathtub                  -5.438652e-03
## amenities_Coffee maker              3.610599e-02
## amenities_Cooking basics            4.225581e-02
## amenities_Dishes and silverware     3.310074e-02
## amenities_Elevator                  1.698072e-01
## amenities_Hot water                -8.905121e-03
## amenities_Internet                  2.836596e-02
## amenities_Kitchen                  -3.281136e-02
## amenities_Private bathroom          1.947149e-01
## amenities_Refrigerator             -7.038374e-02
## amenities_Self Check-In            -5.223783e-02
## amenities_Stove                    -1.871192e-02
## amenities_Toilet paper              9.471573e-03
```

Some of coefficients can not be explained

BOSTON UNIVERSITY

airbnb

Main Results III

Bagging

```
fit.bagging <- bagging(
  formula = y_train~.,
  dd = dd_train,
  nbagg = 100,
  coob = TRUE,
  control = rpart.control(
    minsplit = 2,
    cp = 0.01
  ))
```

MSE Train: 0.2510

MSE Test: 0.2536

Random Forests

```
fit.rndfor <- randomForest(
  y_train~.,
  x_train,
  ntree = 100,
  do.trace = T
)
```

MSE Train: 0.0305

MSE Test: 0.1541

Boosting

```
fit.btree <- gbm(
  f1,
  data = dd.train.sample,
  distribution = "gaussian",
  n.trees = 100,
  interaction.depth = 3,
  shrinkage = 0.1,
  cv.folds = 5
)
```

MSE Train: 0.1802

MSE Test: 0.1850

Boosting

- n.trees = 100, interaction.depth = 1, shrinkage = 0.001, mse_train: 0.4796789; mse_test: 0.48406
- n.trees = 100, interaction.depth = 1, shrinkage = 0.01, mse_train: 0.3406027; mse_test: 0.3454715
- n.trees = 100, interaction.depth = 1, shrinkage = 0.1, mse_train: 0.2251666; mse_test: 0.2283994

- n.trees = 100, interaction.depth = 2, shrinkage = 0.001, mse_train: 0.4711037; mse_test: 0.4750863
- n.trees = 100, interaction.depth = 2, shrinkage = 0.01, mse_train: 0.3064798; mse_test: 0.3099995
- n.trees = 100, interaction.depth = 2, shrinkage = 0.1, mse_train: 0.1902003; mse_test: 0.194123

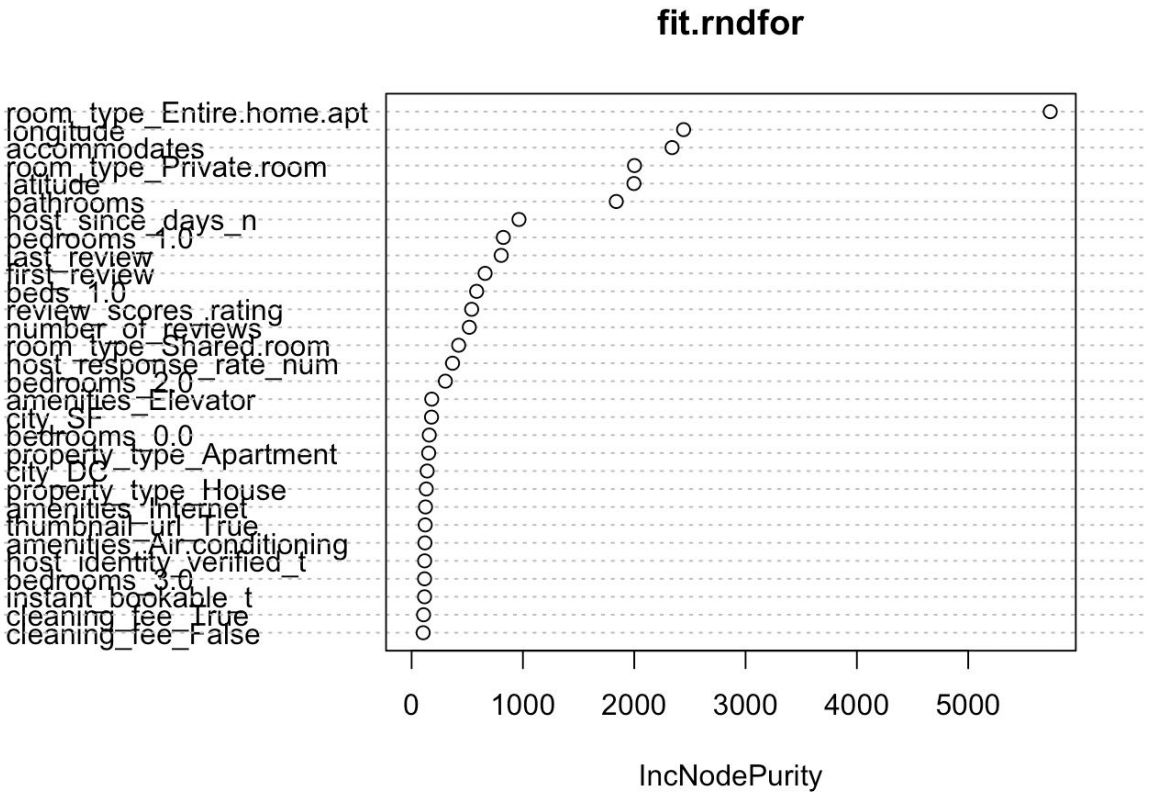
- n.trees = 100, interaction.depth = 3, shrinkage = 0.001, mse_train: 0.4696736; mse_test: 0.4735253
- n.trees = 100, interaction.depth = 3, shrinkage = 0.01, mse_train: 0.2963097; mse_test: 0.299599
- n.trees = 100, interaction.depth = 3, shrinkage = 0.1, mse_train: **0.1802872**; mse_test: **0.1850507**

Challenges

- Find the best parameters
- Interpret the results

Conclusion

- Among all models, Random Forest outputs the best mse for both training and testing.
- Variables most important plotted.



Conclusion

- Most Important Variables
  1. Accommodates
  2. Bathrooms
  3. Beds
- Room type:
  1. Positive: Entire home/apt, Private room
  2. Negative: Tent, Hostel, Shared room, Dorm
- Amenities: Private bathroom, Internet, Elevator
- Bed type:
  1. Positive: Couch, Real Bed, Sofa
  2. Negative: Airbed, Futon

Q & A

BOSTON UNIVERSITY

airbnb