Predicting Airbnb Price in Hong Kong using Regression Tree, Random Forest and Boost

Data Programming with R Project

By

Zidong Li

Agenda

Problem & Objective

Data Source

Data Preprocessing

Exploratory Data Analysis

Prediction Models

Conclusion

Problem & Objective

Problem Definition

Airbnb doesn’t provide free pricing tool. So hosts have to use 3rd party software to get the estimated price

Airbnb competition is in a very high level in Hong Kong with more than 10000 Airbnb listings

Currently no previous research has done on Hong Kong Airbnb listing price

Objective

Building prediction model to predict Airbnb listing price in Hong Kong

Data Source

Collected by Inside Airbnb

(http://insideairbnb.com)

76

Number of variables

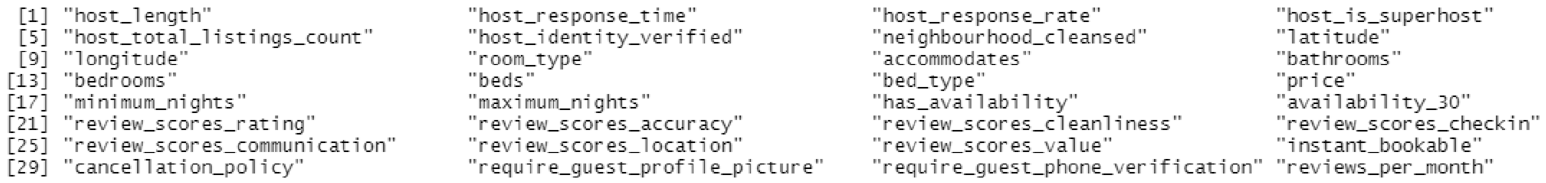
12,569

Number of unique records

Data Preprocessing

Step 1:

Dimension Reduction. Reduced from 76 to 32



```
 [1] "host_length"                 "host_response_time"           "host_response_rate"                 "host_is_superhost"
 [5] "host_total_listings_count"   "host_identity_verified"       "neighbourhood_cleansed"             "latitude"
 [9] "longitude"                   "room_type"                    "accommodates"                       "bathrooms"
[13] "bedrooms"                    "beds"                         "bed_type"                           "price"
[17] "minimum_nights"              "maximum_nights"               "has_availability"                   "availability_30"
[21] "review_scores_rating"        "review_scores_accuracy"       "review_scores_cleanliness"          "review_scores_checkin"
[25] "review_scores_communication" "review_scores_location"       "review_scores_value"                "instant_bookable"
[29] "cancellation_policy"         "require_guest_profile_picture" "require_guest_phone_verification" "reviews_per_month"
```

Step 2:

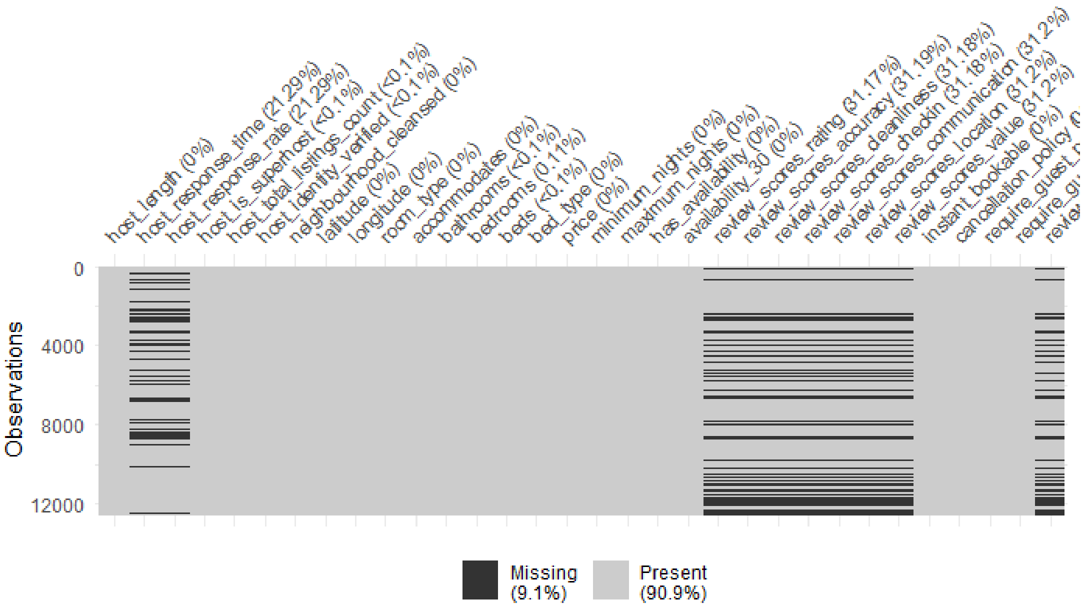
Delete Outlier records with

- Price equals 0 HKD

Step 3:

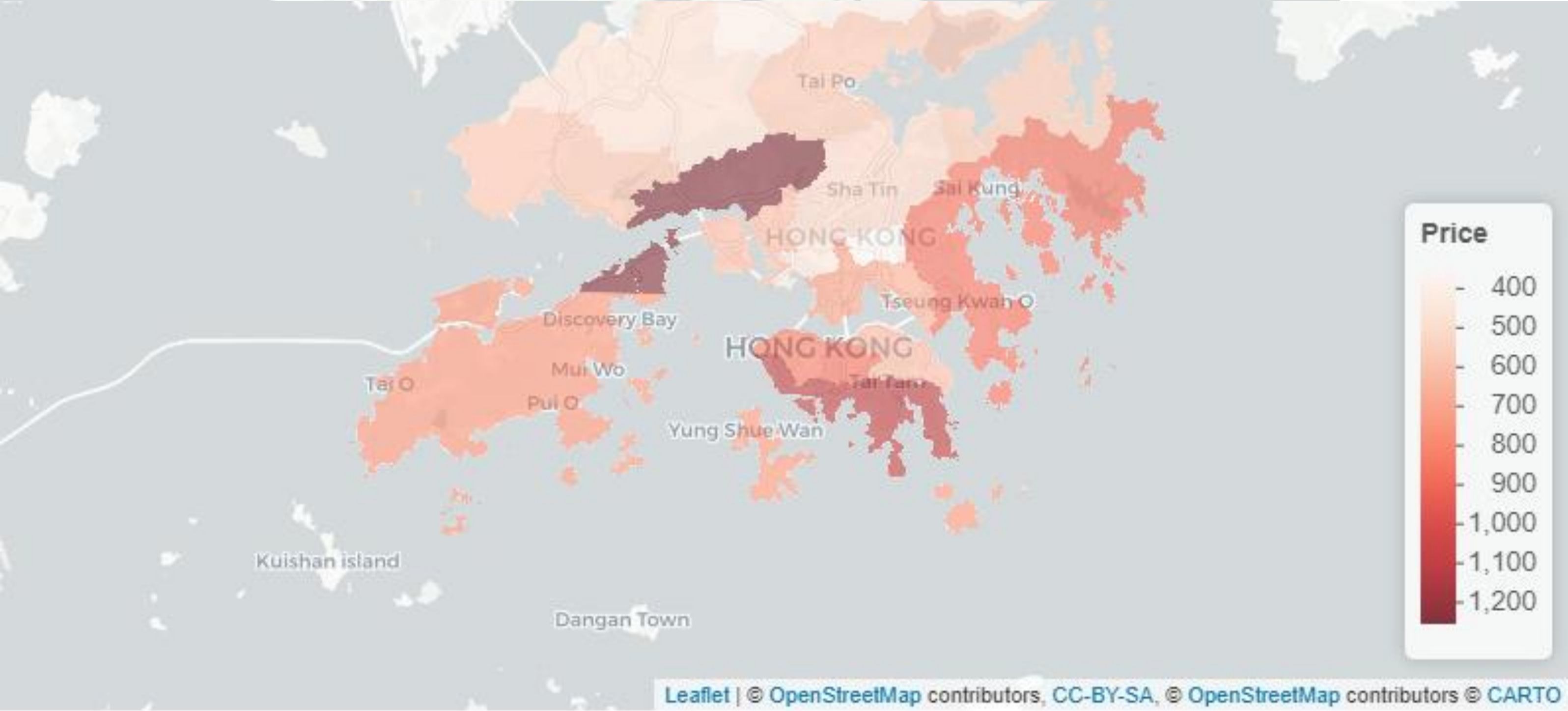
Delete Missing Values.

- Most of the missing values are in review scores and host response
- Delete missing values in review scores
- Delete variables host response time and rate



Exploratory Data Analysis

> “Descriptive Statistics provides information about the scale and type of values that the variable takes as well as tell us possible outliers that may have occurred.”
>
> -Shmueli, Yahav, I., Patel & Lichtendahl, 2016

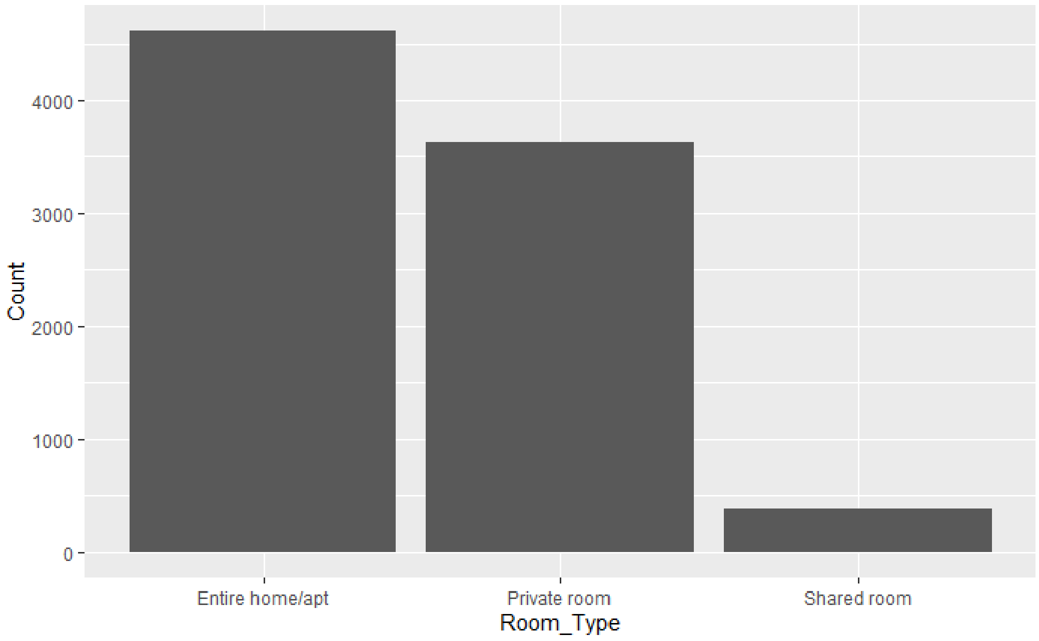


ap of number of listings by 18 regions
st listings are in Yau Tsim Mong and Hong
ng Island, which are tourist attractions

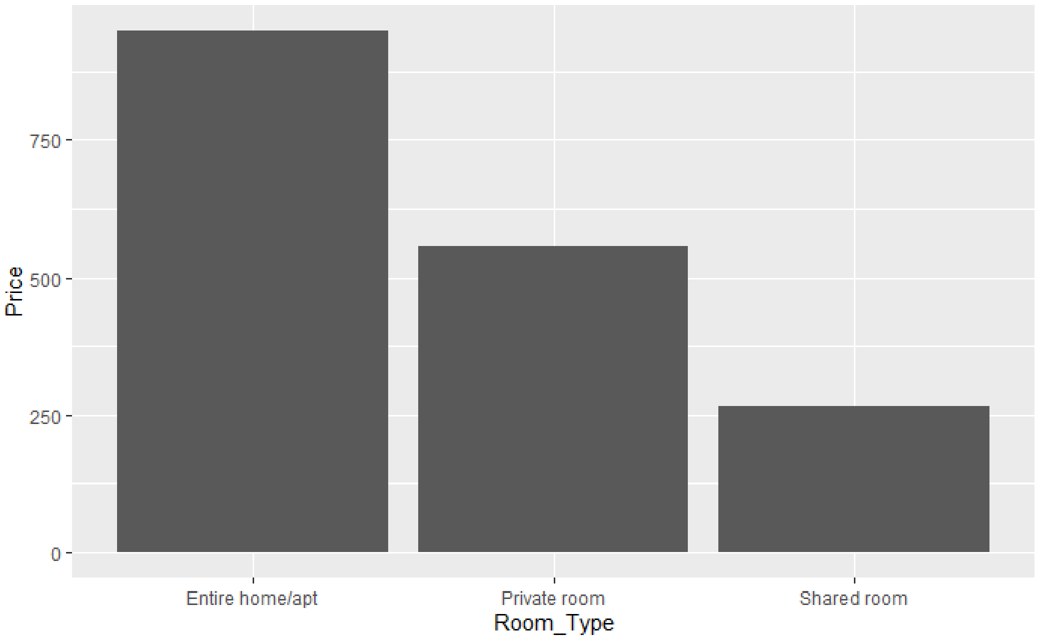
Heatmap of average listing price by 18 regions

- Most expensive regions are Tsuen Wan and South Island, which are not tourist attractions

Exploratory Data Analysis, continued



The number of each Room Type in Hong Kong Airbnb listing

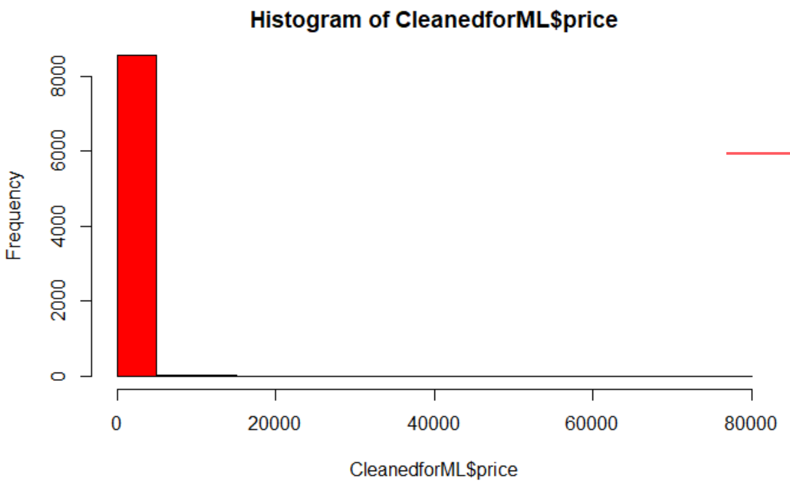


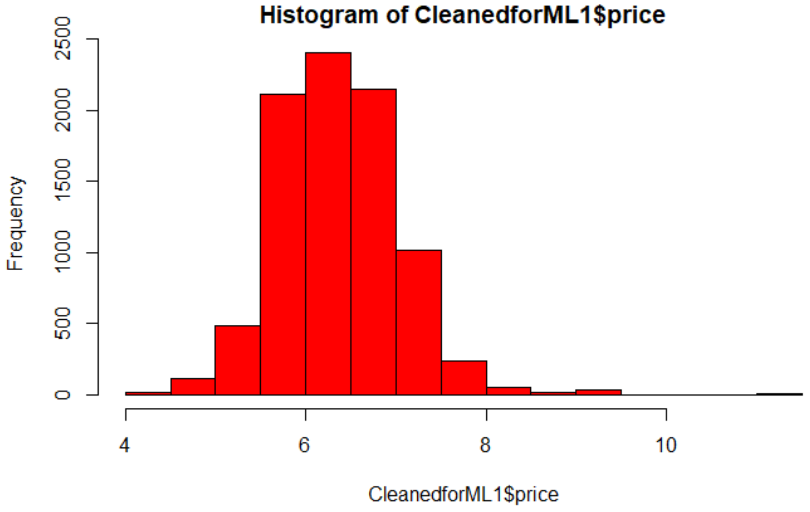
Average Price of Different Room Type

Prediction Models

Data Preprocessing Before Prediction Model

- Creating Dummies for categorical predictors
    - 18 neighbourhoods, different room types, different bed types and different cancellation policies.
- Partition (Train:Test:Validation = 50:30:20)
- Standardization for non-dummy predictors
- Log() for target variable price

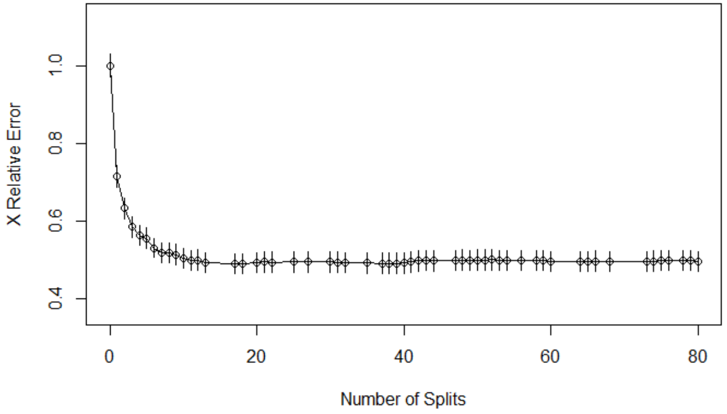




Prediction Models

Model 1: Regression Tree

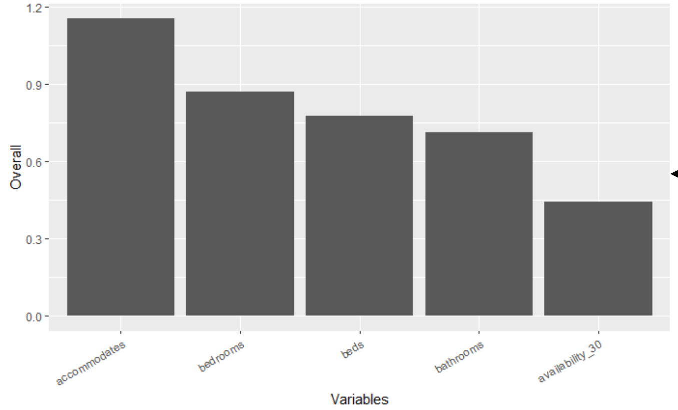
Starting Parameters:
- Cp = 0.001



Revised Parameters:
- Number of decision trees is 16 (cp = 0.031)

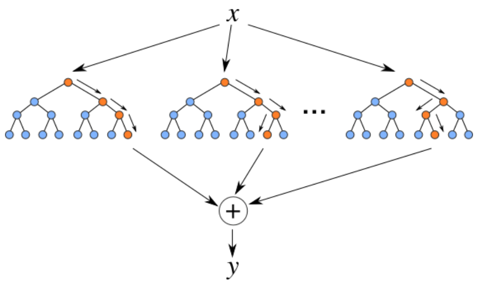
Validation Prediction RMSE: **0.2033,** MAE: **0.3122**

23 out of 56 predictors are significant in models



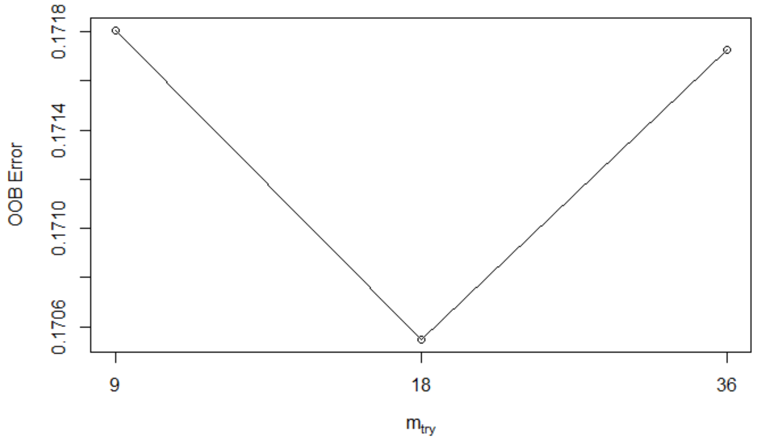
Prediction Models

Model 2: Random Forest



Tuning Parameters using tunerf():

- Number of decision trees is 100



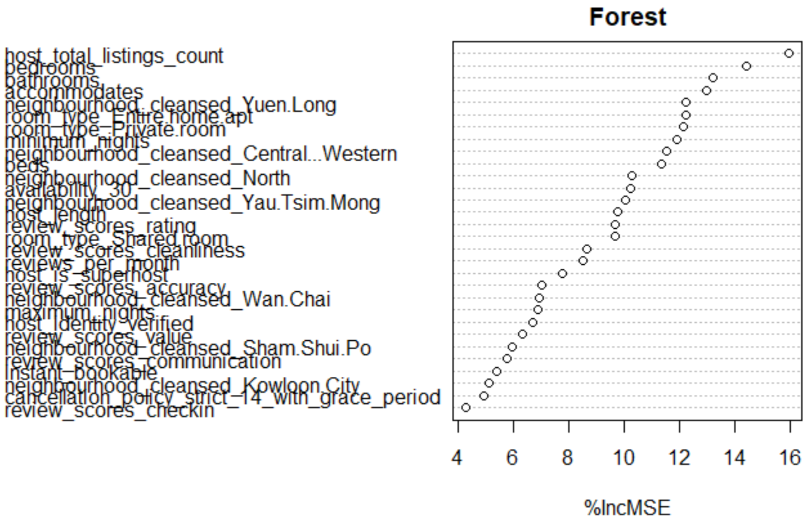
Revised Parameters:

- Number of decision trees is 100
- Mtry = 18

Validation Prediction

RMSE: **0.1035,**

MAE: **0.1953**



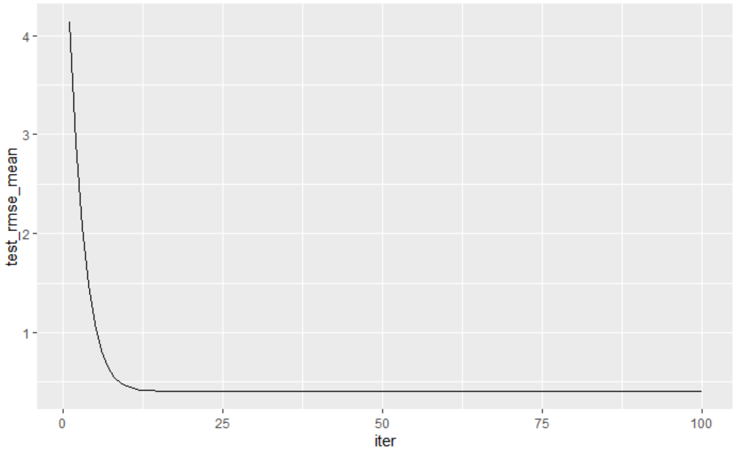
Prediction Models

Model 3: Gradient Boosting

dmlc XGBoost

Tuning Parameters using xgb.cv:

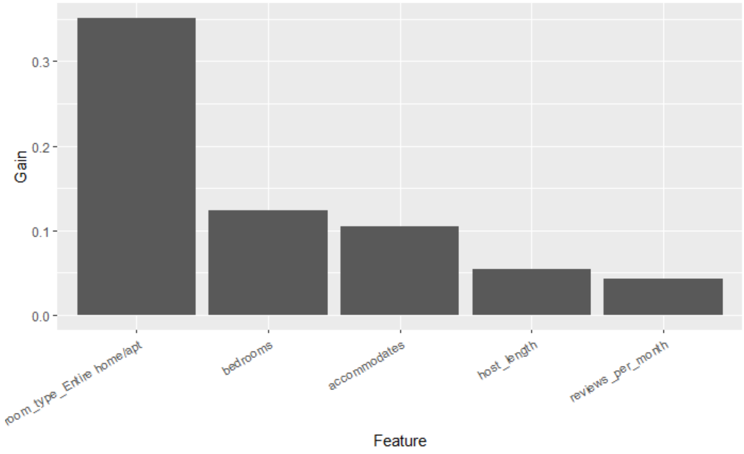
- 5-fold cross validation
- Number of decision trees is 100
- Max-Depth of Tree is 15



Revised Parameters:

- 5-fold cross validation
- Number of decision trees is 12
- Max-Depth of Tree is 15

Validation Prediction
RMSE: 0.1237,
MAE: 0.2170



Conclusion

Model Comparison

Prediction Models	Validation RMSE	Validation MAE
Regression Tree	0.2033	0.3122
Random Forest	0.1035	0.1953
Gradient Boosting	0.1237	0.2170

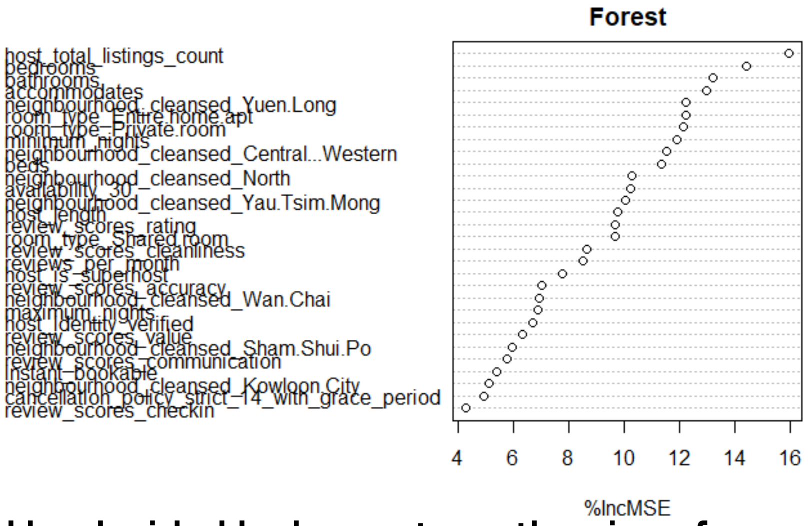
- Using ensemble learning can lead to lower RMSE and Mae (higher accuracy), compared with single tree algorithm
- Between Random Forest and Boosting, Random Forest can give higher accuracy on validation set.

Conclusion

Random Forest Implication and Limitation

Top 10 Important features of random forest model:

- House Types: entire home/apt and private room
- Attributes of house: # of bedrooms, bathrooms, beds and accommodates
- Location: Yuen Long and Central/Western
- Others: Host total listings count and minimum nights



For hosts, when reevaluating Airbnb property price, price of the property should be decided by house type, the size of house, total listings number and minimum nights that they set.

Limitation: Since the model only explains 62% of price, still 38% of price could be explained by other variables that aren't under consideration. For future plan, more variables need to be explored and considered

Thank you!