Predicting Airbnb listing prices for host in London city

Presented by Group 12
Aparna Bhat
Neha Mehta
Olivia Rodrigues
Prakash Chokkalingam
Prathyusha Sathineni

Problem Statement

- Airbnb offers complete independence to its host to price their properties with minimal pointers. Currently, hosts compare similar listings in their neighbourhood so as to come up with competitive price.
- Since the number of hosts are increasing in Airbnb, coming up with the right price so as to remain competitive in a hosts neighbourhood is very important.
- If prices are too low or too high, it is going to affect the bookings of property and eventually Airbnb will be missing a lot of potential income.

Dataset(Source Variables)

Description

- DataSource:https://www.kaggle.comabdmitriy/airbnb#listings.csv
- Last scraped in November 2019
- Records: 85000+
- Variables: 106

Map

Map based on average of Longitude and average of Latitude. Details are shown for Smart Location and Neighbourhood.

Data cleaning

- **We have dropped variables based on :-**
  - Redundancy (e.g. street - smart location)
  - Missing values
  - the number of levels present in the categorical data. (e.g. There were 96 unique categories for calendar_updated, a host might update their calendar for multiple different reasons).

- Special characters like $ and % , we have removed them from the dataset.

- **Handling missing values:**
  - In variables like bathrooms, bedrooms, and beds, we have replaced the missing values with the median value.
  - In pricing variables like security deposit, price, and cleaning fee, we have replaced the missing values with the minimum value - zero ($0).

Data cleaning

- Similarly, in minimum nights and maximum nights variables, we replaced the values greater than 365 with 365, because in a year we have only 365 days.
- In property type, we had 44 categories. So, we have combined the records and formed 3 levels - (House, Apartment, and Other)
- Lastly, we have checked the data types of each variable and converted it into an appropriate data type.
- After data cleaning, our dataset consists of **63000+** records and **30** variables.

Initial Analysis

- Price Distribution: The maximum price per night is $1000.
- Beds: Maximum of the listings have 1 or 2 beds.
- The average price of room types entire home/apt and hotel room is almost equal.
- The entire home/apt has the highest cleaning fee of about $50 compared to that of the other room types.
- Out of 33 neighborhoods, Kensington and Chelsea are having the highest average price value of around $168.
- The maximum number of listings are in the Westminster neighborhood.
- The maximum of the listings is of property type Apartment.
- Most of the hosts are not qualified as super hosts and also their identity is also not verified.

Plots - (link)

Preprocessing

- **Selected Variables**

host_response_rate,host_is_superhost,host_listings_count,host_identity_verified.id,neighbourhood,property_type,accommodates,bedrooms,bathrooms, beds, price, security_deposit, cleaning_fee, guests_included, extra_people, availability_90, number_of_reviews, review_scores_rating, instant_bookable

- **Created dummies for the categorical variables and converted them to factor**

host_is_superhost,host_identity_verified,property_type,instant_bookable

- **Converted some variables in to date time format**

host_since,first_review,last_review

Model 1 - Knn :Regression

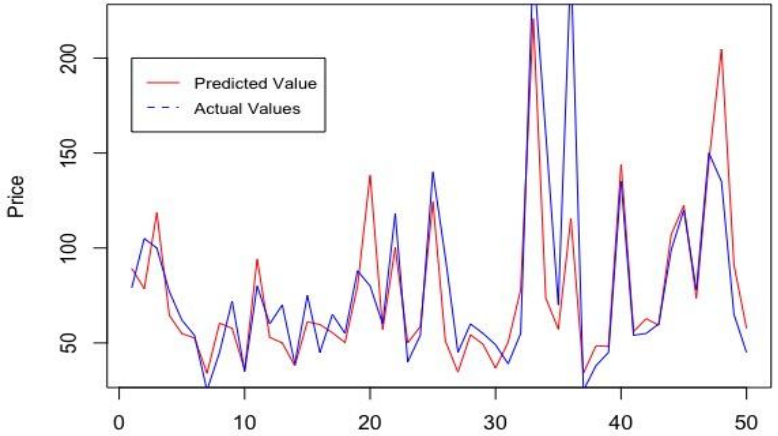
- As the value of k increases, the RMSE on validation data also increases.
- Below table displays the RMSE value for different values of k.

k	5	10	20	30	40	80	200	300
RMSE	61	60.63	60.74	61.27	61	63	65	66

- Based on the RMSE table we choose best k ie 40

  Graph displaying the actual values

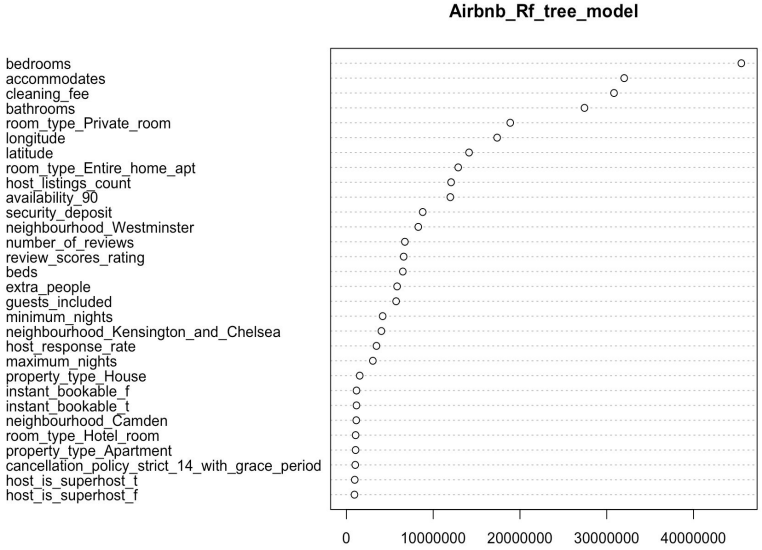
  VS the predicted values(50 records).



Model 2- Regression Trees

- In order to predict the price, we developed regression tree: (Pruned Tree), Bootstrap-Aggregating and Random Forest.
- RMSE for Random forest is lowest.
- Variables: “bedrooms”, “accomodation”, “cleaning fee”
- And “Room_type-private” are important
- Below are the results obtained for different models:

Types of RT	Prune-Tree	Bootstrap Aggregation	Random Forest
RMSE on validation data	66	68	53

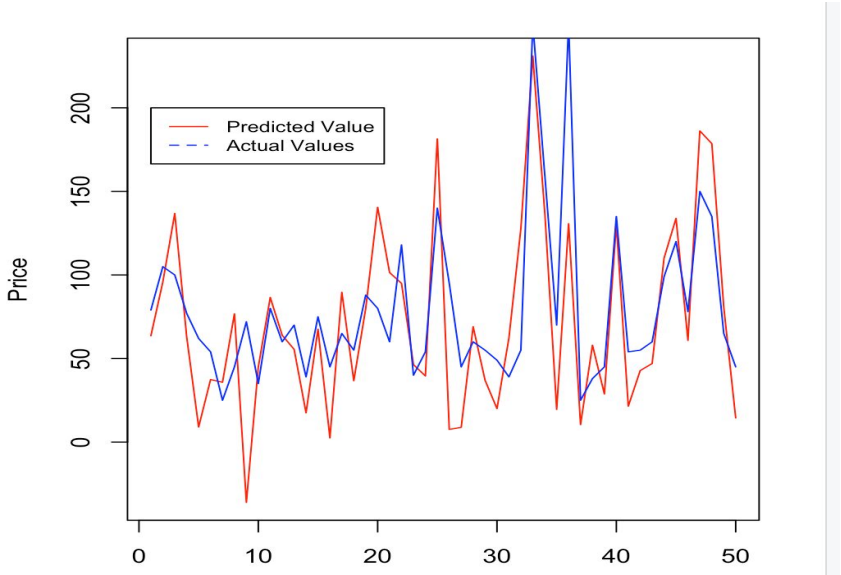


Model 3 - Multiple Linear Regression

Some variables like number_of_reviews , review_scores_rating, instant_bookable had the lowest p value which indicates that those variables are useful.

RMSE : 61

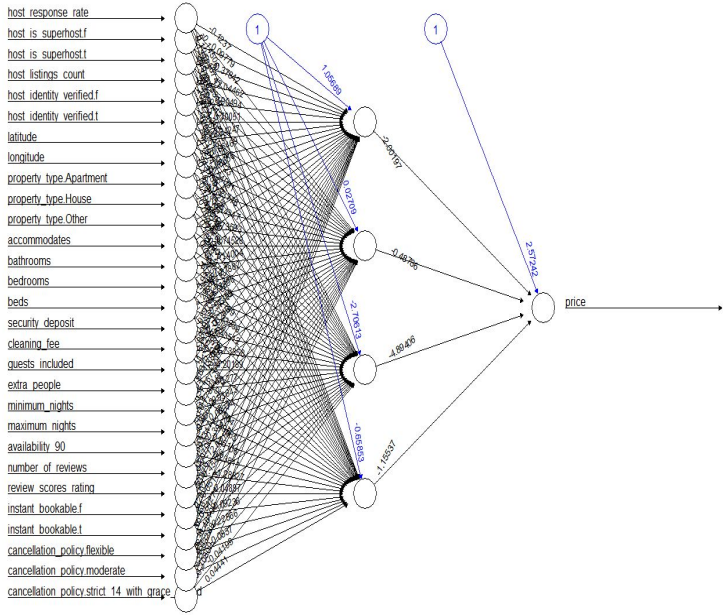
Graph shows the actual values vs predicted value.

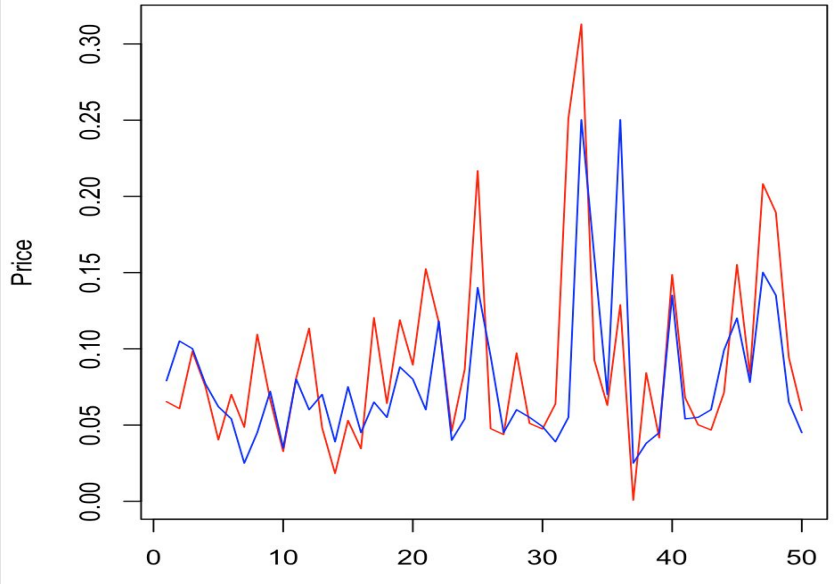


Model 4 - Neural Network

1 Hidden layer with 4 node

RMSE: 77





Analysis of results

- We developed 4 models for prediction of price.
- Multiple Linear Regression, Knn -Regression, Classification Tree and NN.
- If we compare the RMSE(Root Mean Square Error)values of all these model we found that all were in the range of 52-77.
- Amongst , all of these models we found that Random Forest Tree approach provides the best RMSE value on the Validation data.
- However, the models we developed have an minimum rmse of 52$.

Analysis of results

- Our Model provide details regarding on how variables such as “accommodation”, “Cleaning fee”, “room type”, “availability” play an role on deciding in the Price.
- “Superhost” has a very less predictability on the optimal price.
- If the user selects “Private room” or “Entire Apt” then the prices are higher.
- A few neighbourhood locations in London for eg :(Westminster, Kingston and Chelsea) tend to show high Price.

Comparison of model

Based on Accuracy metrics

Model	NN	Multiple Regression	KNN	Regression Tree
RMSE	77	61	61	53

Conclusion and Recommendations

- **Learnings:**
  - Performed exploratory analysis and Visualisation using Tableau on large dataset(80k records).
  - Found best variables which would be helpful in Price prediction for our dataset.
- **Conclusion:**
  - The optimal price arrived through predictive techniques would be useful to increase the occupancy rate of hosts ,thereby increasing the revenue for hosts and in turn for airbnb.
- **Target host groups for recommendation:**
  - Price reduction recommendation would be suggested to non super host with higher than optimal price.(Super host might have a exceptionally good value which justifies the price)
  - Price increase recommendation would be suggested to Super host with lower than optimal price.(Non Super host might have a few gaps in quality and hence we refrain from price increase recommendation).