**IST707 Final Project Report: Airbnb Listings - San Diego**

Zequn Che & Jiyu Feng

04/28/2020

**Introduction**

Airbnb is the abbreviation of “AirBed and Breakfast”. It is an online service that helps tourists to

arrange their travel experiences and homestays. Airbnb was founded in August 2008, in San

Francisco. Users can use filters to find desirable houses and complete online booking processes

through the internet or mobile apps. Airbnb has no real estate or host events on its rental

listings. The company only provides services as brokers from each booking. According to the public media, the company provides tens of millions of unique homestay choices in 192 countries and operates in 65,000 cities. In 2011, services from Airbnb increased 800% incredibly.

For this project, our group will use a dataset about Airbnb listings of San Diego from 2008 to 2016. These data include personal contact information of hosts, details of their rentals, and customers’ reviews.

**Dataset Description**

The dataset is downloaded from the “Opendatasoft” website which retrieves data from a website called “Inside Airbnb”. It is an independent, non-commercial tool that allows people to obtain data and explore how Airbnb is being used in different cities and countries. The dataset contains more than 80 attributes and 5645 records with hosts’ personal information, rental details, and customers’ reviews. For our project, we will preprocess these attributes like removing variables with no predictive values and converting data types for fitting the models.

**Project Goal**

The main goal of this project is to perform analyzing processes to the audience and find insights into the rental industry and sharing economy by using Airbnb data sources.

For the project, we design to use two methods to solve the main goal based on two different standpoints:

1. From the customer’ perspective:

find relationships between prices and other elements

1. From the host’ perspective:

Predict the range of the list price for his or her rental with Airbnb in San Diego

**Data Analysis Method**

For achieve the project goal, we will use these two methods to solve concerns above:

1. Association Rules: finding patterns of the price and other factors

Help the audience to understand why and which attributes influence the level of the price.

1. Decision Tree: predict the price range for the new host

Using classification methods to help hosts to set the appropriate list price in San Diego.

**Data Preprocessing**

The raw data we got has 89 attributes and 6201 records. Most of the attributes are not very valuable for analyzing. For better doing the analysis, we need to remove the attributes which we are not using. Moreover, we will remove some records which have a lot of missing data instead of editing the missing data.

We keep 12 attributes and 4620 records for a high accuracy analysis:

* “Host.Total.Listings.Count” is an attribute of the counting owned by the host which shows the market share. It shows the host’s ability to control the price.
* “Neighbourhood.Cleansed” is an attribute of the neighbourhood which indicates the situation and quality of a certain neighbourhood. It might influence the price.
* “Property.Type” and “Room.Type” intuitively show the types of housing. They would be good attributes to analyze the price.
* “Accommodates” shows the capacity of housing which might be a good attribute for analyzing.
* “Bathrooms”, “Bedrooms”, “Beds” shows the essential infrastructure in the houses. They will be very important attributes to analyzing.
* “Bed.Type” is the type of beds including real beds, air bed, couch, pull-out sofa, futon. This attribute might influence the price as well.
* “Price” is the target attribute which we are going to predict.
* “Cleaning Fee” and “Cancellation Policy” are extra attributes which will increase the costing of the whole service.

For better analyzing the data, we cut the attribute “Price” into two levels. (0,181,Inf) as ("low", "high")

We make the “property type”, “room type”, “Neighbourhood.Cleansed”, “Bed.Type”, “Cancellation.Policy” to factors from strings.

We set all the numeric attributes as numeric for making sure all of them are good for analyzing.

Then we omit all the missing values.

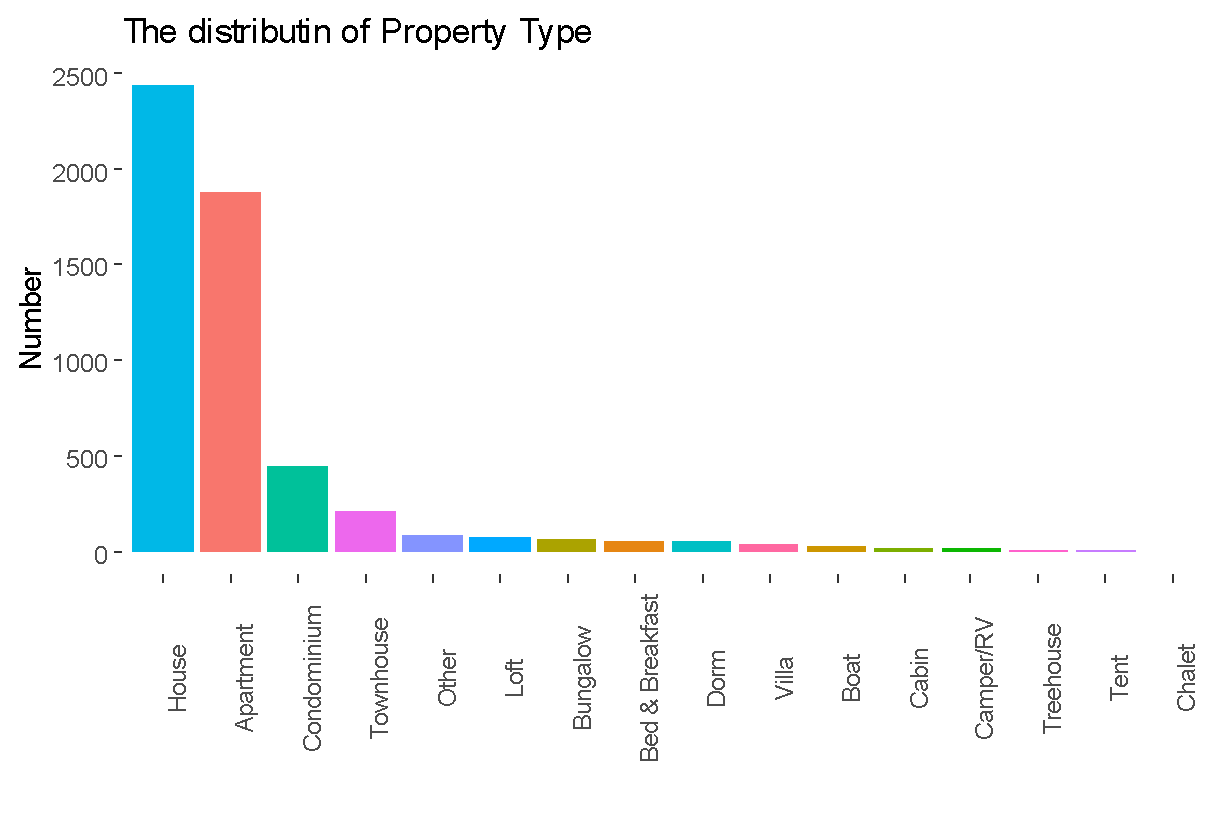
Now the data is well-prepared for future analyzing.

Here is the bulleted list:

* "Host.Total.Listings.Count"
* "Neighbourhood.Cleansed",
* "Property.Type"
* "Room.Type"
* "Accommodates"
* "Bathrooms"
* "Bedrooms"
* "Beds"
* "Bed.Type"
* "Price"
* "Cleaning.Fee"
* "Cancellation.Policy"

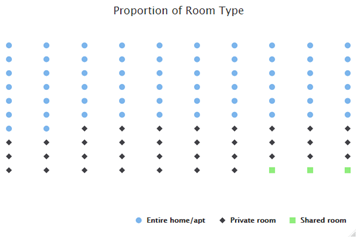
**Data Explore: Visualization(Partial important features)**

* The distribution of property type



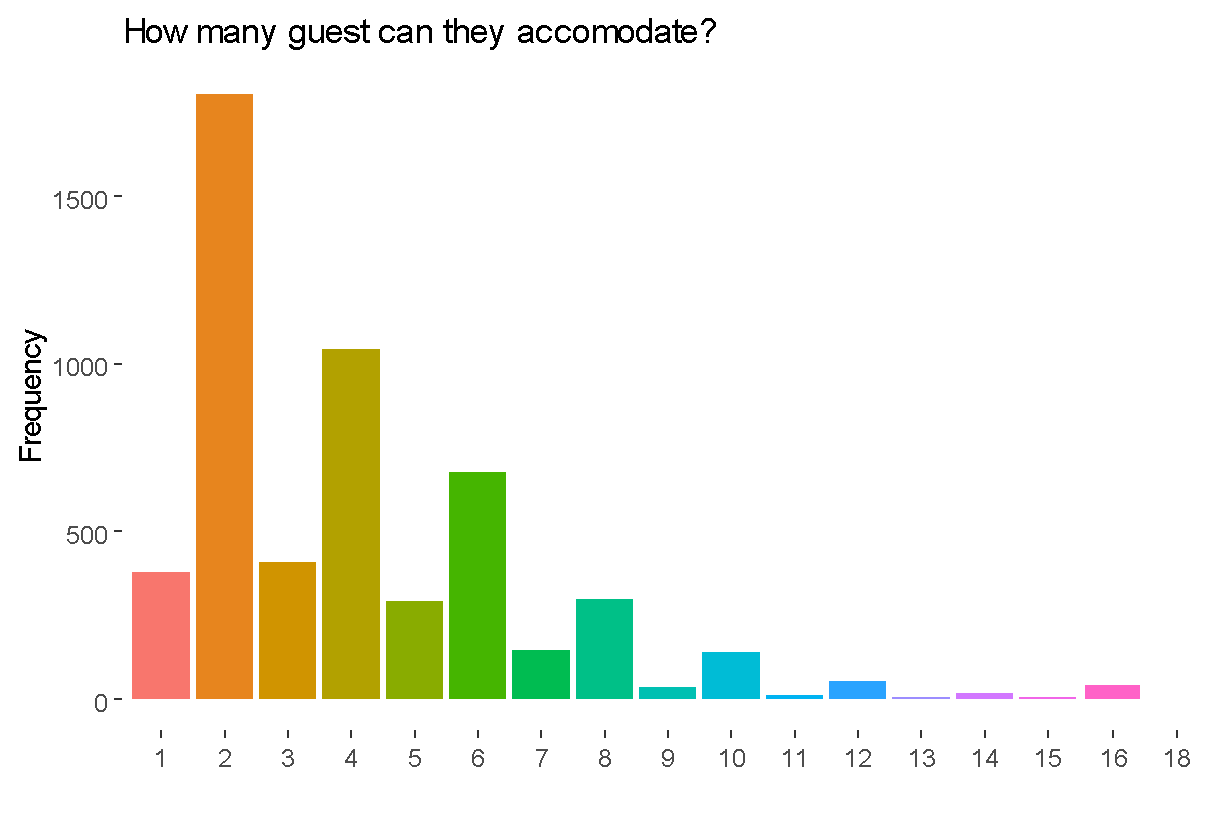
We can see that most rental types in San Diego are house and apartment.

* Proportion of room type



Entire home/apt is the most popular room type for people to rent.

* Accommodation



Most rentals can accommodate 2 or 4 people.

**Data Analysis Method - Association Rule: Relationship between price and other attributes**

For this part, our group uses the association rule to find patterns between the list price and other attributes. After adjusting the parameters and removing redundant rules, we get 680 association rules for us to find insights into the Airbnb list price in San Diego.

If the customer want to rent with a low price ($0 to $181), he or she need to focus on rentals which are:

* The host has no more than 5 listings on Airbnb.
* The property type is a house with the condition that accommodates or guests included are no more than 3.
* Only 1 bed, security fee is $95 or $100 and low cleaning fee such as $20.
* Only one bedroom or one bathroom for use.

For the high list price which is more than $181, the customer should pay attention to these conditions:

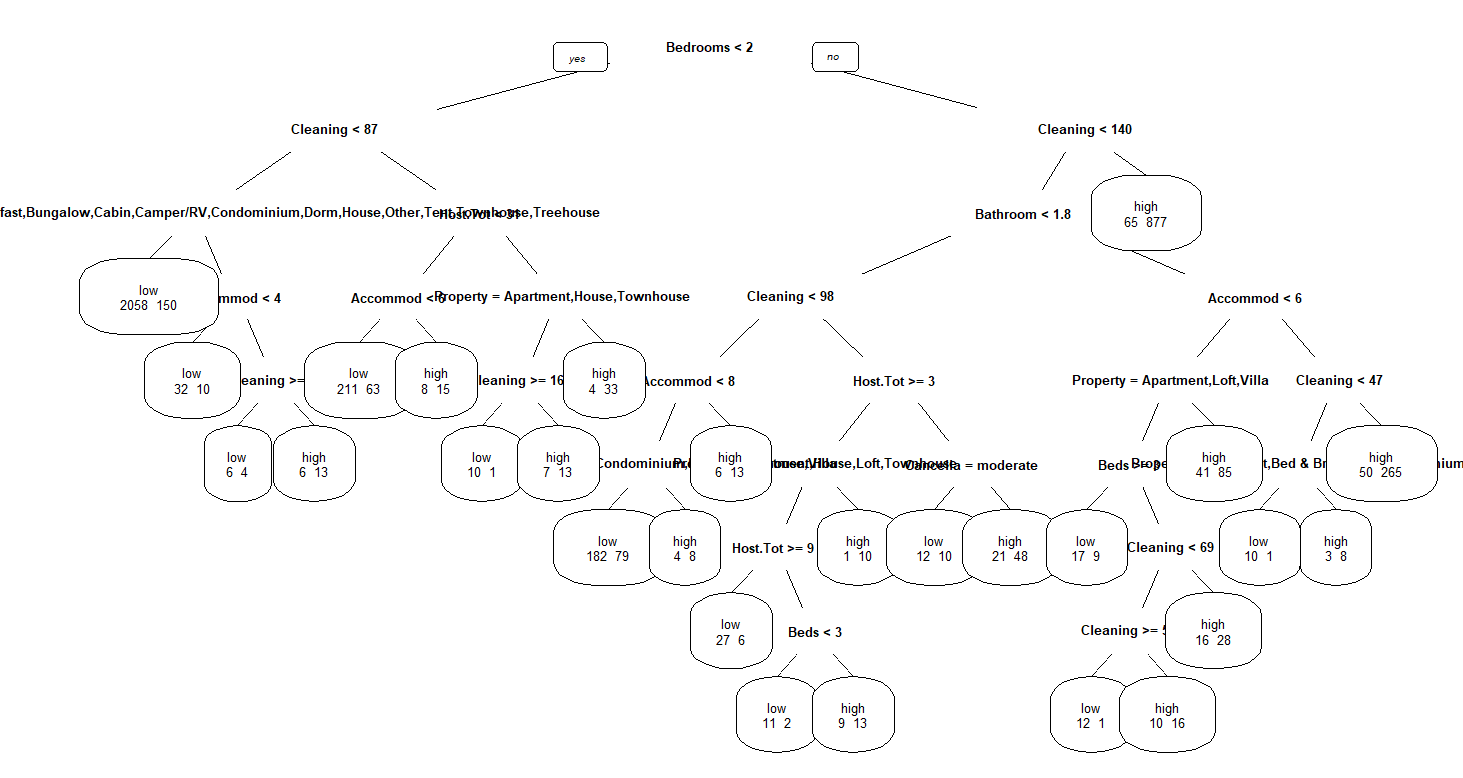
* The rental is located near the Mission Bay.
* Bedrooms and bathrooms which are both more than 3 are provided for use.
* The property type is a house with the condition that more accommodates to rent and more beds provided.
* Very high security fee and cleaning fee.
* Cancellation Policy is strict.

From the association rule, we may conclude that the list price will be considered with the neighbourhood location, security or cleaning fee, rental details such as type and the number of bedrooms or bathrooms for use. The reason for this conclusion is that these elements increase or decrease the management and maintenance costs of each listing in San Diego to some extent.

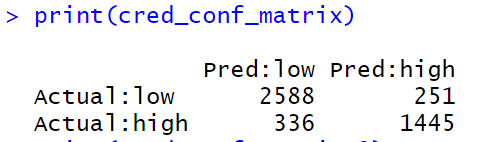
**Data Analysis Method - Decision Tree: Price Prediction**

For making the Decision tree, we removed the attribute “Neighbourhood.Cleansed” which is not a good one for our model.

In the analysis by decision tree model, we mainly use the ‘Rpart’ method. We set the seed at ‘7’, set the cp at 0.001, minsplit at 10, minbucket at 10. The first tree we got is:



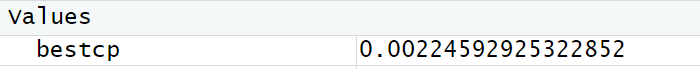
In the Decision tree we get, the model is too overfitting for prediction. So we need to optimize the model.

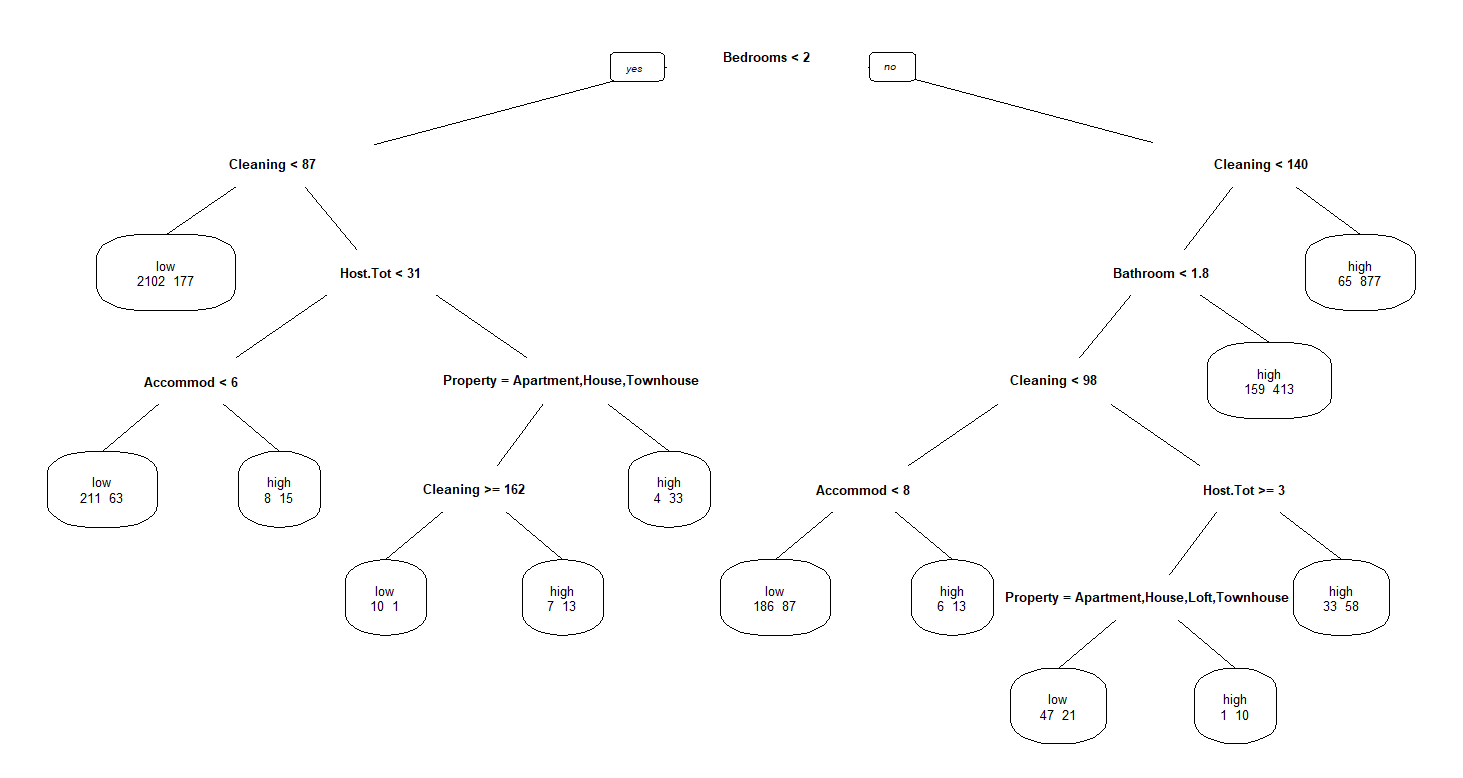


For this decision tree, the Confusion Matrix shows the accuracy is:

(2588+1445)/4620 = 87.29%

Then we pruned the decision tree by best cp to see the difference.





First, if the Bedroom is less than 2, we go left side:

If the cleaning fee is less than 87 dollars, it leads to a low price.

If the cleaning fee is more than or equal to 87 dollars:

If the host own less than 31 houses:

If the accommodate is less than 6, it leads to a low price:

If the accommodate is more than or equal to 6, it leads to a high price

If the host own more than or equal to 31 houses:

If the property type is Apartment/House/Townhouse, it leads to a high price.

If the property type is not Apartment/House/Townhouse:

If the cleaning fee is more than or equal to 162, it leads to a low price;

If the cleaning fee is less than 162, it leads to a high price.

If the bedroom is more than or equal to 2, we go right side:

If the cleaning fee is more than or equal to 140, it leads to a high price;

If the cleaning fee is less than 140:

If the bathroom is more than or equal to 1.8, it leads to a high price;

If the bathroom is less than 1.8:

If the cleaning fee is less than 98 dollars:

If the accommodate is less than 8, it leads to a low price;

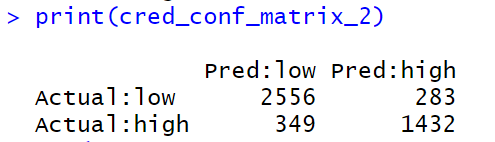
If the accommodate is more than or equal to 8, it leads to a high price.

If the cleaning fee is more than or equal to 98 dollars:

If the host owns less than 3 houses, it leads to a high price.

If the host owns more than or equal to 3 houses: If the property type is Apartment/House/Loft/Townhouse, it leads to a low price.

If the property type is not Apartment/ House/ Loft/ Townhouse, it leads to a high price.



For this pruned decision tree, the Confusion Matrix shows the accuracy is:

(2556+1432)/4620 = 86.32%

The simplification of the pruning slightly reduces the accuracy of the model.

**Conclusion**

For the project, our group analysed listings of Airbnb in San Diego from both the customer and the host perspectives with classification methods. From outputs of two methods, we found that features we selected such as neighbourhood locations, security and cleaning fees, and the number of accommodations could explain the pattern of setting list prices in San Diego. Both the association rule and the decision tree displayed similar rules for setting the price. But the decision tree provided a clear conclusion with more details and reasons and received a high accuracy output. Although our group only used two methods to achieve classification problems for the price analysis, we still comprehend how algorithms we learnt from the class could be used for solving real-world cases.