Econ 7880 Group Project Airbnb

Background introduction

Airbnb:

- **Founded Year:** 2007
- **HeadQuarter**: San Francisco, California
- **Business Lines:**

  Online Marketplace for lodging (for vacation rentals and tourism)

- **Online Booking Platforms: website** and **mobile app**
- **Business Sizes:** 4 million Hosts and over 1 billion guest arrivals

CRISP Models

6 Steps:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

Business Understanding 1

Aim:

- Boost **Revenue** and **Total Margins** during Pandemic Period
- **Retain** existing customers and **acquire** new customers
- Find the **Popular Room Type and Neighbors** to boost revenue
- Find the **place(Longtitute and Latitute)** which can increase price bargaining power of owners
- Provide **home-feeled personalized and customized service** to individual and business travellers

Business Understanding 2

How:

- Use intuitions to find key predicted variables
- Perform Data Proprcessing(ect Removing Missing Values and Normalizing Variables)
- Split the datasets into train and test data
- Construct the linear regression models
- Evaluate the accuracy of the train data and whether it is satisfied with our business goals
- Provie deployment recommendations and state the interesting findings about the models

Data Understanding

Aim:

seek a better understanding of the data and make price predictions.

Original data：

Dataset：48,895 observations with 16 columns (categorical and numeric)

Target Variable: Price

Data Understanding

Neighborhood group

latitude

longitude

Room Type

Minimum number of nights

number of reviews

Comments every month

calculated host listings count

Availability_365

Business Understanding

Difficulties:

- Averaging the listed prices of similar place to set our price. But the market is dynamic and we would want to update the price frequently.
- May miss the competitive advantages

Necessary to find out the main indicators that affect the listing price

Data Preparation

Although we have got the data， we noticed some problems:

- many missing values
- columns needed to be renamed
- columns are complex
- some of data are strings are difficult to predict
- haven't divided training data and testing data

Data Preparation

Rename each column, easier to understand

Create a new table “MY_DF”, because column 1, 2, 3, 4, 6, 13 is unnecessary. So drop them

The first column is the order of each line so we need to remove it.
Convert the list into numerical data：

- Change Strings to 1 or 0 (Dummy Variable) for predictive analytics
- In “NEI_GROUP”, Only “Manhattan” is 1, others are 0
- In “RM_TYPE”, Only “Entire home/apt” is 1, others are 0
- Delete all missing data

Modeling

- Select Model: Linear regression

- Divide the data into training group (70%), testing group (30%);
- Using training dataset multiple regression of price on selected variables,get the parameters;
- get the predicted y_cap ;

Modeling

- **The selected of variables is successful, their effect of price is significant.**
- **Room_Type**: The relationship between price and Room_type is positive, when Room_type increase 1 unit, price will increase 108.12.
- **Neighbourhood_group**: The relationship between price and Neighbourhood_group is positive, when Neighbourhood_group increase 1 unit, price will increase 52.44.

Coefficients:

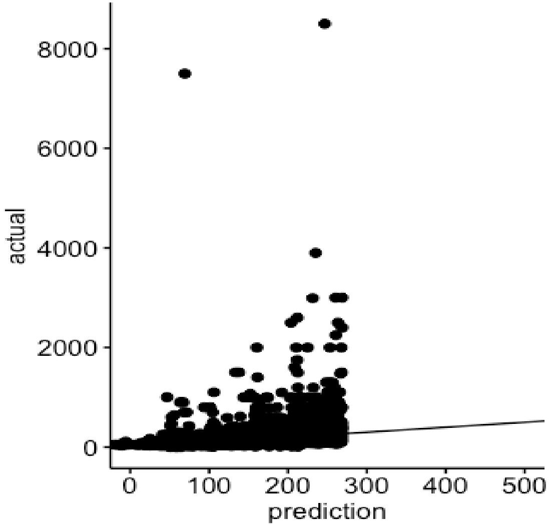
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	54.166351	2.385503	22.706	< 2e-16	***
neighbourhood_group1	52.439767	2.394822	21.897	< 2e-16	***
room_type1	108.122681	2.373094	45.562	< 2e-16	***
minimum_nights	-0.164341	0.068105	-2.413	0.0158	*
number_of_reviews	-0.162313	0.029054	-5.587	2.34e-08	***
reviews_per_month	-1.659973	0.837881	-1.981	0.0476	*
availability_365	0.149155	0.009379	15.904	< 2e-16	***

Evaluation

Around 10.2% of the variance in Price has been explained by these 6 variables.

```
Residual standard error: 193.3 on 27183 degrees of freedom
Multiple R-squared:  0.102,	Adjusted R-squared:  0.1018
F-statistic: 514.4 on 6 and 27183 DF,  p-value: < 2.2e-16
```

There are some outliers affect Model performance.



Evaluation

Model Accuracy: mean squared error (training MSE vs. test MSE)

```
> MSE_train <- mean(result$residuals^2)
> MSE_train
[1] 37336.01
> MSE_test = mean((data$actual - data$pred)^2)
> MSE_test
[1] 27980.41
>
```

MSE of testing data is lower than training data ,so the model is fit better;

Findings

1.Room type indeed postively correlated with the house listing price,and it is high coefficients with 101.3 and it also has low p value:2.2*10^(-17)

2.Neighboorhoods indeed postively correlated with the house listing price, and its coefficients is 44.48 and low p value2.2*10^(-17).

3. Longtitute and Latitute negatively correlated ,the coefficient of longitude is -304.5,the coefficient of latitude is -126.1.

So it means lower latitute and longtitute and may be better

Findings

Postive Variable	Coefficients	Findings
RM_TYPE	88.85	Highly Postively Related.
NEI_GROUP	35.07	
VIEW_PER_MONTH	1.20	Not highly related and ignore this data.
LIST_COUNT	0.99	
AVAIL	0.07	

when RM_TYPE change 1 unit,the listing price will change 88.85%

Findings

Negative Variable	Coefficients	Findings
LONGI	-218.00	Highly Negatively Related.
LATI	-59.00	
MIN_NIGHT	-0.34	Not highly related and ignore this data.
NUM_OF_VIEW	-0.09	

Deployment Recommendations

We noticed that the target variable is listing price, it cannot fully represent the demand side of the market, but we assume that listing price can represent demand side in some degrees:

1.Provide more **Entire home** provided

2.Provide more rooms rented in **Manhattan Region.Manhattan Region** is the most suitable location in New York.

3.Provide more rooms in **Latitude** between **40.7 and 40.8** and Lontitute closer to **74** since most of the tenants prefer to live in these regions.For a **lower latitude** region.Tenants would prefer to rent house near the beach with warm sunshine.