BIA-656A

Instructor: Prof. German Creamer

Airbnb Listings Occupancy Rate Prediction

Xinze Yu, Aiwu Song, Tze-How Lee, Wen-Tsai Yu

1. Problem Introduction

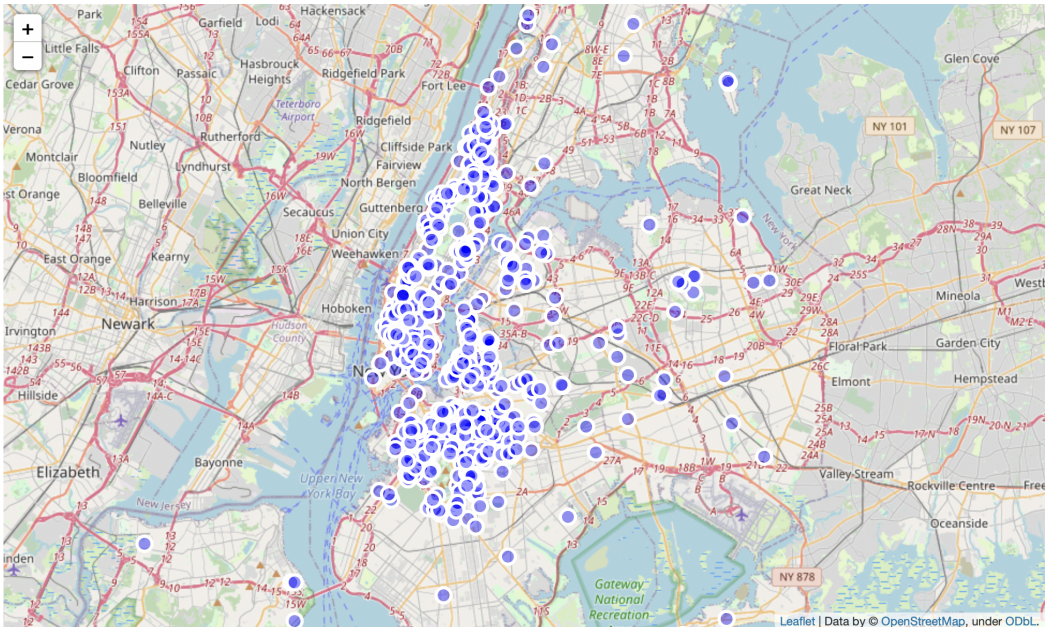


Figure 1. Airbnb listing in NYC Area (Only showing first 1000 listing)

With the development of society and the improvement of people's living standards, more and more people choose to travel for vacation. When planning a trip, vacation rentals often play an important role in travel planning. To reduce travel costs and enjoy better travel quality, people usually book their trips online in advance. Airbnb provides an online platform for people to book top-rated rentals and choose suitable accommodation, including houses, cabins, condos, unique stays, and more for their trip. However, as an Airbnb host, it's difficult to predict the occupancy rate of the listing before putting it online. Therefore, in order to help house owners rent out their properties faster and

reduce management costs, we created an Airbnb listings occupancy rate prediction model for New York City.

Project Objective

We build a model to explore the impact and correlation of housing amenities and price on occupancy rates to evaluate the occupancy rates. With the application of this model, hosts can clearly understand the competitiveness of their listing and adjust the rental price of the house in time to make more profits.

2. Data Understanding

Data source: http://insideairbnb.com/get-the-data

The original data has 38,277 rows and 74 columns, shown in Figure 2.

...	...	...	...	...	...	...	...	
38272	53662330	https://www.airbnb.com/rooms/53662330	20211204143024	2021-12-05	HUGE PRIVATE ROOM, AWESOME LOCATION	Must See! SPECTACULAR ONE FAMILY DUPLEX Home! ...	NaN	https://a
38273	53662542	https://www.airbnb.com/rooms/53662542	20211204143024	2021-12-05	LARGE PRVT ROOM, QUIET RESIDENCE PERFECTLOCATION	Must See! SPECTACULAR ONE FAMILY DUPLEX Home! ...	NaN	https://a0.mu
38274	53662772	https://www.airbnb.com/rooms/53662772	20211204143024	2021-12-05	King Room - Midtown Manhattan	You can pack a lot into a minute in this centr...	NaN	https://a0.mu
38275	53663081	https://www.airbnb.com/rooms/53663081	20211204143024	2021-12-05	King Room - Bryant Park.	NaN	NaN	https://a
38276	53665099	https://www.airbnb.com/rooms/53665099	20211204143024	2021-12-05	★Luxury in the ❤of Bklyn | Fast Wi-Fi | Sleeps...	Stunning newly remodeled apartment in Bushwick...	Bushwick! This neighborhood is loca...	https://a0.mu

38277 rows × 74 columns

Figure 2. Original Data

After preprocess the data has 10,864 rows and 72 columns.

	index	id	listing_url	host_id	host_verifications	neighbourhood_group_cleansed	latitude	longitude	property_typ
0	1	3831	https://www.airbnb.com/rooms/3831	4869	['email', 'phone', 'reviews', 'offline_governm...	Brooklyn	40.684940	-73.957650	Entire gue sui
1	3	5136	https://www.airbnb.com/rooms/5136	7378	['email', 'phone', 'reviews']	Brooklyn	40.662650	-73.994540	Entire rent ur
2	4	5178	https://www.airbnb.com/rooms/5178	8967	['email', 'phone', 'facebook', 'reviews']	Manhattan	40.764570	-73.983170	Private roo in rental ur
3	6	5803	https://www.airbnb.com/rooms/5803	9744	['email', 'phone', 'facebook', 'reviews', 'jum...	Brooklyn	40.668010	-73.987840	Private roo in townhous
4	7	6848	https://www.airbnb.com/rooms/6848	15991	['email', 'phone', 'reviews', 'offline_governm...	Brooklyn	40.709350	-73.953420	Entire rent ur
...	...	...	...	...	...	...	...	...	
10859	38021	53572597	https://www.airbnb.com/rooms/53572597	433203831	['email', 'phone']	Queens	40.766867	-73.924403	Private roo in bungalo
10860	38153	53609786	https://www.airbnb.com/rooms/53609786	49136387	['email', 'phone', 'facebook', 'reviews', 'kba']	Brooklyn	40.710918	-73.959553	Entire rent ur
10861	38164	53612732	https://www.airbnb.com/rooms/53612732	420843290	['email', 'phone']	Brooklyn	40.678150	-73.910910	Private roo in rental ur
10862	38168	53614162	https://www.airbnb.com/rooms/53614162	2368191	['email', 'phone', 'facebook', 'reviews', 'jum...	Brooklyn	40.660757	-73.929937	Entire rent ur
10863	38247	53648074	https://www.airbnb.com/rooms/53648074	433511308	['phone']	Queens	40.591505	-73.795033	Private roo in rental ur

10864 rows × 72 columns

Figure 3. Data After Preprocessing

The dependent variable:

(Y) Availability 30, 60, 90, 365: The availability of the Airbnb property in 30, 60, 90 or 365 days.

The independent variable:

From reviews and price to the house amenities, in total we have 47 independent variables before feature selection.As shown in the following Figure 4.

```
['number_of_reviews',
 'review_scores_rating',
 'reviews_per_month',
 'instant_bookable',
 'calculated_host_listings_count',
 'price',
 '24-hour check-in',
 'Air conditioning',
 'Breakfast',
 'Cable TV',
 'Carbon monoxide detector',
 'Cat',
 'Dog',
 'Doorman',
 'Doorman Entry',
 'Dryer',
 'Elevator',
 'Essentials',
 'Eco friendly',
 'Fire extinguisher',
 'First aid kit',
 'Free parking on premises',
 'Paid street parking off premises',
 'Gym',
 'Hair dryer',
 'Hangers',
 'Heating',
 'Hot tub',
 'Indoor fireplace',
 'Internet',
 'Iron',
 'Keypad',
 'Kitchen',
 'Laptop friendly workspace',
 'Lock on bedroom door',
 'Lockbox',
 'Other pet(s)',
 'Paid parking off premises',
 'Pool',
 'Private entrance',
 'Shampoo',
 'Smart lock',
 'Smoke alarm',
 'Smoking allowed',
 'TV',
 'Washer',
 'Washer / Dryer']
```

Figure 4. Independent Variables

3. Data Preparation

Data Cleaning

In the first step of cleaning data, we remove unnecessary columns such as "host_name“, "host_about“, "neighborhood_overview" and so on, for a total of 22. Next, we remove rows with missing data which have none of them. Lastly, we remove vacant & low occupancy property and only keep those recent and frequently active property, where the conditions are set to ‘availability_365 > 60’ and ‘last_review>'2021-05-21''.

Data Converting

After cleaning data, we convert the data twice. First is binary some variables to 0 or 1 such as amenities, ‘host_is_superhost’ and so on. At this part, the original amenity column lists all the amenities, so we apply re the split the text and create new dummy variables for each amenity like ‘24-hour check-in', ‘Air conditioning’, ‘Breakfast’ etc. Second is convert the number of available days to the ratio. We divide all the availability days by days.

Data Partitioning

The final step in data preparation is data partitioning. We classify the ‘availability’ ratio into two levels: Normal and Good, for the subsequent machine learning. Finally, we set four independent variables Y: ‘availability_30_percent’, ‘availability_60 _percent’, ‘availability_90 _percent’, and ‘availability_365 _percent’; each represents occupancy ratio in a different time scale. Other dependent variables as X.

4. Modeling

In this project, since we have a definite variable Y as the basis for training and testing the model, supervised algorithm is our first choice. We can make use of what we have learned in class to achieve our goals. Many benefits of supervised learning include data clarity and simplicity of training. It also has several drawbacks, such as the inability to learn on its own and hard to handle some of the complex tasks.

We have built several supervised learning models. Among the many models, we built six models: Logistic Regression, LDA, QDA, Random Forest, Neural Network, and SVC. Then we evaluated the results of the target variables with different models.

We use four different available days as our target variables. First, we explore these variables. The different available days columns are divided by their available days, and the results are described as 33%, 50%, and 66%.

count	10864.000000
mean	0.316547
std	0.293242
min	0.000000
33%	0.100000
50%	0.233333
66%	0.400000
max	1.000000

(a) Availability 30-day percentage

count	10864.000000
mean	0.532048
std	0.279471
min	0.000000
33%	0.450000
50%	0.566667
66%	0.666667
max	1.000000

(b) Availability 60-day percentage

count	10864.000000
mean	0.631846
std	0.269114
min	0.000000
33%	0.611111
50%	0.700000
66%	0.766667
max	1.000000

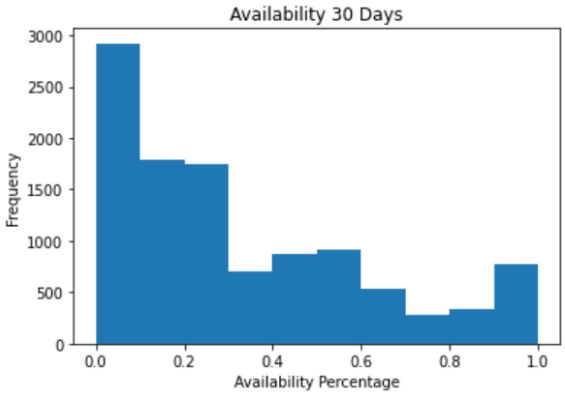
(c) Availability 90-day percentage

count	10864.000000
mean	0.636058
std	0.290087
min	0.167123
33%	0.435616
50%	0.679452
66%	0.887671
max	1.000000

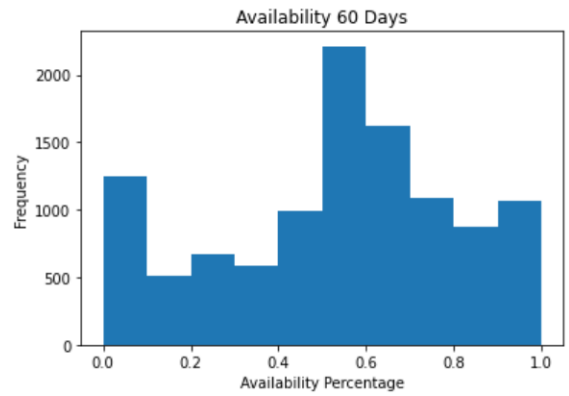
(d) Availability 365-day percentage

Figure 5. The percentage of each of the four Y variables

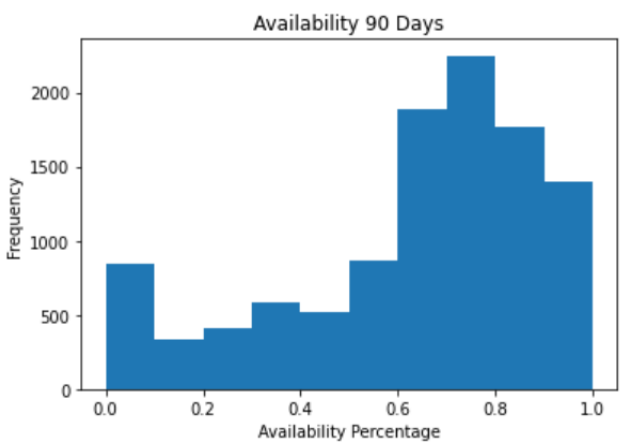
As shown in Figure 5., there are different percentages for different available days, where the available days are 30 with an average minimum of 0.3165, which shows that its data are not well distributed and primarily concentrated in one area. However, it is hard to see the difference between them through the numerical description only.



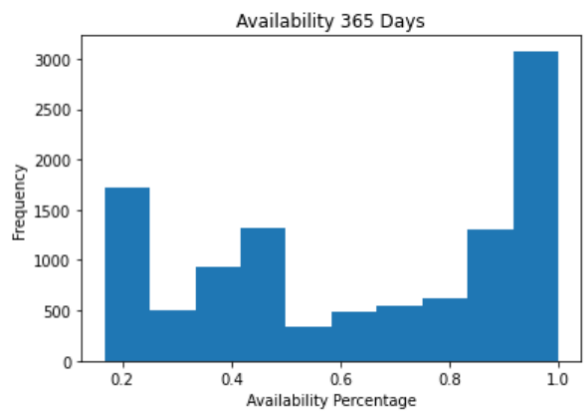
(a) Available 30-day percentage



(b) Available 60-day percentage



(c) Available 90-day percentage



(d) Available 365-day percentage

Figure 6. The graph of each of the four Y variables percentage

The graphs allow us to clearly see the trend of the variables, shown in Figure 6. It can be observed that the chart with 90 days available has the best trend compared to the other

variables. The number of statistics increases as the percentage of data becomes larger. Therefore, we decided to use the available days of 90 as the target variable.

In addition, we can also see in Figure 6 (c) that the data are mostly distributed around 80%. Therefore, we use 80% as the cut-off point, and the data below 80% are labeled as 'Low,' and the data above 80% are labeled as 'High.' Later, we divide the data into a training set and test set and split them into 7:3.

During the modeling process, it was found that the large number of dependent variables X and the small influence of some variables affected the accuracy knot of the overall model. We performed feature filtering on some models and selected the top 10 important features. After that, we compared the results of the same model technique with and without feature filtering.

Finally, after many drafts and models, we will choose the most appropriate modeling technique for this project. This will make it easier for the host to know how the room occupancy rate they want to rent is performing. The host can decide whether to change the price based on whether the final prediction is low or high.

5. Evaluation

Logistic Regression

When we use logistic regression for modeling, we get its confusion matrix and calculate its accuracy.

The confusion matrix is [533 812] [334 1581], and the accuracy is 64.84%. where the accuracy calculus is (533+1581)/3260. However, the result is not satisfactory, and the accuracy is quite low. Therefore, we further analysis on the logistic regression and

filtered out 10 important features as ['24-hour check-in', 'Carbon monoxide detector', 'Dog', 'Doorman', 'Doorman Entry', 'Eco friendly', 'Paid Street parking off premises', 'Internet', 'Laptop friendly workspace', 'Other pet(s)']. Then we re-modeled these 10 features, and its confusion matrix is 7 26 860 2367 with 72.82% accuracy, where the accuracy calculus is (7+2367)/3260. It is observed that the accuracy has improved by 7.98%.

Linear discriminant analysis (LDA)

We perform LDA modeling on the available days is 90, and then we get its confusion matrix and evaluation classification index as shown in the following Table 1.

```
[[ 159  115]
 [ 708 2278]]
              precision    recall  f1-score   support

        High      0.580     0.183     0.279       867
         Low      0.763     0.952     0.847      2393

    accuracy                          0.748      3260
   macro avg      0.672     0.568     0.563      3260
weighted avg      0.714     0.748     0.696      3260
```

Table 1. Confusion matrix and classification report of LDA

According to Table 1, we can get the precision, recall and f1-score where the accuracy is 74.8%. The LDA model evaluation result is above 70%, so we will not filter the features for it.

Quantitative Descriptive Analysis (QDA)

We perform LDA modeling on the available days 90, and then we get its confusion matrix and evaluation classification index.

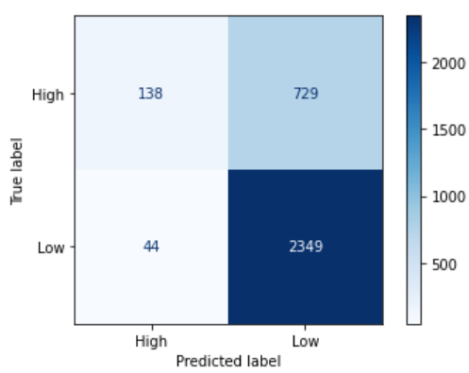
```
[[  47   54]
 [ 820 2339]]
              precision    recall  f1-score   support

        High      0.465     0.054     0.097       867
         Low      0.740     0.977     0.843      2393

    accuracy                          0.732      3260
   macro avg      0.603     0.516     0.470      3260
weighted avg      0.667     0.732     0.644      3260
```

Table 2. Confusion matrix and classification report of QDA

According to Table 2, we can get the precision, recall and f1-score where the accuracy is 73.2%. The QDA model also evaluates above 70%, thus we do not filter features for it.

Random Forest

Next, we model the random forest algorithm in the model using estimators = 100. Then we get its confusion matrix and evaluate the classification index.




```
              precision    recall  f1-score   support

        High      0.758     0.159     0.263       867
         Low      0.763     0.982     0.859      2393

    accuracy                          0.763      3260
   macro avg      0.761     0.570     0.561      3260
weighted avg      0.762     0.763     0.700      3260
```

Figure 7. Confusion matrix and classification report of random forest

As Figure 7 shows, the confusion matrix reveals that the data are mostly clustered in "Low," and the predictions are mostly correct. In addition, the classification report has an accuracy of 76.3%.

When we do feature importance on this model, such as Table x, we find that some features do not affect Y at all. Next, we select the top 10 important features from these

features. The calculated results are ['number_of_reviews', 'review_scores_rating', 'calculated_host_listings_count', 'reviews_per_month', 'price', 'Fire extinguisher', 'First aid kit', 'Hair dryer', 'Iron', 'Kitchen']. Compared with logistic regression, the results are completely different because of the different algorithms. After that, we make another random forest model for these 10 features, and the result is shown in Table 3.

```
[[ 156   87]
 [ 711 2306]]
              precision    recall  f1-score   support

        High      0.642     0.180     0.281       867
         Low      0.764     0.964     0.852      2393

    accuracy                          0.755      3260
   macro avg      0.703     0.572     0.567      3260
weighted avg      0.732     0.755     0.701      3260
```

Table 3. Confusion matrix and random forest classification report after feature selection

In Table 3, the confusion matrix indicates that the data are also mostly concentrated in "low", and the predictions are mostly correct. However, the accuracy decreases to 75.5%, which indicates that there are more than 10 features with significant impact.

Neural Network

We use neural network modeling where the hidden layer is set as (100, 15, 5). The accuracy of the algorithm is 71.6%.

Support Vector Classification (SVC)

```
              precision    recall  f1-score   support

        High       0.67      0.01      0.02       867
         Low       0.74      1.00      0.85      2393

    accuracy                           0.74      3260
   macro avg       0.70      0.50      0.43      3260
weighted avg       0.72      0.74      0.63      3260
```

Figure 8. Classification report of SVC

After we used SVC modeling and obtained an accuracy of 74% from the classification report as Figure 8.

By trying different variables, models, and parameter adjustments, the most suitable model for this project is a random forest without feature selection, with an accuracy of 76.3%.

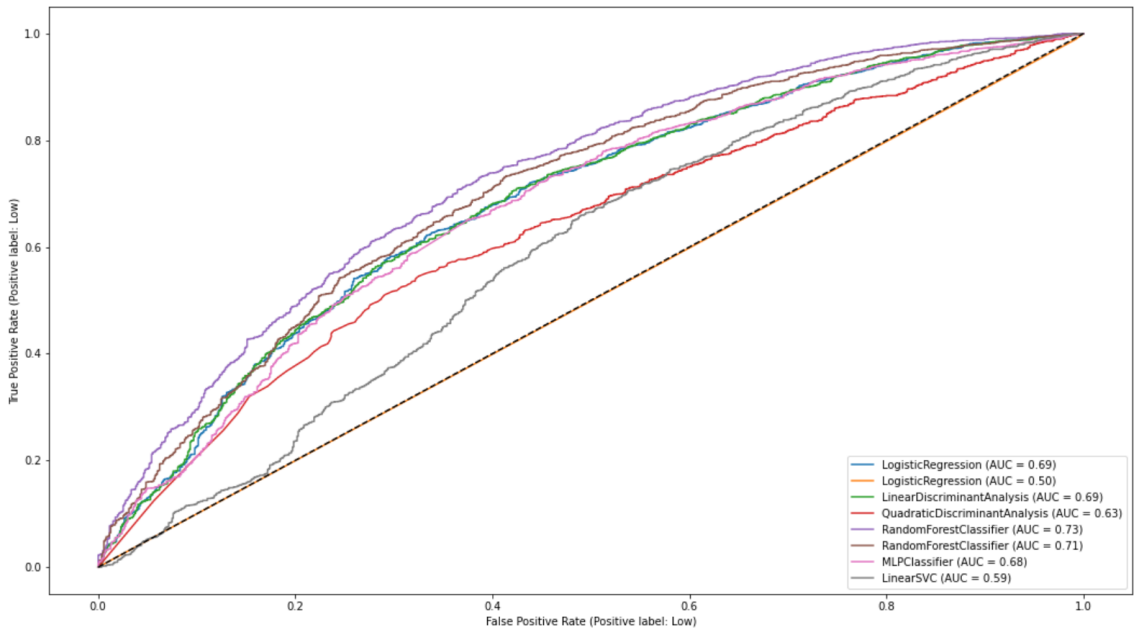


Figure 9. AUC of different models

Finally, we compare all the models as the data are mostly gathered in the "Low" area, therefore we do the ROC curve for the correct prediction rate of "Low". As shown in Figure 9, the logistic regression without feature selection has no discriminative power, and the AUC of other models mostly falls around 0.6~0.7, especially the random forest without feature selection performs the best.

We believe that the introduction of more data sources can further improve the accuracy, which can be combined with public information from local governments, weather data, and even design models to evaluate the quality of photos uploaded by owners (people use photos for comparison and reference when booking accommodation). We can also

conduct more in-depth adjustment and research on the parameters of the model to further improve the stability and accuracy.

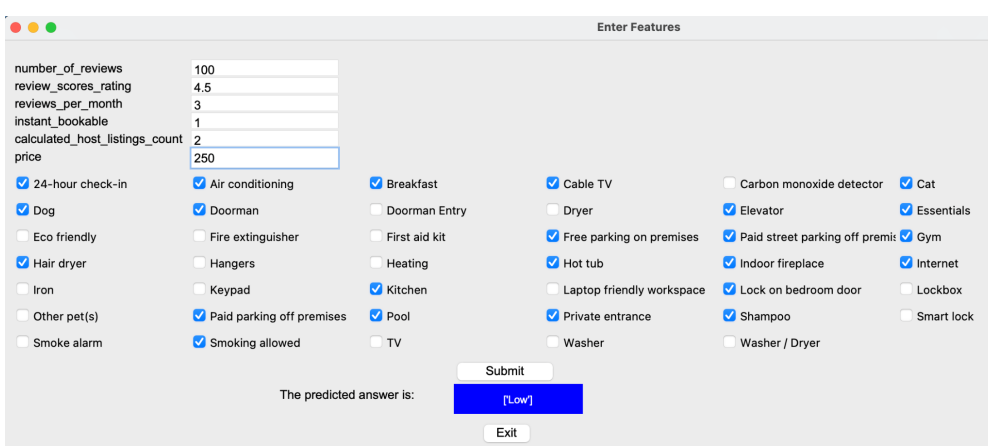
6. Deployment

We designed a UI interface that provides Host input features and then gets occupancy rate prediction results. This interface converts our trained model into a function that outputs actual results

UI Demo (Application included in the upload file)

UI Design:

Enter Features and Return Prediction:



Despite the fact that we are getting the model with 76% accuracy, we may not be able to meet the requirements in real-life scenarios. We believe that introducing more external data, such as crime rates and the number of attractions in the area, could further improve accuracy. From the company's point of view, it's important to focus on customer feedback when this feature is launched, because the tool is designed to help improve the understanding of their own properties, not everyone can accept negative comments, so the presentation of results should be specially designed. In the case of gradually improved accuracy in the early stage, it is also necessary to reach a consensus on communication with customers.

7. References

[1] Inside Airbnb Adding data to the debates, New York City.

http://insideairbnb.com/get-the-data/, May 2022

Appendices:

Project contribution:

Name	% contribution to final solution	Work
Xinze Yu	25%	Data Preparation, Modeling, UI Design, Project Paper
Aiwu Song	25%	Deployment, Poster, Project Paper
Tze-How Lee	25%	Modeling, Poster, Project Paper
Wen-Tsai Yu	25%	Evaluation, Poster, Project Paper