Internet of Things Management
(ISTM6217)

Group 6

Adel Hassen
Hikma Awol

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

Table of Contents

Executive Summary

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in that locale ("Airbnb: Advantages And Disadvantages"). Travelers usually prefer to book an Airbnb because they find it much cheaper than a hotel. A study done by Priceonomics in 2013 found that it was 21 percent cheaper to rent out a whole apartment on Airbnb than get a hotel room, and 49 percent cheaper to rent out a private room. Airbnb currently covers more than 100,000 cities and 220 countries worldwide ("Airbnb: Advantages And Disadvantages").

In this report, our team will discuss the data, methodologies we have chosen to work with and our findings from our Airbnb data analysis. We will be using the various methodologies we learned in class to analyze reviews provided by customers to recommend an Airbnb to other customers. Customers will be able to get a recommendation based on their keyword search. For example, if a customer prioritizes cleanliness, they'll be able to get a suggestion based on a comment made by another user about how clean the place was. This is a content-based recommendation.

Data

Data Description

We obtained our dataset from Kaggle. This is a very large dataset that contains 1, 214,658 rows and 32 columns. It is roughly 1GB in size.

Data Cleaning

We had to clean the data before proceeding with the analysis. We followed the following steps:

- Read in New York and LA listing and Airbnb reviews).
- Remove columns we are not using for this project (by doing so we also were able to reduce the size of our data).
- Inspect the size of each dataframe
- Change column name to "listing_id" (makes merging with airbnb_reviews easier).
- Merge New York listing with reviews
- Merge LA listings with reviews
- Put both DataFrames together using append

- Drop null values

Exploratory Data Analysis

Before we implement advanced data analysis techniques, we explored our data through visualizations and tables. Our exploratory data analysis section is divided into two sections: analysis of the price of a listing and analysis of whether a host is a Superhost.

First, let's take a look at the price of the listings. We looked to see whether we would see significant disparities between different levels of some of the variables in our data. Location is a feature that appears to affect the price of a listing; we see a difference between Airbnb's in New York City vs Los Angeles. We see that the average price of a listing in Los Angeles is higher than that of one in New York. What cities are driving this difference? In our data, we have only Los Angeles as a city in California, and for New York, we have the cities of Manhattan, Brooklyn, and the Bronx. Manhattan has the highest average price at $176.82, but average listing prices in Brooklyn and the Bronx are $133.45 and $73.25 respectively. The average price of an Airbnb in Los Angeles is $130.02.

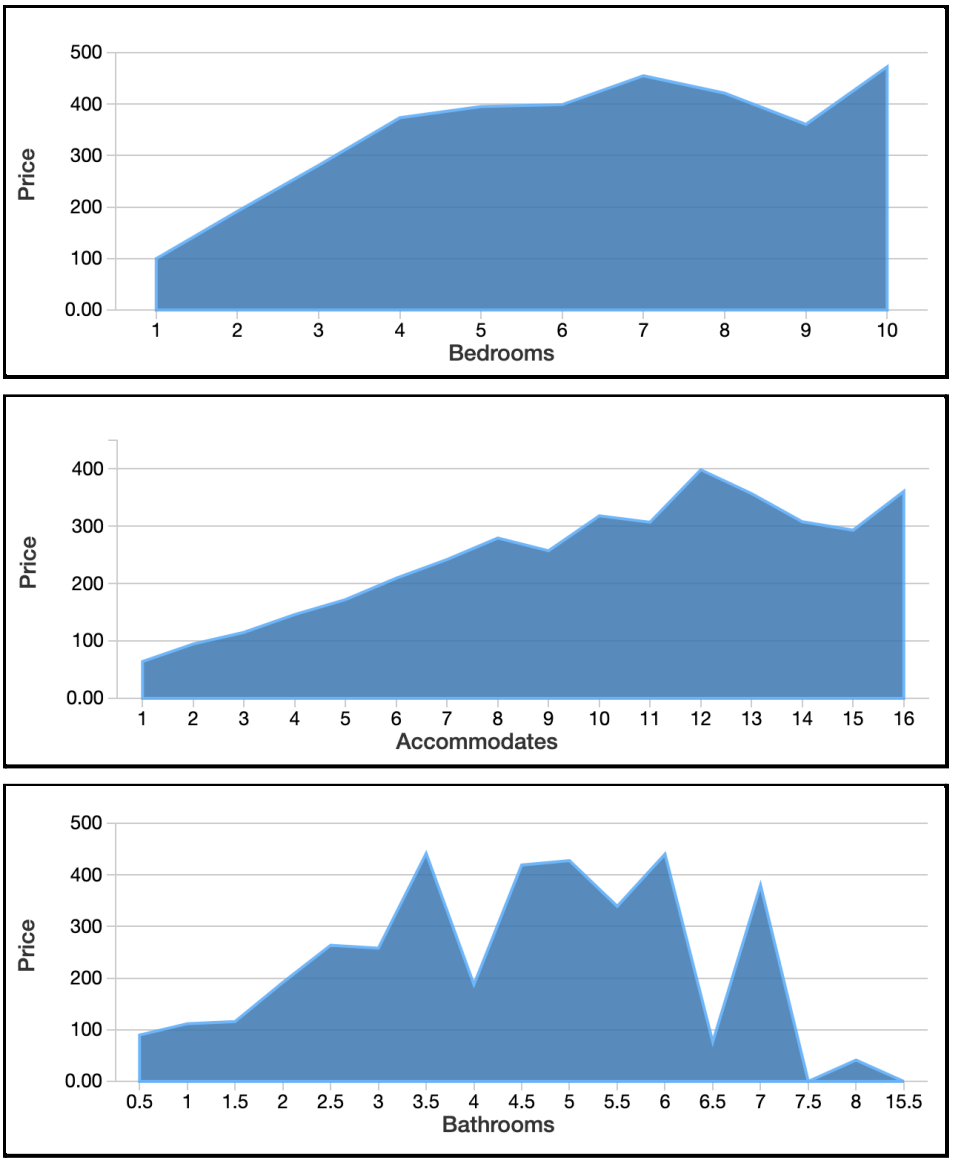
City	State	Price
Manhattan	NY	176.82
Los Angeles	CA	130.02
Brooklyn	NY	113.45
Bronx	NY	73.25

We also took a deeper look at the price distribution by neighborhood; we found that most neighborhoods with the highest average prices are located in California. The average price of Airbnb's in Vermont-Slauson, California is an incredible $568.24. New York's most expensive neighborhood for Airbnbs is Mill Basin with listings averaging $500. We also studied price distribution based on room type. The results of this table are quite straightforward with entire homes/apartments costing more than double a private room. Shared rooms are the cheapest option with an average price of $41.17.

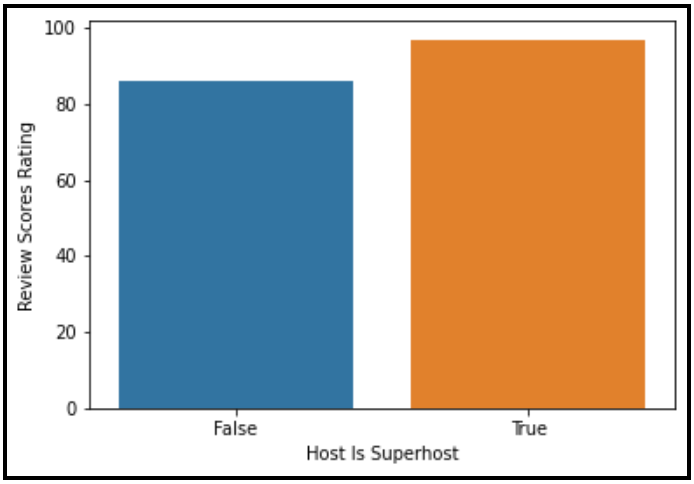
Neighborhood	State	Price
Vermont-Slauson	CA	$ 568.24
Mill Basin	NY	$ 500.00
Unincorporated Santa Susana Mountains	CA	$ 466.00
Avalon	CA	$ 368.87
Flatiron District	NY	$ 346.05
Bel-Air	CA	$ 342.08
Fieldston	NY	$ 339.13
Green Valley	CA	$ 316.25
Mission Hills	CA	$ 295.00
Beverly Crest	CA	$ 276.03

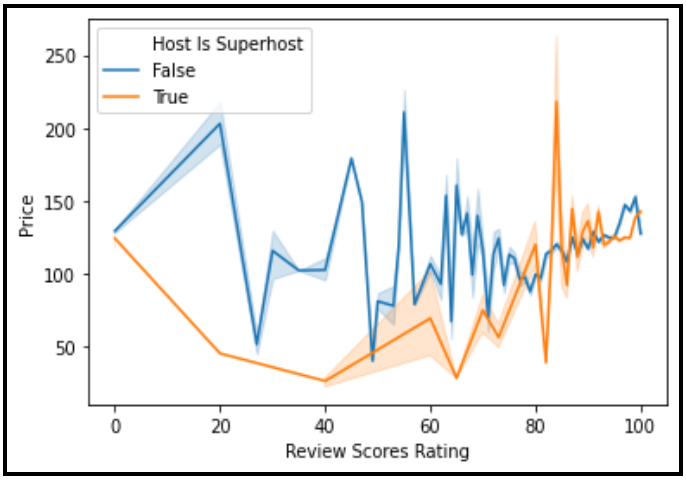
	Room Type	Price
1	Entire home/apt	166.27
2	Private room	74.96
3	Shared room	41.17

Another interesting relationship we looked at was the price fluctuation over different levels of a listing's features. We see price gradually increasing as the number of bedrooms increases and price gradually increases as the number of guests accommodated increases. Price increases with more bathrooms, but this relationship fluctuates. In particular, there is a large drop from 3.5 bathrooms to 4 bathrooms.



The final part of our preliminary research included figures that demonstrated the effect of different features on whether a host is a Superhost. The bar chart shows a clear disparity between the average rating of a Superhost and a non-Superhost. For a host that is Superhost, the average rating is 96.87 compared to 86.1 for a host that is not a Superhost. The comparison of hosts when looking at price is interesting. For ratings with low scores, non-Superhosts tend to charge much higher than Superhosts. At around a rating of 80, the price of both host types follows a similar pattern.





Research Questions

For our project, we decided to work on a dataset of Airbnb reviews guests made after their stay. We also decided to analyze Amenities the host has listed on their websites. Amenities column will be used to analyze and build a model to recommend places to stay for individuals on a keyword of their interest. For example, if a person is interested in staying at a place that has a swimming pool, the keyword “swimming pool” will be used to recommend places for Airbnb users. It is important to set the goals for the project before we move on to the analysis section of the project. The goals and objectives of our project include:

- User linear regression to predict price
- Use a logistic regression model to predict whether the host is a super host or not.
- Build deep learning models to recommend places to an Airbnb user.

Methodologies and Results/Findings

Methodology 1

The first method we used is linear regression. We used this technique to predict the price of a certain Airbnb listing using selected features. The features included are Bedrooms, Accomodates, Bathrooms, Host is Superhost, City, State, and Room Type. Linear regression attempts to model the linear relationship between a dependent variable and one or more

dependent variables. We attempt to fit the equation that minimizes the difference between our actual values and our predicted values. We used the Least Squares Method which fits a model by minimizing the sum of squares of the residuals. To access our model's performance, we will calculate the R-squared value, a measure that quantifies the amount of variation of our dependent variable explained by our model.

The second linear regression model we fit included one additional independent variable, Amenities. The Amenities column is a text field listing out all the features that are included in a listed Airbnb. Since we can not directly incorporate text into our model, we created a data pipeline that tokenized text, removed stopwords, counted term frequency, and returned tf-idf. Tf-idf is short for Term Frequency Inverse Document Frequency. Term Frequency provides a measure of the importance of a term within a document using the number of occurrences. Inverse Document Frequency is calculated by taking the log of the total number of documents divided by the number of documents containing a specified term. Combined, tf-idf is a measure of the importance of a word within a document relative to the overall corpus. We added tf-idf of Amenities as a regressor to see if it would improve our model.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

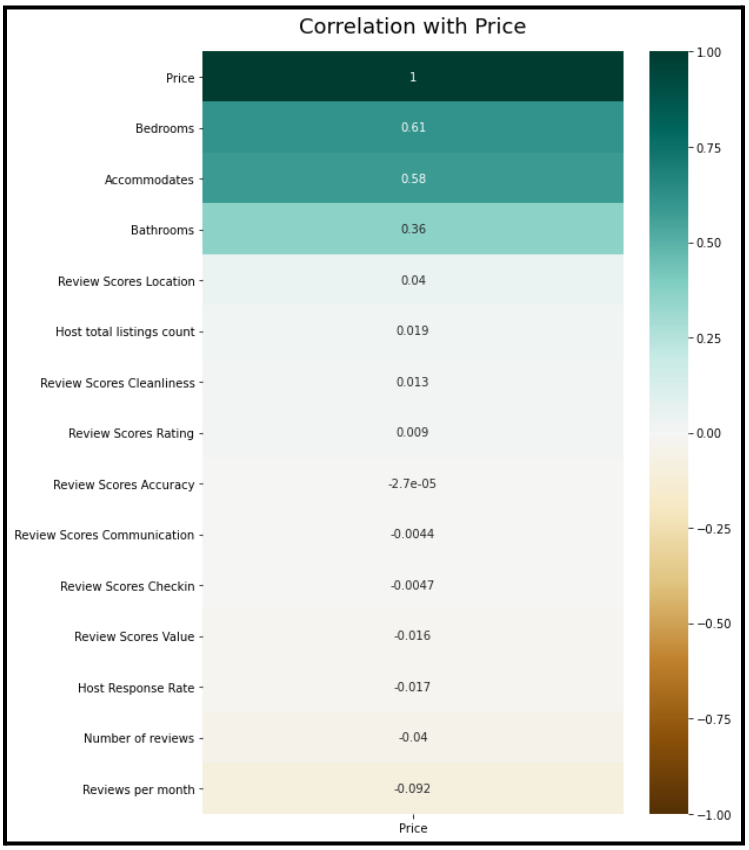
N = total number of documents

Results and Findings 1

From the Correlation plot below, we see that Bedrooms, Accommodates, and Bathrooms are correlated with Price. We also included Host is Superhost, City, State, and Room Type because there appears to be a difference in Price at different levels of those variables. This model achieved an R-squared value of 0.5041 on the test set meaning 50.41% of the variation in Price is explained by our model. This is a poor-performing model. We add Amenities represented using

tf-idf and we see a significant increase in the R-squared value. Now the R-squared value of the test set is 60.70%. The jump in R-squared shows the importance of having Amenities listed for an Airbnb, but at 60.70%, the model is still mediocre. The features included are Bedrooms, Accomodates, Bathrooms, Price, City, State, and Room Type.

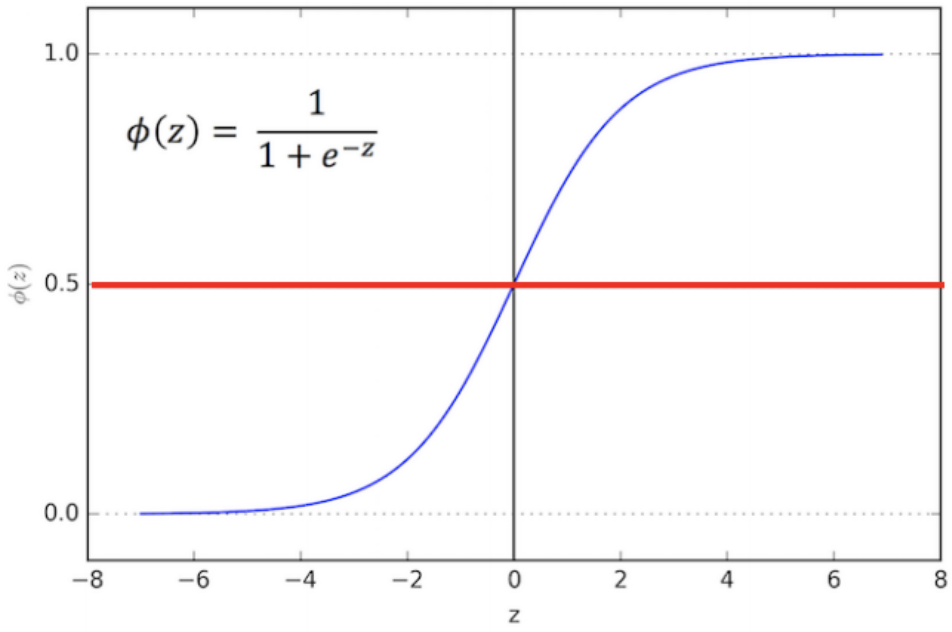
Model	Training R-square	Test R-square
No Amenities	50.50%	50.41%
With Amenities	61.42%	60.70%



Methodology 2

1Next, we used logistic regression to classify whether a host is a Superhost or not. The features included are Bedrooms, Accomodates, Bathrooms, Price, City, State, and Room Type. Logistic

regression is a linear algorithm that assumes a linear relationship between the dependent variable and the independent variables. The differences between the two popular statistical methods are that logistic regression transforms predictions using the sigmoid function and the output of logistic regression is either 0 or 1. The sigmoid function returns values between 0 and 1, and we can set a threshold value (0.5 is a good start) to determine whether a prediction will be classified as 0 or 1. This is why we use logistic regression for binary classification.



Results and Findings 2

The logistic regression without Amenities achieved an accuracy of 67.13%. This isn't necessarily a poor model, but looking at the confusion matrix, we see that it is predicting most hosts to be non-Superhosts. In the confusion matrix, non-Superhost is "positive" and Superhost is "negative". It does a poor job of predicting whether someone is a Superhost because it tends to classify a host as a non-Superhost because of the class imbalance. This will result in a very low specificity score ($\frac{True\ Negative}{True\ Negative + False\ Positive} = \frac{1{,}492}{1{,}492 + 60{,}187} = 2.4\%$). We are committing a type I error in this case. When we add Amenities using tf-idf to the model, the accuracy increases to 78.66%. We also see that the type I error is less prevalent ($\frac{True\ Negative}{True\ Negative + False\ Positive} = \frac{36{,}370}{36{,}370 + 25{,}309} = 59.97\%$).

Model	Accuracy
Logistic Regression No Amenities	67.13%
Logistic Regression With Amenities	78.66%

Logistic Regression No Amenities

	False	True
False	124,274	1,399
True	60,187	1,492

Logistic Regression With Amenities

	False	True
False	111,007	14,666
True	25,309	36,370

Methodology 3

The third methodology we used for this project is a content based recommendation for users. We used the “Amenities” column to recommend a place to stay for an airbnb user. Amenities are listed by the host. Those Amenities might be what the user is looking for. It might be a swimming pool, heater, or free parking. We will use keywords from the Amenities column to suggest stays for the user. In order to proceed with that we first need to tokenize the text. We then remove the stop words from the tokenized texts. We then do the word term frequency to see

how often this term appears in a text. The next step would be to treat TF-IDF features for each text. We also need to fit and train word 2vec. At this point we'll create a word vector for each document. After going through the above steps, we were able to provide a key word and get a business recommendation. We were also able to calculate the cosine similarity between the two vectors (they keyword and the recommended business).

Results and Findings 3

```
+------------+---------+
|  listing_id|Amenities|
+------------+---------+
|newlistingid|  heating|
+------------+---------+
```

	listing_id	similarity	Amenities
1	846799	0.68240005	Internet;Kitchen;Heating
2	17779504	0.6786232	Wireless Internet;Heating
3	245366	0.6786232	Wireless Internet;Heating
4	16661916	0.67059827	Kitchen;Heating;Essentials
5	2677677	0.6684881	Air conditioning;Heating;Family/kid friendly;Essentials
6	2107939	0.6595231	Air conditioning;Heating;Smoke detector;Essentials

The above instance shows one of the results that we got. We used the keyword "heater", and we were able to get airbnb recommendations. From the result we can also see the cosine similarity between the keyword and the businesses that are recommended.

Conclusion

In our linear regression models, we found that bedrooms, number of guests accommodate, and bathrooms have a positive correlation with the Price of an Airbnb. We also found that Airbnb's in California are more expensive on average, but the most expensive city in our data was Manhattan, not Los Angeles. In our logistic regression model, we faced the issue of class imbalance when attempting to classify Airbnb hosts as a Superhost or a non-Superhost. Adding amenities using a tf-idf representation allowed the model to drastically increase its specificity score as well as achieve an overall accuracy of 78.66%. For the last methodology we

used we were to successfully recommend using keyword recommendations. We, however, were not able to calculate the accuracy of our results.

References

"Airbnb: Advantages And Disadvantages". Investopedia, 2021, https://www.investopedia.com/articles/personal-finance/032814/pros-and-cons-using-airbnb.asp.