Data Science Capstone

Prof. Jafari

Tianyi Wang

December 6th, 2018

Airbnb at Dublin

I. Introduction

Nowadays, more and more tourists prefer to try new experiences and book their accommodation in a non-traditional way such as online with Airbnb. People would not stay with a hotel instead of that, they are choosing different places to stay in order to get more experiences with the local environment. Airbnb as a new online accommodation booking platform for tourists immerged from 2008, house owners can post their properties information on the platform, so the tourists can find proper places from those posts. Different with the hotel, those properties don't have standards or star rating by relevant agencies, so the properties' feature, basic information, descriptions by owners and the reviews by customers play the key role for travelers' decision. This online website Airbnb developed rapidly, even though tourists must place their trust in the accommodation owners regarding the quality of their stay.

In this project, the Airbnb in Dublin, Ireland will be focused. Ireland is an island country, tourism is one of the most important economic industry of Ireland, which expected will reach 6.0% of total GDP which is EUR 17.2 billion (USD 19.4 billion) in 2018. In 2017, it was 5.9% of GDP in 2017 and is forecast to rise by 4.2% in 2018, and to rise by 3.8% to EUR26.0 billion (USD29.4 billion), 7.1% of GDP in 2028. (Irish Tourism Industry Confederation). That's the reason why I want to choose Dublin as my target city for my research. In Ireland, most of the

attractions for tourists are at Dublin, especially during the St. Patrick's Day every year, March 17th, the most tourists of the year. Because of the popularity for traveling, lots of local people willing to post their personal properties on Airbnb, also the demand of Airbnb is relatively high in Dublin. In my project, I'm going to use different kinds of Machine learning tools and Natural Language Processing to discover customer motives in choosing to stay with Airbnb and price of the accommodation.

II. Description of the Dataset

• Raw Dataset

Here is the link of the dataset: http://insideairbnb.com/get-the-data.html

In this project, there have two datasets

- Dublin Airbnb Listings Dataset
- Reviews Dataset

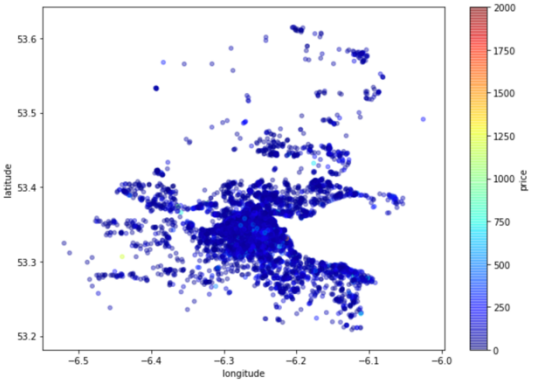

Figure. 1: Dublin Airbnb Listings

I found data on the Airbnb website. The raw dataset has 10020 rows and different variables such as properties' latitudes, longitudes, households' descriptions, prices, bedrooms number, maximum guests' number, and maximum days available for tourists. From all the texts and numbers, I can do what things are important for the tourist who wants to visit Dublin. The second dataset is the reviews dataset by different tourists' review which has 313, 559 rows. it includes username, id, comments, and date. All the different variables and data would be useful to expand my research by using data science tools to relate the content of descriptions to

additional variables relating to the different users' data. In order to get more details of the dataset, please check the code called *Dublin Airbnb Visualization For EDA*. *(Figure.1)*

III. Description of the project

There have two main parts of my project, the first part is using different machine learning tools such as linear regression, logistic regression, random forest to modeling the price and reviews rating. The second part is using Natural Language Processing based on comments from guests and description by the household. I did cluster descriptions of the room by the house owners. By using household's summary of the listing and the comments by the guest, I chose the topic first then did comments comparison in the Latent Dirichlet Allocation and Latent Semantic Indexing Model to predict similarity.

IV. Experimental Setup

1 Basic Modeling for Predict Price and Review Rating:

In order to predict the housing price, I used bathrooms, bedrooms and beds column to predict. With the review scores rating prediction, I used the variables such as review scores accuracy, communication, location, cleanliness, value and check in time. For modeling these two predictions, I used Linear and Logistic Regression, and Random Forest.

2 Natural Language Processing:

Now that we have a processed dataset and an understanding of what we are trying to predict, I focus on converting the description text into useful features for the machine learning model. In Natural Language Processing (NLP), I have to the collect description text to a corpus and it is converted into a Document-Term-Matrix. where each listing is a document containing a matrix of terms. The NLTK and Gensim packages were used to complete this. For text mining, in order to reduce the number of terms and focus on the most important things, non-English and stop words are removed. Words are also lemmatized, and a RegEx tokenizer is used to ignore non-alphanumeric strings. The remaining words are then converted into a bag of words which representation for the Document-Term-Matrix.

I. *Cluster the Room* (two code. One for big four neighborhoods, and other one is in the specific neighborhoods) After data pre-processing steps, I used TFIDF to extract top words that describe each neighborhood. Instead of describing neighborhoods, this notebook attempts a two-step process:

   (1) Depends on how the owner describe the rooms I create four and twelve clusters of

   (2) Take top 20 words that best describe each cluster. We will take a look at clustering based on 4-clusters and 12-clusters.

II. *Predict Similarity*: By using household's summary of the listing and the comments by the guest, I chose the topic first in the household's summary of the listing, then using the model to predict similarity in the comments by the guest. I use Latent Dirichlet Allocation (LDA) to discover topics inherent in the corpus, classify the corpus according to the learned topics and use them as features for the regression model. Then I randomly

pick one review in the review dataset to predict similarity. I choose to compare the Latent Dirichlet Allocation and Latent Semantic Indexing Model to predict similarity.

V. Results

1 Machine Learning by Different Models:

For predict the housing price, I used bathrooms, bedrooms and beds column to predict. Through the feature importance, bedroom (0.66), bathroom (0.17) are more important than bed (0.15). The R square value in Linear regression for modeling the price are relatively low. (train: 0.3; test: 0.1). The random forest model's R squares are really low. (train: 0.42; test: 0.3). I also tried to normalize the data; cut the range of the bedroom and bathroom or change to different variables, but the accuracy still low.

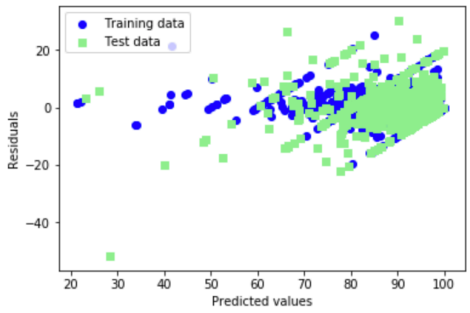

Figure.2: linear regression for the review rating

Compare with modeling the price, the review scores rating prediction is much better. For the feature importance for all these variables, review scores accuracy, location and value are the top three important for the review rating score. In linear Regressions R square value for the training dataset is 0.826 and testing dataset is 0.709. However, the random forest mean square error value is high. (train:9.7; test:17)

2 NLP:

In the data processing, I remove punctuations, emoji; Lemmatizer; Convert each word to its lower case; Tokenization; and did Stopwords removal. For the text mining, I used TfidfVectorizer (TF-IDF) which it weighs down the most frequent words in our data and gives fewer common words more value to produce better results when it is put through a classifier.

1) Cluster the room by house owner's descriptions

I did four clusters by house owner's descriptions in four big neighborhoods which are South Dublin, Fingal, Dublin City, and Dún Laoghaire-Rathdown. You can see from cluster 2, 3 has most listings in Dublin City (Figure.3) and the word clouds showing the top 300 words in the cluster one (Figure.4).

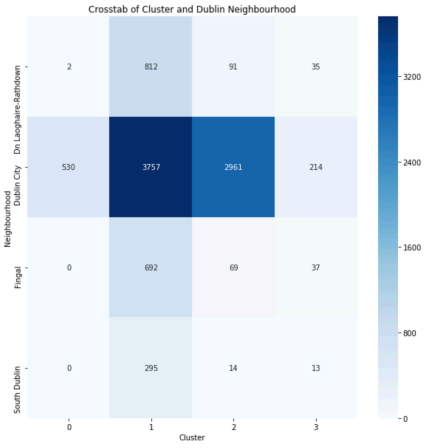

Figure. 3: Crosstab of cluster and Dublin Neighbourhood

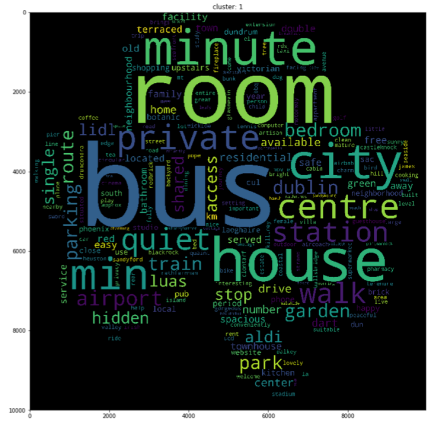

Figure. 4: Top 300 Words Cloud for Cluster 1

In the 12-cluster for the description, I chose to show the top 30 words which has best describe in different clusters. For cluster one and four which has the most number, you can see from figure. 5 St. Patrick; Grafton and Stephen; cathedral these words are most popular in this cluster. You can tell this cluster are in the city center since there has the most famous tourist

place in words from the word cloud. The top 30 words that best describe each listing cluster four (Figure. 6), you can tell it's close to the Phoenix Park area and it's most is townhouse and need transportation (Luas) in the word cloud. I also divided the neighborhood by specific area than the four big neighborhoods. In cluster 4, Docklands (620) and North City Cengtral/O'Connell street (408) has the highest. In cluster 1, Ranelagh and Rathmines (228) philbsborough (183) are having the most numbers in the listings.

Figure.5: Top 30 Words that Best Describe Cluster1

Figure.6: Top 30 Words that Best Describe Describe Cluster four

2) Predict Similarity

```
LDA model
Topic Number 1: 0.047*"min" + 0.033*"city" + 0.031*"centre" + 0.031*"bus" + 0.028*"walk" + 0.023*"park" + 0.022*"dubl
in" + 0.021*"house"
Topic Number 2: 0.028*"dublin" + 0.023*"city" + 0.020*"apartment" + 0.013*"minute" + 0.012*"perfect" + 0.012*"great"
+ 0.012*"centre" + 0.010*"room"
Topic Number 3: 0.061*"room" + 0.036*"double" + 0.036*"bed" + 0.025*"bedroom" + 0.025*"house" + 0.020*"bathroom" + 0.
017*"single" + 0.016*"two"
Topic Number 4: 0.039*"nan" + 0.020*"city" + 0.017*"room" + 0.014*"dublin" + 0.010*"walking" + 0.010*"centre" + 0.008
*"distance" + 0.008*"bed"
Topic Number 5: 0.046*"minute" + 0.039*"walk" + 0.036*"dublin" + 0.022*"city" + 0.016*"house" + 0.014*"located" + 0.0
13*"centre" + 0.013*"bus"
Topic Number 6: 0.029*"de" + 0.012*"e" + 0.011*"le" + 0.011*"n" + 0.011*"centre" + 0.010*"dublin" + 0.009*"mew" + 0.0
09*"park"
Topic Number 7: 0.029*"room" + 0.021*"dublin" + 0.020*"house" + 0.015*"home" + 0.015*"bedroom" + 0.014*"bed" + 0.014*
"city" + 0.013*"large"
Topic Number 8: 0.046*"dublin" + 0.030*"street" + 0.029*"city" + 0.027*"minute" + 0.027*"walk" + 0.026*"bar" + 0.019*
"temple" + 0.019*"apartment"
Topic Number 9: 0.032*"apartment" + 0.028*"dublin" + 0.028*"canal" + 0.027*"walk" + 0.027*"min" + 0.024*"grand" + 0.0
22*"city" + 0.016*"dock"
Topic Number 10: 0.022*"city" + 0.022*"room" + 0.019*"bus" + 0.018*"house" + 0.018*"minute" + 0.017*"min" + 0.014*"ce
ntre" + 0.014*"walk"
Topic Number 11: 0.036*"apartment" + 0.029*"city" + 0.024*"bedroom" + 0.023*"dublin" + 0.022*"minute" + 0.019*"walk"
+ 0.017*"centre" + 0.016*"room"
Topic Number 12: 0.021*"bedroom" + 0.019*"apartment" + 0.018*"dublin" + 0.017*"kitchen" + 0.017*"room" + 0.014*"city"
+ 0.013*"bathroom" + 0.011*"living"
Topic Number 13: 0.048*"place" + 0.022*"couple" + 0.020*"good" + 0.020*"business" + 0.019*"close" + 0.019*"min" + 0.0
17*"solo" + 0.016*"city"
Topic Number 14: 0.032*"centre" + 0.031*"minute" + 0.025*"dublin" + 0.022*"city" + 0.020*"bus" + 0.019*"close" + 0.01
8*"quiet" + 0.016*"airbnb"
Topic Number 15: 0.026*"street" + 0.016*"dublin" + 0.015*"apartment" + 0.012*"city" + 0.011*"spire" + 0.010*"minute"
+ 0.010*"flat" + 0.009*"bar"
```

Figure. 7: LDA Model to predict the Topic

```
--------------------------------
LSI model
Topic Number 1: 0.353*"dublin" + 0.340*"city" + 0.298*"walk" + 0.292*"minute" + 0.247*"min" + 0.229*"centre" + 0.208*
"apartment" + 0.176*"bus"
Topic Number 2: 0.757*"min" + 0.356*"walk" + -0.238*"dublin" + -0.234*"apartment" + -0.140*"bedroom" + 0.137*"bus" +
-0.124*"minute" + -0.120*"room"
Topic Number 3: -0.669*"minute" + 0.334*"min" + -0.312*"walk" + 0.276*"room" + 0.186*"bedroom" + 0.169*"apartment" +
0.165*"bed" + 0.155*"double"
Topic Number 4: 0.450*"room" + -0.448*"dublin" + 0.273*"minute" + -0.236*"apartment" + 0.228*"house" + 0.202*"double"
+ 0.169*"bed" + 0.156*"bathroom"
Topic Number 5: 0.442*"apartment" + -0.347*"city" + 0.345*"walk" + -0.320*"centre" + -0.217*"house" + -0.206*"bus" +
-0.177*"close" + -0.175*"place"
Topic Number 6: 0.625*"place" + -0.269*"dublin" + 0.229*"close" + 0.223*"couple" + 0.211*"good" + 0.200*"business" +
0.192*"solo" + 0.186*"adventurer"
Topic Number 7: -0.561*"dublin" + 0.492*"apartment" + 0.341*"city" + -0.308*"house" + 0.296*"centre" + -0.158*"room"
+ 0.122*"bus" + -0.095*"street"
Topic Number 8: -0.368*"walking" + 0.307*"dublin" + -0.304*"bar" + -0.284*"street" + 0.259*"walk" + -0.251*"distance"
+ 0.232*"place" + -0.219*"temple"
Topic Number 9: -0.377*"house" + 0.357*"room" + 0.337*"bus" + -0.308*"bedroom" + -0.299*"walk" + 0.200*"minute" + 0.1
88*"walking" + -0.180*"centre"
Topic Number 10: -0.321*"city" + -0.310*"centre" + -0.310*"room" + 0.288*"house" + 0.278*"bedroom" + 0.275*"bus" + 0.
258*"walking" + -0.194*"bed"
Topic Number 11: -0.453*"bed" + 0.443*"room" + -0.371*"bedroom" + 0.349*"house" + -0.298*"double" + 0.293*"apartment"
+ 0.128*"located" + -0.121*"minute"
Topic Number 12: 0.383*"bed" + 0.325*"house" + 0.313*"bus" + -0.309*"walking" + -0.258*"home" + -0.226*"distance" + -
0.198*"minute" + -0.170*"room"
Topic Number 13: 0.370*"restaurant" + -0.331*"house" + 0.292*"home" + -0.256*"minute" + -0.240*"city" + 0.184*"pub" +
-0.182*"min" + 0.182*"bus"
Topic Number 14: -0.497*"bed" + 0.307*"bedroom" + 0.295*"street" + -0.268*"walking" + 0.240*"bus" + -0.225*"distance"
+ -0.210*"house" + -0.182*"within"
Topic Number 15: -0.467*"home" + 0.362*"centre" + -0.232*"located" + -0.197*"city" + -0.190*"street" + -0.189*"center
" + 0.158*"dublin" + 0.151*"close"
```

Figure.8: LSI Model to predict the Topic

I use Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) to discover topics inherent in the corpus, classify the corpus according to the learned topics. The figures above are the topics from two different models which I got from the summary column in the listing dataset.

```
: # randomly pick one reviews to predict similarity.
  reviews = pd.read_csv('Dublin_reviews.csv')
  text = reviews.loc[40,'comments']
  print(text)
```

```
Very nice area, with many little cafés and restaurants. For the parisians, I would compare it to the Canal Saint-Mart
in: kind of newly fancy neighbourhood. The place is beautiful! The bed is very confortable, the conforter is very war
m, there is a closet to put all your stuff in the little cozy room. There is also a small private bathroom for the gu
ests which gives you a lot of privacy. We didn't meet Ciara, but she sounds very casual! We will definitely stay with
her again next time we come visit Dublin and we recommend her place.
It is not in the very middle of the city center, but in a walking distance (I would say 15 min from the Trinity). Kin
d of just out of the crowd :-) which was what we were looking for.
Thanks Ciara for hosting us!
```

Figure.9: Example of randomly picked the reviews

```
  if np.issubdtype(vec.dtype, np.int):
```

```
[(4284, 0.88888925), (9290, 0.85169077), (7784, 0.8401889), (5624, 0.82852936), (8408, 0.81923926), (9215, 0.8192299)
, (9295, 0.81021416), (457, 0.805737), (9059, 0.7959621), (8230, 0.7906388)]
Cosy Double Bedroom near ChristChurch Dublin City in Redbrick family owned townhouse within walking distance of City
Centre. Near theatres, venues, shops and tourist attractions including Dublinia, St Patrick's Cathedral, and Trinity
College. Prefer female non smoker to share my space.
```

Figure.10: LAD Model Result for the prediction

```
  if np.issubdtype(vec.dtype, np.int):
```

```
[(7134, 0.8743823), (4008, 0.85864604), (3508, 0.85517937), (7858, 0.8521234), (2984, 0.8519559), (2566, 0.84592015),
(7020, 0.8440836), (1914, 0.8413142), (411, 0.83868736), (3663, 0.83818674)]
Our lovely flat is perfect for 4 people. There is two double room and two bathroom. Only 15 min walking from temple b
ar, you will be staying in a quiet place. Perfect for couples, friends and family
```

Figure.11: LSI Model Result for the Prediction

Then I randomly pick one review in the review dataset to predict similarity. I choose to compare the Latent Dirichlet Allocation and Latent Semantic Indexing Model to predict similarity. Latent Dirichlet Allocation is slightly better accuracy than Latent Semantic Indexing Model. LDA and LSI both describe mathematical models that are designed to be used for

information retrieval such as returning search results. LSI examines the words used in a document and looks for their relationships with other words. LSI has one major weakness which is ambiguity LDA, on the other hand, is a significant extension of LSI. Words are grouped into topics. They can exist in more than one topic which appears all the time. (Hughes)

VI. Summary and Conclusions

In conclusion, from the data processing to build different models, this project helped me to walk through what I have learned in Data Science program. This study could help to find out why people choose or choose not to stay with Airbnb rather than in traditional accommodations, such as hotels, motel etc. According to the results, the transportations and locations for tourists are also really important. Most people are willing to stay around the city center area which is close to everything.

The other difficult part of this project is to decide which variables can be determined to use. I tried a different kind of model for the price and lots of seems doesn't turn out well because the data is not balanced since most of them are one bedroom. Airbnb also has various of different kind of types of housing, the price is made by the household and it has all the range of bedroom (1-10) so it's really hard for me to predict the exact price for the specific area. In addition, in the NLP part, the description for the household and guests' comments has lots of information.

A couple of interesting thoughts to be pursued further research. For future research, 'San Francisco Model' (Inside Airbnb) based on price average length of stay and review rate to get a new variable called yield. The concept of yield is used as a proxy for potential future earnings. Yield is defined as the amount of revenue that a property will earn over a year. They assumed the length of the stay and a relatively conservative review rate of 50% is used to convert No. of reviews into estimated bookings in order to get the value of yield. The calculation is the Average

length of stay * Price * No. of reviews/month * Review Rate. This occupancy model probably would be more reasonable than just use price. In addition, from these datasets, there have lots of text data. I used several columns in the listing dataset converting the description text into useful features for the machine learning model. However, in the review dataset, since there doesn't have a label, so I am not able to do the algorithm to determine how positive or negative a statement/text is. Since there has the Id for the host too for the comments, I could go back the listing dataset to match the comments and to do more research with the rating score to specific households' guests.

Reference:

1. Hughes, John. "Latent Dirichlet Allocation v Latent Semantic Indexing." *Indicium* , 29 Sept. 2011, www.indiciumweb.co.uk/2010/09/latent-dirichlet-allocation-v-latent-semantic-indexing/.

2. "Inside Airbnb. Adding Data to the Debate." *Inside Airbnb*, insideairbnb.com/about.html.

3. "TRAVEL & TOURISM ECONOMIC IMPACT 2018 IRELAND." *Irish Tourism Industry Confederation*, www.wttc.org/-/media/files/reports/economic-impact-research/countries-2018/ireland2018.pdf.