Data Project Proposal

December 2, 2019

Data Project Proposal
Guanzhong Chen
Data 1030
Brown ID: 140268394
GitHub Link: https://github.com/frank07080/Data-1030-Project

1 Introductory and Problem

Airbnb stands for "AirBed and Breakfast." More and more people nowadays choose to live in Airbnbs instead of hotels when they are out traveling. There are reasons why Airbnbs become more popular than hotels when it comes to where to live during traveling. One of them is that you are renting houses from a person that give a total different feeling from hotels. The other one may be that Airbnbs are, in a sense, cheaper than hotels.

Price of Airbnb, then, becomes our target variable especially in city like New York. Different from classification problem, predicting the price of Airbnb is a problem of regression because price is a continuous variable. So why is price important? The reason is that most people would to take price as their first consideration when budgets are taken into their accounts. It would be convenient if people can predict the price of an Airbnb if they are given enough information of that Airbnb.

2 Description of Dataset

The dataset is from Kaggle called "New York City Airbnb Open Data." There are a total of 48895 data points and 16 features in the dataset. The dataset is well-documented, and we can check a description for each feature in Kaggle.

There are several public projects where data has been used. One of them is called "Maps of NYC Airbnbs with Python." The data was used for finding a listing that meets specific criteria for an upcoming trip of the author. The features were used by filtering out unimportant features and finding specific requirements of an interested feature satisfying the author's criteria. For example, the author would want the host to must have more than 10 reviews.

The other one is called "Data Exploration on NYC Airbnb." The data was used for visualizing and analyzing in that project. For example, the feature host IDs was used for visualizing hosts with most listings in New York city. And other features were used for visualizing purposes too.

3 Dataset Preprocessing

There are a total of 5 features we are interested in for now and will be preprocessed.

The first feature is the neighbourhood group in New York city. I have chosen the OneHotEncoder for this categorical feature. The reason is that there is no specific order of the neighbourhood groups. The first 10 rows of this feature is shown below.

```
[[0. 1. 0. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 1. 0. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 1. 0. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 0. 1. 0. 0.]
 [0. 0. 1. 0. 0.]]
```

The next feature is room type of Airbnbs in New York city. I have chosen the OrdinalEncoder for this categorical feature. The reason is that they can be ordered by how good a room is. For example, the entire apartment/room is better than a private room. And a private room is better than a shared room. The first 10 rows of this feature is shown below.

```
[[1.]
 [0.]
 [1.]
 [0.]
 [0.]
 [0.]
 [1.]
 [1.]
 [1.]
 [0.]]
```

The next feature is number of reviews of Airbnbs in New York city. I have chosen the Standard Scaler for this continuous feature. The reason is that some of good Airbnbs will get significantly more reviews than the others. And this follows a tailed distribution. The first 10 rows of this feature is shown below.

```
[[-0.32041358]
 [ 0.48766493]
 [-0.52243321]
 [ 5.53815562]
 [-0.32041358]
 [ 1.13861706]
 [ 0.57745143]
 [ 9.12961566]
 [ 2.12626857]
 [ 3.06902683]]
```

The next feature is reviews per month of Airbnbs in New York city. This is very similar to the last feature. I have chosen the Standard Scaler for this continuous feature. The reason is the same a s the last one. The first 10 rows of this feature is shown below.

```
[[-0.6922205 ]
 [-0.59105534]
 [        nan]
 [ 1.94402463]
 [-0.75768031]
 [-0.46608661]
 [-0.57915356]
 [ 1.24777027]
 [-0.22805093]
 [-0.02572061]]
```

The next feature is availability of a year of Airbnbs in New York city. I have chosen the MinMax Scaler for this continuous feature. The reason is that there is a clear min and max of this feature. The first 10 rows of this feature is shown below.

```
[[1.        ]
 [0.97260274]
 [1.        ]
 [0.53150685]
 [0.        ]
 [0.35342466]
 [0.        ]
 [0.60273973]
 [0.        ]
 [0.51506849]]
```

Combining all five features and put them together into a pandas dataframe, we have the following for the first 10 rows of all features.

	0	1	2	3	4	5	6	7	8
0	0.0	1.0	0.0	0.0	0.0	1.0	-0.320414	-0.692221	1.000000
1	0.0	0.0	1.0	0.0	0.0	0.0	0.487665	-0.591055	0.972603
2	0.0	0.0	1.0	0.0	0.0	1.0	-0.522433	NaN	1.000000
3	0.0	1.0	0.0	0.0	0.0	0.0	5.538156	1.944025	0.531507
4	0.0	0.0	1.0	0.0	0.0	0.0	-0.320414	-0.757680	0.000000
5	0.0	0.0	1.0	0.0	0.0	0.0	1.138617	-0.466087	0.353425
6	0.0	1.0	0.0	0.0	0.0	1.0	0.577451	-0.579154	0.000000
7	0.0	0.0	1.0	0.0	0.0	1.0	9.129616	1.247770	0.602740
8	0.0	0.0	1.0	0.0	0.0	1.0	2.126269	-0.228051	0.000000
9	0.0	0.0	1.0	0.0	0.0	0.0	3.069027	-0.025721	0.515068

[ ]: