Examination of The Factors That Drives The Airbnb Rentals to Success Across New York City

Manxueying Li, Ziwei Zheng

Background

In such a busy and crowded city as New York, more and more people are considering airbnb rentals. Evaluation is critical to the entire airbnb community, enabling tenants to choose their own accommodation plan wisely and enabling landlords to open their homes with confidence, attract tenants and provide intimate accommodation services. The fact that the airbnb business is based on positive reviews is not much of an exaggeration. Each comment helps the guest decide whether to book a certain room or not. The more positive reviews there are, the more landlords earn, according to airbnb.

Data

There are a lot of factors that would affect the popularity of the rent. In this project, we collect a dataset that contains 15 factors that could potentially affect the popularity, including name of the airbnb, the location of the airbnb, its price and availability, etc.
The data is collected from Kaggle
(https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data)

Inspirations

1. What can we learn about different hosts and areas?
2. What can we learn from predictions? (ex: locations, prices, etc)
3. Which hosts are the busiest and why?
4. What features will make a host’s rental popular?

Due to the fact that there is no exact measure of popularity, we will use number of reviews per month to indicate a rental place’s popularity.

Text Processing

Input Text Data:

- “Name” Column, which describes a rental place’s environment by the host.

Preprocessing data:

- Creating BOW and feature vector X
- Creating labels for y (1 for popular and 0 for unpopular)

Text Processing

Methods

- Linear regression :) just to see how poorly it fits the data
- Logistic Regression with solver = ‘liblinear’
- SVM with ‘rbf’ kernel
- Neural Network
  - Hidden layer => activation: ‘relu’, 200 neurons
  - Dropout 50%
  - Output layer => activation: ‘softmax’
  - Compile: loss func: ‘sparse_categorical_crossentropy’, optimizer: ‘Adam’

Data Processing

Input Column(s)	Output Columns(s)
'neighbourhood_group','neighborhood','latitude','longitude','room_type','price', 'minimum_nights', 'calculated_host_listings_count','availability_365'	'reviews_per_month'

Preprocessing data:

- Convert textual columns such as 'neighbourhood' to ascii value
- Split data to ⅔ training set and ¼ test set

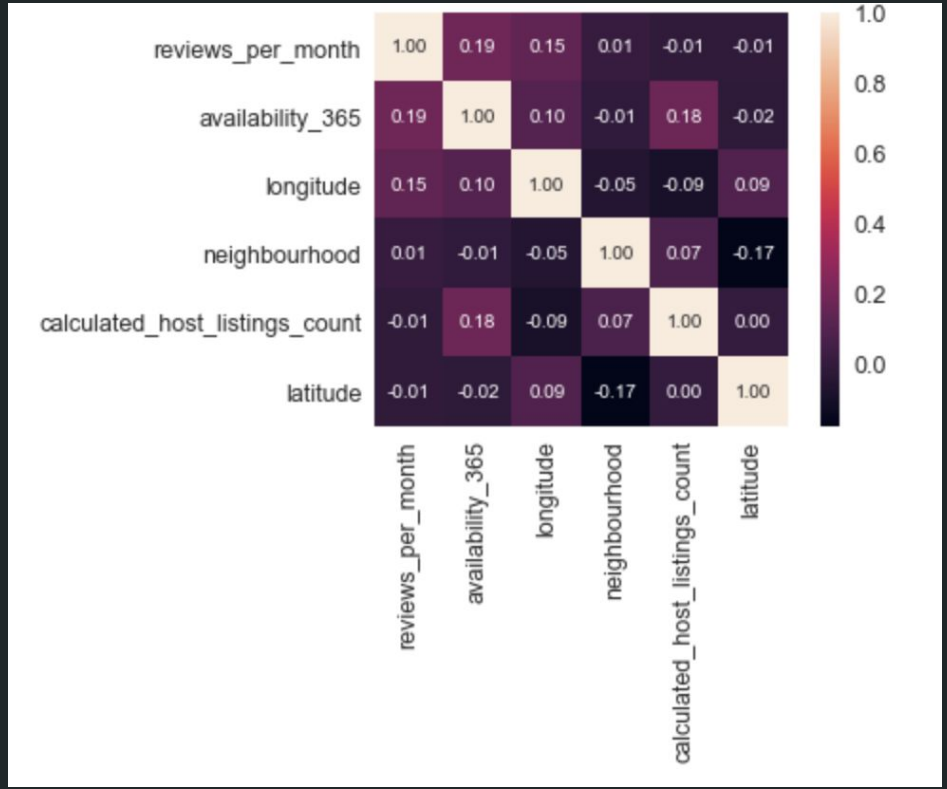
Methods:

- Multi Linear Regression
- Neural Network: Keras Model

Data Processing

Using heatmap, we can find the top features that makes a rental popular based on Numerical data.

Some features in the plot are important for the popularity, with the biggest one being availability in 365 days.



Text Result

Method used	Accuracy on training set	Accuracy on test set	Time taken
baseline(all zeros)	0.998145	0.997450	/
Logistic Regression	0.998145	0.997450	0.9945271015167236
SVM	0.998261	0.997527	204.2654411792755
NN	0.999189	0.996214	606.9619624614716

The result could be improved by using a better parameterized model, eg. logistic regression with different gamma and lambda, and SVM with different gamma and C, as well as different number of neurons with multiple layers. Training the model itself already takes a long time, so we did tried doing so. But they are possible improvements.

Text Result

Since the neural network does not have a good performance as we expected, we tried to condense our feature vector to 500 most frequent words used for names using Tokenizer on Keras and setting the num_words to 500.

```
max_words = 500
tokenize = text.Tokenizer(num_words=max_words, char_level=False)

tokenize.fit_on_texts(Xtr1)
print("word index:\n",tokenize.word_index)

x_train = tokenize.texts_to_matrix(Xtr1)
x_test = tokenize.texts_to_matrix(Xts1)
```

Method used	Accuracy on training set	Accuracy on test set
baseline(all zeros)	0.997991	0.997759
NN with max_words=500	0.999575	0.997914

Data Result

For Multi Linear Regression Model, we evaluated the performance through loss function, has the following result for training set:

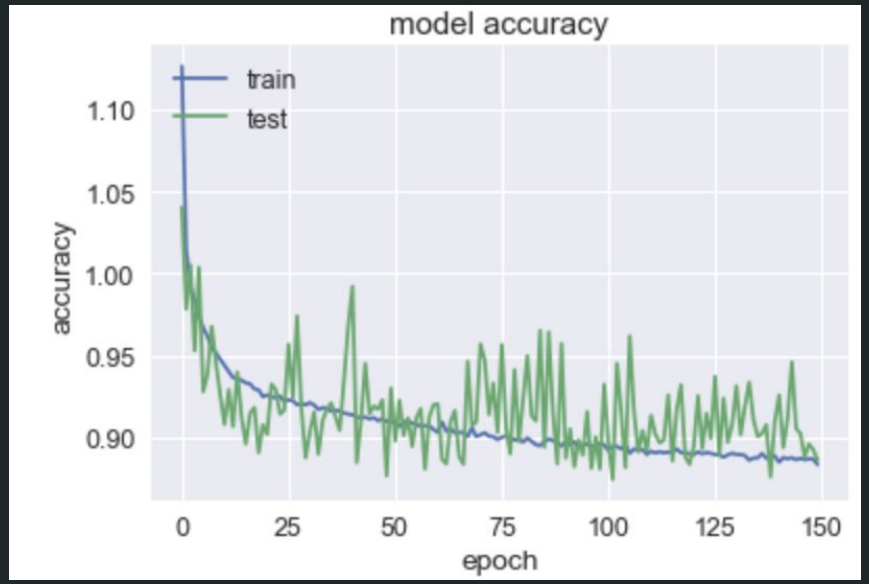
multiple variable loss=4.03e+3

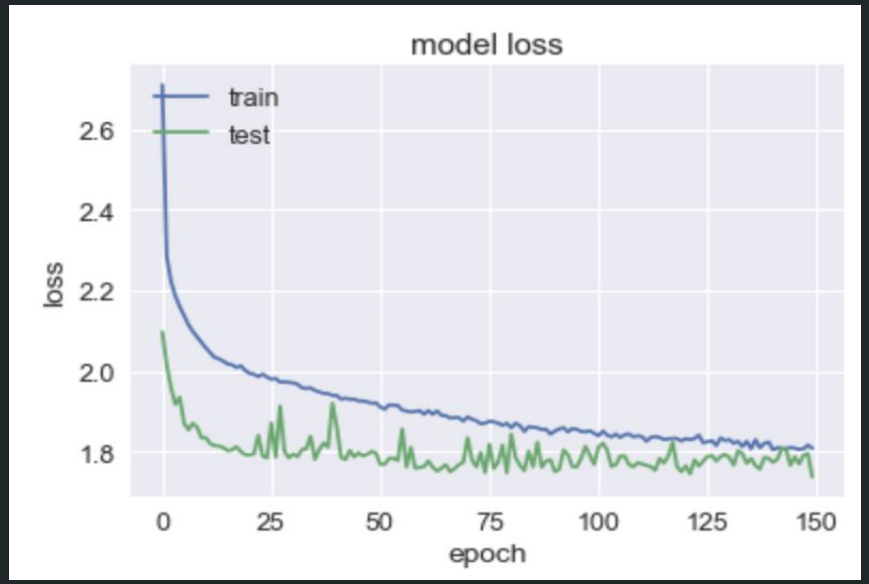
The model has a loss of following for test set:

multiple variable loss=6.72e+04

Since the multi-linear regression model did not perform too well, next we tried to use neural network model to help us predict a rental place’s popularity.

Data Result





The left plot accuracy is calculated through mean absolute error. The right plot shows value loss. This result is not very good but could have possible improvements through such as bigger dataset for the neural network.

Conclusion

Possible thing we could do if we have more time:

1. Try more sophisticated methods in Natural Language Processing on the text
    a. Vader from the NLTK module for sentiment analysis
        i. Using scores from Vader (a neutrality score, a positivity score, a negativity score, an overall score) to transform each word into numerical vectors
    b. Doc2Vec model and TF-IDF model
        i. To find the significance of each word in "name" and to filter words.
    c. A Random Forest classifier on training set
2. Looking at descriptive data and numerical data together and finding out how it affects the output popularity

Questions?