Smart Pricing - Airbnb Price Prediction for New York City

Chang-Han Chang
Indiana University, Bloomington
M.S. in Data Science
cc93@iu.edu

Baekeun Park
Indiana University, Bloomington
M.S. in Intelligent System Engineering
parkbae@iu.edu

Abstract

After the Covid-19 pandemic, a favorable recovery in the real estate market has been predicted by many analysts, and a significant rise in housing sources as well as the high demand for intelligently setting up the housing price for hosts can be expected. The goal of this project is to utilize machine learning techniques to train a Deep Neural Network model to predict the housing price for Airbnb properties in New York City.

1 Introduction

The COVID-19 crisis significantly impacted the residential estate market. In 2020, home sales have dropped to a very low level, nearly to the lowest levels during the financial crisis that began in 2007, since the health concerns and stay-at-home orders led to fewer buyers looking for homes and fewer sellers willing to list their properties or allow strangers to enter their homes during the pandemic. (Gascon and Haas, 2020)

Observed by Gascon and Haas (2020), dispite the real estate activity began to improve and potential buyers started to increase their housing search and purchase activity in the summer, the housing supply did not recover at the same pace, similarily in the rental market, the availability of hosing source in the U.S. has hit a record low (Conor Sen, 2020). However, given the vacancy rates have soared to a unprecedented high, housing source and supply will likely to bounce back hardly after the pandemic. In response to the forseen rise in housing source as well as the high demand of intelligently setting up the price for hosts, this study employs machine learning techniques to train a Deep Neural Network (DNN) model with 3 layers by given parameters such as locations, areas, price, and reviews, to give a prediction on housing price based on the Airbnb properties in the New York City.

Prior to our training, we implemented a mathematical method, 1.5 times of interquartile range (1.5IQR), to identify the outliers in our data, and a statistical method to differentiate the importance between different features, then we remove such outliers as well as the unimportance features to offer a representative sampling of the factors that affect the accuracy of the price prediction. In our study, our DNN model successfully predicted a price by given parameters. However, whether the variance in terms of the accuarcy for the prediction varies positively or negatively, our aim was to first give a prediction then discuss about what factors potentially affect the accuracy.

2 Related Work

In predicting prices for housing particuliarly in neural network models, Shen et al. (2020) proposed a text-based price recommendation system to recommend a acceptable price in terms of reasonable loss in Mean Absolute Error for newly added listings for 4 of the biggest cities in the world using Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and mean shift. Other than text-based analysis, a picture-based method is also applied. For example, Piao et al. (2019) proposed a novel prediction model based on Convolutional Neural Network (CNN) and successfully produced high accuracy predictions, 98.68%. In addition, Xiao and Yan (2019) used Radial Basis Function (RBF) neural network algorithm based on Principal Component Analysis (PCA) to predict housing prices and were able to get a very small loss in ierms of low Mean Squared Error (MSE) during their training.

Besides, many machine learning algorithms and regression techniques have also been applied for price prediction. Li et al. (2016) proposed the machine learning algorithms, Linear Regression

model with Normal Noise (LRNN), and successfully predicted the reasonable price for Airbnb. Also, Durganjali and Pujitha (2019) used various classification algorithms, Logistic Regression, Decision Tree, Naïve bayes, and Random Forest, to predict the house resale price and analyzed the advantages and disadvantages for the algorithms.

All the above studies draw methods from machine learning techniques and algorithms for predecting price for real estate properties, either rental or resale. Different from those methods, the authors of this study use a Deep Neural Network (DNN) model and take a variety of features from a city-specific corpora in New York City as inputs to train the model in order to predict a price for Aibnb properties.

3 Data Sets

This study offers a smart pricing analysis of rental housing market on Airbnb New York City Airbnb Open Data corpora (Denis Gomonov, 2019). The corpora contains the Airbnb listings and metrics in New York City from January 2019 to December 2019 (48,895 records). To identify the relationship between price and properties as well as the housing distribution from different reigons in New York City, we ploted a density map (Figure 1.1) and a distribution map (Figure 1.2). The density map demostrates the regions with relatively higher price, and the distribution map illustrates the listings among different areas throughout the year.

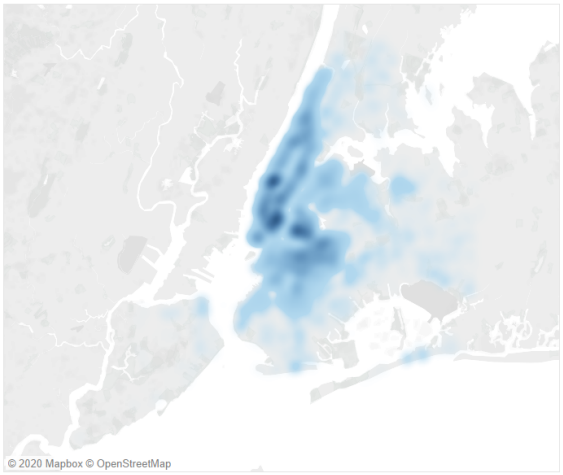

Figure 1.1

To isolate a representative sample of potentially useful features and records, we analyzed Airbnb data, extracting 41,810 valid subsets in terms of

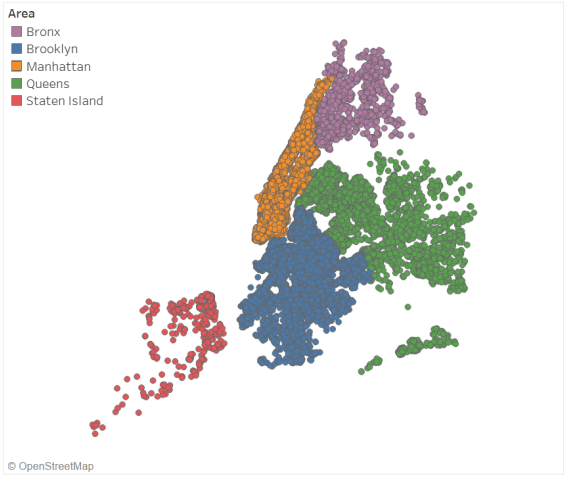

Figure 1.2

the features with high importance (Imp.) of influence to the housing price (Table 1) to decrease the disturbance in our training. A outlier, however, is another factor which significantly affects the training. We then parsed the selection further to isolate only those records that are identified in the 1.5IQR to remove the outliers, price from 16 dollars per night to 233 dollars per night, resulting as the inputs for training our machine learning model. A price distribution plot (Figure 1.3) and a box plot (Figure 1.4) are plotted to illustrates distribution of price at different reigions of the finalized input data.

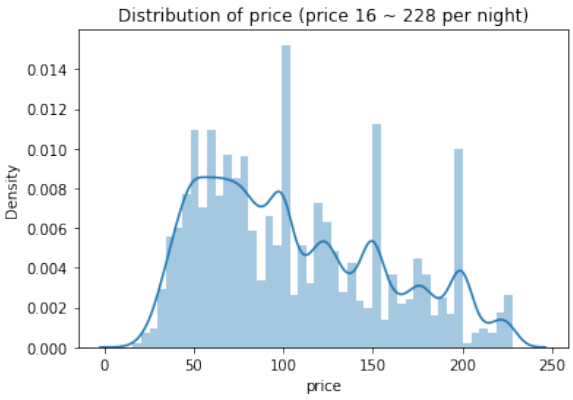

Figure 1.3

In our particular study, in order to feed the input data in our machine learning model, we utilized One-Hot Encoding operation to transform the catagorical features (the object data type in Table 1) into numerical data. Besides, in order to fasten the training speed, a normalization, Min-Max method, was also applied as it has been shown by Jin et al. (2015) to greatly accelerate the training

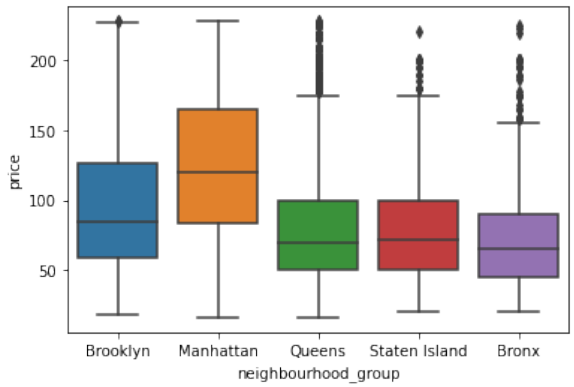

Figure 1.4

for linear neural networks.

Feature	DType	Imp.
id	int64	low
name	object	low
host_id	int64	low
host_name	object	low
neighbourhood_group	object	med
neighbourhood	object	med
latitude	float64	high
longitude	float64	high
room_type	object	high
price	int64	
minimum_nights	int64	med
number_of_reviews	int64	high
last_review	object	low
review_per_month	float64	high
calculated_host_listings_count	int64	med
availability	int64	high

Table 1: Features.

4 Method

4.1 Feature Engineering

In our study, we identified the importance of the features in terms of the influence on price of the Airbnb data by using the Extra Trees Classifier, also known as Extremely Randomized Trees, and labeled them as high, medium (med), and low, as shown in Table 1. Then, we removed the features labeled as low in order to decrease the disturbance during our training. In addition, outliers, which have been proved by Khamis et al. (2005) to negatively contribute to neural network perfromance, defined as the price data outside the range of low whisker and high whisker in our corpora, should be removed. For the price data in the original intact corpora, the interquartile range is the subtraction of the 75^{th} percentile Q_3 and the 25^{th} percentile Q_1. The low whisker of price W_l is the range between 1.5 times the interquartile range IQR and $Q1$, and the high wisker of price W_h is the addition of 1.5 times the interquartile range IQR and $Q3$. An operation followed by the 1.5IQR rule was applied to remove the outliers.

$$W_l = Q_1 - 1.5IQR$$

$$W_h = Q_3 + 1.5IQR$$

4.2 One-Hot Encoding

One-Hot encoding is a process by which categorical variables are converted into a form that could be provided to machine learning and deep learning algorithms to do a better job in prediction. (Harrag and Gueliani, 2020). Our study employs the process to transform the catagorical data (the object data type listed in Table 1) into numerical data in which the legal combinations of values are only those with a single high (1) bit and all the others low (0).

4.3 Min-Max Normalization

Jin et al. (2015) has shown that normalization can speed up the training in linear neural networks. Our study applied the most commonly used normalization method, Min-Max normalization, to transform data. For every feature, the minimum Min value gets converted to 0, the maximum value Max gets converted to 1, and every other value X gets converted to a decimal number N between 0 and 1.

$$N = \frac{X - Min}{Max - Min}$$

4.4 Training

In this study, we exploited TensorFlow and designed a deep neural network with 3 Rectified Linear Unit (ReLU) layers, where the first layer has 233 neurals, the second layer has 128 layers, and the third layer has 1 layer, and the Adam optimizer with learning rate set to 0.01, as our deep learning model (Figure 2) to produce the target output, price.

4.5 Evaluation

To evaluate the robustness of our DNN model, we took the good advantage of one of the statistical indicators, Mean Squared Error (MSE), to get a

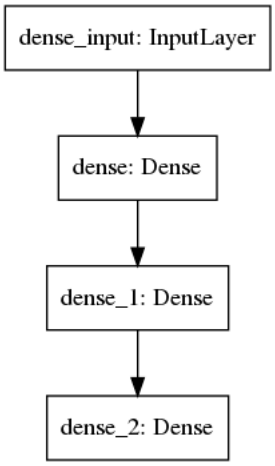

Figure 2

Actual	Prediction	Variance
120	164.925	37.44%
49	65.591	33.86%
100	95.697	-4.30%
150	93.340	-37.77%
148	131.265	-11.31%
89	144.150	28.26%
200	177.138	-11.43%
145	75.507	-47.93%
105	93.256	-11.19%
200	184.603	-7.70%

Table 2: Predictions.

straightforward L2-Norm loss. The benefits of choosing MSE is, even a small loss gets amplified, resulting a clear observation and robust evaluation. MSE is caculated by the fomula as following, where n is the number of data points, Y_i is the i^{th} evaluation data, and $\hat{Y}_i$ is the i^{th} prediction data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

5 Result

The prediction result of our trained DNN model is plotted in Figure 3, where the red line represents the perfect prediction line, and the 10 randomly selected predictions present in Table 2, indicating that a large number of predictions have unacceptable variance, either positive or negative, dispite our attempts at data preprocessing. However, implemtention of this machine learning pipeline gives a valuable ideas of determining features as well as designing the deep learning models at the preliminaries.

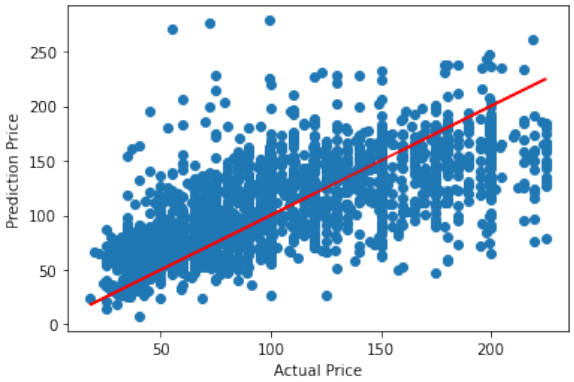

Figure 3

6 Conclusion

The Deep Neural Network (DNN) technique uses a back propagation algorithm to train a model until it converges, which we believe is a proper technique to apply to solve the problems like price prediction. However, many other factors have been identified to largely affect the housing price on Aribnb, such as floor, view, and furniture, and not existing in our data corpora. We observed that listings at higher floor tend to emphsize their beautiful views and have higher price. Similarily, the properties with high-class furniture are observed to have higher price. We believe that besides outliers and the features we removed, many more factors need to be included to train our machine learning model improve the performance in terms of decrease in variance and increase in accuracy.

Source Code

Source code for all experiments and results can be found here `https://github.com/dzcyb0rg68/e503_airbnb_prediction`

Acknowledgments

The authors wish to thank Dr. Ariful Azad for his instruction and supports. The work was part of the course, Intro to Intelligent Systems (E503), in 2020 Fall at Indiana University, Bloomington. Also, all the codes in this study were computed on GH Server in the Luddy School at Indiana University (gh.luddy.indiana.edu).

References

Charles S. Gascon and Jacob Haas (2020, October) The Impact of COVID-19 on the Residential Real Estate Market *Regional Economist.*

https://www.stlouisfed.org/publications/regional-economist/fourth-quarter-2020/impact-covid-residential-real-estate-market

Conor Sen (2020, December) Pandemic Housing Shifts Will Speed Recovery in 2021 *A slumping home market dragged out a rebound from the 2008 recession; this time the industry will help, not hinder, the economy.*. https://www.bloomberg.com/news/articles/2020-12-12/pandemic-housing-shifts-will-speed-2021-economic-recovery

Denis Gomonov 2019 New York City Airbnb Open Data *Airbnb listings and metrics in NYC, NY, USA (2019).* https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Yang Li, Quan Pan, Tao Yang, and Lantian Guo (2016, July) Reasonable price recommendation on Airbnb using Multi-Scale clustering *35th Chinese Control Conference (CCC), Chengdu*. pp. 7038-7041

Lujia Shen, Qianjun Liu, Gong Chen, and Shouling Ji (2020, February) Text-based price recommendation system for online rental houses *in Big Data Mining and Analytics*. vol. 3, no. 2, pp. 143-152

P. Durganjali and M. Vani Pujitha (2019, March) House Resale Price Prediction Using Classification Algorithms *2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India*. pp. 1-4

Lizhong Xiao and Tingrui Yan (2019, November) Prediction of House Price Based on RBF Neural Network Algorithms of Principal Component Analysis *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Shanghai, China*. pp. 315-319

Yong Piao, Ansheng Chen, and Zhendong Shang (2019, August) Housing Price Prediction Based on CNN *2019 9th International Conference on Information Science and Technology (ICIST), Hulunbuir, China*. pp. 491-495

Jian Jin , Ming Li , and Long Jin. (2015, July) Data Normalization to Accelerate Training for Linear Neural Net to Predict Tropical Cyclone Tracks

Azme Khamis, Zuhaimy Ismail, Khalid Haron, and Ahmad T. Mohamm (2005, July) The Effects of Outliers Data on Neural Network Performance *Journal of Applied Sciences*. vol. 5, issue 8, pp. 1394-1398

Fouzi Harrag and Selmene Gueliani (2020, August) Event Extraction Based on Deep Learning in Food Hazard Arabic Texts *Social and Information Networks*.