Examining Listing Price Indicator in Tourism Industry

Case Study from Airbnb

Yuichi Kuriyama

***Abstract*— Thanks to the rapid technological development, people could have more variety options for accommodations in tourism. Airbnb was a breakthrough in terms of the flexibility to provide more affordable and yet amazing experience for users. This case study attempts to demystify the factor contributing to the listing price and construct an ideal predictive model for both users and hosts by examining a reason behind factors. By applying Random Forest to Airbnb accommodation data and investigate the result, my result shows that geographical location from city centre, room quality and the amount of text information were great factors to affect listing price. However, continuous research is essential to construct un-biased predictive model by conducting more comprehensive data collection and incorporating a wide range of diverse variables such as host verification.**

Keywords—Airbnb, Feature Engineering, Random Forest, Geographical Proximity, Natural Language Processing

I. INTRODUCTION

Due to a fierce marketing competition, dramatic change has been observed in the tourism industry in line with the rapid development of online services. For instance, Airbnb is a peer-to-peer online platform that enables property owners to put their vacancy room online so that guests can find an accommodation as a type of sharing economy [1]. Given the popularity, Airbnb has been an alternative route to find affordable rooms in comparison to traditional accommodations such as family hotels. This industrial revolution in tourism also left an impact upon the marketing behaviour and the leasing strategy for the urban property owner [2]. Thanks to the arrival of Airbnb, the property owner can opt the property investment style from long-term rental to on-demand short-time accommodation for vacations. As a result, in order to maximise a stable profit for the property owner and find a best deal for guests, setting an optimal listing price plays a pivotal role in the customer satisfaction and active user retention in the platform.

The aim of this case study is to examine indicators to affect the listing price in Airbnb and construct the predictive model with validation method by analysing several features. Section 2 (Analytical questions and data) clarifies analytical questions to archive the aim and the source of the dataset. Section 3 (Data) explores key characteristics of the dataset along with research interests with detailed description with all considerations from data. Section 4 (analysis) describes each analytical step based upon the coherence of the research questions. Section 5 **(**Findings, reflections and further work**)** critically argues the result with figures and charts. In addition, this section also includes limitations and further lines of enquiry.

II. ANALYTICAL QUESTIONS AND DATA

2.1. Analytical Question

As mentioned in initial motivations above, optimal listing price could be a key factor to further success for the business. Moreover, importance should be placed upon a reasonable explanation about which features are affecting the price.

In order to address the motivation, analytical questions are listed below:

i. How does geographical location affect the price and how much does it contribute to the model?

ii. How does room quality affect the price and how much does it contribute to the model?

iii. How does text information affect the listing price and how much does it contribute to the model?

iv. What is the ideal predictive model for housing price?

These questions require to carefully examine the dataset throughout the exploratory data analysis and feature engineering. Furthermore, this study also attempts to create a predictive model to be deployable for future reference. This analytical question is suitable for the aim and motivation as those factors may be considered as a potential determinant of listing price. For instance, geographical location could be crucial factor as guests might prefer staying the convenient location in terms of sightseeing and access to the public transport. Likewise, room quality could be important as users would like to have a comfortable experience during their stay. Text information also might be a good illustration of determining the price as users can obtain many implications from text information such as an engagement from hosts in the accommodation. Answering the last analytical questions eventually allows to draw a meaningful result and could be useful for Airbnb to improve customer and host experience in a foreseeable future.

2.2. Data

Dataset is extracted from Signate (Japanese data science competition platform) [3]. Although the website is Japanese, dataset is from US. Hence, it does not require any translation for reproduction of this study.

III. DATA (MATERIALS)

An original data contains 55583 columns and 29 columns including a wide range of information about accommodations in five major cities (NY, LA, SF, DC, Chicago and Boston). The original data was split 75% for training data and 25% for test data. This ensures the assessment for the model against unseen data [4].
In order to observe the performance, Columns are classified by 5 categories (room, review, location, host, others).

i. Room:
Number of bathrooms, bedrooms, Accommodatable person for each room, Room type, Amenities

ii. Review:
First review date, Last review day, Number of reviews, Review score rating

iii. Location:
City Name, Latitude, Longitude, Zip code, Neighborhood

iv. Host:
Cancellation policy, Cleaning fee, Profile picture of host, Host identification, Host response rate, Duration of being host.

v. Other:
ID, Description, Name, Thumbnail, y

In other category, Thumbnail column contains information about whether accommodation have thumbnail photo. Furthermore, column 'y' is classified as a target column to predict and provide information about the listing price for each accommodation.

For the sake of the research scope, unrelated variables in regard to the analytical question such as ID and Host information were removed. This allows to focus on the analytical interest in-depth. It seems to be true that the richness of the amount of data with ample information for each accommodation could meet the satisfaction for the research interest. In terms of ethical consideration, although datasets include personal information of accommodation such as zip code, datasets meet ethical consideration given that information is from Airbnb website and open resource and originally datasets was published by Airbnb. Therefore, this data material is satisfactory for this research project.

IV. ANALYSIS

4.1. Data Preparation

Prior to the model construction, several steps must be taken to obtain a meaningful implication. First, the dataset is split 75% for train data and 25% test data. This sensible data strategy enables the final model to estimate a generalisation performance against unseen data [5]. Secondly, Figure 1 shows a right-skewed distribution of listing price from training data. As can be seen, outliers are observed in the very expensive accommodation. To remove this, interquartile range (IQR) measure is applied to detect outliers. More precisely, each value greater than the 3rd quartile enlarged for interquartile range multiplied by 1.5 was treated as an outlier [6].

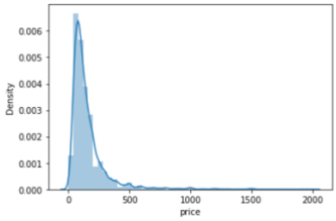

Figure 1

Next step involves with an imputation of missing values. For instance, the number of bedrooms, bathrooms, baths indicate a luck of value needs to be replaced. By having a reasonable assumption and common sense, the imputation was done as follows.

i. The number of bathrooms

- 1 per accommodation

ii. The number of beds

- 1 or 2 accommodates should be 1 bed
- 3 or 4 accommodates should be 2 beds
- Anything beyond 4 should be 3 beds

iii. The number of bedrooms

- Between 1 and 4 accommodates should be 1 bedroom
- More than 4 accommodates should be 2 bedrooms

Last but not least, categorical variables such as city and property type were also transformed to numerical variables for dataset to apply machine learning algorithm for model building on later stage. (Ex: LA:0, DC:1, SF:2, NYC:3, Chicago:4, Boston:5)

4.2. Data Derivation

4.2.1. Zip code

Each accommodation is assigned with zip code showing approximate location. As incorporating zip code without any manipulation might be problematic due to many distinct information, alternative data derivation method was applied to solved this.

Figure 2

Figure 2 provides information about the postal code system in the US with colour map. Generally, the first two digits define the state and borough, and other digits determine more precise area. Therefore, as a part of feature engineering, the first three digits were picked as a feature to see the impact of new feature within manageable number of categories.

4.2.2. Longitude and Latitude

Longitude and Latitude also indicate the geographical location for each accommodation in-depth. As location is commonly accepted factor to influence listing price [7], feature engineering based upon geographical proximity might be effective method when investigating indicators.

Figure 3

Figure 4

Figure 3 illustrates an interactive plot of data point for each accommodation in DC and similarly and Figure 4 shows each accommodation with accommodation price distribution. According to these figures, it implies accommodation around downtown tends to be more luxury in comparison to other counterparts. For this reason, this project attempts to create a feature by calculating a distance from iconic places in city centre in each city. For instance, in the case of DC, direct distance from white house was computed and added as a new feature called 'distance from central'.

4.2.3. Amenities

Availability of amenity might also be a potential indicator to attract the customer. To utlise this information, the number of amenities were counted and incorporated as a new feature.

4.2.4. Description

Description can often indicate detailed information about accommodations. This also include basic facilities, locational information, social engagement from hosts. In addition to counting the number of words in description, several techniques have been implemented to conduct a meaningful data derivation.

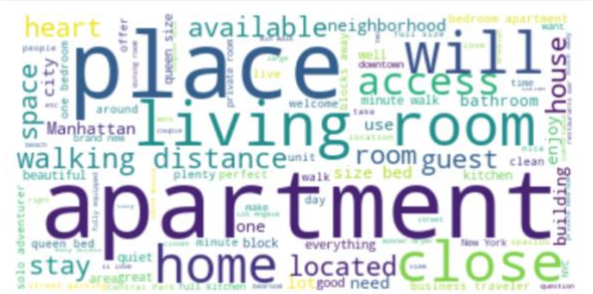

Figure 5

Figure 5 reveals visualising a word representation from description after removing stop word to reduce the noise of textual data [8]. What can be clearly seen in this figure is the general pattern of word used in the description. Although it is understandable to see common words such as room on the top, it is interesting to see an appearance of price-sensitive words.

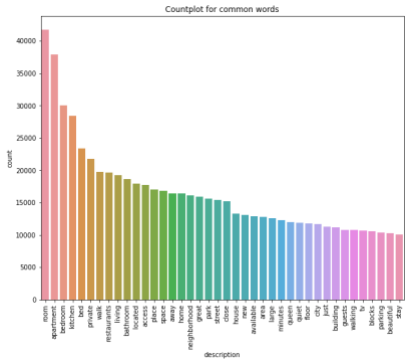

Figure 6

Figure 6 reveals the count plot for top 30 words from description. Based upon those two figures, several new columns were created to detect price-sensitive words such as 'luxury', 'cheap' and 'Manhattan'.

4.2.5. Name

The similar principle could be applicable to name column to create more informative attributes by extracting information from existing attributes.

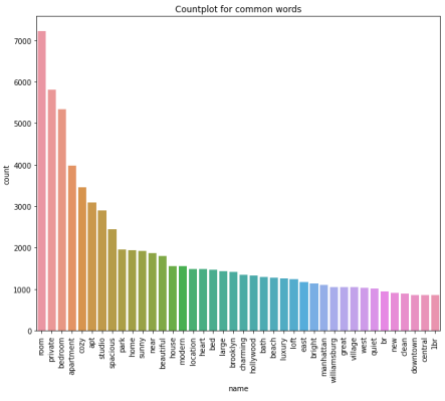

Figure 7

In general, name column indicates some locational and property information to obtain an attention from users when they were browsing the website. Hence, accommodation including some keywords such as Hollywood and Brooklyn were extracted as features.

4.3. Construction of model

In this case study, Random Forest was applied to predict the listing price. Random Forest is particularly suitable for the study not only for the high performance of the model but also for the interpretability. Although linear model is also used to assess feature relevance, due to the inflexibility such as homoscedasticity and independence [9], the difficulty arises for implementation. Hence, as Random Forest is a non-parametric model and provides a heuristic for correcting biased measures of feature importance, called *permutation importance* [9], which aids providing a critical answer for analytical questions. Moreover, hyper-parameter tuning was implemented by Random Search. As Random Search allows to optimise the hyper-parameter effectively with the same computational budget compared with Grid Search [10], this methodology could meet the satisfaction for the research scope. In line with this, 3-fold cross variation was also applied to see an estimation of prediction accuracy [11] due to computational capacity.

4.4. Validation of results

In order to observe the model performance against unseen data, final constructed model was also run by test data. As an evaluation metric, RMSE (Root Mean Square Deviation) were employed in this case study. RMSE is relatively widespread measure to predict a correctness of the model in the regression task. RMSE stems from the idea that the smaller the discrepancy between the actual and predicted values, the better the model fits. The reason that RMSE was chosen over other evaluation metrics is the sensitivity against errors [12]. As RMSE tends to be more sensitive to outliers and yielding a huge error might be more problematic in this task, this choice seems to be appropriate.

V. Finding, Refrelctions and Future work

5.1. Finding

The final model result against test data is shown below.

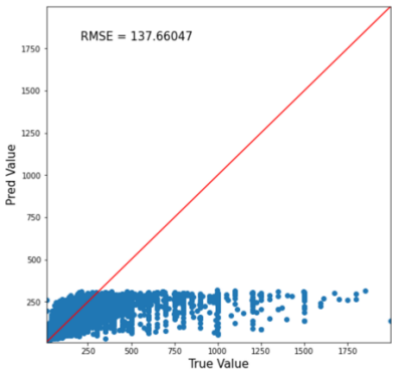

Figure 8

Figure 8 illustrates the result of RMSE and the actual value and predicted value for each accommodation. According to Figure 8, although Random Forest model tends to be robust to the normal price range accommodation around $250, luxury and high-price accommodation above $500 did not show a great prediction. One possible assumption for this is the outlier in the training data. As a preprocessing, outliers were removed because that might potentially affect the model ability. Given Random Forest could not be sensitive to the outlier, Random Forest could have better performance on predicting those price ranges without removing outliers. That being said, it seems to be true that this model is more robust to normal price accommodation. Therefore, constructing a separate model might be more effective measure.

Next, Figure 9 illustrates a feature importance for each variable incorporated in the model after feature engineering. What stands out from Figure 8 is the contribution for room type to the model. As it stated in the hypothesis, room type could indicate several implications such as the size of the room and the atmosphere of the room. Furthermore, what can be clearly seen is the dominance of variables showing the geographical proximity such as distance from central, longitude and latitude. Since distance form central specified more iconic places such as SOHO, it could be easier to capture some information compared with simple latitude and longitude.

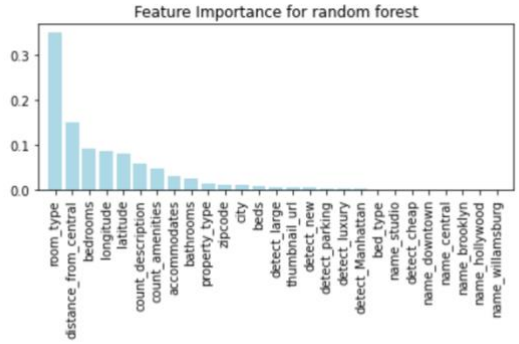

Figure 9

Moreover, text-related features show more interesting results in terms of the feature importance. For instance, counting the word for description and name show approximately 6% and 5% for each to the contribution to the prediction, resulting in the intermediate predictive power. However, for the word detection variables, it has very weak effect for prediction. One implication for this is every host want to mention attractive keywords to grab customer's attention although it does not describe the accommodation precisely due to improving search engine optimisation on Airbnb platform.

5.2. Reflection

For first analytical question, since there are huge demands for a convenient location for users and the fluctuation of land value could affect the listing price, geographical location played an important role in the predictive model according to the figure 8.

Answer for second question was also addressed throughout this project. By imputing a missing value and

label encoding, hidden features were successfully incorporated to the model. The incorporation of room quality variables such as room type, and bedrooms helps distinguish the listing price.

Thirdly, the amount of text information could also be a potential factor for the listing price. Text information gives users a sense of security, stemming from host's attention and engagement for the accommodation. Given that Airbnb is a mere third party to mediate users and hosts, users might be somewhat judgmental about the host. Hence, having a flexible yet long message could provide a comfort zone for users and end up having a high-price due to the popularity.

Lastly, Price prediction should be fair for every different accommodation. Given a high incorrectness for luxury accommodation, it could be better to create the model separately, which enables the prediction more fairly.

5.3. Future Work

Needless to say, there is a dire need of collecting more data especially high price accommodation to create the robust model. At the same time, some variables not including the research scope such as host verification could produce more accurate and equitable results for the analysis. Also, in order to aid business decisions, more sophisticated methods to understand the model such as alpha value could be vital.

[1] Cheng, M. and Jin, X., 2019. What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, 76, pp.58-70.

[2] Krause, A. and Aschwanden, G., 2020. To Airbnb? Factors Impacting Short-Term Leasing Preference. *Journal of Real Estate Research*, 42(2), pp.261-284.

[3] Signate, Airbnb Listing Price Competition available at https://signate.jp/competitions/266

[4] Nguyen, Q., Ly, H., Ho, L., Al-Ansari, N., Le, H., Tran, V., Prakash, I. and Pham, B., 2021. Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. *Mathematical Problems in Engineering*, 2021, pp.1-15.

[5] Xu, Y. and Goodacre, R., 2018. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, 2(3), pp.249-262.

[6] Pupovac, V. and Petrovecki, M., 2011. Summarizing and presenting numerical data. *Biochemia Medica*, 21(2), pp.106-110.

[7] Zhang, Z., Chen, R., Han, L. and Yang, L., 2017. Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach. *Sustainability*, 9(9), p.1635.

[8] Saif, H, Fernandez, M, He, Y and Alani, H (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: *LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings.*, pp. 810–817.

[9] Altmann, A., Toloşi, L., Sander, O. and Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10), pp.1340-1347.

[10] Bergstra, J. and Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, pp.281-305.

[11] Fushiki, T., 2009. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21(2), pp.137-146.

[12] Chai, T. and Draxler, R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), pp.1247-1250.

Section	Word Counts
Abstract	**129**
Introduction	**277**
Analytical Questions and Data	**300**
Data (Materials)	**273**
Analysis	**993**
Findings, reflections, and further work	**600**