Recommendation of Price on Airbnb using Machine Learning

Jae Won Choi

Abstract: ***Airbnb's rental property prices are challenging because they determine how many customers there are. Customers, on the other hand, need to evaluate the offer price with minimal knowledge of the optimal value of the accommodation. This white paper aims to develop a reliable pricing model that uses machine learning and natural language processing techniques to assess prices by providing minimum available information about prices for both real estate owners and customers. Attributes, rooms, and bed features made up the predictors and created the prediction model using a variety of methods, from linear regression to root mean square error evaluation was used for creating the prediction model.***

Keywords* : *Sharing economy, Airbnb, Rental price, Machine learning

I. INTRODUCTION

Recently, the sharing economy emerged in the tourism and lodging industry accommodation sector [1]. Airbnb used sharing economy to connect idle accommodation assets including empty accommodation or apartments, to people who own them and who need temporary accommodation via a digital market place [2]. Price is widely known as very important features to determine continuous success of the accommodation area. In particular, it is important to study the price elements of shared economic accommodation, as the generalization of existing studies of hotel price-setters is limited in the context of shared economies. Many price indicators used in the traditional hospitality industry, including star ratings and corporate partnerships, are not suitable for accommodation in a shared economy and are primarily personal assets used in housing [2]. Thus, a set of new price indicators related to shared economy-based accommodation, such as host economy, special amenities, and certain diversified accommodation characteristics have been identified. In addition, due to the unique characteristics of shared economic accommodation services, particularly idle assets and non-expert business owners, it is useful to review the impact of decision-making factors related to the traditional hospitality industry. For example, since the location of a shared economy lease is not determined in advance by the supplier, the impact of the price on that location is not clear.

The purpose of this study is to identify pricing factors for shared economy-based accommodation products in digital market places (especially at Airbnb.com). The sample data provided by Airbnb.com is examined. Linear regression and split validation with RapidMiner is used to investigate pricing factors for accommodation types, including properties, rooms, and beds. The findings have a significant impact on the design of pricing systems to share a cost-based accommodation provider, such as the pricing recommendation tool recently launched by Airbnb.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author

Jae Won Choi*, College of software, Chungang University, Heugseoglo 84, Dongjaggu, Seoul, Korea,

II. RELATED STUDY

A. The Determinants of the Rental Price of Airbnb

Some study showed that hosts providing rental accommodation at Airbnb.com would charge higher prices if they were typically star stars with higher accommodation [2]. The study initiated an effort to investigate the factors that determine the price of shared economy-based accommodation [3]. However, these studies have been developed primarily in urban data sets and have limited independent parameters for certain aspects, such as host characteristics or location. These research designs limit the understanding of price elements to share accommodation rentals. From our perspective, a global model that controls geographic locations and describes pricing elements in cities around the world can better reflect market conditions. Assuming that tourists from around the world can accommodate Air B&B's list price, the global model reflects the link between price and price elements in a balanced market. You can also examine many determinants, such as host properties, site and property attributes, amenities and services, lease rules, and online review ratings, to better represent the impact of price elements. Although only a few studies have been identified on the pricing factors for a shared economy-based accommodation rental, many studies have been conducted on hotel pricing factors, which provide a benchmark for designing the study and comparing results to the two types of differences. The following sections review relevant research on hotel pricing factors.

B. The Determinants of Hotel Price

The identified hotel pricing elements are divided into five categories: site-specific characteristics, quality signal elements, hotel services and amenities, accommodation specifications and external market elements. The most important site-specific attributes are hotel locations that play an important role in hotel investment [4]. Hotel locations are typically provided in terms of downtown, transportation hubs, major tourist attractions, or distance to the beach.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The results of previous research on the effect of hotel location on price are fairly consistent. Near-focus distances, such as downtown, are generally associated with higher prices.

The second category of hotel price elements consists of quality signal elements defined as "various factors that reduce information asymmetry in the market by providing information about the quality of a product purchased by a buyer" [5]. Researchers identified several hotel quality signal elements, such as star ratings, online customer ratings, and chain partnerships. Using an asterisk rating system, "the same type of accommodation (e.g., hotels, motels and inns") generally depends on "typical physical and service characteristics" and on individuals at government, industry and other personal levels. Organizations that generate ratings vary from country to country. Previous studies have shown that star ratings have a significant positive impact on hotel prices in western and eastern countries [5]. In addition, Yang et al. [5] provides empirical evidence that high customer ratings have a positive impact on hotel prices. Finally, the brand chain partnership has been identified as an important hotel quality signal. Researchers have empirically shown that hotels linked to the brand chain generally charge higher prices [5].

The third category of hotel prices consists of hotel amenities and services. Variables related to amenities and services are listed in several hotel pricing models [6]. Convenience facilities such as minibars, TVs, safes, and hair dryers are generally offered with higher hotel rates. Hotels offering laundry services typically charge lower rates [6]. The higher the room rate, the higher the room rate for providing customers and services such as quick check-out, breakfast and pre-bookings [5]. Inconsistent results were obtained for Internet access. However, this effect has been negative since 2010 due to the generalization of Internet services and the rise of economic hotels [5].

The fourth category of hotel price elements consists of characteristics such as room count, building age and bar presence, parking, fitness center and swimming pool [5]. However, only the parking spaces [7] and fitness centers (Yang et al., 2016) were found to be consistently related to high room rates. Other findings are not clear due to results that do not match previous studies.

The last category of hotel price elements consists of market and industry characteristics. For example, the number and proximity of competitors have been shown to affect hotel prices. Low market access, shown by high flight costs, is associated with a drop in hotel prices [5]. Researchers also conducted demand studies to identify hotel price factors associated with customers' willingness to pay [6].

III. METHODOLOGY

A. Dataset

Air B&B's corresponding variable information comes from the 3rd party website, the international website of the popular Internet platform Kagle (www.kaggle.com) Air B&B provides data based on publicly available information from Air B&B.com. Kagle asked participants to predict how much the rent would be. To help develop the algorithm, organizers provided the right type of accommodation (data stream) for large rental sets.

The effects of 42 variables are examined in the following four categories based on previous literature on shared economy-based accommodation leasing and fourth hotel pricing factor categories: property type, room type, and bed type. These variables are listed and defined in Table 1.

Table I: The variables in each category

Categories	Variables
Rental price	Price
Property types	Apartment, Bed & Breakfast, Boat, Boutique hotel, Bungalow, Cabin, Camper/RV, Castle, Cave, Chalet, Condominium, Dorm, Earth House, Guest suite, Guesthouse, Hostel, House, In-law, Island, Lighthouse, Loft, Other, Parking Space, Serviced apartment, Tent, Timeshare, Tipi, Townhouse, Train, Treehouse, Vacation home, Villa, Yurt
Room types	Entire home/apt, Private room, Shared room
Bed types	Airbed, Couch, Futon, Pull-out sofa, Real bed

B. Analysis Method

This study uses RapidMiner tool to do linear regression and machine learning for the prediction of the rental price of Airbnb.

Linear regression analysis is a technique used for numerical prediction. Linear regression analysis is a statistical method to determine the strength of a relationship between variables. Just as classification gears are commonly used for the prediction of categorical labels, regression techniques are used for the prediction of continuous values. Linear regression analysis is used to fit a linear relationship between quantitative dependent variables Y and K independent variables. The basic model of multiple linear regression analysis is as follows.

$$Y_i = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_K \cdot X_K + \varepsilon_i \qquad (1)$$

As described above, the factors of the facility, the distance to the nearest landmark, and the house price that can be described by the multivariate linear regression model using the nearest landmark popularity. However, due to the actual uncertainty, the uncertainty that can be simulated with noise ΔP should be combined when performing a price prediction.

$$P = A\begin{bmatrix} f \\ d \\ q \end{bmatrix} + distribution(\Delta P) \qquad (2)$$

In this model, we use the facility embedding *fi* representing property types, *d* is room types, and *q* is bed types. In this model, $\beta_0, \beta_1, \beta_2 .... \beta_K$ are the regression coefficients and ε_i is the superorder that occurs in measuring the dependent variable y. Creating a regression model involves finding the regression coefficient $\beta_0, \beta_1, \beta_2 .... \beta_K.$

Generally, the sample from the population is used in the blanch analysis.

Since we do not know the exact regression coefficient β_0, β_1, β_2 .... β_K of the population, we use the least square method (OLS) to estimate it from the given sample data. This method yields an estimate β^*_0, β^*_1, β^*_2 .... $^*\beta_K$ that minimizes the sum of squared deviations between the predicted values Y^* of the model rather than the actual values Y.

$$Y^*_i = \beta^*_0 + \beta^*_1 \cdot X_1 + \beta^*_2 \cdot X_2 + \ldots + \beta^*_K \cdot X_K + \varepsilon_i \qquad (3)$$

OLS can be applied when the dependent variable satisfies the following assumptions.

1. The dependent variable follows the standard normal distribution.
2. The independent and dependent variables have a linear relationship.
3. Each observation is independent of each other.
4. The variability of the value of the dependent variable for each independent variable is the same regardless of the value of the independent variable. This is called homoscedasticity.

If the above assumptions are met, the predicted value is an unbiased value, and the mean square error is the smallest compared to other unbiased estimates. If the regression model is used for prediction, the first assumption mentioned above is not satisfied, and even if the dependent variable has an arbitrary distribution, the estimated value of the prediction can give very good results. This is possible because the data mining approach uses the case for verification separately from the case used in the learning model for verification.

To stay alive in a highly competitive marketplace, many companies are using data mining technology to analyze price forecasts. It is important to build a more effective and accurate rental pricing model to effectively win customers. Statistical and data mining techniques were used to construct a pricing model. Data mining techniques can be used to detect interesting patterns or relationships in data and to predict or classify behavior by fitting models based on available data. In the case where the learning dataset and the test dataset are separated for machine learning, the test dataset must satisfy the following requirements. First, the training dataset and the test dataset must be created in the same format. Second, the test dataset should not be included in the training dataset. Third, the training dataset and the test dataset must be consistent in data. However, it is very difficult to create a test data set that meets these requirements. In data mining, various verification frameworks using one dataset have been developed to solve this problem. This study uses the Split Validation operator provided by RapidMiner to support this. The operator splits the input dataset into a training dataset and a test dataset to support performance evaluation. In this study, we select relative segmentation among the segmentation method parameters of this operator and use 70% of input data as learning data.

IV. RESULTS

A. Linear Regression

The linear regression analysis results are as follows.

Table II : The results of linear regression

Category	Variable	Coefficient	p-value
Property types (33)	Apartment	0.626	0.000
	Bed & Breakfast	0.850	0.000
	Boat	0.821	0.000
	Boutique hotel	1.168	0.000
	Bungalow	0.415	0.000
	Cabin	0.337	0.000
	Camper/RV	0.320	0.000
	Castle	1.265	0.000
	Cave	0.839	0.086
	Chalet	0.705	0.013
	Condominium	0.866	0.000
	Dorm	0.308	0.000
	Earth House	0.986	0.004
	Guest suite	0.582	0.000
	Guesthouse	0.351	0.000
	Hostel	0.256	0.002
	House	0.779	0.000
	In-law	0.434	0.000
	Island	1.368	0.048
	Lighthouse	1.023	0.139
	Loft	0.844	0.000
	Other	0.830	0.000
	Parking Space	1.391	0.045
	Serviced apartment	0.856	0.000
	Tent	0.315	0.054
	Timeshare	1.293	0.000
	Tipi	1.077	0.007
	Townhouse	0.800	0.000
	Train	1.134	0.020
	Treehouse	0.714	0.006
	Vacation home	1.078	0.000
	Villa	1.083	0.000
	Yurt	0.738	0.001
Room types (3)	Entire home/apt	0.710	0.000
	Private room	- 0.146	0.000
	Shared room	- 0.574	0.000
Bed types (5)	Airbed	- 0.023	0.475
	Couch	0.101	0.017
	Futon	- 0.090	0.000
	Pull-out sofa	- 0.027	0.344
	Real bed	0.041	0.009

The analysis revealed that 30 of the 33 variables of property types were significant at the p <0.05 level. All three room types were significant at the p <0.05 level. Especially, private rooms and shared rooms were shown to decrease the rental price. Three of the five-bed types were significant at the p <0.05 level. Especially, a futon was shown to decrease the rental price.For many predictive tasks, this study tries to penalize predictive values that are much farther away than the actual values that are closer to the actual values. To do this, this study can take an average of the square error values, called root mean square error (RMSE). The formula for RMSE is:

$$RMSE = root\{(e_1{}^\wedge 2 + e_2{}^\wedge 2 + \ldots + e_n{}^\wedge 2) / n\} \qquad (4)$$

where n represents the number of rows in the test set. This formula may seem overwhelming at first, but this study is all about:

- Taking the difference between each predicted value and the actual value (or error),
- Squaring this difference (square),
- Taking the mean of all the squared differences (mean), and
- Taking the square root of that mean (root).

So if you read from bottom to top: Square mean square error. In this study, let's calculate the RMSE value of the prediction for the test set. RMSE is approximately $1.367. One of the convenient things about RMSE is that because we take the square root, the unit of RMSE is the same as the predicted value, so it's easy to understand the magnitude of the error. And in this case, it's quite large. This study is still far from accurate predictions.

V. CONCLUSION

This study identifies the factors that determine the price of a shared economy-based accommodation, which is different from the factors that determine the price of a hotel. In the hotel industry, star and chain partnerships have been identified as quality signal elements (Masiero et al., 2015; Saló et al., 2014; Yang et al., 2016). However, the tie-up between stars and chains that can be rented through Air B & B is not relevant. Instead, accommodation usually charges higher prices. Air B&B consumers will be willing to pay premium prices by recognizing the three categories mentioned above as a single quality signal. There is evidence of racial impact on rent prices, but there is not enough empirical evidence of the impact on the availability of profile pictures in previous studies. Several variables have been found to be unique in terms of economy-based accommodation leasing, including the provision of private rooms, public rooms and bedding beds. Studies show that private, public, and bedding beds offer a lower cost of accommodation. This study provides evidence of the effects of other unique variables.

In this study, you plan to explore the price elements for a shared economy-based accommodation lease using a data set with a list of Air B & B. The result is a comprehensive understanding of the pricing elements of the product in the new business model. Based on a limited set of features, including accommodation specifications, the white paper aims to present the best performance model for Air B & B pricing. Get the best results in terms of RMSE using machine learning techniques, including linear regression and neural network and functional importance analysis. Using this methodology, this study provides hidden price-response patterns for real estate leases at different prices. This study contributes to the literature on shared economy by providing a global model that summarizes the price elements of non-traditional accommodation. This study provides insight into how stakeholders such as accommodation rental suppliers can analyze market conditions and improve revenue. The study also designs tools that can be supplied to shared economy-based accommodation rental platforms such as Air B & B and that can be priced according to current pricing factors.Nevertheless, we acknowledge the important limitations of this study. Economic modeling is used to explore data sets and identify associations between different elements and prices. However, social or psychological factors that determine the price of the host are not considered. It is therefore important to carry out a qualitative study to explore the rationale for the pricing of the organizers.

REFERENCES

1. A. Sundararajan, “From Zipcar to the sharing economy,” Harvard Business Review, 2013, Retrieved from https://hbr.org/2013/01/from -zipcar-to-the-sharing-eco/.
2. G. Zervas, D. Proserpio, & J. Byers, “The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry,” Boston U. School of Management Research Paper, (2013-16), 2016, Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2366898
3. N. Pairolero, Assessing the effect of Airbnb on the Washington DC housing market, 2016, Available at http://dx.doi.org/10.2139/ssrn. 2734109.
4. Y. Yang, J. Tang, H. Luo, & R. Law, Hotel location evaluation: A combination of machine learning tools and web GIS. International Journal of Hospitality Management, 2015, 47, 14-24.
5. Y. Yang, N. J. Mueller, & R. R. Croes, Market accessibility and hotel prices in the Caribbean: The moderating effect of quality-signaling factors. Tourism Management, 2016, 56, 40-51.
6. L. Masiero, J. L. Nicolau, & R. Law, A demand-driven analysis of tourist accommodation price: A quantile regression of room bookings. International Journal of Hospitality Management, 2015, 50, 1-8.
   A. Saló, A. Garriga, R. Rigall-I-Torrent, M. Vila, & M. Fluvià, Do implicit prices for hotels and second homes show differences in tourists' valuation for public attributes for each type of accommodation facility? International Journal of Hospitality Management, 36, 2014, 120-129.

AUTHORS PROFILE

Jae Won Choi, Divison of software, College of software, Chungang University, Seoul, The republic of Korea. My interests in studying are data science, artificial inteligence, blockchain, game, etc.