How to succeed as an Airbnb Host
——Evidence from Text analysis and Machine Learning Approach

Literature Review

Yutian Lai

Airbnb, one of the most outstanding marketplace in the sharing economy that facilitates peer-to-peer communication and trust (Airbnb, 2020), provides an alternative accommodation experience for its consumers that cannot be replicated by the conventional hotel industry (Bridges & Vásquez, 2018). With its distinctive operation model and the loosely-regulated sharing-economy condition, it also challenges the theories and practices in consumers' needs identification, which are of great importance in marketing research.

Such a challenge has drawn the attention of many researchers. By analyzing and mining Airbnb data from **insideairbnb.com,** which provides information(including house attributes, consumers' reviews, etc.) about Airbnb's listings around the globe, many researchers have gained unique insight that would help Airbnb refine its business model.

My research project would also employ data from **insideairbnb.com.** I would use datasets of three cities: New York, Tokyo, and London, to identify what customers value most behind the score rating system on Airbnb. More specifically, I want to investigate the attributes that become the driving force for Airbnb properties' ratings, so as to assist both the hosts and Airbnb to condition their service to satisfy guests' needs.

I. Airbnb ratings

Customer review scores are an approximation of the host/property' s performance. It is regarded as essential in business as it helps offer information when consumers make purchasing decisions, thus promoting trade on the Airbnb platform and increasing the trust of buyers. It is also highly valued in academia as it is a good metric to infer the personal and diverse staying experience of consumers (Zervas et al., 2015).

As previous studies on Airbnb ratings suggested, customer review scores are important "signals" consumers refer to when making decisions (Zmud et al., 2010). When the sharing-economy marketplace raises the problem of information asymmetry and lack of trust-building since consumers have difficulty discerning the true quality of properties when the only information source they have is the hosts' descriptions (Tsai & Huang, 2009), antecedent guests' staying experiences become an important indicator that is closely related to the service/quality of the host/property, and can boost trust and affect buyers' final decision (Zmud et al., 2010).

Though significant, the customer review score is not a reliable variable to depend solely on when making purchasing decisions. This is because, according to current literature, that the rating scheme loses its effectiveness in reflecting the intrinsic quality of properties as the rating distribution becomes dramatically skewed towards high scores, possibly owing to lack of standard in accommodation rating, the bias of antecedent ratings and underreporting of negative reviews (Tussyadiah & Zach, 2017; Salganik et al. 2006). Researchers have embarked on the investigation into what is driving behind these rating scores. And the findings imply what attributes are shaping the satisfaction level of consumers are highly context-specific (Aiken & Boush, 2006). One group of researchers find consumers put more emphasis on the practical attributes that relate to the condition of the property, including price, location, amenities and cleanliness, etc. (Bridges & Vásquez, 2018; Tussyadiah & Zach, 2017). The other group of researches argue customers value experiential attributes more, which refer to attributes of the hosts and interaction with the hosts/neighborhood, involving whether the host is local, whether the host is super host, whether the host completes host verification, and host response rate/time, etc. (Xie & Mao, 2017; Festila & Müller, 2017).

Based on preceding research, I aim to incorporate together the factors explored by researchers before and other factors that are beyond current literature (e.g. hosts' age) which might impact review ratings into machine learning algorithms and feature importance analysis, so as to find the decisive factors in consumers' satisfaction-level. Besides, since the literature suggests the importance of "context" (Aiken & Boush, 2006) in determining the ranking of factors that affect the ratings, I would apply the analysis to three different regions(Tokyo, New York, London) and compare the results. Hopefully, this piece of research could add more evidence to the debate of what factors are impacting ratings the most.

II. Host description and text analysis

The advance of social media offers the public the opportunity to be "generators" of text instead of pure "consumers" (Piryani et al., 2017). Airbnb platform displays a tremendous amount of text data written by hosts/consumers for researchers to capture useful insights into the Airbnb business (Piryani et al., 2017). With the advent of corresponding algorithms and technology, such as natural language processing techniques, which move text analysis to a new stage beyond simple statistical analysis (Edwards et al. , 2017), Airbnb text data analysis draws the attention of more and more researchers. They realize text could be supplementary to the analysis of the macro-level attributes as text focuses on the nuanced aspects of Airbnb experiences. For instance, when macro-level analysis claims that “location” would significantly affect review rating, text analysis could answer what specific elements in the "location" attribute comprise its significance and thus helping researchers have a coherent and detailed view of consumers’ accommodation experiences and satisfaction level (Piryani et al., 2017).

Previous studies on Airbnb dataset text mainly focus on analyzing online review comments (Tussyadiah & Zach, 2017; Cheng & Jin, 2019). Researchers believed in the significance and usefulness of online review comments to identity consumers’ accommodation experience and how they make the judgment (Tussyadiah & Zach, 2017). Cheng & Jin (2019) applied text mining and sentiment analysis to Sydney Airbnb review comments and claimed consumers emphasized certain key influencers of their Airbnb experiences, including “host”, “amenities”, and “location.” Host text, on the other hand, attracted insufficient research focus compared with consumer text. Ma et al. (2017) explored how hosts in 12 major U.S. cities could be more trustworthy by strategically disclosing themselves in their profile. This study adopted text length analysis and topic modeling through latent Dirichlet allocation (LDA), aiming to find strategies behind trust-building on Airbnb. Zhang et al. (2018) also extracted textual features from New York host descriptions, and analyzed the information amount, sentiment, readability, and semantic topics of the hosts to find what hosts would like to convey to the consumers.

My research would concentrate on the Airbnb host description analysis. The same as other attributes, the description text influences rating but the link is not fully addressed in past literature. To fill this research gap, I would apply topic modeling through LDA and machine learning models with text input to discover the underlying topics in descriptions to find how hosts with high consumer satisfaction-level are strategically positioning themselves, and hopefully to discover geospatial patterns in host descriptions from the three cities accordingly.

III. Machine learning algorithm used in Airbnb study

One apparent feature of Airbnb data is its “bigness” (Cheng & Edwards, 2019). For instance, in my research, I would deal with three datasets of in total 40,000 properties’ information. Traditional statistical approaches would be unable to capture insights from the data effectively and efficiently (McAbee et al., 2017). Airbnb researchers now become more proficient in using new techniques that are suitable for the representation of "big" data, among which machine learning algorithms are frequently employed (Cheng & Edwards, 2019). Past research suggested researchers mainly used machine learning models for Airbnb price and trust prediction. Kalehbasti et al. (2019) took host/property attributes, and customer reviews as predictors to predict Airbnb price in New York, aiming to aid both the hosts and the guests with pricing and price evaluation in face of the information asymmetry problem (Tsai & Huang, 2009). Zhang et al. (2018) predicted perceived trust in New York based on numeric features (rating score, host response time, number of reviews, etc.), textual features (information readability, amount, etc.), and image features (facial emotions) of the hosts/properties. Both studies adopted multiple machine learning models, including linear regression, support vector regression, tree-based methods, neural networks, and K-means clustering, etc. and analyzed important impacting factors from the best-performing model.

Research using such a computational framework for rating predicting is relatively rare. My research intends to fill the gap, incorporating host/property attributes and host text data as predictors, employing machine learning algorithms used in previous research, to find what factors are significantly affecting review scores.

To conclude, my study contributes to the Airbnb literature through the investigation into the critical but less addressed topic of rating prediction by combining text analysis and machine learning techniques, which have been proved capable of deriving crucial data-driven insights. Through this study, I aim to obtain a holistic and detailed knowledge of the attributes determining the satisfaction-level of consumers and to generate important and useful marketing and academic implications.

References

Aiken, K. D., & Boush, D. M. (2006). Trustmarks, objective-source ratings, and implied investments in advertising: investigating online trust and the context-specific nature of internet signals. *Journal of the Academy of Marketing Science*, *34*(3), 308-323.

Airbnb (2020). About us.
Retrieved
from https://www.airbnb.com.au/about/about-us.

Bridges, J., & Vásquez, C. (2018). If nearly all Airbnb reviews are positive, does that make them meaningless?. *Current Issues in Tourism*, *21*(18), 2057-2075.

Cheng, M., & Edwards, D. (2019). A comparative automated content analysis approach on the review of the sharing economy discourse in tourism and hospitality. *Current Issues in Tourism*, *22*(1), 35-49.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58-70.

Edwards, D., Cheng, M., Wong, I. A., Zhang, J., & Wu, Q. (2017). Ambassadors of knowledge sharing. *International Journal of Contemporary Hospitality Management*.

Festila, M., & Müller, S. (2017). The impact of technology-mediated consumption on identity: The case of Airbnb. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (2019). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. *arXiv preprint arXiv:1907.12665*.

Ma, X., Hancock, J. T., Lim Mingjie, K., & Naaman, M. (2017). Self-disclosure and perceived trustworthiness of Airbnb host profiles. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 2397-2409).

McAbee, S. T., Landis, R. S., & Burke, M. I. (2017). Inductive reasoning: The promise of big data. *Human Resource Management Review*, *27*(2), 277-290.

Piryani, R., Madhavi, D., & Singh, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, *53*(1), 122-150.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *science*, *311*(5762), 854-856.

Tsai, H. T., & Huang, H. C. (2009). Online consumer loyalty: Why e-tailers should seek a high-profile leadership position. *Computers in Human Behavior*, *25*(6), 1231-1240.

Tussyadiah, I. P., & Zach, F. (2017). Identifying salient attributes of peer-to-peer accommodation experience. *Journal of Travel & Tourism Marketing*, *34*(5), 636-652.

Xie, K., & Mao, Z. (2017). The impacts of quality and quantity attributes of Airbnb hosts on listing performance. *International Journal of Contemporary Hospitality Management*.

Zervas, G., Proserpio, D., & Byers, J. (2015). A first look at online reputation on Airbnb, where every stay is above average. *Where Every Stay is Above Average (January 28, 2015)*.

Zhang, L., Yan, Q., & Zhang, L. (2018). A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb. *Decision Support Systems*, *115*, 105-116.

Zmud, R. W., Shaft, T., Zheng, W., & Croes, H. (2010). Systematic differences in firm's information technology signaling: implications for research design. *Journal of the Association for Information Systems*, *11*(3), 1.