Paper 123-2021

Fair Pricing Model for Airbnb in Singapore

Heranshan S/O Subramaniam, Clarence Tay Cheng Wei, Ranice Tan Hui Qi

ABSTRACT

With the rise of the peer to peer (P2P) sharing economy, Airbnb has emerged as a major rival to traditional business in the consumer lodging industry. This study aims to address a significant problem faced by Airbnb host with regards to revenue loss due to inadequate pricing strategy. Though several studies have been conducted, there is still a gap in existing literature with regards to factors influencing Airbnb prices in Singapore. This study used data of Airbnb listings in Singapore from September 2021 to address this by building a fair pricing model for Airbnb in Singapore. Simple linear regression, stepwise forward regression, stepwise backward regression, and K-Nearest Neighbour (KNN) models were built and evaluated. The stepwise forward regression yielded the optimum predictive model with an adjusted R-square of 0.60. Number of bedrooms, room type, location and the presence of a pool were found to be the major factors influencing price. Update of the model with post pandemic data, inclusion of other fixed costs such a cleaning and miscellaneous fees and inclusion of host attributes and customer ratings were identified as areas of future work.

INTRODUCTION

The rise of the sharing economy has created new business models competing with traditional businesses. The sharing economy is defined by the Oxford dictionary as "an economic system in which assets or services are shared between private individuals, typically by means of the internet."[1]. In the accommodation industry this has given rise to P2P rental of short-term accommodations, which were found to be used exclusively as a substitute for existing accommodations, primarily hotels (Guttentag & Smith, 2017). Airbnb is a major player in the P2P accommodation industry, making up approximately 20 percent of consumer lodging expense based on US data (Molla, 2019). P2P accommodation places the burden of pricing on the hosts. This has been identified as a major challenge plaguing Airbnb hosts, who may face revenue loss attributed to inadequate pricing strategy. A study found that adopting price positioning and dynamic pricings yielded positive effects on a listing's revenue performance (Kwok & Xie, 2019). Though several studies have been conducted to evaluate the price influencers for Airbnb rental pricing, there has been a lack of studies focused on Singapore. As such, this study aims to fill that gap by establishing a predictive model that would aid the host in adopting a market calibrated price and enhance their listing's revenue performance.

LITERATURE REVIEW

Several studies have been conducted on the Airbnb business model, with many focused on the price determining factors of the rental accommodations it provides. A study based on three first-tier cities in China found the top five determinants of pricing to be room type, city, distance to tourist attractions, number of pictures posted, and number of amenities provided (Chang & Li, 2021). The study analyzed variables spanning across five categories, namely listing attributes, listing location, host attributes, rental policies and listing reputation. The regression model attained in the study had an adjusted R-square of 0.2072.

Another study spanning across thirty-three cities in thirteen countries of three continents found complexities in the price-determinant relationship for accommodations in the sharing economy (Wang & Nicolau, 2017). The study evaluated 25 variables spanning across the same five categories as the previous study. Ordinary least square found that 24 of the 25 variables were good predictors for price, while quantile regression found that all variables had significant influence on price. Host attributes such as superhost status, larger number of listings and verified host identities led to higher prices. Location was a strong price determinant, with listing attributes, amenities, rental policies, and reviews also significantly influencing prices. Larger accommodation capacity, more bathrooms and bedrooms were associated with higher prices. Customer ratings were deemed a powerful price influencer, with higher average ratings leading to higher prices. The study did not include any social or psychological factors that may have influenced pricing.

METHODOLOGY

DATASET

The dataset used is retrieved from the Inside Airbnb database[2], where the data of Airbnb listings in Singapore was extracted. The data was updated in September 2021, and consists of the location of the listings, host information and

[1] https://www.oxfordlearnersdictionaries.com/definition/english/sharing-economy

[2]http://insideairbnb.com/get-the-data.html

activity, property type, characteristic and amenities, as well as some ratings and reviews.

There are a total of 4221 listings from the data, with price range from $0 to $10286 per night. To prevent distortion of statistical analysis, the outliers in the data were excluded, and the final analysis considers data that cost less than $300 per night, accounting for over 90% of listings.

All data reprocessing, exploratory data analysis, prediction modelling and evaluation were conducted using SAS JMP PRO v16.

DATA PREPARATION

The raw data was extracted and reviewed against the data dictionary provided by Inside Airbnb[3] (Appendix A). The data change log is accessible in Appendix C. The metadata was first explored to ensure all data are in the correct form. Variables that were in the wrong form and modelling type i.e. host_id, host_response_rate and host_acceptance_rate were corrected. Next, the data that are irrelevant or redundant to the purpose of the analysis were removed. This includes personalized descriptions, IDs, URLs, photos or repeated categorical variables that are related to the listing. Some data such as availability in the coming months were also removed in consideration of the covid-19 measures in place.

Variables with significant proportion of missing values were identified using the column viewers and distribution functions. Bathrooms variable which was missing 100% of the data was excluded. All the review scores, which can be used to determine the quality of the listing and enjoyability of previous experiences, were missing 45% of the data and were removed as well. The host response time, rate and acceptance rate which were missing 20-25% of data were also removed.

A univariate data analysis was conducted for the remaining variables to explore the distribution and shape of the data. The price data was left-skewed, containing several outliers. Hence, the rows were filtered to include listings with price less than $300, accounting for over 90% of the data. Variables such as number_of_reviews and individual calculated_host_listings_count were unevenly skewed to the left, with majority having 0 counts. Eventually, the data was removed as without the review score, the number_of_reviews alone may be insufficient to determine the quality of a listing. The individual calculated_host_listings_count was also removed and represented using the total_host_listings_count instead.

Variables with excessive distinct values were also regrouped where appropriate. This includes the bathrooms_text variable which contained 45 categories which was regrouped into 2 categories – ‘Private’ or ‘Shared’. The amenities contained a substantial 3109 categories because they were unique to each listing. Text frequency analysis was first used to identify 16 popular amenities such as TV, pool, Wi-Fi, air-conditioning and kitchen. New columns were inserted for each amenity and formula was then applied to produce Boolean results based on the list of amenities provided. The neighbourhood_cleansed consisted of 44 levels, with some neighbourhoods containing less than 50 counts such as Sungei Kadut, Choa Chu Kang etc. To prevent distortion of the analysis, these neighbourhoods were excluded.

Next, a multivariate analysis was conducted for the continuous variables to identify highly correlated variables that can be interchanged. For each pair of variables that showed high correlations with each other, one was removed and the other was kept in the analysis. This includes host_neighbourhood – neighbourhood_cleansed pair, host_listings_counts – host_total_listings_count and maximum_nights – minimum_maximum_nights etc.

Based on the remaining variables, rows which contained empty cells were removed from the analysis. Lastly, a validation column was added to split the data into training, validation and test set, stratified by the price column. The metadata of the final dataset used can be found in Appendix B.

FINDINGS AND DISCUSSIONS

LINEAR REGRESSION

Understanding that there is more than one variable that affects the rental pricing, a multiple linear regression (MLR) technique was employed to model the linear relationship between the independent variables and the dependent variable, which is an extension of the ordinary least-square regression method.

[3]https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=982310896

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

$y_i =$ dependent variable

$x_i =$ explanatory variables

$\beta_0 =$ y-intercept (constant term)

$\beta_p =$ slope coefficients for each explanatory variable

$\epsilon =$ the model's error term (also known as the residuals)

Figure 1: Linear Regression Formula

The variables are fed into the Fit Model function, with Standard Least Squares chosen in the "Personality" to run the model. From the Effect Summary below, it is observed that a few of the variables has a larger influence on the dependent variable. Eg. Number of bedrooms, location, room type, availability of swimming pool etc.

Table 1: Linear Regression Effect Summary

Source	LogWorth	PValue
bedrooms	20.995	0.00000
neighbourhood_cleansed	14.939	0.00000
room_type	13.241	0.00000
Pool	6.694	0.00000
minimum_nights	5.002	0.00001
accommodates	4.251	0.00006
Kitchen	3.884	0.00013
Dryer	3.536	0.00029
host_response_time	3.315	0.00048
TV	3.145	0.00072
Washer	2.832	0.00147
bathroom_type	2.710	0.00195
host_total_listings_count	2.691	0.00204
Hot Water	2.494	0.00320
Lock	2.101	0.00792
beds	1.786	0.01636
Workspace	1.703	0.01981
Refrigerator	1.103	0.07892
Iron	0.990	0.10233
Microwave	0.621	0.23933
Hair Dryer	0.548	0.28316
Parking	0.432	0.36980
Essentials (Toilet Paper, Soap, Towel, Pillow, Linens)	0.224	0.59686
Air Con	0.194	0.63926
maximum_nights	0.094	0.80572
host_is_superhost	0.039	0.91418
Wifi	0.030	0.93223

Profiler

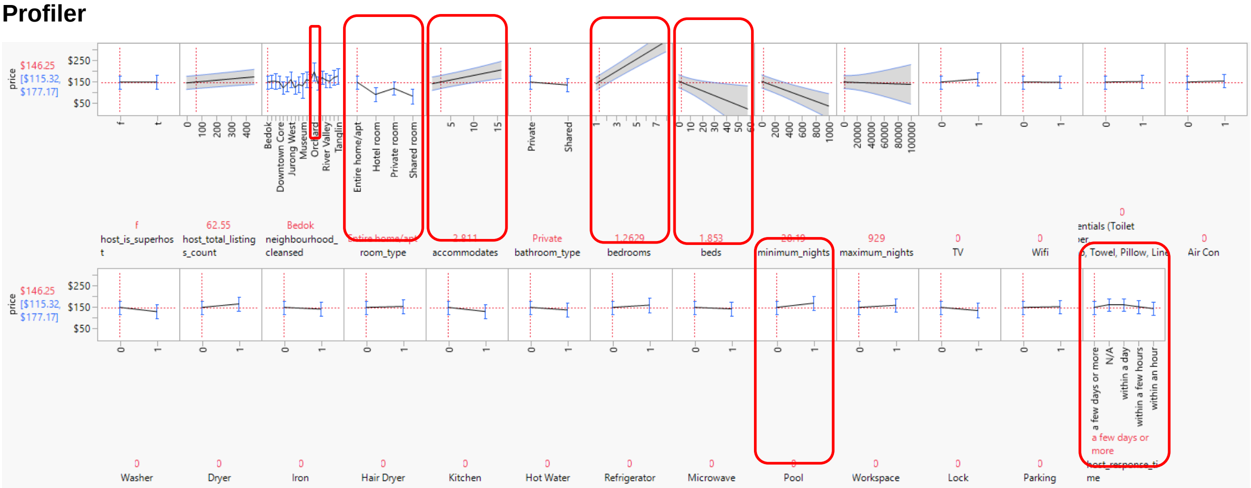

Figure 2: Linear Regression Prediction Profiler

The positive/negative effect can also be seen in the Prediction Profiler. As shown, a place with a larger number of bedrooms, located in Orchard area, rented out as an entire apartment, equipped with TV, Dryer and with swimming pool, would likely lead to a higher rental price. On the other hand, a place with a larger minimum nights required, located at Geylang or Kallang, rented out as a shared place would drive a negative effect on the rental price.

It is interesting to note that a higher number of beds leads to a lower rental price. Upon closer inspection into the data, it is found that these listings belonged to hostels located mainly in Kallang which they might have mistakenly keyed in their capacity instead 1 bed for their listed price.

There were also some variables that were not very intuitive, eg. for host response time, it seems like a faster response leads to a lower pricing. This could be due to the nature that cheaper rentals (likely to be hostels) are more responsive as a business, hence leading to this correlation.

Table 2: Linear Regression Model Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6175	42.826	950
Validation Set	0.5722	45.344	952
Test Set	0.6072	43.451	951

The RSquare of the Test Set is 0.6072, which matches quite closely to the Training Set RSquare of 0.6175.

FORWARD STEPWISE LINEAR REGRESSION

In our next regression method, the same variables from MLR were also fed into the Fit Model with Stepwise chosen in the "Personality" to run the model. The Stopping Rule was set as "Max Validation RSquare" with Direction set as "Forward".

With this, the model starts with the intercept while adding and removing the various independent variables in a step-by-step manner. In this process, significant variables were added, and the process would stop when the maximum RSquare is found for the validation set.

Table 3:3 Forward Stepwise Linear Regression Effect Summary

Source	LogWorth	PValue
bedrooms	20.457	0.00000
room_type{Shared room&Private room&Hotel room-Entire home/apt}	14.074	0.00000
neighbourhood_cleansed{Jurong West&Kallang&Geylang&Outram&Museum&Rochor&Bukit Merah&Newton&Bedok&Singapore River&Marine Parade-Jurong East&Downtown Core&River Valley&Queenstown&Novena&Clementi&Tanglin&Orchard}	7.427	0.00000
Pool	7.331	0.00000
accommodates	5.089	0.00001
minimum_nights	5.043	0.00001
Kitchen	4.750	0.00002
TV	4.196	0.00006
neighbourhood_cleansed{Jurong East&Downtown Core&River Valley-Queenstown&Novena&Clementi&Tanglin&Orchard}	3.799	0.00016
Dryer	3.636	0.00023
neighbourhood_cleansed{Jurong West-Kallang}	3.390	0.00041
neighbourhood_cleansed{Jurong West&Kallang&Geylang-Outram&Museum&Rochor&Bukit Merah&Newton&Bedok&Singapore River&Marine Parade}	3.361	0.00044
bathroom_type	3.206	0.00062
Washer	2.871	0.00135
room_type{Private room-Hotel room}	2.778	0.00167
host_response_time{within an hour&a few days or more-N/A}	2.628	0.00235
host_total_listings_count	2.545	0.00285
Hot Water	2.495	0.00320
neighbourhood_cleansed{Jurong West&Kallang-Geylang}	2.447	0.00358
Lock	2.393	0.00405
beds	2.109	0.00778

Source	LogWorth	PValue
Workspace	1.999	0.01002
room_type{Shared room-Private room&Hotel room}	1.819	0.01517
neighbourhood_cleansed{Queenstown&Novena&Clementi-Tanglin&Orchard}	1.631	0.02337
neighbourhood_cleansed{Outram&Museum-Rochor&Bukit Merah&Newton&Bedok&Singapore River&Marine Parade}	1.523	0.02999
neighbourhood_cleansed{Queenstown-Novena&Clementi}	1.342	0.04550
host_response_time{within an hour&a few days or more&N/A-within a day&within a few hours}	0.598	0.25246

The variables with a larger influence are again bedrooms, room type, specific locations (neighbourhood), availability of swimming pool and etc, which are very similar to the ones produced by previous MLR method. The notable difference was the locations were further broken down into multiple variables, indicating that not all locations are equal in their impact to the dependent variable.

Also, the variables that were previously not that influential in the MLR method (eg. Availability of an iron, WIFI, parking; lower LogWorth variables) were excluded in this stepwise model.

Profiler

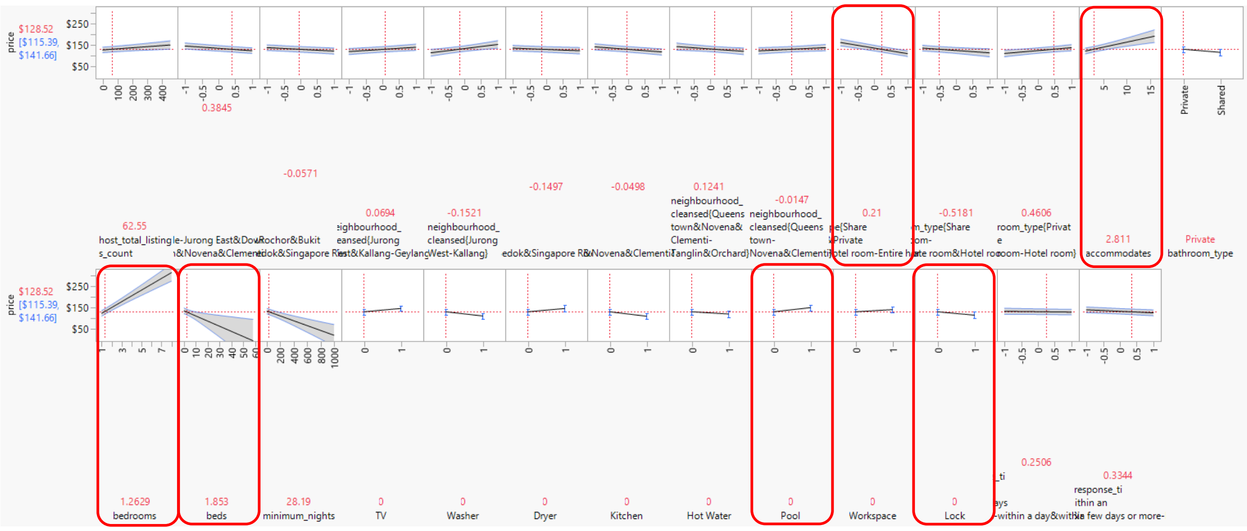

Figure 3: Forward Stepwise Linear Regression Prediction Profile

Another notable observation is that rentals with locks provided seems to drive a lower pricing, which could be attributed to the fact that most of the cheap rentals (likely hostels) tend to provide locks, hence the correlation in the model.

Table 4:4 Forward Stepwise Linear Regression Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6100	43.248	950
Validation Set	0.5728	45.311	952
Test Set	0.6053	43.559	951

The RSquare of the Test Set is 0.6053, which also matches quite closely to the Training Set RSquare of 0.6100.

BACKWARD STEPWISE LINEAR REGRESSION

Like the Stepwise (Forward) method, in the Stepwise (Backward) method, the same variables from and Personality were chosen in the Fit model, with exception of the Direction being set as “Backward” and with all effects added first into the model.

With this, the model starts with all the variables (intercept locked) while removing the various independent variables in a step-by-step manner. In this process, insignificant variables were removed until further removal of variables does not produce the maximum RSquare.

Table 5:5 Backward Stepwise Linear Regression Effect Summary

Source	LogWorth		PValue
bedrooms	21.197		0.00000
room_type{Shared room&Private room&Hotel room-Entire home/apt}	14.605		0.00000
Pool	7.092		0.00000
neighbourhood_cleansed{Jurong West&Kallang&Geylang&Outram&Museum&Rochor&Bukit Merah&Newton&Bedok&Singapore River&Marine Parade-Jurong East&Downtown Core&River Valley&Queenstown&Novena&Clementi&Tanglin&Orchard}	5.481		0.00000
minimum_nights	5.254		0.00001
accommodates	4.480		0.00003
neighbourhood_cleansed{Jurong East&Downtown Core&River Valley-Queenstown&Novena&Clementi&Tanglin&Orchard}	3.979		0.00011
Kitchen	3.974		0.00011
Dryer	3.669		0.00021
TV	3.373		0.00042
neighbourhood_cleansed{Jurong West-Kallang}	3.067		0.00086
neighbourhood_cleansed{Jurong West&Kallang&Geylang-Outram&Museum&Rochor&Bukit Merah&Newton&Bedok&Singapore River&Marine Parade}	3.040		0.00091
room_type{Private room-Hotel room}	3.021		0.00095
Washer	2.906		0.00124
host_total_listings_count	2.786		0.00164
bathroom_type	2.730		0.00186
Hot Water	2.564		0.00273
host_response_time{within an hour&a few days or more-N/A}	2.554		0.00279
neighbourhood_cleansed{Jurong West&Kallang-Geylang}	2.511		0.00308
Lock	2.148		0.00711
room_type{Shared room-Private room&Hotel room}	1.873		0.01340
beds	1.822		0.01507
Workspace	1.821		0.01509
neighbourhood_cleansed{Queenstown&Novena&Clementi-Tanglin&Orchard}	1.746		0.01795
neighbourhood_cleansed{Outram&Museum-Rochor&Bukit Merah&Newton&Bedok&Singapore River&Marine Parade}	1.501		0.03152
neighbourhood_cleansed{Queenstown-Novena&Clementi}	1.387		0.04102
neighbourhood_cleansed{Singapore River-Marine Parade}	1.227		0.05935
Refrigerator	1.101		0.07922
Iron	0.923		0.11931
host_response_time{within a day-within a few hours}	0.904		0.12462
Hair Dryer	0.628		0.23566
neighbourhood_cleansed{Jurong East-Downtown Core&River Valley}	0.623		0.23806
Microwave	0.548		0.28284
neighbourhood_cleansed{Tanglin-Orchard}	0.529		0.29612
Parking	0.508		0.31077
neighbourhood_cleansed{Newton-Bedok}	0.486		0.32680
host_response_time{within an hour-a few days or more}	0.443		0.36033
host_response_time{within an hour&a few days or more&N/A-within a day&within a few hours}	0.331		0.46621
neighbourhood_cleansed{Rochor&Bukit Merah-Newton&Bedok&Singapore River&Marine Parade}	0.121		0.75723
neighbourhood_cleansed{Newton&Bedok-Singapore River&Marine Parade}	0.023		0.94837

The variables with a larger influence are very similar to the ones produced by previous Stepwise (Forward) method. However, this model contains more variables compared to the forward model, which can be attributed to the nature of its conservative method, where it starts with a full model before trimming out the variables.

Profiler

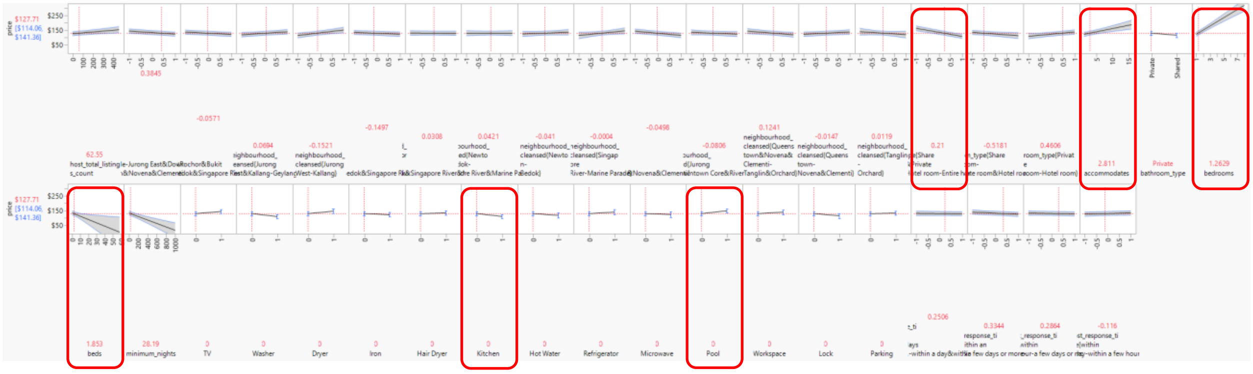

Figure 4: Backward Stepwise Linear Prediction Profiler

Like the previous model where there are some variables that are not intuitive, e.g in this model, having a kitchen or washer ends up with a lower pricing. Again, this can be attributed to the nature that places that tend of have these listed down in their description are likely the cheap hostel rentals, compared to other types of rentals (public/private housing). One learning from this data analysis is that the effectiveness of the regression also depends on how detailed the data is being populated by the host.

*Table 6:*6 *Backward Stepwise Linear Regression Crossvalidation*

Source	RSquare	RASE	Freq
Training Set	0.6168	42.868	950
Validation Set	0.5733	45.283	952
Test Set	0.6081	43.403	951

The RSquare of the Test Set is 0.6081, which matches quite closely to the Training Set RSquare of 0.6168.

K-NEAREST NEIGHBOURS REGRESSION

The K Nearest Neighbours (KNN) method, a supervised learning algorithm, attempts to classify the data based on a similarity measure, by grouping the test data to its nearest neighbours (trained data) by Euclidean distance via a majority voting process, which is influenced by the K-value (the number of nearest neighbours to include in the voting process).

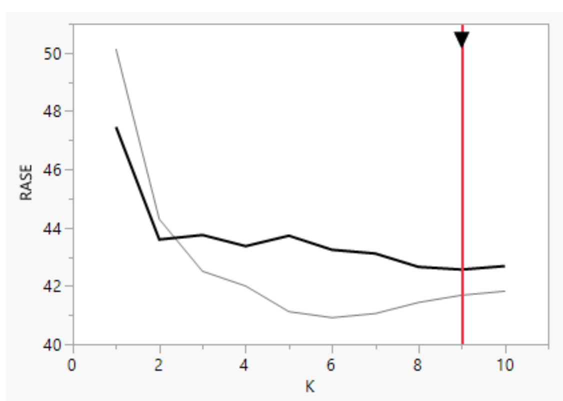

Figure 5: RASE vs K for K-nearest neighbours regression model

Through the cross-validation, the k value is determined to be 9 (odd number; ideal to avoid tie in voting process), which means that for a test data to be classified into certain groups, it must win majority vote within its 9 nearest neighbours. A larger K value does help to smoother the decision boundaries compared to a smaller K value (potentially noisier and will impose higher influence on the outcome).

Table 7:7 K-Nearest Neighbours Model Crossvalidation

DATA_SAMPLING	Predictor	Creator	RSquare
Training	K-NN (9)	K Nearest Neighbors	0.7176
Test	K-NN (9)	K Nearest Neighbors	0.6535
Validation	K-NN (9)	K Nearest Neighbors	0.6233

The RSquare for the Test data works out to be 0.6233, with the highest RSquare value among all the models attempted.

COMPARISON BETWEEN THE MODELS

Performance

Table 8:8 Model Comparison

DATA_SAMPLING	Predictor	Creator	RSquare	RASE	AAE	Freq
Training	K-NN (9)	K Nearest Neighbors	0.7176	36.823	26.195	953
Test	K-NN (9)	K Nearest Neighbors	0.6535	40.793	29.097	952
Validation	K-NN (9)	K Nearest Neighbors	0.6233	42.533	29.877	953
Training	Linear Pred Price	Fit Least Squares	0.6175	42.826	31.028	950
Training	Stepwise (Backward)	Fit Least Squares	0.6168	42.868	31.162	950
Training	Stepwise (Forward)	Fit Least Squares	0.6100	43.248	31.529	950
Test	Stepwise (Backward)	Fit Least Squares	0.6081	43.403	32.183	951
Test	Linear Pred Price	Fit Least Squares	0.6072	43.451	32.245	951
Test	Stepwise (Forward)	Fit Least Squares	0.6053	43.559	32.304	951
Validation	Stepwise (Backward)	Fit Least Squares	0.5733	45.283	33.138	952
Validation	Stepwise (Forward)	Fit Least Squares	0.5728	45.311	33.134	952
Validation	Linear Pred Price	Fit Least Squares	0.5722	45.344	33.202	952

Table 9: Adjusted Rsquare comparison

Adjusted R^2		
Linear	Stepwise (Forward)	Stepwise (Backward)
0.597	0.599	0.600

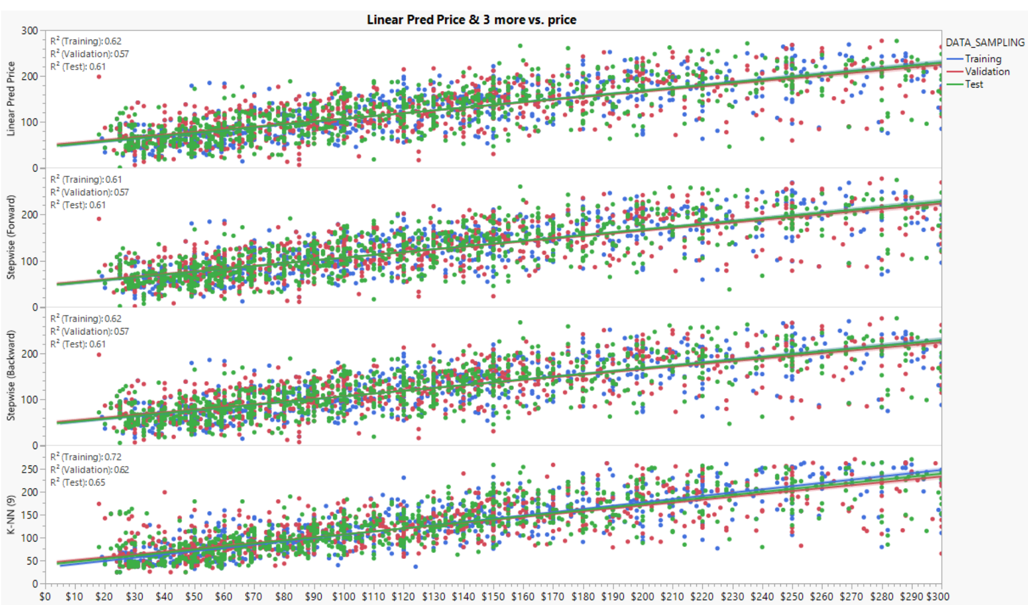

Figure 6: Predicted Price vs. Actual Price for all 4 models

Based on the square root of the mean squared prediction error (RASE), K-NN yielded the best model with the smallest RASE for training, validation, and test. The K-NN model also had the best R-square results, a R-square of 0.65 for test dataset, well above the other models. However, the K-NN model also has the largest difference in R-square between the training and test datasets (-0.06). In comparison, the linear and stepwise regression models have a higher RASE and a lower R-square of about 0.60, but only a R-square difference of – 0.004 to – 0.009 between the training and test datasets. This may be a sign of overfitting of the K-NN dataset to the training data. Therefore, the linear models may be more suitable in this case. Amongst the linear models, the stepwise (backward) regression has the highest adjusted R^2, followed by the stepwise (forward) and the linear model.

Variables

For the linear model, most of the variables make sense and are in line with the other models, except for host_response_time, whereby the price of listings of hosts that responds the fastest tend to be lower than that of those who a few days to response.

The variables that deviate from the other models in the stepwise (forward) regression model are the washer, kitchen and lock amenity, whereby listings with these costs less. Listings without lock tend to cost $15 per night cheaper than those with lock. This may be because lock may be an attribute related to private or shared rooms, which are typically priced lower than entire house or apartments. For washer and kitchen, more investigation may be required as it may need to be separated between shared of private amenity.

In the stepwise (backward) model, a few amenities deviate from the other models and expectation. These include washer, iron, kitchen, hot water, microwave and lock, whereby the listings with these amenities cost $20, $8, $20, $12 $6 and $14 cheaper per night respectively than listings without them.

Complexity

Based on the complexity, the linear regression models are generally less complex than the K-NN model. Amongst the 3 linear models, the least complex is the linear model followed by the stepwise (forward) regression and the stepwise (backward) regression model, which also contains the most variables.

Hence, by balancing the measures used to determine the optimum model, the stepwise (forward) regression model seems to give a good balance between performance, variables, and complexity.

FACTORS OF IMPORTANCE DRIVING AIRBNB PRICES IN SINGAPORE

From the stepwise forward linear regression, the factors with the largest logworth were number of bedrooms, room type, neighbourhood and pool. The other factors had a significantly lower logworth. The observations made are in line with the studies reviewed in the literature review, which indicated that location and accommodation size were strong influences of pricing (Wang & Nicolau, 2017).

Number Of Bedrooms/Accommodation Size

Larger number of bedrooms led to a higher price; An extra bedroom increases listing price of about $27 per night and an additional accommodation size is priced at about $5 per size per night.

Room Type

For room type, entire homes and apartments had the greatest positive influence on price. From the model, hosts charge additional ~$52 per night for entire apartments compared to shared/private rooms/hotel accommodations.

Neighbourhood

Accommodations in the city area had a positive influence on price as well. Listings in prime location such as Orchard, Tanglin, River Valley and the Downtown Core typically cost additional $22 per night compared to its surrounding neighbourhood in Kallang, Newton, Outram and Rochor.

Amenities

Out of all the amenities included, pool had the strongest influence in price. Based on the model, prices typically differ by $15, $15, and $10 per night for the presence of common amenities such as TV, dryer, and workspace respectively. Hosts with pools in their apartment are also able to charge additional $20 per night.

CONCLUSION

The optimal fair-pricing prediction model achieved in terms of performance, complexity and variables was using the forward stepwise linear regression model, achieving an R-square of 0.61. With a predictive model for fair pricing of Airbnb in Singapore, hosts may be better able to adopt a market calibrated price and enhance their listing's revenue performance. At the same time, people who intend to rent accommodations from Airbnb, may also use the fair pricing model ensure that they are fairly compensated for the rates they are being charged.

RECOMMENDATION AND FUTURE WORK

As the data was updated in September 2021, where tourism in the city was heavily affected by the COVID-19 pandemic, the prices may not be representative of prices after the tourism industry recovers. Hence, this model needs to be further updated and improved with latest data to be more effective in its pricing strategy. The amenities data may also need to be fine-tuned to be better representative of the apartments facilities such as by providing a pre-determined list instead of having the hosts list it out by themselves.

The prices of the listings are not inclusive of fixed costs like the cleaning and miscellaneous fees, which may account for a substantial percentage of the accommodation cost. Hence, it may not be representative of the actual prices that is being charged to the renter per night. To improve the model from a customer's perspective, including the additional fixed costs may be more representative to aid a prospective renter in choosing a better priced listing.

The prediction of the model was also unable to sufficiently account for host attributes and customer ratings, which are deemed a powerful price influencer. This may be due to either lack of reviews of the large amount of missing data from the original dataset. It would be recommended to include these data after more has reviews and scores have been collected, so that a robust predictive model can be better established.

REFERENCES

Chang , C., & Li, S. (2021). Study of Price Determinants of Sharing Economy-Based Accomodation Services: Evidence from Airbnb.com. *Journal of Theoretical and Applied Electronic Commerce Research, 16*, 584-601. doi:https://doi.org/10.3390/jtaer16040035

Guttentag, D. A., & Smith, S. L. (2017). Assessing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations. *International Journal of Hospitality Management, 64*, 1-10. doi:10.1016/j.ijhm.2017.02.003

Kwok, L., & Xie, K. L. (2019). Pricing strategies on Airbnb: Are multi-unit hosts revenue pros? *International Journal of Hospitality Management, 82*, 252-259. doi:https://doi.org/10.1016/j.ijhm.2018.09.013

Molla, R. (2019, March 25). *American Consumer Spent More on Airbnb than on Hilton Last Year*. Retrieved from Vox: https://www.vox.com/2019/3/25/18276296/airbnb-hotels-hilton-marriott-us-spending

Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management, 62*, 120-131. doi:http://dx.doi.org/10.1016/j.ijhm.2016.12.007

APPENDIX

APPENDIX A: METADATA OF ORIGINAL DATASET

Name	Label	Role	Level
id	Airbnb ID	ID	continuous
listing_url	Listing URL	URL	nominal
scrape_id	Scraping ID	ID	continuous
last_scraped	Date last scraped (updated)	Date	continuous
name	Name of listing	Text	nominal
description	Description of listing	Text	nominal
neighborhood_overview	Description of neighbourhood	Text	nominal
picture_url	URL of picture	URL	nominal
host_id	ID of host	ID	continuous
host_url	URL of host profile	URL	nominal
host_name	Name of host	Text	nominal
host_since	Date host started hosting	Date	continuous
host_location	Location of host	Text	nominal
host_about	Description of host	Text	nominal
host_response_time	Time host takes to response	Text	nominal
host_response_rate	Rate host takes to response	number	nominal

host_acceptance_rate	Rate host takes to acceptance	number	nominal
host_is_superhost	Is host a superhost?	Text	nominal
host_thumbnail_url	URL of host thumbnail	URL	nominal
host_picture_url	URL of host picture	URL	nominal
host_neighbourhood	Neighbourhood of host	Text	nominal
host_listings_count	Listings of host	number	continuous
host_total_listings_count	Total listings of host	number	continuous
host_verifications	Verification method of hosts	Text	nominal
host_has_profile_pic	Does host have profile picture	Text	nominal
host_identity_verified	Has host identity been verified	Text	nominal
neighbourhood	Neighbourhood of listing	Text	nominal
neighbourhood_cleansed	Neighbourhood of listing (cleaned)	Text	nominal
neighbourhood_group_cleansed	Neighbourhood Region (cleaned)	Text	nominal
latitude	Latitude	number	continuous
longitude	Longitude	number	continuous
property_type	Type of property (Private Room/Entire Unit)	Text	nominal
room_type	Room type	Text	nominal
accommodates	Number of guests accomodated	number	continuous
bathrooms	No of bathrooms		nominal
bathrooms_text	No of bathrooms and type	Text	nominal
bedrooms	No of bedrooms	number	continuous
beds	No of beds	number	continuous
amenities	Types of amenities included	Text	nominal
price	Price per night	number	continuous
minimum_nights	Minimum nights	number	continuous
maximum_nights	Maximum nights	number	continuous
minimum_minimum_nights	Minimum nights	number	continuous
maximum_minimum_nights	Minimum nights	number	continuous
minimum_maximum_nights	Maximum nights	number	continuous
maximum_maximum_nights	Maximum nights	number	continuous
minimum_nights_avg_ntm	Minimum nights	number	continuous
maximum_nights_avg_ntm	Maximum nights	number	continuous
calendar_updated	Calendar updated	Text	nominal
has_availability	Availability	Text	nominal
availability_30	Availaibility in next 30 days	number	continuous
availability_60	Availability in next 60 days	number	continuous
availability_90	Availability in next 90 days	number	continuous
availability_365	Availability in next 365 days	number	continuous
calendar_last_scraped	Calendar last updated	number	continuous
number_of_reviews	Total number of reviews	number	continuous
number_of_reviews_ltm	Number of review ltm	number	continuous
number_of_reviews_l30d	Number of review l30d	number	continuous
first_review	Date of first review	number	continuous
last_review	Date of last review	number	continuous
review_scores_rating	Overall rating	number	continuous
review_scores_accuracy	Accuracy score	number	continuous
review_scores_cleanliness	Cleanliness score	number	continuous
review_scores_checkin	Check-in score	number	continuous
review_scores_communication	Communication score	number	continuous
review_scores_location	Location score	number	continuous
review_scores_value	Value score	number	continuous
license	License available		nominal
instant_bookable	Is location instant bookable	Text	nominal

calculated_host_listings_count	Host listing counts	number	continuous
calculated_host_listings_count_entire_homes	Host listing counts for entire homes	number	continuous
calculated_host_listings_count_private_rooms	Host listing counts for private rooms	number	continuous
calculated_host_listings_count_shared_rooms	Host listing counts for shared rooms	number	continuous
reviews_per_month	Reviews per month	number	continuous

APPENDIX B: METADATA OF FINAL DATASET

Name	Label	Original/ Calculated	Role	Level
host_is_superhost	Is host a superhost?	Original	Text	nominal
host_response_time	Time host takes to response	Original	Text	nominal
host_total_listings_count	Total listings of host	Original	Number	continuous
neighbourhood_cleansed	Neighbourhood of listing (cleaned)	Original	Text	nominal
room_type	Room type	Original	Text	nominal
accommodates	Number of guests accomodated	Original	Number	continuous
bathroom_type	Whether bathroom is shared or private	Calculated	Text	nominal
bedrooms	No of bedrooms	Original	Number	continuous
beds	No of beds	Original	Number	continuous
minimum_nights	Minimum nights	Original	Number	continuous
maximum_nights	Maximum nights	Original	Number	continuous
TV	Presence of TV	Calculated	Number	nominal
Wifi	Presence of wifi amenity	Calculated	Number	nominal
Essentials	Presence of essentials amenity	Calculated	Number	nominal
Air Con	Presence of air conditioning	Calculated	Number	nominal
Washer	Presence of washer	Calculated	Number	nominal
Dryer	Presence of dryer	Calculated	Number	nominal
Iron	Presence of iron	Calculated	Number	nominal
Hair Dryer	Presence of hair dryer	Calculated	Number	nominal
Kitchen	Presence of kitchen	Calculated	Number	nominal
Hot Water	Presence of hot water	Calculated	Number	nominal
Refrigerator	Presence of refrigerator	Calculated	Number	nominal
Microwave	Presence of microwave	Calculated	Number	nominal
Pool	Presence of pool	Calculated	Number	nominal
Workspace	Presence of workspace	Calculated	Number	nominal
Lock	Presence of lock	Calculated	Number	nominal
Parking	Presence of parking	Calculated	Number	nominal
price	Price per night	Original	Number	continuous

APPENDIX C: DATA PREPARATION CHANGE LOG

No.	Variable Name	Issue	Action
1	host_id	Data should not be in continuous modelling type.	Change to nominal data type.
2	host_response_rate	Data should not be character and nominal modelling type.	Change to numeric and continuous modelling type.
3	host_acceptance_rate	Data should not be character and nominal modelling type.	Change to numeric and continuous modelling type.
4	id	Data irrelevant to analysis.	Hide and exclude data.
5	last_scraped	Data irrelevant to analysis.	Hide and exclude data.
6	listing_url	Data irrelevant to analysis.	Hide and exclude data.
7	host_url	Data irrelevant to analysis.	Hide and exclude data.
8	host_name	Data irrelevant to analysis.	Hide and exclude data.
9	host_location	Data irrelevant to analysis.	Hide and exclude data.

10	scrape_id	Data irrelevant to analysis.	Hide and exclude data.
11	picture_url	Data irrelevant to analysis.	Hide and exclude data.
12	host_verifications	Data irrelevant to analysis.	Hide and exclude data.
13	host_has_profile_pic	Data irrelevant to analysis.	Hide and exclude data.
14	host_thumbnail_url	Data irrelevant to analysis.	Hide and exclude data.
15	host_picture_url	Data irrelevant to analysis.	Hide and exclude data.
16	host_neighbourhood	Data irrelevant to analysis.	Hide and exclude data.
17	host_identity_verified	Data irrelevant to analysis.	Hide and exclude data.
18	latitude	Data irrelevant to analysis.	Hide and exclude data.
19	longitude	Data irrelevant to analysis.	Hide and exclude data.
20	has_availability	Data irrelevant to analysis considering data was taken during covid-19 period.	Hide and exclude data.
21	availability_30	Data irrelevant to analysis considering data was taken during covid-19 period.	Hide and exclude data.
22	availability_60	Data irrelevant to analysis considering data was taken during covid-19 period.	Hide and exclude data.
23	availability_90	Data irrelevant to analysis considering data was taken during covid-19 period.	Hide and exclude data.
24	availability_365	Data irrelevant to analysis considering data was taken during covid-19 period.	Hide and exclude data.
25	neighbourhood	Data irrelevant to analysis.	Hide and exclude data.
26	calendar_updated	Data irrelevant to analysis.	Hide and exclude data.
27	calendar_last_scraped	Data irrelevant to analysis.	Hide and exclude data.
28	license	Data irrelevant to analysis.	Hide and exclude data.
29	name	Data irrelevant to analysis.	Hide and exclude data.
30	neighbourhood_overview	Data irrelevant to analysis.	Hide and exclude data.
31	host_about	Data irrelevant to analysis.	Hide and exclude data.
32	host_id	Data irrelevant to analysis.	Hide and exclude data.
33	host_since	Data irrelevant to analysis.	Hide and exclude data.
34	first review	Data irrelevant to analysis.	Hide and exclude data.
35	last review	Data irrelevant to analysis.	Hide and exclude data.
36	description	Data irrelevant to analysis.	Hide and exclude data.
37	instant_bookable	Data irrelevant to analysis.	Hide and exclude data.
38	number_of_reviews_ltm	Data irrelevant to analysis considering data was taken during covid-19 period.	Hide and exclude data.
39	number_of_review_l30d	Data irrelevant to analysis considering data was taken during covid-19 period.	Hide and exclude data.
40	host_neighbourhood	High correlatiion with neighbourhood_cleansed	Hide and exclude data. Keep neighbourhood_cleansed.
41	host_listing_counts	High correlatiion with host_total_listing_counts	Hide and exclude data. Keep host_total_listing_counts.
42	minimum_minimum_nights	High correlation with minimum_nights	Hide and exclude data. Keep minimum_nights
43	maximum_minimum_nights	High correlation with minimum_nights	Hide and exclude data. Keep minimum_nights
44	minimum_avg_ntm	High correlation with minimum_nights	Hide and exclude data. Keep minimum_nights
45	minimum_maximum_nights	High correlation with maximum_nights	Hide and exclude data. Keep maximum_nights.
46	maximum_maximum_nights	High correlation with maximum_nights	Hide and exclude data. Keep maximum_nights.
47	maximum_avg_ntm	High correlation with maximum_nights	Hide and exclude data. Keep maximum_nights.

48	bathrooms	Data is missing	Hide and exclude data.
49	review_scores_rating	Data has 1811/4221 missing rows.	Hide and exclude data.
50	review_scores_accuracy	Data has 1866/4221 missing rows.	Hide and exclude data.
51	review_scores_cleanliness	Data has 1865/4221 missing rows.	Hide and exclude data.
52	review_scores_checkin	Data has 1866/4221 missing rows.	Hide and exclude data.
53	review_scores_communication	Data has 1865/4221 missing rows.	Hide and exclude data.
54	review_scores_location	Data has 1867/4221 missing rows.	Hide and exclude data.
55	review_scores_value	Data has 1867/4221 missing rows.	Hide and exclude data.
56	reviews_per_month	Data has 1811/4221 missing rows.	Hide and exclude data.
57	host_response_time	Data has 765/4221 missing rows.	Hide and exclude data.
58	host_acceptance_rate	Data has 1052/4221 missing rows.	Hide and exclude data.
59	property_type	Similar to room type	Hide and exclude property_type. Keep room_type.
60	number_of_reviews	Data is disproportionately skewed-left, with majority having less reviews. Number of reviews without review score irrelevant to analysis.	Hide and exclude data.
61	calculated_host_listings _count	Repeated from total_host_listings_count	Hide and exclude data.
62	calculated_host_listings _count_entire_homes	Data is disproportionately skewed left. Use total_host_listings_count instead.	Hide and exclude data.
63	calculated_host_listings _count_private_rooms	Data is disproportionately skewed left. Use total_host_listings_count instead.	Hide and exclude data.
64	calculated_host_listings _count_shared_rooms	Data is disproportionately skewed left. Use total_host_listings_count instead.	Hide and exclude data.
65	bathrooms_text	Contains too many categories -45, may be difficult for analysis.	Recode into 'private' and 'shared' in 'bathroom_type' column
66	amenities	Contains 3109 categories, will be difficult for analysis	Using text frequency analysis, choose top amenities and recoded into 16 Boolean columns.
67	host_is_superhost	Contains 8 rows of missing data	Hide and exclude rows with missing data
68	host_total_listings_count	Contains 8 rows of missing data	Hide and exclude rows with missing data
69	bathrooms_text	Contains 29 rows of missing data	Hide and exclude rows with missing data
70	bedrooms	Contains 451 rows of missing data	Hide and exclude rows with missing data
71	beds	Contains 78 rows of missing data	Hide and exclude rows with missing data
72	neighbourhoods_cleansed	Several neighbourhoods such as Mandai, Pionner contains < 50 counts of data. May not be sufficient for analysis.	Hide and exclude categories with < 50 counts of data.
73	neighbourhood_groups_cleansed	Contains close correlation of estimates to neighbourhoods_cleansed data, resulting in biased in model.	Hide and exclude data. Keep neighbourhoods_cleansed.