Discover Factors Which Have Effects on Airbnb's Stakeholders by Using Python
Using Sydney Airbnb as an Example

Ziqi Wan

[1] *Finance and Business analytics, University of Sydney, 2006 NSW, Australia*
[*] *University of Sydney's e-mail: zwan8416@uni.sydney.edu.au*

ABSTRACT
This report is aimed at the analysis of Sydney's Airbnb data to provide advice to related stakeholders. Data processing, feature engineering, and model building methods were utilized to realize that endeavour.
A reliable dataset can only be formed when firstly using the data cleaning process. Linear regression, advanced non-parametric model, and model stacking are subsequently established to predict the price. According to the above analysis, insights and quantitative advice to Airbnb's stakeholders are drawn.

Keywords: *Airbnb analysis, data analysis, model building.*

1.EDA & FEATURE ENGINEERING

1.1. Data Types Identification

To begin, by utilizing the python function, 'requires license' and the other two columns have constant values. 'require guest profile picture' and the remaining four columns have one main value and some sparse values. As a result, these columns are needed to be deleted since they have no significant effects on price. Then, because data type in 'zipcode' is the object, this column is identified in dabl as dirty floats. The 'to numberic' function is used to convert this type to numeric data. Furthermore, the dollar sign of 'price' should be deleted for future analysis and model building.

To provide a more meaningful context for the field 'Host since,' the maximum value was chosen as the basis for calculating the number of days the host has been registered. Thus, 'Host since' then is replaced with float, which refers to the host's experience. The same approaches are applied on columns including 'Last review' and 'first review'.

Lastly, after total inspection of the dataset, it was determined that 'host response rate' and 'host acceptance rate' should be treated as floats rather than strings. By removing '%', these columns are reclassified to floats types

1.2. Multicollinearity Removing

To examine multicollinearity inside the dataset, 'data profiling package' are used. Due to the observed multicollinearity, related columns were deleted based on the outputs.

1.3. Check for Missing Values

1.3.1. Delete columns:

Both the 'Square feet' and 'monthly discount' columns have more than 90% null. As a result, these two columns are deleted.

1.3.2. Fill with 0:

'0' represent non-existent based on assumption. So, 'weekly discount', and remaining four columns' missing values are placed with '0'.

1.3.3. Fill with certain texts:

By assuming non-numeric features whose meaning can be concisely summarised from other inputs, bias is eliminated. To replace the missing value, the following techniques were used:

- 'Full area accessible' are filled in missing values in 'access' column.

- Based on assumption that the majority of the hosts are from Australia, 'AU' are filled in missing values in 'host location' column.
- Assuming that the default method of contacting hosts is through the Airbnb application, the missing values in the 'interaction' column are filled in with 'Contact via Airbnb'.

1.3.4. Fill with 'None':

Due to the Non-numerical features whose inputs cannot be inferred, the null values of 'space' and remaining six columns are filled by 'None'

1.3.5. Fill with mean:

On the assumption that the mean represents the averaged value, the nulls in the 'Cleaning fee perc' and other three columns are filled with the mean of corresponding inputs.

1.3.6. Fill with 0/1:

'Host about' and 'neighbourhood overview' take the form of lengthy unregularized texts. Additionally,

assuming that information about text differentiation does not contribute to price changes, whether or not such information is provided will convey the information to the customer. As a result, dummy encoded methods were applied to the 'Host about' and 'neighbourhood overview' columns, where 1 represents existence of information and 0 represents missing values.

1.3.7. Fill with Iterative Imputation:

Based on the assumption that 'Review scores checking', 'review scores location' and other five columns are correlated, Iterative Imputer is applied to predict null values by modelling a function of other features and repeating this process multiple times.

2. MODEL BUILDING

2.1. Independent variables selection

Using the above feature engineering and data processing methods, a clean and suitable dataset can be obtained. There is total 83 features discovered following the data cleaning process, due to overfitting problems, it is unwise to select all 83 features in model building. On the other hand, using too few independent variables is also unsuitable because of underfitting problems. Thus, it is crucial to select the optimal number of independent for the model building.

The assumption is that the independent variable can be chosen only if the absolute correlation between the response variable and the independent variable is greater than 0.2. Therefore, 'host listings count', and other 13 columns are selected to represent independent variables (X).

2.2. Transformation of independent variables

Both log transformation and min-max transformation approaches are applied to make all independent variables scalable.

3. MODEL ESTIMATION

3.1. Ordinary Least Squares Regression

Ordinary Least Squares is the most fundamental linear estimation method and is frequently used as a benchmark method. The linear regression method is a reasoning method that can be used to analyze many variables simultaneously and has high degree of interpretability. This linear estimation focuses on empirical risk minimization as defined below.

$$\hat{\theta} = \underbrace{argmin}_{\theta} \{\frac{1}{n} * \sum_{i=1}^{n} L(y_i, f(x_i; \theta)\} \quad \text{Eq(1)}$$

The results of the below function can be obtained by minimizing the training error, which is represented by the difference between the estimated and actual values.

$$\underbrace{min}_{\beta} \frac{1}{2}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad \text{Eq(2)}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \text{Eq(3)}$$

To avoid bias that OLS could potentially experience, several assumptions should be made:

Assumption 1: The actual regression model is linear $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$

Assumption 2: The population mean of error is zero

Assumption 3: the error term is homoscedastic and has a normally distributed probability distribution.

'sklearn.linear_model' package should be applied to create the OLS model.

There are 14 coefficients of independent variables. The 'bedrooms' variable has the highest positive coefficient. The result shows that when all other variables remain constant, the renting price will increase by approximately $20.09 for each additional unit in the 'bedroom' variable. The least negative coefficient is for 'calculated host listings count private rooms'. The result indicates that when other variables keep constant, the renting price will increase by approximately $0.37 for each additional unit increase in 'bedroom' variable.

3.2. Regularized risk minimizations

Although OLS regression has interpretability advantages, overfitting or underfitting problems may

occur. To solve this issue, the penalty factor $\lambda\Omega(\theta)$ should be applied. Therefore, the whole model can be balanced between complexity and error, and regularized risk can be reduced.

$$\hat{\theta} = \underbrace{argmin}_{\theta} \left\{ \frac{1}{n} * \sum_{i=1}^{n} L(y_i, f(x_i;\ \theta) + \lambda\Omega(\theta) \right\} \quad \text{Eq(4)}$$

Where $\Omega(\theta)\ stands\ for\ model\ complexity$

$\lambda\ stands\ for\ hyperparmeter$

3.3. Elastic net

The ridge and lasso regression can be combined by utilizing Elastic net methods, the methods would take both β_j^2 and $abs(\beta_j)$ into account. It will not only lead to sparsity, but also improve the model's ability to deal with the correlation between predictors:

$$\widehat{\beta_{EN}} = \underbrace{argmin}_{\theta} \{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^{\ 2} + \lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha) abs(\beta_j)) \quad \text{Eq(5)}$$

λ is a vital part in the equation. The equation of penalized estimation is the same as the OLS function when λ is equal to 0. Compared to the coefficient estimated by OLS, coefficients of the estimation would reduce when λ increases. The estimation coefficient would be 0, and the penalty factor will dominate when λ approaches infinity.

Ridge regression can be represented by penalty factor of β_j^2 , which causes the parameters estimated by ridge regression to slightly shrink from the parameters of OLS. Ridge regression can be used to avoid perfect correlation. However, because this type of regression does not select variables, the model contains many irrelevant features and thus fails to optimize. In addition, because the original coefficients from OLS are small in this case, the improvement of the model is not obvious,

Lasso regression can be represented by penalty factor of $abs(\beta_j)$, which shrinks the coefficients and removes the variables that shrink to zero because of the shrinkage. This process, referred to as feature selection, results in a sparse model.

The 'enet.l1_ratio_' function can be used to obtain the Lasso regression.

According to the result, Lasso regression outperforms Ridge in this case since the EN model assigns significant proportion of 99% to Lasso. This result is reasonable, ridge regression performs better when the coefficients of the model features have similar scales. However, Lasso Regression is better suited for models with multiple features with large coefficients, but with other coefficients near zero.

3.4. Advanced non-parametric model

The decision tree is a supervised machine learning algorithm that makes predictions based on a series of questions. Under this model, the prediction of each test case that falls into the area is constant. Low bias and high variance is related to this model, which usually leads to overfitting problems.

There is no need to use transformed variables to build this model, as these kinds of models are not sensitive model. Therefore, use the untransformed selected independent variables and choose to adjust the parameters to find the best prediction of the model. The random forest model is chosen among non-parametric models because it performs better than tree-based models, with a lower root mean square error and a larger R-squared.

3.5. Model stacking

Model stacking is a type of "ensemble learning" in which multiple models' predictions are combined to improve the overall model performance rather than selecting the best performing model. Therefore, noise, bias, and variance can be minimized. In this report, model stacking was used to combine the five models mentioned previously. The results show that stack model has the lowest RMSE and highest R-square. Thus, stack model is the best among all mentioned models.

Table.1. accuracy check

	RMSE	R^2
OLS_MODEL	0.441745	0.636539
RIDGE	0.441751	0.636529
LASSO	0.441760	0.636515
ENET	0.441760	0.636515
TREE	0.424944	0.663660
RF	0.388006	0.719592
STACK	0.387098	0.720902

3.6. Model Validation & Evaluation

Model stacking is chosen as the final model due to its well-performed RMSE and R-square. Multiple attempts have been conducted to improve the model's performance by varying the independent variables' combinations. A moderate score is chosen because the higher rating would potentially lead to the problem of overfitting and a lower rating would cause underfitting. The corresponding moderate score appears to represent the best complexity of the model best.

The stack model was selected as the final model due to its good RMSE and R-squared. To improve the performance of the model by changing different

combinations of independent variables, many attempts have been made to generate good predictions. The medium score is chosen because a higher score may cause overfitting, while a lower score may cause underfitting.

4. DATA MINING

4.1. Insights from Review Scores

The analysis of review scores reveals the relationship between score levels and price levels. There is a positive correlation between cleanliness and price based on 'review scores cleanliness' columns. This demonstrates that hosts can increase the price of the house by improving its cleanliness, thereby increasing the house's profit margin. In addition, 'review scores communication' were also evaluated to discover the relationship between effective communication and housing price levels. Although there are some fluctuations in the overall relationship, it shows a slightly positive correlation. This demonstrates the importance of effective communication between hosts and customers in enabling the landlord to increase the relevant price level. Using effective communication can promote hosts to obtain higher ratings, thereby increasing their attractiveness to new customers and increasing home prices.

Additionally, the 'review score location' metric reveals that a higher location level is associated with a higher price level in the hosts. Yang pointed out that in hotel's favourable location provides the host with a competitive advantage over competitors in the short and long term [1]. Furthermore, Lee shows that the properties of the house's location including tourist attraction, convenience and etc. would affect the customer's hotel choice [2]. Masiero also pointed out that there is a correlation between willingness to pay and location attributes [3]. The higher the hedonic value a customer obtains, the more the customer pays. Based on EDA results, and lords will find excellent investment opportunities in areas near North Sydney and Bondi Beach.

4.2. Insights from Hosts Related Variables

According to the EDA process, hosts that participate in authentication typically have more pricing power than hosts that do not participate in authentication. Therefore, it is vital to provide landlord certification for Airbnb's houses. In addition, there is no direct cost associated with becoming a super host. However, EDA revealed a positive correlation between becoming a super host and review ratings. It is generally believed that customer satisfaction can be transformed into customer relationships with hosts; therefore, it will benefit to continuously evolve into their hospitality and enthusiasm to promote interaction between the host and the guest. The appearance of texts such as "support", "happy" and "answer" can be interpreted as positive interaction between hosts and consumers; these words are frequently associated with favourable price levels. There are numerous ways to enhance the interactive effect, including using attractive property photographs and promptly responding to guests' needs.

Additionally, there is a correlation between weekly discount and price; a larger weekly discount corresponds to a higher price level. If the landlord wants to increase competitiveness, developing weekly discounts as promotional activities will be a wise choice to defeat the others and achieve success. Ogden-Barnes and Minahan emphasize the importance of using direct discounts as a strategic to succeed in large-scale competition [4].

4.3. Insights from Property Related Variables

From the analysis of Airbnb's housing types, most of the listings are detached houses and apartments. Compared with other listings, their prices rank among the top two. The monetary policy decision recently announced on November 3, 2020, will lower the target cash rate to 0.1% [5]. Although the economy is currently in recession, due to the gradual decline in interest rates and property values, this may be an advantageous time for landlords to invest in new properties.

Under the current circumstances, fewer people are requesting new houses, leading to a drop in property prices. However, there is a phenomenon about mismatching of some properties' price since their intrinsic value is not affected. For landlords who willing to take risks, they can borrow money to invest in new properties to get more income in the future. In addition, due to the relaxation of travel restrictions between states and countries, Australia's demand for Airbnb is likely to rebound after Covid-19. In return, landlords will get higher income.

There is a relationship between the price level and room type. For those hosts that could alter the number of bedrooms and bathrooms. To increase income, they could increase the number of bedrooms and bathrooms of their property. If hosts are unable to increase the number of bedrooms and bathrooms in their properties, they can increase the number of beds. The more accommodates, the higher the collected revenue.

5. CONCLUSION

Stacking model was chosen as the final model to make analysis about the attractiveness of the Airbnb. Airbnb's host need to take the room's cleanliness, location, discount level, types in to account. There are potential limitations in this report, Model building, feature engineering and data processing are made based on several subjective assumptions. It is most likely that

some of them would be violated, resulting in a bias in this result.

REFERENCES

[1] Y. Yang, J. Tang, H. Luo, R. LawHotel location evaluation: A combination of machine learning tools and web GIS. International Journal of Hospitality Management, 2015, pp. 14-24

[2] K.-W. Lee, H.-b. Kim, H.-S. Kim, D.-S. Lee. The determinants of factors in FIT guests' perception of hotel location. Journal of Hospitality and Tourism Management, 2010, pp. 167-174

[3] Masiero, L. Understanding hotel location preference of customers: Comparing random utility and random regret decision rules. (Report). Tourism Management, 73, 2019, pp.83–93. DOI: https://doi.org/10.1016/j.tourman.2018.12.002

[4] S. Ogden-Barnes, S. Minahan, S. Sales promotion decision making concepts, principles, and practice (First edition.). Business Expert Press, 2015.

[5] Statement by Philip Lowe, Governor: Monetary Policy Decision | Media Releases ,2020. DOI: https://www.rba.gov.au/media-releases/2020/mr-20-28.html