Data warehousing and mining

Prediction of the price of Airbnb accommodation listings

Table of Contents

[Sharing Economy 2](#_Toc139470809)

[Background 2](#_Toc139470810)

[Objective 3](#_Toc139470811)

[Methods Applied 4](#_Toc139470812)

[Data Pre-processing 4](#_Toc139470813)

[Exploratory Data Analysis (EDA) with Data Visualization 6](#_Toc139470814)

[Building the Accommodation Prediction Model 6](#_Toc139470815)

[Sentiment analysis 7](#_Toc139470816)

[Results 9](#_Toc139470817)

[Exploratory Data Analysis (EDA) 9](#_Toc139470818)

[Prediction Model 12](#_Toc139470819)

[Sentiment Analysis 12](#_Toc139470820)

[Discussion and Error Analysis 15](#_Toc139470821)

[Challenges and Problems 16](#_Toc139470822)

[References 17](#_Toc139470823)

# Sharing Economy

## Background

From rides (e.g., Uber, Lyft) to accommodation (e.g., Airbnb, HomeAway), the sharing economy has become a popular approach in the tourism and accomodation industry which to a large extent has attracted aa large number of guests and hosts alike. According to (Toader, et al., 2020), the sharing economy has facilitated an increase in the players involved in tourism through generating a viable alternative to the traditional services, thus enabling tourists to customize their trips and in the process enrich their individual and collective experiences. Ideally, the introduction of short-term home rental markets through schemes such as peer-to-peer accommodation provisions which in the recent years have been popularized with the introduction of online platforms and the smartphone revolution, have seen the growth of a new sector in economies around the world (Gyódi & Nawaro, 2021). Indeed, a report by (Airbnb, 2019) showed that in 2019, on average there were two million guests who stayed in accommodations provided via the Airbnb platform every night, a figure that has translated to up to 1 billion guests and 4 million hosts as of 2022 (Airbnb, 2022).

While the aspect of home-sharing economy is lucrative, the question of pricing and customer experience has attracted considerable interest among various parties including but not limited to researchers, economists, hosts, and guests.

#### Factors that influence the price of Airbnb listings

Price is considered a primary factor that most clients observe when selecting a lodging (Zhang, et al., 2017). In their study on the determinants of price in Airbnb listings, (Toader, et al., 2020) argue that hosts ought to propose the price of a listing based on the characteristics of the listing and their corresponding involvement. Other models have been proposed regarding the pricing of accomodation facilities.

(Tang, et al., 2019) propose a hedonistic pricing model which is based on the assumption that the price of a commodity is a reflection of a customer’s evaluation of the commodity's various attributes. As such, the overall price is the cumulative total of a consumer’s willingness to pay for each of the characteristics of a given listing (Tang, et al., 2019).

(Zhang, et al., 2011; Kim, et al., 2018) support the argument that location is a main factor in the pricing of Airbnb listings. Previous research shows that guests tend to value various location aspects like the underlying distance of the listing from the city center, ease of access to public transportation, as well as proximity to attractions (Yang, et al., 2018).

## Objective

The current study seeks to propose a data mining model for the prediction of the price of listings in Barwon South West, Vic, Victoria, Australia. This will include the implementation of various models including Xgboost, Artificial Neural Networks (ANNs), Random Forest, and Decision Trees. The best model, based on the corresponding performance measured using an error metric, will then be proposed for deployment. Besides, the current study seeks to evaluate the sentiments of the comments of users based on their experience.

Primarily, an accurate estimation of the determinants of Airbnb listing is essential not only for the management of tourism activities, but also for urban planning (Gyódi & Nawaro, 2021). Thus, the significance of the current study.

# Methods Applied

To address the two research objectives as stipulated above, the research process has been split into 4 tasks i.e.: *Data Pre-processing, Exploratory Data Analysis (EDA) with Data Visualization, Building the Accommodation Prediction Model,* and *Advanced tasks ：Sentiment analysis*. This section provides an overview of the methods adopted for each of the tasks and a justification of the choices that were made in the process.

### Data Pre-processing

Data processing involved applying treatment measures to the data to improve the quality of the data using methods such as feature selection and handling missing observations.

#### Feature Selection

Data used in this study was split into reviews and listing data which could be merged using the *listing ID* for the reviews data and *id* for the listings data. The listings data contained 6 attributes while the listings data contained 74 attributes, totaling to 80 attributes with 256758 observations. In practice, having many attributes introduces high dimensionality in the data which (Choudhury, 2019) describes as the curse of dimensionality that has been observed to lead to the overfitting of predictive models (Zhang, 2014). To minimize the effect of high dimensionality, a feature selection approach was adopted which involved selecting variables that:

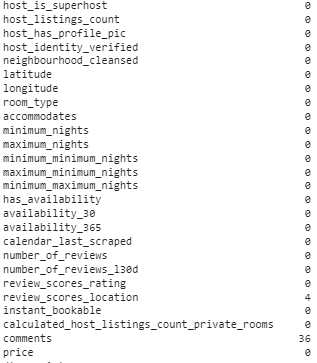
1. Variables that contain at least 1 observation
2. Categorical variables with less than 30 categories
3. Variables with correlation less than 0.6 i.e., no multicollinearity
4. Not an id variable

Following this method, the number of attributes were reduced to 26 variables which is a considerable compared to 80 attributes in the original data.

#### Handling missing observations

The following table shows the number of missing observations per attribute in the variables that were retained following the feature selection process.

Table 1: Number of missing observations per variable



As shown from table 1 above, only the *review\_scores\_location* and *comments* attributes had missing observations i.e., 4 and 36 entries were missing translating to approximately 0.002% and 0.014% missing observations which is relatively low and as such, the rows with the missing observations were excluded from the final dataset.

#### Cleaning Prices

Originally, the prices of the listings were recorded in currency form which during importing the data into the analysis environment were designated as object/strings. To convert the *price* attribute into float for use as the *target* attribute, the entries were stripped off the $ symbol and converted into floats.

### Exploratory Data Analysis (EDA) with Data Visualization

The primary objective for EDA was to examine the distribution of the price which was conducted using a *box plot*. Similarly, other objectives of the EDA were to examine the distribution of accommodation distribution on maps which was accomplished using Plotly’s *scatter mapbox*, summarize the number of accommodations and the mean price of accommodations each market/ each region respectively both of which were done using bar plots.

During examination of distribution of the number of accommodations per region, the *neighborhood* attribute was used to denote the market of the underlying listings.

### Building the Accommodation Prediction Model

As noted earlier, 4 models were proposed to address the prediction objective of the current study i.e., estimation of the price of listing. The proposed models include XGboost, Artificial Neural Networks (ANNs), Random Forest, and Decision Trees. The process of modeling was split into 3 main steps.

#### Splitting data

Essentially, the goal was to select a model that performs relatively well when predicting the price of new Airbnb listings. To this end, the data was split into train and test sets using a ratio split of 80:20. Therefore, the resulting models were trained on 205374 observations and tested on 51344 observations.

#### Model Implementation and Evaluation

This stage involved training the models on the training data and evaluating its performance using the test data. During evaluation of the models, the RMSE metric was proposed.

##### Root Mean Squared Error (RMSE)

Theoretically, the RMSE is the standard deviation of the residuals (prediction errors) i.e., the normalized distance between a vector of predicted values and the vector of actual values. The RMSE is obtained by:

RMSE = 

where yi (y1,…,yn) are the actual values and i (1,…,n)are the predicted values.

#### Model selection

Model selection was based on the RMSE scores of each model. The model with the lowest RMSE was then selected as the best model given the low error score and hence suitable for use in deployment.

### Sentiment analysis

Sentiment analysis involved using NLP methods to detect sentiments in the comments of reviews left by guests. In this research, the Valence Aware Dictionary and Sentiment Reasoner (VADER) tool was used which is a lexicon and rule-based sentiment analysis tool that is meant to analyze sentiments expressed by users through text. The tool generates various polarity scores which were used to determine the sentiment of the comments. Using the compound polarity score, the comments were categorized into positive, neutral, and negative depending on whether the score was greater than 0.05, 0, or less than - 0.05 respectively.

Subsequently, the comments related to the positive and negative comments were used to determine factors related to the likeable and dislikable nature of accommodations using a word cloud.

# Results

## Exploratory Data Analysis (EDA)

#### Distribution of price

Figure 1 below shows the distribution of the price of listings in the Barwon South West, Vic, Victoria region.



Figure 1: Distribution of accomodation price

From figure 1 above, the median price of listings is approximately $200 with some listings going for a price of up to $2750 while the least priced listing was $20.

#### Geographical Distribution of accommodations

Figure 2 below provides an overview of the distribution of the accommodations in the Barwon South West, Vic, Victoria region.

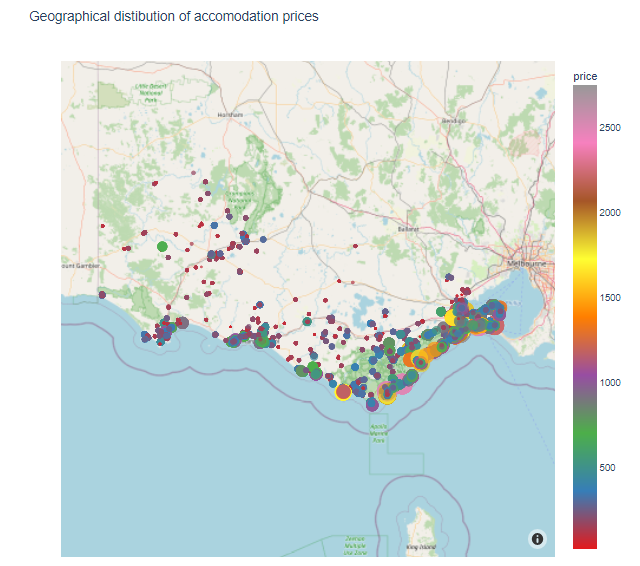


Figure 2: geographical distribution of accomodation prices

Most of the accommodations as shown in figure 3 below surround the Great Otway National Park indicting that most guests are tourists who visit the national park.

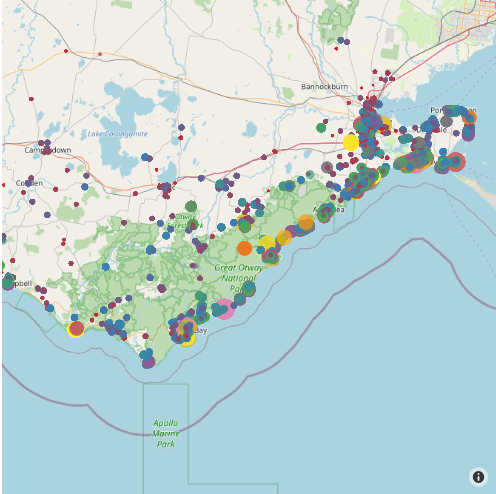
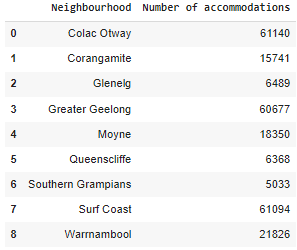


Figure 3: Closer look at the region

#### Number of accommodations in each region

Most of the accomodation listings were in *Colac Otway, Surf Coast*, and the *Greater Geelong* which contributed to over 180,000 listings i.e., 71.24% of the listings in Barwon South West (*see table 2*).

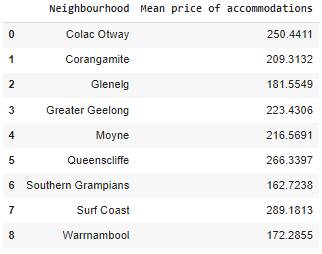
Table 2: Number of accommodations per neighborhood



#### Mean price of accommodations in each market/ each region

Overall, as observed in table 3 below, *Surf Coast* had the highest average price of approximately 289.18 while listings in *Southern Grampians* had the least price of approximately 162.72.

Table 3: mean price of listings per neighborhood



## Prediction Model

Table 4 below shows the RMSE of each model on the test data with the random forest model noted to have the lowest RMSE hence the best performing model.

Table 4: Model performance

|  |  |
| --- | --- |
| Model | RMSE |
| Xgboost | 144.900599 |
| ANN | 7516.068805 |
| Random Forest | 9.780527 |
| Decision Trees | 15.423102 |

## Sentiment Analysis

Using the VADER tool, the distribution of the number of comments related to the different sentiment types of 50,000 random sentiments is given in table 5 below.

Table 5: Sentiment distribution

|  |  |
| --- | --- |
| Sentiment | Number of comments |
| Positive | 47767 |
| Neutral | 1866 |
| Negative | 367 |

#### Factors that influence the liking of an accomodation

Figure 4 below shows the most used words by guests who liked an accomodation where like is assumed based on the positive nature of a comment.



Figure 4: Distribution of the words used in positive comments

From figure 4 above, it is noted that perfect location, short walk, great ocean view, and great host were the most used terms implying that the attitude and nature of the host as well as both the view offered by the accomodation and location influence the perception of guests on whether they liked their stay.

#### Factors that influenced the disliking or an accomodation

Interestingly, host and location are among the main factors that influence the negative perception of an experience by guests. This indicates that a negative attitude and inconvenient locations could negatively impact a client’s perception of the quality of their stay in a given accomodation (*see figure 5*).

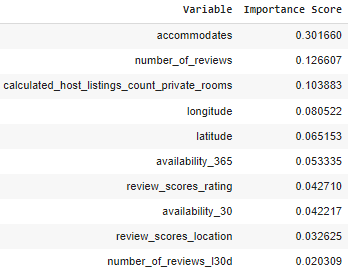


Figure 5: words used to express negative sentiment

# Discussion and Error Analysis

As noted in the model results section, the random forest model had the lowest error rate indicating that compared to the other models, the model had the best prediction accuracy. Therefore, the random forest model was used to determine the most important factors when predicting the price of a listing (*see table 5*).

Table 6: top 10 most important variables in predicting price



As noted in table 5 above, *accommodates*, *number of reviews*, and *calculated hosting listings number of private rooms* are the top 3 most important factors that influence the price of accomodation listings. From figure 6 below, it is observed that *accommodates* has a correlation score of 0.56 while both the *number of reviews* and *calculated hosting listings number of private rooms* have a negative association with price indicating that an increase in the *accommodates* of a listing increases the price of a listing while an increase in either *calculated hosting listings number of private rooms* corresponds to a decrease in the price of the underlying listing.

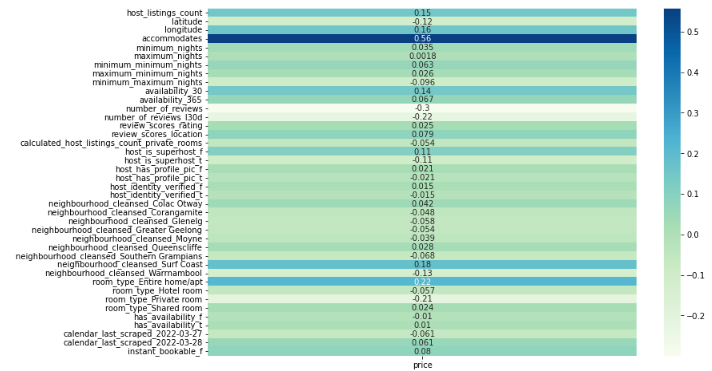


Figure 6: Correlation scores of the features relative to listing price

During implementation, the best model was not optimized i.e., most of the parameters were used by default. This might have affected the performance of the model since optimizing the hyperparameters of data mining models is often known to improve its performance.

# Challenges and Problems

While the entire analysis process had few challenges such as selection of models for the predictive modeling process, and the choice of features to retain. However, the main challenge encountered during the analysis is the resource intensive requirement of determining the sentiments of each comment where, running the VADER model on the entire preprocessed data took approximately 55 minutes which is relatively long. To solve this issue, a sample of 50,000 comments was obtained which took the model about 9 minutes to detect the sentiments.

# References

Airbnb, 2019. *Fast facts,* [Online] Available at: https://news.airbnb.com/fast-facts/  
[Accessed 02 June 2022].

Airbnb, 2022. *News.* [Online] Available at: https://news.airbnb.com/fast-facts/  
[Accessed 02 June 2022].

Choudhury, A., 2019. *Curse Of Dimensionality And What Beginners Should Do To Overcome It,* s.l.: Analytics India Magazine.

Gyódi, K. & Nawaro, Ł., 2021. Determinants of Airbnb prices in European cities: A spatial econometrics approach. *Tourism Management,* 86(2021).

Gyódi, K. & Nawaro, Ł., 2021. Determinants of Airbnb prices in European cities: A spatial econometrics approach. *Tourism Management,* 86(2021), p. 104319.

Kim, J., Jang, S., Kang, S. & Kim, S., 2018. Why are hotel room prices different? Exploring spatially varying relationships between room price and hotel attributes. *Journal of Business Research,* 1(2018), pp. 1-12.

Tang, L., Kim, J. & Wang, X., 2019. Estimating spatial effects on peer-to-peer accommodation prices: Towards an innovative hedonic model approach. *International Journal of Hospitality Management,* 81(2019), pp. 43-53.

Toader, V., Negru, A. L., Bode, O. R. & Rus, R. V., 2020. Analysis of price determinants in the case of Airbnb listings. *Economic Research-Ekonomska Istraživanja,* 1(2021).

Yang, Y., Mao, Z. & Tang, J., 2018. Understanding guest satisfaction with urban hotel location. *Journal of Travel Research,* 57(2018), pp. 243-259.

Zhang, H. et al., 2011. Modeling hotel room price with geographically weighted regression. *International Journal of Hospitality Management,* 1(2011).

Zhang, Z., 2014. Too much covariates in a multivariable model may cause the problem of overfitting. *J Thorac Dis,* 6(9), pp. 196-197.

Zhang, Z., Chen, R. J. C., Han, L. D. & Yang, L., 2017. Key Factors Affecting the Price of Airbnb Listings:A Geographically Weighted Approach. *Sustanability,* 9(1635), pp. 2-13.