Prediction of Airbnb Price in Amsterdam

Table of Content

1.0 Introduction

Airbnb is a platform marketplace that helps people looking for a place to stay like tourists or vacationers with people who have extra rooms or houses to rent out. Users can search for accommodations in various areas across the world, such as apartments, houses, villas, and even unusual venues such as treehouses and yurts. The study's goal is to predict the price of Airbnb using machine learning. Various machine learning algorithms are used in this study to forecast the airbnb pricing. The data set included variables such as total listing price, room_type, value indicating whether the host is a superhost, indicator whether the listing is for multiple rooms or not, business indicator, overall guest satisfaction, distance from city centre, longitude and latitude coordinates for location identification, and so on. All of this data is used in the research to gain an understanding of how variables affect the price of the listing, which in turn determine pricing approaches for best profitability. To build predictive models, the research employs machine learning approaches, linear regression, which includes gradient boosting regression and multi linear regression, which are widely used in this study area.

The study's goal is to develop dependable machine learning algorithms that can accurately forecast the price of an Airbnb listing. The approach is intended to aid Airbnb hosts in making price decisions. The value of Airbnb pricing prediction is found in the benefits it gives to both Airbnb hosts and guests. Accurate pricing prediction can assist hosts in pricing their listings appropriately, enhancing income and occupancy rates. This can also assist them in maintaining market competitiveness and attracting new clients. Accurately predicting the price of an Airbnb listing can help hosts establish appropriate prices for their listings, increasing revenue and occupancy rates. As a result, a precise Airbnb price forecast can be advantageous to hosts.

Prior studies have attempted to develop models to predict Airbnb prices and reduce host losses. One such study, conducted by Ye and Qian et al. (2018) used regression models and Gradient Boosting Machine to estimate booking probability and provide personalised pricing recommendations based on host goals. Similarly, researchers at Stanford University employed machine learning and natural language processing techniques to construct a price prediction model using various approaches, including linear regression, tree-based models, support-vector regression, K-means clustering, and neural networks. Their study also utilised variable feature selection strategies and neural networks to improve accuracy. Another study by Laura Lewis focused on investigating price drivers in London Airbnb listings using XGBoost and neural network models, combining her hosting experience with exploratory data analysis. Additionally, a smaller study by Yang et al. (2018) used linear regression to examine the relationship between market accessibility and hotel pricing in the Caribbean, including user ratings and hotel classes as contributing factors.

1.1 Hypothesis

Null hypothesis (H0):
The distance and number of people staying in the Airbnb influence the price of the listing.

Alternative hypothesis (H1) :
The price of Airbnb not only influence by distance and number of person staying

1.2 Researches Question

- Assist hosts in setting ideal prices for their listings, which can enhance revenue and occupancy rates.
- Accurately predicts Airbnb prices and helps hosts optimise their pricing strategies.

1.3 Preview of Main Results

The study aimed to assist hosts in setting ideal prices for their Airbnb listings to enhance revenue and occupancy rates. The results showed that the price of Airbnb is not only influenced by distance and number of people staying, but also includes location and booking day. Both GBR and multilinear regression techniques were used to analyse a wide range of variables, including categorical and continuous variables, and to capture non-linear relationships between variables. The findings demonstrated that these techniques can model the complex relationships between different factors that influence Airbnb prices, and can be applied to large datasets. Therefore, the study suggests that the model can predict the price based on the dataset, allowing hosts to set a suitable price to attract users to book the room. The results found that the Airbnb price is influenced by distance from the city centre, location, booking day, and number of people staying.

2.0 Data Set

Data was collected from the website https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities, which sources its datasets from a research paper with DOI number https://doi.org/10.1016/j.tourman.2021.104319. The dataset contains information on Airbnb prices in various European cities. This study focuses specifically on Amsterdam, using data for both weekends and weekdays, which was combined into a single dataframe with 1,103 unique listings. The aim of this study is to identify the significant factors that influence the listing price, both within and outside of the host's control.

The dataset provides a comprehensive overview of Airbnb prices in Amsterdam, taking into account various attributes such as room type, cleanliness and satisfaction ratings, bedrooms, distance from the city centre, and more. Using spatial econometric methods, this study analyses and identifies the determinants of Airbnb prices across Amsterdam. The dataset includes various variables such as total price of the listing, room type, host superhost status, multiple room indicator, business indicator, overall guest satisfaction rating, number of bedrooms, distance from city centre, and latitude and longitude coordinates for location identification. The hope is that this dataset will provide insight into the impact of social dynamics and geographical factors on global markets and pricing strategies for optimal profitability.

3.0 Research Methodology

The regression method, which is used to assess and model the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictors or explanatory variables), is the major research method. Based on the Amsterdam Airbnb dataset, we used the Gradient Boosting Regressor (GBR) to develop a predictive model to estimate Airbnb prices. In addition, we employ Mean Absolute Error (MAE) to assess the accuracy and performance of our model.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

n = the total number of data points
xi = the predicted value
yi = the actual value

The Mean Absolute percentage Error (MAPE) is frequently used in Airbnb pricing prediction because it provides a percentage error rate that is simple to understand and analyse, particularly for non-technical stakeholders. A MAPE of 10%, for example, indicates that anticipated prices are 10% off from actual prices on average. MAPE formulas are shown below.

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

$$n = number\ of\ fitted\ points$$

$$A_t = actual\ value$$

$$F_t = forecast\ value$$

Furthermore, MAPE is especially effective when the size of the mistakes is significant. A modest absolute inaccuracy (e.g., $10) in Airbnb pricing may be insignificant for a high-priced home but significant for a low-priced one. Using MAPE, we can assess the model's accuracy regardless of the absolute price level.

3.1 Gradient Boosting Regression

$$F(x) = \beta O + \Sigma|i = 1\ to\ M|\beta i * hi(x)$$

F(x) = Prediction for the target variable based on the input features x
B0 = Intercept term
βi = Coefficients of the regression model
hi(x) = Prediction of the ith regression tree for the input features x

GBR is well-known for its capacity to deal with non-linear connections between variables, which are common in real-world data. Airbnb price data is likely to contain complex, non-linear connections between variables such as property location, number of people capacity, and time of year. GBR can successfully capture these associations and use them to produce accurate predictions.

The publication "Predicting the price of Airbnb listings using Gradient Boosting Machines" by D. Wijaya and D. Lee (2019) is one example of research that employs GBR to estimate Airbnb prices. They used GBR to forecast Airbnb rates based on factors such as property location, person capacity, and many more. GBR outperformed other machine learning algorithms, according to their findings.

3.2 Multi Linear Regression

Multiple linear regression is a technique for determining a linear relationship between an output variable and its predictors. Below shows the multi linear regression equation :

$$MLR = b_0 + b_b x_1 + \ldots + b_n x_n + \xi$$

Where y is the dependent variable, b is the slope-intercept, … are the regression 0 b1 bn coefficients, x_1 ... x_n are the independent variables, and ε is the error term. Using feature engineering and variable selection, primarily by obtaining the log of 'price'. Any variables that showed multicollinearity were eliminated from the model.

R^2 value is a statistical measure of the proportion of variability in the predicted variable explained by the regression model.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2}$$

We selected multi-linear regression as a statistical technique for estimating Airbnb prices because it is widely used. Furthermore, it is a straightforward and understandable model that allows us to study the relationship between each independent variable and the dependent variable, which is the price. This is especially important in the context of Airbnb pricing, where we want to understand how different aspects of the house, such as location, number of guests staying, and so on, influence the price. Furthermore, multi-linear regression can handle both categorical and continuous variables, making it appropriate for the complicated nature of Airbnb pricing data, which contains a variety of data types.

Multi-linear regression can be used to assess the effectiveness of more advanced machine learning methods such as Gradient Boosting Regression (GBR). We can establish whether the increased complexity of the GBR model is justified and whether it produces a significant gain in prediction accuracy by comparing its performance to that of the multi-linear regression model.

Using statistical approaches, GBR and Multi-Linear Regression examine the relationship between predictor variables (such as location, booking day, and number of people staying and etc) and the target variable (the Airbnb listing price). They then take this relationship and utilise it to build a model that can estimate the price of a new listing based on the predictor factors. A host, for example, can use GBR or Multi-Linear Regression to input information about their listing, such as location, time of week, and number of people staying, and the model will generate a predicted pricing for the listing. The host can then use this estimated price to set a competitive and reasonable price for their listing.

In conclusion, GBR and Multi-Linear Regression can help hosts decide how much to charge for their Airbnb listing by providing a data-driven approach to understanding the elements that influence listing fees. By studying these variables and developing a predictive model, hosts may make more informed price decisions, leading in increased occupancy rates and more profitable listings.

4.0 Results

The objective of this research is to identify the factors that influence the price of Airbnb listings in Amsterdam and use this information to develop a predictive model to assist Airbnb hosts with pricing decisions. Two models were employed in the study, namely the Gradient Boosting model and the Multi Linear Regression model, with the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) metrics used to measure error. The Amsterdam Airbnb Dataset was used for the study, which included data for both weekends and weekdays.

During the feature selection process, OLS Regression was used to identify the factors that have a significant influence on Airbnb prices. The results of the analysis showed that all variables, except for the room_private and entirehome_apt variables, are highly correlated with each other, as indicated by the Variance Inflation Factor (VIF). Since room_shared, room_private, and entirehome_apt are all part of the same category (i.e., room type), they were dropped from the model due to the high possibility of multicollinearity.

	Variable	VIF
0	room_shared	1.071673e+07
1	room_private	6.194538e+05
2	person_capacity	1.846036e+00
3	bedrooms	1.862919e+00
4	dist	1.408229e+00
5	lng	1.068270e+00
6	lat	1.386005e+00
7	week_time	1.016503e+00
8	entirehome_apt	6.034967e+05

Diagram 1.0 shows the VIF

Furthermore, we used Ordinary Least Squares Regression (OLS) to estimate the coefficients of linear regression equations that characterise the connection between one or more independent quantitative variables and a dependent variable. The R-squared value in Diagram 2.0 is 0.566, indicating that the model explains 56.6% of the variability in the dependent variable, while the remaining 43.4% is unexplained and could be attributed to other factors not included in the model.A -squared value of 0.566 is considered moderate to good. It implies that the independent variable(s) have a moderate to significant influence on the dependent variable, although there could be other factors influencing the dependent variable as well.

Furthermore, the adjusted R-squared value of 0.564 indicates that the independent factors in the model explain about 56.4% of the variation in the dependent variable after adjusting for the number of independent variables and sample size. It also implies that, after controlling for the number of independent variables and sample size, the independent factors in the model collectively explain a moderate to substantial amount of the variation in the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                realSum   R-squared:                       0.566
Model:                            OLS   Adj. R-squared:                  0.564
Method:                 Least Squares   F-statistic:                     337.4
Date:                Fri, 14 Apr 2023   Prob (F-statistic):               0.00
Time:                        14:31:08   Log-Likelihood:                -14040.
No. Observations:                2080   AIC:                         2.810e+04
Df Residuals:                    2071   BIC:                         2.815e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.4056      0.050      8.169      0.000       0.308       0.503
entirehome_apt   -47.4022      5.886     -8.053      0.000     -58.945     -35.859
room_shared      211.7200     25.893      8.177      0.000     160.941     262.499
room_private      48.5530      5.971      8.131      0.000      36.843      60.263
person_capacity  102.1118      5.977     17.085      0.000      90.391     113.833
bedrooms         101.0560      8.416     12.007      0.000      84.551     117.561
dist             -55.1413      2.588    -21.303      0.000     -60.217     -50.065
lat            -2361.9644    283.173     -8.341      0.000   -2917.298   -1806.631
week_time         34.8306      9.174      3.797      0.000      16.839      52.822
lng              356.1511    121.207      2.938      0.003     118.452     593.851
==============================================================================
```

Diagram 2.0 shows OLS Regression Results

Below shows the regression plot before fine tuning by using the gradient boosting regression.

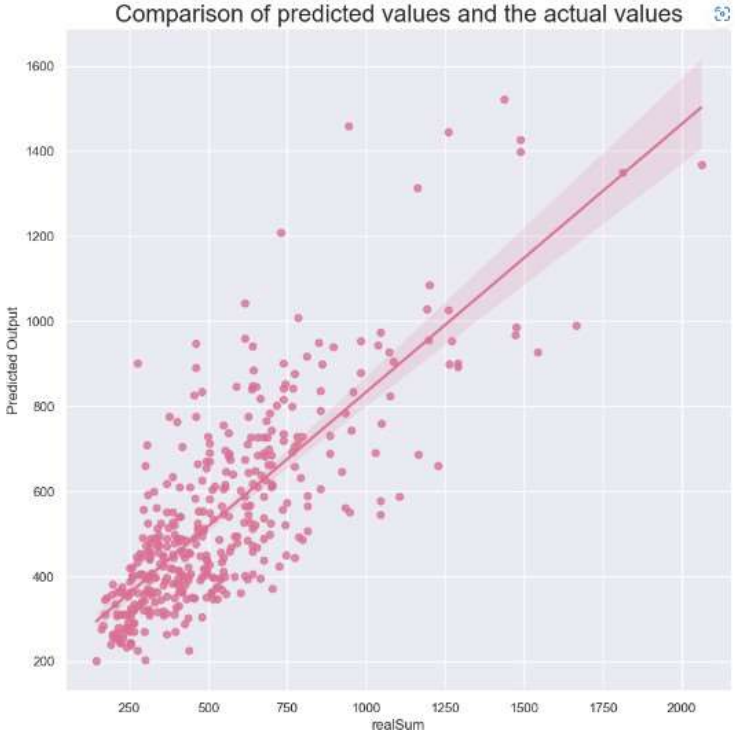

Diagram 3.0 shows the regression plot for Gradient Boosting Regression before fine tuning

The table below shows the results of MAE where the MAE for Gradient Boosting Regression is 130 while the MAE for multi linear regression is 150.

	Model	Mean Absolute Error	Mean Squared Error	MAPE
0	Multi Linear Regession	150	39110	0
1	Support Vector Regressor	173	67649	0
2	K Nearest Regressor	163	54117	0
3	PLS Regression	150	39110	0
4	Decision Tree Regressor	173	75631	0
5	Gradient Boosting Regressor	131	31520	0

Table 1.0 shows the results of MAE for all model before fine tuning

Following the selection of appropriate variables, we run a series of linear regression models to determine the optimal model for predicting Airbnb prices. According to our findings, the gradient boosting regression model and multi regression model outperform the K Nearest Regression method, Decision Tree Regression method, PLS Regression, and Support Vector Regression.

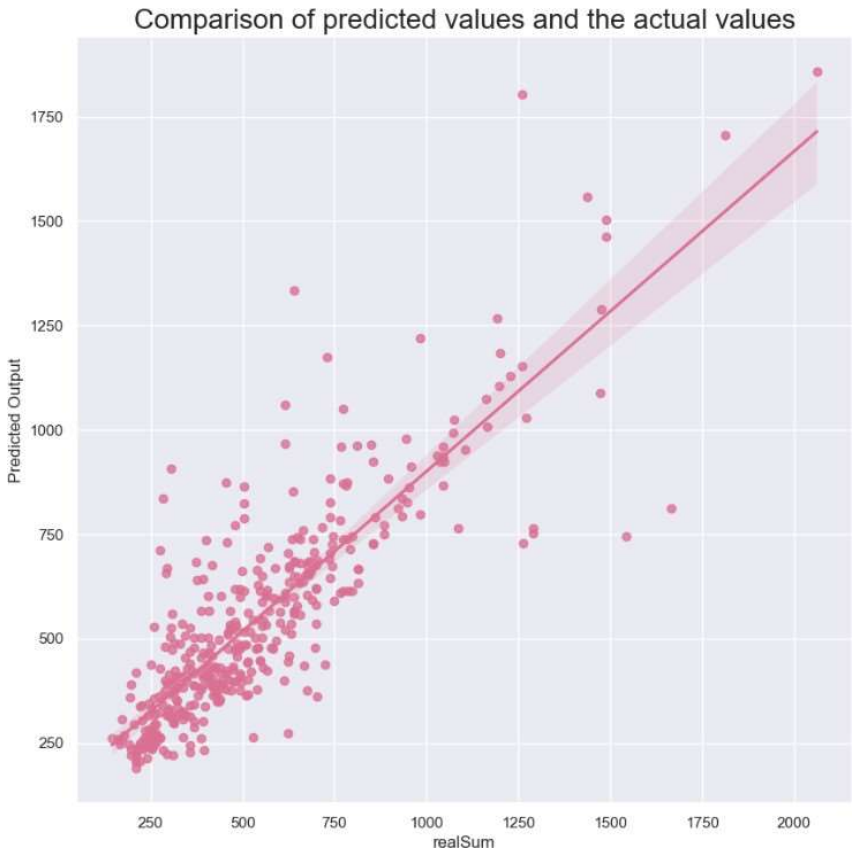

Diagram 4.0 shows regression plot of Gradient Boosting Regression after fine tuning

Using Gradient Boosting Regression, the regression plot shows the relationship between realSum and predicted output. As we can see, there may be outlier, but based on our observations, it

is not an error, but rather a higher cost of renting in the listing. The Gradient Boosting regression graphic outperforms the Multi Linear Regression plot. It demonstrates that the model better represents the relationship between the independent factors and the dependent variable (price). The most important pricing predictors are the number of people, capacity, location, distance from the city core, and week time.

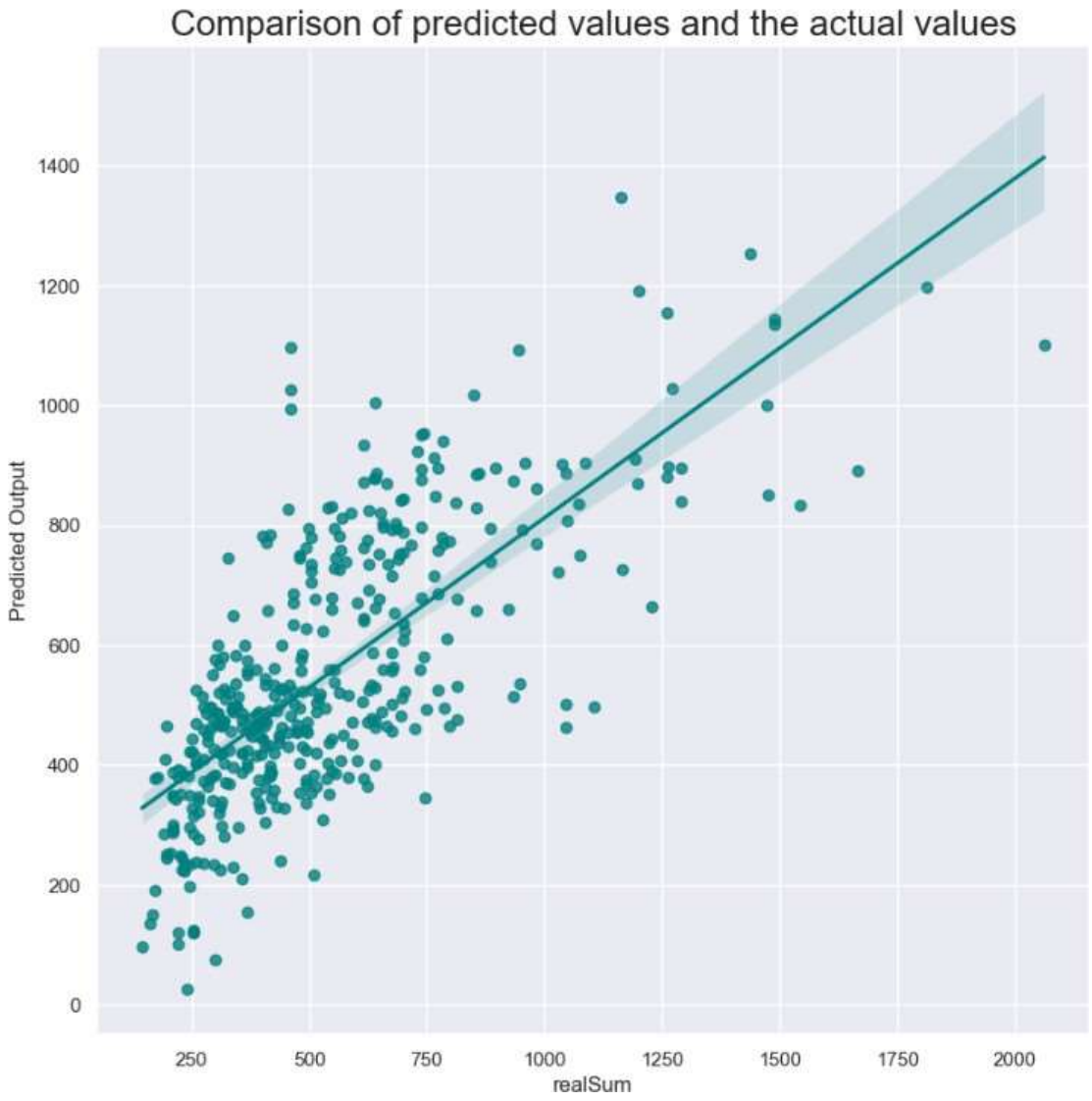

Diagram 5.0 shows the regression plot of Multi Linear Regression after fine tuning

The regression plot shows the relationship between realSum and predicted output using Multi Linear Regression. As we can see there might be some outlier but through our observations it is not an error but it might be a higher price of rental in the listing. As compared to other model, this is the second method that can be considered a better plot as it captured a few of the the relationship between the independent variables and the dependent variable (price) but Gradient Boosting regression plot is considered the best among these two methods.

	Model	Mean Absolute Error	Mean Squared Error	MAPE
0	Gradient Boosting Regressor	100.239955	23463.802165	21.101389

Table 2.0 : MAE for Gradient Boosting Regression

The Mean Absolute Error (MAE) is 100.24, according to table 2.0. According to the statistics, the price range of the Airbnb listings is $100 to $2,500, thus an MAE of 100 may be regarded as appropriate because it represents only a small percentage of the price range. It is appropriate given the wide price range of Airbnb offerings. Because this model is used as a reference and guideline for price setting, and the ultimate decision is dependent on other criteria such as market trends, guest feedback, or competition, an MAE of 100 may be sufficient.

	Model	Mean Absolute Error	Mean Squared Error	MAPE
0	Multi Linear Regressor	150.001663	39110.787133	0.311814

Table 3.0 : MAE for Multi Linear Regression

We employed a few approaches to check the inaccuracy to ensure the model's accuracy. According to table 3.0, the Mean Absolute Error (MAE) is 150.00, which is greater than that of Gradient Boosting Regression. In the case of Multi Linear Regression, the MAPE is 0.311, indicating that predicted prices are 31.1% off from real prices on average. MAPE of 0.311 is fair because it represents only a small fraction of the price range.

In conclusion, two models were used to predict Airbnb prices. Based on the analysis, it can be concluded that the Gradient Boosting Regression model is the best model to predict Airbnb prices among the models tested. The model shows that the number of people, capacity, location, distance from city centre, and week time are significant predictors of price. The MAE of 100.24 may be considered reasonable, as it represents only a small percentage of the price range. The Multi Linear Regression model is also a good model, but it has a higher MAE of 150.00 compared to the Gradient Boosting Regression model. Overall, the analysis suggests that the Gradient Boosting Regression model can be a useful tool for Airbnb hosts to set prices for their listings.

5.0 Discussion

The purpose of this study is to determine which features of the airbnb influence the price of Airbnb in order to predict the price to help the host of Airbnb in making decisions of the pricing. Based on the key findings of the prediction of Airbnb prices using GBR and Multi-Linear Regression, it can be concluded that the price of an Airbnb listing is heavily influenced by the location, weektime, distance from the city centre, and the number of people staying in the Airbnb. These factors can have a significant impact on the price of the listing, and they should be taken into account when setting the price of an Airbnb listing.

Additionally, it can be seen that the GBR model outperforms the Multi-Linear Regression model in terms of prediction accuracy, as evidenced by the lower Mean Absolute Error (MAE) of 100 for GBR compared to 150 for Multi-Linear Regression. This suggests that the GBR model is better suited to predicting Airbnb prices than the Multi-Linear Regression model, and may be a more useful tool for hosts and Airbnb hosts looking to accurately price their listings. Overall, the findings of this study suggest that hosts and Airbnb managers should pay close attention to the location, weektime, distance from the city centre, and the number of people staying in the Airbnb when setting prices for their listings. They should also consider using advanced machine learning algorithms such as GBR to accurately predict prices and maximise their revenue potential.

As for this project the limitation of this study is that the variables are closely related to each other. The main limitation of analysing data using machine learning with only one state or city dataset is that the results and insights obtained may not be generalizable to other regions or contexts. This is because different regions may have unique characteristics, such as different demand patterns, tourist attractions, and cultural preferences, which may affect the Airbnb prices differently. For example, if a machine learning model is trained only on the Amsterdam Airbnb dataset, it may not be accurate for predicting Airbnb prices in other cities or countries. This is because Amsterdam may have unique characteristics that are not representative of other regions, such as high demand due to its popularity as a tourist destination.

Moreover, using only one state or city dataset may also limit the diversity of the training data, which could lead to overfitting of the model. Overfitting occurs when the model is too closely fitted to the training data, which can lead to poor generalisation to new data. To avoid overfitting, it is important to use a diverse range of data points from different regions, time periods, and contexts. Therefore, to ensure that machine learning models accurately predict Airbnb prices across different regions and contexts, it is important to use a diverse range of data points from different cities and countries. Additionally, it is important to carefully evaluate the model's performance on unseen data from different regions and contexts to ensure that it is generalizable and not overfit to the training data.

6.0 Conclusion

According to the discussion, it concluded that the hypothesis can be **rejected** as the results show that the price of Airbnb is not only influenced by distance and number of people staying. It also includes the location and booking day too. The purpose of this research is to assist hosts in setting ideal prices for their listings, which can enhance revenue and occupancy rates. Based on the result, it can be concluded that the model can predict the price based on the dataset to show that it can be applied to assist the host to set a suitable price to ensure to attract the user from booking the room. The use of the Gradient Boosting Regression helps to improve the analysing the dataset as both techniques can identify which features or variables are the most important predictors of Airbnb prices. This information can help hosts to make decisions about how to price their listings.

Besides, both GBR and multilinear regression can handle a wide range of variables, including categorical and continuous variables, and can capture non-linear relationships between variables. This flexibility allows them to model the complex relationships between different factors that influence Airbnb prices. Other than that, both techniques can be applied to large datasets, making them useful for analysing and predicting prices for thousands of Airbnb listings. This scalability is particularly important in the context of Airbnb, where there are millions of listings worldwide, and accurate pricing is essential for hosts and guests alike. The results find that the Airbnb price is influenced by distance from the city centre, location, booking day and number of people staying . Through multi-linear regression methods and gradient boosting methods, it's clear that Airbnb prices are not only influenced by the distance and number of people staying but it also includes the distance, number of people staying, location and booking day .

We discovered that our dataset is insufficient, which leads to overfitting when a model is too complicated and matches the training data too closely, resulting in poor generalisation and inaccurate predictions on new, unknown data.In the absence of sufficient data, the model may attempt to fit noise in the data rather than the underlying patterns, resulting in overfitting. In such circumstances, the model may perform well on training data but not so well on test or new data. It should be highlighted, however, that the final decision on price setting should be based on other variables such as market trends, guest feedback, or competition.

As a result, employing a more comprehensive and up-to-date dataset is one strategy to increase the accuracy of Airbnb price forecasts. By integrating more important information in the dataset, the accuracy of Airbnb price prediction can be increased. The distance to major sites, public transit alternatives, surrounding restaurants and stores, and the number of bedrooms and bathrooms are all potential features to consider. These characteristics can provide useful information on the appeal and value of a specific property. Furthermore, discovering the relationship between current features in the dataset is another technique to improve the accuracy of Airbnb pricing forecasts. This is possible by employing advanced machine learning techniques like clustering and dimensionality reduction. These strategies can aid in the discovery of hidden patterns and correlations in data, which can then be utilised to produce more accurate predictions.

7.0 References

Laura Lewis (2019). Predicting airbnb prices with machine learning and deep learning,

Peng Ye, Julian Qian, Jieying Chen, Chen-hung Wu, Yitong Zhou, Spencer De Mars, Frank Yang, and Li Zhang.(2018). Customised regression model for airbnb dynamic pricing. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 932–940.

Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, and Hoormazd Rezaei.(2019) Airbnb price prediction using machine learning and sentiment analysis. arXiv: Learning

Wijaya, D., & Lee, D. (2019). Predicting the price of Airbnb listings using Gradient Boosting Machines. Journal of Hospitality and Tourism Technology, 10(2), 232-247. doi: 10.1108/JHTT-11-2018-0122

Y. Li, Q. Pan, T. Yang, and L. Guo (2016) "Reasonable price recommendation on airbnb using multiscale clustering," in Control Conference (CCC), 2016 35th Chinese, pp. 7038–7041, IEEE

Y. Yang, N. J. Mueller, and R. R. Croes (2016) "Market accessibility and hotel prices in the caribbean: The moderating effect of quality-signalling factors," Tourism Management, vol. 56, pp. 40–51