8/11/2020

Yu Ying Cheng

U1920016D

Practical Assignment

BC3409

**Contents**

[**1.** **Analysis of dataset:** 2](#_Toc55759250)

[**1.1 Price Analysis:** 4](#_Toc55759251)

[**1.2 Room Type Analysis:** 5](#_Toc55759252)

[**1.3 Location/Neighbourhood Analysis:** 6](#_Toc55759253)

[**1.4 Room Availability Analysis:** 12](#_Toc55759254)

[**2.** **Results:** 13](#_Toc55759255)

[**2.1 Linear Regression:** 13](#_Toc55759256)

[**2.2 Decision Tree Model:** 14](#_Toc55759257)

[**2.3 Random Forest Model:** 16](#_Toc55759258)

[**2.4 XgBoost model:** 17](#_Toc55759259)

[**2.5 Analysis and comparison of results:** 18](#_Toc55759260)

[**3.** **Applications and Recommendations:** 19](#_Toc55759261)

[**4.** **References:** 20](#_Toc55759262)

# **Analysis of dataset:**

Statistical Analysis and visualisation:

Total number of Listings: 7908

There are a total of 17 features and variables present in the dataset given. Some of these features will be identified as critical and will be analysed. Their relationship with price will also be investigated, while other variables will be deemed less significant and will not be included in the machine learning algorithms.

|  |  |
| --- | --- |
| Variable | Data |
| Name | Name of listing |
| id | Id number |
| host\_id | Id number |
| host\_name | Name of host |
| neighbourhood\_group | East/North/Central/West/North-East |
| neighbourhood | Neighbourhood Region |
| latitude | Latitude no. between 1-2 |
| longitude | Longitude no. between 103-104 |
| room\_type | Private Room/(Entire home/apt)/ Shared room |
| price | Price figure |
| minimum\_nights | No. of nights |
| number\_of\_reviews | Numbers of reviews for listing |
| last\_review | Date of last review |
| reviews\_per\_month | Number of reviews per month |
| calculated\_host\_listings\_count | Number of listings host has |
| availability\_365 | Availability of listing out of 365 |

**Correlation between variables:**

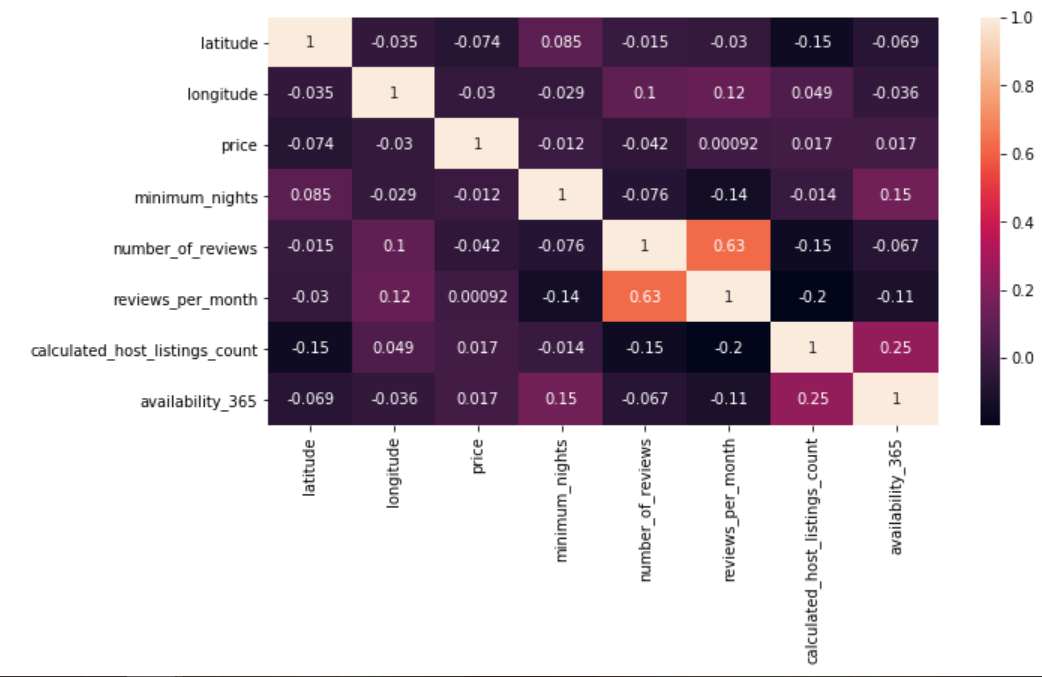


Figure : Heatmap for significant continuous variables

The heatmap above in Figure 1.0 shows the correlation between selected continuous variables. Among the given continuous variables, I decided to omit insignificant and irrelevant variables like listing id, name and host\_id. From the result shown above, the correlation between individual variables are generally very low, with the exception of obvious correlations between number of reviews and reviews per month. The correlation of price and other variables are identifiably low (<0.1), and the analysis of individual variables will be analysed and elaborated later on in the report.

## **1.1 Price Analysis:**

Figure :Breakdown of price range

Based figure 1, it is evident that most of the listings (~76%) are below $200 per night, while only a small percentage (<3%) are listed at more than $500 per night.

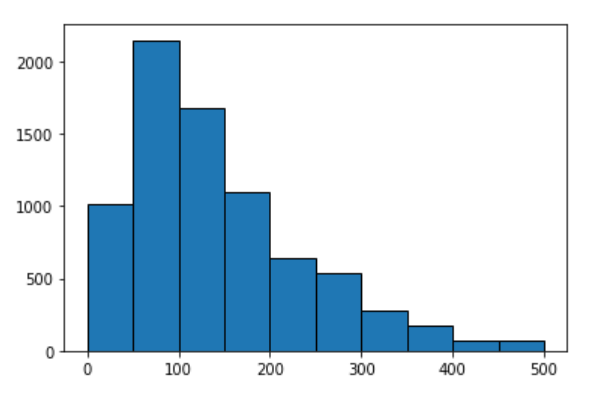
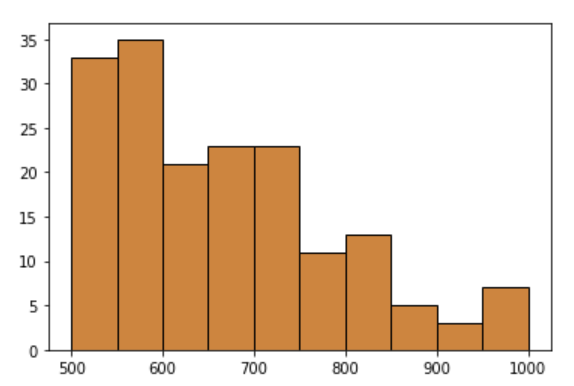
 

Figure 3: Listings with price between 0-500 Figure 4: Breakdown of listings with price between 0-500

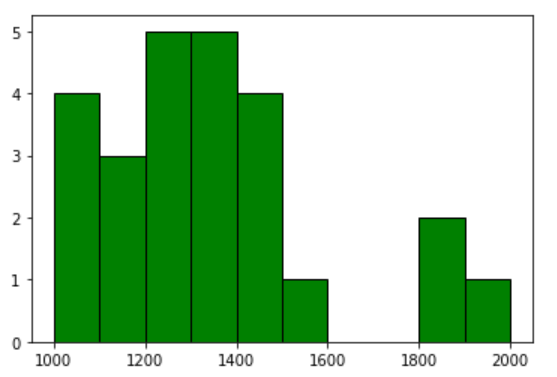
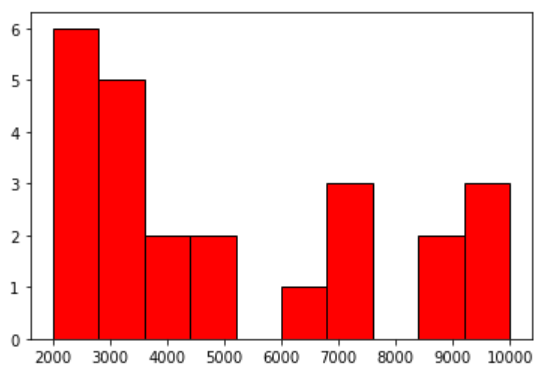
 

Figure 5: Listings with price between 1000-2000 Figure 5: Listings with price >2000

Based on the figures displayed above, majority of listings fall below the <$500 SGD per night category, with listings between the range of $50-$100 SGD per night being the most common (>2000 listings). Beyond the price of $500 SGD, the number of listings are relatively scarce and rare, with the total number of listings that exceed $1000 SGD being fewer than 100 (out of a total of 7908 listings).

Thus, while we can predict that most Airbnb hosts will list their properties at a price of below $500, these hosts can also expect to face more competition and alternatives in this price range, as customers on a budget will have more options to choose from. It can be logically inferred that listings in the lower price range are likely to be more popular and affordable for customers and tourists, thus there is a larger supply and availability of cheaper listings to cater to the demand. On the other hand, hosts with listings in the pricier range can probably expect their listings to be less popular and are targeted at a wealthier audience. We can also guess that pricier listings provide resources and facilities that are unavailable in cheaper listings.

## **1.2 Room Type Analysis:**

Figure 6: Breakdown of Room Type

As illustrated in figure 6, the large majority of listings (~95%) are of the ‘Private Room’ or ‘Entire home/apt’ apartment type. Only 5% of the listings are of the ‘Shared Room’ apartment type. It can be estimated that hosts should aim to design their apartments to be under the first 2 mentioned categories above, as customers are more likely to prefer a private apartment and housing, rather than to share a room/apartment with strangers.

Figure 7: Average Price for Each Room type

Figure 7 shows the average price of listing for each room type. Generally, an entire home/apartment will be the most expensive listing, followed by private room type and then shared room type. An obvious identification and guess to account for this observation will the size of each apartment type. An entire home will normally be the most spacious and largest in terms of apartment area. A private room and shared room will be much smaller, and hosts can expect to earn less for private rooms and shared rooms.

## **1.3 Location/Neighbourhood Analysis:**

The table below shows the number of listings in each neighbourhood in Singapore.

|  |  |
| --- | --- |
| **Neighbourhood** | **Listings in neighbourhood** |
| Ang Mo Kio | 58 |
| Bedok | 373 |
| Bishan | 57 |
| Bukit Batok | 65 |
| Bukit Merah | 470 |
| Bukit Panjang | 34 |
| Bukit Timah | 131 |
| Central Water Catchment | 34 |
| Choa Chu Kang | 63 |
| Clementi | 102 |
| Downtown Core | 428 |
| Geylang | 994 |
| Hougang | 109 |
| Jurong East | 118 |
| Jurong West | 153 |
| Kallang | 1043 |
| Lim Chu Kang | 1 |
| Mandai | 3 |
| Marina South | 1 |
| Marine Parade | 171 |
| Museum | 63 |
| Newton | 134 |
| Novena | 537 |
| Orchard | 136 |
| Outram | 477 |
| Pasir Ris | 71 |
| Punggol | 43 |
| Queenstown | 266 |
| River Valley | 362 |
| Rochor | 536 |
| Sembawang | 41 |
| Sengkang | 67 |
| Serangoon | 69 |
| Singapore River | 175 |
| Southern Islands | 17 |
| Sungei Kadut | 5 |
| Tampines | 6 |
| Tanglin | 210 |
| Toa Payoh | 101 |
| Tuas | 1 |
| Western Water Catchment | 4 |
| Woodlands | 66 |
| Yishun | 53 |

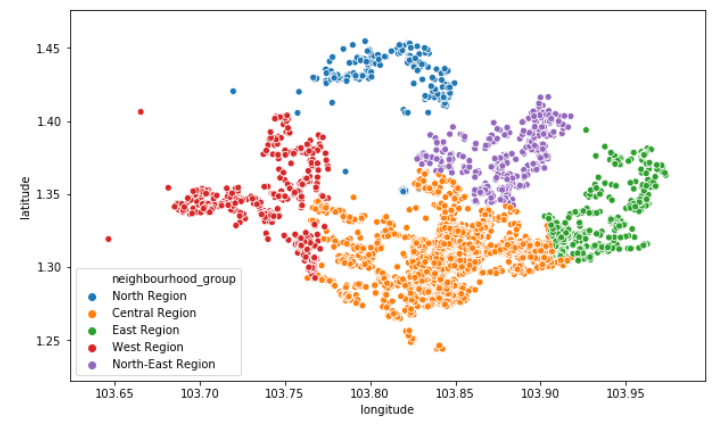


Figure 8: Spread of listings by region

Figure 9: Listings in Neighbourhood

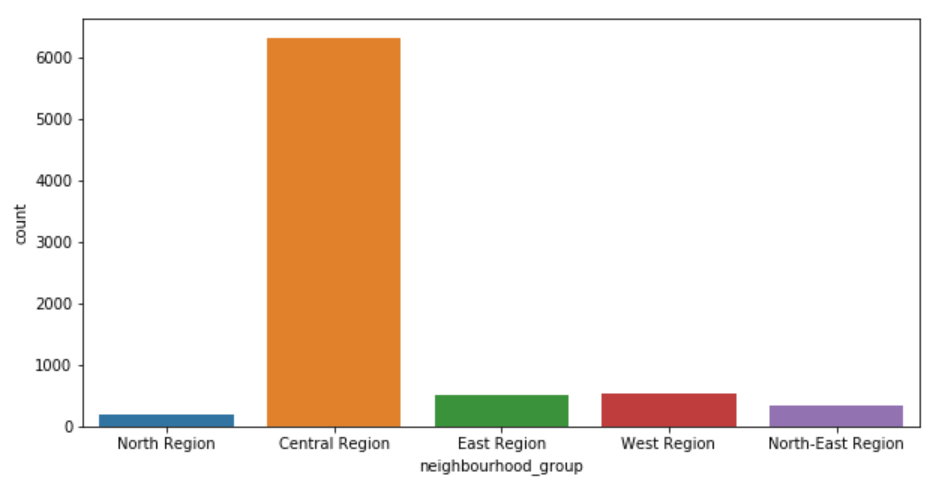


Figure 10: Listings in each Region

The table above, figures 9 & 10 show the breakdown of listings in each neighbourhood. Figure 10 shows us the number of listings in each of the 5 regions in Singapore while figure 9 shows the spread of listings based on their longitude and latitude values. We can see that the overwhelming majority of listings are present in the Central Region, while the North Region represents the fewest number of listings. In terms of the individual neighbourhood regions, Kallang and Geylang are the two most popular regions for host listings (>1000 & ~1000 respectively). On the contrary, regions like Tuas, Marine South and Lim Chu Kang only have 1 listing each.

**Guess for observations:**



Figure 11: Locations of Major hotels in Singapore



Figure 12: Locations of attractions in Singapore

By conducting a simple search of locations of 137 major hotels in Singapore and plotting them on Google Maps (Figure 11), we realise that most of them are located in the Central regions as well. From this observation, we can infer that many of the Airbnb listings are catered to tourists or people seeking a staycation in Singapore, and customers will see Airbnb apartments as direct alternatives to hotels in the region. By further investigating the reason for hotel’s location being centralised in the Central regions, we realise that many of the tourist attractions and frequented places of interest by visitors are located in the Central regions (Figure 12). Thus, we can make an inference that listings in these regions can command a higher price due to the convenience and close proximity to attractions offered to occupants.

**Analysis of location of listing and listing price**

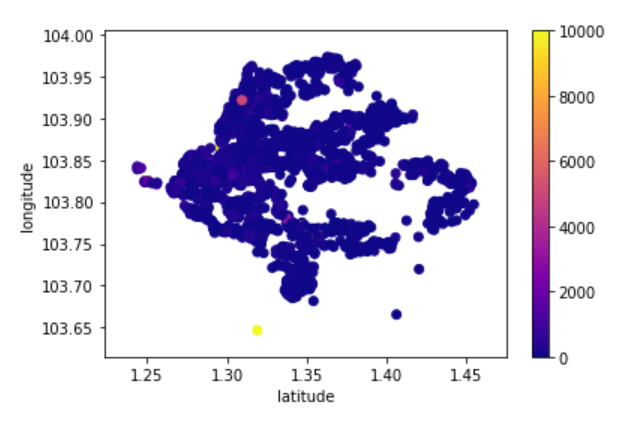


Figure 13: Listing location for all prices

As seen in Figure 9, most of the listings are concentrated in the central and west regions. Listings in the east and north regions are relatively scarcer.

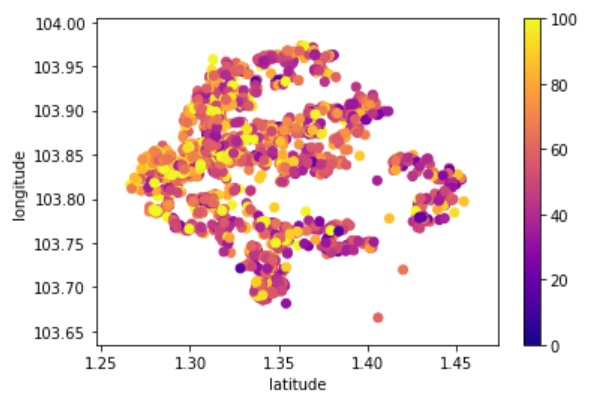
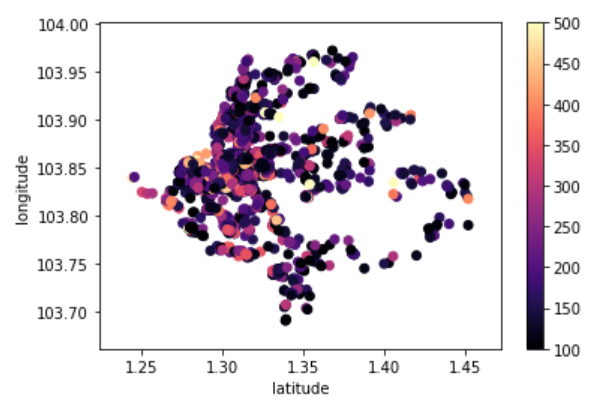
 

Figure 14: Listings Below $100 SGD Figure 15: Listings Between $100-$500 SGD

Based on Figure 10, which shows the locations of listings below $100 SGD and Figure 11, which shows locations of listings between $100-$500 SGD, the pricing of listings on the East and North regions are generally on the cheaper and more affordable end, while the pricing of listings in the Central and West region are relatively pricier.

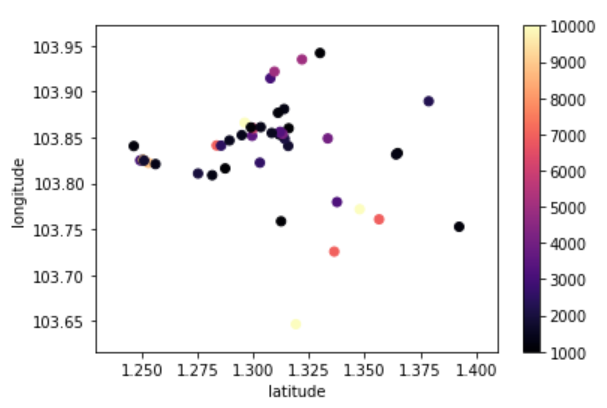
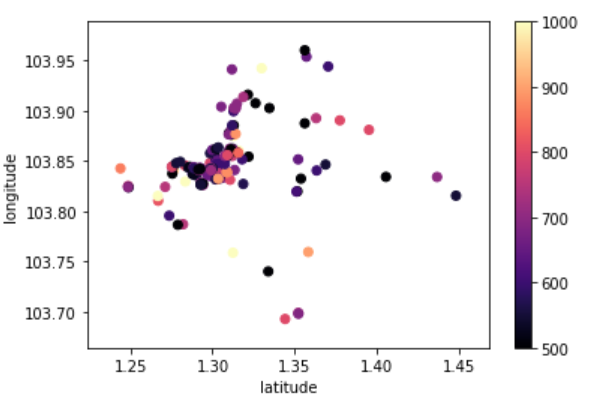


Figure 16: Listings Between $500-$1000 SGD Figure 17: Listings >$1000 SGD

The observation is further exemplified in figures 12 and 13, as the more expensive listings are concentrated in the West and central regions. In figure 13, the ultra-expensive regions (>$1000 SGD) are almost non-present in the East and North-East regions, as there are only a few listings located there.

**Prediction:**

Based on observation detailed above, we can expect hosts to list their apartments at a lower price range if their apartments are in the North-East and East regions, and at a higher price range if their apartments are at located in the Central and West regions.

## **1.4 Room Availability Analysis:**

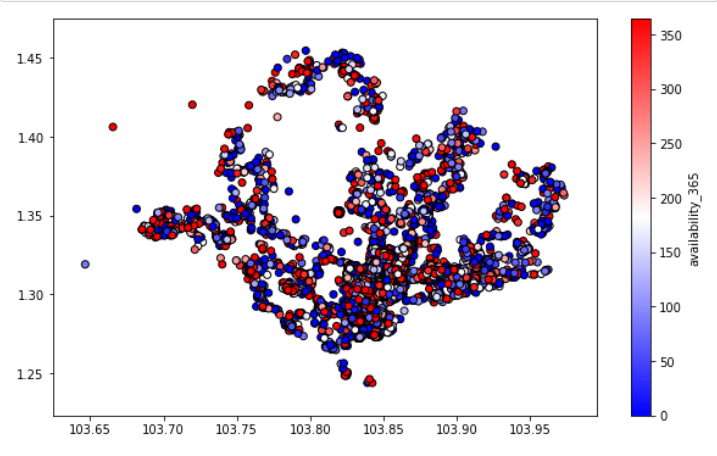


Figure 18: Locations of listings and their availability

By plotting the locations of Airbnb listings and their availability throughout the year, we realise that while the listings of rooms that are available throughout the year(or most) are spread evenly across Singapore. However, listings that are less available are concentrated in the East and Central regions, and we can see that the blue points on Figure 18 in the West and North regions in Singapore are scarcer.

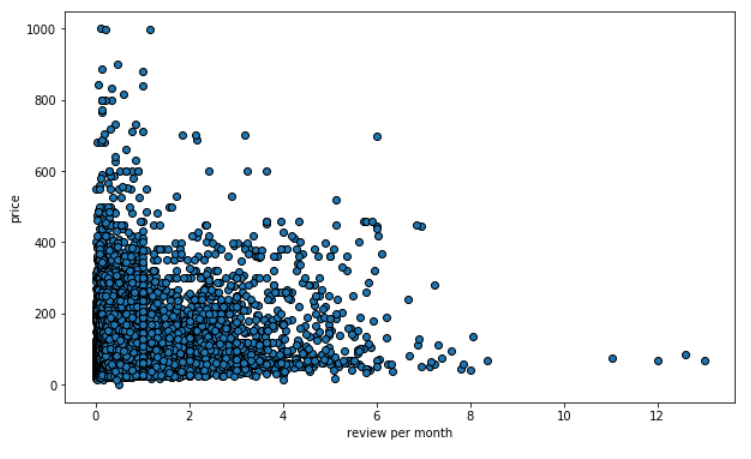


Figure 19: Listing Price against Listings review per month

In figure, we can see the points plotted for listings with price of less than $1000. From this diagram, we can see that there is not a very significant pattern or correlation between the listing’s price and listing’s reviews per month.

# **Results:**

The models selected for this project are Linear Regression, Decision Tree, Random Forest and XgBoost.

## **2.1 Linear Regression:**

Linear regression (LR) is a linear approach to modelling the relationship between a scalar variable (in this case the price of listings on Airbnb) and the other explanatory variables which have been investigated in the earlier sections. Linear regression is extremely useful and practical in prediction and forecasting events like the one in our case, as LR can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After the development of the LR model, the fitted model can be used to make a prediction of host listings for this project.

**Assumptions for Linear Regression:**

**Weak Exogeneity:** Predictor variables for this project are fixed values, rather than random variables and that these variables are error-free. We assume that these variables are accurate in their representation and are not biased in any ways.

**Constant Variance:**  We assume that the different values of the response variable (i.e. price) have the same variance in their errors, regardless of the values of the predictor variables. This assumption may not entirely hold true for this project due the presence of many predictor variables and difference in their nature.

**Independence of errors:** We assume that errors in predictor variables (for example location and apartment size) are uncorrelated with each other. This means that the presence of error for one variable will not result in an immediate impact and effect upon another predictor variable.

**Performance of Linear Regression**

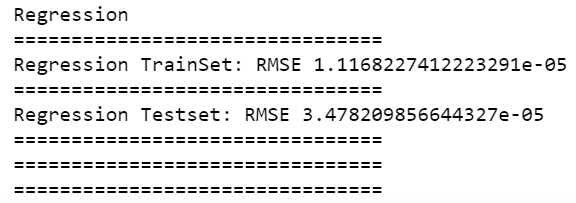


Figure 20: RMSE of Linear Regression Model

By running the script using Scikit learn, we have obtained an RMSE value of 1.12e-05 for the training set and 3.48e-05 for the test set.

## **2.2 Decision Tree Model:**

Decision Trees are a type of supervised learning method used for both classification and regression tasks, with the aim of creating a model that predicts the outcome of a target variable. In this case we are performing a regression task in trying to predict the appropriate price of a listing based on the presence of selected predictor variables.

One hypothesis and prediction I have prior to carrying out the project is that decision tree will be one of the top performers, at least when compared to linear regression. Decision Tree is especially suitable in situations like ours, where features and outcome is nonlinear or where features interact with each other. Tree based models split the data multiple times according to certain cut-off values in the features and through this process, different subsets of the dataset are created, with each instance belonging to one subset.

The Classification and Regression Trees (CART) algorithm is one of the most popular algorithm for decision trees.

**Advantages of Decision Tree:**

* Decision tree generally requires less effort for data preparation during pre-processing
* Decision tree does not require normalisation of data, although I have performed normalisation on the predictor variables as a standard before passing in the data for the different models to establish consistency.
* Visualisation of Decision tree is easy and intuitive, as the data ends up in distinct groups that are often easier to understand than points on a multi-dimensional hyperplane as in linear regression.

**Disadvantage of Decision Tree:**

* Small changes in the data can cause large change in the structure of decision tree, causing instability. Each split depends on the parent split and if a different feature is selected as the first split feature, the entire tree structure changes. Confidence in the model will be compromised.
* Decision tree normally requires more time to train the model, although the dataset is still relatively small in for this project and minimal time was involved in obtaining results.

**Performance of Decision Tree Model:**

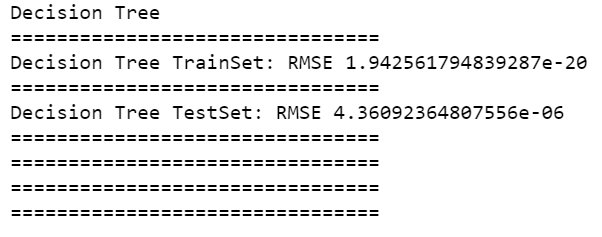


Figure 21: RMSE values for Decision Tree

RMSE values obtained for Decision Tree model on training set data is extremely small at 1.94e-20, while the RMSE values for the test set is larger, albeit still smaller compared to LR model at 4.36e-06.

## **2.3 Random Forest Model:**

Random Forest is an ensemble learning method for classification and regression tasks by constructing a multitude of decision trees at training time and outputting the class or mean/average prediction of individual trees.

In Random Forests, each node in the decision tree works on a random subset of features to calculate the output. The Random Forest then combines the output of individual decision trees to generate the final output.

The main supposed advantage that random decision forests have over decision trees is that they correct for decision trees’ habit of overfitting to their training set. This further explains the extremely small values of the RMSE obtained for random forests model.

**Advantages of Random Forest:**

* Random forest is a model that is widely applicable and suitable for different tasks and needs. It can handle binary, categorical, and numerical features. However, in our case, I have already normalised all the predictor variables in the dataset before the input into the models.
* Random Forests is great with multi or high dimensional data since it works with subsets of data
* Random Forests generally is faster to train than decision trees due to the model only working on a subset of features in this model. Prediction speed is significantly faster than training speed as the user can supposedly save generated forests for future uses.
* Each decision tree has a high variance, but low bias. As the model averages all the trees in random forest, we average the variance as well so that we have a low bias and moderate variance model.

**Disadvantages of Random Forest:**

* Random Forest models are not easily interpretable

**Performance of Random Forest Model:**

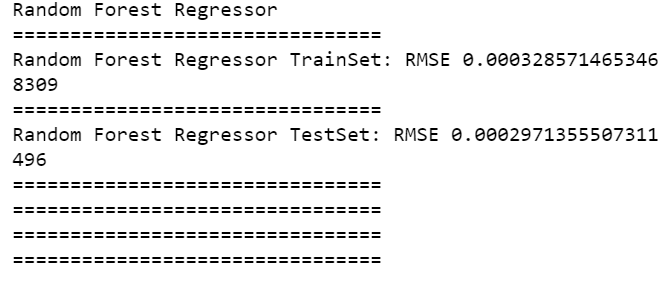


Figure 22: RMSE values of Random Forest

The RMSE values obtained for TrainSet and TestSet are 0.000329 and 0.000297 respectively.

## **2.4 XgBoost model:**

XGBoost stands for extreme gradient boosting; it is an implementation of gradient boosted decision trees designed for speed and performance. It is an extremely popular and successful model and is the go-to algorithm for competition winners for predictive modelling problems. The reason for XGBoost’s success and popularity is the fact that the model is an ensemble technique where new models are added to correct errors made by existing models. Models are added sequentially until no further improvements can be made. It uses a gradient descent algorithm to minimise the loss when adding new models. Thus, the model is more accurate compared to single decision tree models.

**Advantages of Xgboost:**

* Regularisation: XgBoost has built-in Lasso Regression and Ridge Regression regularisation which prevents model from overfitting.
* Cross Validation: the model allows users to run a cross-validation at each iteration of boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run.

**Disadvantages of XgBoost:**

* Boosting algorithms are sensitive to outliers since every classifier is obliged to fix the errors in the predecessors.

**Performance Of XgBoost:**

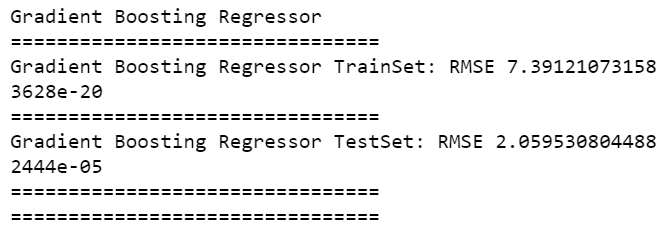


Figure 23: RMSE values for XgBoost

The RMSE values obtained for XgBoost on the TrainSet is 7.39e-20, while the values on the TestSet data is 2.06e-0.5.

## **2.5 Analysis and comparison of results:**

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE for TrainSet** | **RMSE for TestSet** |
| **Linear Regression** | 1.12e-05 | 3.48e-05 |
| **Decision Tree** | 1.94e-20 | 4.36e-06 |
| **Random Forest** | 0.000329 | 0.000297 |
| **XgBoost** | 7.39e-20 | 2.06e-0.5 |

Based on results tabulated as shown above, we can see that Decision Tree performed the best on this dataset for Airbnb listings, while Random Forest performed the worst.

# **Applications and Recommendations:**

The general results for the models tested on the datasets are favourable and it Airbnb may actually consider using the Decision Tree model in order to obtain price listing prediction in their operations. However, I suggest that all 4 models be used to obtain an averaged aggregate price score in the event that Airbnb does take up on this suggestion. This aggregation can be obtain by taking the weighted average of the price suggestion from each model. For example, a 40% weightage can be assigned to the best performing model (Decision Tree), and 25% to the second-best model etc. This way, we can minimise the effect and possibility of outliers which strains a particular model.

The practical implications and applications of this project can be explored in 2 main functions for Airbnb: an additional suggestion feature targeted at customers and a prices recommendation feature for hosts. This will be elaborated further subsequently.

1. **Customer Feature on App/Website:**

This feature will serve as a new front end feature that will indicate to the customers whether the current listings is priced at a more expensive or cheap rate compared to the expected result generated from our models. Furthermore, if we were to improve our model even further, we can even assign weights to the importance of the respective different predictor variables. This will enable the users to know the reasons as to why a listing may be priced higher or cheaper compared to similar listings or the expected pricing determined by our models. For example, location may appear to be a bigger determinant than the number of reviews for a listing. Thus, we can display the reasons as to why a host will price his/her apartment at a higher price. This will allow customers to make a more informed decision and will invariably lead to higher ratings on the Airbnb website, resulting in possibly an increase in revenue and improvement in brand loyalty to Airbnb due to this new feature.

1. **Backend Price Recommendation for hosts:**

This function will serve to recommend a price for a host’s listing before he/she actually publicises it on the Airbnb website. Based on the features and background of the host’s listing, the model will generate a suggested price for the host that is based on the performance of past listings. This will allow the host to better price their listings, and more importantly give first-time hosts a good gauge of a price to list.

# **References:**

* [**https://christophm.github.io/interpretable-ml-book/tree.html**](https://christophm.github.io/interpretable-ml-book/tree.html)
* [**https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a#:~:text=Decision%20tree%20often%20involves%20higher,regression%20and%20predicting%20continuous%20values**](https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a#:~:text=Decision%20tree%20often%20involves%20higher,regression%20and%20predicting%20continuous%20values)**.**
* [**https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#:~:text=Each%20node%20in%20the%20decision,to%20generate%20the%20final%20output.&text=The%20Random%20Forest%20Algorithm%20combines,to%20generate%20the%20final%20output**](https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#:~:text=Each%20node%20in%20the%20decision,to%20generate%20the%20final%20output.&text=The%20Random%20Forest%20Algorithm%20combines,to%20generate%20the%20final%20output)**.**
* [**http://theprofessionalspoint.blogspot.com/2019/03/advantages-of-xgboost-algorithm-in.html**](http://theprofessionalspoint.blogspot.com/2019/03/advantages-of-xgboost-algorithm-in.html)