

**Module 6 Final Project: Report**

**ALY6110: Data Management and Big Data**

|  |
| --- |
| **Name** |
| **Kavish Shah** |
| **Tanusha Erpula** |

1. **Dimension of the Dataset**

We are using One-dimensional datasets to give insights into Airbnb Listings in Melbourne, Victoria, Australia. The dimension of the dataset is 18,237\*75

Listings include some of the names of the columns that are id, name, host\_id, host\_name, neighbor’s group, neighborhood, latitude, longitude, room\_type, price, minimum\_nights, number of reviews, last\_reviews, reviews\_per\_month, calculated\_host\_lisitngs, availability\_365, number\_of\_reviews\_Itm, and license.

1. **Introduction**

**Data types**

Numeric Values: - 44

Categorical Values: -31

• **Rationale of the Dataset:** To get a good analysis of the business side, we have taken a wanted successful company Airbnb (a rental system). The dataset of Airbnb contains a rich amount of information that can help recognize great deterministic features. We have considered the most popular city in Australia which have 74. interrelated columns that can provide meaningful insights. We have all the main elements of Airbnb’s business process from the host to customer reviews. Since this dataset is huge, we might add or reform our goals initially mentioned.

**Tools Used**: - Python Language, Tableau.

**Data Cleaning**

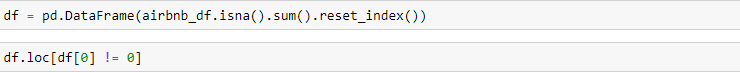
Checking for Null values:

Graphical user interface, text

Description automatically generated

The above command helps to check the null values in the dataset.

Checking for N/A values which are not zero:

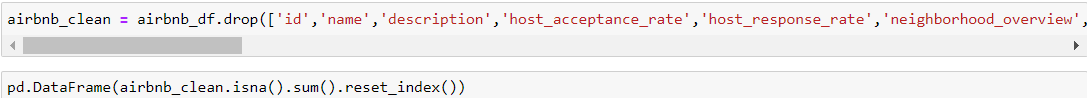


Graphical user interface

Description automatically generated with medium confidence’

This table describes the values with N/A values without zeros.

Dropping unnecessary values:



Graphical user interface

Description automatically generated with medium confidence

The above table shows the columns with unnecessary values.



Text, letter

Description automatically generated

**3. Questions to Investigate:**

* What is the price distribution of Airbnb in different neighborhoods in Australia?
* What is the total number of rooms distributed in the neighborhood of Australia?
* What is the relationship between neighborhood & Ratings based on price & room type?
* What is the relationship between availability with the price of the property in the neighborhood of Australia?
* What is the total number of Airbnb listed according to the neighborhood of Australia?
* What is the Correlation plot to understand the relationship between each variable?

**Exploratory Data Analysis**

### Price distribution for avability\_60 concerning the geographical locations

A picture containing scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

The above scatterplot depicts the price distribution of the properties according to the geographical locations in Australia.

### Total number of rooms distributed over the Australian neighborhood

Text

Description automatically generated

Chart, histogram

Description automatically generated

The above bar graph explains the number of room distributions available over the Australian neighborhood.

### Relationship between neighborhood & Ratings based on price & room type

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

The scatterplot displays the relationship between ratings based on room type and ratings with the neighborhood in Australia.

### Relationship between availability & price

Graphical user interface

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

The plot explains the relationship between the price of the property with the availability in the neighborhoods of Australia.

### Relationship between room type & availability

A picture containing chart

Description automatically generated

Chart, bar chart

Description automatically generated

This bar graph explains the relationship between room type and availability in the neighborhoods of Australia.

### Bar plot denoting the availability based on property type

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

The bar plot explains the availability of the property type in Australia.

### Total number of Airbnb listed according to neighborhood

Text

Description automatically generated

Table

Description automatically generated with low confidence

The bar plot displays the number of listings according to the neighborhood.

### Correlation plot to understand the relationship between each variable

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

**Analysis**

A picture containing chart

Description automatically generated

We have created a basic analysis using Tableau for the Airbnb dataset. I have used various measures and records to build multiple worksheets and create a dashboard. My dashboard consists of 5 worksheets and each worksheet displays a certain question. My first worksheet tries to answer the question ‘Highest Rating for Entire Home/Apt with clean Neighborhood’. This worksheet shows the highest rating for a certain type of property with a clean vicinity. The highest rating is 49.27 for an entire home/apt property type in Melbourne.

Chart, bar chart

Description automatically generated

The second worksheet tries to answer the question ‘Entire Bungalow available which accommodates a maximum number of people in different parts of Australia ‘. This worksheet gives the number of maximum occupants in Kew, Victoria, Australia. The maximum occupancy can be 8 in an entire bungalow house type.

Chart, treemap chart

Description automatically generated

The third worksheet answers the question ‘Houses/apt with no of bathrooms availability all around the year’. This worksheet shows that 1 bathroom type is the maximum number available throughout the year.

Chart, bubble chart

Description automatically generated

The fourth worksheet answers the question ‘Average price of various room types in Australia’. This sheet explains the prices of the various room types with their average price. This sheet concluded that the entire house/apt property type has the highest average price of $263.9 and the lowest average price is $109.7 for a shared room.

Chart

Description automatically generated with medium confidence

The fifth worksheet answers the question ‘Host Response time for a minimum no of nights. This worksheet works to display how long a host takes to respond depending on the days of occupancy of the rooms. This concludes that generally, hosts are taking few or more days to respond if the occupancy time is less compared to the host’s response time to maximum days of occupancy.

**Dashboard**

**Graphical user interface

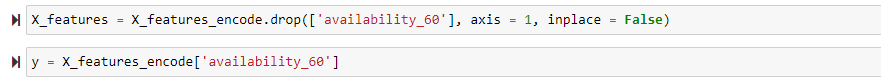
Description automatically generated**

we created a dashboard titled ‘Basic analysis of Airbnb dataset’ with all the worksheets embedded in it. We have added annotations to the worksheets to highlight each worksheet’s conclusions.

**Predictive Modeling**

For prediction, we have taken the target variable as the availability of Airbnb for 60 days in a different neighborhood in Australia. After cleaning the dataset taken the remaining columns are taken as Features.

### Splitting features and labels into X\_features & y variables



### Checking if data is normally distributed

Chart, histogram

Description automatically generated

### Using a log to normalize the data, the graph is shown below:



Chart, histogram

Description automatically generated

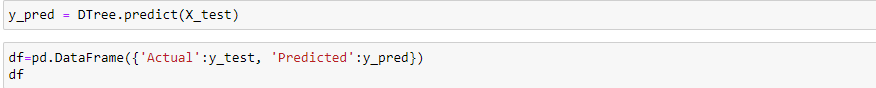
### **Decision Tree Regressor**

It is possible to visualize choices and all of their potential outcomes, including outcomes, input costs, and utilities, using a decision-making tool named a decision tree.

The classification algorithms group includes the decision-tree algorithm. It works with output variables that are categorized and continuous.

Graphical user interface, text, application

Description automatically generated



Graphical user interface, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

### **Ridge Regression**

Any dataset that exhibits multicollinearity can be analyzed using the model's tuning technique known as ridge regression. This technique carries out L2 regularization. Estimated values differ much from real values when the problem of multicollinearity arises, least-squares are unbiased, and variances are significant.

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

The above graph

**Advantages and Disadvantages of Models**

Advantages

* Decision trees take less work to prepare the data during pre-processing than other methods do.
* A decision tree does not require the normalization of data and scaling of data as well
* Additionally, the construction of a decision tree is not significantly impacted by missing values in the data.
* Technical teams and stakeholders can understand a decision tree model very quickly.

Disadvantages

* A slight change in the data can result in a big change in the decision tree's structure, which can lead to instability.
* Because of its intricacy and lengthier training period, decision tree training is relatively expensive.
* When compared to other algorithms, a decision tree's calculations can occasionally become significantly more complex.
* For the Decision Tree algorithm, regression applications and continuous value forecasts are insufficient.

**Model Comparison:**

Table

Description automatically generated

The above table explains the Random mean square error (RMSE), mean square error (MSE), and variance score. Among all the models Random Forest Regressor gives an accuracy of 98% whereas the decision tree regressor gives an accuracy of 96.5%.

**Business Recommendation**

By analyzing the dataset, it can be inferred that as the number of features in Airbnb are increasing the price is decreasing. Hence Airbnb can opt for a pricing model wherein the prices increase with an increase in the number of features.

**Conclusion:**

According to the above analysis, we can successfully conclude that the Random Forest Regressor worked the best and gave a variance score of 0.98 whereas the Ridge Regression model gave a variance score of 0.90. Hence the values predicted by Random Forest Regressor gave us close to accurate results. It is because the random forest uses many different samples in decision trees and that results in low variance. It can be concluded that the predicted value of the availability of Airbnb is close to the actual value.

# Bibliography

Airbnb. (2022, October 16). *Inside Airbnb .* Retrieved from http://insideairbnb.com/get-the-data/

*Create a Dashboard*. Online Help. (n.d.). Retrieved October 15, 2022, from https://help.tableau.com/current/pro/desktop/en-us/dashboards\_create.htm

Tableau Community Forums. (n.d.). Retrieved October 15, 2022, from <https://community.tableau.com/s/idea/0874T000000HBAwQAO/detail>

*Sklearn.linear\_model.Ridge*. scikit. (n.d.). Retrieved October 29, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.Ridge.html

*Sklearn.tree.decisiontreeregressor*. scikit. (n.d.). Retrieved October 29, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html