BUSAN 302
Group Project Report

Personalising guest experiences at Airbnb's

Dataset: Airbnb

Lab group Fri 8-10-E

Hajra Iqbal - hiqb867
Jennifer Wang - jwan594
Deema Sabbagh - dsab252
Quyet Xuan Dinh - dquy728
Jingyi Wang - jwan479

Table of contents

Executive Summary

In our project, we developed a machine learning model that provides personalised accommodation recommendations based on previous guest experiences. Our project began with comprehensive data preprocessing, which included the identification of variables that did not provide useful insight into guest satisfaction, leading to the removal of variables like the attraction index and restaurant index. We also eliminated rows with null values and standardised columns to accommodate variations in value ranges. After looking at the data, the most suitable model, a logistic regression, was used to categorise guest satisfaction into categories like "very bad," "bad," "not bad," "good," and "excellent." The results highlighted the model's effectiveness in predicting guest satisfaction, with moderate precision in the "excellent," "bad," "good," and "very bad" categories. However, the "not bad" category exhibited lower precision. The model excelled in recalling instances of "excellent" but had challenges with "bad," "good," and "not bad." The overall accuracy of the model was 84%, indicating its ability to classify guest satisfaction. Nevertheless, variations in performance across satisfaction categories implied that while it performed well in predicting extreme satisfaction levels, it encountered challenges in predicting intermediate levels of satisfaction. We concluded that this is because of the biassed dataset. Our analysis revealed the complex nature of guest satisfaction, suggesting it is influenced by multiple variables rather than a single variable. The limitations in predicting intermediate satisfaction levels are attributed to the complex and subjective nature of guest satisfaction. To enhance the model's performance, we propose expanding the dataset to have a fairer and more balanced sample and integrating text-based guest reviews and feedback to provide a richer understanding of guest satisfaction and potentially improve prediction accuracy, especially for intermediate satisfaction levels.

Introduction

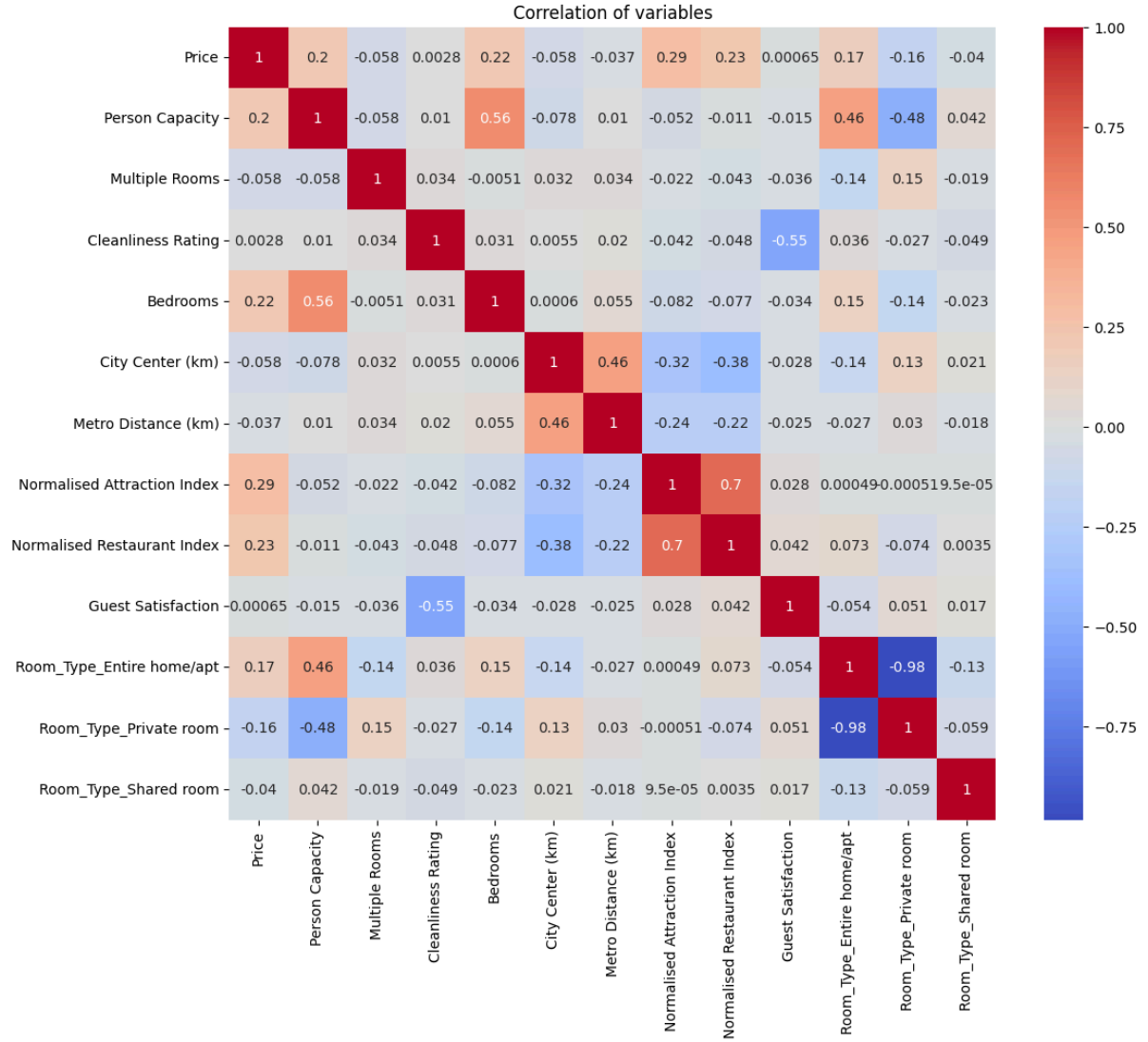
In this report, we present the development and evaluation of a machine learning model aimed at enhancing guest satisfaction in the accommodation industry. Our project leverages data preprocessing and logistic regression to predict guest satisfaction levels. We investigate the complex and multifaceted nature of guest satisfaction and identify areas where the model excels and encounters challenges. This analysis guides our recommendations for potential model improvements, offering insights into the factors that shape guest satisfaction within Airbnb. We will present an analysis of where our model excels and where it faces challenges, providing valuable insights into the details of what makes an Airbnb stay truly satisfying.

Descriptive Statistics

We chose variables to investigate based on our assumptions of their effectiveness and relevance when it came to predicting guest satisfaction. The variables we chose included price, person capacity, multiple rooms, cleanliness rating, bedrooms, city center (km), metro distance (km), normalised attraction index, normalised restaurant index, room type (entire room/ apt), room type (private room) and room type (shared room). We chose these variables as we believed that each variable would have some influence on guest satisfaction ratings. We believed that guests would be affected in all aspects of the Airbnb when it came to prices, and whether the Airbnb was able to meet the price expectations would influence their decisions on whether they were satisfied or not. For person capacity, multiple rooms, bedrooms, and the three room type variables, we chose these because we thought that it would reflect the Airbnb's flexibility and ability to accommodate guests and the Airbnb's space, which would then influence guest satisfaction. In terms of cleanliness rating, we believed that this would definitely influence guest satisfaction as it is usually one of the first things guests pay attention to when renting an Airbnb, and a low cleanliness will lead to a negative impact thus low guest satisfaction. For city center and metro distance, they intertwine with each other in some aspects. Guests who are closer to the city center and a metro would usually be required to pay a little more in terms of price for Airbnb's, therefore this would be relevant as this would lead to an impact on guest satisfaction. Normalised attraction and restaurant indexes represents the popularity or quality of various restaurants and attractions in different cities. The team chose to select them instead of the standard indexes because a normalised index is typically adjusted/scaled to facilitate easier comparison and analysis in the context of machine learning.

Correlations

```
#heatmap
test.dropna()
plt.figure(figsize=(12, 10))
corr_matrix = test[['Price', 'Person Capacity', 'Multiple Rooms', 'Cleanliness Rating',
                    'Bedrooms', 'City Center (km)', 'Metro Distance (km)',
                    'Normalised Attraction Index', 'Normalised Restaurant Index',
                    'Guest Satisfaction','Room_Type_Entire home/apt',
                    'Room_Type_Private room', 'Room_Type_Shared room']].corr()
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", cbar=True)
plt.title("Correlation of variables")
```



A heatmap was generated to view the correlation between variables.

As we can see from the heatmap, both the attraction index and restaurant index have very low correlation to the guest satisfaction rating. Therefore these two variables were dropped.

However, most of the variables have similarly low correlation ratings against the guest satisfaction. None of them were strongly related to guest satisfaction. We looked into why this could be and we thought that it could be because individual variables would not have a large effect on the guest satisfaction and instead it is the combination of variables that contribute strongly to the rating. For example, the weakest correlation of the price and the guest satisfaction (correlation coefficient of 0.00065) could mean that just because guests are paying a higher price for the airbnb it does not mean that they are more likely to be satisfied with it. Their rating on the property depends on multiple factors such as the cleanliness, bedrooms and distance from the city centre.

Before running the logistic regression model, we cleaned the data by removing rows that contained null values. We also encoded the 'guest satisfaction' column as this was non numeric.

```
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
test['Guest Satisfaction'] = encoder.fit_transform(test['Guest Satisfaction'])
```

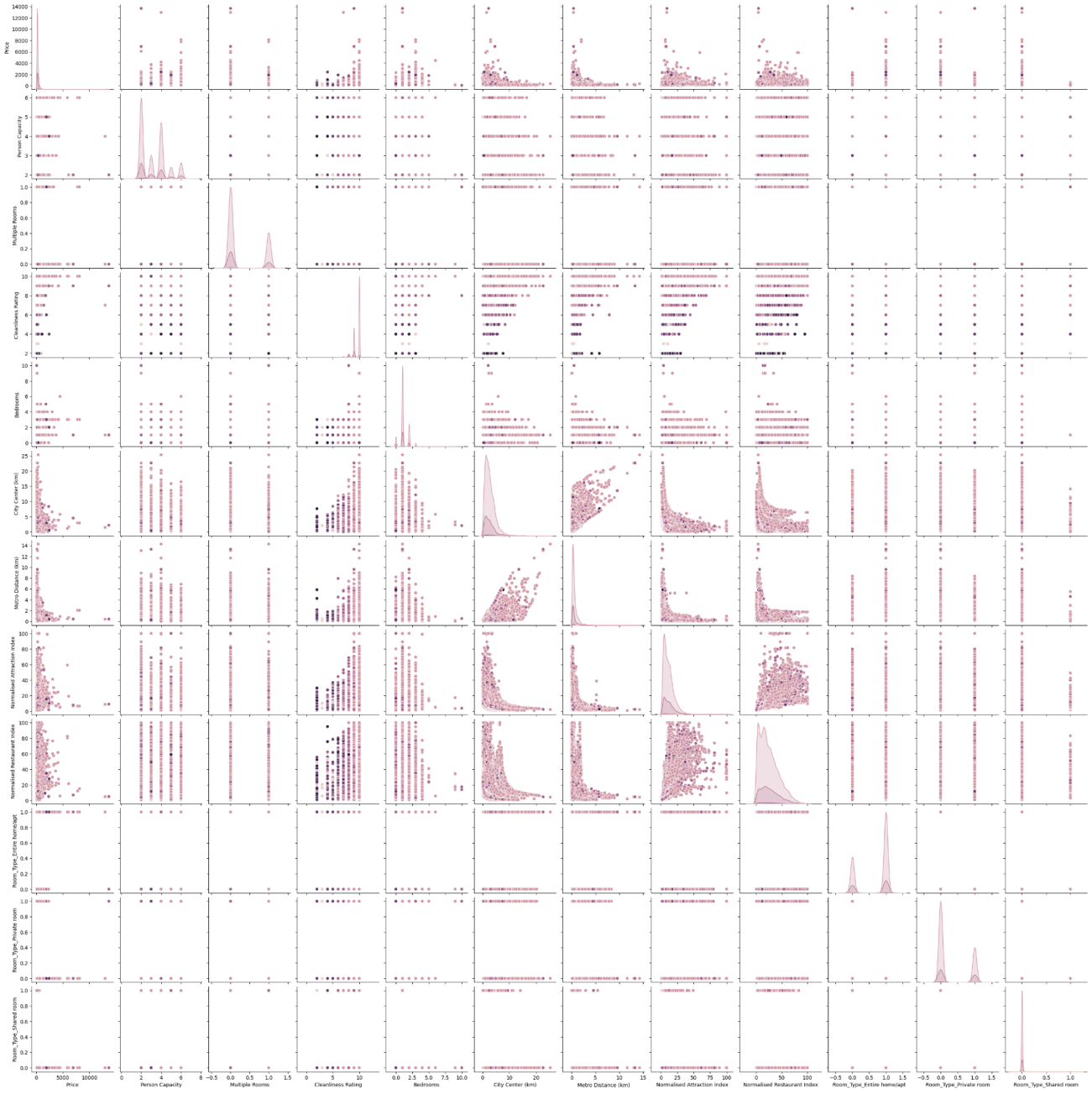
Additionally, we standardised the columns. The reason for this was because there was a vast difference in the range of different variables. For example, the person capacity was in a range of 1-6 while the distance from price ranged in double to triple digit numbers. This would have made the model favour and assume that the price carries higher weight due to the different scales.

```
#standardise data
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()
scaler.fit(x)
x = scaler.transform(x)
x = pd.DataFrame(x)
x.columns = ['Price', 'Person Capacity', 'Multiple Rooms', 'Cleanliness Rating', 'Bedrooms',
             'City Center (km)', 'Metro Distance (km)', 'Normalised Attraction Index',
             'Normalised Restaurant Distance', 'Room_Type_Entire home/apt', 'Room_Type_Private room',
             'Room_Type_Shared room']
x
```

Pairplot

```
#pairplot
test.dropna()
sns.pairplot(test[['Price', 'Person Capacity', 'Multiple Rooms', 'Cleanliness Rating',
                   'Bedrooms', 'City Center (km)','Metro Distance (km)',
                   'Normalised Attraction Index', 'Normalised Restaurant Index',
                   'Guest Satisfaction','Room_Type_Entire home/apt',
                   'Room_Type_Private room',  'Room_Type_Shared room']])
plt.figure(figsize=(6,4))
plt.show()
```



Proportion of classes in the dataset.

Bad: 82/41714 = 0.002% (0)
Excellent: 33012/41714 = 79% (1)
Good: 8024/41714 = 19.2% (2)
Not bad: 505/41714 = 0.012% (3)
Very bad: 91/41714 = 0.0022% (4)

The pairplot shows the correlations between independent variables, coloured by the guest satisfaction classes. The imbalance of the data has resulted in the dominance of the Excellent class, as it takes up most of the data in the dataset.

For variables such as Price, City Centre (km), Metro Distance (Km), Normalised Attraction Index, Normalised Restaurant Index, Room_Type_Shared room, the data distributions have the same characteristics of right-skewness. Each class does have the same distribution as

each other and there are no distinctive patterns from any classes. This indicates complex relationships between variables that are non-linear.

Through the pairplot, we can conclude that the independent variables individually do not have any strong effect on the target variable. But because of the huge bias towards the “Excellent” class, the plot does not show any distinct patterns between the 5 classes. The “Excellent” class dominates the dataset.

Model Results

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)

logistic_model = LogisticRegression()
logistic_model.fit(x_train, y_train)
```

```
#defining x and y
#removing rows with null values and removing irrelevant columns
test_cleaned = test.dropna() #price, city center and metro distance have nulls
x = test_cleaned.drop(['Business','Attraction Index', 'Restaurant Index', 'Guest Satisfaction'], axis=1).values
y = test_cleaned[['Guest Satisfaction']].values
```

A confusion matrix and classification report was carried out to evaluate accuracy of predictions by the model.

```
from sklearn.metrics import classification_report, confusion_matrix
```

```
[19] #confusion matrix

     cnf_matrix = confusion_matrix(y_test, y_pred)
     cnf_matrix


     array([[   3,    1,   10,    1,    3],
            [   0, 6832,  194,    0,    0],
            [   0, 1051,  588,    4,    1],
            [   1,   23,   80,    5,    1],
            [   0,    1,    4,    3,   11]])
```

For class 1 (excellent), there are 6832 true positives, whereas for class 4 (very bad), there are 11 true positives. Therefore this shows that the model can much more easily predict the excellent class.

Additionally, class 3 (not bad) was incorrectly predicted as class 2 (good) 1051 times. This can be something to look into as it is a large number. A possibility for this is that variables that guests may consider as ‘not bad’ or ‘good’ may be similar, showing the difficulty for the model to distinguish classes based on the variables.

```
y_pred = logistic_model.predict(x_test)


classification_rep = classification_report(y_test, y_pred)


print("Classification Report:\n", classification_rep)
```

```
Classification Report:
               precision    recall  f1-score   support

           0       0.75      0.17      0.27        18
           1       0.86      0.97      0.91      7026
           2       0.67      0.36      0.47      1644
           3       0.38      0.05      0.08       110
           4       0.69      0.58      0.63        19

    accuracy                           0.84      8817
   macro avg       0.67      0.42      0.47      8817
weighted avg       0.82      0.84      0.82      8817
```

Through the classification report of the model, we can see how the model performs in predicting Guest Satisfaction with the chosen independent variables.

Precision score:
The precision score shows out of total predicted positives, how many are actually positive. This is given for each of the 5 classes of guest satisfaction ratings that the model is trying to predict.

- The class that has the highest accuracy is “Excellent” with a precision score of 0.86 which indicates that it has predicted the class correctly 86% of the time, a very good percentage. However, this is because the model was trained on an unbalanced dataset where 79% of the values were ‘excellent’, making it more likely to accurately predict this class.
- “Bad”, “Good” and “Very Bad” classes have decent precisions with 0.75, 0.67 and 0.69 respectively. As the numbers indicate that the accuracies are over 60%, it shows that the model is still relatively accurate in predicting these classes.
- The class that has the lowest precision is “Not Bad” with 0.38. By only predicting the class 38% of the time right, it shows that the model struggles in accurately predicting this class. This is expected as only a small percentage of the training data in the guest satisfaction ratings are classified as not bad (0.012%).

Recall score:
A recall score shows out of the total actual positives, how many are predicted correctly as positive. This is also given for each of the 5 classes of guest satisfaction ratings.

- The class that has the highest recall score is “Excellent” with 0.97 which indicates that it has identified 97% of the actual “Excellent” instance, which it has captured most of the actual instances in the class. Again this was expected due to over representation of this class in the dataset.
- “Very Bad” has a moderate recall with 0.58. It indicates that it has correctly identified 58% of the actual “Very Bad” instances.

- “Bad”, “Good” and “Not bad” have low recall with 0.17, 0.36 and 0.05 consecutively, which indicates that they have missed most of the actual instances of their classes.

Below is an indication of the proportion of classes in the dataset.
Bad: 82/41714 = 0.002% (0)
Excellent: 33012/41714 = 79% (1)
Good: 8024/41714 = 19.2% (2)
Not bad: 505/41714 = 0.012% (3)
Very bad: 91/41714 = 0.0022% (4)

F1 Score: Indicates the balance of precision and recall

- The class that has the best balance between precision and recall is “Excellent” with an f1 Score of 0.91. It indicates that the model can predict accurately the class with a high proportion of the predicted instances and correctly identified most of the actual cases, maximising true cases and minimising the false cases.
- “Good” and “Very Bad” have decent balances with f1 scores of 0.47 and 0.63 consecutively.
- “Bad” and “Not Bad” have low balance with f1 scores of 0.27 and 0.08 consecutively. It indicates that the models have trouble in predicting and identifying the actual cases for these two classes, especially for “Not bad” with both precision and recall are well below 50% of the total instances of the class.

Overall:
The model shows an accuracy of 0.84 which indicates that 84% of the total 8817 cases are correctly classified. However, although the model results are good, this model cannot be trusted due to the highly unbalanced dataset. This is a large factor that contributes to the reliability of the model.

Some of the low f1 scores indicate that there are some classes where the model struggles to identify the cases, especially for “Bad” and “Not bad”, which the model missed most of the actual instances.

Discussion / Conclusion

Although the accuracy rate of the classification has 84% overall accuracy rate, the results also show some clear limitations. The model's performance varied across satisfaction categories, performing best for 'excellent' and worst for 'not bad'. This suggests that the model is good at predicting extreme ratings but needs help to predict moderate levels of satisfaction accurately. This is because imbalanced datasets in this regression models are challenging the performance of machine learning models. In regression, skewed target distributions can affect prediction accuracy.

From the pairplot and heatmap, the low correlation between individual variables and guest satisfaction implies that guest satisfaction is a complex, multifaceted concept influenced by a variety of other factors. It is not entirely dependent on any one variable, which means that it only takes the interaction of multiple factors to produce satisfaction.

The results of the analysis show that although the logistic regression model can make fairly accurate predictions of most guest satisfaction, it has limitations (intermediate satisfaction). This may be due to the complexity and subjectivity of guest satisfaction, as well as the fact that the model is more likely to predict things with distinct characteristics.

The local as well as the global context has to be taken into account to ensure predictive accuracy on the basis of the model itself. Factors such as local culture and tourist preferences can vary greatly from place to place. Understanding these local dynamics is therefore critical to the effective application of the model. The regulations and policies of the location of the homes in Airbnb affect operations. Guest satisfaction can also be heavily impacted and therefore a key factor in the success of a host. The model is also affected by global travel industry trends that influence travellers' expectations and the factors they take into account when assessing satisfaction. At the same time global economic conditions, including downturns or upturns, can affect predicted traveller satisfaction. This is because guest preferences and satisfaction criteria may change during different periods of economic conditions.

In order to improve the performance of the model, there are some measures that can be of some help. The dataset can be resampled with more balanced data in the bad, very bad and not bad classes. More data can help the model learn the underlying patterns more effectively. The dataset should consider the inclusion of text-based guest reviews and feedback, as these unstructured data sources may contain detailed information about guest satisfaction to improve understanding of intermediate satisfaction characteristics and prediction accuracy.

In conclusion, the model's current performance is good but shows some limitations due to the data being biassed and the model made less impact on the individual variables than we would have expected. By taking the recommendations, the accuracy of the model has the potential to improve and provide more insight into what factors determine guest satisfaction in the context of Airbnb.