Hotel Booking Cancellations Prediction Report

BA WITH R (BUAN 6356.003)

GROUP NUMBER: 10

- **LIKITH VINAYAKA GIRIDHAR (LXG210034)**
- **PRAFUL PATIL (PVP220001)**
- **SAI ROHIT BOGGARAPPU (SXB220120)**
- **RAHUL KINTALI (RXK210144)**

Hotel Booking cancellations prediction Report.

Business context:

The dataset includes information from two distinct hotels - a Resort hotel and a City hotel - both situated in Portugal, a country in southern Europe. As per the publication available at (https://www.sciencedirect.com/science/article/pii/S2352340918315191), the Resort hotel (H1) is situated in the Algarve region, while the City hotel (H2) is located in Lisbon. These two locations are approximately 280 km apart by car and both are situated on the north Atlantic coastline. The dataset pertains to bookings scheduled to arrive between July 1st, 2015 and August 31st, 2017.

Content Exploration of our dataset:

The dataset at hand contains a target variable labelled "is_canceled" which represents binary classes 0 and 1, where 0 denotes "Not Cancelled" and 1 denotes "Cancelled". The remaining variables in the dataset provide additional details about the bookings, including lead time, arrival date, length of stay, number of guests, type of meal, country of origin, market segment, distribution channel, room type, deposit type, and more. These variables may have numerical, categorical, or binary values.

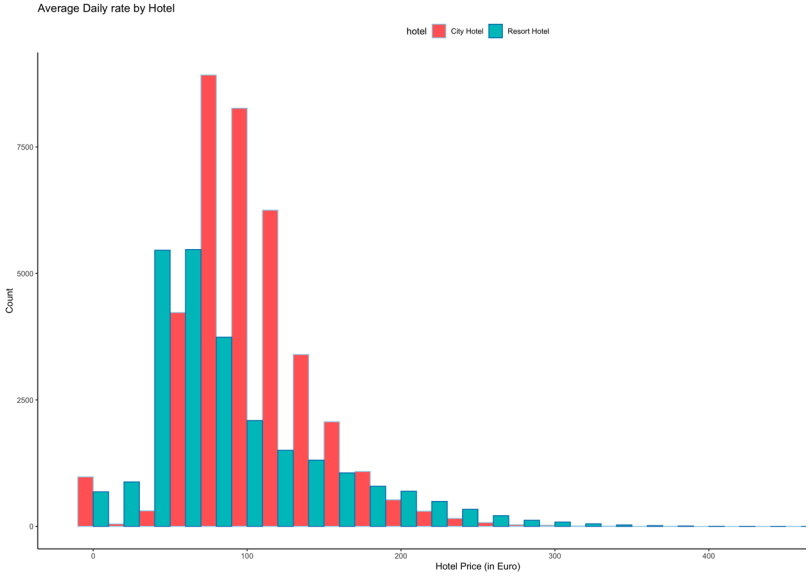
The dataset is presented in a tabular format, with rows representing individual bookings and columns representing attributes of the bookings. This layout allows for easy organization and analysis of the data. The dataset has the potential to be valuable for analysing patterns in hotel bookings, predicting future bookings, and identifying factors that contribute to cancellations. By examining the various attributes of the bookings, insights can be gained into guest behaviour, booking trends, and potential strategies for optimizing hotel operations and revenue management.

Hotel Booking cancellations prediction Report.

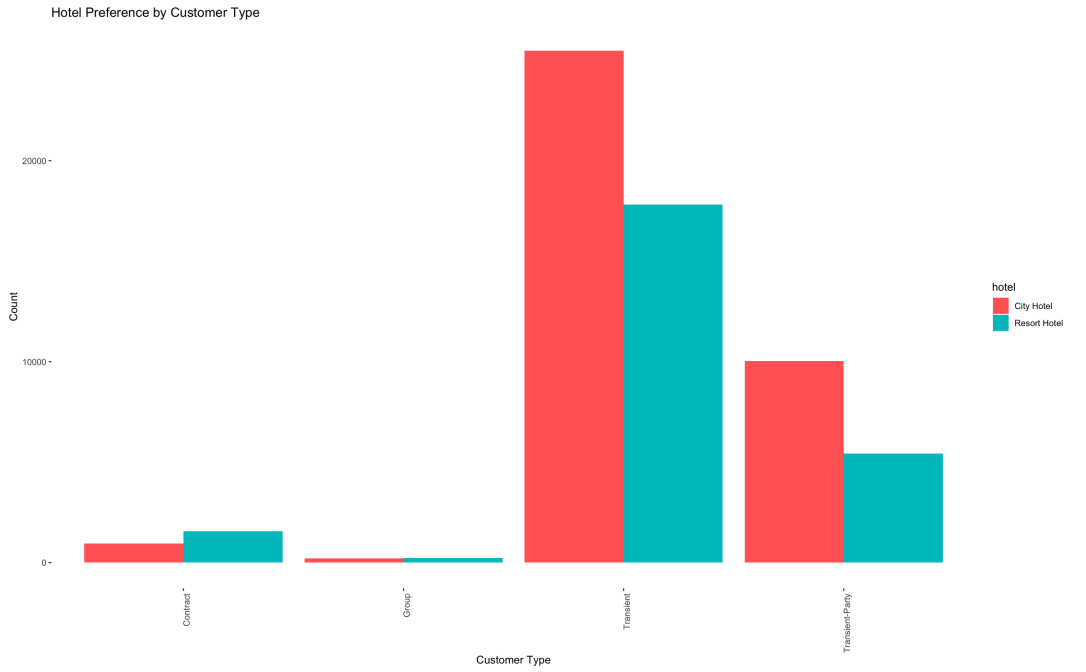
Following are the questions we decided to ask of the dataset:

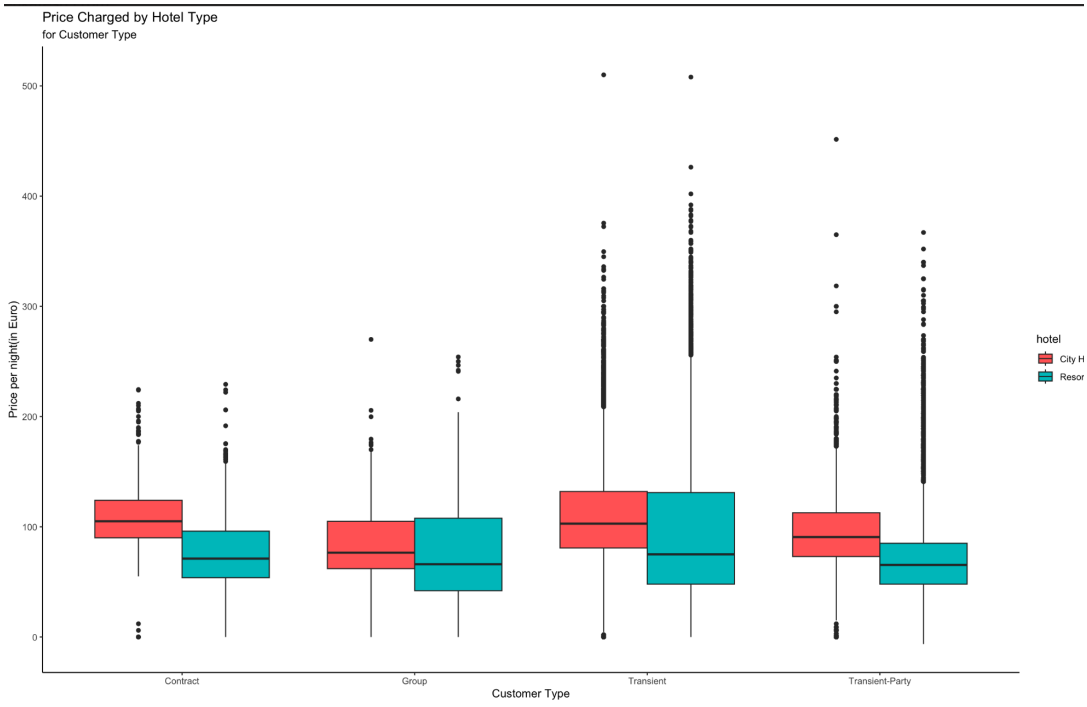
1. **Where do the guests come from?**
   Finding. As part of our analysis, we subset the data to include the countries which has more than 1500 reservation request otherwise including all the country with few or occasional request to avoid the graph from being clumsy. Our findings revealed that Portugal had the highest number of tourists among all the countries examined and the other countries where the highest number of tourists arrived were from Portugal's neighboring countries in Europe.

2. **How much do guests pay for a room per night?**
   Finding. When compared to Resort hotel, people at city hotel tend to pay higher mainly because City hotels are often located in prime areas of urban centers and are commonly chosen by business travelers who require proximity to offices, conference centers, and other business-related facilities. The highest number of guests are from transient customer type bookings and lowest being for group customer type bookings. This trend can be seen for both the hotel types. Prices charged by both the hotel types for different customer types remains the same. The spread of outliers is extremely large for Transient and Transient-Party



Hotel Booking cancellations prediction Report.

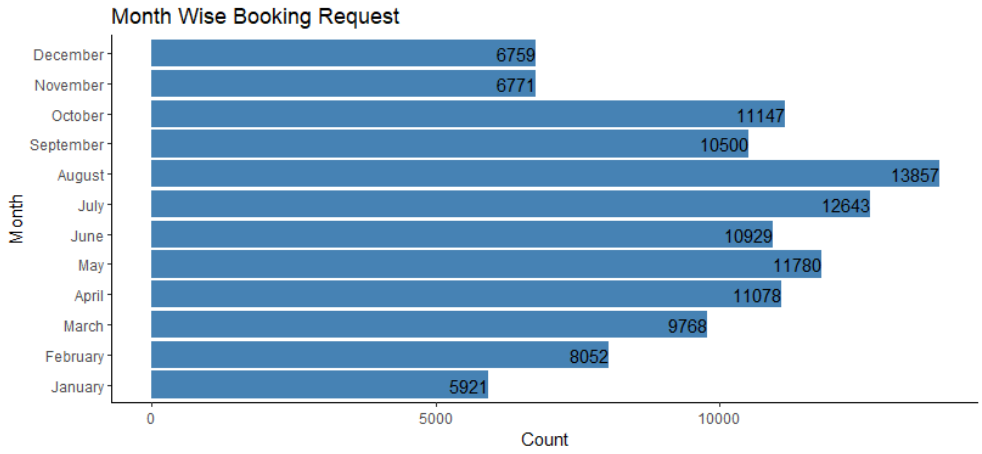


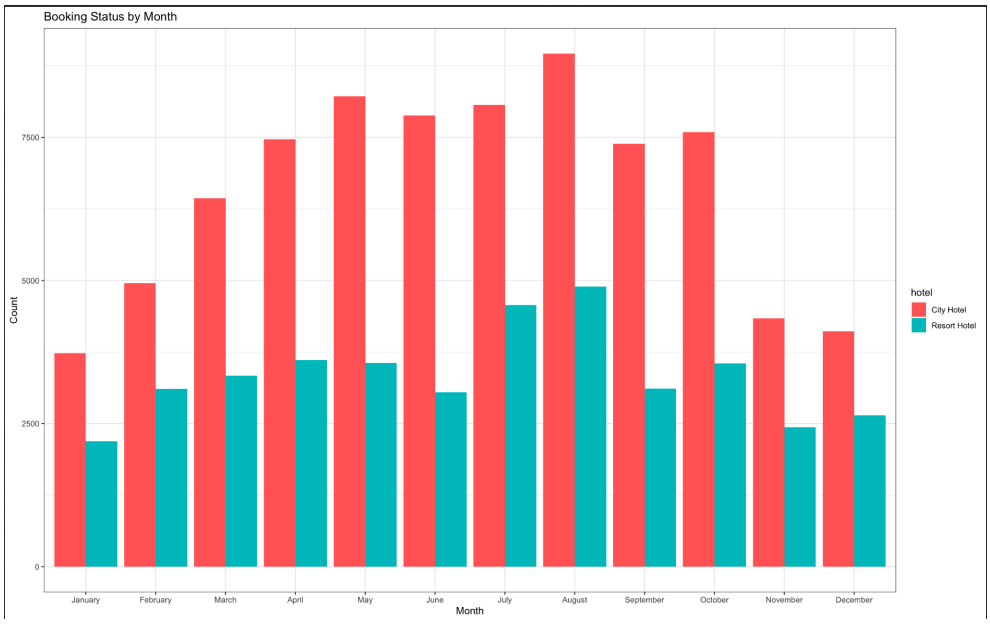


3. **Which is the Busiest Month?**

   Findings. For the City hotel, we can see that the bookings consistently remain around 7500 or more from April to October, and the bookings drop during the winter months. This trend is very similar in the Resort hotel, but the bookings hover around half of the bookings seen in the City hotel during the same period.

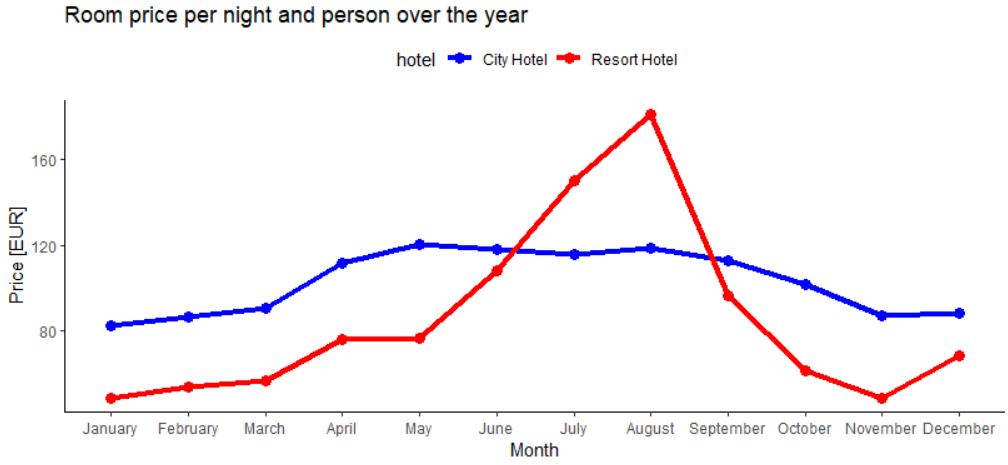
Hotel Booking cancellations prediction Report.





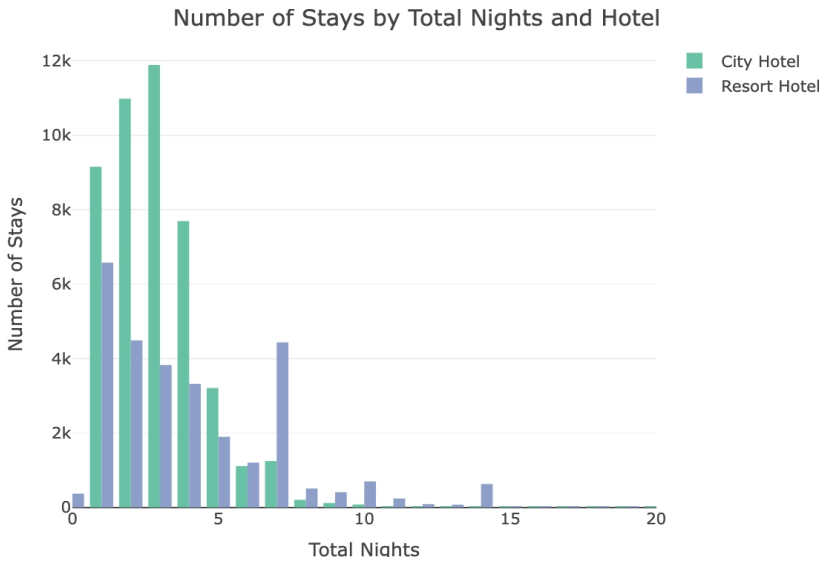
4. **How does the price per night vary over the year?**

   Finding. We can notice that this is a peak and valley plot. The price per night is maximum in August and Minimum in January and November for resort hotel. For Resort hotel the price hikes from May to August and drops till November. This is also because of the points mentioned in finding 3.



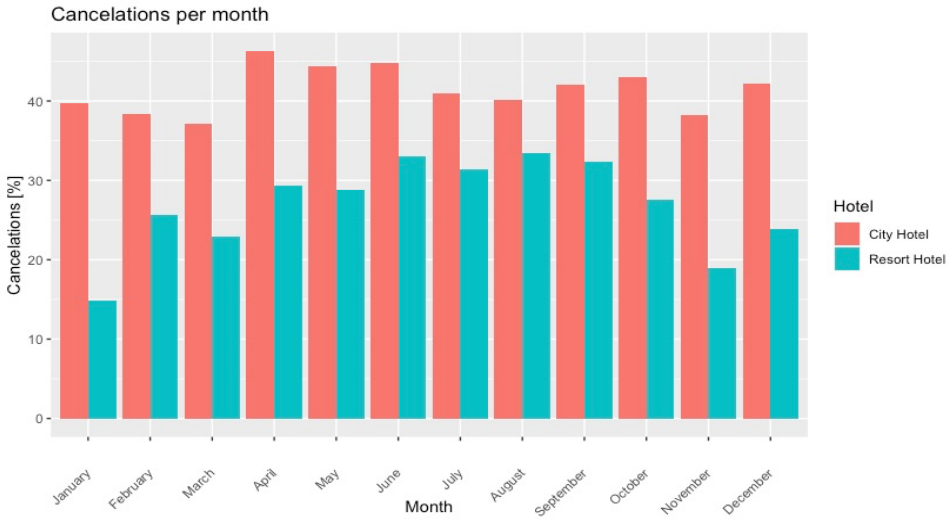
Hotel Booking cancellations prediction Report.

5. **How long do people stay at the hotels?**
   Finding. Based on the analysis of the dataset, it is evident that most guests tend to stay for approximately 3 nights on average at the hotel. Furthermore, the highest number of guests stay for a duration ranging from 0 to 5 days, indicating that the most common length of stay falls within this timeframe.



6. **How many bookings were canceled?**
   Finding. As per the analysis of the dataset, it was observed that a significant portion of the total bookings, accounting for 37%, were canceled. Out of the total cancellations, 28% were from the resort hotel bookings, while a higher percentage of 42% were from the city hotel bookings. This indicates that the city hotel experienced a higher proportion of cancellations compared to the resort hotel. This information provides valuable insights into the cancellation rates for each hotel type, which can be utilized for further analysis, revenue forecasting.

7. **Which month have the highest number of cancelations?**
   Finding. It is apparent that the cancellations for the months of April, May, June, September, October, and December are relatively similar, with April having the highest number of cancellations, albeit by a slightly larger margin.



Hotel Booking cancellations prediction Report.

Tasks Performed:

1. **Explored Interesting Libraries:**
   - **Dplyr**: Used for creating pipelines to build subsets of data frames and data frame summary.
   - **Tidyr**: Used in conjunction with dplyr to clean data.
   - **Plotly**: Used to create interactive plots (annotations)
   - **FactoMineR**: Exploratory data analysis methods to summarize, visualize and describe data.
   - **Countrycode**: Used for converting country names and codes from one format to another.
   - **Lubridate**: To manipulate date.
   - **Mice**: Used for imputing data.

2. **Data cleaning and manipulation for Visualization:**
   - Convert character columns to factors
   - Eliminate NA and other undefined values – Omit missing Records
   - Some rows contain entries with zero adults, children, and babies, so we deleted these records as they have Zero guests.
   - Meal column having “Undefined” values have been replaced by Mode value.
   - Records with “Undefined” values in “market_segment” and “distribution_channel” have been deleted.

3. **Data Visualization:**
   - In addition to the above data cleaning and manipulation we have used various libraries such as dplyr, ggplot 2, tidy, plotly, FacotoMineR, countrycode, lubridate to extract the required subsets from the dataframe and plot various charts such as Bar charts, Histograms, Boxplots, Line chart, horizontal bar chart, stacked bar chart, heat maps etc to draw various insights from the data.

4. **Data Pre-Processing and Cleaning for building the model:**
   - Plotting Heat Map to check for correlation in numeric data: We observe that none of the numeric variables are highly correlated to the target variable “is_canceled”
   - Based on Intuition, we remove some of the irrelevant columns such as “agent”, “company”, “reservation_status_date”
   - We also remove the categorical variable “reservation_status” because of 100% correlation with our target variable “is_canceled”
   - We replace missing values with the mode for categorical variables, we impute children=”0”, country=”Unknown”, meal=”SC”
   - We replace missing values of numeric variables with their median, for variables “lead_time” and “required_car_parking_spaces”

5. **Feature Engineering:**
   - We combine arrival_date_year, arrival_date_month, arrival_date_day_of_month to form a new variable called arrival_date

Hotel Booking cancellations prediction Report.

- Create a new variable called total_guests by adding the number of adults, children, and babies.
- Then we convert the categorical variables to factors

6. **Principal Component Analysis:**
   - We perform Principal Component Analysis on numeric variables.
   - We merge the PCA values with the categorical variables.
   - We use the first 3 components which capture majority of the data.

7. **Building the Model:**
   - We split the dataset into training and testing set with the split factor of 0.7 for training.
   - In Logistic Regression Model we observe the following performance metrics when we test on the test data:

Confusion Matrix:

	Actual		
Prediction		0	1
	0	21287	6315
	1	1216	6943

By Class Statistics:

Accuracy	: 0.7894
No Information Rate	: 0.6293
Kappa	: 0.5099
Sensitivity	: 0.9460
Specificity	: 0.5237
Pos Pred Value	: 0.7712
Neg Pred Value	: 0.8510
Prevalence	: 0.6293
Detection Rate	: 0.5953
Detection Prevalence	: 0.7718
Balanced Accuracy	: 0.7348
F1 Score	: 0.6483634
AUC	: 0.7348

The accuracy on prediction on training data and testing data is 0.78 which is the same as the accuracy obtained on testing data. In general, these metrics suggest that the model has good sensitivity but poor specificity, meaning that it is better at correctly identifying positive cases than negative cases.

Hotel Booking cancellations prediction Report.

- We also build a **Random Forest Model** of 500 trees we observe the following performance metrics when we test on the test data:

  Type of random forest: classification

  Number of trees: 500
  No. of variables tried at each split: 4
  OOB estimate of error rate: 16.52%

Confusion matrix:

	Actual		
Prediction		0	1
	0	21012	4308
	1	1491	8950

By Class Statistics:

Accuracy	: 0.8378401
Kappa	: 0.636591
Sensitivity	: 0.6750641
Specificity	: 0.9337422
Pos Pred Value	: 0.8571976
Neg Pred Value	: 0.8298578
Prevalence	: 0.3707391
Detection Rate	: 0.2502726
Detection Prevalence	: 0.2919661
Balanced Accuracy	: 0.8044031
F1 Score	: 0.7553061
AUC	: 0.8046913

The accuracy obtained on training data is 0.908 and for test data it is 0.8378 which is a very good score. In general, the metrics suggest that the model has good sensitivity and specificity, meaning that it is able to correctly identify both positive and negative cases. The positive predictive value is high at 0.9356, indicating that when the model identifies a case as positive, it is correct most of the time. The negative predictive value is also relatively high at 0.8944, indicating that when the model identifies a case as negative, it is correct most of the time. The prevalence of the positive cases in the dataset is relatively low at 0.3708. The detection rate is 0.2989, indicating that the model correctly identified almost one-third of all cases. The detection prevalence is 0.3195, meaning that the model identified 31.95% of the cases as positive. Finally, the balanced accuracy of the model is 0.8867, which is a relatively high value representing the overall accuracy of the model.

Hotel Booking cancellations prediction Report.

- **Comparing the two Models:**

  Based on the obtained metrics, the Random Forest model appears to perform better than the Logistic Regression model for the hotel booking cancellation prediction problem.

1. **Accuracy:** The Random Forest model has an accuracy of 0.8378401, which is higher than the accuracy of 0.7894 for the Logistic Regression model. This suggests that the Random Forest model is able to correctly predict the outcomes more accurately overall.

2. **Kappa:** The Kappa score for the Random Forest model is 0.636591, which is higher than the Kappa score of 0.5099 for the Logistic Regression model. A higher Kappa score indicates better agreement between predicted and actual outcomes.

3. **Sensitivity:** The Random Forest model has a sensitivity (true positive rate) of 0.6750641, which is lower than the sensitivity of 0.9460 for the Logistic Regression model. Sensitivity measures the ability of a model to correctly identify positive cases. In this case, the Logistic Regression model has higher sensitivity, suggesting it is better at identifying actual positive cases.

4. **Specificity:** The Random Forest model has a specificity (true negative rate) of 0.9337422, which is significantly higher than the specificity of 0.5237 for the Logistic Regression model. Specificity measures the ability of a model to correctly identify negative cases. The higher specificity of the Random Forest model suggests it is better at identifying actual negative cases.

5. **Positive Predictive Value:** The Random Forest model has a higher positive predictive value (precision) of 0.8571976 compared to the positive predictive value of 0.7712 for the Logistic Regression model. This indicates that the Random Forest model is better at correctly predicting positive cases.

6. **Negative Predictive Value:** The Negative Predictive Value for the Random Forest model is 0.8298578, which is higher than the value of 0.8510 for the Logistic Regression model. Negative Predictive Value measures the ability of a model to correctly predict negative cases.

7. **F1 Score:** The F1 score for the Random Forest model is 0.7553061, which is higher than the F1 score of 0.6483634 for the Logistic Regression model. The F1 score is a measure of the trade-off between precision and recall, and a higher F1 score indicates a better balance between the two.

8. **AUC:** The Random Forest model has a higher AUC (Area Under the Curve) of 0.8046913 compared to the AUC of 0.734823287033679 for the Logistic Regression model. AUC is a measure of the model's ability to correctly classify cases, with a higher AUC indicating better performance.

Hotel Booking cancellations prediction Report.

Results and Conclusions

Based on the above metrics, the Random Forest model appears to be the better choice for the hotel booking cancellation prediction problem, as it has higher accuracy, Kappa score, specificity, positive predictive value, negative predictive value, F1 score, and AUC compared to the Logistic Regression model.

Future Work

- Performing oversampling or under sampling on the target variable and re-running the models to analyse performance.
- Check the dependence of target variable on each level of each of the categorical variables. Then select the necessary categories and one hot encode. Finally perform PCA on the one hot encoded categorical sub data frame.
- Explore Hyperparameter tuning to further tune and improve the performance of the model.