Hotel Booking System

Group -18

Project Report

Venkata Sasank Jonnalagadda

Rahasya Chandan

+1 612-450-3414

+1 832-914-5502

Jonnalagadda.ven@northeastern.edu

chandan.r@northeastern.edu

Percentage of Effort Contributed by J.v. Sasank: 50%

Percentage of Effort Contributed by Rahsya Chandan : 50%

Signature of J.v. Sasank: J.v.Sasank

Signature of Rahasya Chandan: Rahasya Chandan

Submission Date: 12/09/2022

Problem Setting:

Hotel industry is one of the businesses which is growing rapidly and in order to attract the customers, the owners are providing various schemes and offers like offering complimentary breakfast, free car parking, free cancellation policy etc. In order for them to make necessary decisions to improve their business, they should study the data they collect from the customers and take necessary actions. In this project we aim to solve a few business case problems and predict the solutions that help both the customers and owners.

Problem Definition:

For the purpose of this Project, we will be looking into:

1. The probability that a customer makes a cancellation.
2. Does the cancellation depend on the external factors - such as the lead time before the booking, booking changes made, the deposit type, and the customer type.
3. Finding out if being a repeat customer changes the probability that a customer does not cancel?

Data Sources:

The"Hotel Booking Demand" dataset an be found on Kaggle - https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

It is originally from the article Hotel Booking Demand Datasets

Data Description:

This data set contains booking information for a city hotel and a resort hotel, and it includes 32 attributes and 119391 instances that could predict the probability of the customer making a cancellation based on external factors.

It has information and variable names such as when the booking was made, length of stay, the lead time before the booking was made, the arrival date, if the booking was cancelled, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

Data Cleaning

Before data visualization, we perform a few data pre-processing steps:

- **Checking for null values**: We found that ‘children’, ‘country’,’ agent’ and ‘company’ had null values. We filled the null values with 0 for the columns ‘children’ and ’ agent’. (Used domine knowledge to replace Null values with 0)
- **Removing Features:** We removed the column ‘company’ as it is unwanted and does not help us with our target variable
- **Renaming Columns:** We renamed our ‘adr’ column to ‘average_daily_rate’ to make it more meaningful.

Before

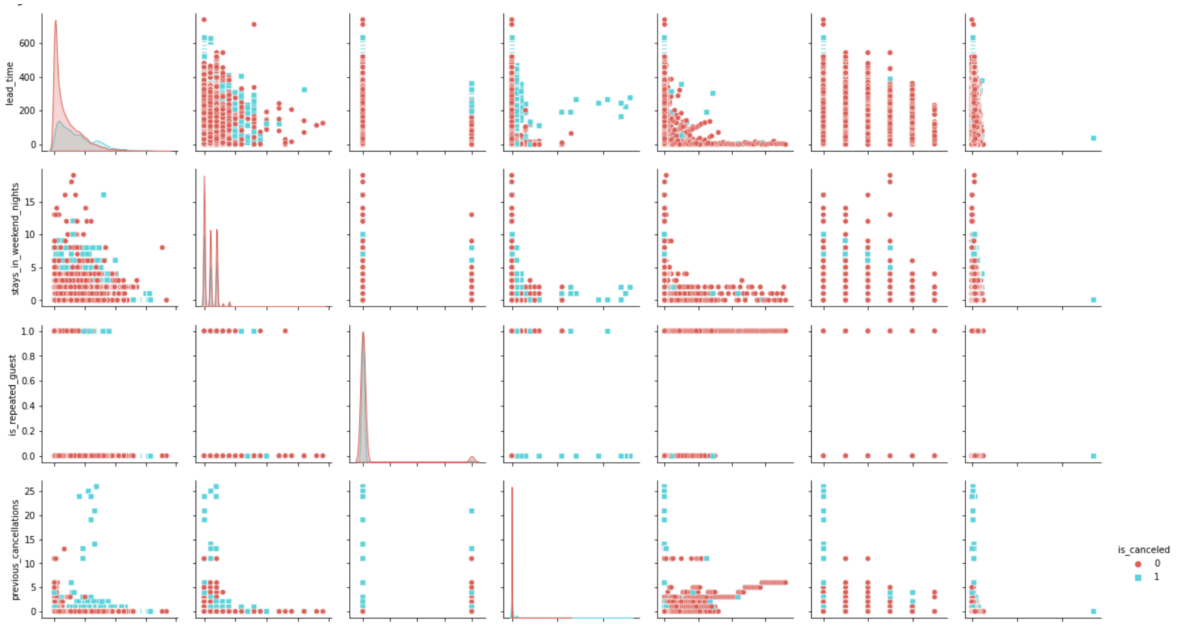
```
df.isna().sum()
```

```
hotel                               0
is_canceled                         0
lead_time                           0
arrival_date_year                   0
arrival_date_month                  0
arrival_date_week_number            0
arrival_date_day_of_month           0
stays_in_weekend_nights             0
stays_in_week_nights                0
adults                              0
children                            4
babies                              0
meal                                0
country                           488
market_segment                      0
distribution_channel                0
is_repeated_guest                   0
previous_cancellations              0
previous_bookings_not_canceled      0
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
agent                           16340
company                        112593
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
reservation_status_date             0
dtype: int64
```

After

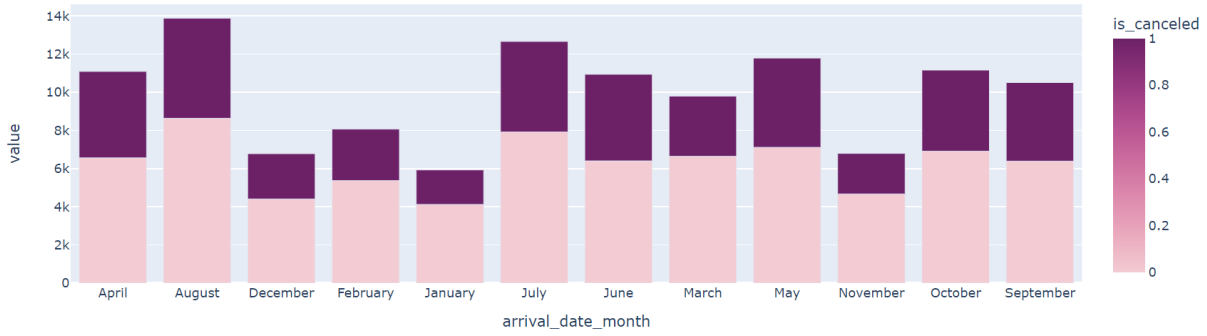
```
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date_year               0
arrival_date_month              0
arrival_date_week_number        0
arrival_date_day_of_month       0
stays_in_weekend_nights         0
stays_in_week_nights            0
adults                          0
children                        0
babies                          0
meal                            0
country                         0
market_segment                  0
distribution_channel            0
is_repeated_guest               0
previous_cancellations          0
previous_bookings_not_canceled  0
reserved_room_type              0
assigned_room_type              0
booking_changes                 0
deposit_type                    0
agent                           0
days_in_waiting_list            0
customer_type                   0
average_daily_rate              0
required_car_parking_spaces     0
total_of_special_requests       0
reservation_status              0
reservation_status_date         0
dtype: int64
```

Data Visualization



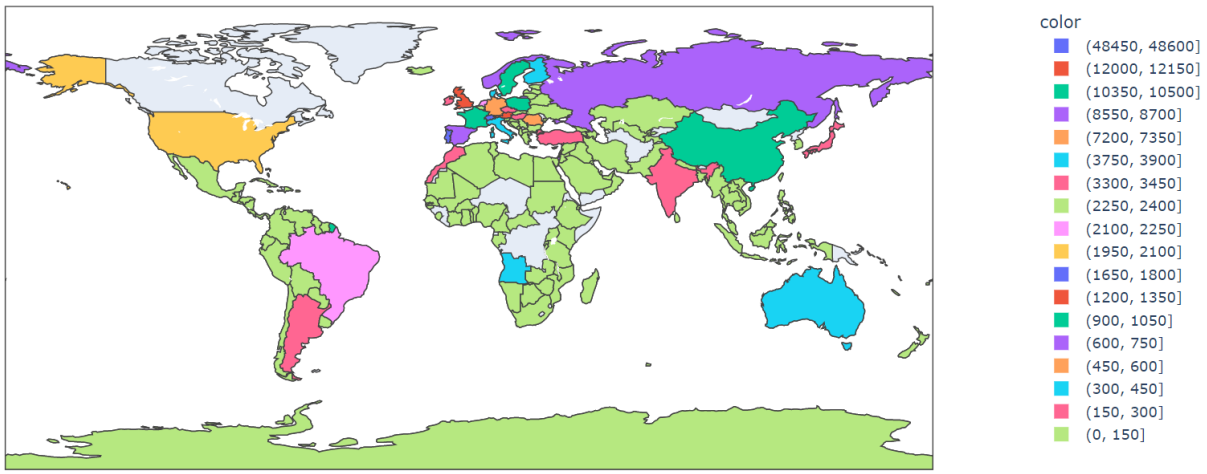
We perform a pair plot for a few selected variables to see the trend and patterns of the variables in the dataset.

On plotting the ‘Cancellation Overview’, we could see that a higher percentage of bookings have not been cancelled compared to the booking that were cancelled.



The above graph shows the total number of bookings in that particular month (Duration of 3 years). The stacked bar chart also shows the number of bookings that were cancelled as well as the number of bookings that were not cancelled. We can see that August has the highest number of cancellations.

Choropleth Map:

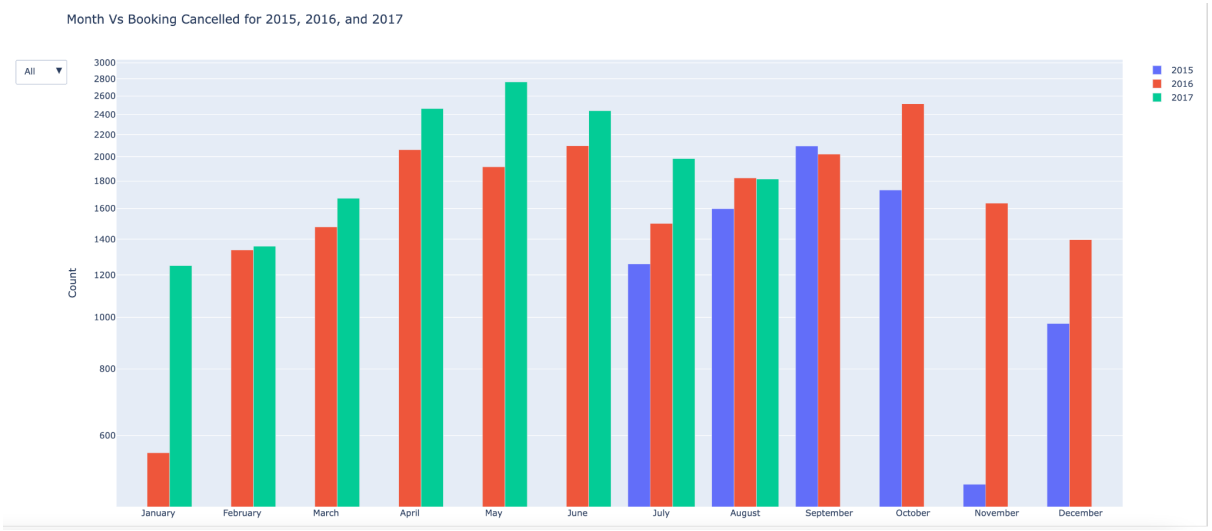


The Choropleth map shows the number of customers from each country. Since there are a wide range of numbers, we had to split into bins in order to have a meaningful interpretation from the choropleth map.We could observe that Portugal is the country with the highest number of customers.

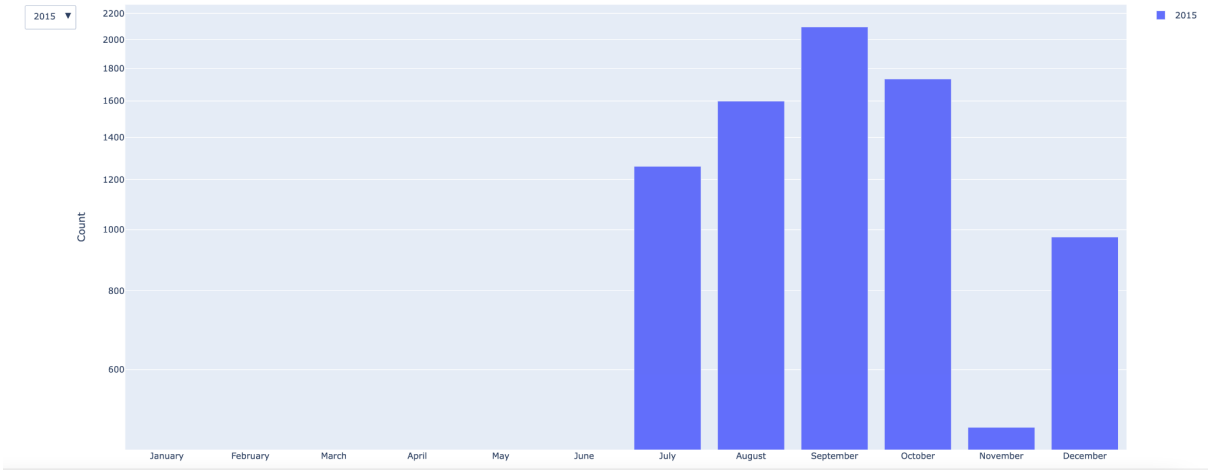
Bookings Cancelled for year 2015, 2016 and 2017:

We have plotted an interactive graph, so that we can see the number of cancellations for all three years and months simultaneously or for each year individually. We can also click on each month and see the number of bookings cancelled for each year. Looking at the graph,we

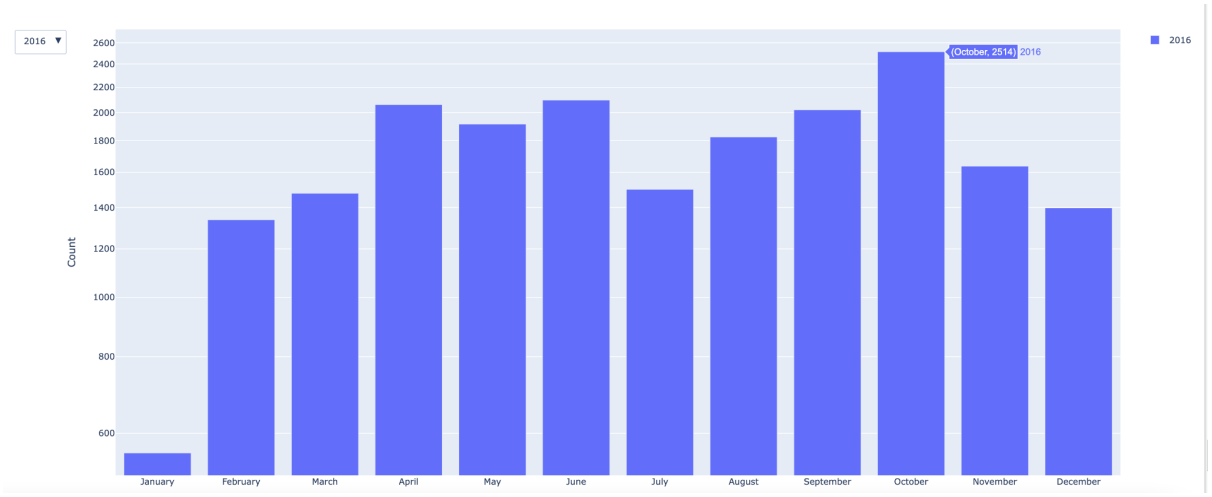
can see in the dataset that the year 2015 has only 5 months of data , 2016 has all 12 months and 2017 contains 8 months of data.



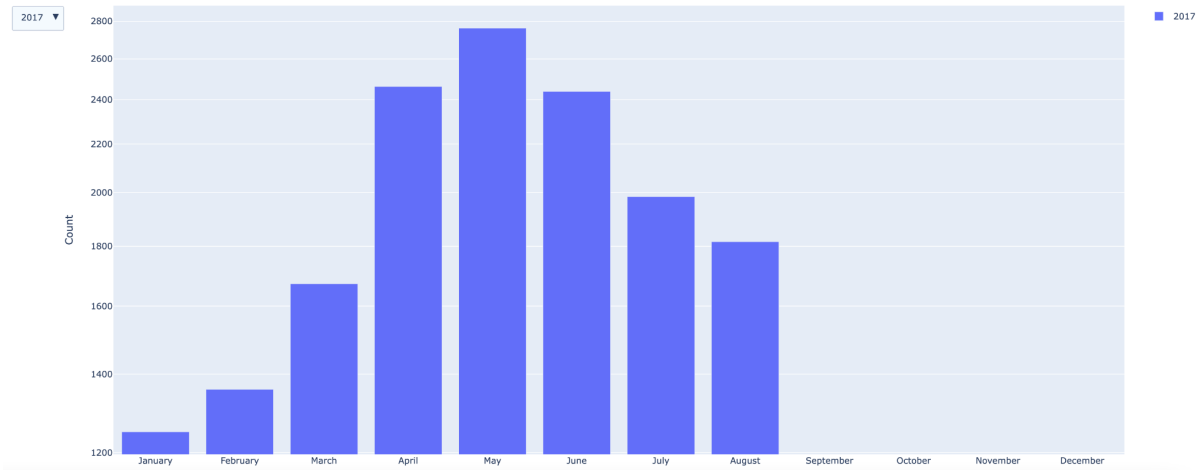
Booking Cancelled for 2015:



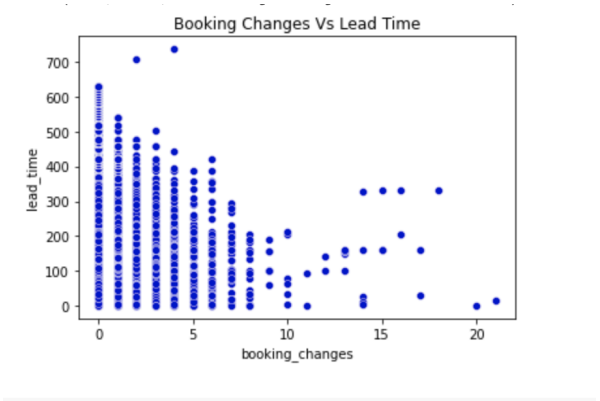
Booking Cancelled for 2016:

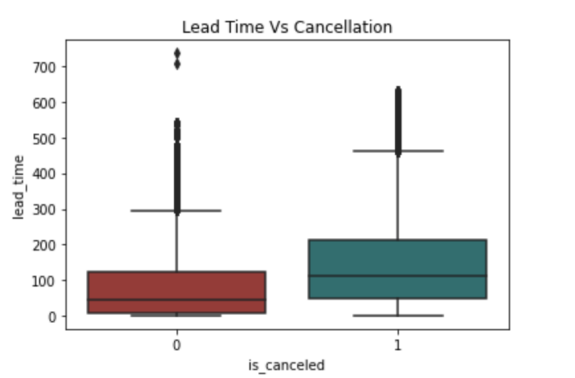


Booking Cancelled for 2017:



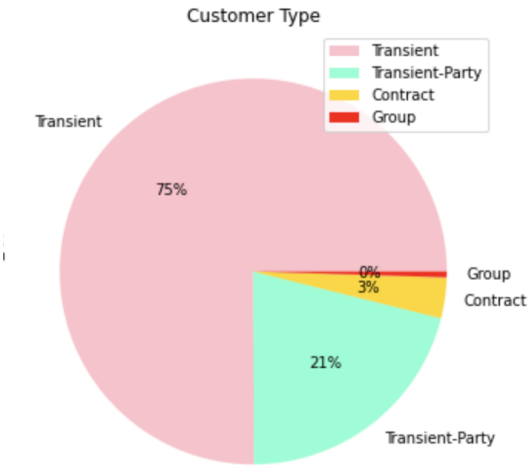
By looking at the above three graphs for all three years, we see that the number of booking cancellations is the lowest during January and December. We can interpret that, this is probably because the months of January and December are usually when customers go on vacation/holiday.

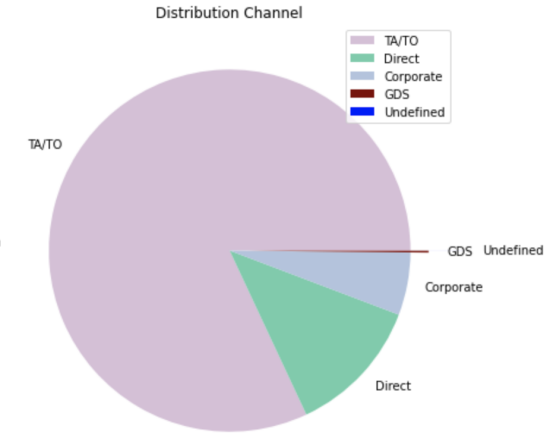




On plotting ‘Booking Changes Vs Lead time’, we see that bookings that are made in advance have lesser changes in their booking. While bookings that are more immediate in the future, have far more changes made.

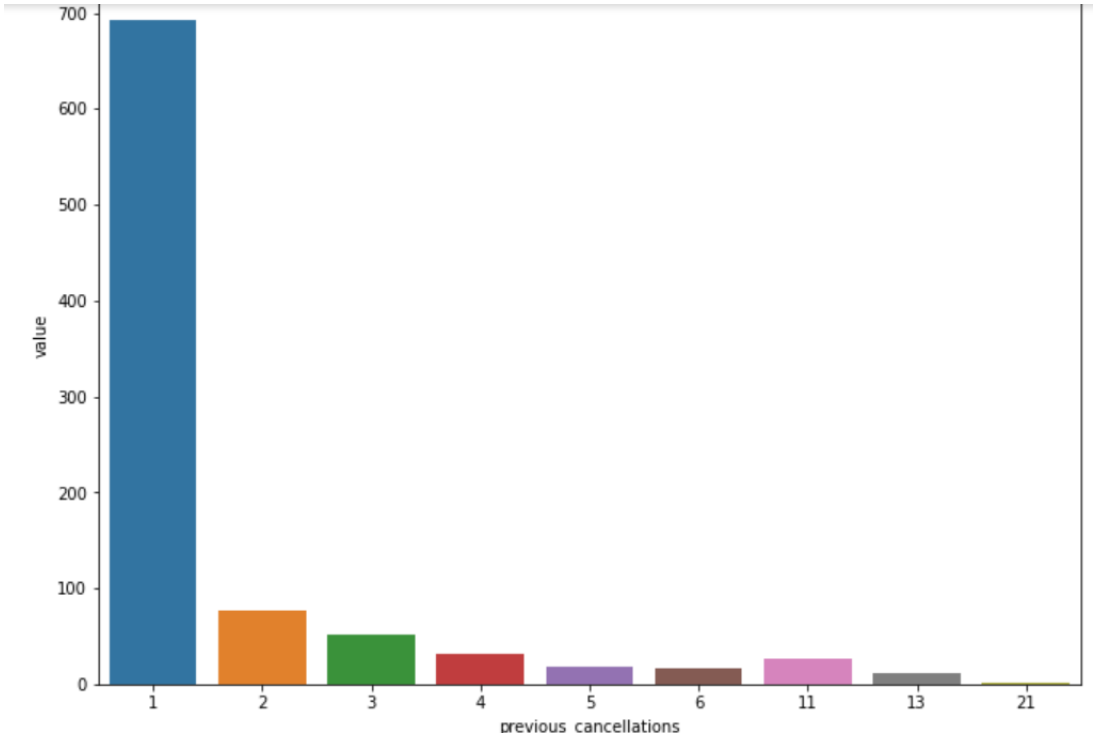
On plotting ‘Lead time Vs Cancelled’, we see for bookings that are cancelled, the higher the the time before the booking is made then there is a more chance the booking is cancelled. Looking at the graph, we also see that they are outliers and plan to remove them during data processing by checking lead times that are outside the upper and lower bounds and then setting that values null and then dropping those values.





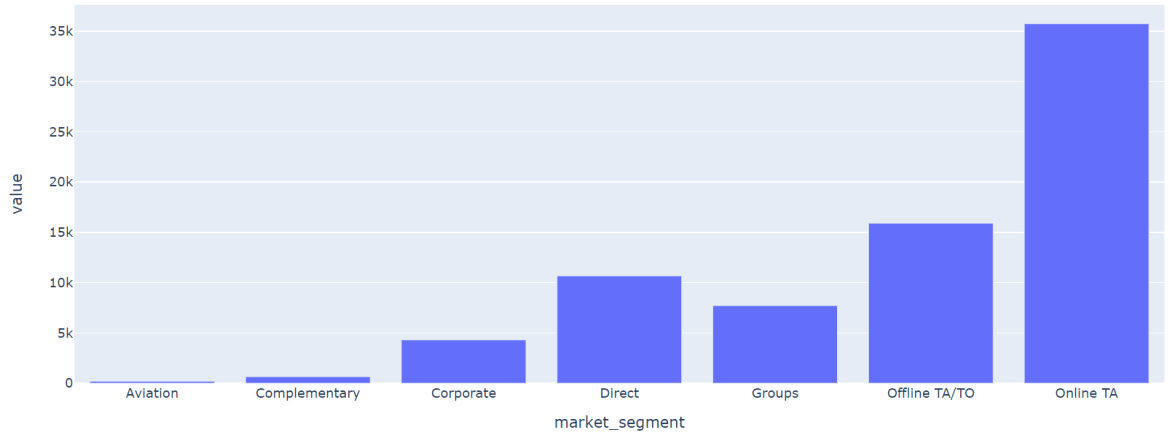
By looking at the 'Customer Type' pie chart, we see that most bookings by customers are transient. The transient customers are not part of a group or contract and are not related to any other transient booking. We also see that almost no bookings are made through a group. The 'Distribution Channel' pie chart shows that most of the bookings are made by travel agents or tour operators.

Number of times repeated customers cancelled their bookings



The above bar graph, represents the number of times the repeated customers cancelled their bookings. We could see that most of the repeated customers cancelled their bookings once.

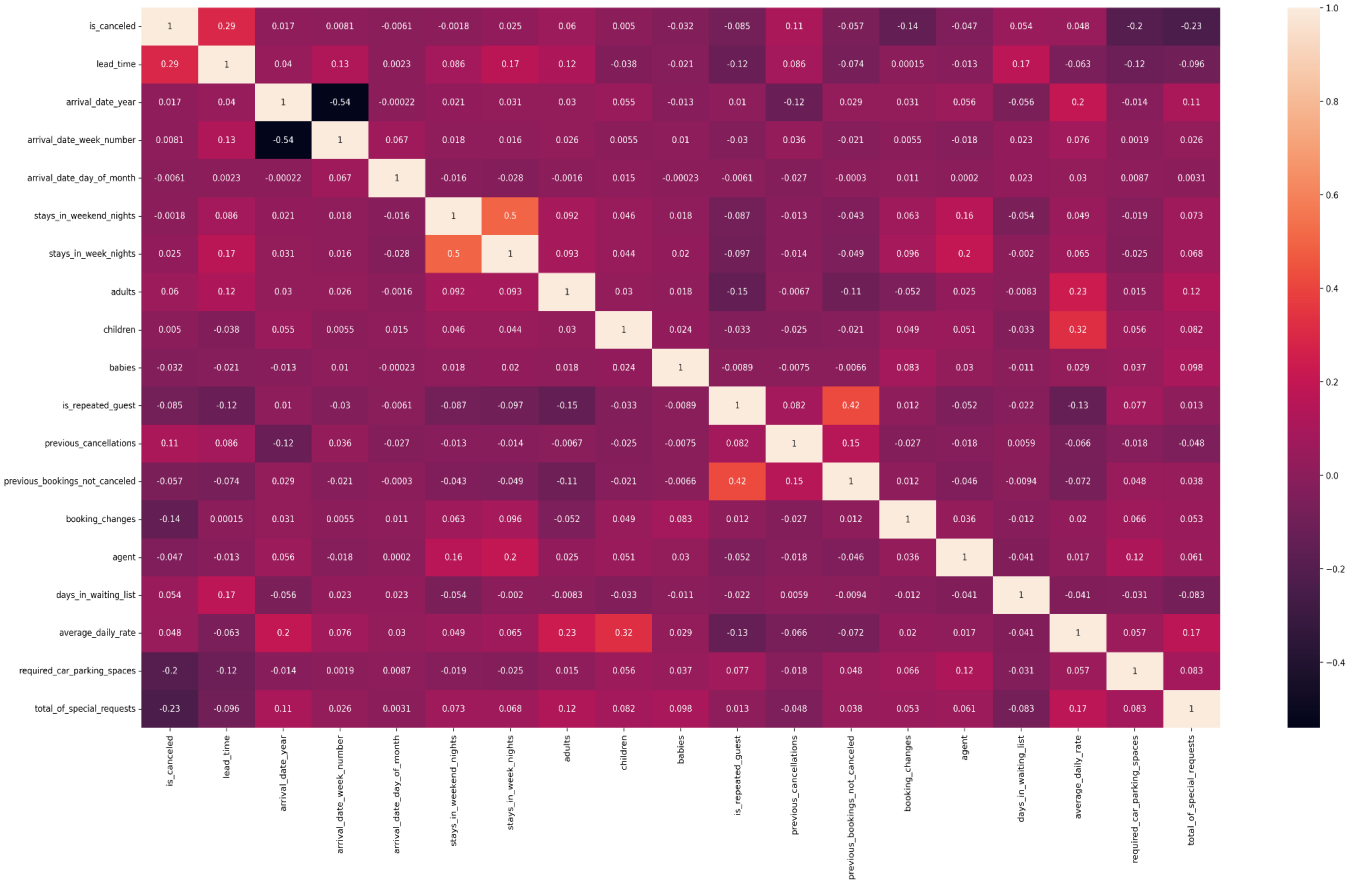
Bookings cancellation in respective market segments



Looking at the 'Booking Cancelled Vs Market Segment' graph, we see that most bookings cancelled by customers are through online tour operators who have made the hotel booking.

Correlation analysis:

Correlation Heatmap:



We can see by looking at the correlation heat map, column ‘is_canceled’ has high positive correlation with lead_time and a lower negative correlation with total_of_special_requests. ‘is_canceled’ also has a positive correlation with columns ‘previous_bookin_not_cancelled’.

```
(df1.corr())
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies
is_canceled	1.000000	0.293123	0.016660	0.008148	-0.006130	-0.001791	0.024765	0.060017	0.005036	-0.032491
lead_time	0.293123	1.000000	0.040142	0.126871	0.002268	0.085671	0.165799	0.119519	-0.037613	-0.020915
arrival_date_year	0.016660	0.040142	1.000000	-0.540561	-0.000221	0.021497	0.030883	0.029635	0.054636	-0.013192
arrival_date_week_number	0.008148	0.126871	-0.540561	1.000000	0.066809	0.018208	0.015558	0.025909	0.005515	0.010395
arrival_date_day_of_month	-0.006130	0.002268	-0.000221	0.066809	1.000000	-0.016354	-0.028174	-0.001566	0.014553	-0.000230
stays_in_weekend_nights	-0.001791	0.085671	0.021497	0.018208	-0.016354	1.000000	0.498969	0.091871	0.045794	0.018483
stays_in_week_nights	0.024765	0.165799	0.030883	0.015558	-0.028174	0.498969	1.000000	0.092976	0.044203	0.020191
adults	0.060017	0.119519	0.029635	0.025909	-0.001566	0.091871	0.092976	1.000000	0.030440	0.018146
children	0.005036	-0.037613	0.054636	0.005515	0.014553	0.045794	0.044203	0.030440	1.000000	0.024030
babies	-0.032491	-0.020915	-0.013192	0.010395	-0.000230	0.018483	0.020191	0.018146	0.024030	1.000000
is_repeated_guest	-0.084793	-0.124410	0.010341	-0.030131	-0.006145	-0.087239	-0.097245	-0.146426	-0.032858	-0.008943
previous_cancellations	0.110133	0.086042	-0.119822	0.035501	-0.027011	-0.012775	-0.013992	-0.006738	-0.024729	-0.007501
previous_bookings_not_canceled	-0.057358	-0.073548	0.029218	-0.020904	-0.000300	-0.042715	-0.048743	-0.107983	-0.021072	-0.006550
booking_changes	-0.144381	0.000149	0.030872	0.005508	0.010613	0.063281	0.096209	-0.051673	0.048952	0.083440
agent	-0.046529	-0.012640	0.056463	-0.018244	0.000202	0.161427	0.195135	0.024994	0.050581	0.030266
days_in_waiting_list	0.054186	0.170084	-0.056497	0.022933	0.022728	-0.054151	-0.002020	-0.008283	-0.033271	-0.010621
average_daily_rate	0.047557	-0.063077	0.197580	0.075791	0.030245	0.049342	0.065237	0.230641	0.324853	0.029186
required_car_parking_spaces	-0.195498	-0.116451	-0.013684	0.001920	0.008683	-0.018554	-0.024859	0.014785	0.056255	0.037383
total_of_special_requests	-0.234658	-0.095712	0.108531	0.026149	0.003062	0.072671	0.068192	0.122884	0.081736	0.097889

Looking at the correlation analysis table, we see how strongly the variables are related, i.e, if they have a positive or negative relationship. The column ‘is_canceled’ has a positive correlation with ‘lead_time’, ‘previous_cancellation’, and ‘average_daily_rate’. While ‘lead_time’ has a negative correlation with ‘is_repeated_guest’ and ‘‘previous_booking_not_cancelled’. The result matches up with the correlation heat map we plotted during the data visualisation section.

Data Processing

Removing Outliers:

```
company                          112593
lead_time                          3005
country                             488
hotel                                 0
reserved_room_type                    0
assigned_room_type                    0
booking_changes                       0
deposit_type                          0
agent                                 0
days_in_waiting_list                  0
previous_cancellations                0
customer_type                         0
adr                                   0
required_car_parking_spaces           0
total_of_special_requests             0
reservation_status                    0
previous_bookings_not_canceled        0
is_repeated_guest                     0
is_canceled                           0
distribution_channel                  0
market_segment                        0
meal                                  0
babies                                0
children                              0
adults                                0
stays_in_week_nights                  0
stays_in_weekend_nights               0
arrival_date_day_of_month             0
arrival_date_week_number              0
arrival_date_month                    0
arrival_date_year                     0
reservation_status_date               0
dtype: int64
```

After performing the function to remove the outliers, we see that there are 3005 observations that are considered to be outliers. We saw the number of observations was too high to remove from the dataset, as such we decided to make the decision to not remove the outliers.

Calculating the statistics of the variables:

	is_canceled	lead_time	stays_in_weekend_nights	stays_in_week_nights	adults	children	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	average_daily_rate
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000
mean	0.370416	104.011416	0.927599	2.500302	1.856403	0.103886	0.031912	0.087118	0.137097	101.831122
std	0.482918	106.863097	0.998613	1.908286	0.579261	0.398555	0.175767	0.844336	1.497437	50.535790
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-6.380000
25%	0.000000	18.000000	0.000000	1.000000	2.000000	0.000000	0.000000	0.000000	0.000000	69.290000
50%	0.000000	69.000000	1.000000	2.000000	2.000000	0.000000	0.000000	0.000000	0.000000	94.575000
75%	1.000000	160.000000	2.000000	3.000000	2.000000	0.000000	0.000000	0.000000	0.000000	126.000000
max	1.000000	737.000000	19.000000	50.000000	55.000000	10.000000	1.000000	26.000000	72.000000	5400.000000

	country	arrival_date_month	is_canceled	stays_in_weekend_nights	stays_in_week_nights	adults
0	PRT	August	0	0	2	2

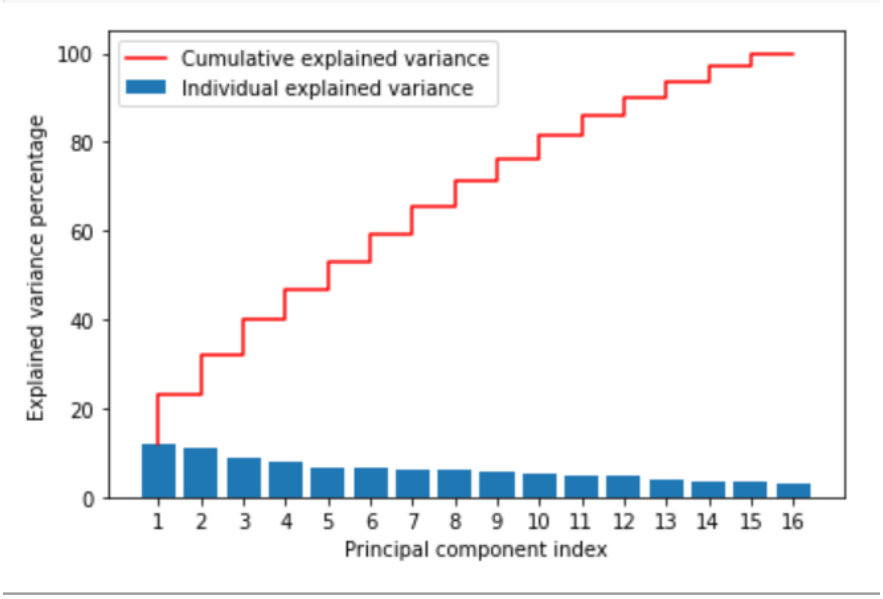
When we performed statistical analysis on our dataset, below are the few findings

- August is the month which had most number of bookings
- Percentage of bookings cancelled is less when compared to the bookings with no cancelation
- Most of the customers prefer to stay less number of nights during the weekends
- Maximum of Average Daily Rate is $54,000 while minimum is $-6.3

- Most of the customers are from the country Portugal

Dimension Reduction

PCA



We have performed PCA on our dataset to determine the variance distribution among the columns. The above graph illustrates the individual variance to the cumulative variance. From the above graph, we can draw that inorder for us to obtain an accuracy of about 90%, we have to take into consideration the maximum number of variables.

Hence, we observed that there are many important columns in our dataset which are of datatype String. Therefore, we converted the values of string variables by assigning them the numeric variables.

After performing this exercise, we implemented models on our dataset.

Exploration of Candidate Data Mining Models

Below are the 4 models, which we implemented to find the best fit for our dataset and problem statement.

1. Logistic Regression

Logistic regression is a simple but effective algorithm that is used for binary classification tasks. We are using it to predict if an event such as the hotel booking being cancelled is occurring or not.

Below are the results which we got after implementing this model on our dataset.

- **Test Accuracy: 75%**

For this model, we have used standard scaling to standardise the data and have plotted the ROC curve graph. We found the Area under the curve to be 0.71.

2. K-NN Classifier:

The K-NN algorithm works in such a way that similar data points in the trained data are grouped to predict the test data. The grouping of the similar data points are done based on the Euclidean/ Manhattan distances.

We have implemented this model on our dataset and below are the results

- **Test accuracy: 82.2%**

In this process, in order to determine the value of K, we have plotted the Accuracy Vs K graph and came to the conclusion that K=11 would be the optimal value.

3. Decision Tree:

The decision tree classifier algorithm is used to make the model by building the decision tree. The decision tree consists of two nodes, the leaf node and the decision node. The decision nodes contain multiple branches and ultimately help make a decision, while the leaf nodes do not have any more branches as it contains the output of the decision.

Below are the results which we got after implementing this model on our dataset.

- **Test Accuracy: 82.8%**

For this model, we have plotted the decision tree, but the graph is too large and needs to be scaled to fit.

4. Random Forest:

Random forest is the type of algorithm which uses multiple decision trees to predict the class. The output of the model would be the class, which most of the decision trees select. However, the individual decision trees in a random forest model must be better than a random classifier.

Below are the results which we got after implementing this model on our dataset.

- **Test Accuracy: 87%**

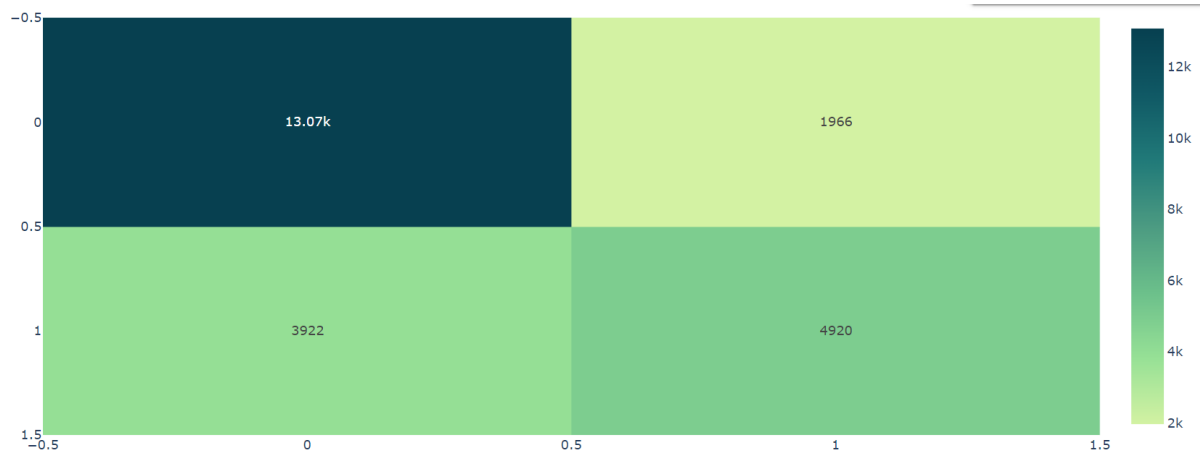
We have used **n_estimators** as **50,75,100** and **max_depth** as **1,2,5** and found out that **max_depth: 5**, **n_estimators: 75** are the best fit parameters.

From the above four models, the Random forest model has the highest accuracy.

Performance Evaluation

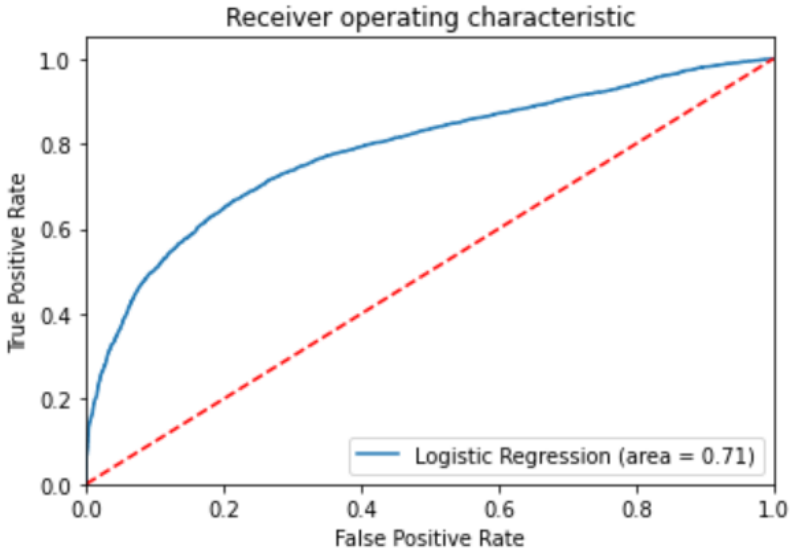
1. Logistic Classifier:

Confusion Matrix:



From the confusion matrix, we find the values of Sensitivity, Specificity, Accuracy and F1-Score. Sensitivity value of 0.86 shows that our model was able to predict 86% of the True Positive values, whereas the specificity value of 0.41 indicates that our model has correctly predicted 41% of True Negative values.

ROC Curve:

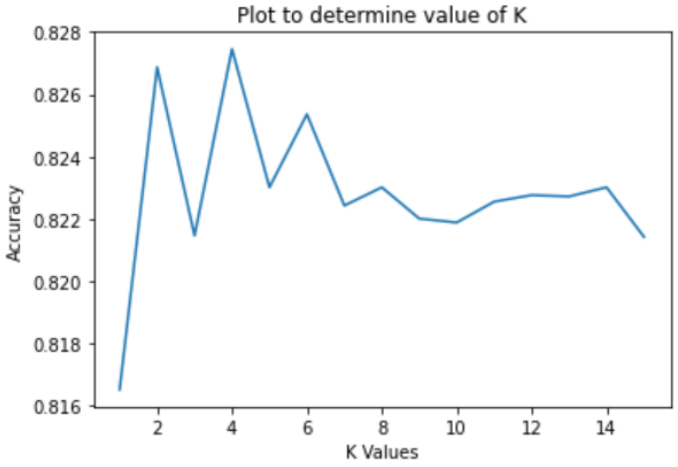


Evaluation Metrics for Logistic Regression :

Precision	Recall	Accuracy	Sensitivity	Specificity	f1-score
0.77	0.87	0.75	0.86	0.55	0.82

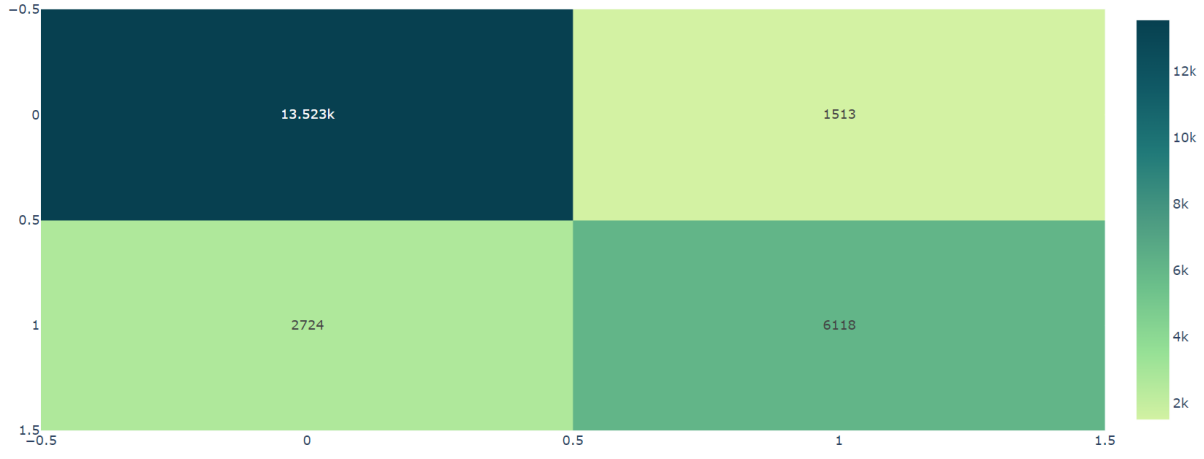
2. K-NN Classifier:

Inorder to determine the value of K, we have plotted the Accuracy Vs K graph, and came to a conclusion that K=11, would be a desired value for our model.



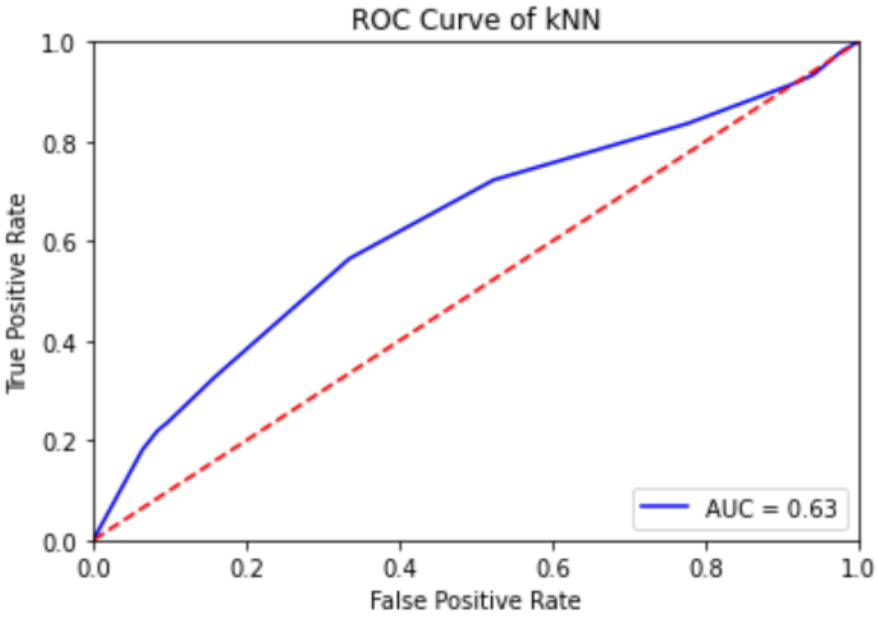
Below confusion matrix, is used to derive the values of Sensitivity, Specificity, Accuracy and F1-Score. Sensitivity value of 91% says that our model has correctly predicted True Positive values and Specificity of 62.64% says that our model has correctly predicted True Negative values.

Confusion Matrix:



ROC Curve:

From the ROC plot below, we can see that the area under the curve is equal to 0.61 and thus this indicates that the model is a bad classifier.

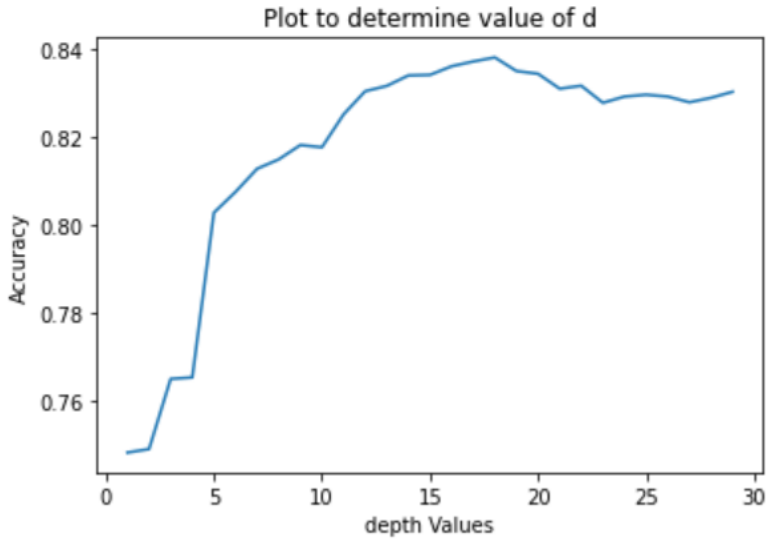


Evaluation Metrics for KNN:

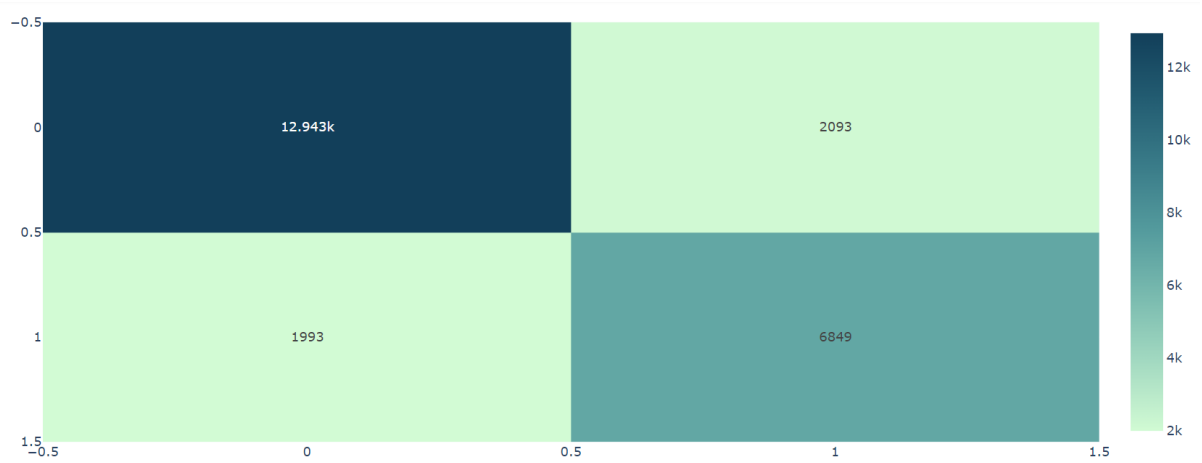
Precision	Recall	Accuracy	Sensitivity	Specificity	f1-score
0.83	0.90	0.82	0.89	0.69	0.86

3. Decision Tree:

Inorder to determine the value of max depth, we have plotted the Accuracy Vs depth graph, and came to a conclusion that max depth = 22, would be a desired value for our model.



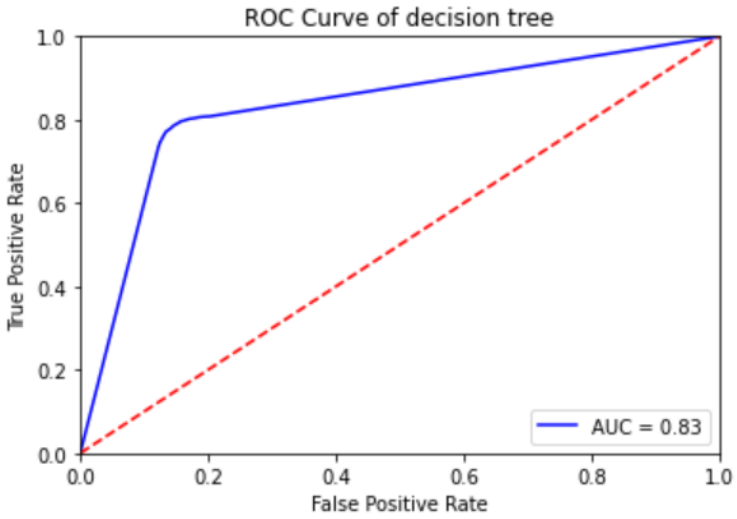
Confusion Matrix:



Sensitivity value of 0.86 shows that our model was able to predict 86% of the Cancelled bookings correctly, whereas the specificity value of 0.74 indicates that our model has correctly predicted 74% of bookings that were not cancelled.

ROC Curve:

From the ROC plot below, we can see that the area under the curve is equal to 0.81 and thus this indicates that the model is a good classifier.



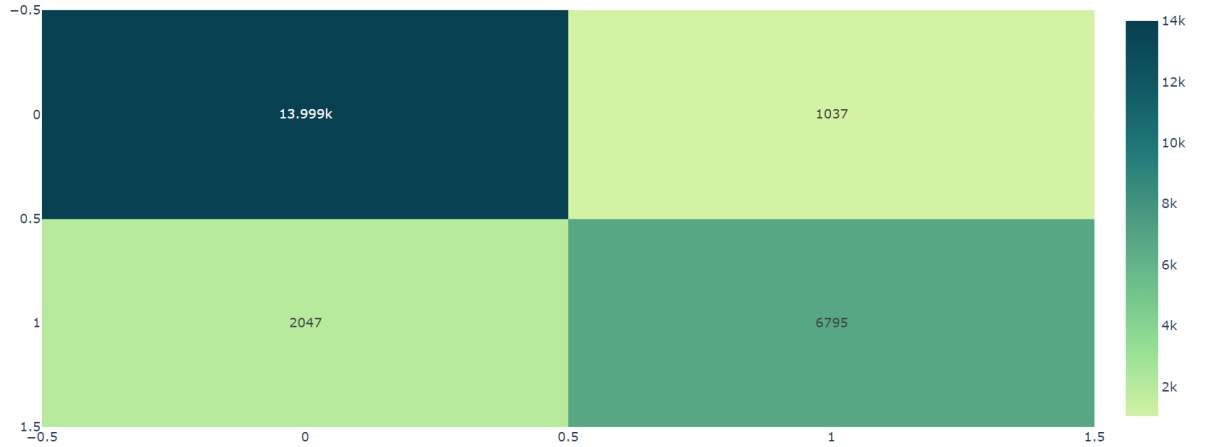
Evaluation Metrics for Decision Tree:

Precision	Recall	Accuracy	Sensitivity	Specificity	f1-score
0.87	0.86	0.83	0.86	0.77	0.86

4. Random Forest:

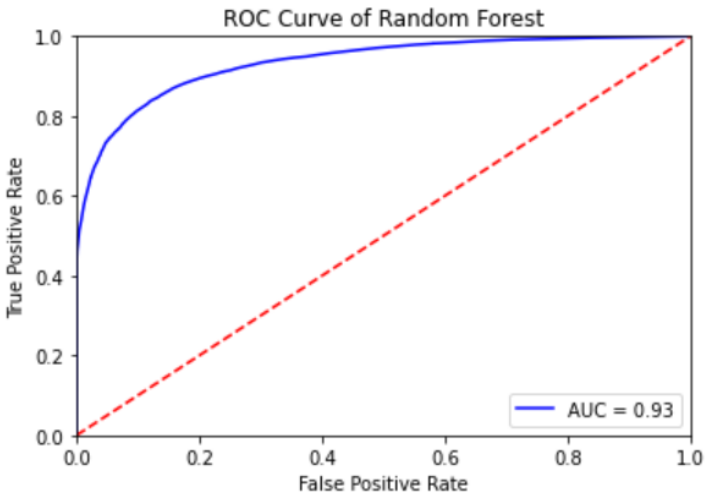
As mentioned above, we have used **n_estimators** as **50,75,100** and **max_depth** as **1,2,5** and found out that **max_depth: 5**, **n_estimators: 75** are the best fit parameters.

Confusion Matrix:



ROC Curve:

From the ROC plot below,we can see that the curve is close to the top left, which indicates that the model is a good classifier.



Sensitivity value of 0.93 shows that our model was able to predict 93% of the Cancelled bookings correctly, whereas the specificity value of 0.76 indicates that our model has correctly predicted 76% of the True Negative values. Below is the Evaluation Metrics for random forest.

Evaluation Metrics for Random Forest:

Precision	Recall	Accuracy	Sensitivity	Specificity	f1-score
0.87	0.93	0.87	0.93	0.76	0.90

Project Results:

- Among the four implemented models, Random Forest model gives the highest accuracy (87%) when applied on our dataset. It also has the highest F1 score with 90%. This model also gives highest sensitivity (93%), which shows that it is very good at classifying the true positive values, which in our case is "Is Cancelled".
- Decision Tree is the next best model with an accuracy of 83% .This model gives both F-1 score and Sensitivity of 86%.
- KNN is the next best model with an accuracy of 82%. However, this model gives good sensitivity i.e 89%, which shows that it is very good at classifying the true positive values, which in our case is "Is Cancelled".
- Logistic Regression is found to be the least performing model, with an accuracy of only 75%.

Below is the table which summarises various evaluation metrics for different models:

Model	Precision	Recall	Accuracy	Sensitivity	Specificity	F-1 Score
Logistic Regression	0.77	0.87	0.75	0.86	0.55	0.82
KNN	0.83	0.90	0.82	0.89	0.69	0.86
Decision Tree	0.87	0.86	0.83	0.86	0.77	0.86
Random Forest	0.87	0.93	0.87	0.93	0.76	0.90

Impact of Project Outcomes:

The goal of this project was to see if cancellation of a booking depends on external factors like the lead time, booking, booking changes made, the deposit type, and the customer type. We find out through the course of the project using the correlation heat map and other visualisations, that none of external factors except for lead time have an effect on a booking being cancelled. We also see that if a customer is a repeat customer, there is more of a probability that the customer will cancel their booking. From all the classification models, the Random Forest model performs the best with a higher accuracy, sensitivity, specificity and F1-score. The Random Forest model would be the best one to used to check the probability that a customer makes a cancellation