Title: Hotel booking demand

DatasetLink:
https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

Group: Harshit Jain, Juhi Patel, Kshitij Mahajan, Aarib Mohammed

Description of project goals:
The group has selected the hotel booking demand dataset from Kaggle. In this dataset, there are 31 variables that describe the bookings of customers at resort and city hotels in Portugal between the 1st of July of 2015 and the 31st of August 2017. We will be looking at how these predictors play a role in behaviors that cause booking cancellation and other predictors that will allow customers to select better times to travel. We will be using descriptive analytics and machine learning algorithms to understand patterns and trends in this dataset that may lead to booking cancellations. Through the machine learning algorithms we would be aiming to create a model to predict if a guest will actually come. This can help a hotel to plan things like personel and food requirements. Also hotels will have contingency plans in place for guests which the ML model is not very optimistic about. Maybe some hotels can also use such a model to offer more rooms than they have to make more money. This surely has an economic impact as it would make the hotels more profitable and efficient. The group found this interesting as each one of us have been to hotels in our life and would benefit the travelers as well as it would increase the inventory of rooms for them.

Studying the trends in this dataset will allow us to understand patterns, trends, and anomalies in the data. We can use the patterns to understand behaviors behind booking cancellation, customer satisfaction, best seasons to book, and more. We have used different machine learning models to assess the patterns behind booking cancellation. Our team believes that studying cancellation behavior in hotel guests will allow new insights into changes that need to be made in demand-management decisions that lead to overall increase in revenue. All plots and data tables can be referenced in the appendix below.

Exploratory analysis:

Data cleaning and preprocessing

- All the null values were checked and were filled with zeros after analysis
- The column ‘company had’ had mostly null values. Due to all the null values this column was dropped.
- All the rows in adults and children columns having 0 simultaneously were deleted as there can't be a case when the booking was done by a ghost!
- Label encoding was done to convert all the categorical variables into numerical variables.
- `market_segment', 'distribution_channel', 'reserved_room_type', 'assigned_room_type', 'Deposit_type', 'customer_type', 'reservation_status', 'reservation_status_date' were categorical columns that were converted into numerical.
- The data had around 75000 not canceled bookings compared to 40000 canceled bookings. So to balance the dataset, the SMOTE technique of over sampling was applied to balance the dataset and avoid overfitting.
- The test-training split was set to 20 percent.
- Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. VIF was used to remove the multicollinear features and led to the removal of the following columns(All columns with VIF>5 were removed one by one till a point where there were no values above 5). The feature was reduced from 31 to 22 using VIF (Appendix X).
- 'Arrival_date_year','reserved_room_type','distribution_channel','reservation_status_date','reservation_status' were some of the columns that were removed.

Insights from EDA

The team has looked into basic statistics about the dataset to look into patterns. We first looked into the hotel types. This analysis will answer the questions below.

- What is the distribution of resort hotels and city hotels?

The dataset contains 79,330 observations for the city hotel and 40,060 resort hotels (Appendix A).

- How many bookings were canceled?
  There are about 44,000 canceled reservations and 75,000 not canceled bookings in the dataset. A surprisingly huge number of cancellations (is_canceled=0 for not canceled bookings and 1 for canceled bookings).

  See Appendix A & B to see plots on observations collected for Resort hotels and City hotels and their respective cancellations.

- Which month observes the highest customer arrival rate?
  We can see the highest customer arrival in the month of August. The month of May and July come close second.This is understandable as these coincides with the summer vacations. The least customer arrival happens in the winter season presumably because it is extremely cold.The months are numbered according to the order which they come over the year with 1 as January and 12 as December (Appendix C).

- Where do the guests come from?
  We also looked at where the guests came from geographically. We see that people travel from all over the world to stay at these hotels. However, most guests are from Portugal and other European countries (Appendix D).

- Cancellations by repeated guests:
  When looking at repeated guests, we found that the repeated guests cancellation is nearly negligible. All the cancellations are done by first time visitors. Also the proportion of repeated guests is significant (Appendix E).

- How does the price vary per night over the year?
  Conducting further analysis, we found the winter season had lower prices per night whereas the months of July and August are extremely expensive for both types of hotels. This falls in direct resonance with the customer arrival rate we explored above. Also the resort hotels are

cheaper than city hotels for most part of the year but in July and August resort prices skyrocket.The specific prices can be referenced in Appendix F and G.

- How many customers visit each hotel type all over the month?
  Looking at the graph, it can be seen that the city hotels have more customers in all months. Considering this, resort hotels seem to be a little closer to city hotels in summer. (Appendix H)

- What are the number of cancellations over the year and how does it look when bifurcated by hotel types?
  An important interpretation can be made by examining three graphics (Appendix H, I and J) together. Fewer customers come in the winter months, so when we look at the cancellation rates, it is quite normal that it appears less in the winter months. The point to be noted on these months is that the cancellation rates of city hotels are almost equal to other months even in winter. The fact that the total cancellation rates of the winter months are low is that the cancellation rates of the resort hotels are low in these months. In short, the possibility of cancellation of resort hotels in winter is very low. This information can be a very important factor when predicting 'is_canceled'.

- How does the booking look by market segmentation?
  We also look at bookings by market segmentation. We see that online travel agents make up about 47.3% of the bookings. The second highest being offline travel agents at 20.3%. Airlines pay approximately twice as much. The high prices paid by airlines might be due to lead time for bookings from Aviation being very short: Mean 4 days vs. 104 days for other bookings. Or airlines NEED a place for their crews to stay. Lastly, airline personnel usually get one room per person - more total rooms required compared to families. Reference Appendix K, O and P.

- How does the booking look by market segmentation?
  The people tend to have to stay longer in resorts.For the city hotel there is a clear preference for 1-4 nights.For the resort hotel, 1-4

nights are also often booked, but 7 nights also stand out as being very popular (Appendix M)

- Tell me something about the footfall of customers.
  Looking at the graph in Appendix H, it can be seen that the city hotels have more customers in all months when compared to resort hotels. Considering the proportion, resort hotels seem to catch up to city hotels in summer. This may be due to the abundance of traveling in the summer due to better climate and summer break from school.

<u>Solutions and insights:</u>
Models

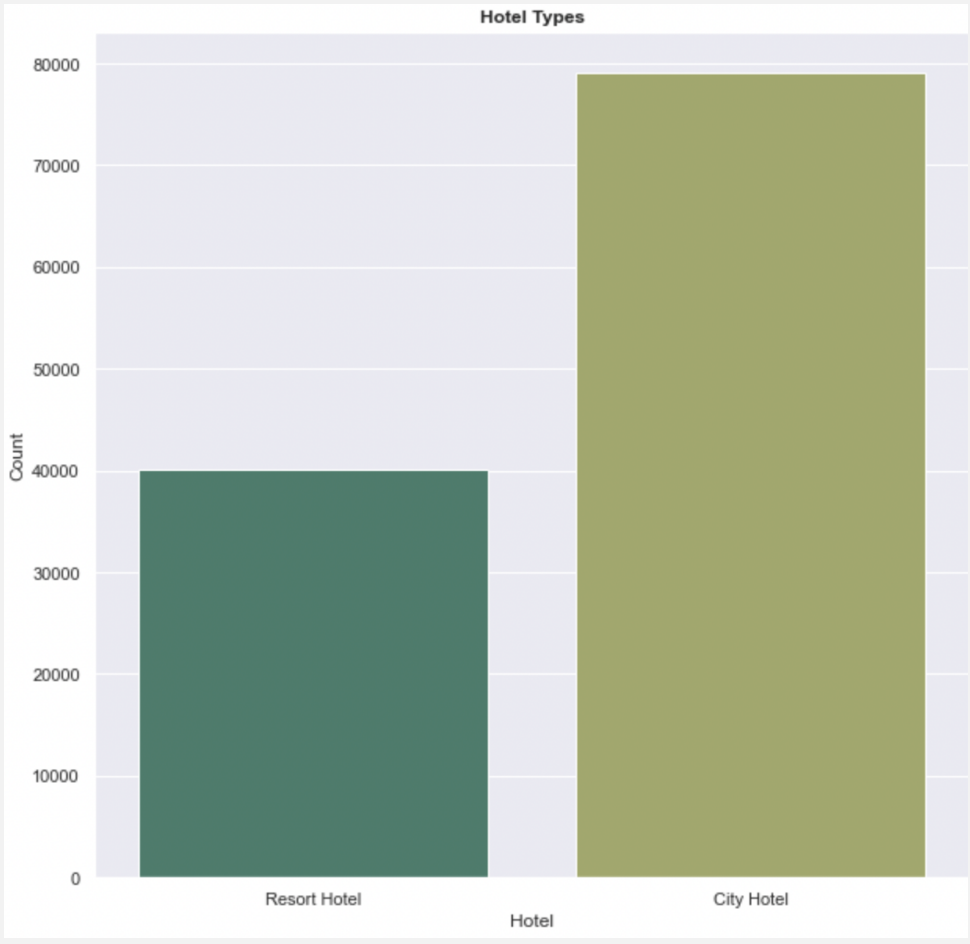
- Logistic Regression (Appendix R)
  - The model was run on the dataset and achieved an accuracy of around 77 percent.
  - The ROC curve was plotted with the area under the curve being 0.84
- K Nearest Neighbors (Appendix S)
  - The model was run on the dataset and achieved an accuracy of around 76 percent.
  - The ROC curve was plotted with the area under the curve being 0.84
- Decision Tree (Appendix T)
  - The model was run on the dataset and achieved an accuracy of around 82 percent.
  - The ROC curve was plotted with the area under the curve being 0.89
- Random Forest (Appendix U and V)
  - The model was run on the dataset and achieved an accuracy of around 88.3 percent.
  - We were able to see a performance improvement from 86.5 to 88.3 percent this difference may seem marginal but as the dataset tends to grow and would have variety of values that is when the performance difference would become quite apparent
  - The feature importance graph was plotted and had predictable features like average daily rate, deposit type, number of special requests, agent through which the booking took place. But what

was the most surprising that the lead time was the most important feature.
    - The ROC curve was plotted with the area under the curve being 0.94
    - The feature importance and performance improvement comparison was done for all models but only random forest observations are shared due to space and time constraints.
- Naive Bayes (Appendix W)
    - The model was run on the dataset and achieved an accuracy of around 53 percent.
    - The ROC curve was plotted with the area under the curve being 0.78.

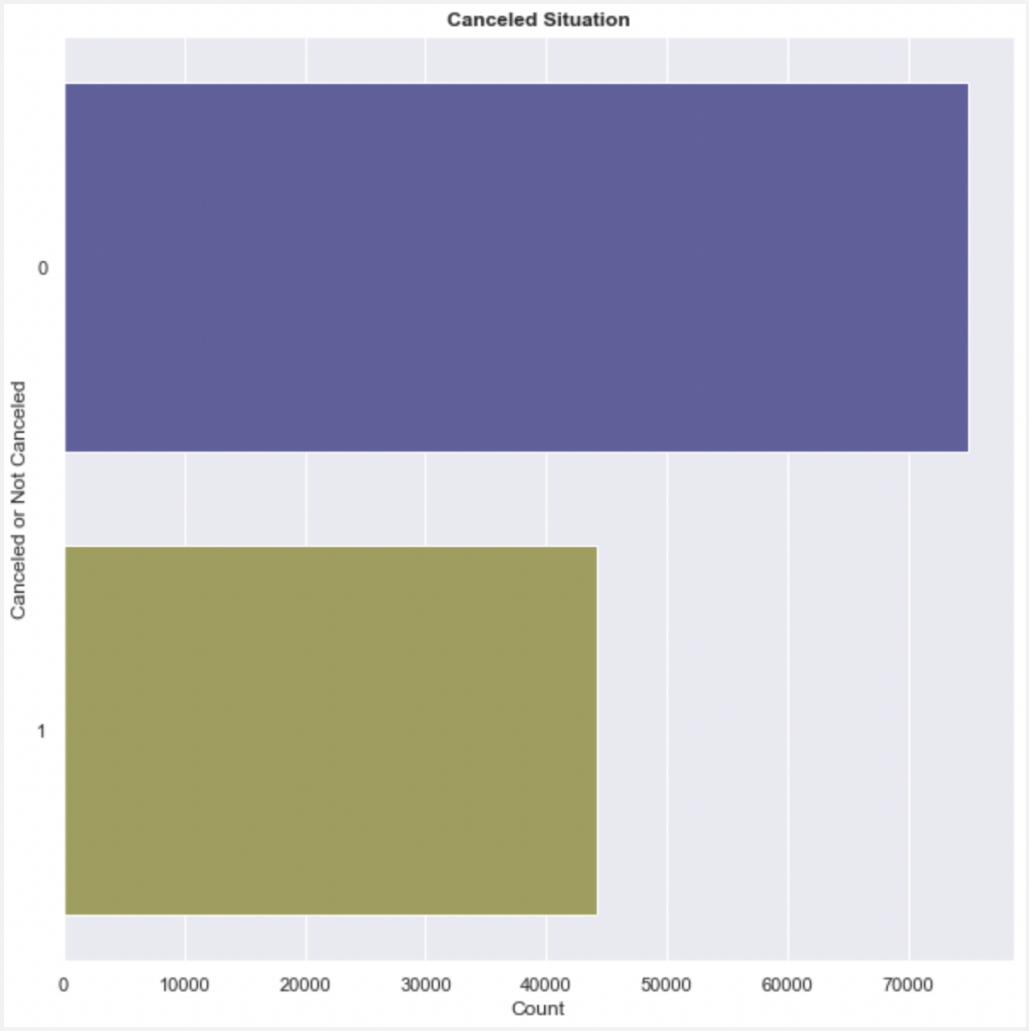
- Conclusion: The accuracy was the highest with the Random Forest Classifier. The score was 88.3%. Therefore this model was chosen.

Appendix

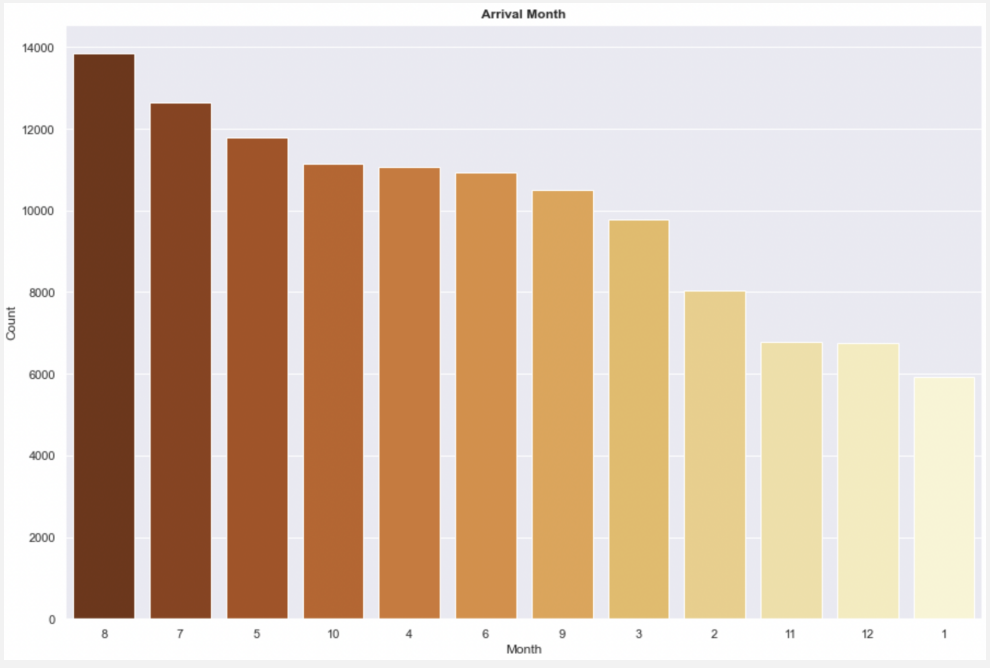
Appendix A


Hotel Types
80000
70000
60000
50000
40000
30000
20000
10000
0
Count
Resort Hotel
City Hotel
Hotel

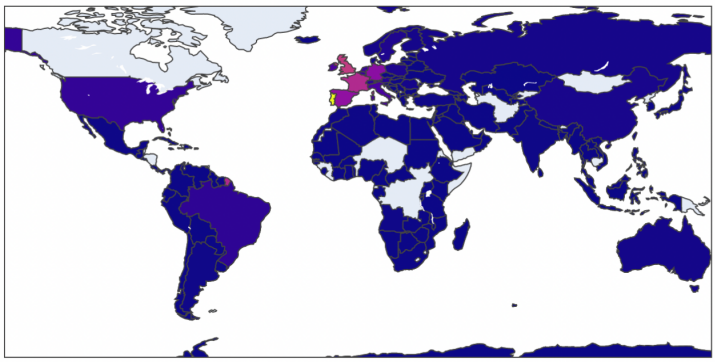
Appendix B

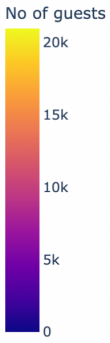

Canceled Situation
Canceled or Not Canceled
0
1
0
10000
20000
30000
40000
50000
60000
70000
Count

Appendix C

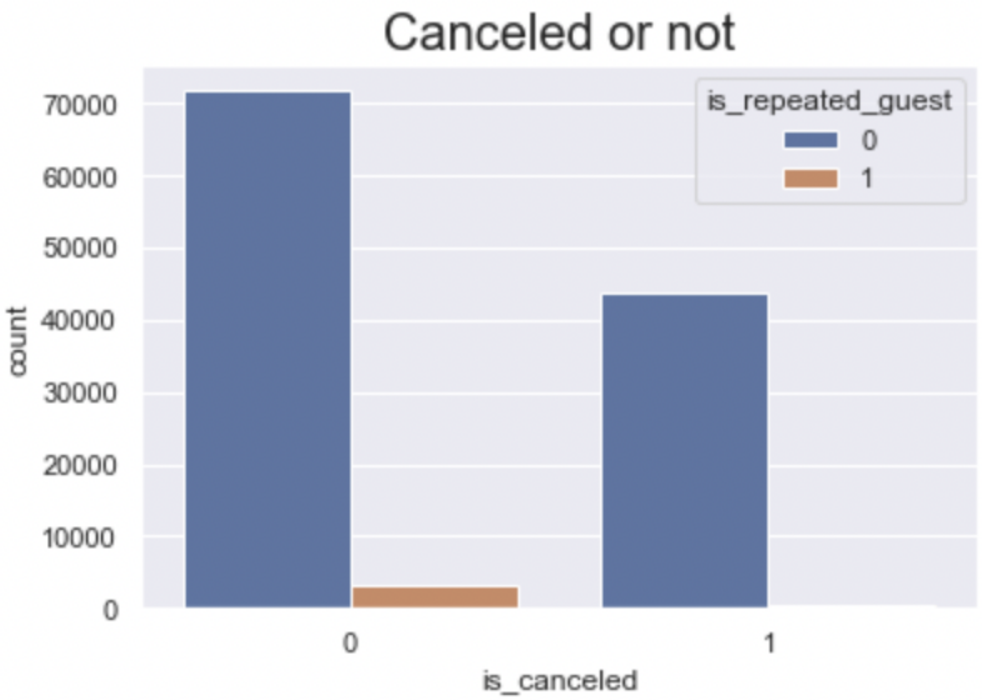


Appendix D

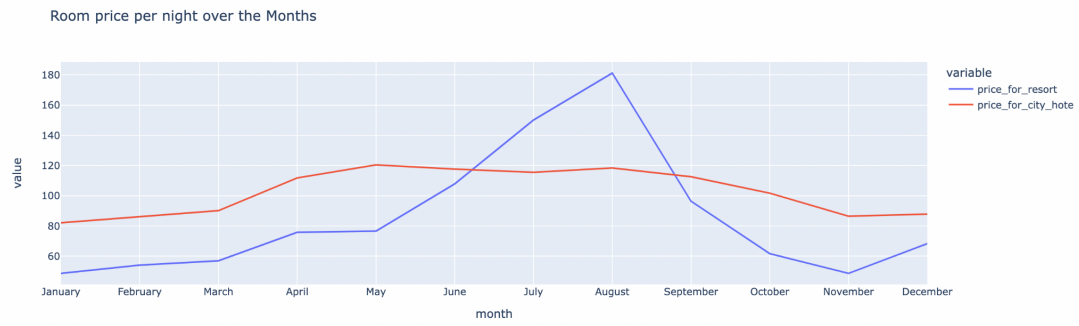




Appendix E



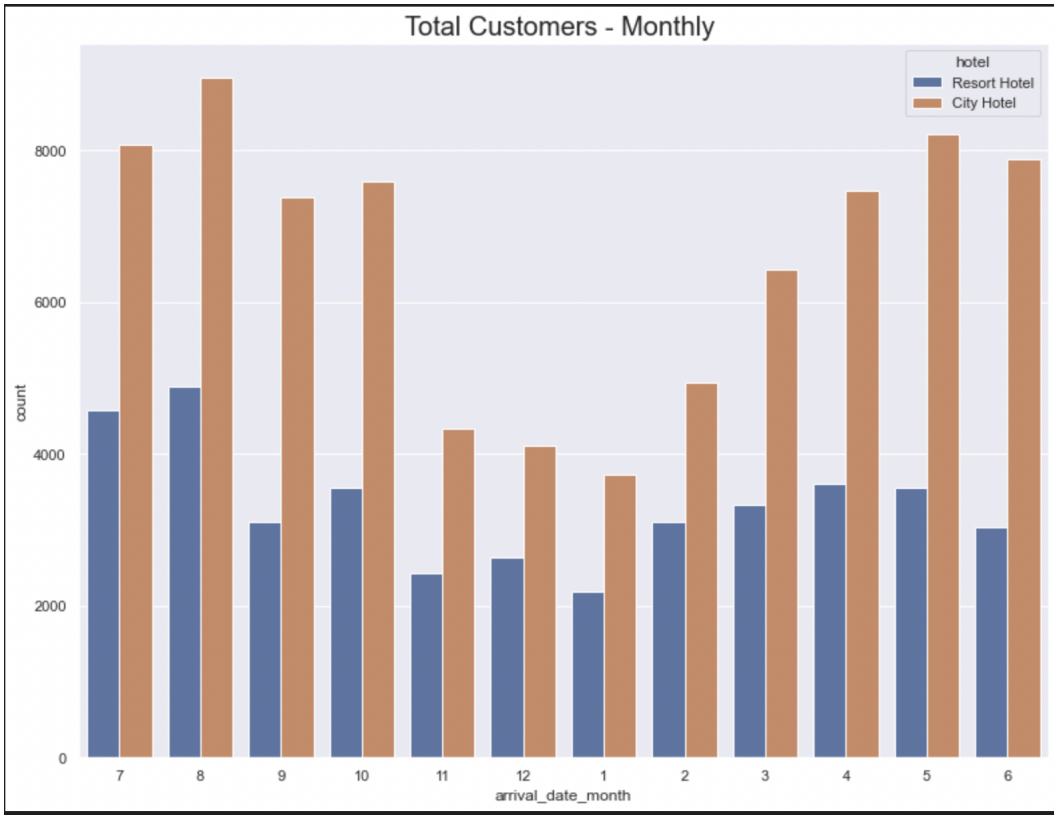
Appendix F



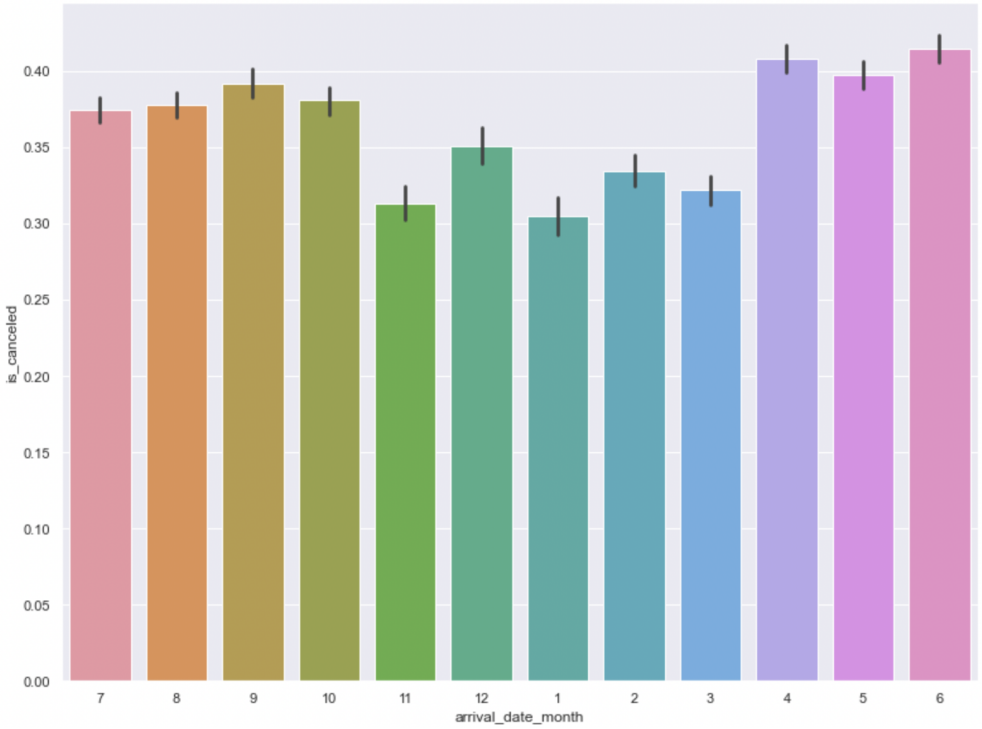
Appendix G

	month	price_for_resort	price_for_city_hotel
0	April	75.867816	111.856824
1	August	181.205892	118.412083
2	December	68.322236	87.856764
3	February	54.147478	86.183025
4	January	48.708919	82.160634
5	July	150.122528	115.563810
6	June	107.921869	117.702075
7	March	57.012487	90.170722
8	May	76.657558	120.445842
9	November	48.681640	86.500456
10	October	61.727505	101.745956
11	September	96.416860	112.598452

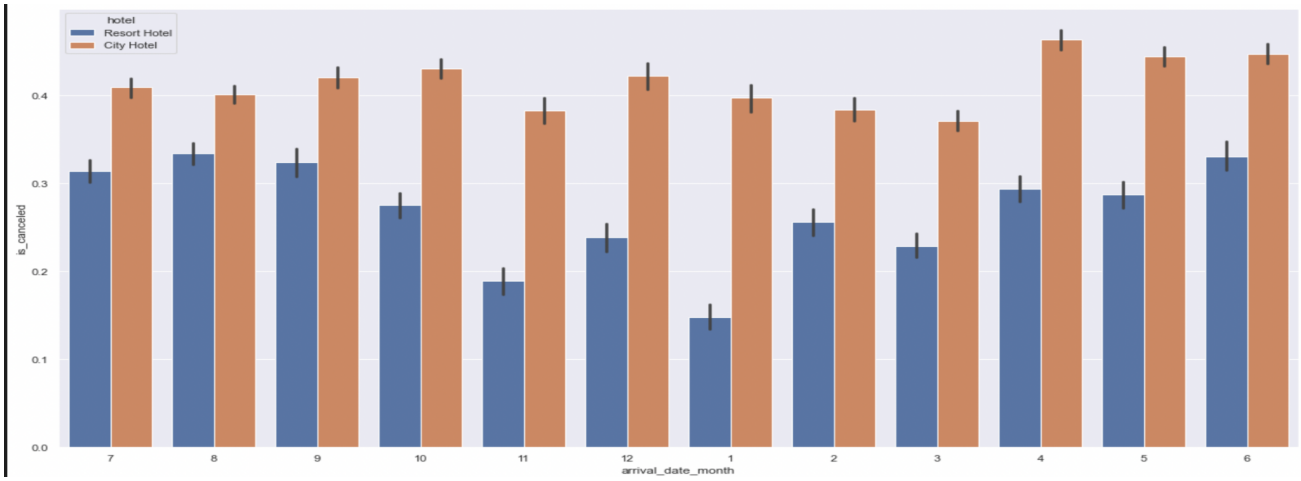
Appendix H



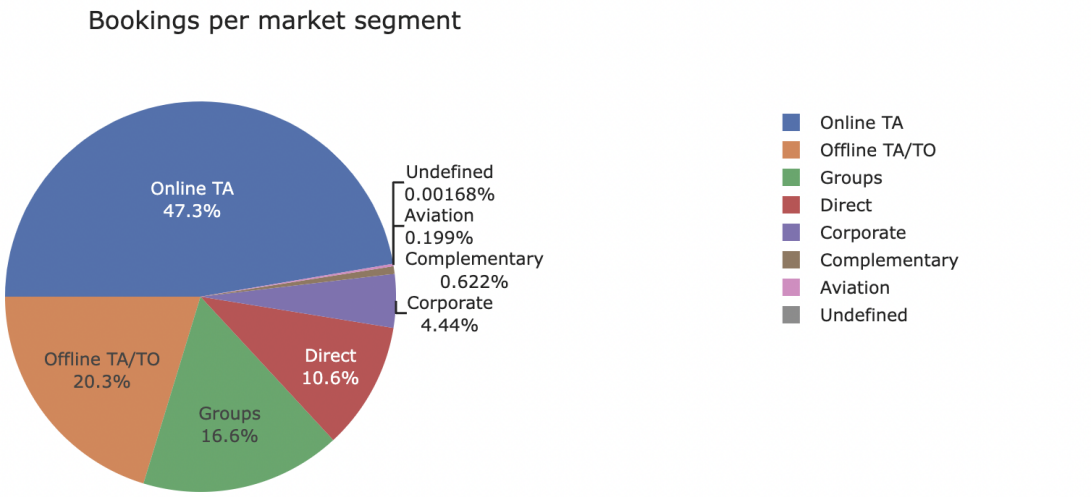
Appendix I



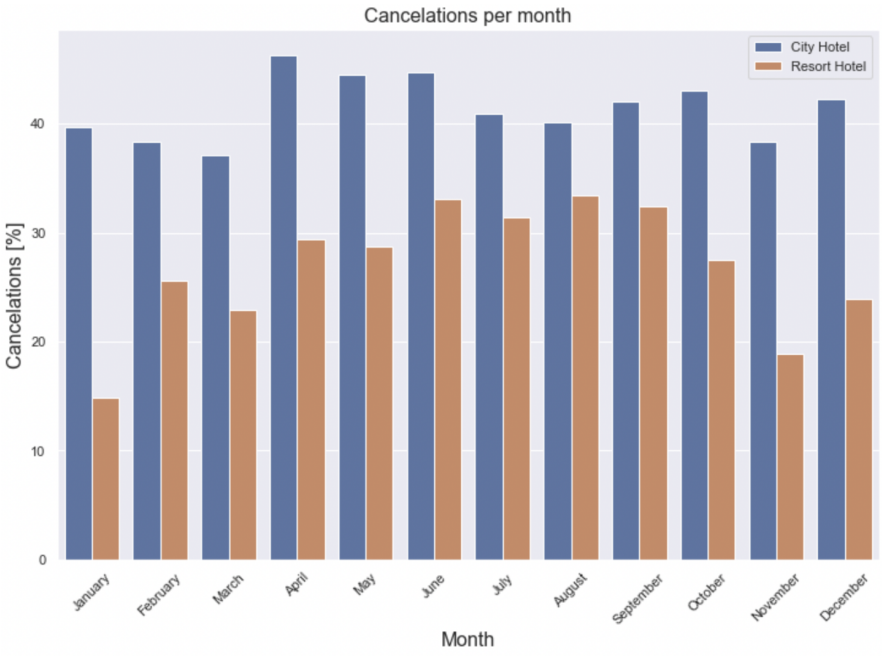
Appendix J



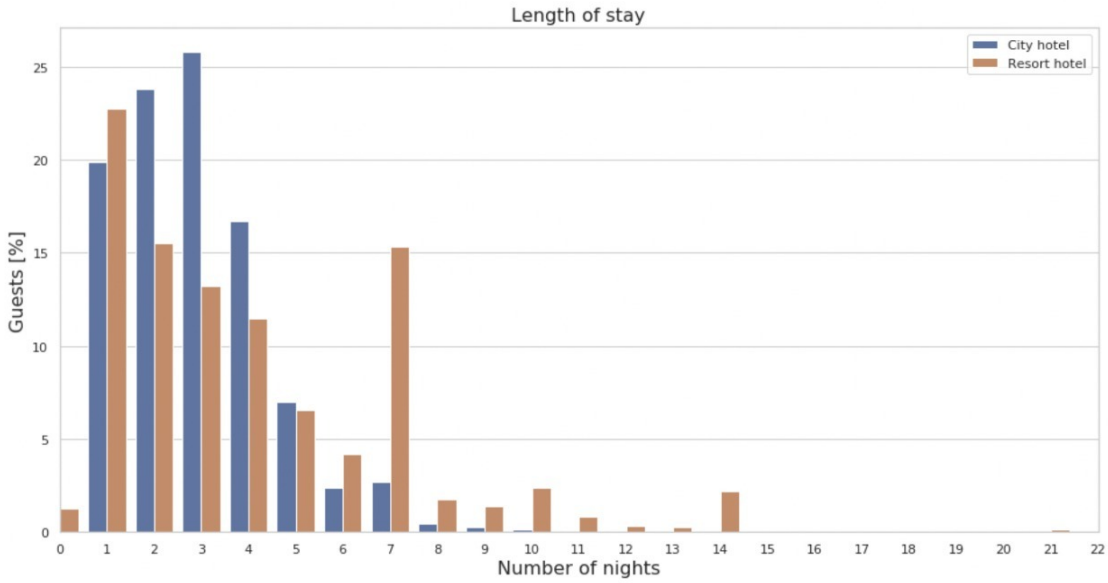
Appendix K


Bookings per market segment
Online TA
47.3%
Offline TA/TO
20.3%
Groups
16.6%
Direct
10.6%
Corporate
4.44%
Complementary
0.622%
Aviation
0.199%
Undefined
0.00168%
Online TA
Offline TA/TO
Groups
Direct
Corporate
Complementary
Aviation
Undefined

Appendix L


Cancelations per month
City Hotel
Resort Hotel
Cancelations [%]
0
10
20
30
40
January
February
March
April
May
June
July
August
September
October
November
December
Month

Appendix M


Length of stay
City hotel
Resort hotel
Guests [%]
0
5
10
15
20
25
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Number of nights

Appendix N

```
Total bookings canceled: 44,199 (37 %)
Resort hotel bookings canceled: 11,120 (28 %)
City hotel bookings canceled: 33,079 (42 %)
```

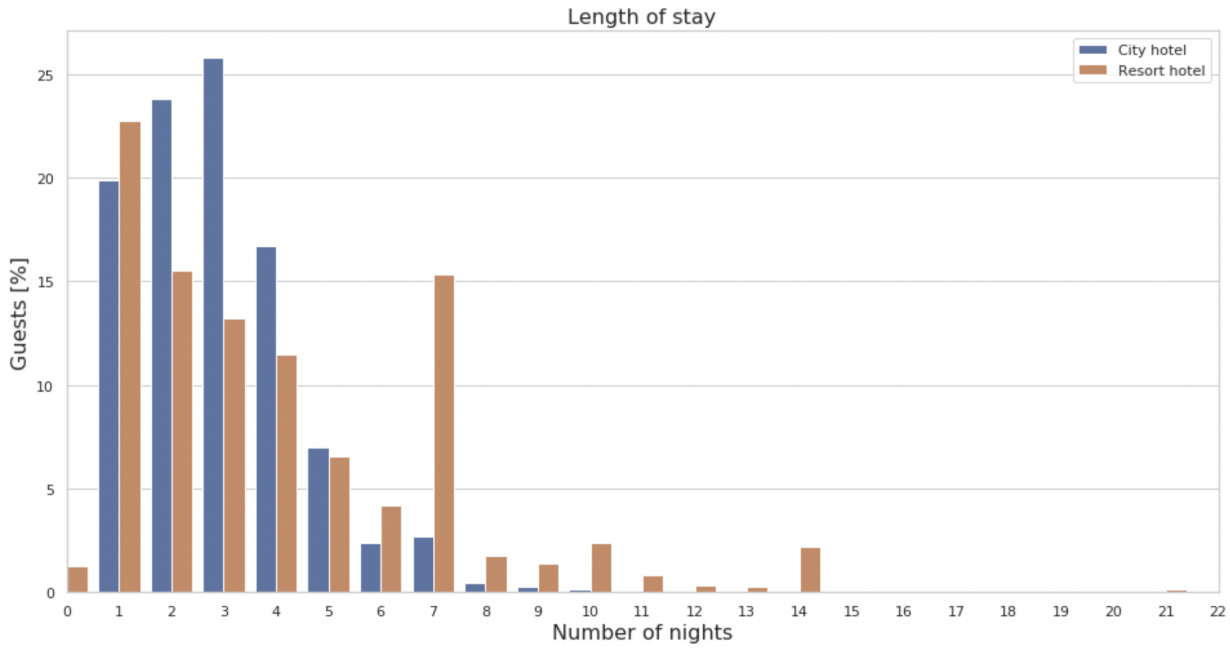
Appendix O

	is_canceled	adults	lead_time	adr_pp
count	235.000000	235.000000	235.00000	235.000000
mean	0.221277	1.012766	4.47234	100.270979
std	0.415992	0.112503	4.61629	20.376689
min	0.000000	1.000000	0.00000	0.000000
25%	0.000000	1.000000	1.00000	95.000000
50%	0.000000	1.000000	3.00000	95.000000
75%	0.000000	1.000000	6.00000	110.000000
max	1.000000	2.000000	23.00000	193.500000

Appendix P

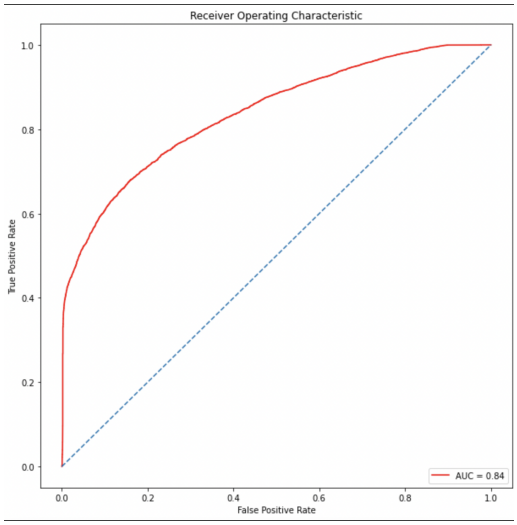
	is_canceled	adults	lead_time	adr_pp
count	118975.000000	118975.000000	118975.000000	118971.000000
mean	0.371061	1.860878	104.306031	55.038212
std	0.483091	0.574499	106.888885	29.016998
min	0.000000	0.000000	0.000000	-3.190000
25%	0.000000	2.000000	18.000000	37.440000
50%	0.000000	2.000000	69.000000	49.500000
75%	1.000000	2.000000	161.000000	66.000000
max	1.000000	55.000000	737.000000	2700.000000

Appendix Q



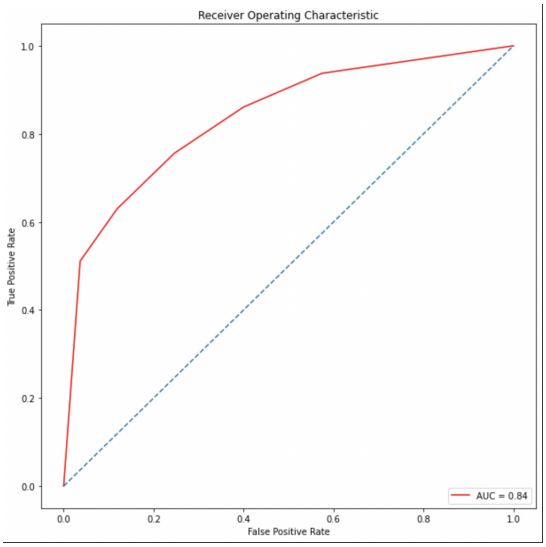
Appendix R Logistic Regression

```
Accuracy Score: 0.7686855129603222

Confusion Matrix:
 [[12028  2891]
 [ 2624  6299]]
Model: Logistic Regression

Classification Report is:
              precision    recall  f1-score   support

           0       0.82      0.81      0.81     14919
           1       0.69      0.71      0.70      8923

    accuracy                           0.77     23842
   macro avg       0.75      0.76      0.75     23842
weighted avg       0.77      0.77      0.77     23842
```



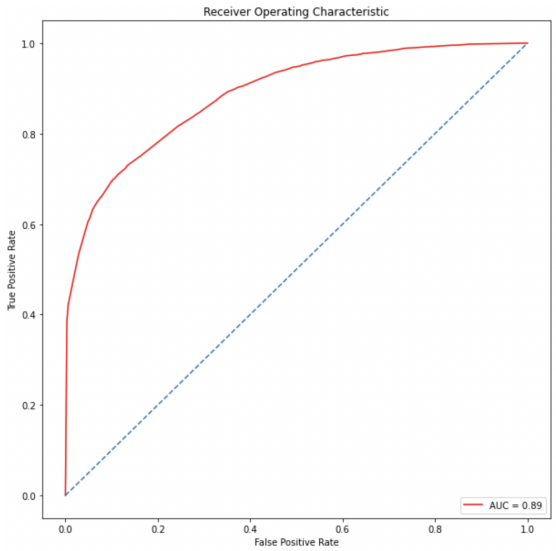
Appendix S KNN

```
Accuracy Score: 0.7557251908396947

Confusion Matrix:
 [[11287  3632]
 [ 2192  6731]]
Model: KNeighboursClassifier

Classification Report is:
              precision    recall  f1-score   support

           0       0.84      0.76      0.79     14919
           1       0.65      0.75      0.70      8923

    accuracy                           0.76     23842
   macro avg       0.74      0.76      0.75     23842
weighted avg       0.77      0.76      0.76     23842
```



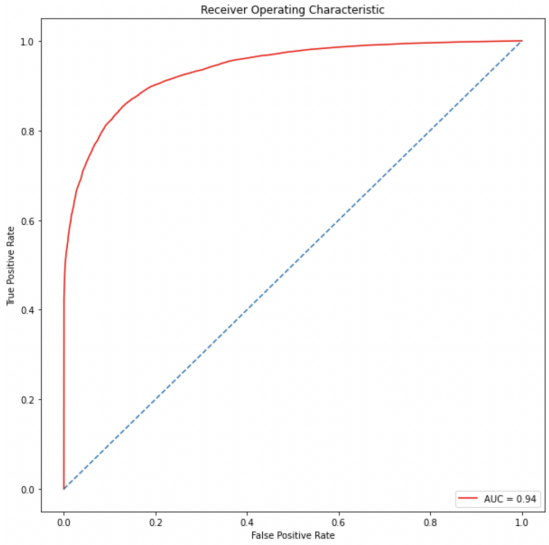
Appendix T Decision Tree

```
Accuracy Score: 0.8172552638201493

Confusion Matrix:
 [[13091  1828]
 [ 2529  6394]]
Classification Report is:
              precision    recall  f1-score   support

           0       0.84      0.88      0.86     14919
           1       0.78      0.72      0.75      8923

    accuracy                           0.82     23842
   macro avg       0.81      0.80      0.80     23842
weighted avg       0.82      0.82      0.82     23842
```



Appendix U Random Forest

```
Accuracy Score: 0.8835129175853892

Confusion Matrix:
 [[10952  1125]
 [ 1675 10285]]
Model: Random Forest

Classification Report is:
              precision    recall  f1-score   support

           0       0.87      0.91      0.89     12077
           1       0.90      0.86      0.88     11960

    accuracy                           0.88     24037
   macro avg       0.88      0.88      0.88     24037
weighted avg       0.88      0.88      0.88     24037
```

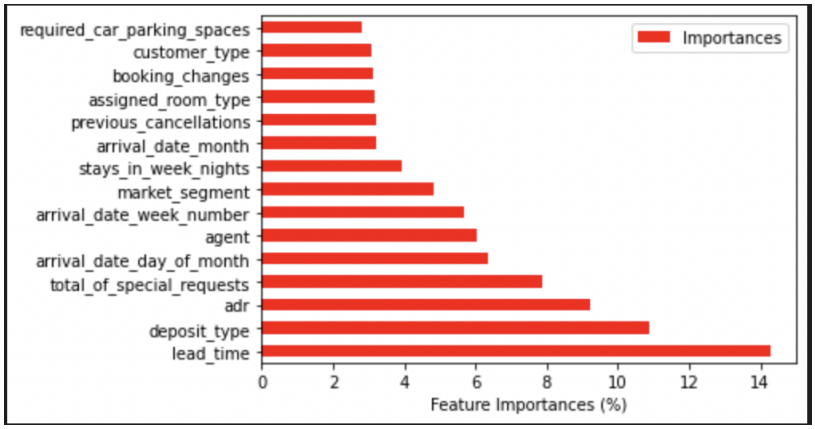


The above random forest model was ran on a balanced dataset using the SMOTE technique.

```
Accuracy Score: 0.868970723932556

Confusion Matrix:
 [[13557  1362]
 [ 1762  7161]]
Model: Random Forest

Classification Report is:
              precision    recall  f1-score   support

           0       0.88      0.91      0.90     14919
           1       0.84      0.80      0.82      8923

    accuracy                           0.87     23842
   macro avg       0.86      0.86      0.86     23842
weighted avg       0.87      0.87      0.87     23842
```

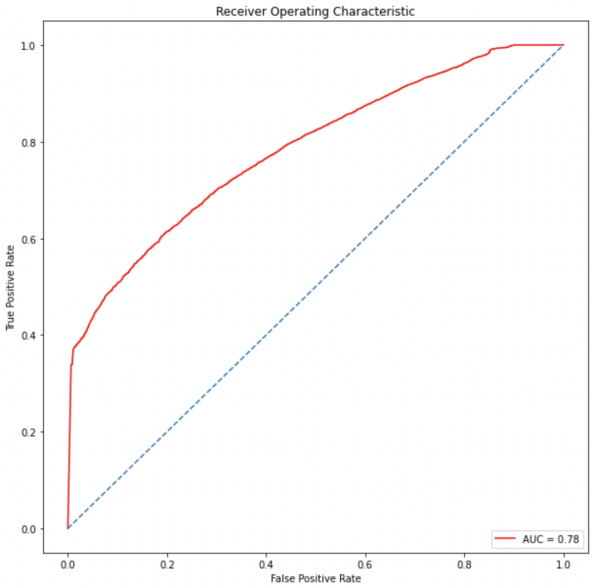
The above random forest model was run on an unbalanced dataset.

Appendix V



Appendix W Naive Bayes

```
Accuracy Score: 0.5266336716718396

Confusion Matrix:
 [[ 4259 10660]
 [  626  8297]]
Model: Gaussian Naive Bayes

Classification Report is:
              precision    recall  f1-score   support

           0       0.87      0.29      0.43     14919
           1       0.44      0.93      0.60      8923

    accuracy                           0.53     23842
   macro avg       0.65      0.61      0.51     23842
weighted avg       0.71      0.53      0.49     23842
```



Appendix X

	Variable	VIF
0	hotel	2.288345
18	agent	1.984457
17	deposit_type	1.725487
15	assigned_room_type	1.516963
21	adr	1.479576
24	reservation_status	1.470392
11	market_segment	1.446603
6	stays_in_week_nights	1.424952
1	lead_time	1.409652
5	stays_in_weekend_nights	1.347643
25	reservation_status_date	1.305080
12	is_repeated_guest	1.284147
23	total_of_special_requests	1.279754
8	children	1.252064
14	previous_bookings_not_canceled	1.236054
3	arrival_date_week_number	1.213870
2	arrival_date_month	1.151122
20	customer_type	1.139957
7	adults	1.134173