Hotel Booking Cancellation Prediction

INN Hotels: Supervised Learning Classification

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

- The hotel can take the following actions for the customers who have booked more than 151 days in advance:
  - Set up a system that can send a prompt (like an automated email or app notification) to the customers 90 days before the arrival date asking for a re-confirmation of their booking and any changes they would like to make in their bookings.
  - Remind guests about imminent deadlines 1 month prior to the date of arrival.

  The response given by the customer will give the hotel ample time to re-sell the room or make preparations for the customers' requests.

- Stricter cancellation policies can be adopted by the hotel.
  - The bookings where the average price per room is high, and there were special requests associated should not get a full refund as the loss of resources will be high in these cases.
  - Ideally, the cancellation policies should be consistent across all market segments but as noticed in our analysis high percentage of bookings done online are cancelled. The booking cancelled online should yield less percentage of refund to the customers.

  The refunds, cancellation fees, etc should be highlighted on the website/app before a customer confirms their booking to safeguard guests' interest.

Executive Summary

- We saw in our analysis that bookings, where the total length of stay was more than 5 days, had a higher chance of getting canceled. The hotel can restrict booking duration up to 5 days only with an option for customers to book an extension in case they are willing to extend their stay. Such restrictions can be strategized by the hotel to generate additional revenue and can be relaxed for certain market segments (like Corporate and Aviation) such that it does not hamper their experience with the hotel.

- In the months of December and January cancellation to non-cancellation ratio is low. Customers might travel to celebrate Christmas and New Year. The hotel should ensure that enough human resources are available to cater to the needs of the guests.

- October and September saw the highest number of bookings but also a high number of cancellations. This should be investigated further by the hotel.

- Improving the experience of repeated customers.
  - A loyal guest is usually more profitable for the business because they are more familiar with offerings from the hotel they have visited before.
  - Attracting new customers is tedious and costs more as compared to a repeated guest.
  - A loyalty program that offers - special discounts, access to services in hotels, etc for these customers can help in improving their experience

Business Problem Overview and Solution Approach

- INN Hotels Group has a chain of hotels in Portugal. They are facing problems with the high number of booking cancellations.

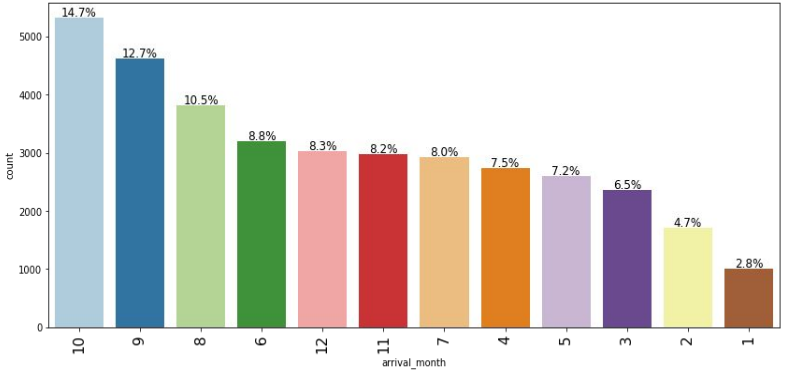
- A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

- The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior.

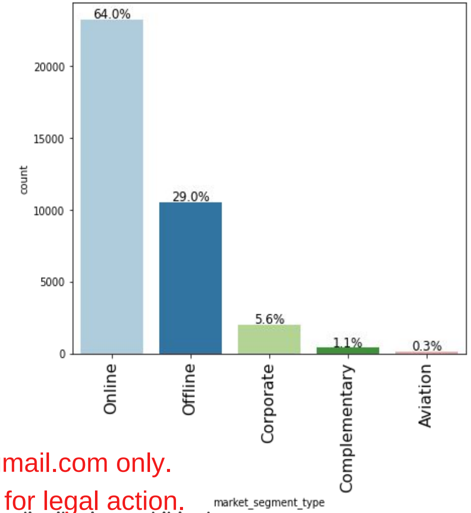
- The task at hand is to analyze the data provided, to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

EDA Results

- October is the busiest month for the hotel followed by September.
- 14.7% of the bookings were made in October.

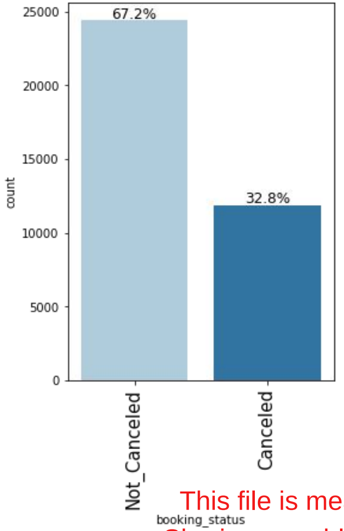


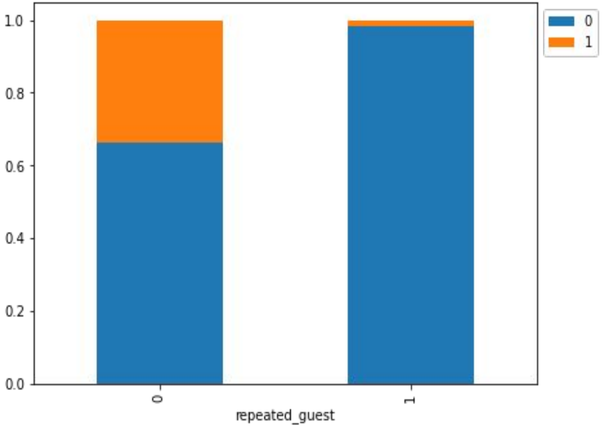
- 64% of the hotel bookings were made online followed by 29% of the bookings which were made offline.



EDA Results

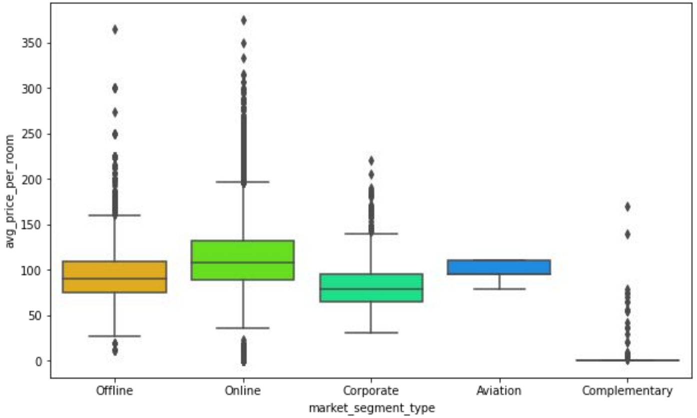
- 32.8% of the bookings were canceled by the customers.
- There are very few repeat customers but the cancellation among them is very less.
- This is a good indication as repeat customers are important for the hospitality industry as they can help in spreading the word of mouth.



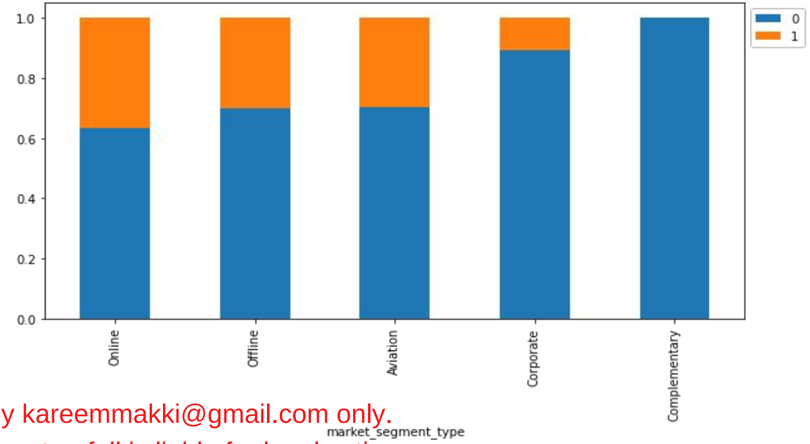


EDA Results

- Rooms booked online have high variations in prices.
- The offline and corporate room prices are similar.
- The complementary market segment gets the rooms at very low prices, which makes sense.

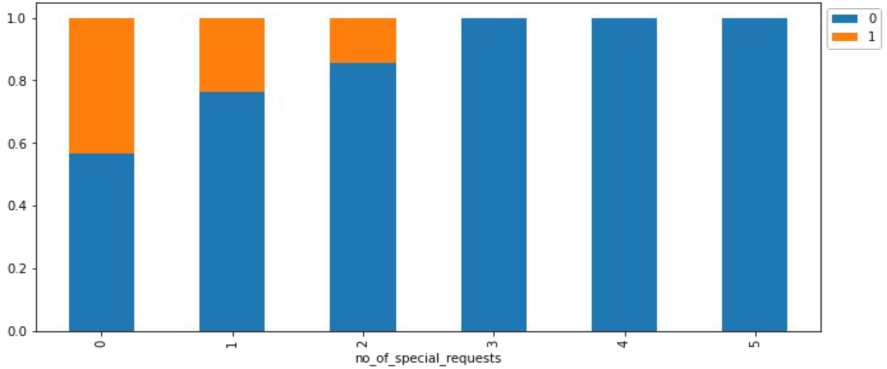


- Around 40% of the online booking were canceled.
- Bookings made offline are less prone to cancellations.
- Corporate segments show very low cancellations.

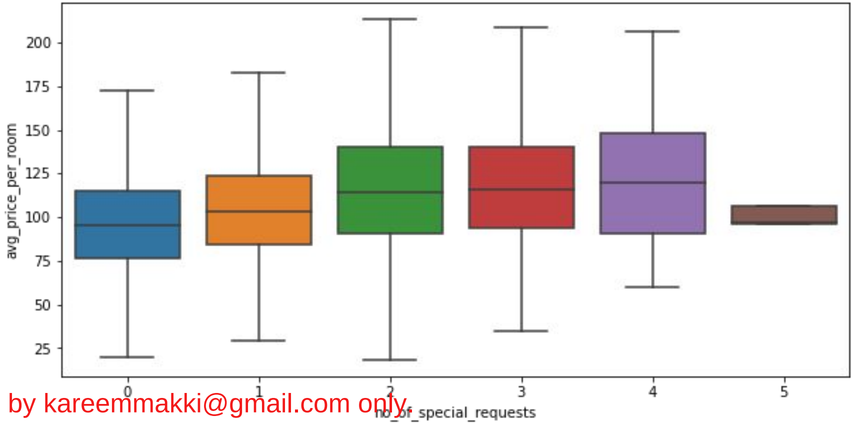


EDA Results

- There's a negative correlation between the number of special requests from the customer and the booking status.
- If a customer has made more than 2 requests there's a very high chance that the booking will not be canceled.



- The median prices of the rooms where some special requests were made by the customers are slightly higher than the rooms where customers didn't make any requests.



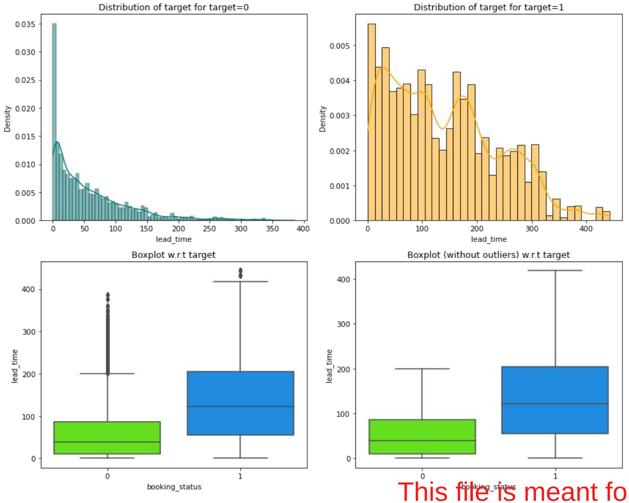
EDA Results

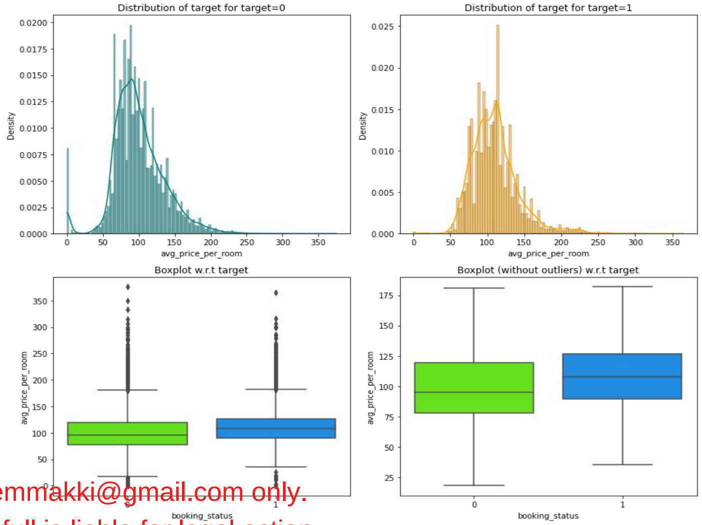
- The distribution of price for canceled and bookings which were not canceled is quite similar.

- The prices for the canceled bookings are slightly higher than the booking which were not canceled.

- There's a big difference in the median value of lead time for bookings that were canceled and bookings that were not canceled.

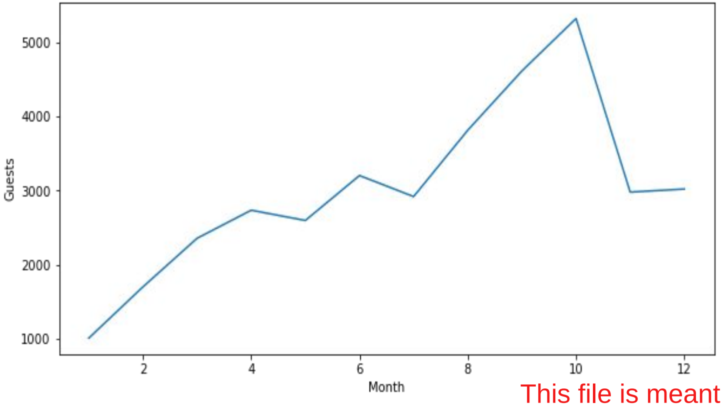
- The higher the lead time higher is the chances of a booking being canceled.

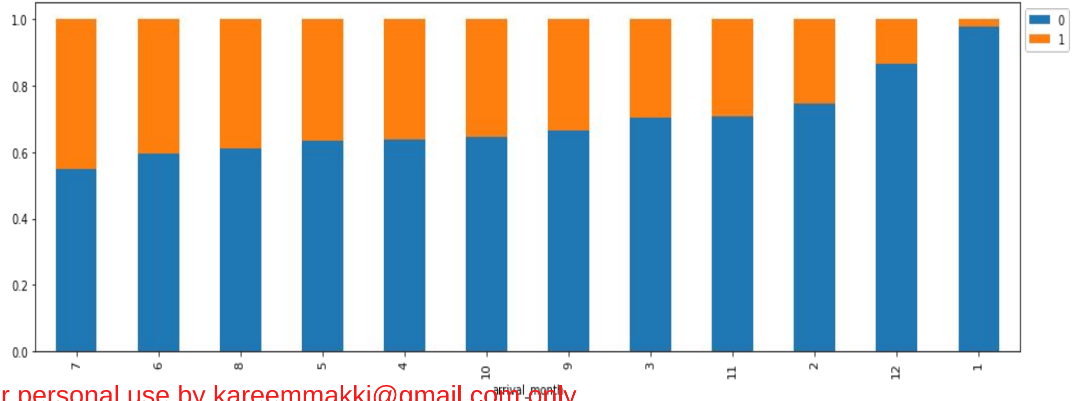




EDA Results

- The trend shows the number of bookings remains consistent from April to July and the hotel sees around 3000 to 3500 guests.
- Most bookings were made in October - more than 5000 bookings but 40% of these bookings got canceled.
- Least bookings were canceled in December and January - customers might have travelled to celebrate Christmas and New Year.





Model Performance Summary

- We want to predict whether a booking will be cancelled or not using the information provided to us.
- We will use the F1 Score as the performance metric for our model because
  - If we predict that a booking will not be canceled and the booking gets canceled then the hotel will lose resources and will have to bear additional costs of distribution channels.
  - If we predict that a booking will get canceled and the booking doesn't get canceled the hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be canceled. This might damage the brand equity.
  - F1 score will help us minimize both false positives and false negatives
- The Logistic Regression and Decision Tree models indicates that the most significant predictors of booking status are:
  - Lead Time
  - Number of special requests
  - Average price per room

Model Performance Summary

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
Logistic regression	0.80	0.80	0.63	0.63	0.73	0.72	0.68	0.67
Logistic Regression (threshold = 0.37)	0.79	0.79	0.73	0.73	0.66	0.66	0.70	0.70
Logistic Regression (threshold = 0.42)	0.80	0.80	0.69	0.70	0.69	0.69	0.69	0.69
Decision Tree	0.99	0.87	0.98	0.81	0.99	0.79	0.99	0.80
Decision Tree - Pre Pruning	0.83	0.83	0.78	0.78	0.72	0.72	0.75	0.75
Decision Tree - Post Pruning	0.89	0.86	0.90	0.85	0.81	0.76	0.85	0.80

Best performing model

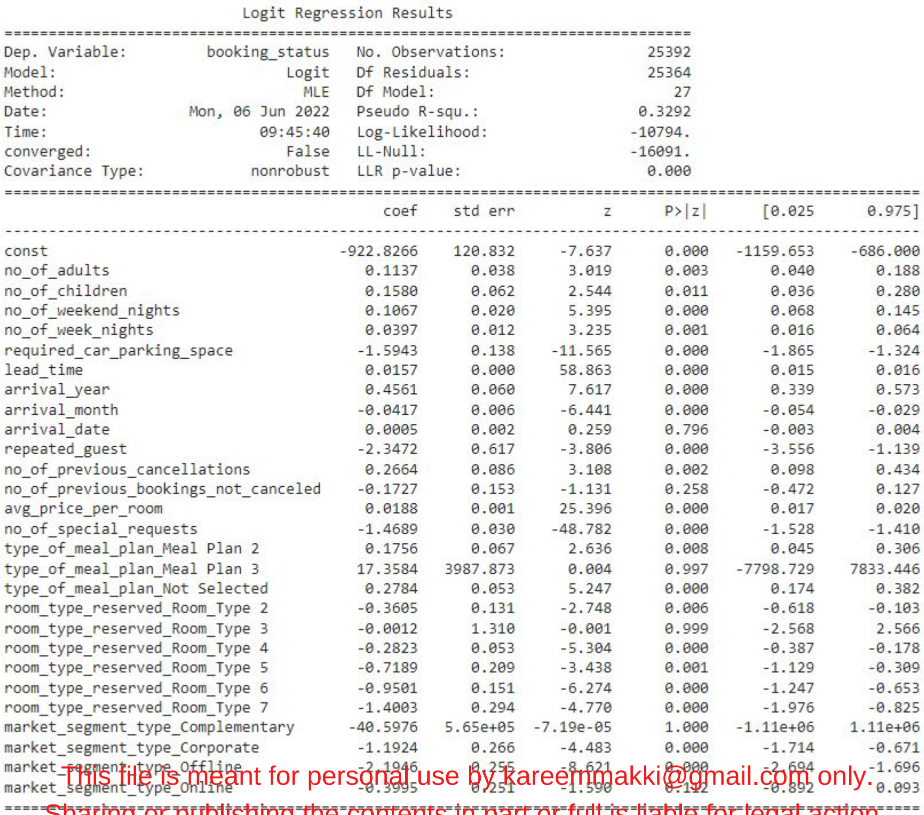
APPENDIX

Data Background and Contents

- The data contains information about 36275 booking records.
- The characteristics include number of adults, number of children, average price per room, type of meal plan selected, number of special requests from a customer, and more.
- Some columns (like average price per room, number of children) have some extreme and irregular values, which warrants an anomaly check.
- The values in the average price per room column which were greater than or equal to 500 were capped to the upper whisker value using the IQR method.
- In the no of children column, there were high values like 9 and 10, which were replaced with 3.
- There were quite a few outliers in some other columns of the data too. However, they were not treated since they are proper values.

Model Building - Logistic Regression

The model with default parameters is as shown below



```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25364
Method:                           MLE   Df Model:                           27
Date:                Mon, 06 Jun 2022   Pseudo R-squ.:                  0.3292
Time:                        09:45:40   Log-Likelihood:                -10794.
converged:                      False   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                 -922.8266    120.832     -7.637      0.000   -1159.653    -686.000
no_of_adults                             0.1137      0.038      3.019      0.003       0.040       0.188
no_of_children                           0.1580      0.062      2.544      0.011       0.036       0.280
no_of_weekend_nights                     0.1067      0.020      5.395      0.000       0.068       0.145
no_of_week_nights                        0.0397      0.012      3.235      0.001       0.016       0.064
required_car_parking_space              -1.5943      0.138    -11.565      0.000      -1.865      -1.324
lead_time                                0.0157      0.000     58.863      0.000       0.015       0.016
arrival_year                             0.4561      0.060      7.617      0.000       0.339       0.573
arrival_month                           -0.0417      0.006     -6.441      0.000      -0.054      -0.029
arrival_date                             0.0005      0.002      0.259      0.796      -0.003       0.004
repeated_guest                          -2.3472      0.617     -3.806      0.000      -3.556      -1.139
no_of_previous_cancellations             0.2664      0.086      3.108      0.002       0.098       0.434
no_of_previous_bookings_not_canceled    -0.1727      0.153     -1.131      0.258      -0.472       0.127
avg_price_per_room                       0.0188      0.001     25.396      0.000       0.017       0.020
no_of_special_requests                  -1.4689      0.030    -48.782      0.000      -1.528      -1.410
type_of_meal_plan_Meal Plan 2            0.1756      0.067      2.636      0.008       0.045       0.306
type_of_meal_plan_Meal Plan 3           17.3584   3987.873      0.004      0.997   -7798.729    7833.446
type_of_meal_plan_Not Selected           0.2784      0.053      5.247      0.000       0.174       0.382
room_type_reserved_Room_Type 2          -0.3605      0.131     -2.748      0.006      -0.618      -0.103
room_type_reserved_Room_Type 3          -0.0012      1.310     -0.001      0.999      -2.568       2.566
room_type_reserved_Room_Type 4          -0.2823      0.053     -5.304      0.000      -0.387      -0.178
room_type_reserved_Room_Type 5          -0.7189      0.209     -3.438      0.001      -1.129      -0.309
room_type_reserved_Room_Type 6          -0.9501      0.151     -6.274      0.000      -1.247      -0.653
room_type_reserved_Room_Type 7          -1.4003      0.294     -4.770      0.000      -1.976      -0.825
market_segment_type_Complementary      -40.5976   5.65e+05  -7.19e-05      1.000   -1.11e+06    1.11e+06
market_segment_type_Corporate           -1.1924      0.266     -4.483      0.000      -1.714      -0.671
market_segment_type_Offline             -2.1946      0.255     -8.621      0.000      -2.694      -1.696
market_segment_type_Online              -0.3995      0.251     -1.590      0.112      -0.892       0.093
========================================================================================================
```

Model Assumptions

- Multicollinearity was checked by using VIF
  - All numerical variables with VIF > 5 were dropped
  - VIF for the constant and dummy variables were ignored

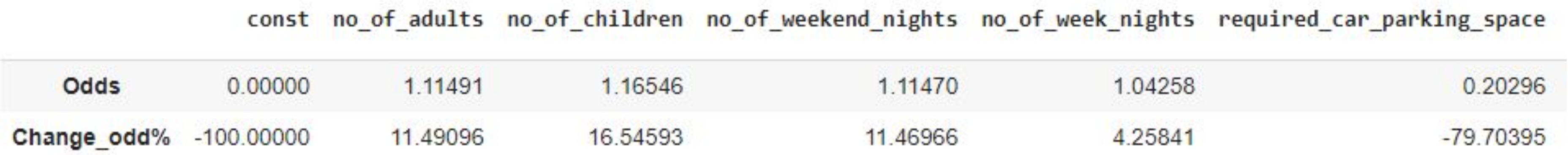
- Dropping high p-value variables

  It has been done using the following procedure

  - Build a model, check the p-values of the variables, and drop the column with the highest p-value.
  - Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
  - Repeat the above two steps till there are no columns with p-value > 0.05.

Coefficients interpretation

- Coefficients of required_car_parking_space, arrival_month, repeated_guest, no_of_special_requests and some others are negative, an increase in these will lead to a decrease in chances of a customer canceling their booking.
- Coefficients of no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, lead_time, avg_price_per_room, type_of_meal_plan_Not Selected and some others are positive, an increase in these will lead to a increase in the chances of a customer canceling their booking.

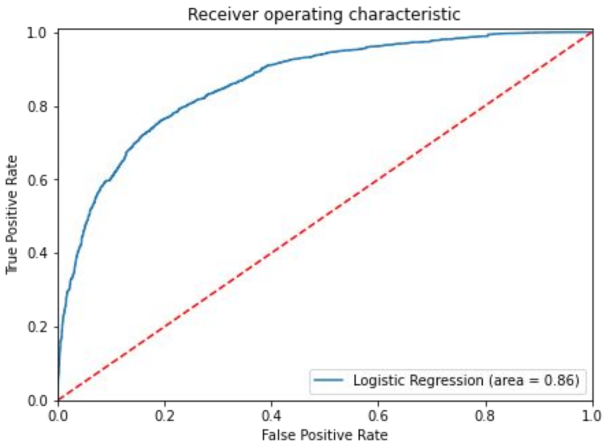


	const	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space
Odds	0.00000	1.11491	1.16546	1.11470	1.04258	0.20296
Change_odd%	-100.00000	11.49096	16.54593	11.46966	4.25841	-79.70395

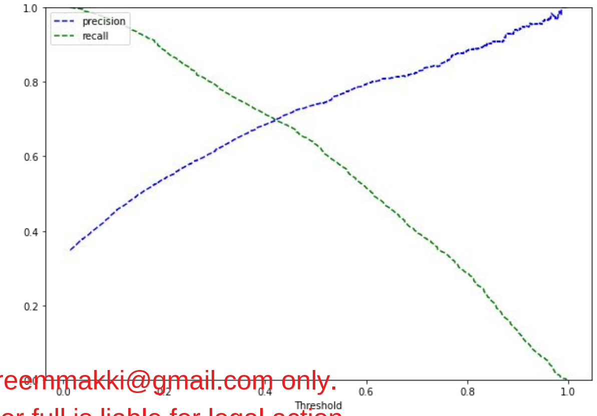
- no_of_adults: Holding all other features constant a 1 unit change in the number of children will increase the odds of a booking getting cancelled by 1.11 times or a 11.49% increase in the odds of a booking getting cancelled.
- no_of_children: Holding all other features constant a 1 unit change in the number of children will increase the odds of a booking getting cancelled by 1.16 times or a 16.54% increase in the odds of a booking getting cancelled.etc.

Model Performance Evaluation and Improvement - Logistic Regression

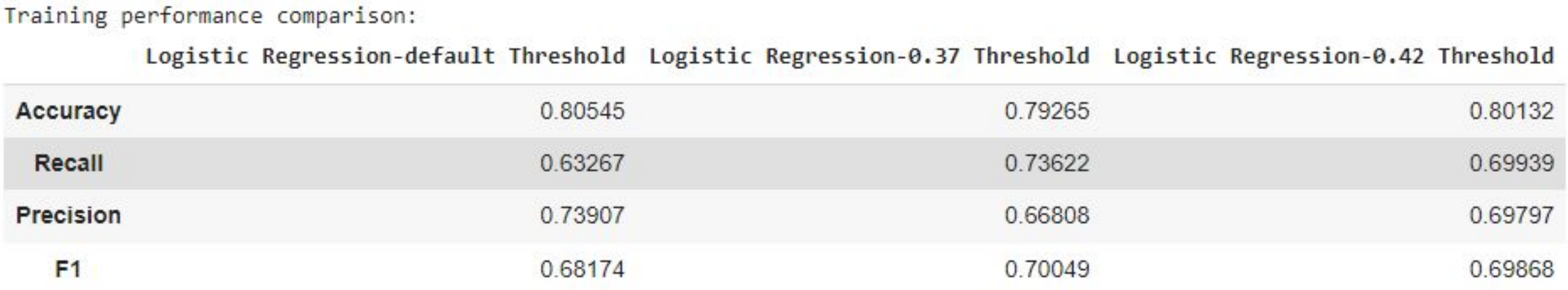
- The optimal threshold obtained using AUC-ROC Curve is 0.37



- The optimal threshold obtained using the Precision-Recall curve is 0.42

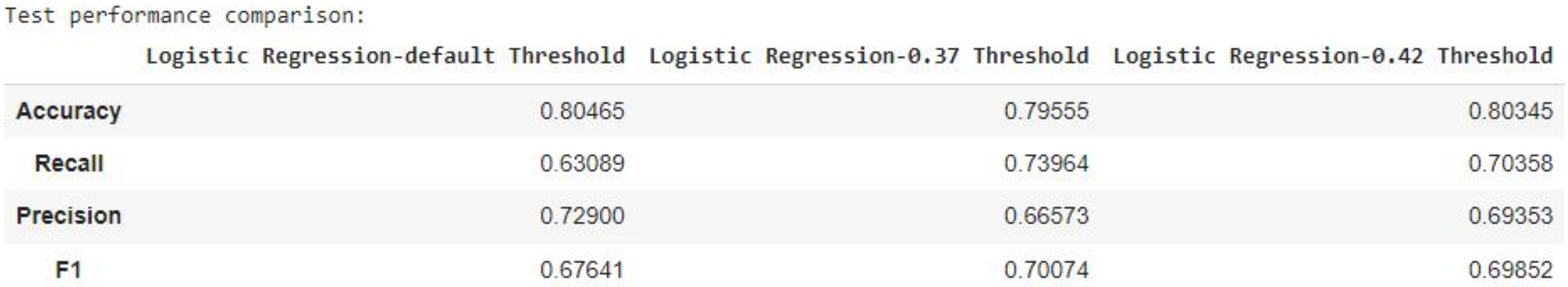


Model Performance Evaluation and Improvement - Logistic Regression



Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868



Test performance comparison:

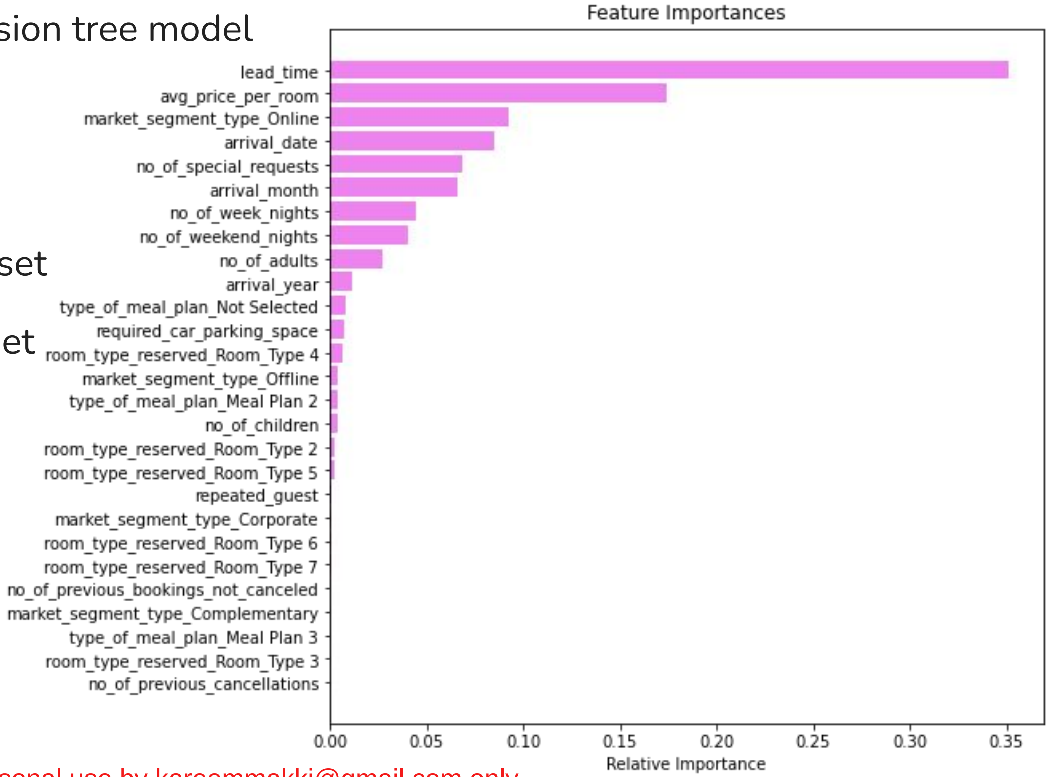
	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

Model Performance Evaluation and Improvement - Logistic Regression

- **Using the model with default threshold the model will give a low recall but good precision score - ** The hotel will be able to predict which bookings will not be cancelled and will be able to provide satisfactory services to those customers which help in maintaining the brand equity but will lose on resources.
- **Using the model with a 0.37 threshold the model will give a high recall but low precision score -** The hotel will be able to save resources by correctly predicting the bookings which are likely to be cancelled but might damage the brand equity.
- **Using the model with a 0.42 threshold the model will give a balance recall and precision score -** The hotel will be able to maintain a balance between resources and brand equity.

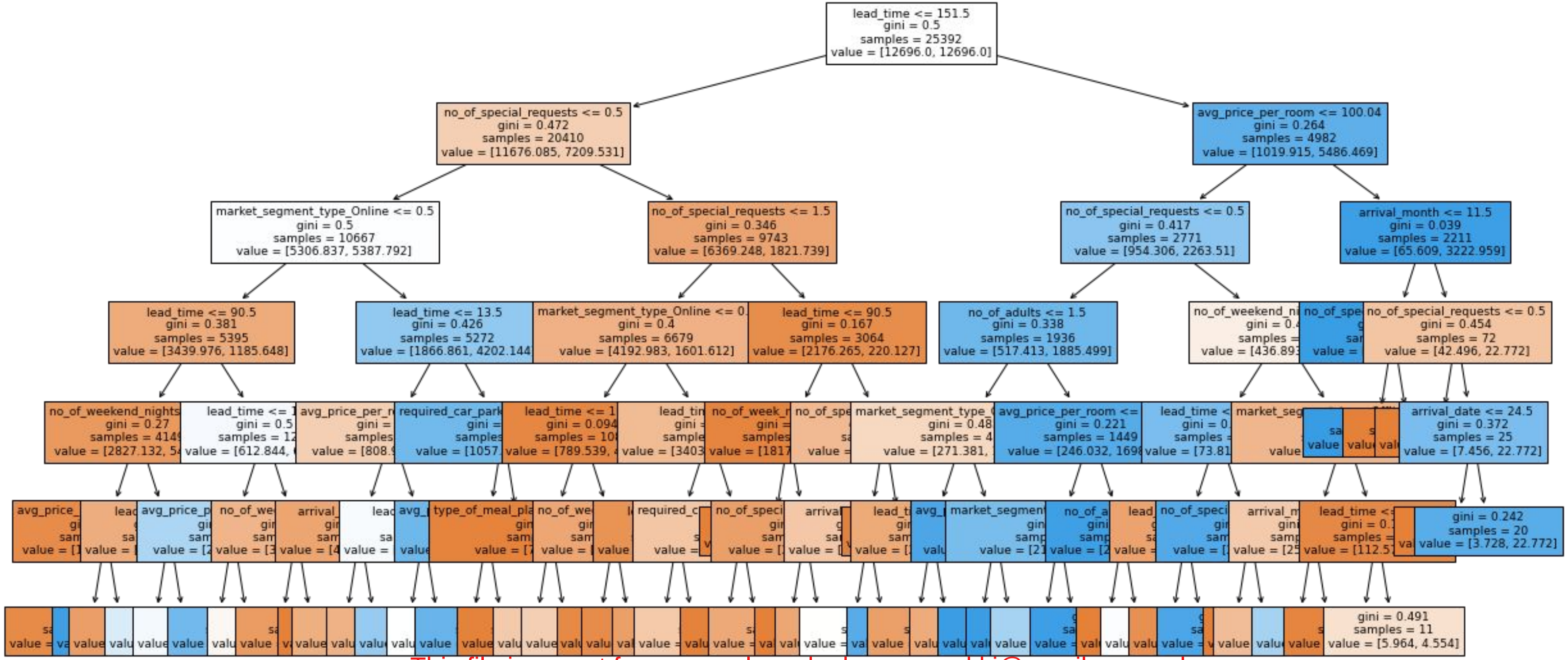
Model Building - Decision Tree

- Following are the steps to build the decision tree model
  - Data preparation
  - Model building
  - Evaluate the performance on train set
  - Evaluate the performance on test set
  - Check for important features



Model Performance Evaluation and Improvement - Decision Tree

- Model performance improvement using Pre-Pruning
- Visualization of pre pruned tree is as follows



Model Performance Evaluation and Improvement - Decision Tree

The rules obtained from the decision tree can be interpreted as:

- The rules show that lead time plays a key role in identifying if a booking will be cancelled or not. 151 days has been considered as a threshold value by the model to make the first split.

Bookings made more than 151 days before the date of arrival:

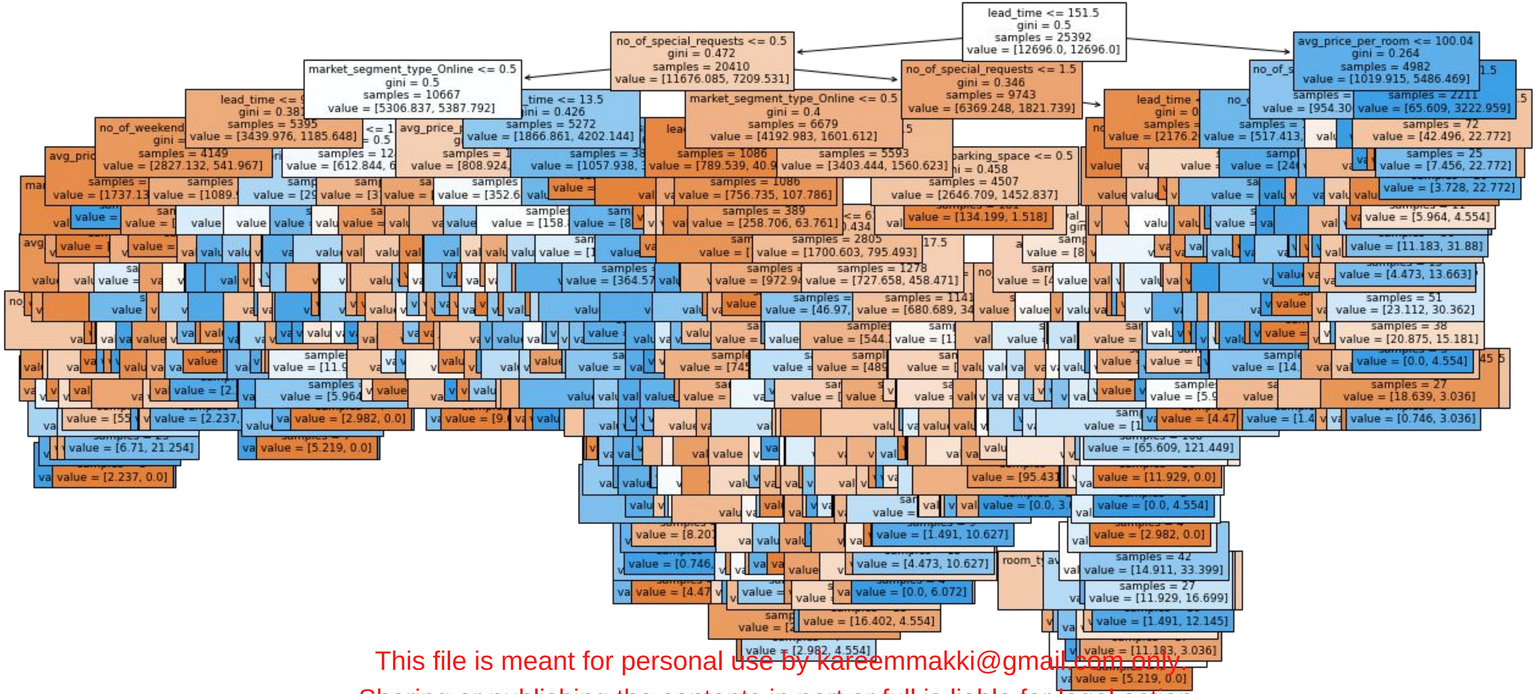
- If the average price per room is greater than 100 euros and the arrival month is December, then the booking is less likely to be cancelled.
- If the average price per room is less than or equal to 100 euros and the number of special request is 0, then the booking is likely to get canceled.

Bookings made under 151 days before the date of arrival:

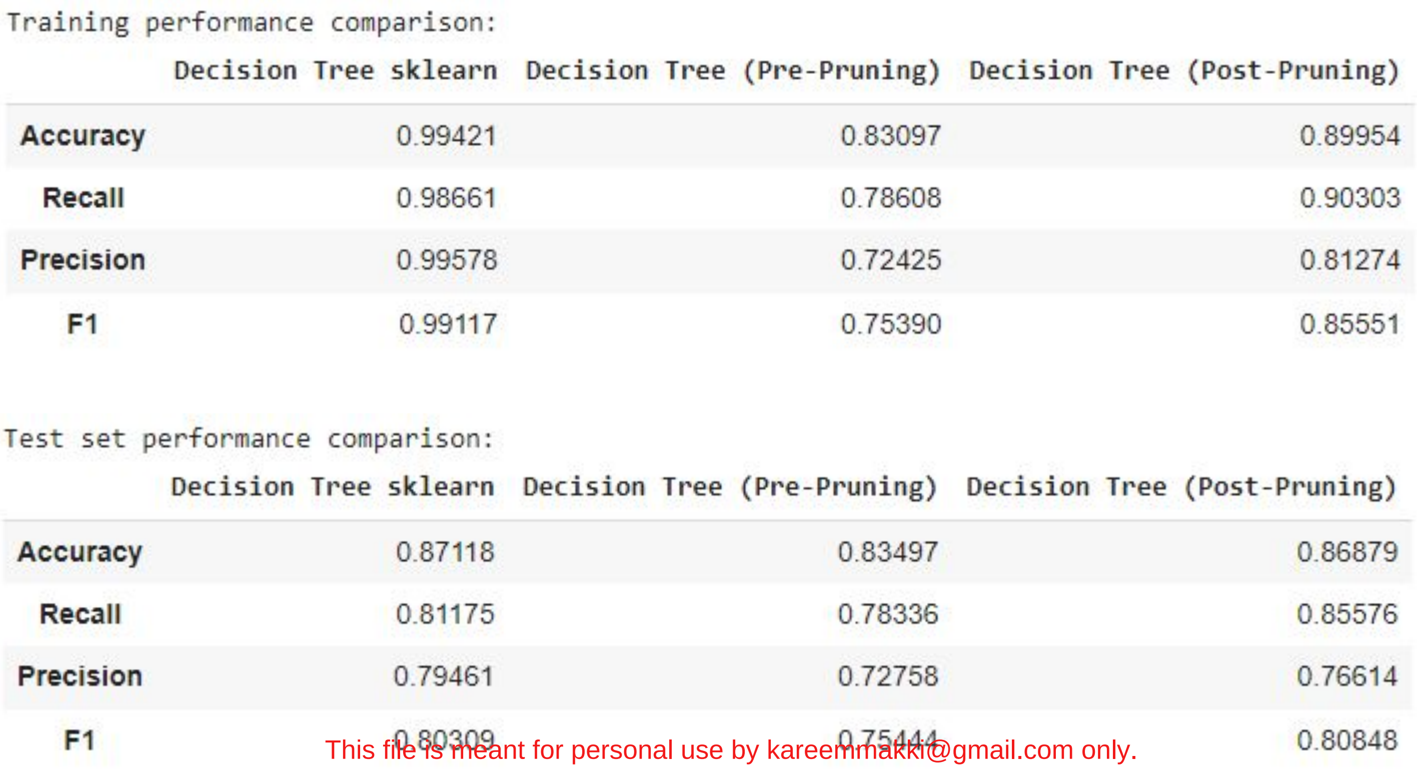
- If a customer has at least 1 special request the booking is less likely to be cancelled.
- If the customer didn't make any special requests and the booking was done Online it is more likely to get canceled, if the booking was not done online, it is less likely to be canceled.

Model Performance Evaluation and Improvement - Decision Tree

- Model performance improvement using Cost Complexity Pruning (Post Pruning)
- Visualization of post pruned tree is as follows



Model Performance Evaluation and Improvement - Decision Tree



Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89954
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81274
F1	0.99117	0.75390	0.85551

Test set performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86879
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

Model Performance Evaluation and Improvement - Decision Tree

- The decision tree model with default parameters is overfitting the training data and is not able to generalize well.
- The pre-pruned tree has given a generalized performance with balanced values of precision and recall.
- The post-pruned tree is giving a high F1 score as compared to other models but the difference between precision and recall is high.
- The hotel will be able to maintain a balance between resources and brand equity using the pre-pruned decision tree model.

greatlearning
Power Ahead

Happy Learning !