INN Hotels SLC Project Business Presentation

Contents

1. Business Problem Overview and Solution Approach
2. Data Overview
3. Exploratory Data Analysis (EDA) – Univariate Analysis
4. Exploratory Data Analysis (EDA) – Correlation Matrix
5. Exploratory Data Analysis (EDA) – Bivariate and Multivariate Analysis
6. Logistic Regression Model Performance Summary
7. Decision Tree Model Performance Summary
8. Actionable Insights and Recommendations

Business Problem Overview and Solution Approach

- Core Business Idea - 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 adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

  The cancellation of bookings impact a hotel on various fronts:
  - Loss of resources (revenue) when the hotel cannot resell the room.
  - Additional costs of distribution channels by increasing commissions to help sell these rooms.
  - Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
  - Human resources to plan for the guests.
- Problem to tackle - The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist have 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.

Data Overview

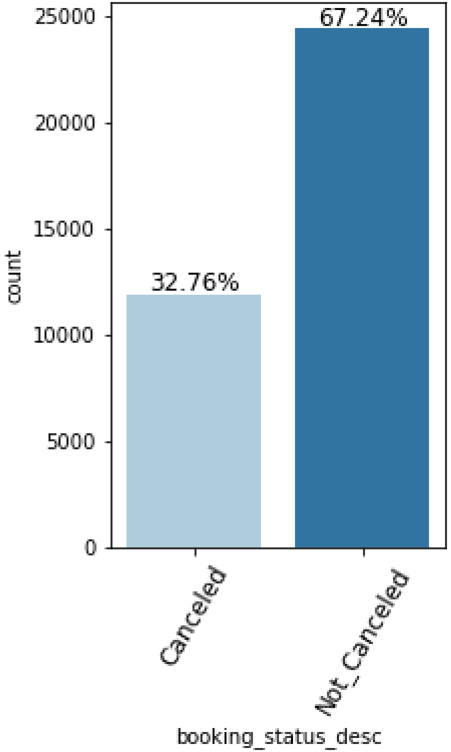
- The data contains the different attributes of Hotel room booking.

Data Dictionary	
Field Name	Description
Booking_ID	the unique identifier of each booking
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
no_of_week_nights	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
type_of_meal_plan	Type of meal plan booked by the customer:
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)
room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
arrival_month	Month of arrival date
arrival_date	Date of the month
market_segment_type	Market segment designation.
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
no_of_special_requests	Total number of special requests made by the customer
booking_status	Flag indicating if the booking was canceled or not.

Exploratory Data Analysis (EDA)

- Univariate Analysis - Dependent Variable – booking status

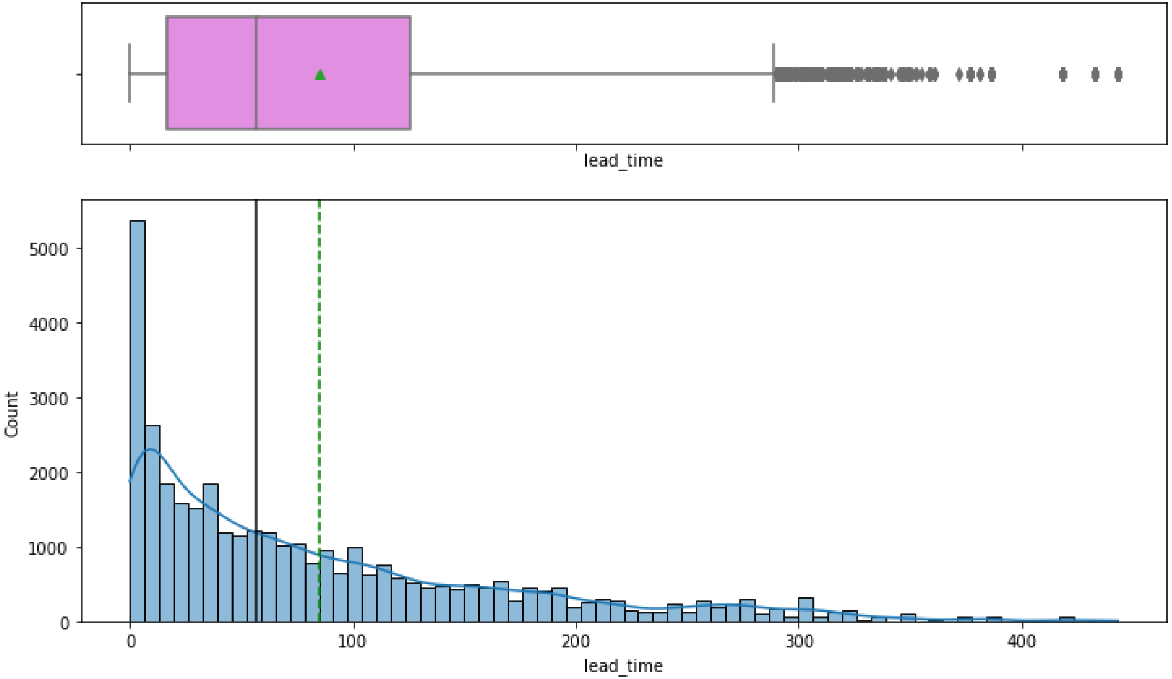
- The field booking status has two unique values - 'Not Canceled' and 'Canceled'.
- There are 67% of observations with booking status = 'Not Canceled' and 33% of observations with booking status = 'Canceled'



Exploratory Data Analysis (EDA)

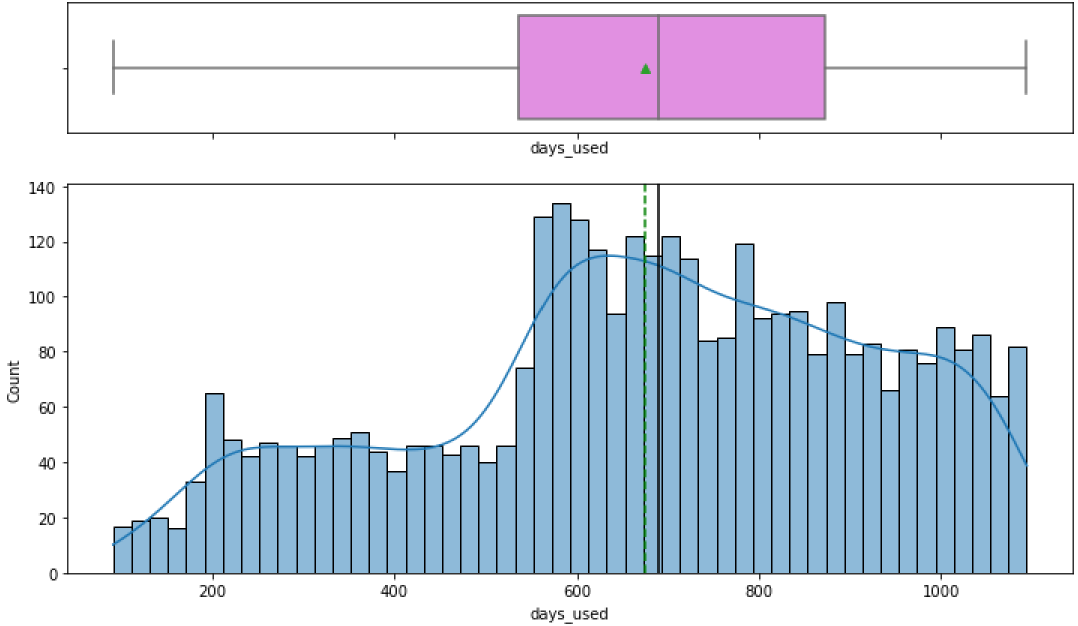
- Univariate Analysis - Independent Variable – lead-time

- There are plenty of outliers in the data as displayed in box plot.
- The mean is greater than median, that indicates the distribution is skewed to the right.
- About 68% of the values for 'lead-time' are from 15 to 130 days.



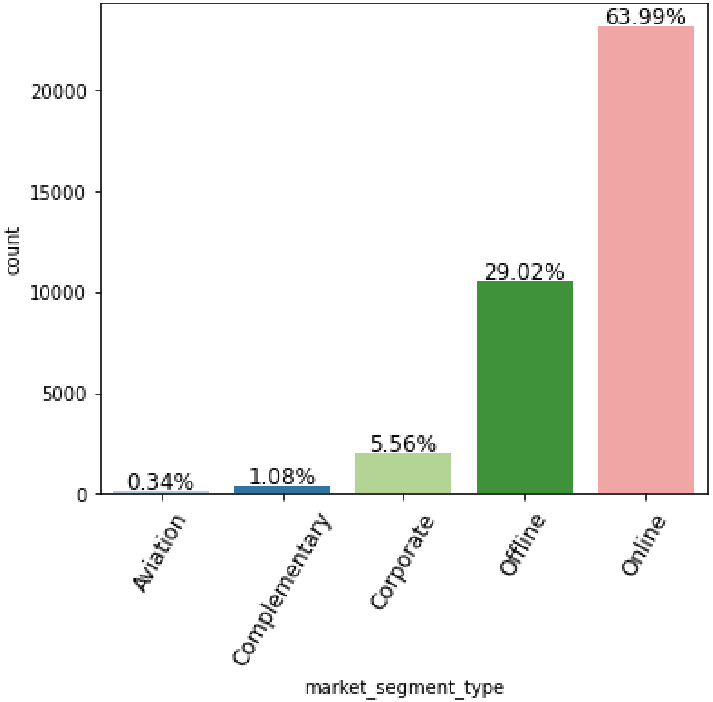
Exploratory Data Analysis (EDA)

- Univariate Analysis - Independent Variable – Days Used
  - There are no outliers in the data as displayed in box plot.
  - The mean is less than median, that indicates the distribution is skewed to the left.
  - About 68% of the values for 'days_used' are from 536 to 872 days.



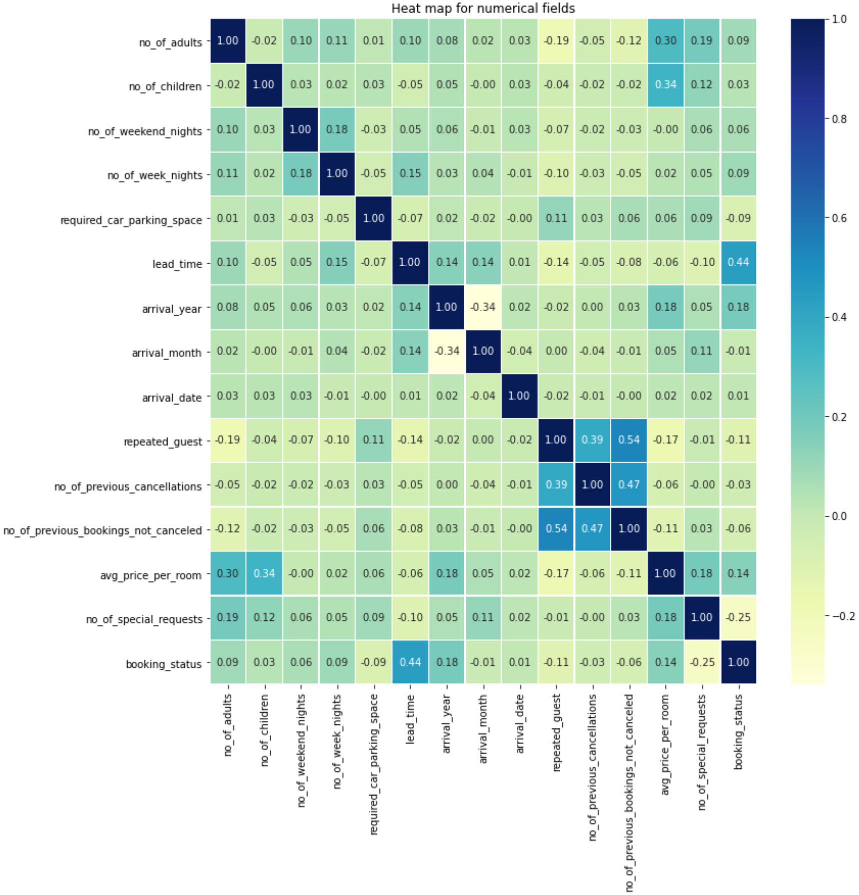
Exploratory Data Analysis (EDA)

- Univariate Analysis - Independent Variable – market_segment_type
  - The field 'market_segment_type' has 5 unique values, Online booking constitute to 64% of the data.
  - 29% of customers opted for Offline booking.



Exploratory Data Analysis (EDA) - Correlation Matrix

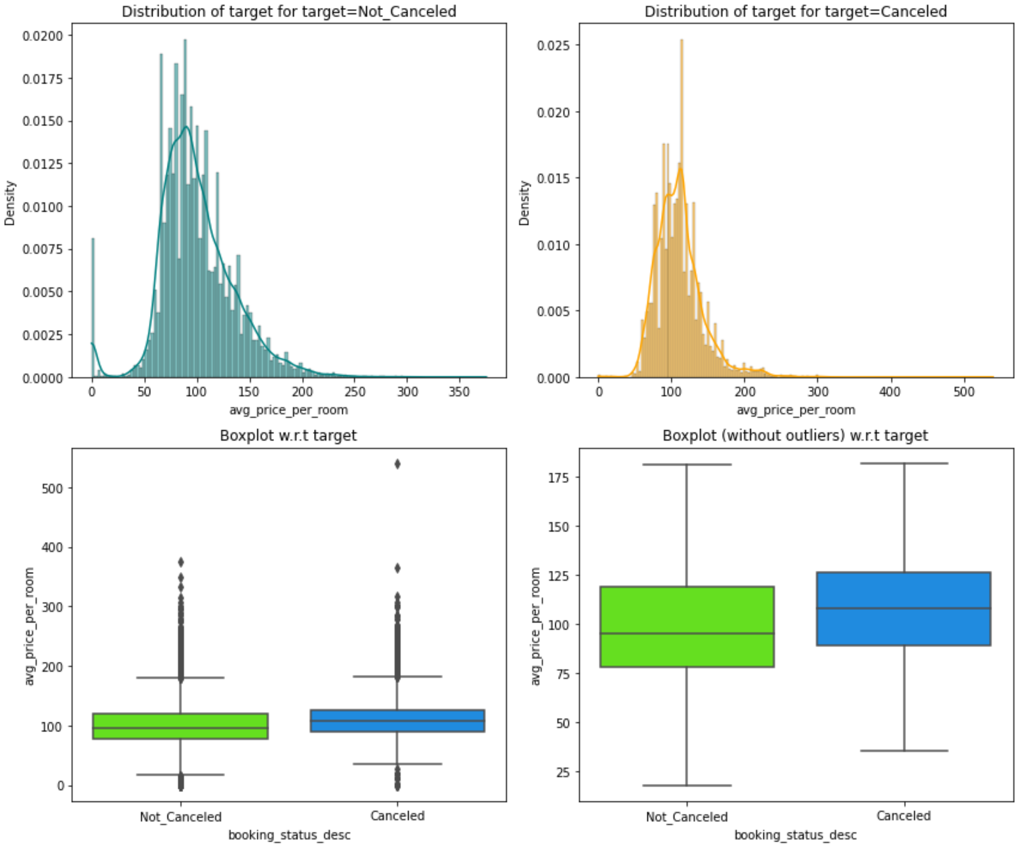
1. There is a positive correlation (0.44) between lead-time and booking status. In other words, the bookings with greater lead time has more chances of getting cancelled.
2. There is a negative correlation (-0.25) between no_of_special_requests and booking status. Hence the booking with fewer number of special requests are likely to get cancelled.



Exploratory Data Analysis (EDA)

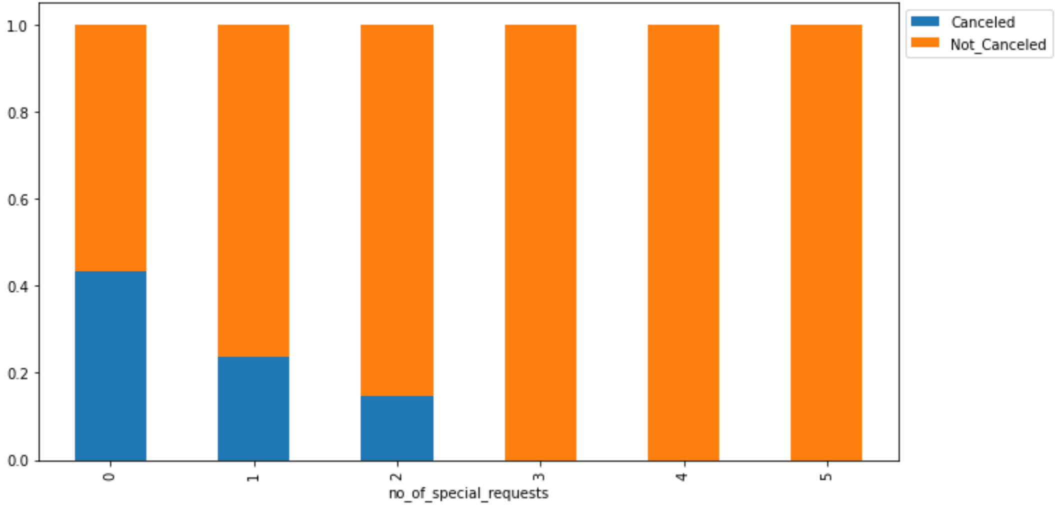
- Bivariate Analysis
  – avg price per room
  vs. booking status

  - As observed in heatmap, the average price of the room does not play a major factor in deciding the booking status.
  - Being said that it can be observed that the median avg room price for a not-cancelled booking is less than 100 euros whereas the median avg room price for a cancelled booking is more than 100 euros.



Exploratory Data Analysis (EDA)

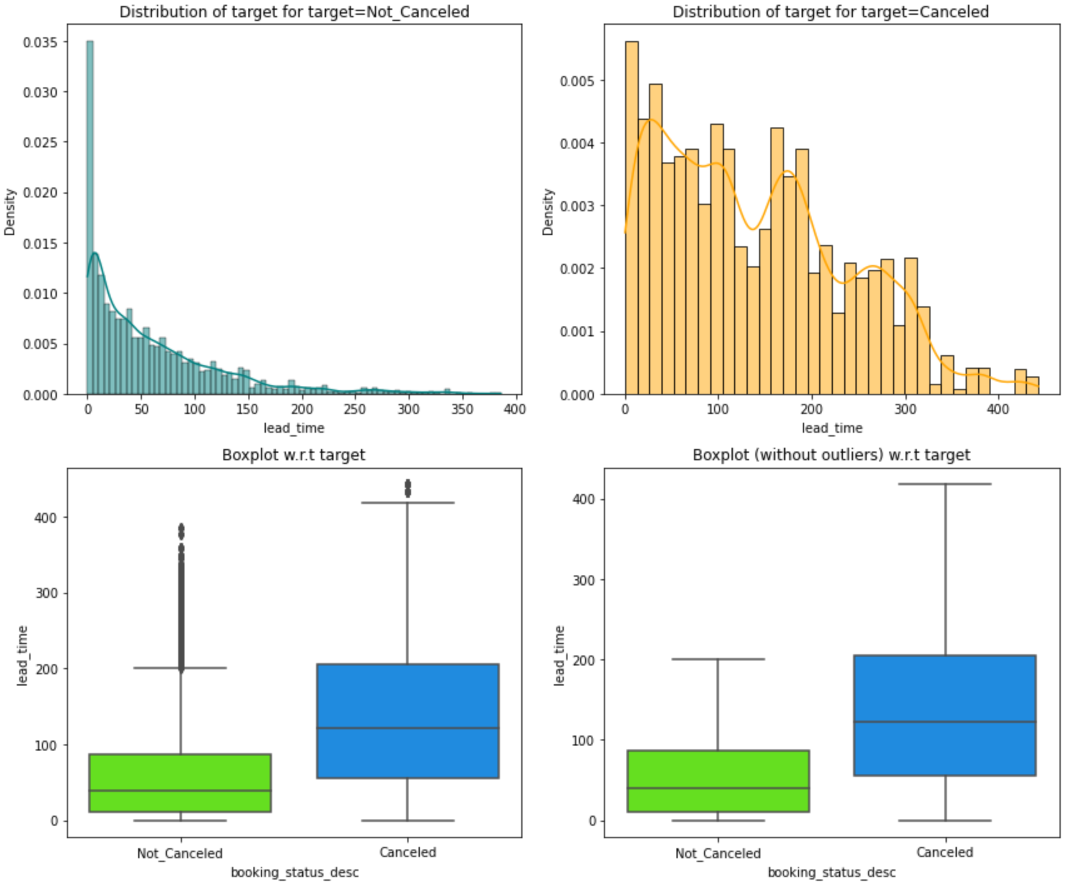
- Bivariate Analysis – no of special requests and booking status
  - As observed in heatmap, the no_of_special_requests does play a major factor in deciding the booking status.
  - Bookings have less cancellation for no_of_special_requests 3, 5 and 5, but the there are only a few hundreds of observations for that condition.



Exploratory Data Analysis (EDA)

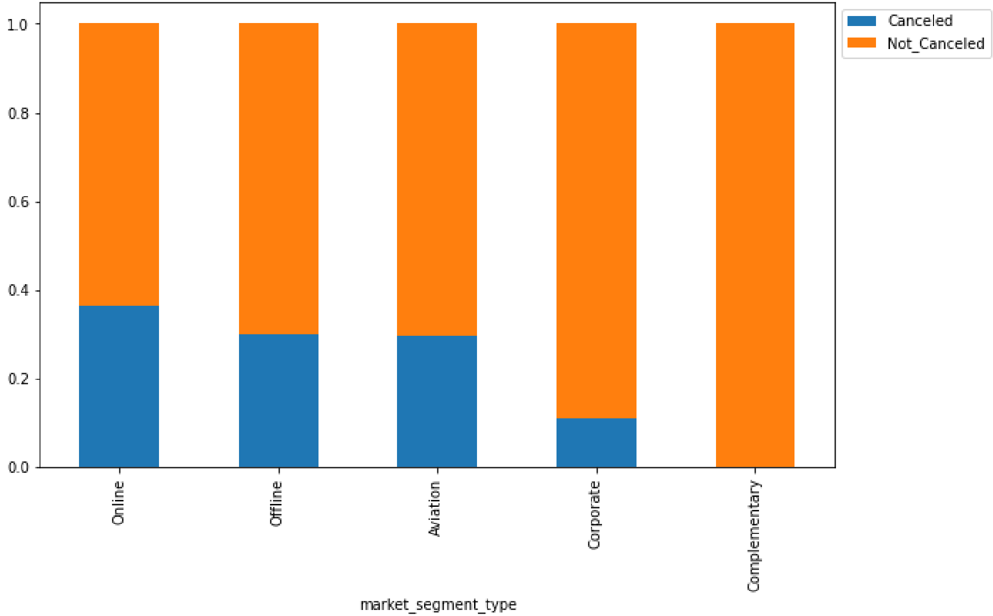
- Bivariate Analysis –
  lead time and booking status

  - As observed in heatmap, booking with higher lead-time has a higher chance of getting cancelled.
  - The median lead-time for a not-cancelled booking is less than 50 days whereas the median lead-time for a cancelled booking is more than 100 days.



Exploratory Data Analysis (EDA)

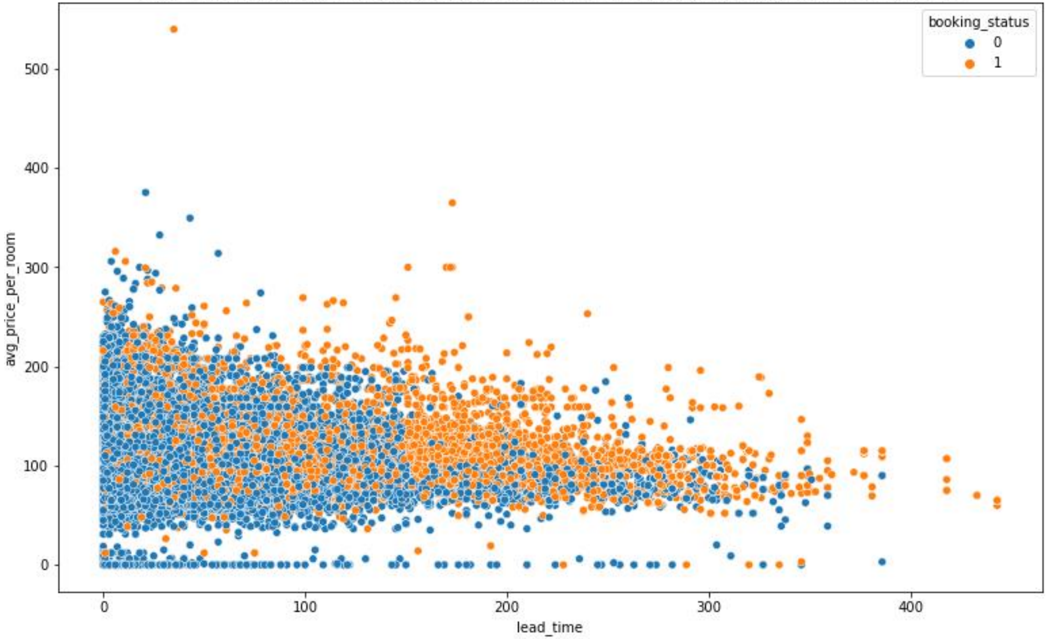
- Bivariate Analysis – market segment type and booking status
  - As observed in heatmap, the market_segment_type has a minor influence in deciding the booking status.
  - None of the complementary bookings are cancelled, which is significant.



Exploratory Data Analysis (EDA)

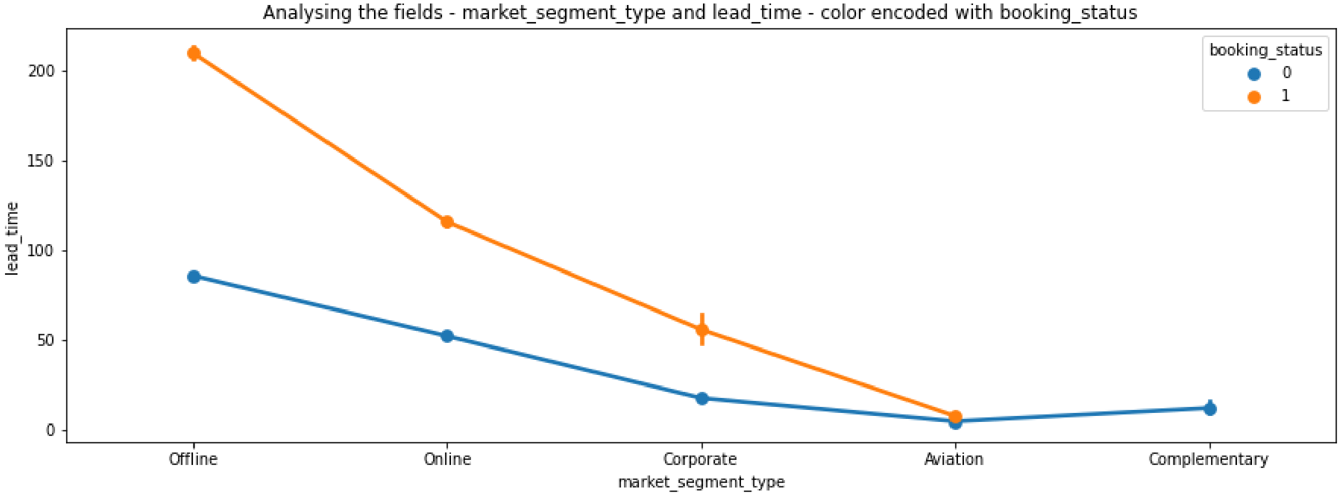
- Multivariate Analysis – lead-time, avg price per room and booking status
  - As we can see, the higher the lead time greater has the chance of getting cancelled (1 stands for cancellation). In other words, the bookings with least lead-time has less chance to getting cancelled.
  - There is no clear relationship between average price per room and booking status, but we can see that the lower the average price, there is less chance of getting cancelled.
  - Also notice that the average price is low when lead-time is high.

Multivariate analysis of the fields using box plot- lead_time, avg_price_per_room and booking_status



Exploratory Data Analysis (EDA)

- Multivariate Analysis – market segment type and lead-time - color encoded with booking status
  - As we can see, there are no cancellations for complementary bookings, but since that does not impact the revenue, it can be ignored.
  - The difference in average lead time between cancelled and not-cancelled bookings is high for offline market segment, meaning the offline bookings are done well in advance and is getting cancelled.
  - The difference in average lead time between cancelled and not-cancelled bookings is lower for online and corporate market segments and is almost zero for aviation segment.

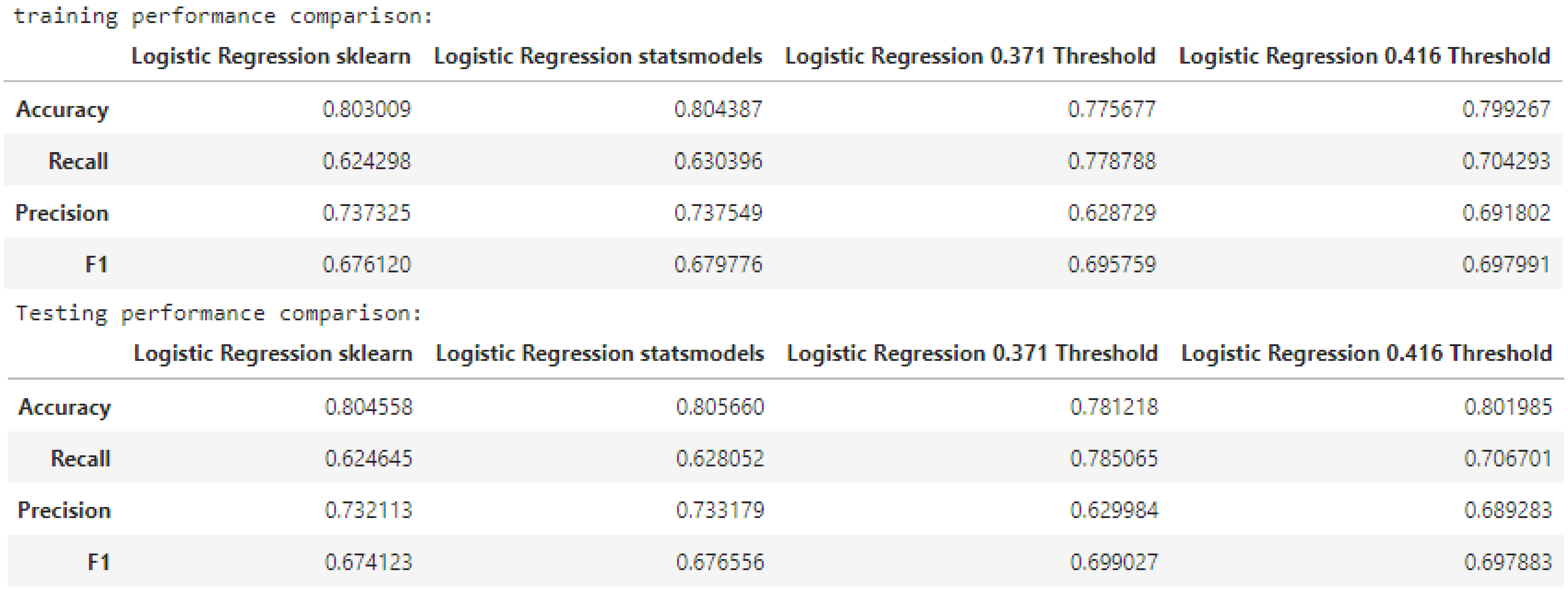


Logistic Regression Model Performance Summary

- All the models are giving a generalized performance on training and test set.
- The highest recall is 78.5% on the testing set with 0.317 threshold.
- Using the model with sklearn and statsmodels (default threshold) will give a low recall but good precision scores - This model will help the hotel save resources but will incorrectly predict potential non-cancelling customers.
- Using the model with 0.317 threshold the model will give a high recall but low precision scores - This model will help the hotel identify non-cancelling customers effectively, but the cost of resources will be high.
- Using the model with 0.416 threshold the model will give a balance recall and precision score - This model will help the hotel to maintain a balance in identifying potential non-cancelling customers and the cost of resources.

Logistic Regression Model Performance Summary

- Summary of key performance metrics for training and test data in tabular format for comparison



```
training performance comparison:
```

	Logistic Regression sklearn	Logistic Regression statsmodels	Logistic Regression 0.371 Threshold	Logistic Regression 0.416 Threshold
Accuracy	0.803009	0.804387	0.775677	0.799267
Recall	0.624298	0.630396	0.778788	0.704293
Precision	0.737325	0.737549	0.628729	0.691802
F1	0.676120	0.679776	0.695759	0.697991

```
Testing performance comparison:
```

	Logistic Regression sklearn	Logistic Regression statsmodels	Logistic Regression 0.371 Threshold	Logistic Regression 0.416 Threshold
Accuracy	0.804558	0.805660	0.781218	0.801985
Recall	0.624645	0.628052	0.785065	0.706701
Precision	0.732113	0.733179	0.629984	0.689283
F1	0.674123	0.676556	0.699027	0.697883

Logistic Regression Model Performance Summary

Summary of most important factors used by the ML model for prediction

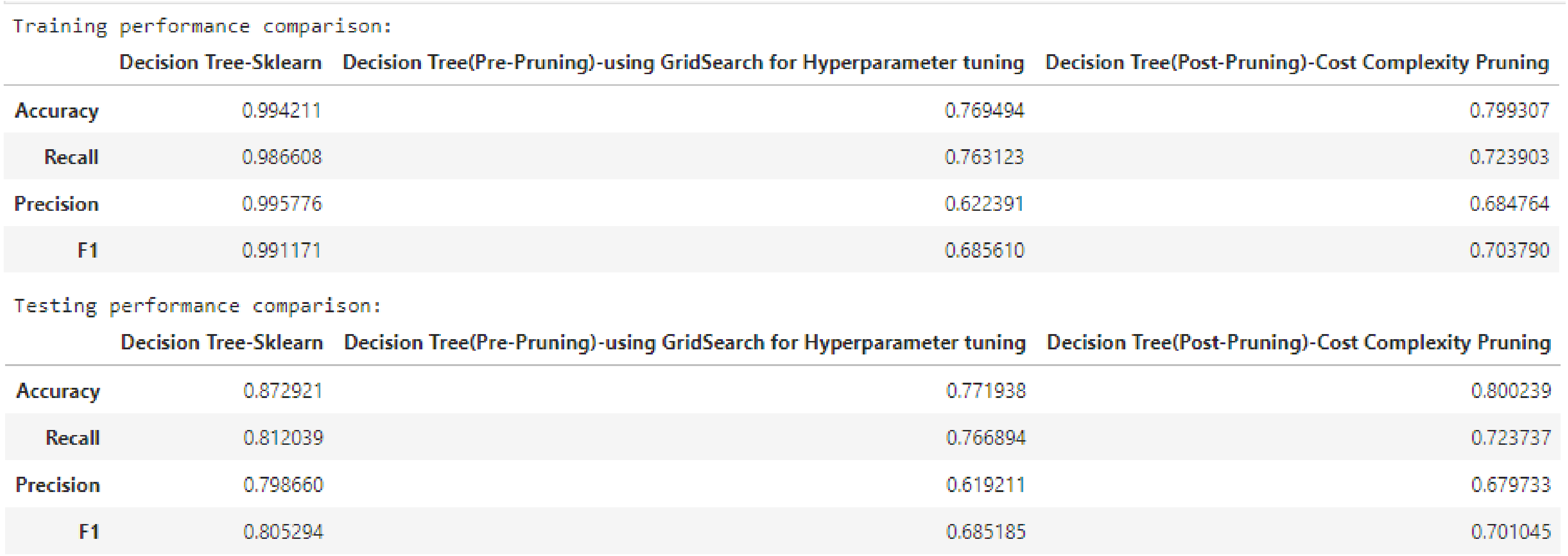
1. Repeated Guest has a negative correlation to cancellation. So, if the guest is a regular customer, there is a less chance of cancellation.
2. Market segments Offline and Corporate has negative correlation to cancellation. So, if the guest is from Offline or Corporate market segment, there is a less chance of cancellation.
3. Required_car_parking_space has a negative correlation to cancellation. So, if the guest has requested for a parking space, there is a less chance of cancellation.
4. Just like decision tree model, there is lesser chance of cancellation if the hotel can accommodate special requests from customers.
5. No_of_previous_cancellations has a positive correlation to cancellation. So, if the guest has cancelled before, there are chances that he/she will cancel again.

Decision Tree Model Performance Summary

- The sklearn model is an overfit model, but the other 2 pruned models are giving a generalized performance on training and test set.
- The highest recall is 76.7% on the testing set using GridSearch for Hyperparameter tuning.
- Using the model with sklearn is overfit and is hard to interpret business rules from it.
- Using the model with using GridSearch for Hyperparameter tuning will give a high recall but low precision scores - This model will help the hotel identify non-cancelling customers effectively, but the cost of resources will be high.
- Using the model with Cost Complexity Pruning will give a balance recall and precision score - This model will help the hotel to maintain a balance in identifying potential non-cancelling customers and the cost of resources.

Decision Tree Model Performance Summary

- Summary of key performance metrics for training and test data in tabular format for comparison



Training performance comparison:

	Decision Tree-Sklearn	Decision Tree(Pre-Pruning)-using GridSearch for Hyperparameter tuning	Decision Tree(Post-Pruning)-Cost Complexity Pruning
Accuracy	0.994211	0.769494	0.799307
Recall	0.986608	0.763123	0.723903
Precision	0.995776	0.622391	0.684764
F1	0.991171	0.685610	0.703790

Testing performance comparison:

	Decision Tree-Sklearn	Decision Tree(Pre-Pruning)-using GridSearch for Hyperparameter tuning	Decision Tree(Post-Pruning)-Cost Complexity Pruning
Accuracy	0.872921	0.771938	0.800239
Recall	0.812039	0.766894	0.723737
Precision	0.798660	0.619211	0.679733
F1	0.805294	0.685185	0.701045

Decision Tree Model Performance Summary

Summary of most important factors used by the ML model for prediction

1. Lead time is a major factor that influences the cancellation of a hotel booking. Higher the lead time of booking, higher the chance of cancellation.
2. Online market segment also influences the cancellation of a hotel booking. If the market segment is online, higher the chance of cancellation than any other segments.
3. The more the number of special requests are accommodated by hotel, the lesser the chance of cancellation.
4. Bookings with lower average price per room have lesser chance of getting cancelled. Higher the price, higher the chance of getting cancelled.

Business Insights and Recommendations

1. Even though customers with higher lead time has a good chance of cancellation, INN Hotels should allow customers to book early as filling hotel rooms is also important for revenue. Now the catch is to find the customers who will cancel and have them cancel as early as possible so that the hotel can market that room again at a higher price. INN hotels should send weekly reminders (email and texts to customers about their advance booking and remind them about the consequences of not cancelling the room in advance.
2. INN hotels should try to accommodate as many special requests as possible from customers. This will drive customer satisfaction and will reduce last minute cancellations.
3. INN hotels should try to give coupons/discounts for bookings with less lead time as these bookings with lower average prices have less chances of getting cancelled.
4. INN Hotels should try to negotiate deals with Corporates to have them book the hotel for their staff, as Corporate market segment have less chances of getting cancelled.
5. INN Hotels should initiate a reward point system to reward regular customers and provide them with room upgrades. Repeated Guests have less chances of cancelling their booking.

greatlearning
Power Ahead

Happy Learning !