Star Hotels – Project

Machine Learning Solution For Predicting Cancellation

Contents

Business Problem Overview and Solution Approach

- ## Context of Business Problem

  Large no of cancellation is a huge impact for business as it impacts in multiple ways on their revenues. 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 for hotels. Such losses are particularly high on last minute cancellations. The new technologies such as online booking has changed the customer's booking possibilities and how hotel handles the cancellations.

- ## 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. Star Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and seeking a predictive model which will predict which booking is going to be cancelled in advance, which will help them formulating profitable policies for cancellation and refunds.

- ## Financial implications

  The cancellation of bookings impact a hotel on various fronts:
  1. Loss of resources (revenue) when the hotel cannot resell the room.
  2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
  3. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
  4. Human resources to make necessary arrangements for the guests.

- ## How to use ML model to solve the problem

  With a good robust predictive model where the model can most accurately predict the cancellations from the booking in place, help business with an idea of what will be the % of bookings going to be cancelled, so that they can make policies which can lower their cancellation rate, and also help them prepare well ahead to re-sell those rooms so that they can avoid revenue loss.

Data Overview

- Data Dictionary

Variables	Description
1. no_of_adults	Number of adults
2. no_of_children	Number of Children
3. no_of_weekend_nights	Number of weekend nights in booking
4. no_of_week_nights	Number of weeknights in booking
5. type_of_meal_plan	Meal plan booked by customer
6.required_car_parking_space	Customer required parking space or not
7. room_type_reserved	Room type booked by customer
8. lead_time	No of days between booking and arrival date
9. arrival_year	Year of arrival date
10. arrival_month	Month of arrival date
11. arrival_date	Date of arrival date
12. market_segment_type	Market segment designation of booking originated
13. repeated_guest	Customer repeated guest or not
14.no_of_previous_cancellation	No of previous booking cancelled by customer
15.no_of_previous_booking_not_cancelled	No of previous bookings by customers which are not cancelled
16. avg_price_per_room	Average price per day of the reservations
17. no_of_special_requests	Total no of special requests made
18. booking_status	Flag indicating the booking is cancelled or not

Observations	Variables
56926	18

Notes:

1. Dataset looked consistent with the dictionary provided.
2. There are ~ 25% duplicate values
3. There are no null/missing values across all variables.

Missing value counts	
no_of_adults	0
no_of_children	0
no_of_weekend_nights	0
no_of_week_nights	0
type_of_meal_plan	0
required_car_parking_space	0
room_type_reserved	0
lead_time	0
arrival_year	0
arrival_month	0
arrival_date	0
market_segment_type	0
repeated_guest	0
no_of_previous_cancellations	0
no_of_previous_bookings_not_canceled	0
avg_price_per_room	0
no_of_special_requests	0
booking_status	0

Duplicate value counts
14350

Data Overview contd..

- Brief description of significant manipulations made to raw data

1. Hidden missing value such as no booking nights, average room price being 0 for online bookings are treated with mode and median based on market segment type.

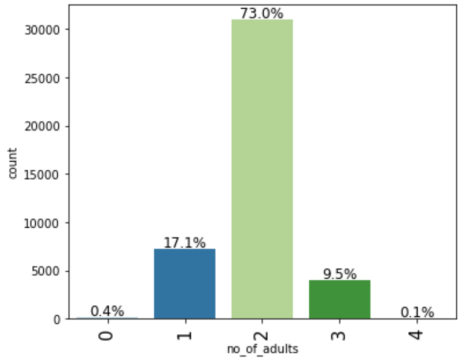
2. Outliers' treatment was done on the average room price per night, no other variables were treated as they are discreet in nature.

3. We have replaced the outliers in the average room price with their lower and upper whisker value as 3 depending on sides of their presence. Whisker as 3 was chosen to avoid wrongly updating complementary bookings as with 1.5 whisker those bookings were considered outliers.

4. We have extracted two new features stay_duration (no of total days in booking) and net_cancelled (if customer has higher no of previous bookings cancelled than not cancelled)

Exploratory Data Analysis – Univariate Analysis

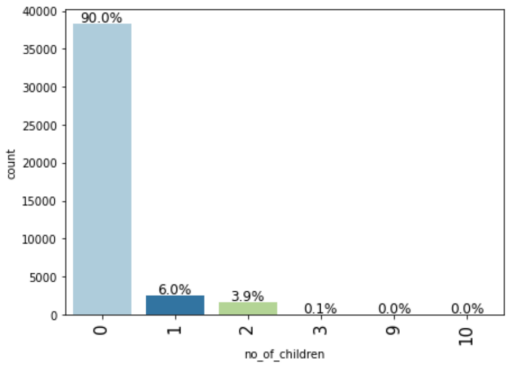
Proportion of adults per booking



Notes:

1. ~73% bookings are made for 2 adults and are most popular
2. Only ~.1% of bookings are made for 4 adults

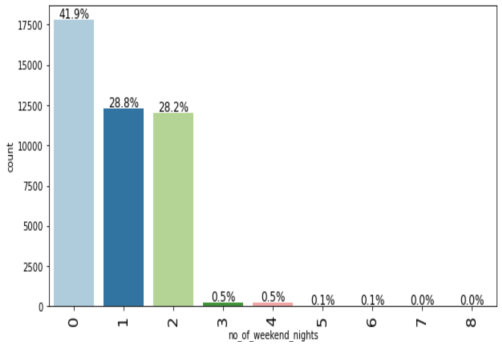
Proportion of children per booking



Notes:

1. 90% of bookings are not having any children occupancy
2. There are bookings with high numbers of children occupancy but in very less numbers with 9 children we have 2, and with 10 we have 1 booking

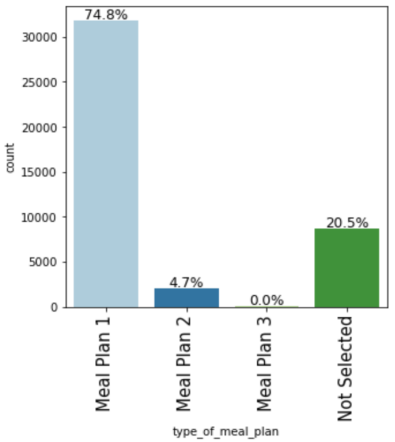
Proportion of weekend nights



Notes:

1. ~42% of bookings are not made for weekends
2. ~57% bookings are having either 1 or 2 weekend nights

Proportion of Meal plans

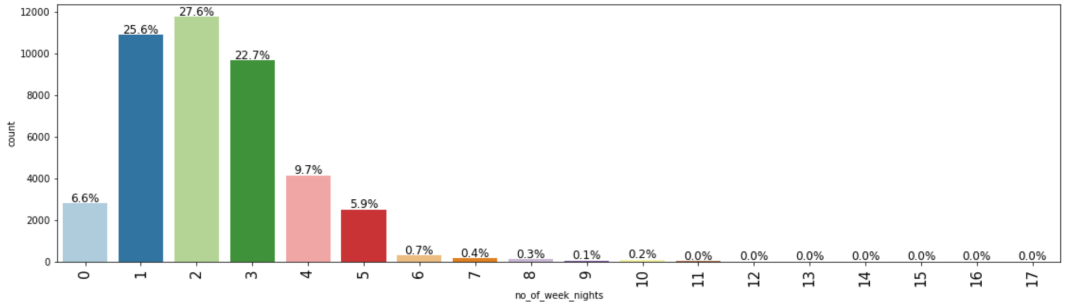


Notes:

1. ~75% bookings chose Breakfast, which is meal plan 1
2. ~20% bookings are made with no meal plan selected

Exploratory Data Analysis – Univariate Analysis

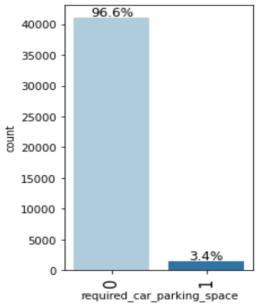
Proportion of weeknights per booking



Notes:

1. ~27% of bookings are made at least with 2 weeknights
2. Bookings are made for weeknights as high as 17

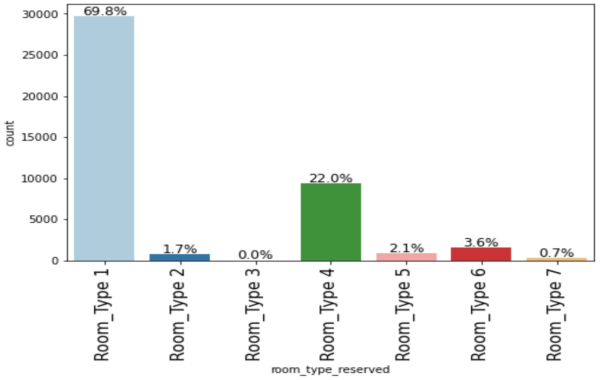
Required Car parking



Notes:

1. ~97% of bookings doesn't have any requirement for car parking.

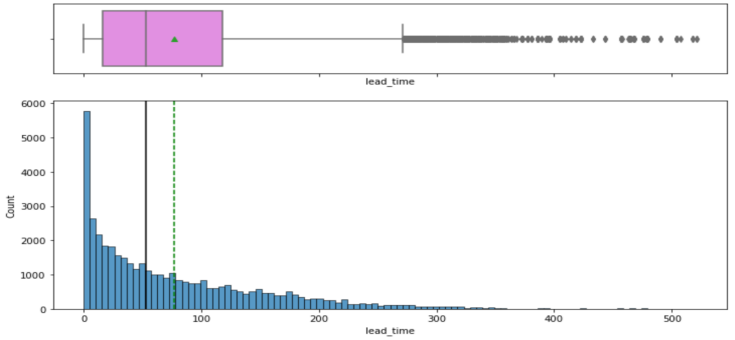
Proportions of room type



Notes:

1. ~70% bookings are for Room type1.
2. Room type4 is next popular room

Distribution of lead time for bookings

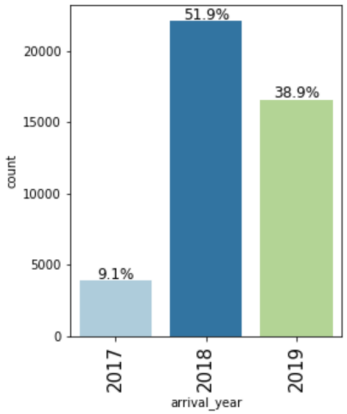


Notes:

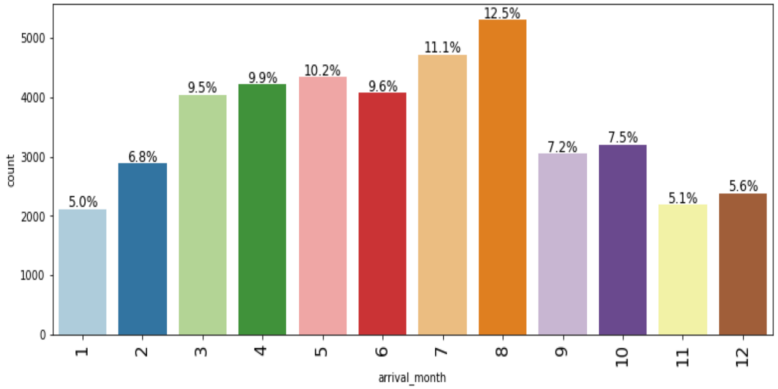
1. Majority bookings are made with 0 to 200 days lead time, with average ~70 days
2. Bookings are with lead time as high as ~500

Exploratory Data Analysis – Univariate Analysis

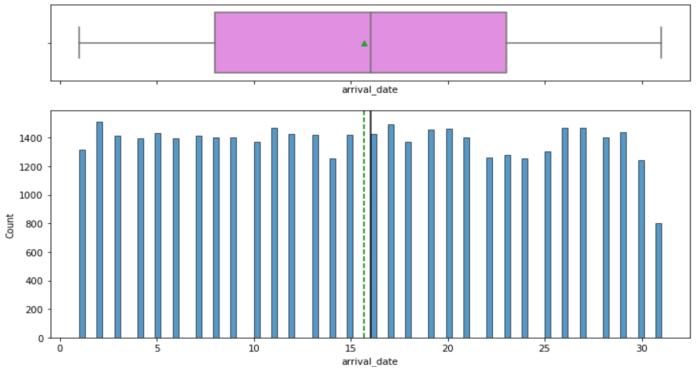
Arrival Year



Arrival Month



Arrival Date

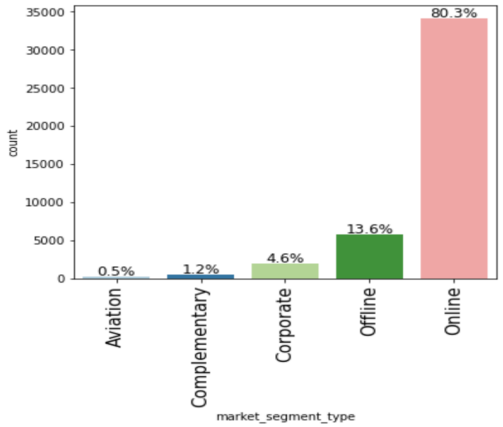


Notes:

1. Bookings are from 2017 to 2019 which are present in dataset. With 2018 sees a peak in demand, then 2019 demand has fallen
2. August is the busiest month, with ~70% of bookings are for March to August
3. Bookings are consistent pretty much for every day of the month

Exploratory Data Analysis – Univariate Analysis

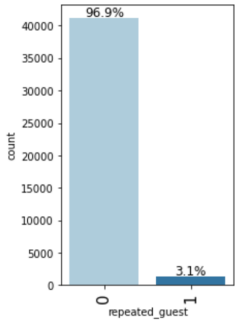
Market Segment proportions



Notes:

1. ~80% bookings are made online
2. Aviation segment shows the lowest booking contribution

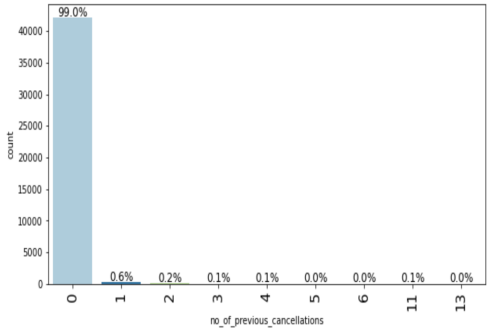
Repeated guest proportions



Notes:

1. ~97% are not repeated guests

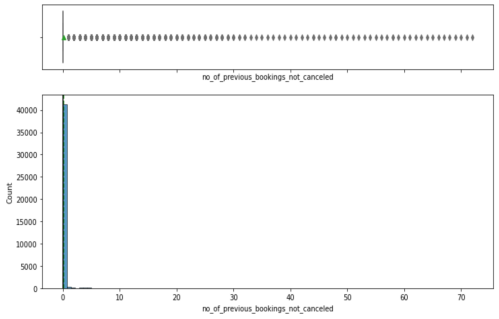
Previous Cancellations



Notes:

1. ~99% of guests doesn't show any previous cancellation
2. Previous cancellation for guests are as high as 13

Previous Not Cancelled Bookings

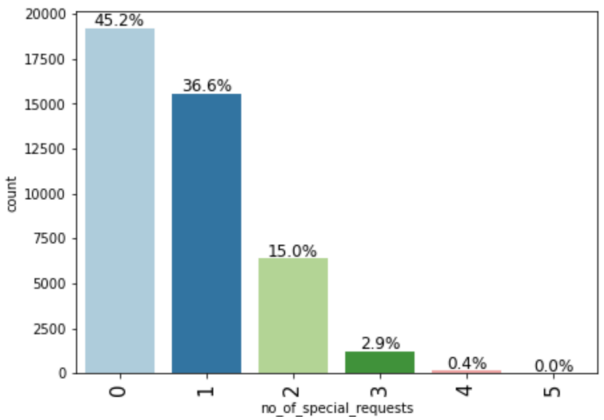


Notes:

1. Majority guests shows no previous booking cancelled, which make sense as most are non repeated guests
2. Previous not cancelled booking shows as high as ~70

Exploratory Data Analysis – Univariate Analysis

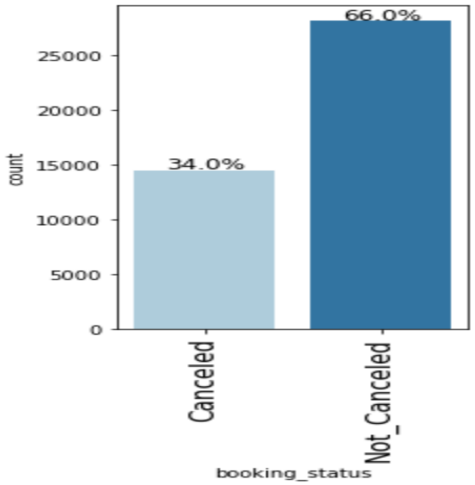
Proportions of special requests



Notes:

1. ~45% bookings are made no special requests
2. ~36% of bookings having 1 special request
3. There are up to 5 special requests made for a booking

Proportions of special requests

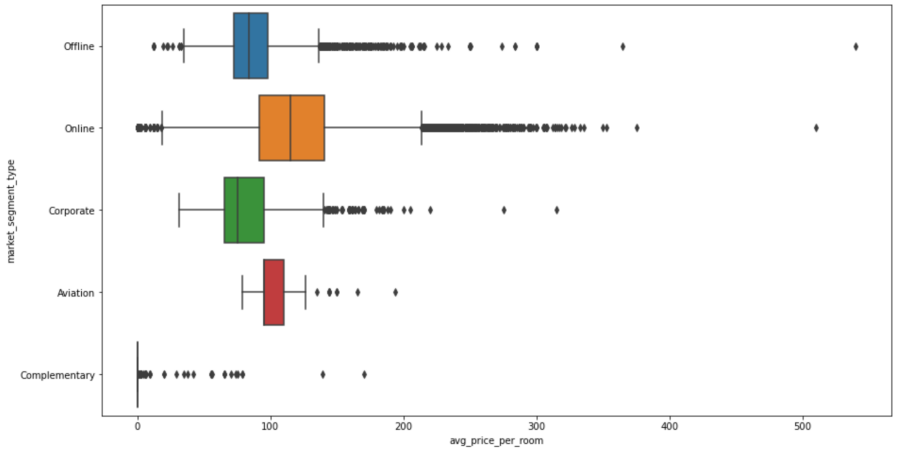


Notes:

1. 34% of bookings in the dataset are Cancelled
2. 66% of bookings are not cancelled

Exploratory Data Analysis – Bivariate Analysis

Average room price vs Market segment

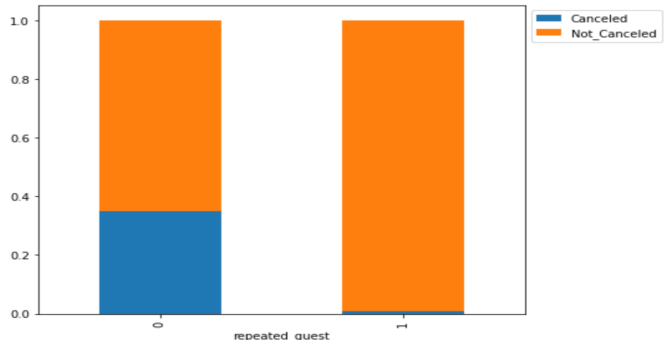


Notes:

1. Average room price vary based on market segment
2. Online has the highest mean average room price.
3. Corporate gets the lowest average room price, other than complementary where room price is always 0.

Repeated Guest vs Booking Status

booking_status	Canceled	Not_Canceled	All
repeated_guest			
All	14487	28089	42576
0	14477	26784	41261
1	10	1305	1315



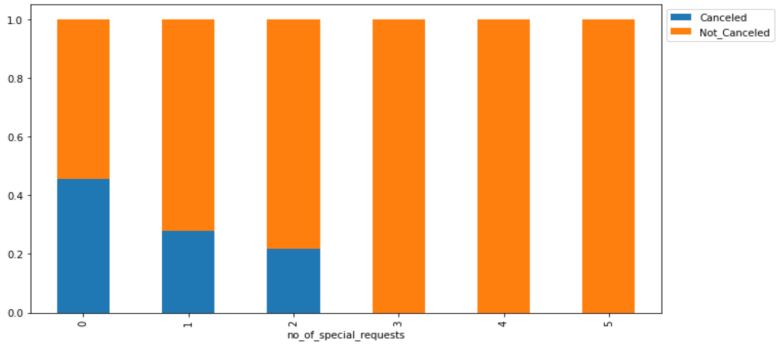
Notes:

1. Cancellation among repeated guests (denoted as 1) are very less
2. ~35% of bookings are cancelled among non repeated guests

Exploratory Data Analysis – Bivariate Analysis

Special Requests vs Booking Status

no_of_special_requests	Canceled	Not_Canceled	All
All	14487	28089	42576
0	8752	10476	19228
1	4346	11225	15571
2	1389	4992	6381
3	0	1230	1230
4	0	150	150
5	0	16	16

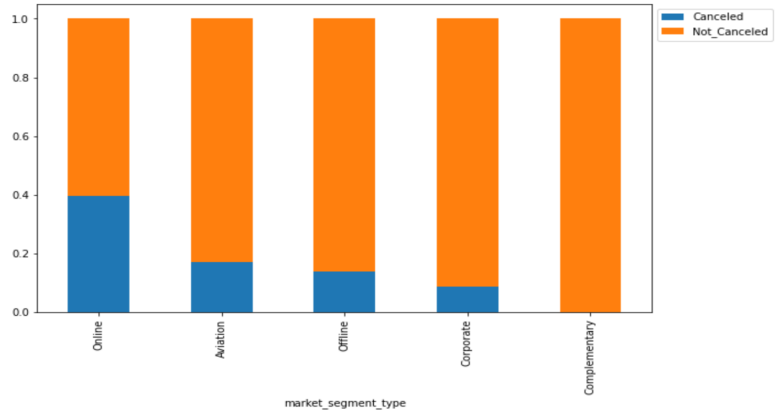


Notes:

1. Bookings with no special requests have the highest cancellation rate, close to ~50%
2. With a greater number of special requests (>3) the bookings are not likely to be cancelled

Market Segment vs Booking Status

booking_status	Canceled	Not_Canceled	All
market_segment_type			
All	14487	28089	42576
Online	13483	20686	34169
Offline	804	4973	5777
Corporate	167	1772	1939
Aviation	33	162	195
Complementary	0	496	496

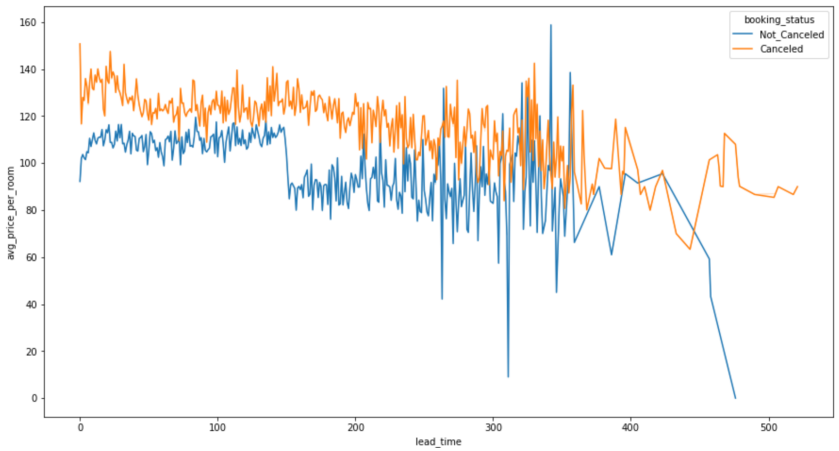


Notes:

1. In Online market segment we can see the highest cancellation rate.
2. Complementary bookings are not likely to be cancelled.

Exploratory Data Analysis – Bivariate Analysis

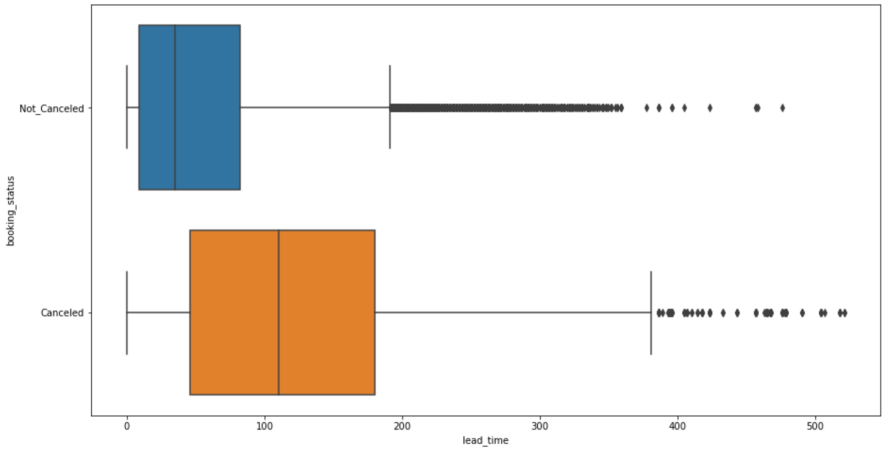
Lead Time vs Average room price



Notes:

1. Room price being dynamic it varies across bookings; however, we see with lesser than ~150 days lead time the average room price is higher
2. Cancelled bookings does have higher average room price than bookings which are not cancelled.

Lead Time vs Booking Status

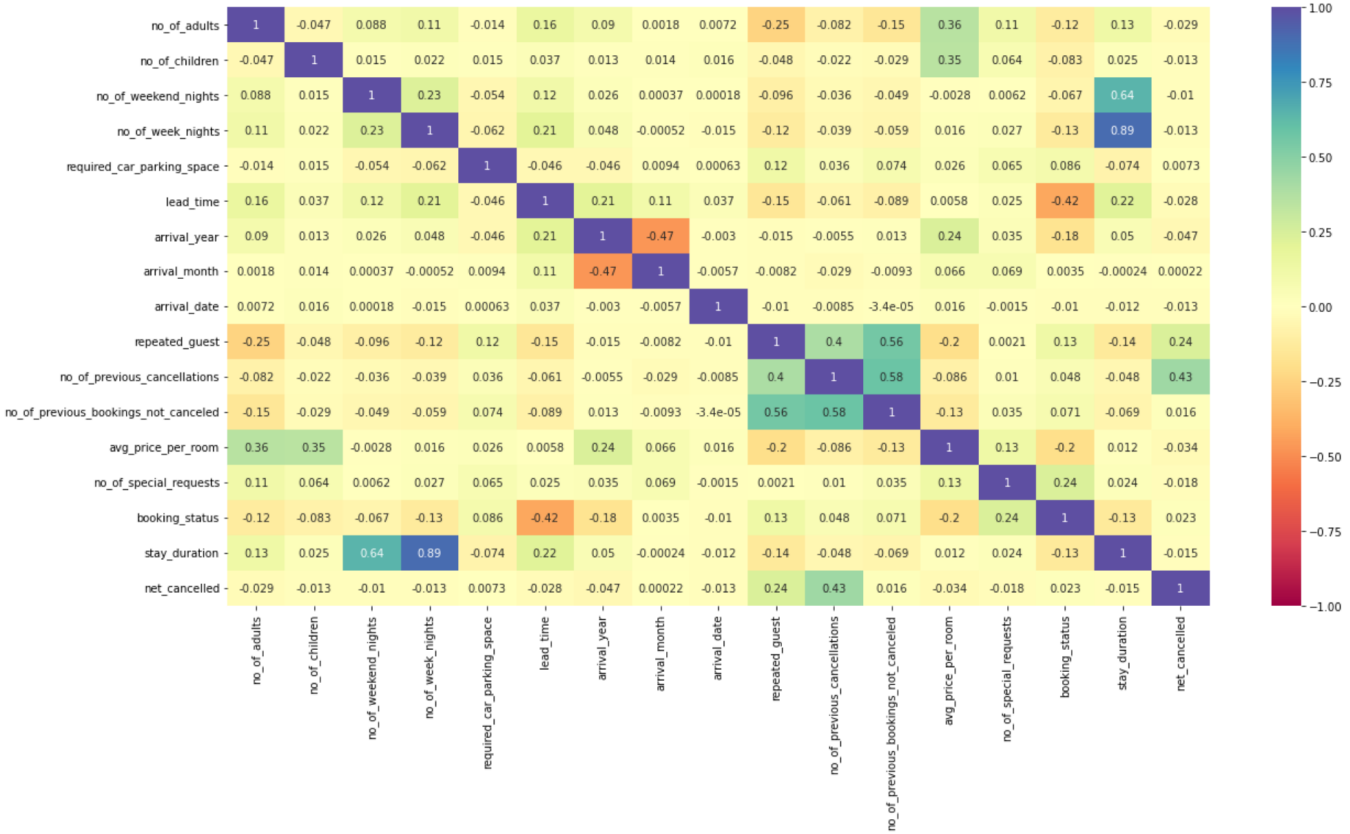


Notes:

1. Majority of bookings which are cancelled has higher average lead time.

Exploratory Data Analysis – Bivariate Analysis

Heatmap of correlation matrix



Notes:

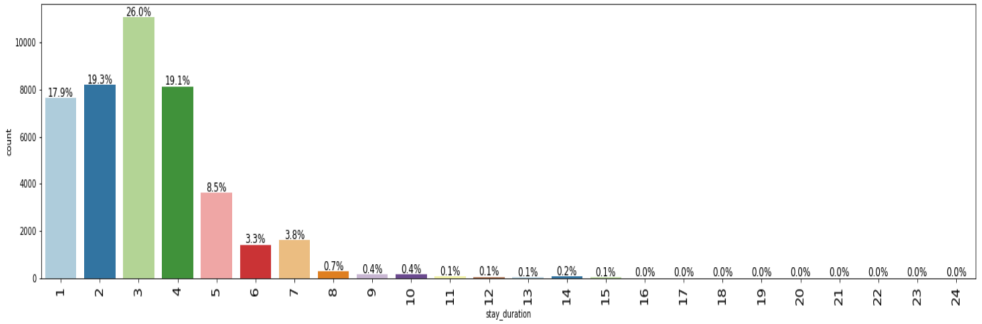
1. No of weekend nights and no of weeknights shows high correlation with the newly derived feature, which make sense since it's derived from them. We will keep it for model creation and then drop based on relevance.

2. None of the features shows high correlation with booking status, except lead_time shows average correlation with booking status

3. no of previous booking not canceled shows average positive correlation with repeated guests and no of previous cancellation, .56, .58 respectively

Exploratory Data Analysis – New Features

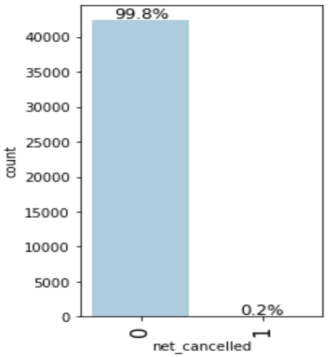
Proportions of Stay Duration



Notes:

1. Guests have made bookings ranges from 1 to 24 days.
2. ~80% of bookings are for 1 to 4 days of stays.
3. 3 days stay is the most popular booking.

Proportions of Net cancelled



Notes:

1. We can observe only ~.2% of guests have prior cancellations which are more than the prior bookings which are not cancelled

Model Performance Summary – Logistic Regression

Approach

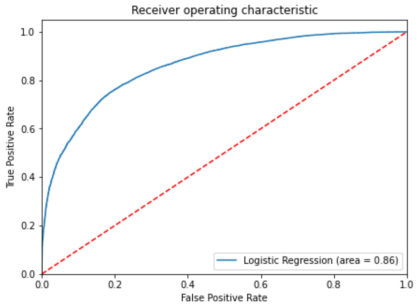
- Explored statsmodel Logistic regression and DecisionTreeClassifier for creating models for the solution.
- 30% of randomly chosen data are kept for testing model performance.
- Observation count for model training 29803, and testing 12773, with proportions of target variable similar to original dataset
- Model performance was checked against Recall, as for this problem we are trying to minimize our False negatives, which means identifying a booking will not be cancelled where in reality it would be cancelled

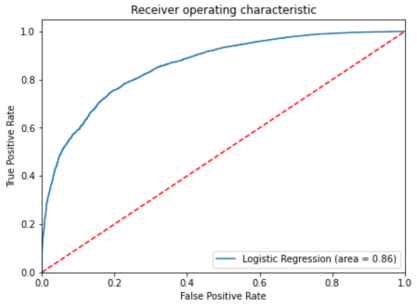
Model and Parameters

- Initial model with Logistic regression from statsmodel were prepared with all variables.
- This model is used for further tuning and dropping features based on significance using VIF and p-value.
- This model was further explored with ROC AUC curve, Precision Recall curve to explore the threshold tuning and if model recall score can be improved.

Parameters	Coef	P-value
no_of_children	-0.1142	0.016
no_of_weekend_nights	-0.0551	0.002
no_of_week_nights	-0.0780	0.000
required_car_parking_space	1.3366	0.000
lead_time	-0.0166	0.000
arrival_month	0.0381	0.000
repeated_guest	2.9020	0.000
no_of_previous_cancellations	-0.2181	0.022
avg_price_per_room	-0.0169	0.000
no_of_special_requests	1.3122	0.000
type_of_meal_plan_Meal Plan 2	0.1847	0.019
type_of_meal_plan_Not Selected	-0.3117	0.000
room_type_reserved_Room_Type 2	0.2982	0.019
room_type_reserved_Room_Type 4	0.1529	0.000
room_type_reserved_Room_Type 5	0.2548	0.023
room_type_reserved_Room_Type 6	0.5618	0.000
room_type_reserved_Room_Type 7	0.8363	0.000
market_segment_type_Corporate	-1.5812	0.000
market_segment_type_Online	-2.2058	0.000

- All the features having p-value under 0.05(significance level), So all these features are significant for deriving price for used phones
- Positive coefficients denotes increase the chance of booking being cancelled with the increase in value for the corresponding attributes, and negative coefficients denotes decrease the chance of booking being cancelled with increase in value of the corresponding attributes





Performance of the Models

Training performance comparison:

	Logistic Regression default Threshold	Logistic Regression-0.69 Threshold	Logistic Regression-0.58 Threshold
Accuracy	0.792974	0.770761	0.791632
Recall	0.884123	0.748221	0.838382
Precision	0.817251	0.886867	0.844822
F1	0.849373	0.811666	0.841590

Test set performance comparison:

	Logistic Regression default Threshold	Logistic Regression-0.69 Threshold	Logistic Regression-0.58 Threshold
Accuracy	0.789869	0.767557	0.789243
Recall	0.880661	0.743611	0.836562
Precision	0.815161	0.885116	0.842370
F1	0.846646	0.808217	0.839456

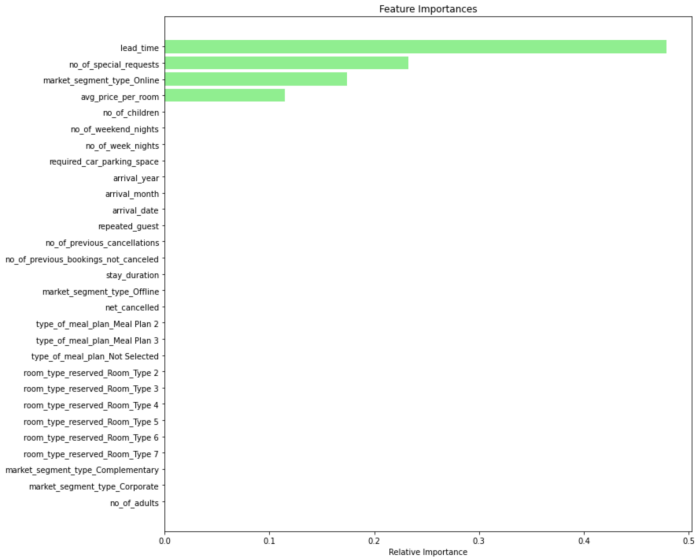
Notes:

- We can see model with default threshold gives us the best recall which is our key metric to evaluate performance based on our business problem
- With ~88% recall value we can say that our model is correctly predicting if a booking will be cancelled 88% of the time.

Model Performance Summary – Decision Tree

Model and Parameters

- Initial model was prepared with default parameters and using DecisionTreeClassifier.
- This model is used for further tuning using GridSearchCV for pre pruning the tree as it was overfitted.
- Pre pruned decision tree performed really well and in generealized manner hence post pruning was not explored.
- Best parameters for pre pruned decision tree is
  criterion='entropy',
  max_depth=5,
  min_impurity_decrease=0.01



- lead time is the most important feature for predicting cancellation.
- no of special requests, market segment online, average room price are some features which are important for the pruned decision tree

Performance of the Models

Models	Data used	Recall value
Decision Tree – Default	Train	1.00000
	Test	0.827053
Decision Tree – Pre pruned	Train	0.99756
	Test	0.995008

Notes:

- We can see model with best parameters identified by GridSearchCV gave us the best recall value.
- Pre pruned Decision tree model gave us similar performance in both train and test set, so we can consider it's not overfitted and considered as final.

Conclusion on Models

- Both Logistic Regression from statsmodel and Pre pruned decision tree models are performing in generalized way with high recall value. However, pre pruned decision tree model gives us the best performance among them.
- Hence, to address the business problem in hand we recommend to use the pre pruned decision tree model.

Business Insights and Recommendations

Insights

- Online bookings have highest cancellation rate, might be a cancellation fee after a fee free time period should be implemented to target to lower the cancellation.
- Cancellation fee should be prorated based on no of lead days remain before arrival date at the time of cancellation
- Analyzing what kind of special requests are made and offering some complementary with the booking could lower the cancellation counts on the bookings no special requests made.
- For repeated guests' cancellation is lower, however offering some loyalty program could further reduce the cancellation
- Machine Learning model should be used to predict the cancellation and based on capacity and no of predicted cancellation rooms can be sold with enough lead time so that discounted price can be avoided.

Recommendations

- Capturing cancellation reason and analyzing further can lead to other findings which could be useful for making policies around cancellation.
- Online has the majority booking rate, capturing geo location and analyzing the pattern of cancellation could be further useful.
- Providing discounted rate, or bonus stays, should be explored for the time period other than March to August.
- Using the Machine Learning model will be beneficial to predict cancellations, however the rooms should be sold based on capacity and no of predicted cancellation.
- Over a time period the data should be collected, and overbooked data should be further analyzed to find the correct ratio of overbooking.

greatlearning
Power Ahead

Happy Learning !