Capstone Project

Hotel Booking Cancellation Prediction.

Overview

1. Problem Definition
2. Data Overview
3. EDA
4. Model summary
5. Model performance table
6. limitations
7. business Recommendations

greatlearning

Problem Definition

- A Hotel Group chain is facing problems with the high number of booking cancellations.
- We 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 cancelled in advance and help in formulating profitable policies for cancellations and refunds.

Data Overview

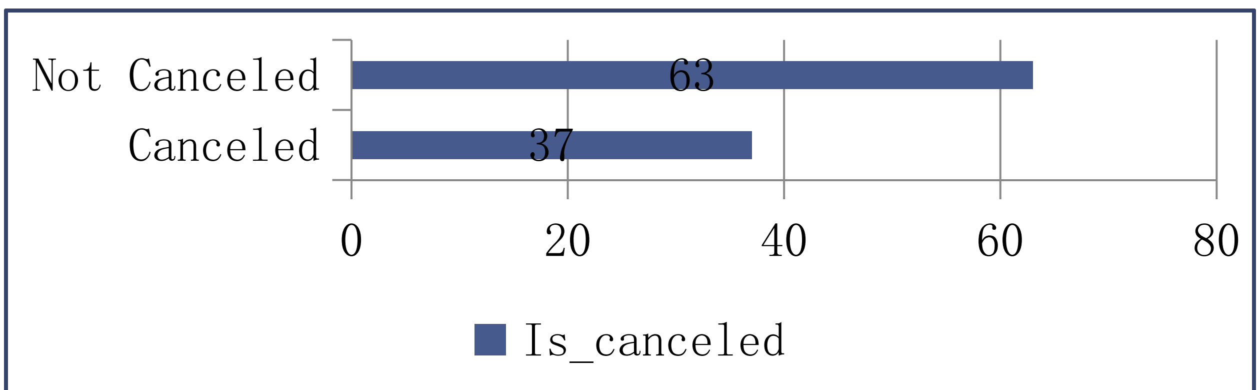
- The dataset consists of **119390 Rows** and **32 Columns.**
- The data is spread through a time period of **3 years** with customer arrival year ranging between **2015** and **2017.**
- Each row of the data set represent a booking instance created by a customer and all the related details.
- Our Target Variable is **'is_canceled'**
- Classification of Data Types of the Attributed present inside the dataset.

greatlearning

Data Overview

Column Data Type	*Present Data Types*	*Actual Data Types*
Categorical	12	19
Numerical	20	13

Data Balance/ Imbalance w.r.t Target Variable



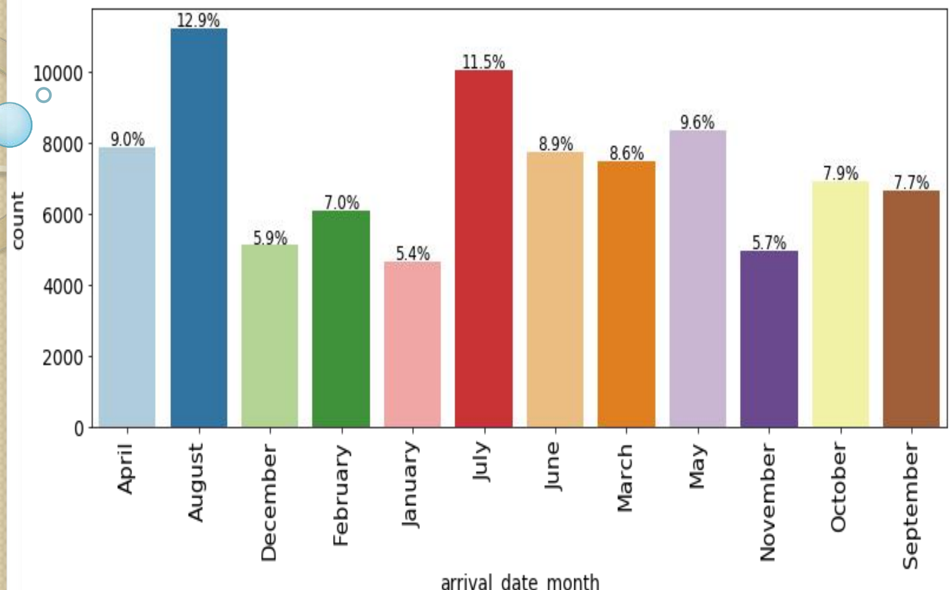
Data Overview

Existing Data Types and Null Values

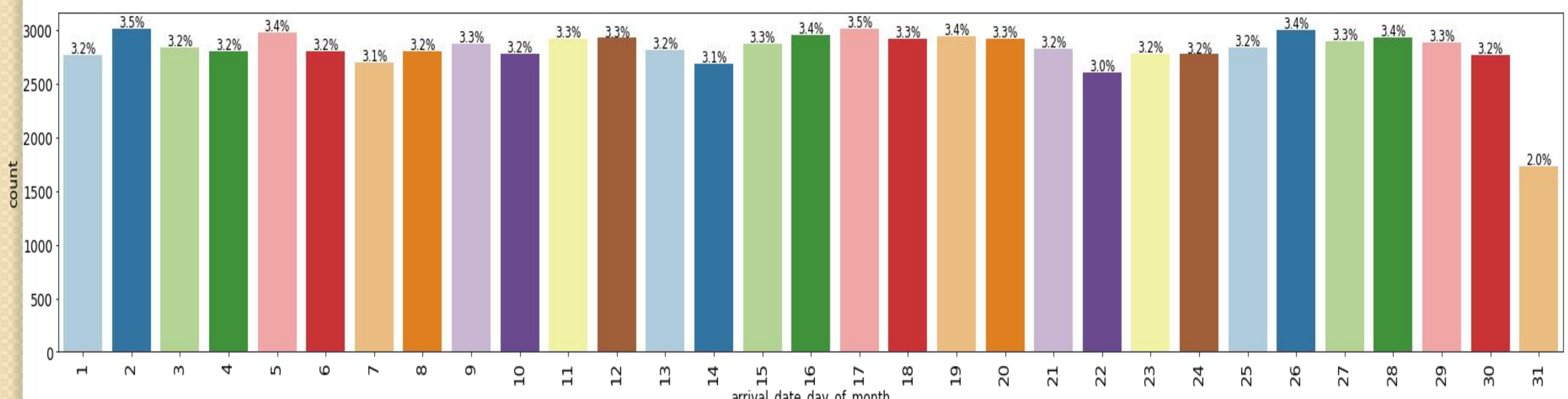
Attribute Name	*Null Values*	*Percentage Null Values*
Company	112593	94.31
Agent	16340	13.69
Country	488	0.41
Children	4	0

Number of Duplicate Rows: 31994

EDA

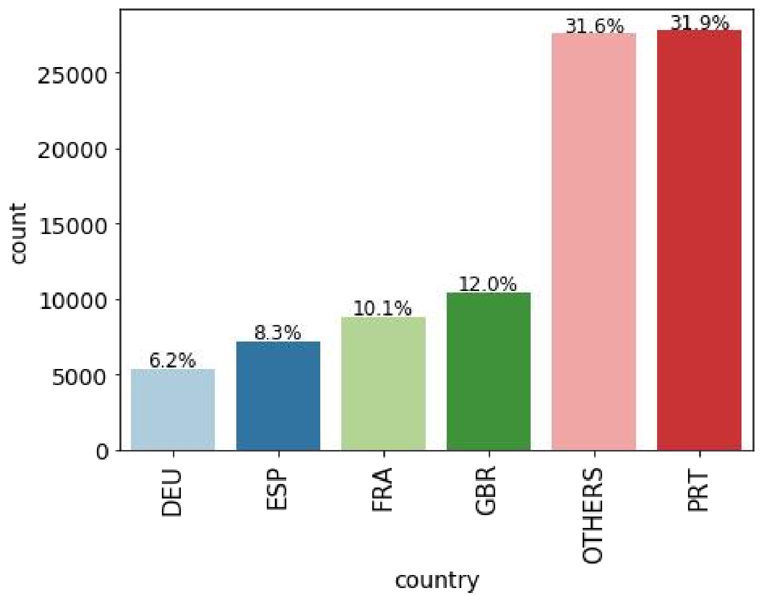


1. More number of people prefer to visit during July and August.
2. November, December and January seems to be the off season where around 5% people prefer to visit.

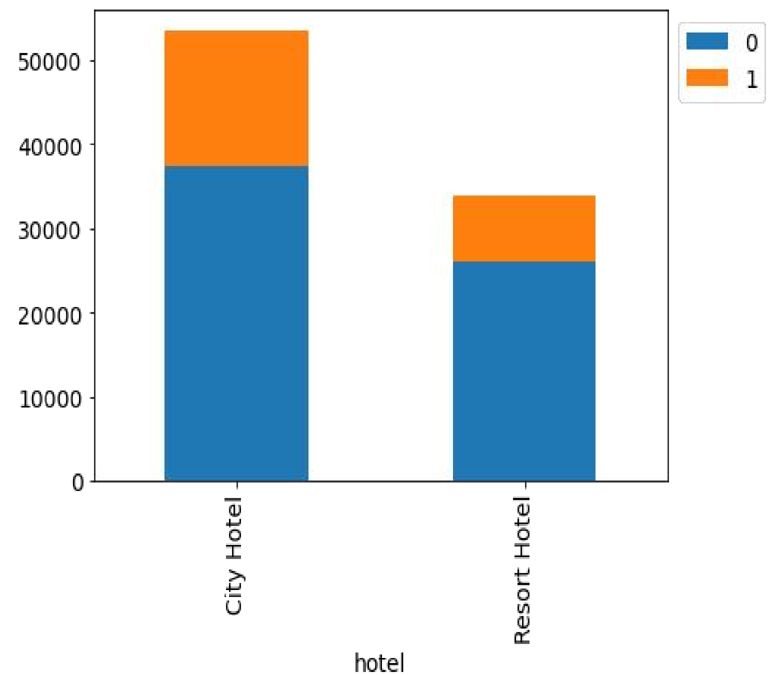


1. Slightly less percentage of people arrived to the hotel on 31st.

EDA

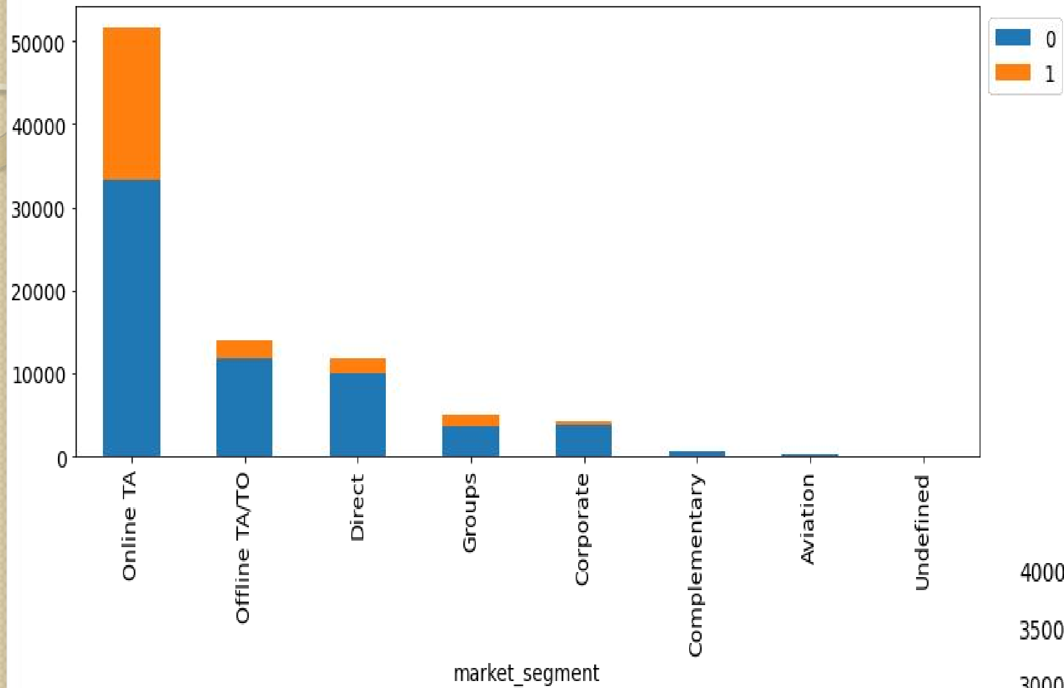


Most of the tourist came from Portugal,France, Germany, Spain and Great Britan.

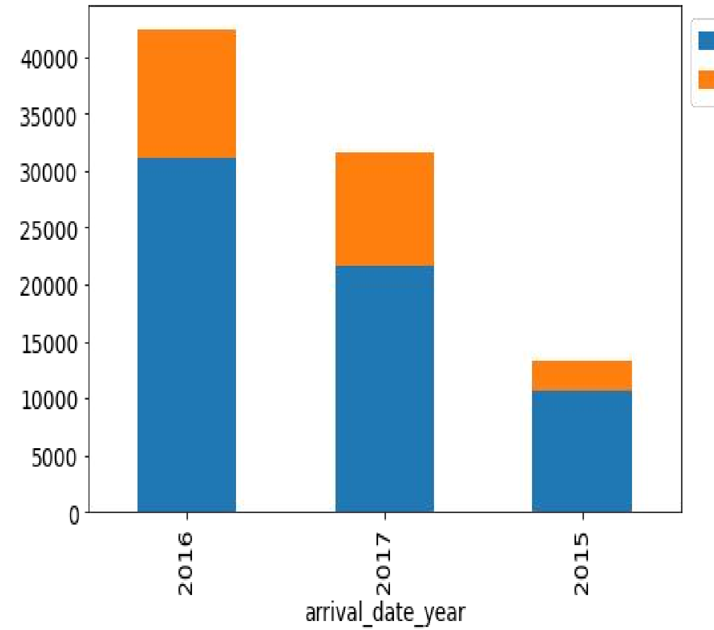


City Hotel has more No. of cancellation compare to Resort Hotel

EDA

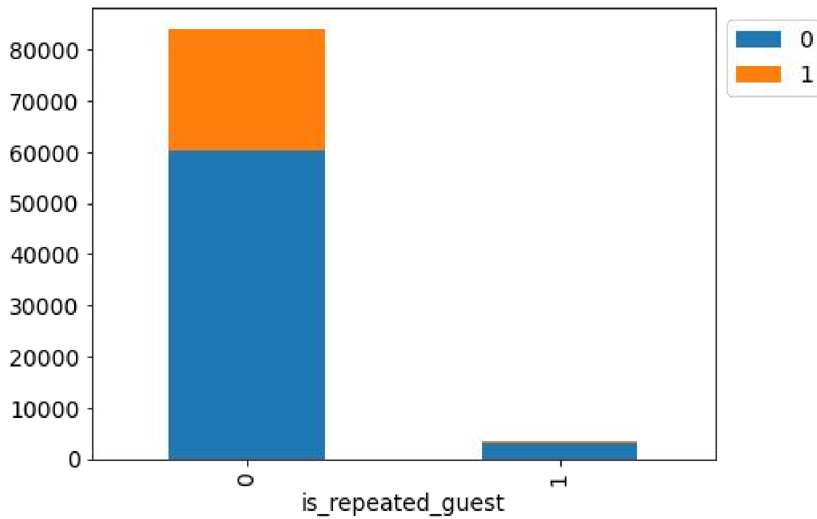


Online TA is most preferred market segment for both City Hotel and Resort Hotel



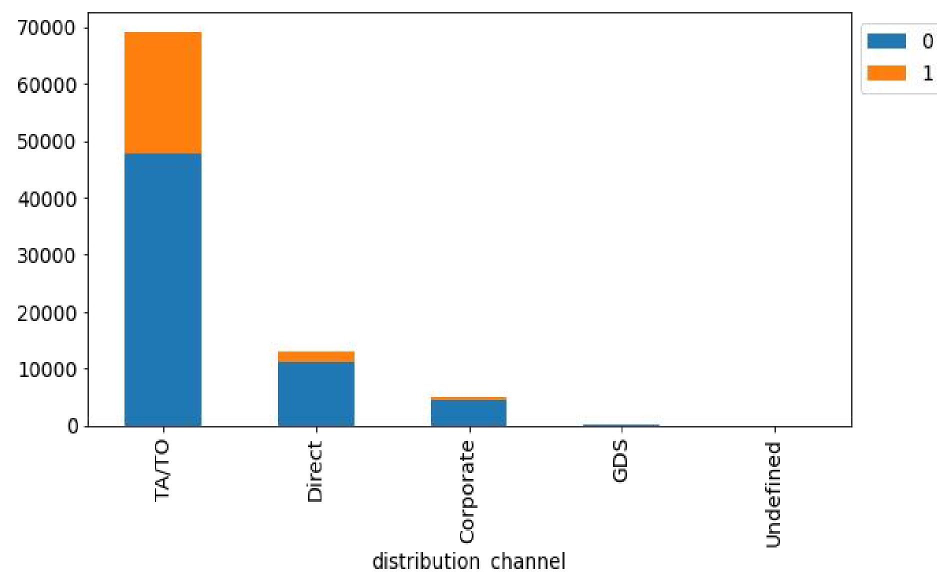
2016 is the most arrival date year in comparison with 2017 and 2015

EDA



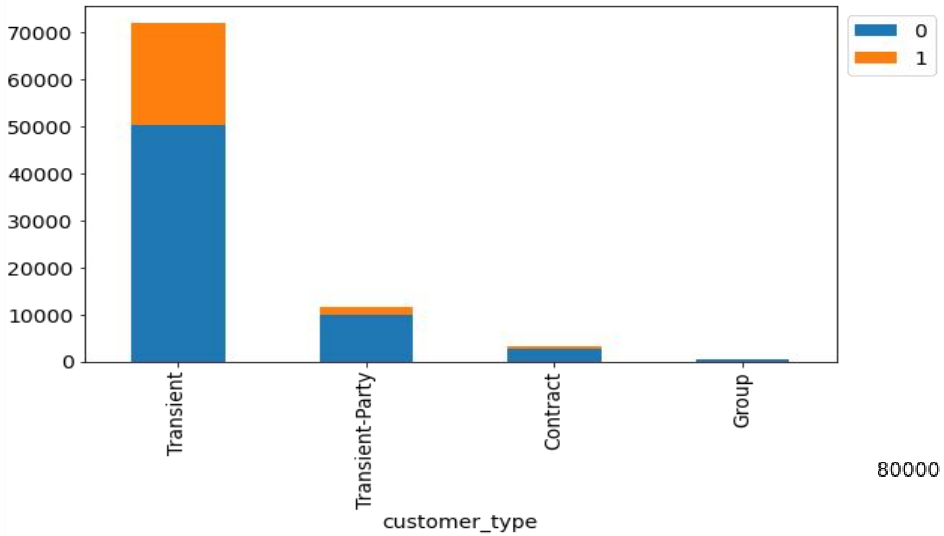
No. of repeated guest are very few in both City Hotel and Resort Hotel

TA / TO is most preferred Distribution channel

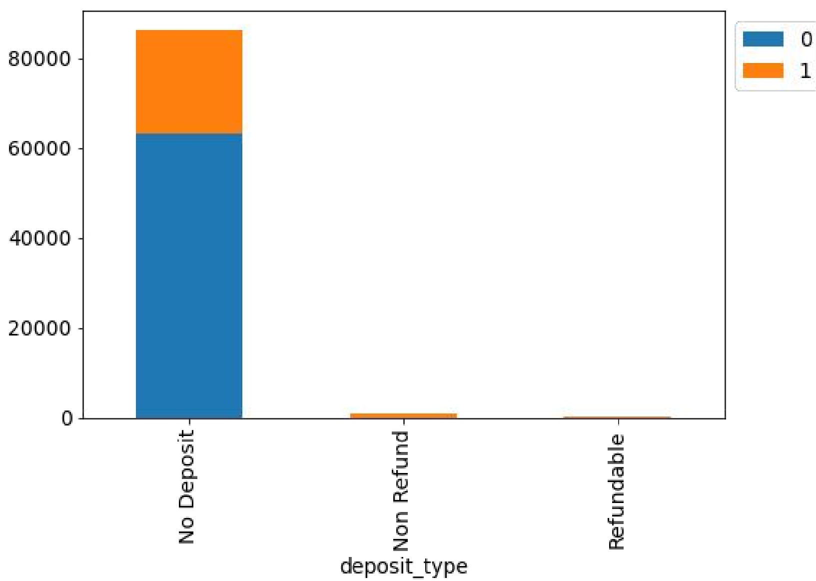


greatlearning

EDA

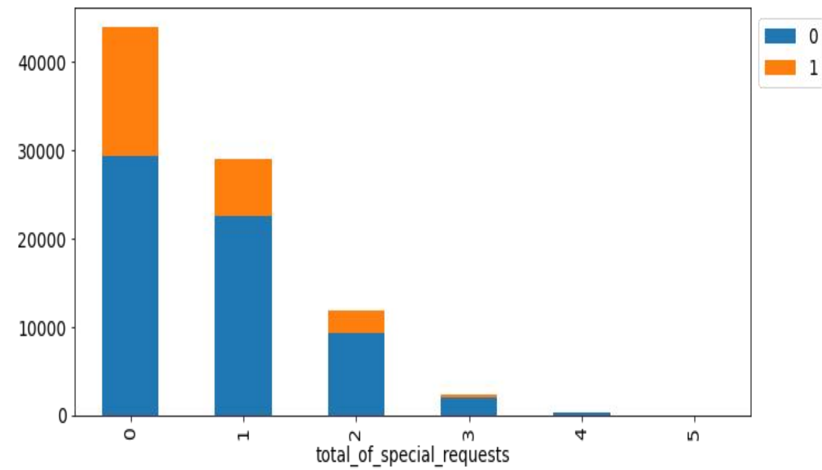


Cancellations in case of Transient bookings are higher



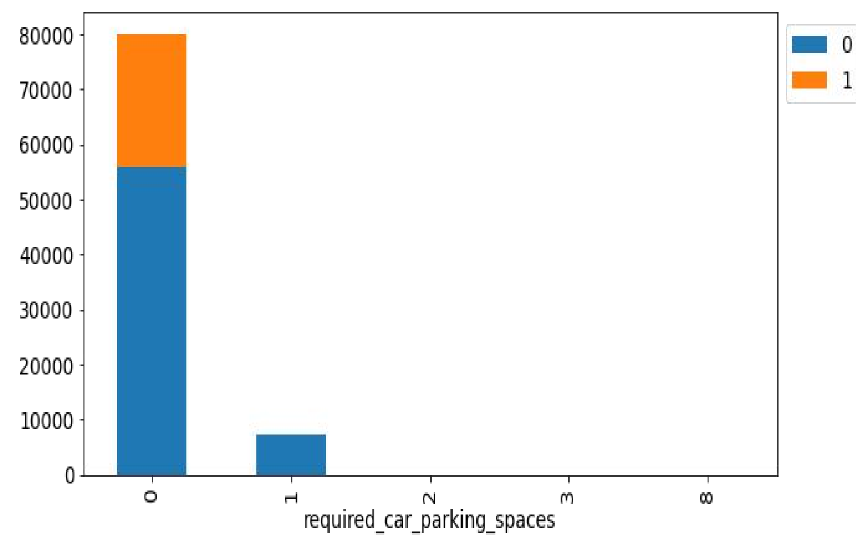
Most people have done no deposit booking and and max cancelled the booking are from it as well.

EDA

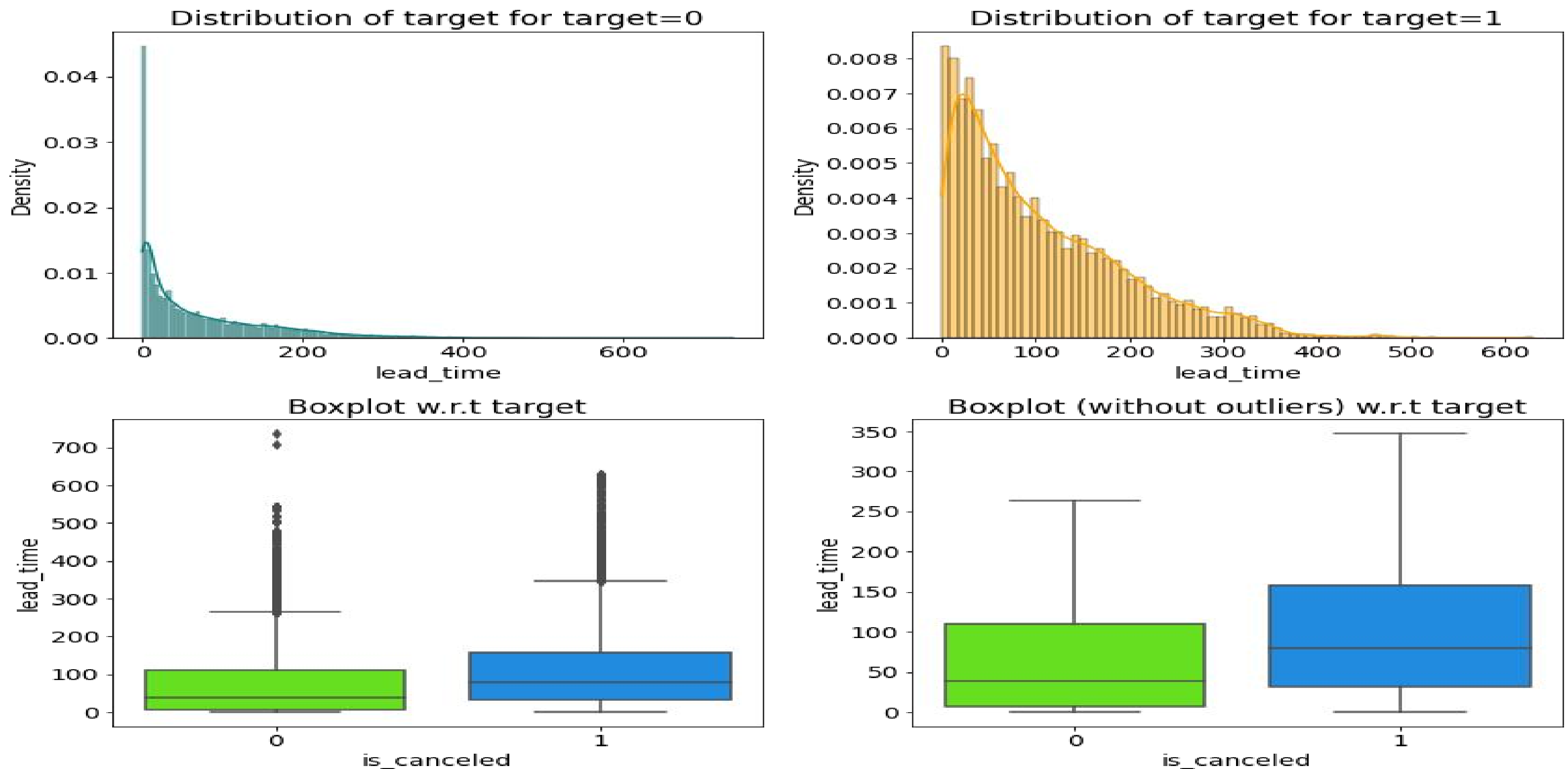


1. As number of special requests rises chances of cancellation drops
2. Most of the bookings are without any special request

1. Cancellations are only in the case where car parking space is not required

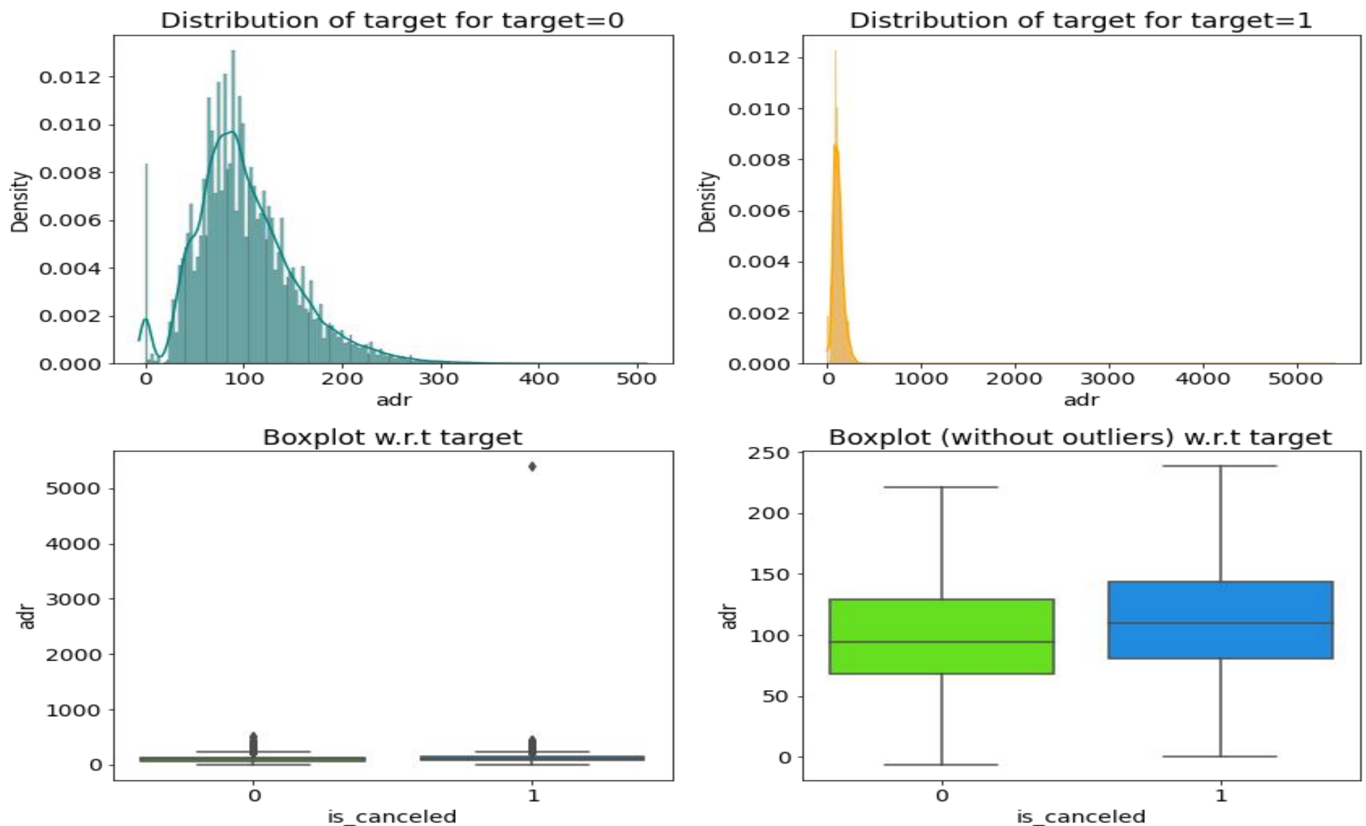


EDA



1. Distribution of both cancelled and no-cancelled are highly skewed to the right.
2. Median of lead_time in case of cancelled bookings is higher than not-cancelled bookings.
3. Most of the bookings in both cases has zero lead time which seems to be the case of same day arrival/cancellation at the hotel

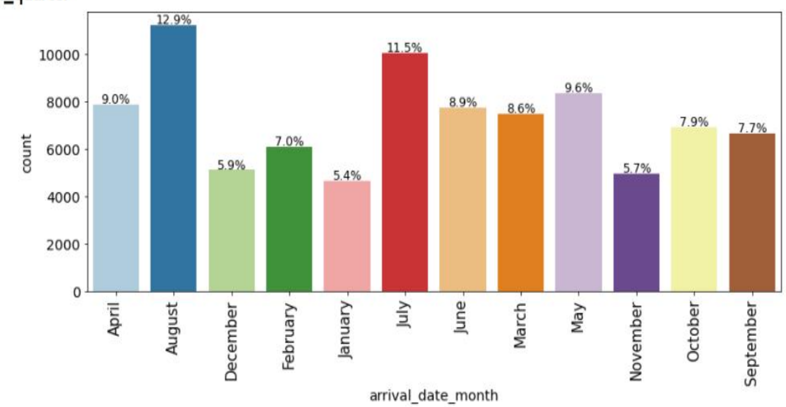
EDA



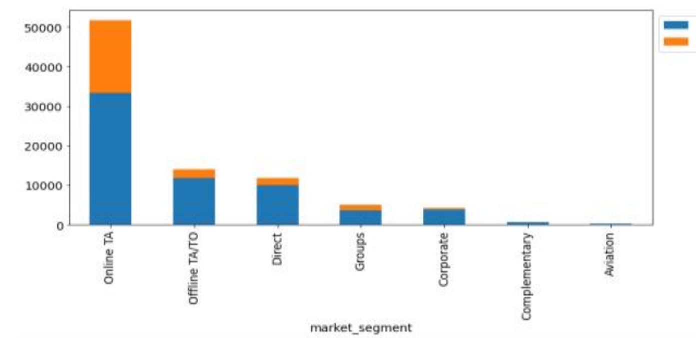
1. Median of Average Daily rate for cancelled booking is comparitively higher than median for not-cancelled booking.
2. Presence of large number of outliers in the average daily rate.
3. Average Daily rate is zero for some cases. It might be a case of complimentary bookings

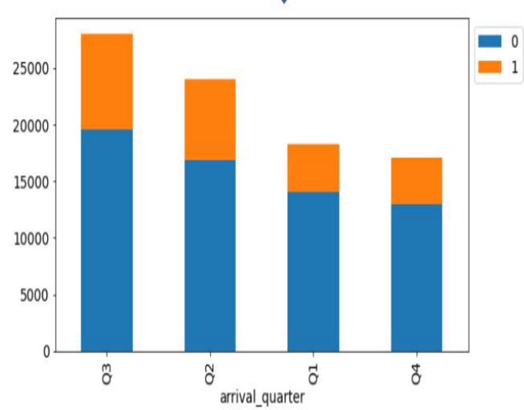
Derive new Features_

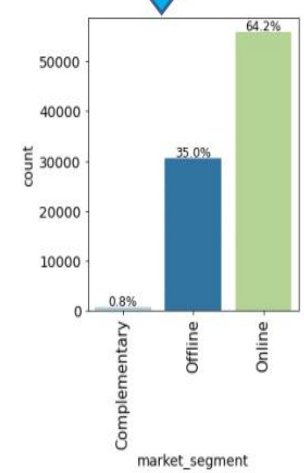
arrival_quarter



market_segment

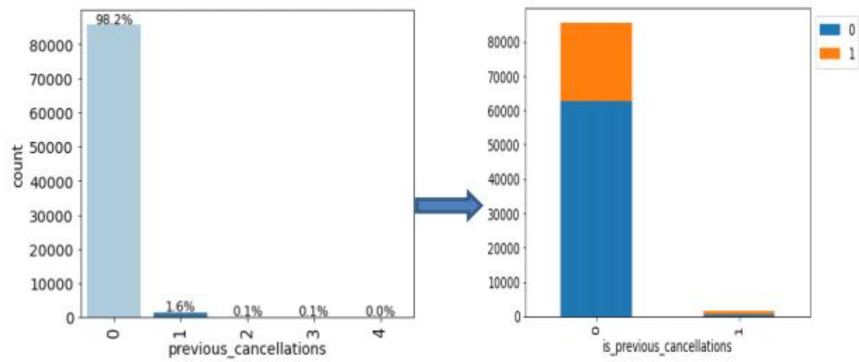




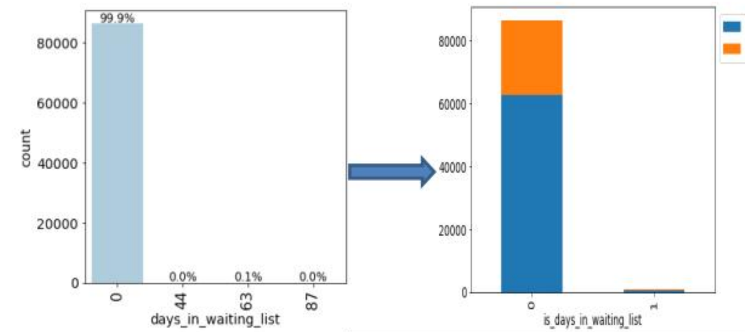


Derive new Features_

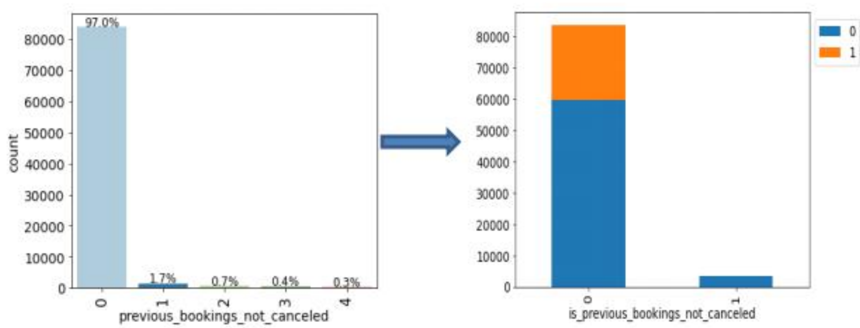
is_previous_cancellations



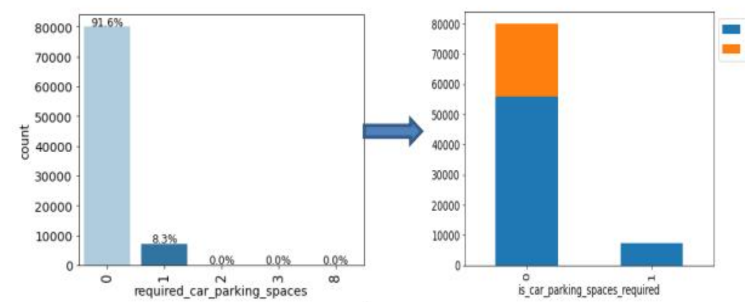
Is_days_in_waiting_list



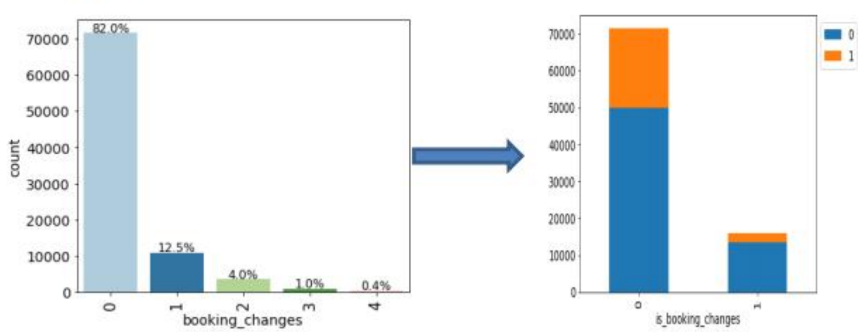
is_previous_bookings not cancelled



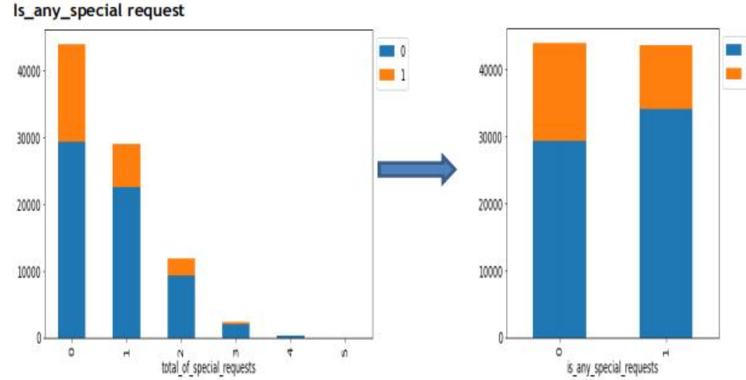
Is_car_parking_space_required



is_booking Changes

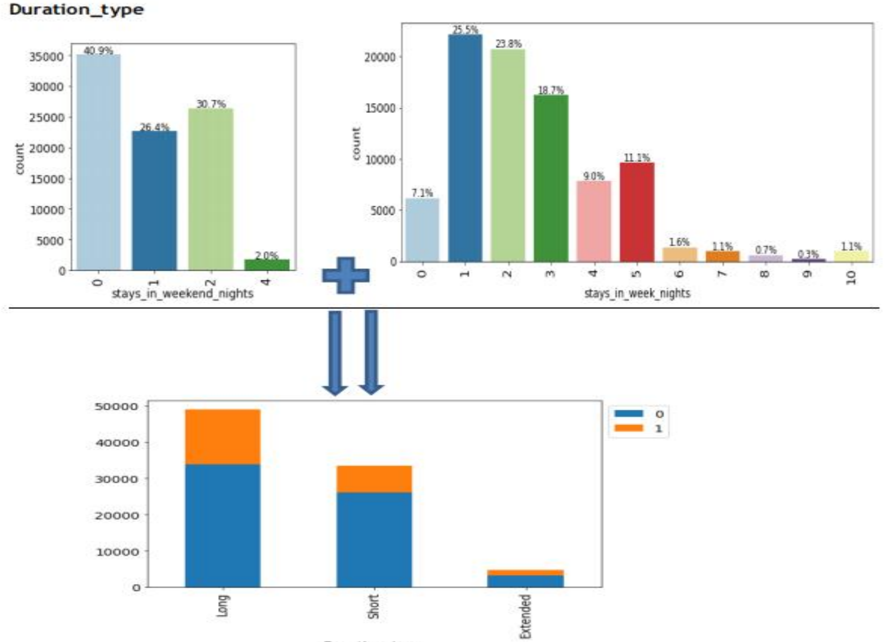


Is_any_special request



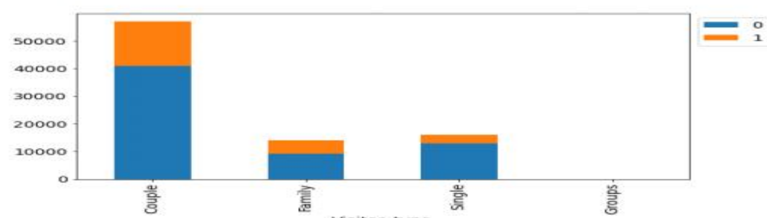
Derive new Features_

Duration_type



- Total_nights = stays_in_weekend_nights + stays_in_week_nights
- Duration_type :
    1. Short: Total_nights <3
    2. Long: Total_nights >=3 and <=7
    3. Extended: Total_nights >7

Visitor_type



Model comparison table Before SMOTE

Model		Accuracy	Precision	Recall	F1 Score	ROC AUC	COHEN'S KAPPA
Logistic Regression	Train	0.73	0.76	0.01	0.01	0.50	0.0089
	Test	0.727	0.77	0.01	0.01		
Decision Tree	Train	0.84	0.73	0.65	0.69	0.75	0.51
	Test	0.81	0.68	0.61	0.64		
Random Forest	Train	0.90	0.86	0.75	0.80	0.765	0.56
	Test	0.84	0.75	0.61	0.67		
KNN	Train	0.83	0.64	0.91	0.75	0.726	0.46
	Test	0.74	0.63	0.58	0.61		
LGBM Classifier	Train	0.84	0.76	0.63	0.69	0.77	0.56
	Test	0.84	0.74	0.62	0.68		

Inferences:

- From above metrics we can clearly see that LGBM classifier and Random Forest Classifier is giving out the best results
- ROC AUC is similar between them

Model comparison table After SMOTE

Model		Accuracy	Precision	Recall	F1 Score	ROC AUC	COHEN'S KAPPA
Logistic Regression	Train	0.76	0.75	0.79	0.77	0.734	0.417
	Test	0.74	0.52	0.70	0.60		
Decision Tree	Train	0.85	0.82	0.88	0.85	0.72	0.46
	Test	0.78	0.58	0.68	0.63		
Random Forest	Train	0.91	0.86	0.96	0.91	0.80	0.557
	Test	0.81	0.61	0.80	0.69		
KNN	Train	0.88	0.83	0.95	0.89	0.748	0.432
	Test	0.74	0.52	0.77	0.62		
LGBM Classifier	Train	0.85	0.82	0.88	0.85	0.80	0.555
	Test	0.81	0.61	0.80	0.69		

Results have improved after applying smote since there's more data for the machine to learn pattern for cancellations.

Hyperparameter Tuning:

For tuning the hyperparameter of various models we used GridSearchCV.

Below is the best parameter for Random Forest Classifier which is so far the best model:

'criterion': 'gini', 'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200

Inference:Random forest is the best model

Model Performance Summary

- The Random forest and Decision Tree models indicates that the most significant predictors of booking status are:

1. Lead Time
2. market_segment_online
3. Number of special requests

Confusion matrix

Confusion matrix of Random Forest Classifier(Best Model)

Limitations

- While working on the dataset, we noted few limitations which impacted the expectation of our result
- • There were evidences of minority class been already up sampled/synthetically sampled.
- Possible use of SMOTE and over sampling applied previously. Dataset was not oriented towards the actual scenario of hotels in India.
- • We could see further scope of improvement. we could see that our developed model had
- some restrictions to reach its full capacity in generating the desired results.

greatlearning

Business Recommendations

- In our dataset, there are mostly cases of No-deposit as deposit type and cancellations are also higher in that case.
- Hotel Managers should avoid such type of bookings during on season or they should draft new policies for No-deposit bookings to avoid No-shows. Cancellations are only in the case where car parking space is not required. So, bank managers should not hold bookings of customers who actually require car parking spaces.
- 70% data belongs to the five countries. Therefore, if we want to increase the number of customers from other countries, we can optimize SEO from other sources based on the place, community, and language.

Business Recommendations

- There are very few repeated guests visited to the hotel. Hotel managers need to focus on
- increasing repeated customers. Retaining old visitors is much affordable than acquiring new ones.
- Booking channel origin makes a huge amount of difference to whether a guest cancels or not, and the data consistently comes out in favor of direct bookings over OTAs.
- There were higher cancellations on bookings via OTAs. So, direct bookings avoid the chances of commissions taken by different travel portals thus helping in generating more revenue

Thanks!