Predicting Hotel Booking Cancelation

By Melody Rarasati

BACKGROUND

Over the years, the hotel industry has changed with a majority of bookings now made through third parties such as Booking.com, agoda, traveloka, etc.

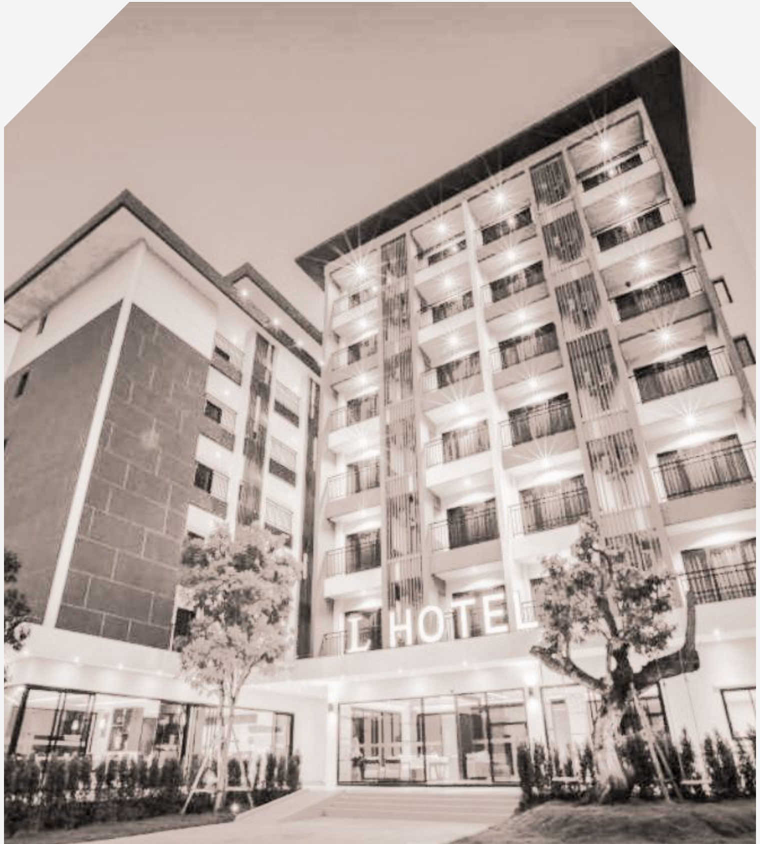
Those Online Travel Agencies (OTA) have transformed cancellation policies from a footnote at the bottom of the page to the main selling point in their marketing campaigns (source). As a result, customers have become accustomed to free cancellation policies.

This increase in booking cancellation makes it harder for hotels to accurately forecast, leading to non-optimized occupancy and revenue loss.

PROBLEM CAUSED BY BOOKING CANCELLATION

Not Optimized Occupancy

Operational Problems (such as over or understaffing)



Revenue Lost

Decrease Customer Satisfaction and Online Reputation Score

OBJECTIVES

Gain insights about the customers (and hopefully reasons why they cancel their reservation)

Build a classification model to predict whether or not a booking will be canceled with the highest accuracy possible.

ABOUT THE DATASET

This data set consists of 119,390 observations and holds booking data for a city hotel and a resort hotel in Portugal from 2015 to 2017. It has 32 variables which include reservation and arrival date, length of stay, canceled or not, the number of adults, children, or babies, the number of available parking spaces, how many special guests, companies, and agents pushed the reservation, etc.

Dataset source:

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

WORKFLOW

Load Data	01
Exploratory Data Analysis	02
Data Preprocessing	03
Modelling	04
Hyperparameter Tuning	05
Conclusion and Recommendation	06

01 DATASET INFORMATION

31 features
1 target → is_canceled

Numerical:

is_canceled
lead_time
arrival_date_year
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
booking_changes
agent
company
days_in_waiting_list
adr
required_car_parking_spaces
total_of_special_requests

Categorical:

hotel
arrival_date_month
meal
country
market_segment
distribution_channel
reserved_room_type
assigned_room_type
deposit_type
customer_type
reservation_status
reservation_status_date

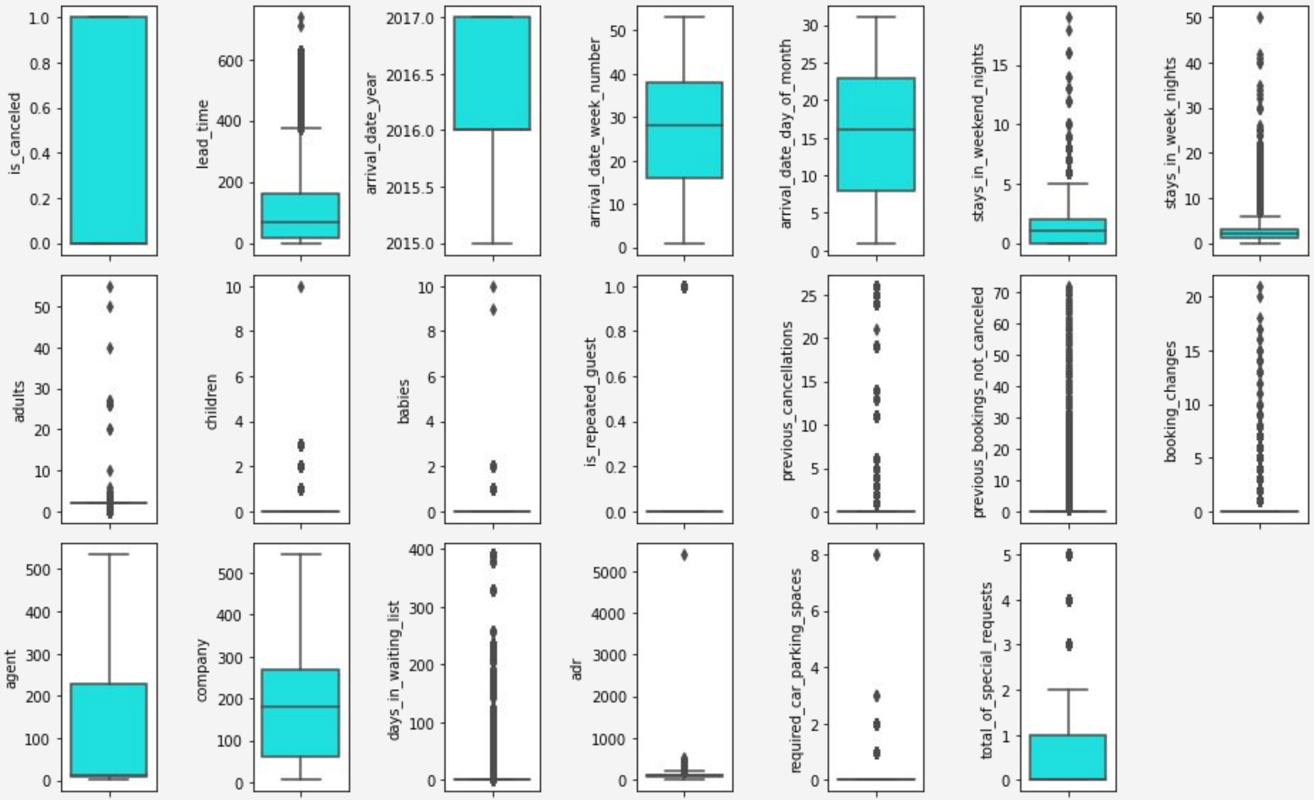
02 Exploratory Data Analysis

Statistical Analysis

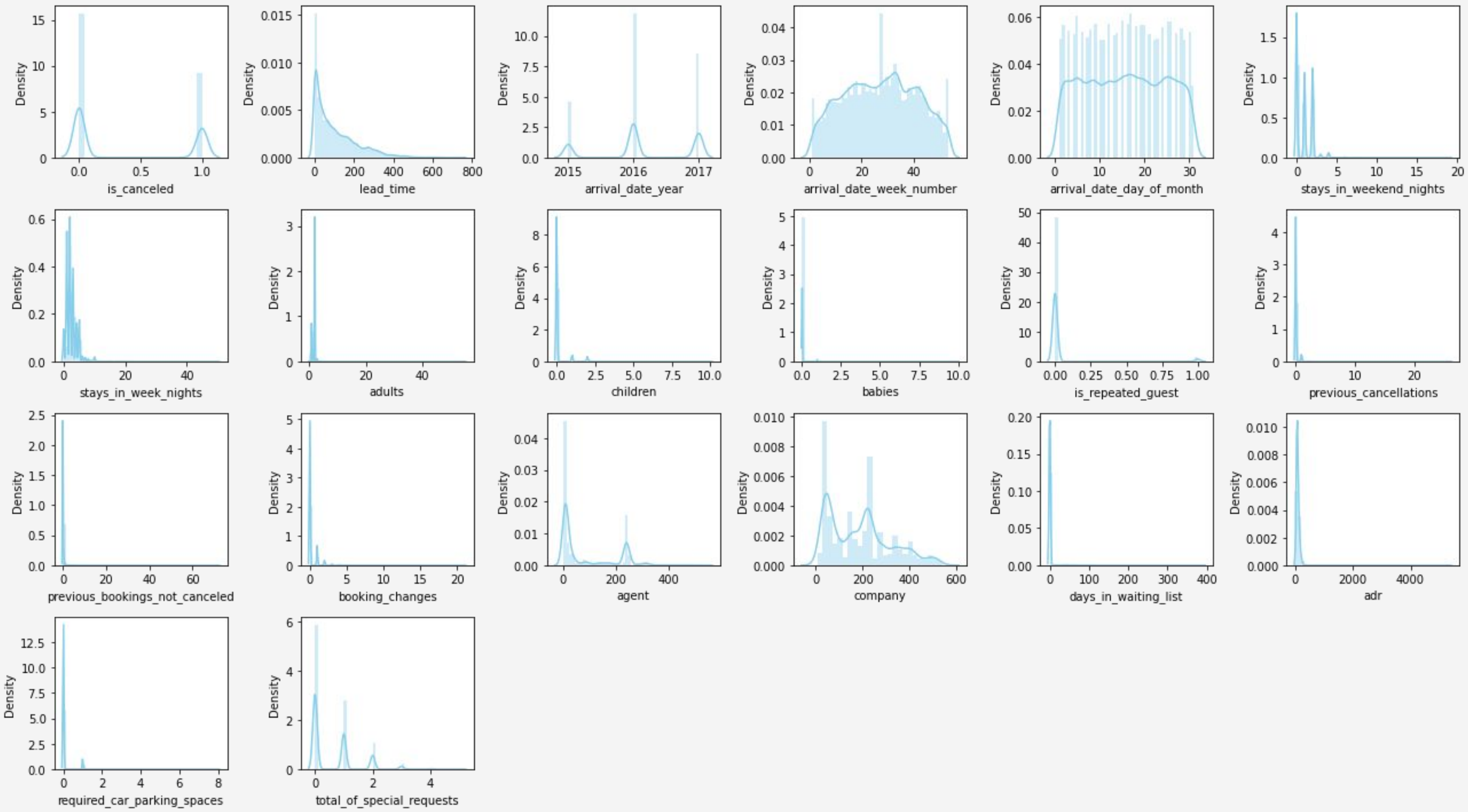
- The adr column (average daily rate) have minimum of -6.38 and a maximum of 5400. A negative ADR could be possible if a hotel had to compensate a guest for some reason. While those numbers are surprising, we do not have enough information to assure that those observations are not accurate datapoints.
- Min value in adults column is 0 is weird because a reservation should be made by at least one adults. Will be processed later in data preprocessing section

	count	mean	std	min	25%	50%	75%	max
is_canceled	119390.0	0.370416	0.482918	0.00	0.00	0.000	1.0	1.0
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0
arrival_date_year	119390.0	2016.156554	0.707476	2015.00	2016.00	2016.000	2017.0	2017.0
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0
children	119386.0	0.103890	0.398561	0.00	0.00	0.000	0.0	10.0
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0
is_repeated_guest	119390.0	0.031912	0.175767	0.00	0.00	0.000	0.0	1.0
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0
agent	103050.0	86.693382	110.774548	1.00	9.00	14.000	229.0	535.0
company	6797.0	189.266735	131.655015	6.00	62.00	179.000	270.0	543.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	126.0	5400.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0

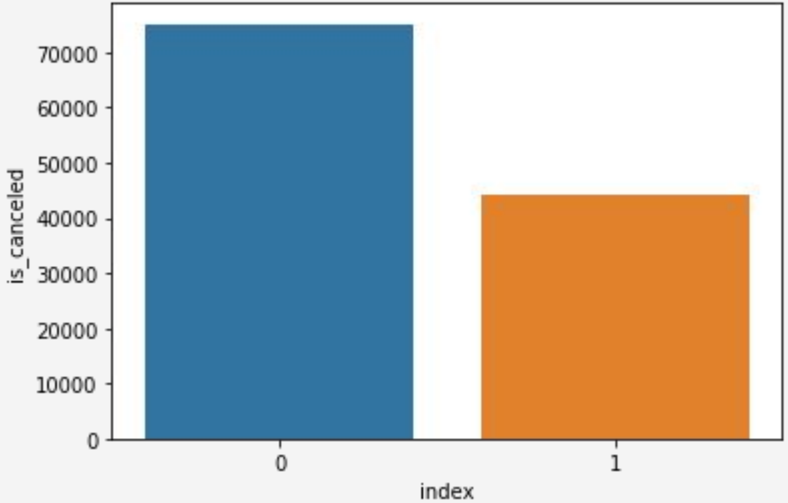
Boxplot to Detect Outlier



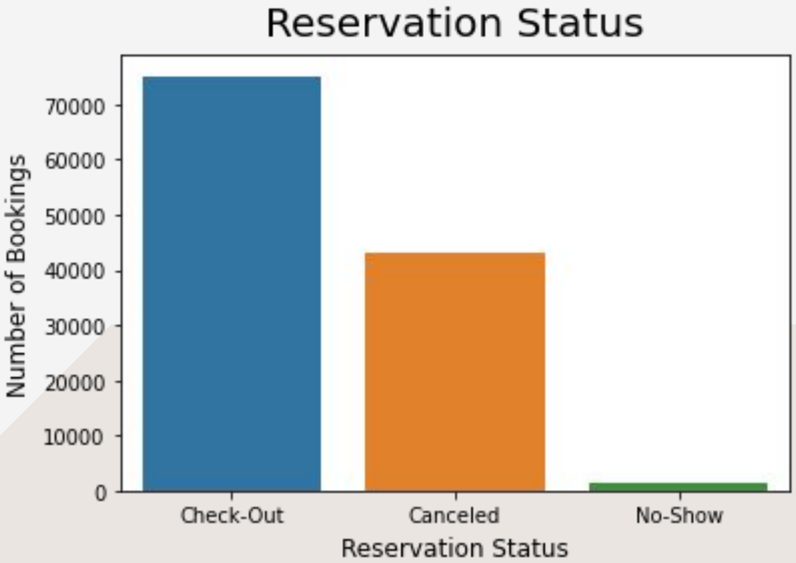
Distribution Form



Canceled Booking

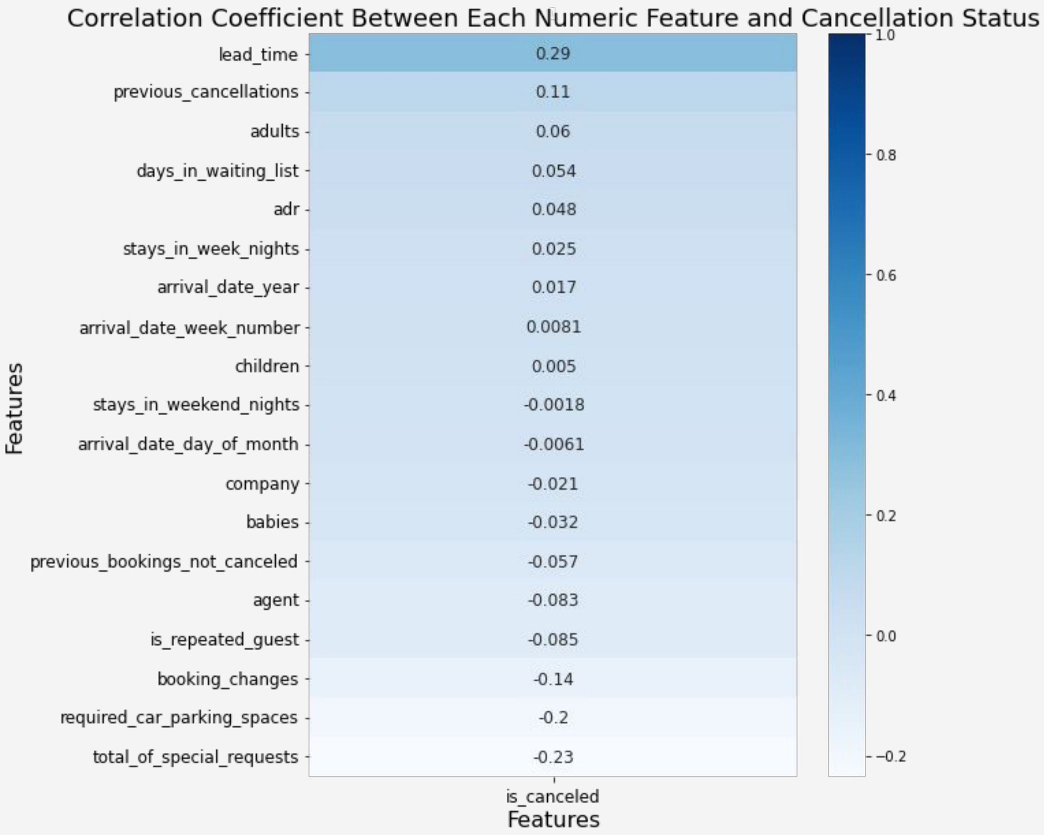


In terms of the target variable, the number of canceled booking (is_canceled = 1) is lower than not canceled booking. But, the imbalance condition is NOT severe (63% : 37%)



Majority of bookings are canceled prior to arrival

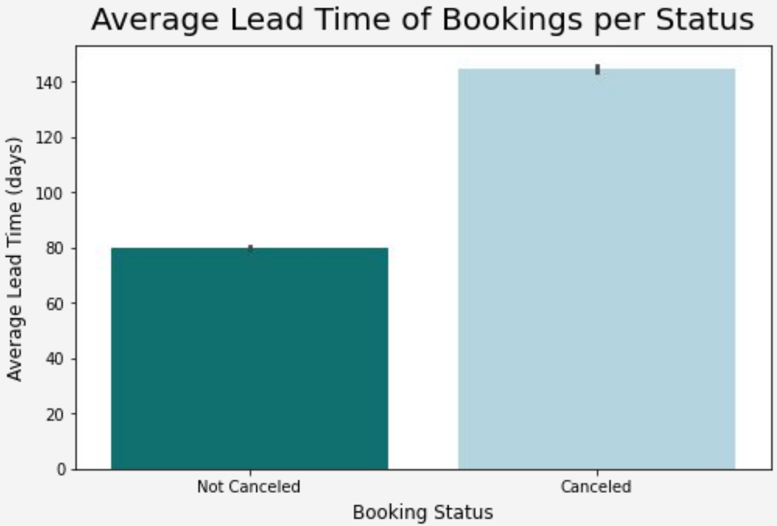
Feature Correlation



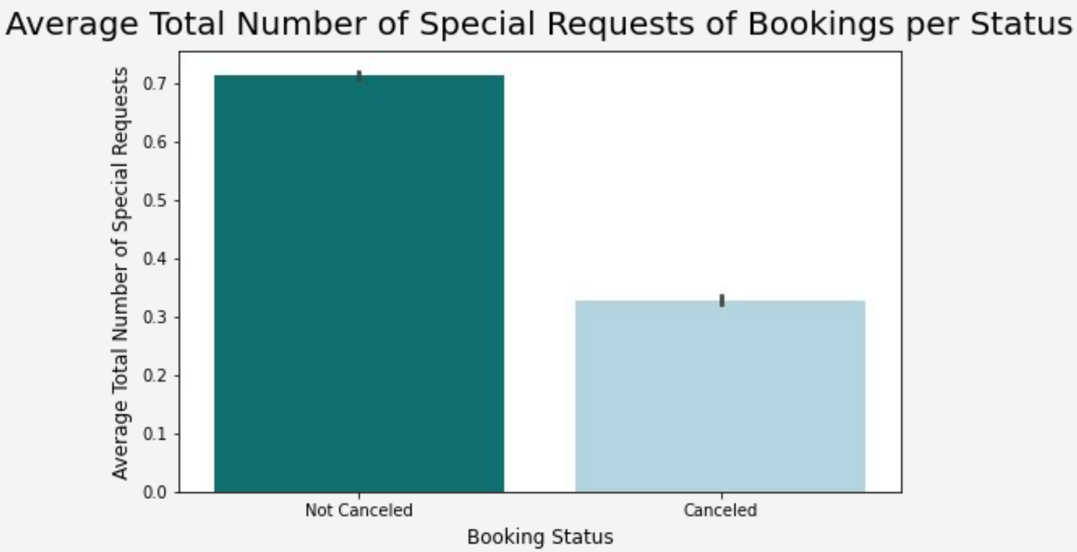
- lead_time is the most highly correlated feature with whether or not a booking is canceled.
- total_of_special_requests is the second feature with the strongest correlation to target feature.
- The number of required_car_parking_spaces is the third feature with the strongest correlation.

Lead Time

Canceled bookings have a longer lead time on average. It makes sense that as the number of days between when the booking is made and the supposed arrival date increases, customers have more time to cancel the reservation and there is more time for an unforeseen circumstance derailing travel plans to arise.

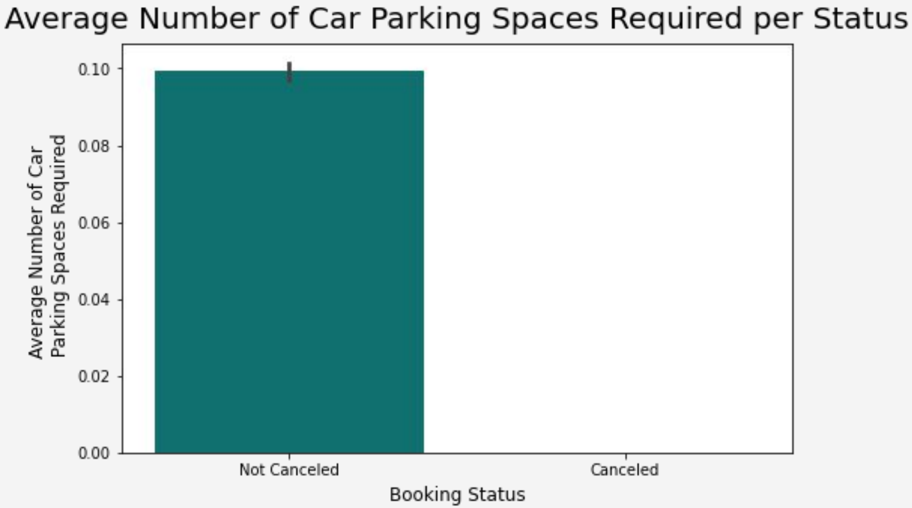


Special Request



Customers who cancel their bookings make on average fewer special requests. As the number of special requests made increases, the likelihood that a booking is canceled decreases. This suggests that engagement with the hotel prior to arrival and feeling like their needs are heard may make a customer less likely to cancel their reservation.

Car Parking Spaces Required



On average, customers who do not cancel their bookings tend to require more parking spaces. Similarly to the number of special requests, it would make sense that the more a customer engages with the hotel (by putting in a request for a parking spot), the less likely they are to cancel. It is also fair to think that by the time a guest is thinking about where they will park their car, they are most likely pretty commited to their destination. Finally, thinking about this from the hotel perpective, it is possible that not many hotels around have a parking. As a result, the need for a parking space would limit the customer in their hotel options and make them less likely to cancel. More information would be required from the hotel directly to confirm this theory. However, if true, this suggests that adding parking spaces could be a way to help reduce cancellations.

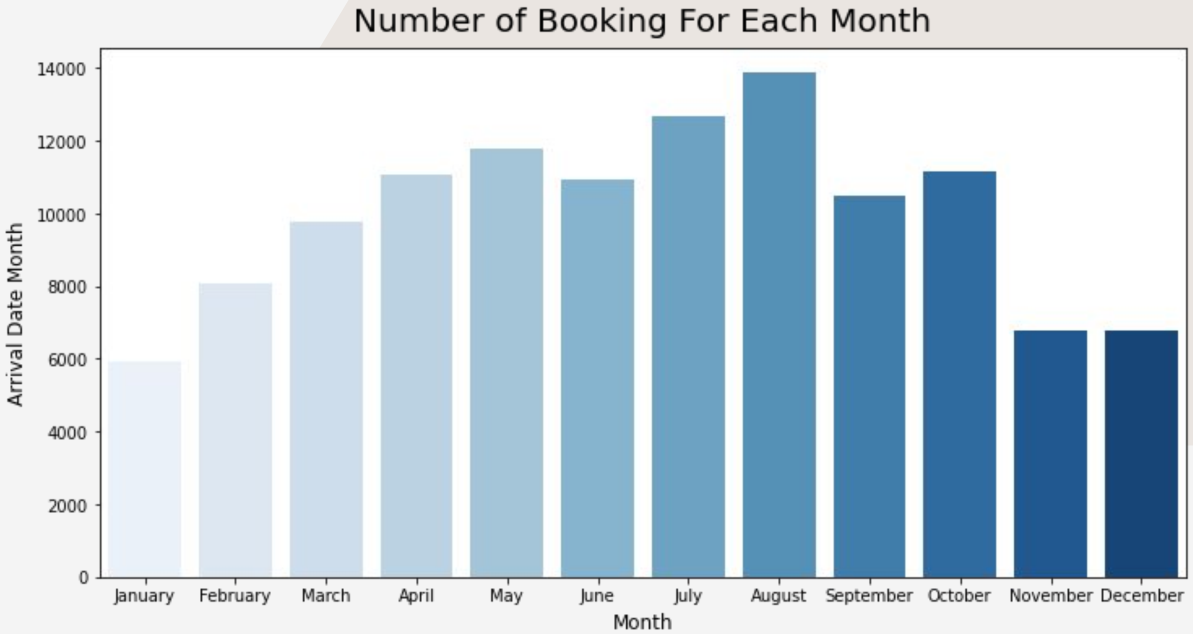
Booking For Each Month

Season In Europe

Months	Min Temperature	Season
DEC - FEB	14°c	Summer
SEP - NOV	7°c	Autumn
MAR - MAY	7°c	Winters
JUN - AUG	2°c	Spring

Based on number of bookings, we can divide the year into three seasons:

Peak season : June - August
Shoulder season : March - May and September-October
Off-season : November - February



03 DATA PREPROCESSING

	feature	missing_value	percentage
0	company	112593	94.31
1	agent	16340	13.69
2	country	488	0.41

Missing Value

- the agent and company features were not included in the model.
- country is small in portion 0.41%, we can simply drop them.

Duplicate Data

The dataset contains 31971 duplicates (27% of data). However, it is possible that multiple bookings with the same features were made on the same day. Since we do not have a feature such as "booking ID", we cannot say for sure that those are true duplicates which makes deleting those "duplicates" questionable. So we will keep the "duplicates"

Clean not logical data

In EDA we found that Min value in adults column is 0. It is weird because a reservation should be made by at least one adults. So we will take only data that have minimal 1 adult.

Data Encoding

Before Encoding

```
[ ] cat_df = df_hotel[categoricals]
    cat_df.head()
```

	hotel	arrival_date_month	meal	market_segment	distribution_channel	reserved_room_type	deposit_type	customer_type
0	Resort Hotel	July	BB	Direct	Direct	C	No Deposit	Transient
1	Resort Hotel	July	BB	Direct	Direct	C	No Deposit	Transient
2	Resort Hotel	July	BB	Direct	Direct	A	No Deposit	Transient
3	Resort Hotel	July	BB	Corporate	Corporate	A	No Deposit	Transient
4	Resort Hotel	July	BB	Online TA	TA/TO	A	No Deposit	Transient

After Encoding

```
[ ] cat_df.head()
```

	hotel	arrival_date_month	meal	market_segment	distribution_channel	reserved_room_type	deposit_type	customer_type
0	0	7	0	0	0	0	0	0
1	0	7	0	0	0	0	0	0
2	0	7	0	0	0	1	0	0
3	0	7	0	1	1	1	0	0
4	0	7	0	2	2	1	0	0

04 Modelling

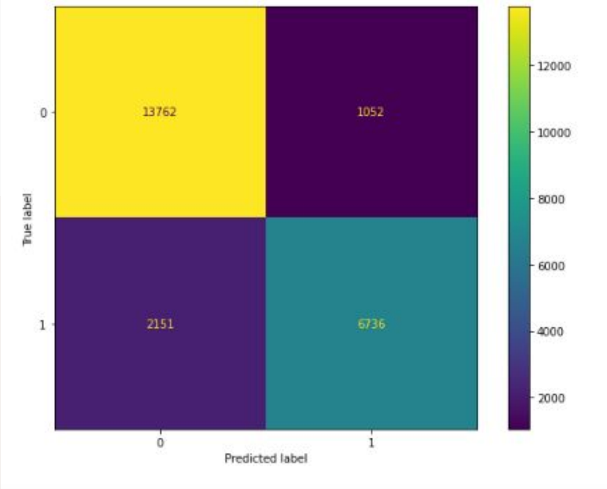
In order to predict Hotel Booking Cancellation by the given data, I will try several classification models. I will use Logistic Regression, KNN, Decision Tree, and Random Forest to model the data.

The imbalance of data in target column is_canceled is not severe (63:37), so I will use accuracy as the scoring parameter.

The dataset is splitted into train data (80%) and test data (20%).

Model Evaluation Comparison

	Model	Accuracy Score	ROC AUC Score
3	Random Forest Classifier	0.864858	0.843474
2	Decision Tree Classifier	0.820261	0.809936
0	Logistic Regression	0.777267	0.734080
1	KNN	0.771022	0.743176



Random Forest Classifier outperform other models.

Accuracy is the number of correctly predicted data points out of all the data points.

Area Under Curve (AUC) score represents the degree or measure of separability.

05 Hyperparameter Tuning

	Model	Accuracy Score Before Tuning	ROC AUC Score Before Tuning	Accuracy Score After Tuning	ROC AUC Score After Tuning
3	Random Forest Classifier	0.864858	0.843474	0.768997	0.693046
0	Logistic Regression	0.777267	0.734080	0.777267	0.734080
1	KNN	0.771022	0.743176	0.771022	0.743176
2	Decision Tree Classifier	0.820261	0.809936	0.780473	0.751907

- The metrics Scoring in base model is higher than tuned model. It means the model before hyperparameter tuning already have the best parameter combination.
- But still hyperparameter tuning is an essential part of controlling the behaviour of a machine learning model

06 Conclusion and Recommendation

- The best model to predict hotel booking cancellation is Random Forest Classifier without hyperparameter tuning. This model classifies whether or not a booking will be canceled with 86% accuracy. As a result, this model would allow hotels to more accurately forecast their occupancy, manage their business accordingly, and increase their revenue.
- There are 3 features that highly correlated with booking cancellation, that is lead_time, total_of_special_requests, and required_car_parking_spaces.
- As the number of special requests made increases, the likelihood that a booking is canceled decreases. This suggests that engagement with the hotel prior to arrival and feeling like their needs are heard may make a customer less likely to cancel their reservation.
- On average, customers who do not cancel their bookings tend to require more parking spaces. Similarly to the number of special requests, it would make sense that the more a customer engages with the hotel (by putting in a request for a parking spot), the less likely they are to cancel. It is also fair to think that by the time a guest is thinking about where they will park their car, they are most likely pretty commited to their destination. Finally, thinking about this from the hotel perpective, it is possible that not many hotels around have a parking. As a result, the need for a parking space would limit the customer in their hotel options and make them less likely to cancel. More information would be required from the hotel directly to confirm this theory. However, if true, this suggests that adding parking spaces could be a way to help reduce cancellations.
- Spring time is the peak season for hotel where occupancy level is high.

THANK YOU

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution.