Introduction

cancelation cancelation cancelation

this is one of the biggest problems that faces the tourism industry which make this industry more risky than other investments , predicting the real demand is a real challenge for the managers in this field and helps them improve their profits, decrease the risk, and be always ready with enough facilities

This dataset contains two datasets with hotel demand data. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Both **hotels are located in Portugal: H1 at the resort region of Algarve and H2 at the city of Lisbon**
our main aim__from this project is to discover the data ,knowing the distributions of our features and try to discover the realationships between them __focusing the problem of high_cancelation_rate in order to find advices to decrease this rate or even predict it accordings the different features

about the dataset

-this dara is extracted from hotels' Property Management System (PMS) SQL
-some of the variables were engineered from other variables from different database tables. The data point time for each observation was defined as the day prior to each booking׳s arrival
-some features are engineered from different variables -this data is **documented** and supplied with the paper -all informations about the dataset are collected from sciencedirect
-PDF file with all informations about the data and its features is attached

```
#importing necessary libraries ans loading the dataset
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
#loading the data
hotels=pd.read_csv('hotel_bookings.csv')
```

```
#taking a look
hotels.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type
0	Resort Hotel	0	342	2015	July	27	1	0	0	2	...	No Deposit
1	Resort Hotel	0	737	2015	July	27	1	0	0	2	...	No Deposit
2	Resort Hotel	0	7	2015	July	27	1	0	1	1	...	No Deposit
3	Resort Hotel	0	13	2015	July	27	1	0	1	1	...	No Deposit
4	Resort Hotel	0	14	2015	July	27	1	0	2	2	...	No Deposit

5 rows × 32 columns

```
#a vertical view for the data
hotels.sample(3).T
```

	99865	91886	7571
hotel	City Hotel	City Hotel	Resort Hotel
is_canceled	0	0	1
lead_time	5	93	78
arrival_date_year	2016	2016	2016
arrival_date_month	October	June	August
arrival_date_week_number	43	27	34
arrival_date_day_of_month	16	26	15
stays_in_weekend_nights	2	1	1
stays_in_week_nights	2	0	4
adults	1	2	2
children	0	0	0
babies	0	0	0
meal	BB	SC	BB
country	LUX	BRA	PRT
market_segment	Groups	Online TA	Online TA
distribution_channel	TA/TO	TA/TO	TA/TO
is_repeated_guest	0	0	0
previous_cancellations	0	0	0
previous_bookings_not_canceled	0	0	0
reserved_room_type	A	A	C
assigned_room_type	A	A	C
booking_changes	1	0	0
deposit_type	No Deposit	No Deposit	No Deposit
agent	21	9	242

	99865	91886	7571
company	NaN	NaN	NaN
days_in_waiting_list	0	0	0
customer_type	Transient-Party	Transient	Transient
adr	66.5	85.5	219
required_car_parking_spaces	0	0	0
total_of_special_requests	0	1	0
reservation_status	Check-Out	Check-Out	Canceled
reservation_status_date	2016-10-20	2016-06-27	2016-05-30

checking the data types ,duplicates , nulls, uniques, and outliers

```
#checking the nulls
hotels.isnull().sum()
```

```
hotel                                 0
is_canceled                           0
lead_time                             0
arrival_date_year                     0
arrival_date_month                    0
arrival_date_week_number              0
arrival_date_day_of_month             0
stays_in_weekend_nights               0
stays_in_week_nights                  0
adults                                0
children                              4
babies                                0
meal                                  0
country                             488
market_segment                        0
distribution_channel                  0
is_repeated_guest                     0
previous_cancellations                0
previous_bookings_not_canceled        0
reserved_room_type                    0
assigned_room_type                    0
booking_changes                       0
deposit_type                          0
agent                             16340
company                          112593
days_in_waiting_list                  0
customer_type                         0
adr                                   0
required_car_parking_spaces           0
total_of_special_requests             0
reservation_status                    0
reservation_status_date               0
dtype: int64
```

```
#checking for the dtptes and null_values of the data
hotels.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   hotel                           119390 non-null  object 
 1   is_canceled                     119390 non-null  int64  
 2   lead_time                       119390 non-null  int64  
 3   arrival_date_year               119390 non-null  int64  
 4   arrival_date_month              119390 non-null  object 
 5   arrival_date_week_number        119390 non-null  int64  
 6   arrival_date_day_of_month       119390 non-null  int64  
 7   stays_in_weekend_nights         119390 non-null  int64  
 8   stays_in_week_nights            119390 non-null  int64  
 9   adults                          119390 non-null  int64  
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64  
 12  meal                            119390 non-null  object 
 13  country                         118902 non-null  object 
 14  market_segment                  119390 non-null  object 
 15  distribution_channel            119390 non-null  object 
 16  is_repeated_guest               119390 non-null  int64  
 17  previous_cancellations          119390 non-null  int64  
 18  previous_bookings_not_canceled  119390 non-null  int64  
 19  reserved_room_type              119390 non-null  object 
 20  assigned_room_type              119390 non-null  object 
 21  booking_changes                 119390 non-null  int64  
 22  deposit_type                    119390 non-null  object 
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64  
 26  customer_type                   119390 non-null  object 
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64  
 29  total_of_special_requests       119390 non-null  int64  
 30  reservation_status              119390 non-null  object 
 31  reservation_status_date         119390 non-null  object 
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

```
#checking the unique values of our data
hotels.nunique()
```

```
hotel                              2
is_canceled                        2
lead_time                        479
arrival_date_year                  3
arrival_date_month                12
arrival_date_week_number          53
arrival_date_day_of_month         31
stays_in_weekend_nights           17
stays_in_week_nights              35
adults                            14
children                           5
babies                             5
meal                               5
country                          177
market_segment                     8
distribution_channel               5
is_repeated_guest                  2
previous_cancellations            15
previous_bookings_not_canceled    73
reserved_room_type                10
assigned_room_type                12
booking_changes                   21
deposit_type                       3
agent                            333
company                          352
days_in_waiting_list             128
customer_type                      4
adr                             8879
required_car_parking_spaces        5
total_of_special_requests          6
reservation_status                 3
reservation_status_date          926
dtype: int64
```

```
#checking the duplicates
hotels.duplicated().sum()
```

```
31994
```

```
#checking some descriptive statistics of the numeric variables
hotels.describe()
```

	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	ba
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119386.000000	119390.00
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.927599	2.500302	1.856403	0.103890	0.007949
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613	1.908286	0.579261	0.398561	0.097436
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.000000	0.000000
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.000000	0.000000
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.000000	0.000000
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000	50.000000	55.000000	10.000000	10.000000

it seems that we have smoe outliers observations in the data specially in ADR feture so we have to check

```
#checking for outliers
hotels.adr.nlargest()
```

```
48515     5400.0
111403     510.0
15083      508.0
103912     451.5
13142      450.0
Name: adr, dtype: float64
```

```
hotels.babies.nlargest()
```

```
46619    10
78656     9
264       2
6719      2
7896      2
Name: babies, dtype: int64
```

```
hotels.stays_in_week_nights.nlargest()
```

```
14038     50
14037     42
101794    41
9839      40
33924     40
Name: stays_in_week_nights, dtype: int64
```

```
hotels.adults.nlargest(20)
```

```
2173    55
1643    50
1539    40
1917    27
1962    27
1587    26
1752    26
1884    26
2003    26
2164    26
2228    20
2418    20
2417    10
2229     6
2231     5
2419     5
125      4
354      4
1023     4
6116     4
Name: adults, dtype: int64
```

```
hotels.children.nlargest()
```

```
328      10.0
6748      3.0
7666      3.0
16360     3.0
18745     3.0
Name: children, dtype: float64
```

we can note that:

> - we have two features that miss alot of data 'company' & 'agent'
> - some other features that have few unique values would be better to convert their type to **category**
> - we have a big number of **duplicates** in our data **but** whereas the variables don't include any unique identification column or names and this data is professionally gathered which dosen't allow this huge number of duplicates so we can freely assume that this observations are just repeated observations for different guests or groups
> - we can add a coloumn of the **total_nights** stayed which will be useful in our analysis
> - the data has some outliers observation that would be better to drop to improve the sense of the statistics and any predction-models we can generate from the data
> - its pretty clear that we have some **outliers** in the data and we have to handle to make sense to our plots and models

data wrangling

```
#dropping the columns with missing values
hotels.drop(['company','agent'],axis=1,inplace=True)
```

```
#converting the type of some columns to categorical
hotels[['meal','arrival_date_month','deposit_type','reservation_status','market_segment','distribution_channel',"is_canceled"]]=hotels[['meal','arrival_date_month','deposit_type','reservation_status','market_segmen
```

```
hotels.is_canceled.cat.rename_categories(['confirmed','canceled'],inplace=True)
hotels.is_canceled.cat.reorder_categories(['confirmed','canceled'],inplace=True)
```

```
#creating the total _nights feature
hotels['total_nights']=hotels.stays_in_weekend_nights+hotels.stays_in_week_nights
```

```
#filling the few missing values of country coloumn
hotels.country.fillna(method='ffill',inplace=True)
```

```
#dropping the outlier point from the data
outlier=hotels[(hotels.adr==5400)|(hotels.children==10)|(hotels.babies>5)|(hotels.adults>4)|(hotels.total_nights>30)].index
hotels.drop(outlier,axis=0,inplace=True)
```

```
# dropping 4 rows with nan values in choldren coloumn
hotels.dropna(axis='rows',inplace=True)
```

rechicking the data det after wrangling

```
#recheckingthe data
hotels.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 119342 entries, 0 to 119389
Data columns (total 31 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119342 non-null  object
 1   is_canceled                     119342 non-null  category
 2   lead_time                       119342 non-null  int64
 3   arrival_date_year               119342 non-null  int64
 4   arrival_date_month              119342 non-null  category
 5   arrival_date_week_number        119342 non-null  int64
 6   arrival_date_day_of_month       119342 non-null  int64
 7   stays_in_weekend_nights         119342 non-null  int64
 8   stays_in_week_nights            119342 non-null  int64
 9   adults                          119342 non-null  int64
 10  children                        119342 non-null  float64
 11  babies                          119342 non-null  int64
 12  meal                            119342 non-null  category
 13  country                         119342 non-null  object
 14  market_segment                  119342 non-null  category
 15  distribution_channel            119342 non-null  category
 16  is_repeated_guest               119342 non-null  int64
 17  previous_cancellations          119342 non-null  int64
 18  previous_bookings_not_canceled  119342 non-null  int64
 19  reserved_room_type              119342 non-null  object
 20  assigned_room_type              119342 non-null  object
 21  booking_changes                 119342 non-null  int64
 22  deposit_type                    119342 non-null  category
 23  days_in_waiting_list            119342 non-null  int64
 24  customer_type                   119342 non-null  object
 25  adr                             119342 non-null  float64
 26  required_car_parking_spaces     119342 non-null  int64
 27  total_of_special_requests       119342 non-null  int64
 28  reservation_status              119342 non-null  category
 29  reservation_status_date         119342 non-null  object
 30  total_nights                    119342 non-null  int64
dtypes: category(7), float64(2), int64(16), object(6)
memory usage: 23.6+ MB
```

```
#some descriptive statistics about the numeric variables in dataset after dropping outliers
hotels.describe()
```

	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
count	119342.000000	119342.000000	119342.000000	119342.000000	119342.000000	119342.000000	119342.000000	119342.000000	119342.000000	119342.000
mean	103.985211	2016.156751	27.165298	15.799316	0.925173	2.494319	1.853396	0.103844	0.007793	0.031917
std	106.841221	0.707368	13.603898	8.780857	0.984473	1.862291	0.488726	0.397599	0.089350	0.175779
min	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.000000	0.000000	0.000000
50%	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.000000	0.000000	0.000000
75%	160.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.000000	0.000000	0.000000
max	737.000000	2017.000000	53.000000	31.000000	10.000000	22.000000	4.000000	3.000000	2.000000	1.000000

data is clean and ready to get discovered

```
#preparing for plotting
sns.set()
color=sns.color_palette()[0]
sns.set_style('whitegrid')
sns.set_context('talk')
palette='PuBuGn_r'
```

```
#the propotion of each hotel kind that has been reserved
hotels.hotel.value_counts(normalize=True)
```

```
City Hotel      0.664611
Resort Hotel    0.335389
Name: hotel, dtype: float64
```

> two third of the reservation was for the city hote and one third fot the resort hotel

```
hotels.is_canceled.value_counts(normalize=True)
```

```
confirmed    0.629661
canceled     0.370339
Name: is_canceled, dtype: float64
```

> the overall ratio of cancelation is 37% which tells that cancelations is a real problem
>
> let's check the cancelation ratio of each hotel

```
#cancelation average for city hotel
hotels[hotels.hotel=='City Hotel'].is_canceled.value_counts(normalize=True)
```

```
confirmed    0.582745
canceled     0.417255
Name: is_canceled, dtype: float64
```

```
#cancelation average of resort hotel
hotels[hotels.hotel=='Resort Hotel'].is_canceled.value_counts(normalize=True)
```

```
confirmed    0.72263
canceled     0.27737
Name: is_canceled, dtype: float64
```

**the overall cancelation rate is 37% and this percentage is realy big and need to deep understand to get good predicted
this percentage is differ between the two hotels , the city hotel has around 1.5 times worser rate**

some univariate visualizations to recognize the distribution of features

```
plt.figure(figsize=(10,6))
plt.subplot(1,2,1)
hotels.hotel.value_counts().plot.pie(explode=(0,.05),startangle=90,autopct='%1.1f%%')
plt.title('booking propotion')
plt.subplot(1,2,2)
hotels[hotels.is_canceled=='confirmed'].hotel.value_counts().plot.pie(explode=(0,.05),startangle=90,autopct='%1.1f%%')
plt.title('distribution of the total cancelation propotion')
```

```
Text(0.5, 1.0, 'distribution of the total cancelation propotion')
```

png

city hotel is most booked with 66.5% from all bookings against 35.5% for resort hotel but also it has the biggest percentage of the total cancelation (74.5%)

let's check the distribution of some numeric features

```
plt.figure(figsize=(8,8))
sns.histplot(hotels.adr,kde=True,discrete=True)
plt.xlim(0,250)
plt.title('distribution of the average dialy rate')
```

```
Text(0.5, 1.0, 'distribution of the average dialy rate')
```

png

```
plt.figure(figsize=(8,8))
sns.histplot(hotels.lead_time,kde=True)
plt.title('the distribution of the lead_time')
```

```
Text(0.5, 1.0, 'the distribution of the lead_time')
```

png

```
plt.figure(figsize=(8,8))
sns.histplot(hotels.total_nights,color=color)
plt.xlim(0,17)
plt.title('distribution of total_nights booked')
```

```
Text(0.5, 1.0, 'distribution of total_nights booked')
```

png

- **the three plots above say that the distribution of the ADR is multimodal with multi peaks can also consider as right skewed distribution which is the favorite for the investorsin the long_term investment**
- **while the distributions of the lead_time and the total_nights are straight and clearly positive skewed (the mean is right of the peak)**

some conditional plots makes the image more obvious

```
plt.figure(figsize=(8,6))
sns.histplot(data=hotels,x="is_canceled", hue='hotel',bins=5,stat='frequency',multiple='dodge',element='bars',palette='Blues',shrink=.8)
plt.title('the bookings be canceled or confirmed')
```

```
Text(0.5, 1.0, 'the bookings be canceled or confirmed')
```

png

from the graph above we can recognize that the cancelation is real problem specially for city hotel

what about the the monthes , do we have to predict more cancelations in specific monthes or seasons .. let's discover the distribution of the bookings throw the different monthes

```
#rordering the monthes category for better plotting
cat=['January','February','March','April','May','June','July','August','September','October','November','December']
hotels['arrival_date_month']=hotels.arrival_date_month.cat.reorder_categories(cat)
sns.catplot(data=hotels,x='arrival_date_month',hue='is_canceled',palette=palette,kind='count',height=6)
plt.xticks(rotation=45)
plt.title('cancelation in different monthes of the year')
```

```
Text(0.5, 1.0, 'cancelation in different monthes of the year')
```

png

its clear that august and julie are our top_season when we expect the double number of guests than other monthes like january and december but they are also the **top_season of the cancelation**

what about the ADR the (the daily average rate) what or how much can we expect in the different times of the year

```
plt.figure(figsize=(8,6))
plt.title('the estimated daily average rate per occupied room with standars_deviation per month')
plt.xticks(rotation=45)
sns.barplot(data=hotels,x='arrival_date_month',y='adr',hue='hotel',palette=palette,ci='sd')
```

```
<AxesSubplot:title={'center':'the estimated daily average rate per occupied room with standars_deviation per month'}, xlabel='arrival_date_month', ylabel='adr'>
```

png

> - as expected the ADR of the top_season_monthes is the greatest

```
#plotting boxplots for ADR in the different years
plt.figure(figsize=(8,6))
sns.violinplot(data=hotels,y='adr',x='arrival_date_year',hue='hotel',palette= 'BuPu' ,split=True,inner='quartile')
plt.ylabel('AVERAGE DAILY RATE')
plt.xlabel('YEAR OF ARRIVAL')
plt.title('the daily average of the different years')
```

```
Text(0.5, 1.0, 'the daily average of the different years')
```

png

> **even with the high rate of cancelation** in augest an july but both hotels can expect to get their best rate in these monthes

let's go on discovering the relationdhip between cancelations and other features

```
sns.catplot(data=hotels,x='distribution_channel',hue='is_canceled',palette=palette,kind='count',height=6)
plt.title('distribution of distribution channel according cancelation')
```

```
Text(0.5, 1.0, 'distribution of distribution channel according cancelation')
```

png

> travel agents & tour operators are have the biggest reservation ratio but also the biggest cancelation ratio while the direct bookings has the best cancelation ratio

```
sns.catplot(data=hotels,x='market_segment',hue='is_canceled',palette=palette,kind='count',height=6)
plt.xticks(rotation=45)
plt.title('the market segments and cancelations')
```

```
Text(0.5, 1.0, 'the market segments and cancelations')
```

png

> - online travel_agents are th bigest source for the reservations of our destination, while the direct_reservation has the least cancelation ratio

```
sns.catplot(data=hotels,x='deposit_type',hue='is_canceled',palette=palette,kind='count',height=6)
plt.title('distribution of deposit type according cancelation')
```

```
Text(0.5, 1.0, 'distribution of deposit type according cancelation')
```

png

> - booking with no deposit is the most common type
> - almost all of nun_refund deposit type reservation were canceled that was totaly unexpected

what is the most common type of the customers

```
g=sns.catplot(data=hotels,x='customer_type',hue='is_canceled',palette=palette,kind='count',height=6)
plt.xticks(rotation=45)
```

```
(array([0, 1, 2, 3]),
 [Text(0, 0, 'Transient'),
  Text(1, 0, 'Contract'),
  Text(2, 0, 'Transient-Party'),
  Text(3, 0, 'Group')])
```

png

> - the most common type of customers is the transient travelers that's may be why the most of bookings are from one to three nights

which hotel has the higher ADR

```
#the distribution of the average daily rate of both hotels
sns.displot(x=hotels.adr,height=6,hue=hotels.hotel,kind='kde')
plt.title('the distribution of the average daily rate of both hotels')
plt.xlim(0,400)
```

```
(0.0, 400.0)
```

png

> it's clear that the distribution of ADR is right skewed and the resort hotel has better rate

```
#rordering the monthes category for better plotting
cat=['January','February','March','April','May','June','July','August','September','October','November','December']
hotels['arrival_date_month']=hotels.arrival_date_month.cat.reorder_categories(cat)
#plotting
sns.displot(x=hotels.lead_time,kind='kde', hue=hotels.is_canceled,fill=True,height=5.5)
plt.title('density distribution of the lead_time')
plt.xlim(0,500)
```

```
(0.0, 500.0)
```

png

> it's a positive skewed distribution the mean lead_time is greater than the most common lead_time

which country book the destination the most

```
countries=hotels.country.value_counts().nlargest(10)
plt.figure(figsize=(8,6))
sns.barplot(y=countries.index, x=countries.values, alpha=0.8)
plt.title('the destination most booking countries')
```

```
Text(0.5, 1.0, 'the destination most booking countries')
```

png

> the most pepole reserving the destination are native citizens

is there a relationship between the previous cancelation of the customer and the probability that he cancel again

```
sns.jointplot(data=hotels,x='is_canceled',y='previous_cancellations',height=6)
```

```
<seaborn.axisgrid.JointGrid at 0x22616882fa0>
```

png

- we see from this graph that the pepole how had more than 6 previous cancelations will often cancel their future reseravtion thoese pepole may be not serious with their reservations

what about the total nights booked and the ADR and the probability of the reservation to be canceled

```
sns.jointplot(data=hotels,x='adr',y='total_nights',hue='is_canceled',height=7,marker='.',s=100)
plt.title('total_nights , ADR and cancelation probability')
```

```
Text(0.5, 1.0, 'total_nights , ADR and cancelation probability')
```

png

- from the graph above we see that reservations with middle daily average rate have more chance to be canceled specially these reservations of more than 15 nights

let's make a simple prediction model

```
#preparing our features for the model
hotels.reset_index()
hotels_cat=pd.get_dummies(hotels[['hotel','meal','is_repeated_guest','market_segment','customer_type','arrival_date_week_number','country','reserved_room_type']])
hotels_num= hotels[['total_nights','adults','children','previous_cancellations','lead_time']]
X=hotels_cat.join(hotels_num,how='inner')
y=hotels.is_canceled
```

```
#fitting the model
x_train,x_test,y_train,y_test=train_test_split(X,y,test_size=.1)
clf=LogisticRegression(C=1000,solver='sag',max_iter=10000)
clf.fit(x_train,y_train)
```

```
LogisticRegression(C=1000, max_iter=10000, solver='sag')
```

```
#model accurancy score
print('the accuracy score of the prediction model is:  ',clf.score(x_test,y_test))
```

```
the accuracy score of the prediction model is:   0.7703393380812735
```

let's plot the the relationship between some important features

```
g = sns.PairGrid(hotels[['is_canceled','lead_time','distribution_channel','deposit_type','previous_cancellations','adr','arrival_date_day_of_month','total_nights']])
g.map(sns.scatterplot,alpha=.5)
```

```
<seaborn.axisgrid.PairGrid at 0x22616759880>
```

png

we can see the correlations between the numeric features

```
corr=hotels.corr()
corr.style.background_gradient(cmap='coolwarm').set_precision(2)
```

	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_gue
lead_time	1.00	0.04	0.13	0.00	0.09	0.17	0.13	-0.04	-0.02	-0.12
arrival_date_year	0.04	1.00	-0.54	-0.00	0.02	0.03	0.05	0.06	-0.01	0.01
arrival_date_week_number	0.13	-0.54	1.00	0.07	0.02	0.02	0.03	0.01	0.01	-0.03
arrival_date_day_of_month	0.00	-0.00	0.07	1.00	-0.02	-0.03	0.00	0.01	0.00	-0.01
stays_in_weekend_nights	0.09	0.02	0.02	-0.02	1.00	0.48	0.11	0.05	0.02	-0.09
stays_in_week_nights	0.17	0.03	0.02	-0.03	0.48	1.00	0.11	0.05	0.02	-0.10
adults	0.13	0.05	0.03	0.00	0.11	0.11	1.00	0.04	0.03	-0.17
children	-0.04	0.06	0.01	0.01	0.05	0.05	0.04	1.00	0.03	-0.03
babies	-0.02	-0.01	0.01	0.00	0.02	0.02	0.03	0.03	1.00	-0.01
is_repeated_guest	-0.12	0.01	-0.03	-0.01	-0.09	-0.10	-0.17	-0.03	-0.01	1.00
previous_cancellations	0.09	-0.12	0.04	-0.03	-0.01	-0.01	-0.01	-0.02	-0.01	0.08
previous_bookings_not_canceled	-0.07	0.03	-0.02	-0.00	-0.04	-0.05	-0.13	-0.02	-0.01	0.42
booking_changes	0.00	0.03	0.01	0.01	0.05	0.08	-0.06	0.05	0.09	0.01
days_in_waiting_list	0.17	-0.06	0.02	0.02	-0.05	-0.00	-0.01	-0.03	-0.01	-0.02

	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_gues
adr	-0.07	0.21	0.08	0.03	0.06	0.07	0.30	0.34	0.03	-0.14
required_car_parking_spaces	-0.12	-0.01	0.00	0.01	-0.02	-0.02	0.02	0.06	0.04	0.08
total_of_special_requests	-0.10	0.11	0.03	0.00	0.07	0.07	0.15	0.08	0.11	0.01
total_nights	0.16	0.03	0.02	-0.03	0.75	0.94	0.13	0.05	0.03	-0.11

conclusions

> - **the cancelation and its prediction is a real problem for the tourism industry and good understandig for this problem and the features that related with will be very useful to decrease the investments's risk of this important industry**
> - **deep understanding of these relationships between cancelations and any features that related with will help managers to improve the confirmation rate of their reservations from the dataset we could find some possible relationships between the cancelations and some features even they are not strong or certain**
> - there is a possible relationship between the customers with previous cancelation and the future cancelations
> - the cancelations increses when the adr is in its common mode specially when the total nights booked are more than 15 nights
> - we couldn't find any clear and strong realationship between any of the features and the cancelation may be we have to include more factors in the dataset we can set a question in cancelation form asking customers to set a reason , that will be more helpful
> - by deeper research we can build a stronger model in future to predict the actual confirmed reservations which will help the managers be ready for the real demand
> - by deeper research we can advice the managers to improve choosing their agents and working together with them to improve the occupation of their hotels
> - most customers are from Portugal , nice but the managers must give extra attention for the international advertising and international distribution channels

resoures

> the site of sincedirect the dataset source
> pandas library doucumentation
> seaborn documentaion
> matplotlib documentation
> stack overflow