HOTEL BOOKING CANCELATION PREDICTION

Hà Trịnh Trung
Trần Vũ Hoàng Tuấn
Đỗ Minh Hiển
Phạm Đình Tân

Business Understanding

Mô tả bài toán: xây dựng một mô hình dự đoán để xác định xem liệu đặt phòng khách sạn có bị hủy hay không, điều này rất quan trọng đối với các khách sạn vì việc hủy ảnh hưởng đến doanh thu và kế hoạch hoạt động.

Report

01

DISCOVERY

02

Analyze

03

Model
Prediction

04

Suggestion

01

DISCOVERY

Information about dataset

```
Data columns (total 36 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
 32  name                            119390 non-null  object
 33  email                           119390 non-null  object
 34  phone-number                    119390 non-null  object
 35  credit_card                     119390 non-null  object
```

Bộ dữ liệu gồm có:

- 119390 dòng
- Cột: Tập dữ liệu chứa 36 cột, thể hiện nhiều thuộc tính khác nhau liên quan đến việc đặt phòng khách sạn.

- Missing Values:
- Cột children có 4 giá trị bị thiếu
- .Cột country có 488 giá trị bị thiếu.
- Cột agent có 16.340 giá trị bị thiếu.
- Cột company có một số lượng đáng kể các giá trị bị thiếu, tổng cộng là 112.593.

Information about dataset

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
name	0
email	0
phone-number	0
credit_card	0

dtype: int64

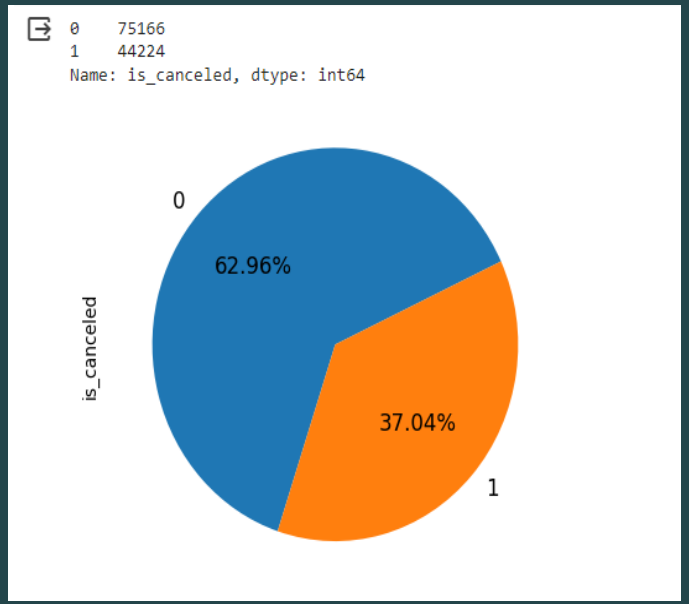
Trong cột company, tỷ lệ phần trăm giá trị bị thiếu là hơn 90%, vì vậy loại bỏ các cột của company

Trong cột children và cột agent, tỷ lệ phần trăm giá trị bị thiếu lớn hơn 20%, vì vậy hãy điền chúng bằng KNNImputer

Thay thế giá trị null trong cột country bằng giá trị thường xuyên nhất

children	0.003350
country	0.408744
agent	13.686238
company	94.306893

LABEL PERCENTIVE



75166 cột có nhãn là 0 tương đương 62,96% non-churned

44224 cột có nhẫn là 1 tươwng đương 37,04% churned

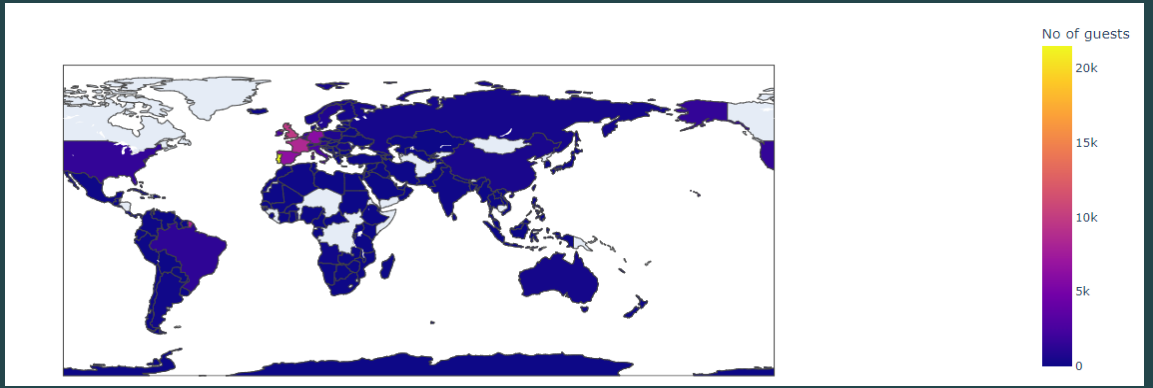
02

ANALYZE

EDA

	country	No of guests
0	PRT	21492
1	GBR	9676
2	FRA	8481
3	ESP	6391
4	DEU	6069
...	...	...
160	BHR	1
161	DJI	1
162	MLI	1
163	NPL	1
164	FRO	1

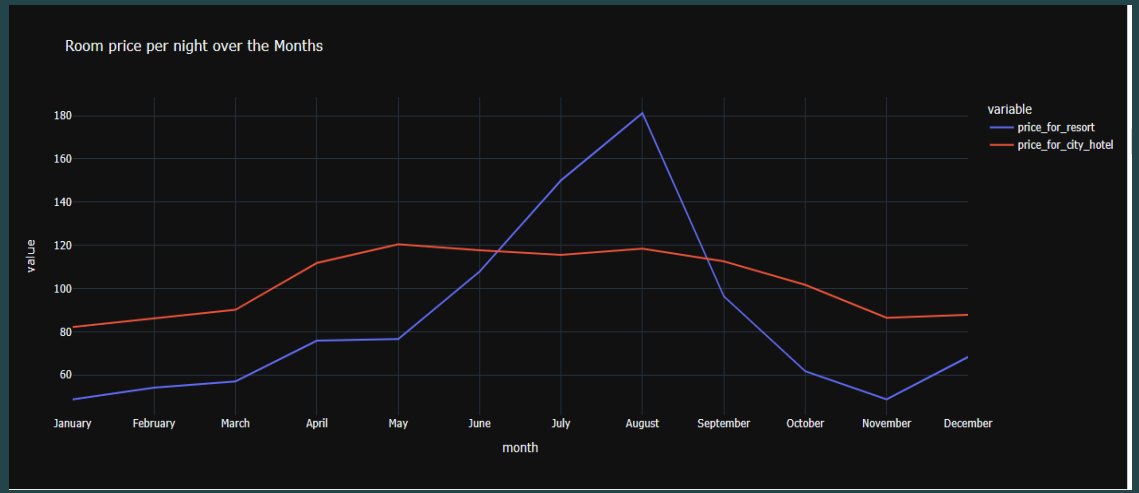
Hầu hết khách đến từ Bồ Đào Nha và các nước khác ở Châu Âu.
giải pháp: tập trung marketing và đưa ra các khuyến mại tại Bồ Đào Nha và các nước ở châu âu



EDA

	month	price_for_resort	price_for_city_hotel
0	January	48.708919	82.160634
1	February	54.147478	86.183025
2	March	57.012487	90.170722
3	April	75.867816	111.856824
4	May	76.657558	120.445842
5	June	107.921869	117.702075
6	July	150.122528	115.563810
7	August	181.205892	118.412083
8	September	96.416860	112.598452
9	October	61.727505	101.745956
10	November	48.681640	86.500456
11	December	68.322236	87.856764

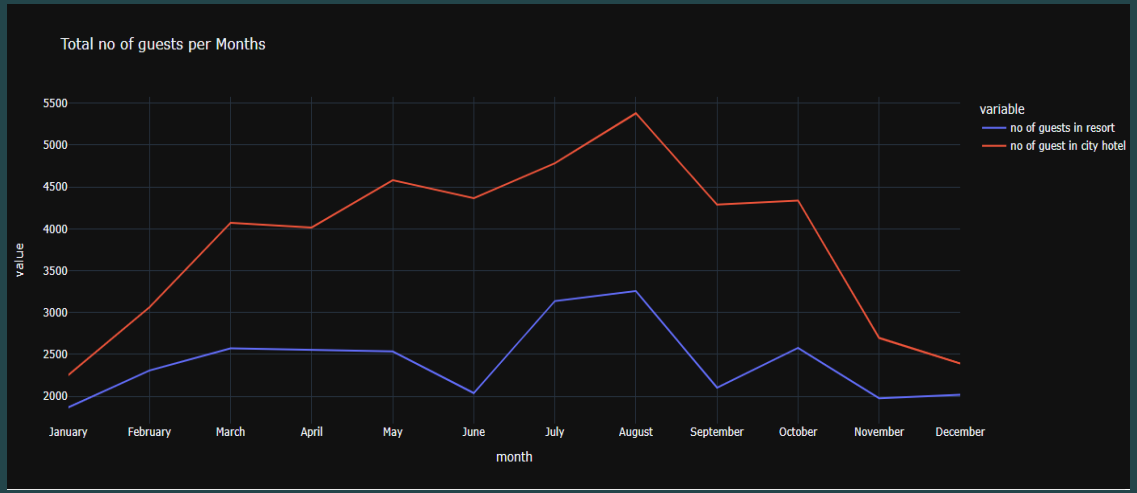
giá ở resort_hotel cao hơn nhiều vào mùa hè (từ tháng 5-8) và giá city_hotel ít biến động hơn và đắt nhất vào mùa Xuân Thu



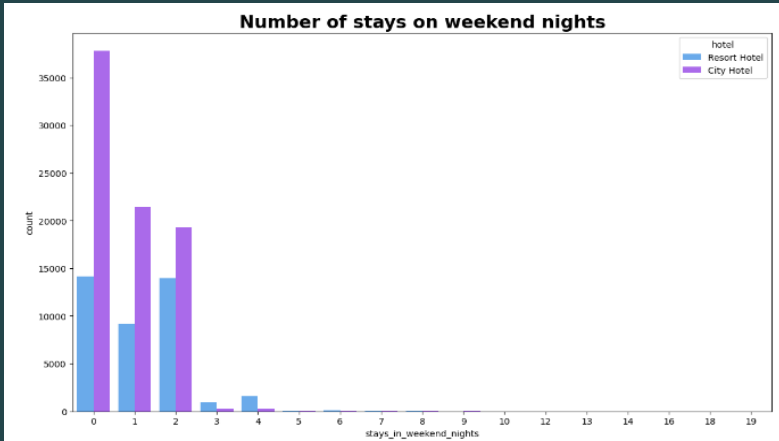
EDA

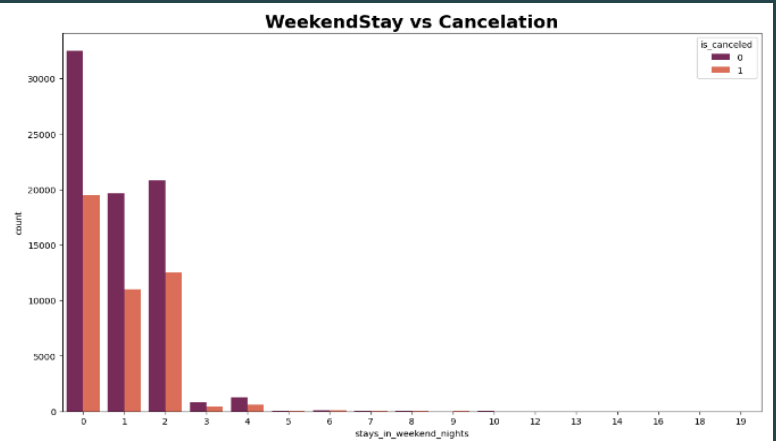
	month	no of guests in resort	no of guest in city hotel
0	January	1868	2254
1	February	2308	3064
2	March	2573	4072
3	April	2550	4015
4	May	2535	4579
5	June	2038	4366
6	July	3137	4782
7	August	3257	5381
8	September	2102	4290
9	October	2577	4337
10	November	1976	2696
11	December	2017	2392

city-hotel có nhiều khách hơn vào mùa xuân và mùa thu, khi giá cũng cao nhất, vào tháng 7 và tháng 8 lượng khách ít hơn, mặc dù giá thấp hơn. Lượng khách đến nghỉ tại resort-hotel giảm nhẹ từ tháng 6 đến tháng 9, đây cũng là thời điểm giá cao nhất. Cả hai khách sạn đều có ít khách nhất vào mùa đông



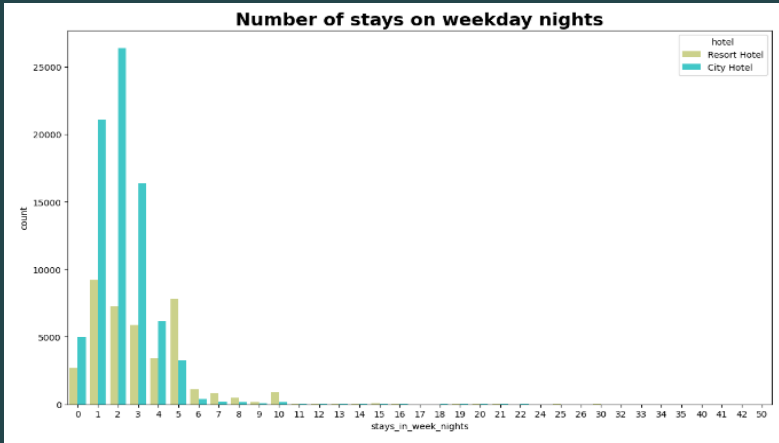
EDA

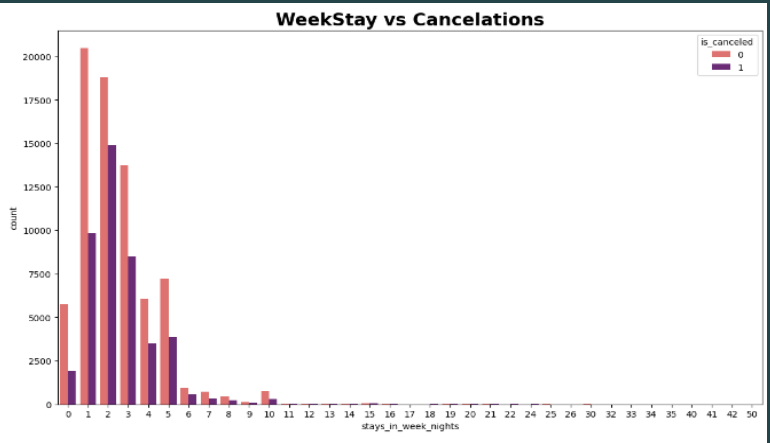




Trong biểu đồ đầu tiên, chúng ta có thể thấy hầu hết các đêm cuối tuần đều được đặt ở Khách sạn City
Biểu đồ thứ hai cho thấy hầu hết các đêm cuối tuần đã đặt đều không bị hủy

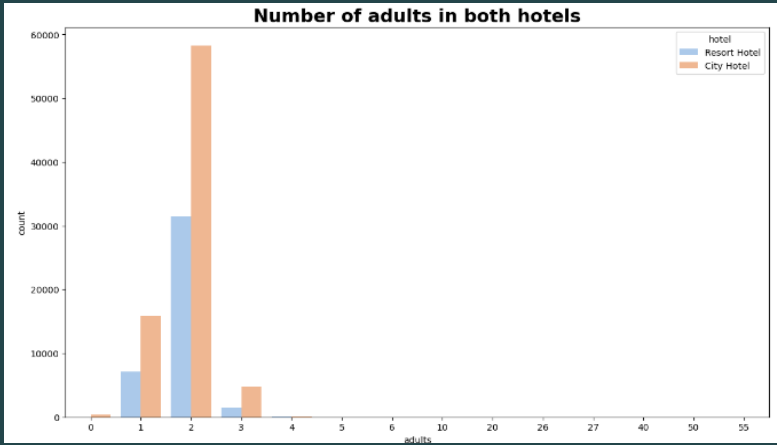
EDA

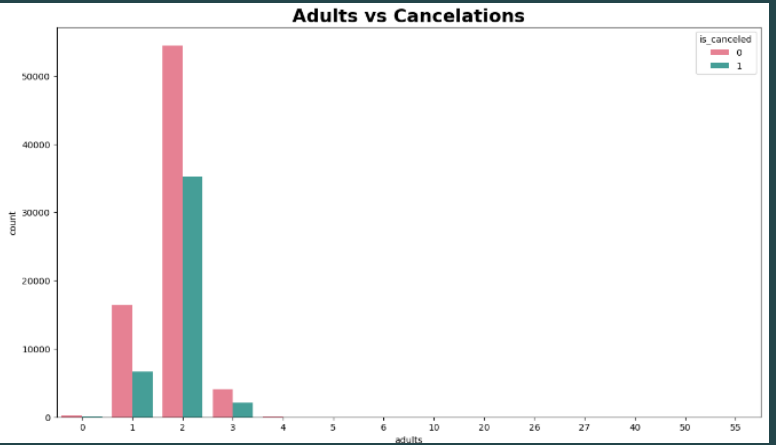




số đêm nghỉ trong tuần ở khách sạn City nhiều hơn
Số lần hủy được ghi nhận ít hơn

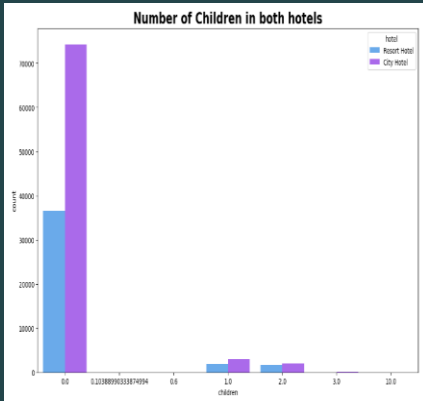
EDA

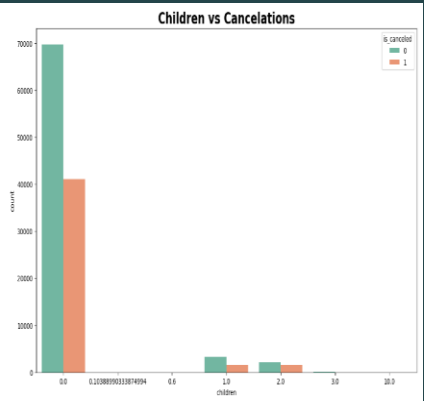




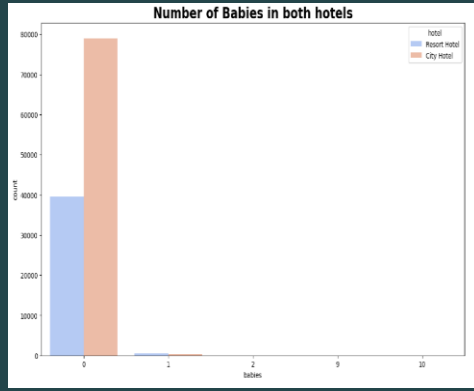
Số lượng người lớn là 2 người nhiều hơn và ưa thích khách sạn ở thành phố hơn là khách sạn nghỉ dưỡng, trên thực tế, hơn một nửa số du khách thậm chí đã hủy đặt phòng

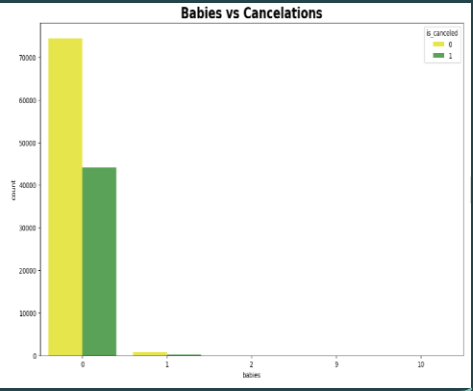
EDA



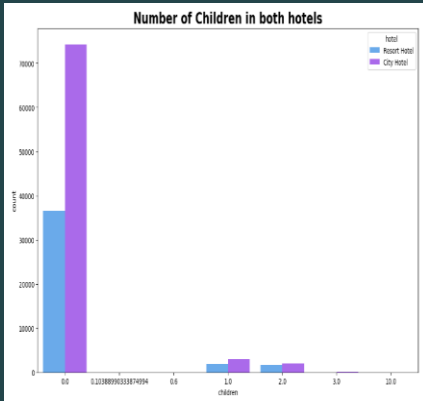


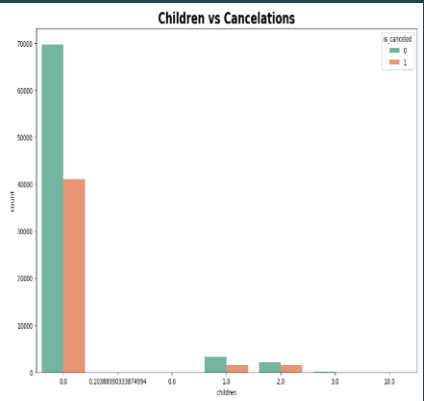
Hầu hết du khách đến theo cặp mà không có trẻ em/trẻ sơ sinh và ưa thích khách sạn ở Thành phố hơn là khách sạn nghỉ dưỡng
du khách có 1 hoặc 2 con cũng thích khách sạn ở thành phố hơn



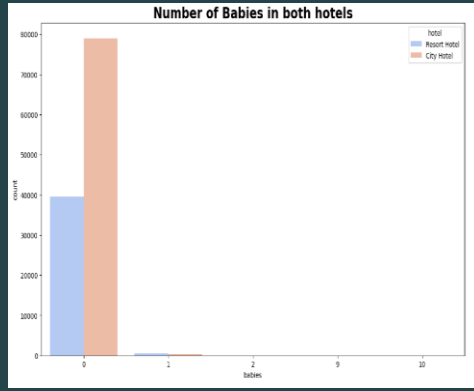


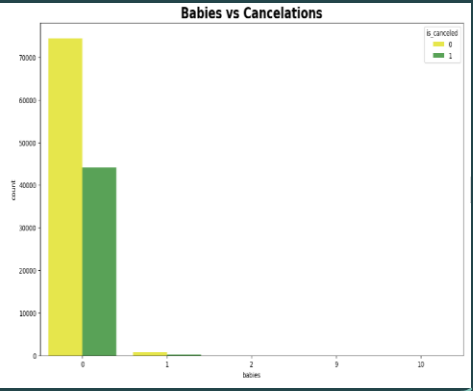
EDA



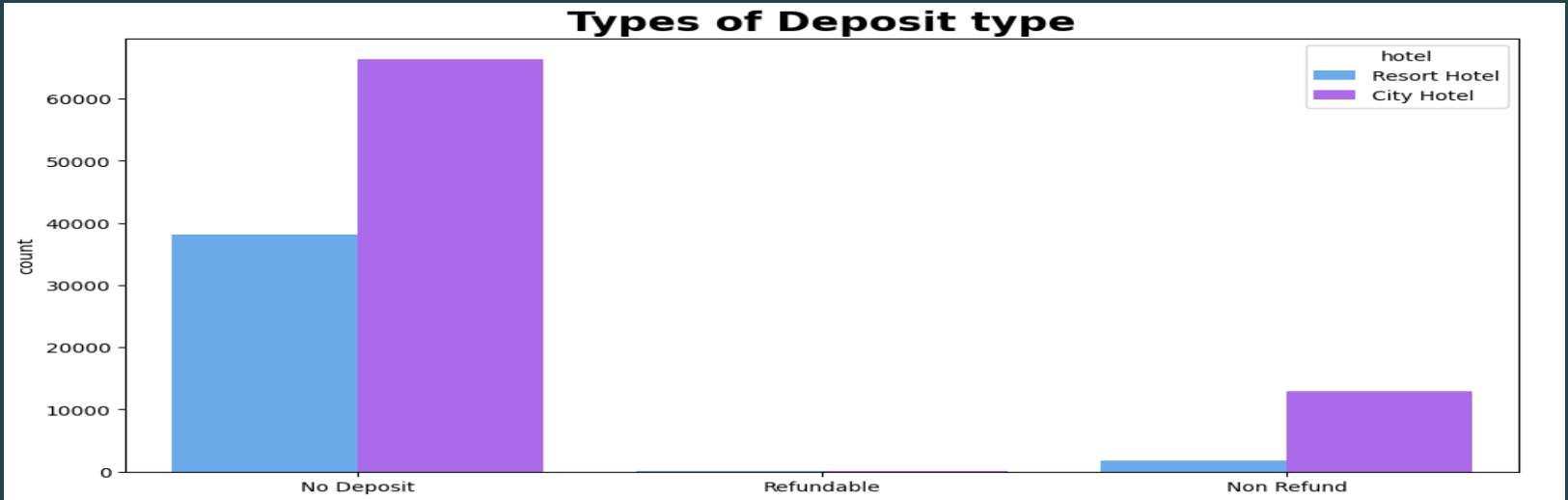


Hầu hết du khách đến theo cặp mà không có trẻ em/trẻ sơ sinh và ưa thích khách sạn ở Thành phố hơn là khách sạn nghỉ dưỡng
du khách có 1 hoặc 2 con cũng thích khách sạn ở thành phố hơn



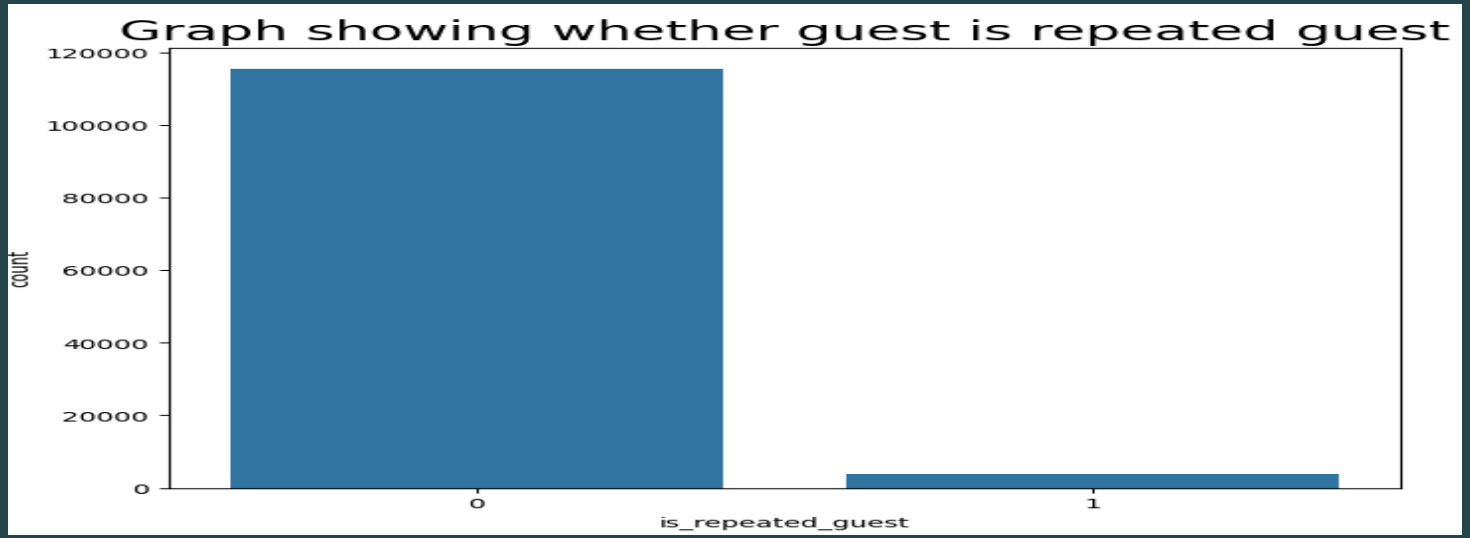


EDA



Không có khoản đặt cọc nào cho khách sạn Thành phố trong khi các Khu nghỉ dưỡng có một số khoản đặt cọc.
Không đặt cọc có thể dẫn đến việc hủy đặt phòng

EDA

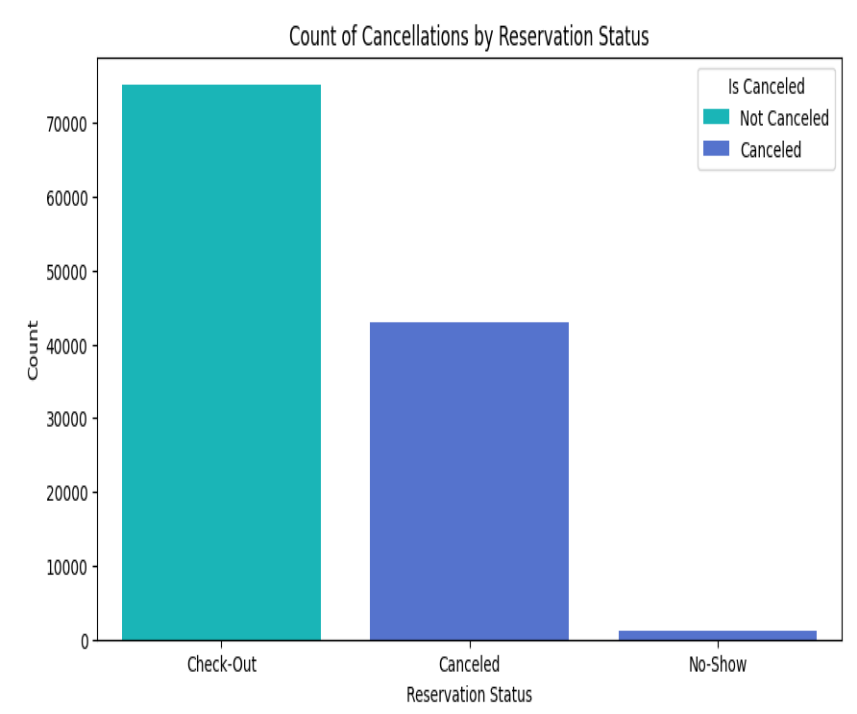


Số lượng khách quay lại thấp
Cần nhắm mục tiêu đến những khách thường xuyên vì họ đã đặt phòng trước đó.

03

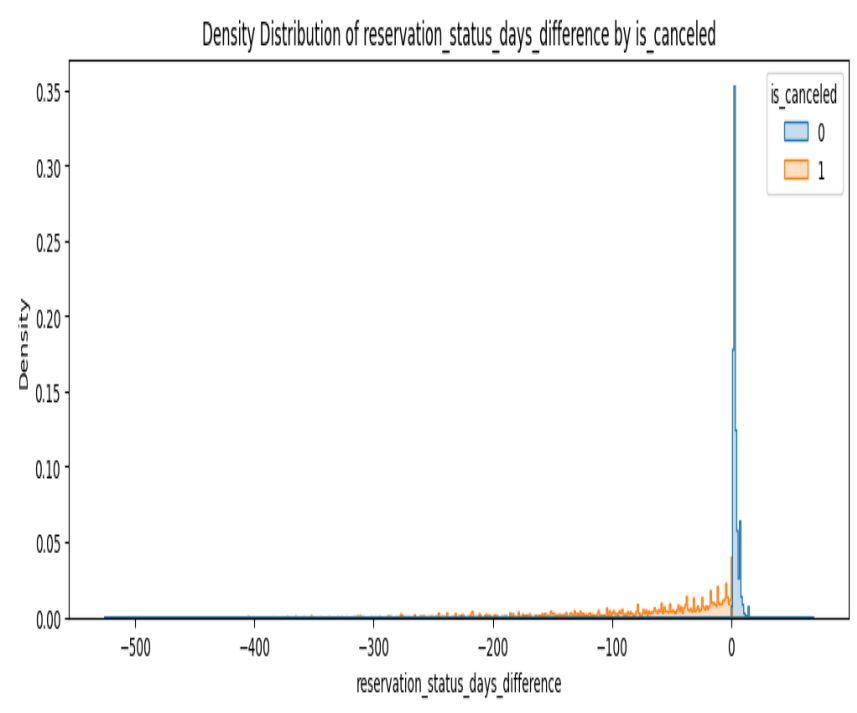
MODEL PREDICTION

DATA PREPOSESSING



Biểu đồ cho thấy rõ rằng tất cả các đặt phòng có trạng thái đặt phòng là "Đã hủy" hoặc "Không đến" thực sự đã bị hủy (is_canceled= 1) và tất cả các đặt phòng có trạng thái đặt phòng là "Trả phòng" đều không bị hủy (is_canceled = 0). Điều này xác nhận rằng tính năng booking_status có liên quan trực tiếp đến biến mục tiêu is_canceled và việc đưa nó vào mô hình sẽ gây rò rỉ dữ liệu. Vì vậy, điều cần thiết là phải loại bỏ tính năng reservation_status trước khi huấn luyện mô hình.

DATA PREPOSESSING



booking_status_days_difference thể hiện số ngày tính từ ngày đến đến ngày mà trạng thái đặt phòng được cập nhật lần cuối.
Giá trị dương cho biết trạng thái đặt phòng đã được cập nhật sau ngày đến. Vì is_canceled luôn bằng 0 đối với các hàng này, điều đó cho thấy rằng những lượt đặt phòng này không bị hủy và trạng thái đặt phòng đã được cập nhật (ví dụ: thành 'Trả phòng') sau khi khách đến. Giá trị âm cho biết trạng thái đặt phòng đã được cập nhật trước ngày đến. Vì is_canceled luôn là 1 đối với các hàng này, điều đó cho thấy rằng các lượt đặt chỗ này đã bị hủy trước ngày đến. Điều này cho thấy mối quan hệ rõ ràng giữa booking_status_days_difference và tính năng is_canceled. Nếu đặt phòng bị hủy, trạng thái đặt phòng thường được cập nhật trước ngày đến. Ngược lại, nếu đặt phòng không bị hủy, trạng thái đặt phòng thường được cập nhật sau ngày đến.
Do đó, nên bỏ qua booking_status_date và booking_status_days_difference khỏi mô hình. Cái sau bắt nguồn từ cái trước và cả hai đều dẫn đến rò rỉ dữ liệu, điều này có thể dẫn đến một mô hình không khái quát tốt cho dữ liệu mới:

DATA PREPOSESSING

	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	119390.0	0.103893	0.398557	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	119390.0	75.543395	106.916090	1.00	7.00	9.000	152.0	535.0
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

dữ liệu nhiễu
-adr: hể hiện số tiền trung bình mà khách phải trả cho một phòng mỗi ngày. Trong tập dữ liệu này, ADR có giá trị tối thiểu là -6,38. Tỷ lệ âm không có ý nghĩa gì, nó chỉ ra những lỗi tiềm ẩn hoặc những trường hợp đặc biệt.
-adults: Giá trị tối thiểu là 0, ngụ ý đặt chỗ mà không có người lớn. Đây có thể là lỗi nhập dữ liệu trừ khi có những trường hợp chính đáng khi chỉ có trẻ em hoặc trẻ sơ sinh đặt phòng.
-hildren and babies: Cả hai đều có giá trị tối đa là 10, có vẻ cao bất thường đối với một lần đặt phòng. Đây có thể là một lỗi ngoại lệ hoặc lỗi nhập dữ liệu tiềm ẩn,

BUILDING MODEL

Accuracy Score of Logistic Regression is : 0.7942177585409925
Confusion Matrix :
[[13713 1157]
[3740 5187]]
Classification Report :

	precision	recall	f1-score	support
0	0.79	0.92	0.85	14870
1	0.82	0.58	0.68	8927
accuracy			0.79	23797
macro avg	0.80	0.75	0.76	23797
weighted avg	0.80	0.79	0.79	23797

Accuracy Score of Decision Tree is : 0.8260284909862587
Confusion Matrix :
[[12779 2091]
[2049 6878]]
Classification Report :

	precision	recall	f1-score	support
0	0.86	0.86	0.86	14870
1	0.77	0.77	0.77	8927
accuracy			0.83	23797
macro avg	0.81	0.81	0.81	23797
weighted avg	0.83	0.83	0.83	23797

Accuracy Score of KNN is : 0.7724923309660882
Confusion Matrix :
[[12759 2111]
[3303 5624]]
Classification Report :

	precision	recall	f1-score	support
0	0.79	0.86	0.82	14870
1	0.73	0.63	0.68	8927
accuracy			0.77	23797
macro avg	0.76	0.74	0.75	23797
weighted avg	0.77	0.77	0.77	23797

Accuracy Score of Random Forest is : 0.8649829810480313
Confusion Matrix :
[[13817 1053]
[2160 6767]]
Classification Report :

	precision	recall	f1-score	support
0	0.86	0.93	0.90	14870
1	0.87	0.76	0.81	8927
accuracy			0.86	23797
macro avg	0.87	0.84	0.85	23797
weighted avg	0.87	0.86	0.86	23797

Accuracy Score of Ada Boost Classifier is : 0.8157750977013909
Confusion Matrix :
[[13915 955]
[3429 5498]]
Classification Report :

	precision	recall	f1-score	support
0	0.80	0.94	0.86	14870
1	0.85	0.62	0.71	8927
accuracy			0.82	23797
macro avg	0.83	0.78	0.79	23797
weighted avg	0.82	0.82	0.81	23797

SUGGESTION

Đa số khách sạn được đặt là khách sạn trong thành phố. Chắc chắn cần phải dành quỹ nhắm mục tiêu nhiều nhất cho những khách sạn đó.

tỷ lệ hủy cao có thể là do chính sách không đặt cọc cao.

nên nhắm mục tiêu vào các tháng từ tháng 5 đến tháng 8. Đó là những tháng cao điểm do đang là mùa hè.

Đa số khách đến từ Tây Âu đặc biệt là bồ đào nha . nên chi một khoản ngân sách đáng kể cho những chính sách khuyến mại và chiến dịch marketing

Vì không có khách quay lại nên nhắm mục tiêu quảng cáo đến khách để tăng lượng khách quay lại.

Các chiến lược để chống lại tỷ lệ hủy phòng cao tại khách sạn
-Đặt mức giá không hoàn lại, Thu tiền đặt cọc và thực hiện các chính sách hủy cứng nhắc hơn.
-Khuyến khích đặt phòng trực tiếp bằng cách đưa ra giảm giá đặc biệt

Thank YOU

For Your Attention