Predict Hotel Reservation Cancellations

Background

Dự án "Predict_Reserve" là một nghiên cứu về dữ liệu đặt phòng khách sạn nhằm dự đoán việc hủy đặt phòng. Đựa trên bộ dữ liệu Hotel_Reserve trên Kaggle

Dự án "Predict_Reserve" mang lại những thông tin quan trọng từ dữ liệu đặt phòng khách sạn thông qua phân tích khám phá, phân cụm dữ liệu và mô hình dự đoán. Kết quả của dự án có thể được áp dụng để cải thiện quy trình đặt phòng và tối ưu hóa trải nghiệm khách hàng.

Table Of Content

Collecting Data

Dữ liệu được lưu vào Driver và tải lên thông qua thư viện **pandas**

```
#link data
link = 'https://drive.google.com/file/d/1UprIC1PBsCUx3XHhAfDmTkoe4phtByFZ/view?usp=sharing'
```

```
#load data
path = 'https://drive.google.com/uc?export=download&id='+link.split('/')[-2]
df_reserve = pd.read_csv(path,encoding= 'unicode_escape')
```

Exploring Data Analyst

- Phân tích và đánh giá các biến Numerical, Date_time, Categorical và Target

  :

- Đánh giá dữ liệu bị thiếu

- Đánh giá tương quan giữa các biến

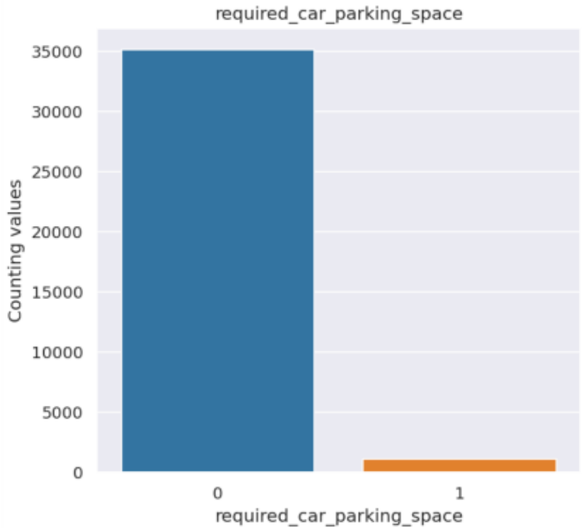
Exploring Data Analyst

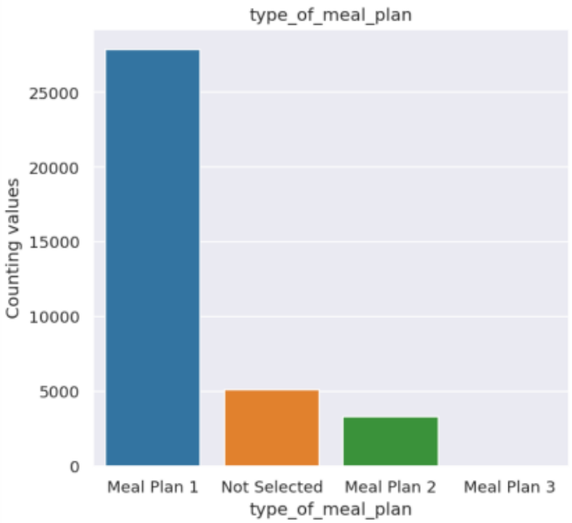
Categorical Feature

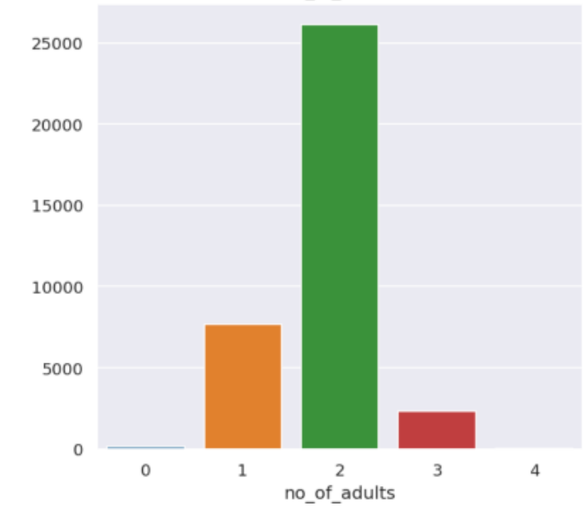
- Loại phòng 1 được đặt nhiều nhất, nhiều hơn rất nhiều so với tất cả các loại phòng khác. Có thể nói rằng giá trung bình của loại phòng 1 là 103,423539 euro hoặc 103,42 euro * 80 ruble/euro = 8273,6 ruble.
- Hầu như không ai sử dụng chỗ đậu xe.
- Trong số các gói bữa ăn, họ chọn gói 1 hoặc không chọn gói bữa ăn, ít hơn là chọn gói 2 và không ai chọn gói 3.
- Về số đêm nghỉ cuối tuần và hàng tuần, không có gì bất thường, tức là hầu hết mọi người đặt khách sạn trong một tuần cùng một cặp cuối tuần.
- Trong số những người đến, có khả năng cao là các cặp M+W không có con.

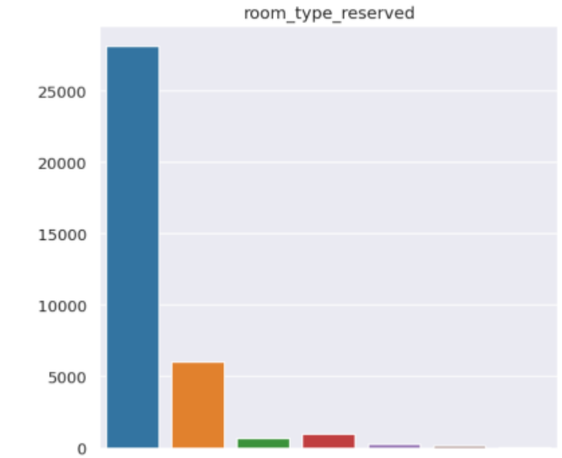
Categorical Feature

```
# Look at the categorical data
a = 1
for x in df[categorical]:
  plt.subplot(4,3,a)
  plt.gca().set_title(x)
  sns.countplot(x = x, palette = 'tab10', data = df)
  plt.ylabel('Counting values')
  a += 1
```



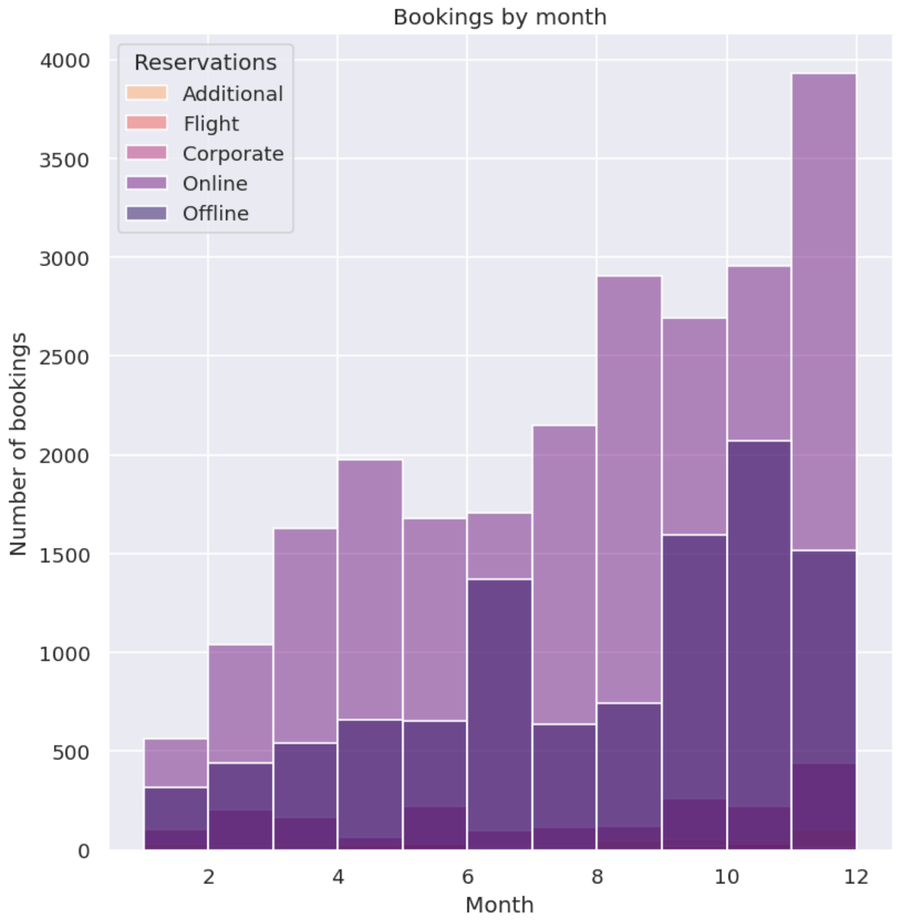






Exploring Data Analyst

Thời gian đặt phòng cao điểm là từ tháng 8 đến tháng 12. Khoảng thời gian này rơi vào mùa đông, có thể không phải là thời điểm lý tưởng để du lịch nhưng lại là thời điểm hoàn hảo để lên kế hoạch cho những chuyến du xuân hè sắp tới.

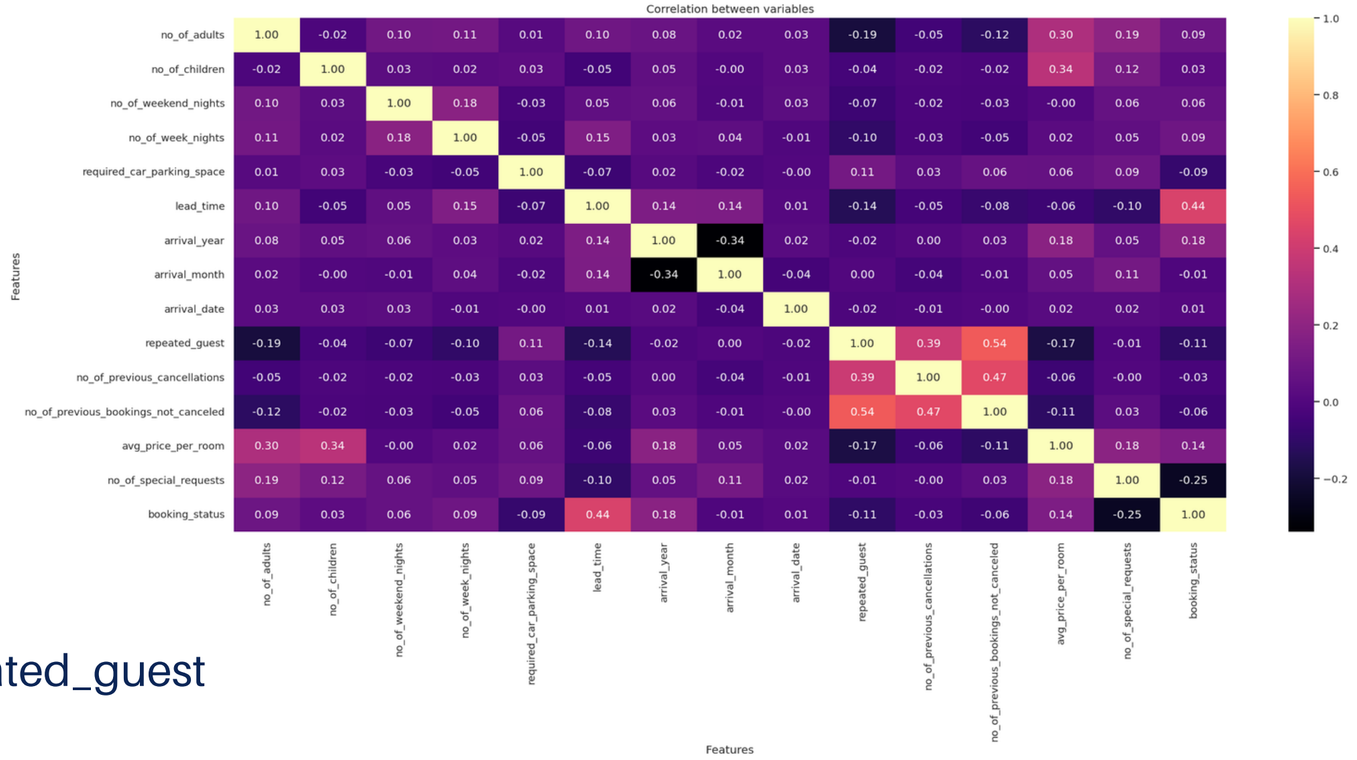


Cancellation Rate

```
#Calculate Cancellation Rate
value = df["booking_status"].value_counts()
rate = value.sum()
value*100/rate
```

```
Not_Canceled    67.236389
Canceled        32.763611
```

The Correlation Between Features

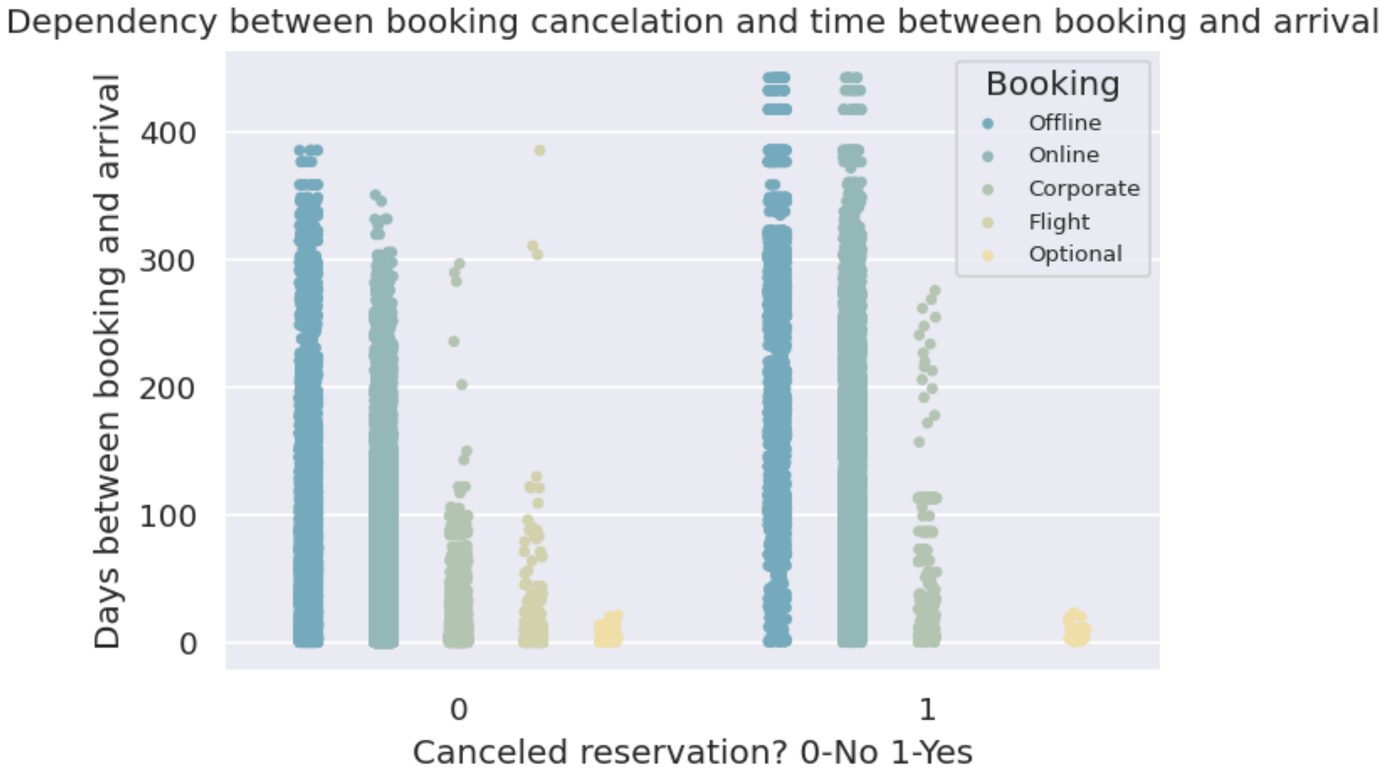


Highlight correlation is seen

#lead_time / booking_status

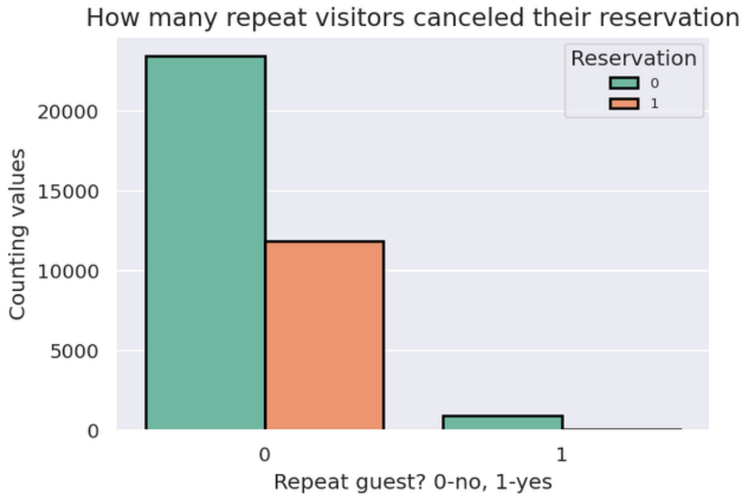
#no_of_previous_cancellations / repeated_guest

#no_of_previous_bookings_not_canceled / repeated_guest



Đặt phòng trực tiếp (offline) thông thường được coi là đáng tin cậy hơn vì ít có khả năng bị hủy bỏ hoặc ảnh hưởng so với đặt phòng trực tuyến. Đáng ngạc nhiên là số ngày giữa ngày đặt phòng và ngày đến không có tác động lớn đến khả năng hủy đặt phòng. Mặc dù có thể dường như việc đặt phòng càng sớm, càng có khả năng bị hủy bỏ do các yếu tố cá nhân như thay đổi kế hoạch, v.v., nhưng trong thực tế, điều này không có nhiều tác động.

Khoảng giá xung quanh 100 euro là nơi diễn ra số lượng đặt phòng cao nhất. Bên cạnh đó, rõ ràng có nhiều trường hợp hủy đặt phòng bắt đầu từ khoảng một trăm ngày và một trăm euro. Do đó, chúng ta có thể nói rằng có một sự tương quan giữa giá phòng và số ngày từ ngày đặt phòng đến ngày đến. Nói cách khác, càng đắt đỏ phòng và càng lâu thời gian giữa việc đặt phòng và đến nơi, càng cao khả năng hủy đặt phòng.



Từ biểu đồ, có thể nhận thấy rằng khách hàng quay lại có số lần hủy ít hơn khách hàng mới.

Ngoài ra, cũng có thể thấy rõ rằng có một tỷ lệ đáng kể giữa khách hàng mới và khách hàng thường xuyên.

Rất ít lượt đặt phòng "thành công" được thực hiện liên tiếp bởi một khách hàng, điều này có thể nói lên sự trung thành và tin tưởng của họ đối với thương hiệu.

Preprocessing Data

- Xoá bỏ các biến datetime và Booking_ID

  :

- Chuyển các biến categorical thành dạng object

- Sử dụng get_dummies trong thư viện pandas để one-hot encoding

Segmentation

Segment 0:

- chủ yếu là kênh online
- meal plạn 0
- có mức avg_price cao nhất

Segment 1:

- chủ yếu là Corporate
- tỷ lệ cancelation rất thấp

Segment 2:

- tập trung nhiều nhất kênh online và 1 phần ở offline
- tỷ lệ cancelation cao
- meal plan 3 và 0
- có mức avg_price trung bình
- có thể là các cặp đôi trẻ, không có children

Segment 3:

- chủ yếu là kênh offline và một phần Corporate
- meal plan 0 và 1
- có mức avg_price trung bình
- có thể là các cặp đôi trẻ, không có children

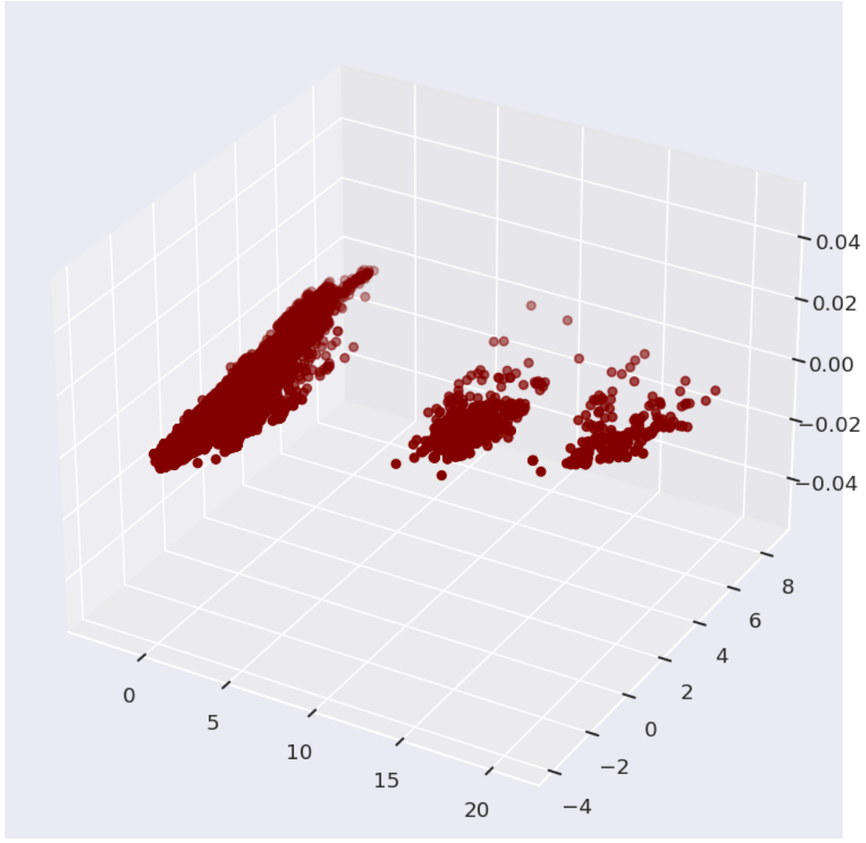
Segmentation

Step by Step

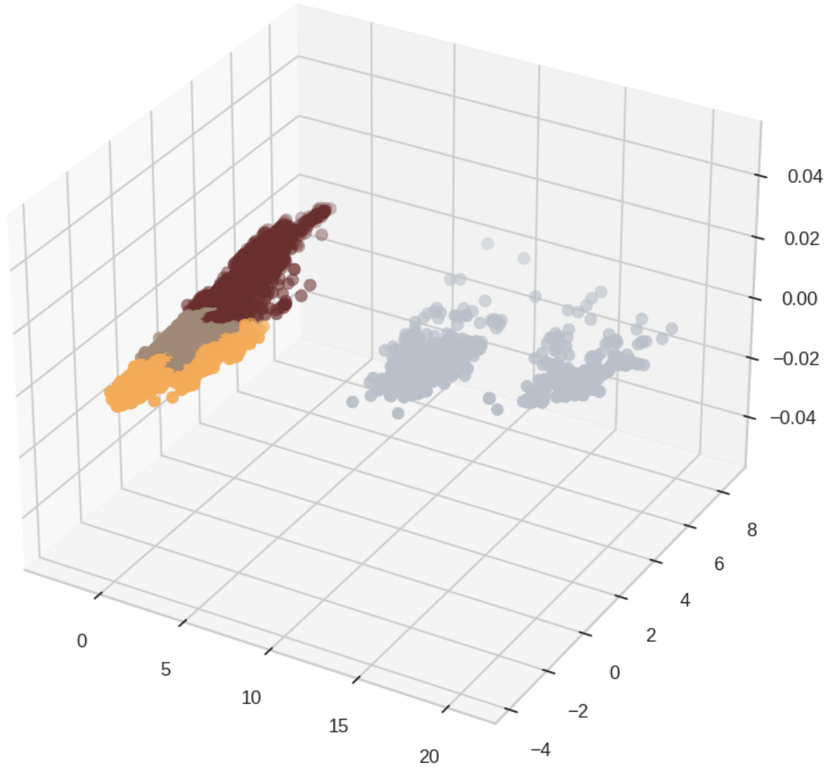
- Scaled Data Using StandardScaler
- Giảm Chiều Dữ Liệu Với Số Lượng Thành Phần Chính Là 2
- Sử Dụng Mô Hình K-Means Để Phân Cụm Dữ Liệu
- Gán Label Và Phân Tích Khám Phá

Input - Output

A 2D Projection Of Data In The Reduced Dimension



The Plot Of The Clusters



Model & Results

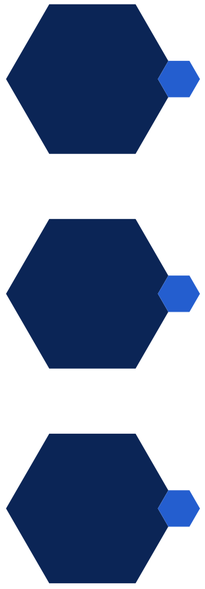
:

	Model	Accuracy Score	F1 score	Precision	Recall
0	Random forest	0.876413	0.806000	0.840048	0.774605
0	KNN	0.811265	0.691684	0.754173	0.638758
0	Logistic regression	0.803179	0.670563	0.753022	0.604380
0	Naive Bayes	0.385647	0.518300	0.350141	0.997228

Mô Hình Random Forest Cho Hiệu Suất Tốt Nhất Trong Số Các Mô Hình

Predict

Step by Step



Data Cleaning

Data Preprocessing

Model

Một số cách để cải thiện tỷ lệ cancelations

Chiến lược giá cả: Với việc phòng loại 1 được đặt nhiều nhất và giá trung bình xung quanh 100 euro, bạn có thể xem xét tăng giá cho loại phòng này để tối đa hóa lợi nhuận. Đồng thời, bạn cũng nên duy trì một số lượng phòng có giá thấp hơn để thu hút khách hàng giá trị và duy trì sự cạnh tranh.

Quản lý hủy đặt phòng: Với phát hiện rằng có sự tương quan giữa giá phòng và số ngày từ ngày đặt phòng đến ngày đến, bạn có thể áp dụng các chính sách hủy linh hoạt. Nếu khách hàng đặt phòng loại phòng đắt đỏ hoặc đặt trước lâu hơn, bạn có thể áp dụng các điều khoản hủy phù hợp để giảm thiểu việc hủy đặt phòng và đảm bảo thu nhập ổn định.

Tăng cường tiếp thị cho gói bữa ăn: Với phát hiện rằng hầu như không ai chọn gói bữa ăn 2 hoặc 3, bạn có thể tăng cường tiếp thị và quảng bá cho những lợi ích của các gói bữa ăn này. Cung cấp ưu đãi đặc biệt hoặc mở rộng lựa chọn bữa ăn có thể thu hút khách hàng và tăng doanh thu từ dịch vụ ăn uống.

Hướng đến cặp đôi không có con: Với giả định rằng hầu hết người đến là các cặp M+W không có con, bạn có thể tạo ra các gói dịch vụ và trải nghiệm hướng đến đối tượng này. Cung cấp các gói ưu đãi lãng mạn hoặc các hoạt động giải trí dành riêng cho cặp đôi có thể thu hút và tạo ra trải nghiệm đáng nhớ cho khách hàng.

Thank You For Listening!

Read More In Notebook