**Project Proposal**

**Hotel Analysis**

Tên đề tài: Phân tích hành vi đặt phòng khách sạn và các yếu tố ảnh hưởng đến việc huỷ đặt phòng

**Mục đích và kết quả:**

**Mục đích:** Phân tích dữ liệu đặt phòng khách sạn để hiểu các yếu tố nào khiến khách huỷ phòng.  
Khám phá mối quan hệ giữa các biến như thời gian đặt, loại khách, loại phòng, kênh đặt phòng, quốc tịch,… với hành vi huỷ

**Kết quả:**

Hiểu rõ yếu tố nào làm tăng xác suất huỷ phòng

(Tuỳ chọn) Mô hình dự đoán hỗ trợ ra quyết định đặt cọc hoặc ưu tiên khách hàng

Biết được nguyên nhân hủy đặt phòng, qua đó giúp cho doanh nghiệp giảm thiểu tỉ lệ hủy đặt phòng.

**Nguồn dữ liệu:**

Link data set: <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>

Mô tả: Bộ dữ liệu này chứa thông tin đặt phòng cho khách sạn thành phố và khách sạn nghỉ dưỡng, bao gồm thông tin như thời điểm đặt phòng, thời gian lưu trú, số người lớn, trẻ em và trẻ sơ sinh, số chỗ đậu xe có sẵn, cùng nhiều thông tin khác.

**Cấu trúc dữ liệu:**

Bao gồm 32 cột:

|  |  |
| --- | --- |
| Cột | Mô tả |
| hotel | Loại khách sạn |
| Is cancled | Đặt phòng bị hủy hay không (1: hủy, 0: không) |
| Lead\_time | Số ngày từ khi đặt đến ngày đến khách sạn |
| arrival\_date\_year | Năm đến khách sạn |
| arrival\_date\_month | Tháng đến khách sạn |
| arrival\_date\_week\_number | Tuần trong năm khi đến |
| arrival\_date\_day\_of\_month | Ngày trong tháng khi đến |
| stays\_in\_weekend\_nights | Số đêm lưu trú vào cuối tuần |
| stays\_in\_week\_nights | Số đêm lưu trú vào ngày thường |
| adults | |  | | --- | |  |  |  | | --- | | Số lượng người lớn trong đặt phòng | |
| children | |  | | --- | |  |  |  | | --- | | Số trẻ em đi cùng | |
| babies | |  | | --- | | Số trẻ sơ sinh đi cùng | |
| meal | Loại bữa ăn được đặt kèm |
| country | Quốc gia của khách hàng |
| market\_segment | Kênh thị trường |
| distribution\_channel | Kênh phân phối đặt phòng |
| is\_repeated\_guest | Khách có phải là khách quen không (1: có, 0: không) |
| previous\_cancellations | Số lần hủy đặt phòng trước đây của khách |
| previous\_bookings\_not\_canceled | Số lần đặt phòng trước đó không bị hủy |
| reserved\_room\_type | Loại phòng đã đặt ban đầu |
| assigned\_room\_type | Loại phòng thực tế được giao |
| booking\_changes | Số lần thay đổi thông tin đặt phòng |
| deposit\_type | Loại tiền đặt cọc |
| agent | ID của đại lý đặt phòng |
| company | ID của công ty đặt phòng |
| days\_in\_waiting\_list | Số ngày trong danh sách chờ |
| customer\_type | Loại khách hàng |
| adr | Giá phòng trung bình mỗi đêm (Average Daily Rate) |
| required\_car\_parking\_spaces | Số chỗ đậu xe khách yêu cầu |
| total\_of\_special\_requests | Tổng số yêu cầu đặc biệt của khách |
| reservation\_status | Trạng thái đặt phòng |
| reservation\_status\_date | Ngày cập nhật trạng thái đặt phòng |

**Kế Hoạch Phân Tích:**

**Giai đoạn 1: Chuẩn bị và làm sạch dữ liệu**

Xử lý missing values (thiếu quốc tịch, agent…)

Chuyển đổi định dạng thời gian và phân loại dữ liệu

**Giai đoạn 2: Khám phá dữ liệu (EDA)**

Tính tỷ lệ huỷ phòng chung và theo nhóm (theo quốc gia, kênh, tháng)

Phân tích mối quan hệ giữa huỷ phòng và các biến như:

lead\_time, deposit\_type, customer\_type, booking\_changes, etc.

**Giai đoạn 3: Mô hình hóa**

Áp dụng **Logistic Regression**, **Random Forest**, hoặc **XGBoost** để dự đoán huỷ đặt phòng (is\_canceled = 1).

Đánh giá độ chính xác bằng **confusion matrix, ROC-AUC, F1-score**.

**Trực quan hóa dữ liệu:**

Pie chart: Tỷ lệ hủy đặt phòng tổng thể

Clustered column chart: Mối quan hệ giữa lead\_time và hủy đặt phòng

Clustered column chart: Tỷ lệ hủy theo tháng

Bar chart: Tỷ lệ hủy theo quốc gia

Stacked bar chart: Loại phòng đặt và loại phòng được giao

100% stacked column chart: Tỷ lệ hủy theo loại đặt cọc

Ma trận tương quan

Clustered column: Trẻ em, em bé ảnh hưởng đến huỷ

Histogram: Tỷ lệ hủy theo thời gian lưu trú

**Giai đoạn 4: Trình bày kết quả**

Tóm tắt các yếu tố ảnh hưởng rõ rệt đến việc huỷ phòng

Đưa ra mô hình dự đoán tỉ lệ cancel tốt hơn

Đề xuất giải pháp cho khách sạn (ví dụ: yêu cầu đặt cọc với nhóm có lead\_time cao)

**Lộ trình thực hiện**  
1. Thu thập và làm sạch dữ liệu bằng Python.  
2. Phân tích dữ liệu khám phá (EDA) và trực quan hóa bằng Python  
3. Xây dựng mô hình hồi quy tuyến tính đa biến.  
4. Phân tích thống kê và rút ra insight.  
5. Hoàn thiện báo cáo cuối cùng và chuẩn bị thuyết trình.

**Thuyết trình**  
- Trình bày kết quả phân tích và các insight quan trọng từ dữ liệu.