HOTEL BOOKING ANALYSIS USING PYSPARK

[I.GIỚI THIỆU BÀI TOÁN 2](#_Toc163145513)

[II. NGHIÊN CỨU LIÊN QUAN 2](#_Toc163145514)

[III. PHƯƠNG PHÁP 3](#_Toc163145515)

[3.1 Các kỹ thuật nền tảng 3](#_Toc163145516)

[3.2 Đề xuất phương pháp 4](#_Toc163145517)

[*3.2.1* *Xây dựng mô hình Decision Tree để phân tích đặt phòng khách sạn* 4](#_Toc163145518)

[*3.2.2* *Xây dựng mô hình Random Forest để phân tích đặt phòng khách sạn* 5](#_Toc163145519)

[*3.2.3* *Xây dựng mô hình XGBoost để phân tích đặt phòng khách sạn* 5](#_Toc163145520)

[IV. THỰC NGHIỆM 6](#_Toc163145521)

[4.1 Miêu tả dữ liệu 6](#_Toc163145522)

[4.2 Tiền xử lý dữ liệu 8](#_Toc163145523)

[4.3 Các độ đo đánh giá hiệu năng 8](#_Toc163145524)

[4.4 Các tham số và môi trường cài đặt 9](#_Toc163145525)

[*4.4.1* *Tham số* 9](#_Toc163145526)

[*4.4.2* *Môi trường cài đặt* 9](#_Toc163145527)

[4.5 Các phương pháp cơ sở 9](#_Toc163145528)

[4.6 Phân tích so sánh các kết quả 9](#_Toc163145529)

[*4.6.1* *Decision Trees* 9](#_Toc163145530)

[*4.6.2* *Random Forest* 11](#_Toc163145531)

[*4.6.3* *XGBoost* 13](#_Toc163145532)

[V. KẾT LUẬN 14](#_Toc163145533)

# 

# I.GIỚI THIỆU BÀI TOÁN

Trong ngành du lịch và khách sạn, quản lý đặt phòng là một trong những vấn đề quan trọng nhất. Khi tiến hành đặt phòng khách sạn, khách hàng  không chỉ cần xem xét yếu tố giá cả mà còn phải quan tâm đến thời điểm đặt phòng và thời gian lưu trú để có trải nghiệm tốt nhất. Một câu hỏi thường gặp là khi nào là thời điểm tốt nhất trong năm để đặt phòng khách sạn? Thực tế, giá và sự sẵn có của phòng thường biến đổi theo mùa, theo thời tiết hay sự kiện đặc biệt nào đó, như lễ Giáng sinh, Tết Nguyên đán hoặc triển lãm. Việc phân tích thời gian lưu trú của khách hàng cũng là một phần quan trọng của quản lý đặt phòng. Các khách hàng có thể muốn đặt phòng vào những ngày trong tuần hoặc tháng trong năm khi giá cả thấp nhất hoặc khi thời tiết lý tưởng. Bằng cách nắm bắt được các xu hướng và biến động trong ngành, sử dụng phương pháp dự báo và phân tích dữ liệu, các hệ thống quản lý phòng tích hợp công nghệ nhân tạo và học máy để ước tính chính xác về tần suất và lượng đặt phòng,  tỷ lệ huỷ phòng trong tương lai, giúp các nhà quản lý khách sạn có thể tối ưu hóa lịch đặt phòng và tăng cường doanh thu. Có nhiều công cụ và phương pháp để sử dụng phân tích dữ liệu quản lý đặt phòng khách sạn, nhưng  trong  khuôn khổ bài nghiên cứu này, nhóm chúng tôi sử dụng PySpark để phân tích dữ liệu lớn và xử lý dữ liệu phân tán, kết hợp sức mạnh của Apache Spark với linh hoạt và tiện lợi của ngôn ngữ lập trình Python để trả lời những câu hỏi trên và giúp các nhà quản lý khách sạn đưa ra dự đoán về nhu cầu đặt phòng của khách hàng.

# II. NGHIÊN CỨU LIÊN QUAN

1. Nghiên cứu “Hệ thống dự đoán giá vé máy bay và giá khách sạn sử dụng thuật toán học máy tổng hợp” của 4 tác giả Tejal Dimble, Nikita Pandey, Harshada Narkhede, Ruturaj More Students, tập trung vào việc sử dụng dữ liệu vé máy bay và phát triển các mô hình dự đoán thay đổi giá vé theo thời gian, nhằm giúp người tiêu dùng tiết kiệm chi phí. Nghiên cứu cũng mở rộng sang dự đoán giá phòng khách sạn, sử dụng các mô hình thống kê và học máy. Dữ liệu từ Smith Travel Research được sử dụng và áp dụng các mô hình nâng cao như ANFIS. Nghiên cứu này mô tả hai phương pháp khác nhau được áp dụng để mô hình hóa sự thay đổi giá, khi tiến hành phân tích so sánh các thuật toán theo hai phương pháp này, được áp dụng trên dữ liệu thực trên thế giới đã được thực hiện, tìm ra thuật toán hiệu quả nhất để dự đoán giá vé máy bay và khách sạn. Kết quả cho thấy việc sử dụng dữ liệu lịch sử và mô hình hóa có thể giúp dự đoán xu hướng giá và tiết kiệm chi phí cho người tiêu dùng.
2. Nghiên cứu “So sánh và phân tích các mô hình học máy để dự đoán việc hủy đặt phòng khách sạn” của 4 tác giả Yiying Chen, Chuhan Ding, Hanjie Ye, Yuchen Zhou sử dụng dữ liệu thông tin đặt phòng từ cả hai loại khách sạn( khách sạn resort và khách sạn bình dân) ở Bồ Đào Nha cùng với thông tin tương ứng về khách hàng để dự đoán khả năng hủy đặt phòng của khách hàng. Có ba phương pháp dự đoán khác nhau được đề xuất, bao gồm hồi quy logistic, k -Nearest Neighbor (k - NN) và CatBoost, kết quả nghiên cứu chỉ ra rằng CatBoost là phương pháp phù hợp nhất cho mục đích này. CatBoost được chọn vì hiệu quả, độ chính xác cao và chi phí thấp hơn so với các phương pháp khác. Bằng việc tập trung vào hiệu quả và tiết kiệm trong việc dự đoán khả năng hủy đặt phòng, kết quả bài nghiên cứu này tạo cơ sở cho việc quản lý doanh thu và tài nguyên trong ngành khách sạn trong tương lai.
3. Nghiên cứu “ Dự đoán lòng trung thành của khách hàng cho ngành khách sạn bằng phương pháp học máy” của 5 tác giả Iskandar Zul Putera Hamdan, Muhaini Othman, Yana Mazwin Mohmad Hassim, S. Marjudi, Munirah Mohd Yusof, dự án nghiên cứu sử dụng học máy để dự đoán lòng trung thành của khách hàng khách sạn. Ba thuật toán được chọn là rừng ngẫu nhiên, hồi quy logistic và cây quyết định. Các thuật toán được phát triển bằng ngôn ngữ lập trình Python và kết quả được đánh giá bằng các số liệu như Ma trận nhầm lẫn, độ chính xác, thu hồi và điểm F1. Kết quả cuối cùng cho thấy rằng hồi quy logistic, cây quyết định và rừng ngẫu nhiên đạt được độ chính xác lần lượt là 57,83%, 71,44% và 69,91%.

# III. PHƯƠNG PHÁP

* 1. Các kỹ thuật nền tảng

Trong lĩnh vực phân tích đặt phòng khách sạn, các nền tảng kỹ thuật khác nhau đóng vai trò then chốt trong việc thu thập, xử lý và phân tích dữ liệu để rút ra những hiểu biết sâu sắc có thể hành động. Những nền tảng này bao gồm một loạt các công nghệ và công cụ được thiết kế để xử lý sự phức tạp của dữ liệu đặt phòng khách sạn và hỗ trợ các chủ khách sạn và các bên liên quan đưa ra quyết định sáng suốt. Dưới đây, chúng tôi đi sâu vào các nền tảng kỹ thuật quan trọng giúp phân tích đặt phòng khách sạn:

* Hệ thống quản lý tài sản khách sạn (PMS):

+ Các nền tảng PMS như Opera, Amadeus và Fidelio đóng vai trò là trung tâm tập trung để quản lý hoạt động của khách sạn, bao gồm dữ liệu đặt phòng, hồ sơ khách và quản lý doanh thu. Chúng cung cấp những hiểu biết có giá trị về tỷ lệ lấp đầy, hình thức đặt phòng và dự báo doanh thu.

* Đại lý du lịch trực tuyến (OTA):

+ Các OTA hàng đầu như Expedia, Booking.com và agoda cung cấp kho dữ liệu đặt phòng khổng lồ, đánh giá của khách và chi tiết về giá cả. Việc truy cập dữ liệu OTA cho phép chủ khách sạn theo dõi xu hướng thị trường, đánh giá hiệu suất của đối thủ cạnh tranh và tối ưu hóa chiến lược giá.

* Hệ thống quản lý quan hệ khách hàng (CRM):

+ Các nền tảng CRM như Salesforce, HubSpot và Zoho tạo điều kiện thuận lợi cho việc tương tác và giữ chân khách hàng bằng cách lưu trữ liên hệ, sở thích và tương tác của khách. Bằng cách tận dụng dữ liệu CRM, khách sạn có thể cá nhân hóa các chiến dịch tiếp thị, hợp lý hóa hoạt động liên lạc của khách và thúc đẩy lòng trung thành.

* Công cụ xử lý và phân tích dữ liệu:

+ Các công cụ như Python (với các thư viện như Pandas, NumPy và Scikit-learn), Lập trình R và Apache Spark cho phép các chủ khách sạn xử lý trước, phân tích và trực quan hóa dữ liệu đặt phòng. Từ phân tích dữ liệu thăm dò đến mô hình dự đoán, những công cụ này hỗ trợ việc ra quyết định dựa trên dữ liệu trong quản lý doanh thu của khách sạn.

* Kỹ thuật học máy và học sâu:

+ Các thuật toán học máy bao gồm hồi quy, phân loại và phân cụm được áp dụng để dự đoán hành vi đặt phòng, phân khúc hồ sơ khách hàng và tối ưu hóa giá phòng. Các mô hình học sâu như mạng lưới thần kinh nâng cao độ chính xác trong dự đoán và khám phá các mẫu phức tạp trong dữ liệu đặt chỗ.

* Công cụ trực quan hóa dữ liệu:

+ Các công cụ trực quan hóa như Tableau, Power BI và Matplotlib (bằng Python) hỗ trợ trình bày dữ liệu đặt chỗ thông qua bảng điều khiển, biểu đồ và đồ thị tương tác. Trình bày trực quan cho phép chủ khách sạn xác định nhanh các xu hướng, mô hình và sự bất thường trong dữ liệu đặt phòng.

Những nền tảng kỹ thuật này tạo thành nền tảng phân tích đặt phòng khách sạn, trao quyền cho chủ khách sạn những hiểu biết sâu sắc có thể hành động để tối ưu hóa doanh thu, nâng cao trải nghiệm của khách và duy trì tính cạnh tranh trong ngành khách sạn năng động. Bằng cách khai thác sức mạnh của dữ liệu và phân tích nâng cao, các khách sạn có thể thích ứng với xu hướng thị trường đang phát triển, dự đoán nhu cầu của khách và thúc đẩy tăng trưởng kinh doanh một cách hiệu quả.

* 1. Đề xuất phương pháp
     1. *Xây dựng mô hình Decision Tree để phân tích đặt phòng khách sạn*

1. Giới thiệu

* Trong phần này, chúng tôi đề xuất sử dụng thuật toán Decision Tree để dự đoán mô hình đặt phòng khách sạn. Decision Tree cung cấp một cách tiếp cận trực quan để hiểu các mẫu dữ liệu, khiến chúng phù hợp để phân tích các quy trình ra quyết định phức tạp như đặt phòng khách sạn.

1. Phương pháp luận

* Xử lý trước dữ liệu: Chuẩn bị tập dữ liệu bằng cách mã hóa các biến phân loại, xử lý các giá trị bị thiếu và chia tỷ lệ các đặc điểm số nếu cần. Điều này đảm bảo dữ liệu phù hợp để đào tạo mô hình Decision Tree.
* Đào tạo mô hình: Huấn luyện mô hình Decision Tree trên tập dữ liệu được xử lý trước. Mô hình sẽ tìm hiểu các mẫu trong dữ liệu để dự đoán liệu việc đặt phòng khách sạn có bị hủy hay không dựa trên các đặc điểm khác nhau như ngày đặt phòng, thời gian lưu trú và nhân khẩu học của khách hàng.
* Đánh giá mô hình: Đánh giá mô hình Decision Tree đã được đào tạo bằng cách sử dụng các số liệu thích hợp như độ chính xác, độ chính xác, khả năng thu hồi và điểm F1. Bước này đảm bảo hiệu suất của mô hình và xác định mọi vấn đề tiềm ẩn như trang bị quá mức.
* Điều chỉnh siêu tham số: Tinh chỉnh các siêu tham số của mô hình Decision Tree bằng cách sử dụng các kỹ thuật như tìm kiếm dạng lưới hoặc tìm kiếm ngẫu nhiên để cải thiện hiệu suất của nó hơn nữa.

1. Kết quả dự kiến

* Bằng cách sử dụng thuật toán Decision Tree để phân tích đặt phòng khách sạn, chúng tôi dự đoán sẽ xác định được các yếu tố chính, ảnh hưởng đến việc hủy và không hủy đặt phòng. Mô hình Decision Tree sẽ cung cấp thông tin chuyên sâu về hành vi của khách hàng, giúp ban quản lý khách sạn tối ưu hóa chiến lược đặt phòng và giảm thiểu tình trạng hủy đặt phòng.
  + 1. *Xây dựng mô hình Random Forest để phân tích đặt phòng khách sạn*

1. Giới thiệu

* Random Forest  là một kỹ thuật học tập tổng hợp kết hợp nhiều Decision Tree để nâng cao độ chính xác của dự đoán. Trong bối cảnh phân tích đặt phòng khách sạn, Random Forest mang lại hiệu suất vượt trội bằng cách giảm thiểu tình trạng trang bị quá mức và nắm bắt các mối quan hệ dữ liệu phức tạp.

1. Phương pháp luận

* Xử lý trước dữ liệu: Tương tự như phương pháp Decision Tree, xử lý trước tập dữ liệu đặt phòng khách sạn để đảm bảo nó phù hợp cho việc đào tạo mô hình Random Forest.
* Đào tạo mô hình: Huấn luyện mô hình Random Forest trên tập dữ liệu được xử lý trước. Nhóm Decision Tree sẽ phân tích chung dữ liệu, đưa ra những dự đoán chắc chắn về kết quả đặt phòng khách sạn.
* Đánh giá mô hình: Đánh giá hiệu suất của mô hình Random Forest bằng cách sử dụng các số liệu đánh giá tiêu chuẩn. So sánh kết quả với mô hình Decision Tree để đánh giá sự cải thiện về độ chính xác dự đoán.
* Điều chỉnh siêu tham số: Tinh chỉnh các siêu tham số của mô hình Random Forest để tối ưu hóa hiệu suất của nó hơn nữa. Các kỹ thuật như xác thực chéo có thể được sử dụng cho mục đích này.

1. Kết quả dự kiến

* Mô hình  Random Forest dự kiến sẽ hoạt động tốt hơn mô hình Decision Tree bằng cách tận dụng phương pháp học tập tổng thể. Nó sẽ đưa ra những dự đoán chính xác hơn về việc hủy đặt phòng khách sạn, cho phép ban quản lý khách sạn đưa ra quyết định sáng suốt và cải thiện sự hài lòng của khách hàng.
  + 1. *Xây dựng mô hình XGBoost để phân tích đặt phòng khách sạn*

1. Giới thiệu

* XGBoost (Tăng cường độ dốc cực cao) là một thuật toán tăng cường độ dốc mạnh mẽ được biết đến với hiệu suất vượt trội trong các tác vụ lập mô hình dự đoán. Bằng cách xây dựng mô hình XGBoost để phân tích đặt phòng khách sạn, chúng tôi mong muốn tận dụng khả năng của nó để dự đoán chính xác kết quả đặt phòng.

1. Phương pháp luận

* Xử lý trước dữ liệu: Xử lý trước tập dữ liệu đặt phòng khách sạn tương tự như các phương pháp trước đó để đảm bảo khả năng tương thích với mô hình XGBoost.
* Đào tạo mô hình: Đào tạo mô hình XGBoost trên tập dữ liệu được xử lý trước. Khung tăng cường độ dốc của XGBoost sẽ liên tục cải thiện hiệu suất của mô hình bằng cách giảm thiểu lỗi dự đoán.
* Đánh giá mô hình: Đánh giá hiệu suất của mô hình XGBoost bằng cách sử dụng các số liệu đánh giá tiêu chuẩn và so sánh nó với các mô hình Decision Tree và Random forest.
* Điều chỉnh siêu tham số: Tinh chỉnh các siêu tham số của mô hình XGBoost bằng cách sử dụng các kỹ thuật như tìm kiếm lưới hoặc tối ưu hóa Bayesian để nâng cao độ chính xác dự đoán của nó hơn nữa.

1. Kết quả dự kiến

* Chúng tôi dự đoán mô hình XGBoost sẽ thể hiện hiệu suất dự đoán vượt trội so với cả mô hình Decision Tree và Random forest. Khả năng xử lý các mối quan hệ dữ liệu phức tạp và tối ưu hóa độ chính xác dự đoán khiến nó trở thành lựa chọn lý tưởng để phân tích đặt phòng khách sạn, giúp đưa ra quyết định tốt hơn và mang lại sự hài lòng cho khách hàng.

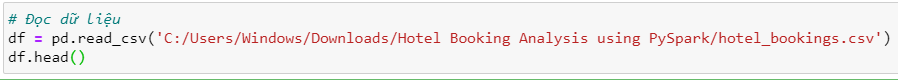
Các phương pháp được đề xuất này nhằm mục đích tận dụng các thuật toán học máy tiên tiến để phân tích dữ liệu đặt phòng khách sạn một cách hiệu quả, cung cấp những hiểu biết có giá trị để tối ưu hóa chiến lược đặt phòng và nâng cao trải nghiệm của khách hàng.

IV. THỰC NGHIỆM

* 1. Miêu tả dữ liệu
* Ngữ cảnh: Tập dữ liệu này cung cấp bức tranh chi tiết về từng hồ sơ đặt phòng, cho phép phân tích hành vi của khách, hình thức đặt phòng và hiệu suất của khách sạn.
* Mục tiêu: Để phân tích hành vi đặt phòng khách sạn và hiểu rõ hơn về các yếu tố ảnh hưởng đến việc đặt phòng và quản lý các đặt phòng trong ngành công nghiệp lưu trú.
* Bộ dữ liệu bao gồm thông tin:
  + Tên khách sạn.
  + Trạng thái phòng đã đặt đã bị hủy bỏ hay không được miêu tả tại cột is\_canceled.
  + Thời gian từ lúc đặt phòng cho đến ngày đến, thông tin về ngày đến.
  + Số đêm lưu trú vào cuối tuần và trong tuần.
  + Thông tin về loại phòng, thông tin về đặt cọc, thông tin về yêu cầu đặc biệt, thông tin về giá đặt phòng trung bình mỗi đêm
  + Thông tin về khách hàng: Số lượng người lớn, trẻ em, em bé,....
  + Thông tin về trạng thái đặt phòng: Trạng thái hiện tại của đặt phòng, bao gồm có thể là "Confirmed" (Xác nhận), "Canceled" (Hủy bỏ), hoặc "Check-Out" (Ra khỏi khách sạn).

|  |  |  |
| --- | --- | --- |
| Trường dữ liệu | Mô tả | Kiểu dữ liệu |
| Hotel | Tên khách sạn | String |
| Is\_canceled | Chỉ số cho biết liệu đặt phòng có bị hủy bỏ hay không | Integer |
| Lead\_time | Thời gian từ thời điểm đặt phòng cho đến ngày đến nơi | Integer |
| Arrival\_date\_year | Năm của ngày đến | Integer |
| Arrival\_date\_month | Tháng của ngày đến | String |
| Arrival\_date\_week\_number | Số tuần của ngày đến trong năm | Integer |
| Arrival\_date\_day\_of\_month | Ngày của ngày đến trong tháng | Integer |
| Stays\_in\_weekend\_nights | Số đêm lưu trú vào cuối tuần | Integer |
| Stays\_in\_week\_nights | Số đêm lưu trú vào các ngày trong tuần | Integer |
| Adults | Số lượng người lớn | Integer |
| Children | Số lượng trẻ em | String |
| Babies | Số lượng em bé | Integer |
| Meal | Loại bữa ăn | String |
| Country | Quốc gia của khách hàng | String |
| Market\_segment | Đoạn thị trường | String |
| Distribution\_channel | Kênh phân phối | String |
| Is\_repeated\_guest | Chỉ số cho biết khách hàng có là khách quen hay không | Integer |
| Previous\_cancellations | Số lần hủy bỏ đặt phòng trước đó | Integer |
| Assigned\_room\_type | Loại phòng được gán | Integer |
| Booking\_changes | Số lần thay đổi đặt phòng | String |
| Deposit\_type | Loại tiền đặt cọc | String |
| Agent | Mã đại lý đặt phòng | Integer |
| Company | Mã công ty đặt phòng | String |
| Days\_in\_waiting\_list | Số ngày trong danh sách chờ | String |

* 1. Tiền xử lý dữ liệu
* Đọc dữ liệu:



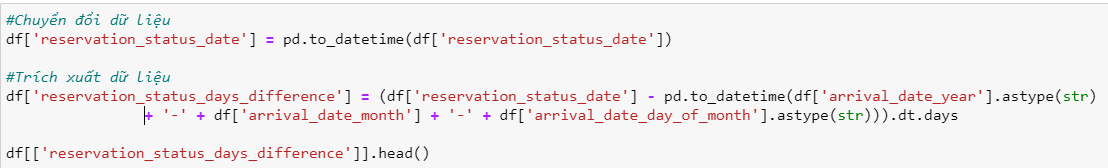
* Thông tin dữ liệu và thống kê:



* Tỷ lệ hủy đặt phòng:



* Chuyển đổi dữ liệu:



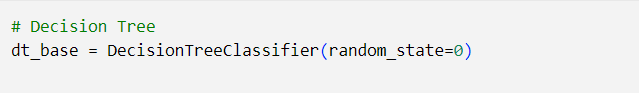
* 1. Các độ đo đánh giá hiệu năng

Trong nghiên cứu này chúng tôi sử dụng bốn chỉ số bao gồm accuracy, precision, recall và F1 score để đánh giá mô hình.

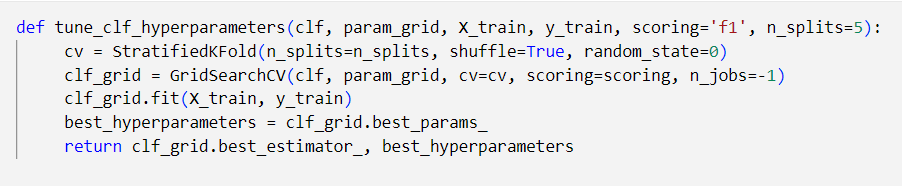
* Precision là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu trên tổng dự đoán khớp mẫu và dự đoán đúng nhưng sai.
* Recall là tỷ lệ giữa số lượng mẫu được dự đoán khớp mẫu với tổng dự đoán khớp mẫu và dự đoán sai nhưng đúng.
* Accuracy là độ chính xác của mô hình, là tỷ lệ giữa số lượng mẫu được phân loại chính xác trên cho tổng số mẫu của tập dữ liệu huấn luyện.
* F1 score là trung bình trọng số của Precision và Recall.

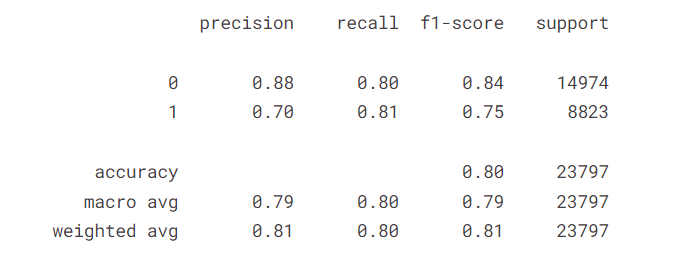
Trong đó:

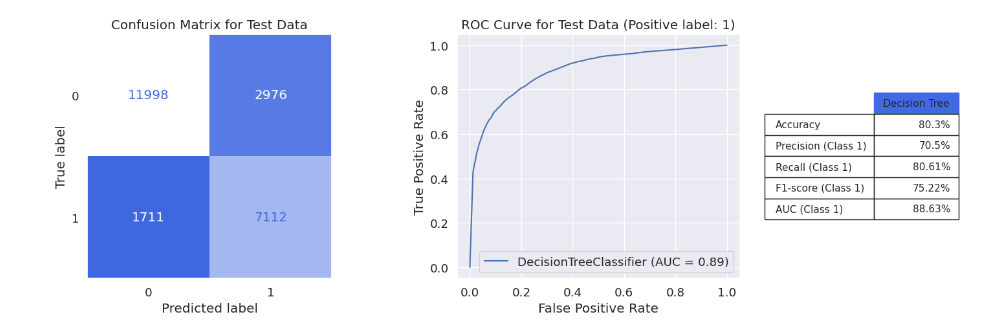
* + TP (True Positive): tổng số trường hợp dự báo khớp mẫu đúng.
  + TN (True Negative): tổng số trường hợp dự báo khớp mẫu sai.
  + FP (False Positive): tổng số trường hợp dự báo các quan sát thuộc mẫu đúng nhưng tính thành sai.
  + FN (False Negative): tổng số trường hợp dự báo các quan sát thuộc mẫu sai nhưng tính thành đúng.
  1. Các tham số và môi trường cài đặt
     1. *Tham số*
* Sử dụng tham số parameters , model, criterion, random\_state, max\_depth, min\_samples\_leaf.
* Siêu tham số hyperparameter.
* Tham số n\_estimators, learning\_rate, subsample.
  + 1. *Môi trường cài đặt*
* Ngôn ngữ lập trình Python: Python là ngôn ngữ lập trình phổ biến cho việc xây dựng và huấn luyện các mô hình học máy. Sử dụng các thư viện như scikit-learn, XGBoost, và LightGBM để triển khai các thuật toán phân loại và tinh chỉnh mô hình.
* Công Cụ Phát Triển: Sử dụng các công cụ phát triển tích hợp (IDEs) như Jupyter Notebook, Google Colab, hoặc PyCharm để viết, chạy và debug code
* Thư viện scikit-learn: cung cấp công cụ xây dựng và huấn luyện mô hình.
* Thư viện pandas: cung cấp cấu trúc dữ liệu và công cụ phân tích dữ liệu
* Thư viện numpy: hỗ trợ các phép tính số học.
* Thư viện matplotlib và seaborn: trực quan hóa dữ liệu bằng các biểu đồ.
* Xgboost và lightgbm: xử lý dữ liệu
  1. Các phương pháp cơ sở
* Mô hình học máy :
  + Decision Trees
  + Random Forest
  + XGBoost
  1. Phân tích so sánh các kết quả
     1. *Decision Trees*
* Định nghĩa mô hình Decision Trees



* Điều chỉnh siêu tham số







*Ma trận nhầm lẫn và đường cong ROC cho thử nghiệm*

Suy luận:

Hiệu suất của mô hình DT của chúng tôi trên dữ liệu thử nghiệm như sau:

Độ chính xác là 80,3%, có nghĩa là mô hình dự đoán chính xác liệu lượt đặt chỗ có bị hủy hay không trong khoảng 80% thời gian.

Độ chính xác cho loại 1 (hủy) là 70,5%, nghĩa là 71% số lượt đặt chỗ mà mô hình dự đoán là đã hủy đã thực sự bị hủy.

Tỷ lệ thu hồi cho loại 1 là 80,61%, có nghĩa là mô hình đã xác định chính xác 81% số lần hủy thực tế.

Điểm F1 cho loại 1 là 75,22%, là giá trị trung bình hài hòa của Độ chính xác và Thu hồi.

AUC (Diện tích dưới đường cong ROC) cho lớp 1 là 88,63%, có nghĩa là mô hình có khả năng phân biệt đối xử tốt giữa hai lớp.

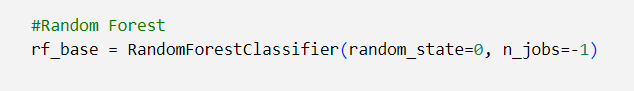
Ma trận nhầm lẫn cho thấy vẫn còn một số Kết quả dương tính giả và m tính giả, nhưng mô hình đang thực hiện tương đối tốt việc giảm thiểu chúng.

Ngoài ra, mô hình không bị trang bị quá mức vì các giá trị số liệu cho tập hợp thử nghiệm và tập huấn luyện gần nhau, cho thấy rằng mô hình đang khái quát hóa tốt dữ liệu không nhìn thấy được.

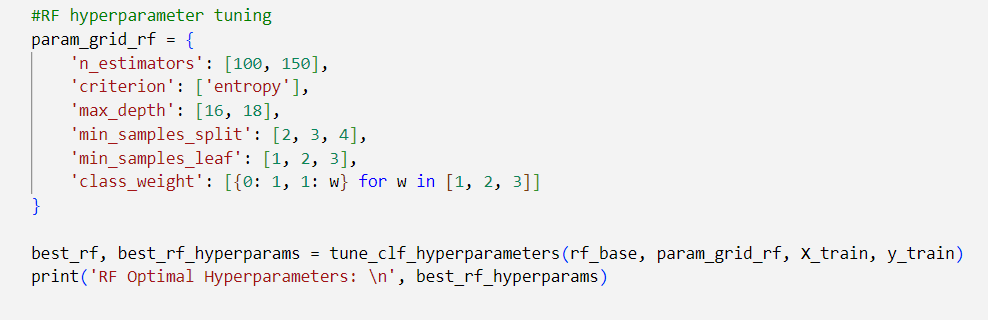
Nhìn chung, mô hình này đang hoạt động tốt trên dữ liệu thử nghiệm và dường như là một mô hình tốt để dự đoán số lần hủy, mặc dù vẫn còn chỗ để cải thiện.

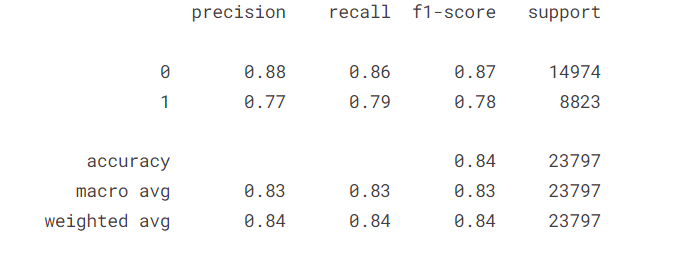
* + 1. *Random Forest*

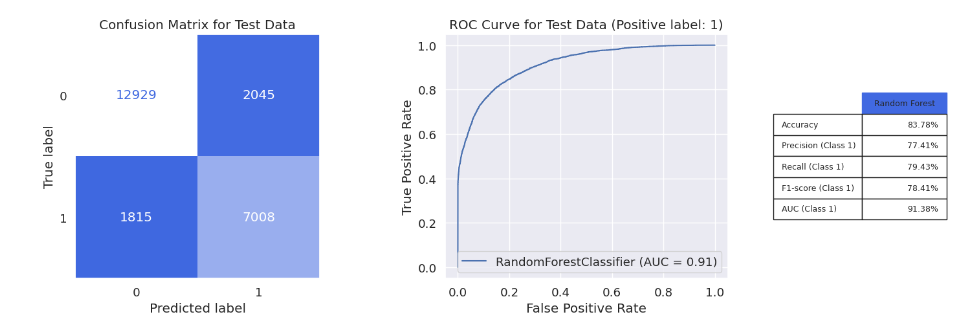
- Định nghĩa mô hình Random Forest:



* Điều chỉnh siêu tham số:







Suy luận:

Hiệu suất của mô hình Random Forest (RF) của chúng tôi trên dữ liệu thử nghiệm như sau:

Độ chính xác là 83,78%, có nghĩa là mô hình dự đoán chính xác liệu lượt đặt chỗ có bị hủy hay không trong khoảng 84% thời gian.

Độ chính xác cho loại 1 (hủy) là 77,41%, nghĩa là khoảng 78% số lượt đặt chỗ mà mô hình dự đoán là đã hủy đã thực sự bị hủy.

Tỷ lệ thu hồi cho loại 1 là 79,43%, có nghĩa là mô hình đã xác định chính xác khoảng 80% số lần hủy thực tế.

Điểm F1 cho loại 1 là 78,41%, là giá trị trung bình hài hòa của Độ chính xác và Thu hồi.

AUC (Diện tích dưới đường cong ROC) cho lớp 1 là 91,38%, có nghĩa là mô hình có khả năng phân biệt đối xử rất tốt giữa hai lớp.

Ma trận nhầm lẫn cho thấy có một số Kết quả dương tính giả và m tính giả, nhưng mô hình đang làm rất tốt việc giảm thiểu chúng.

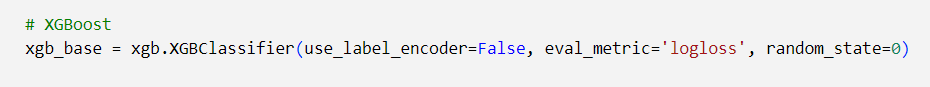
Ngoài ra, mô hình không bị trang bị quá mức vì các giá trị số liệu cho tập hợp thử nghiệm và tập huấn luyện gần nhau, cho thấy rằng mô hình đang khái quát hóa tốt dữ liệu không nhìn thấy được.

So với mô hình Decision Trees (DT), mô hình Random Forest (RF) cho thấy sự cải thiện đáng kể về tất cả các chỉ số hiệu suất, đặc biệt là trong việc giảm số lượng Kết quả dương tính giả. Điều này cho thấy mô hình Rừng ngẫu nhiên tốt hơn trong việc cân bằng sự cân bằng giữa Độ chính xác và Thu hồi, dẫn đến điểm F1 cao hơn.

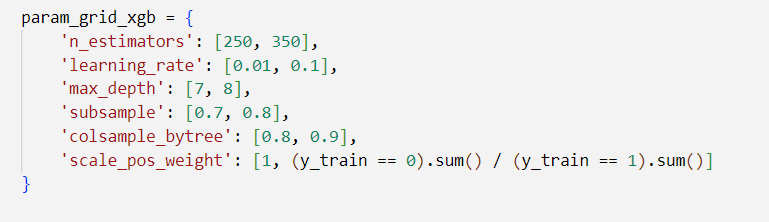
Nhìn chung, mô hình này đang hoạt động tốt trên dữ liệu thử nghiệm và dường như là một mô hình tốt để dự đoán số lần hủy, mặc dù vẫn còn chỗ cần cải thiện, đặc biệt là trong việc giảm số lượng

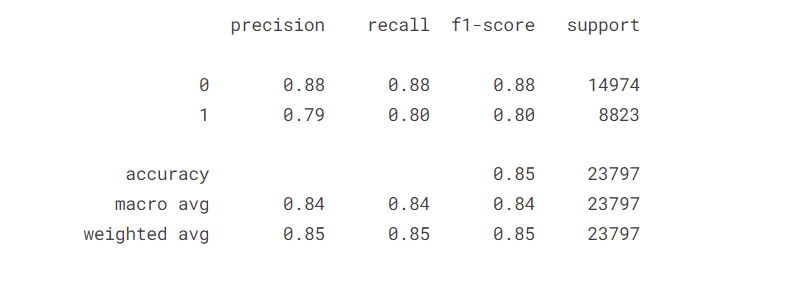
* + 1. *XGBoost*

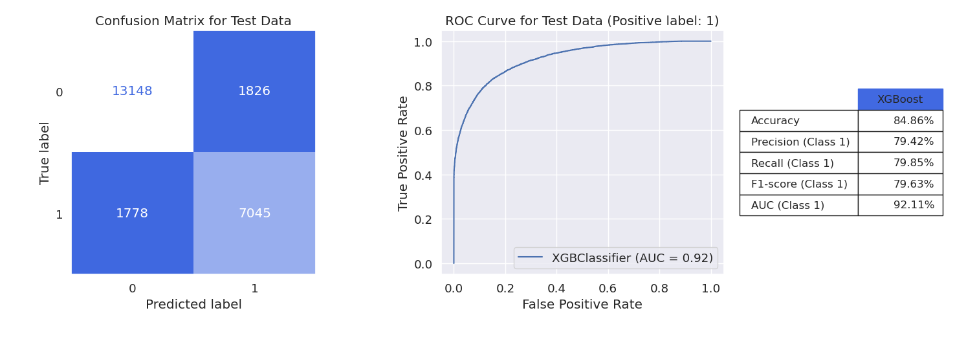
- Định nghĩa mô hình XGBoost:



* Điều chỉnh siêu tham số:







*Ma trận nhầm lẫn và đường cong ROC cho thử nghiệm*

Suy luận:

Hiệu suất của mô hình XGBoost (XGB) của chúng tôi trên dữ liệu thử nghiệm như sau:

Độ chính xác là 84,9%, có nghĩa là mô hình dự đoán chính xác liệu lượt đặt chỗ có bị hủy hay không trong khoảng 85% thời gian.

Độ chính xác cho loại 1 (hủy) là 79,4%, nghĩa là khoảng 80% số lượt đặt chỗ mà mô hình dự đoán là đã hủy đã thực sự bị hủy.

Tỷ lệ thu hồi cho loại 1 là 79,8%, có nghĩa là mô hình đã xác định chính xác khoảng 80% số lần hủy thực tế.

Điểm F1 cho loại 1 là 79,6%, là giá trị trung bình hài hòa của Độ chính xác và Thu hồi.

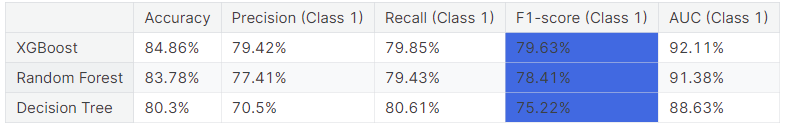
AUC (Diện tích dưới đường cong ROC) cho lớp 1 là 92,1%, có nghĩa là mô hình có khả năng phân biệt đối xử rất tốt giữa hai lớp.

Mô hình không phù hợp quá mức vì các giá trị số liệu cho tập hợp thử nghiệm và tập huấn luyện gần nhau, cho thấy rằng mô hình đang khái quát hóa tốt đối với dữ liệu không nhìn thấy được.

So với mô hình Random Forest (RF), mô hình XGBoost (XGB) cho thấy sự cải thiện một chút về tất cả các chỉ số hiệu suất. Mặc dù sự cải thiện không đáng kể nhưng vẫn tốt hơn trong việc cân bằng sự đánh đổi giữa Precision và Recall, dẫn đến điểm F1 cao hơn. Điều này chỉ ra rằng mô hình XGBoost là mô hình tốt hơn để dự đoán số lần hủy.

Nhìn chung, mô hình này đang hoạt động tốt trên dữ liệu thử nghiệm và dường như là một mô hình tốt để dự đoán số lần hủy.

Kết quả: Dưới đây là bảng so sánh các chỉ số đánh giá hiệu năng giữa 3 mô hình Decision Trees, Random Forest, XGBoost



Như chúng tôi đã giải thích trước đó, số liệu quan trọng nhất cho dự án này là điểm F1 . Điểm F1 cao cho thấy sự cân bằng giữa việc giảm thiểu chi phí liên quan đến kết quả âm tính giả (đặt trước quá nhiều phòng dự kiến việc hủy sẽ không xảy ra) và kết quả dương tính giả (dự đoán số lượt hủy khi khách thực sự đến), điều này rất quan trọng để khách sạn tối ưu hóa lượng phòng trống trong khi giảm thiểu tình trạng đặt trước quá nhiều và khách hàng không hài lòng.

# V. KẾT LUẬN

Trong bối cảnh ngành du lịch đang phát triển mạnh mẽ và sự cạnh tranh ngày càng gay gắt, việc quản lý đặt phòng trong khách sạn đòi hỏi sự chính xác và hiệu quả. Dựa trên nghiên cứu và phân tích sâu về phân tích hủy đặt phòng khách sạn, chúng tôi đã rút ra những kết luận quan trọng và đề xuất giải pháp để tối ưu hóa quản lý đặt phòng.

Trong quá trình nghiên cứu, chúng tôi đã so sánh hiệu suất của ba mô hình phân loại phổ biến: Decision Trees, Random Forest và XGBoost. Chúng tôi nhận thấy rằng điểm F1 là thước đo quan trọng nhất để đánh giá hiệu suất của các mô hình này, bởi nó cân bằng giữa việc dự đoán chính xác số lượng phòng cần đặt và việc tránh dự đoán quá nhiều phòng mà không có hủy.

Kết quả của chúng tôi cho thấy mô hình XGBoost đã đạt được điểm F1 cao nhất, chỉ ra rằng nó là công cụ tốt nhất cho việc dự đoán đặt phòng khách sạn. XGBoost không chỉ có khả năng dự đoán chính xác mà còn có thể tối ưu hóa lợi ích từ việc quản lý phòng trống và tối ưu hóa doanh thu. Ngoài ra, mô hình Random Forest cũng đạt được hiệu suất tốt và có thể được xem xét là một lựa chọn phụ thuộc vào yêu cầu cụ thể của khách hàng và mục tiêu kinh doanh.

Với những phát hiện này, chúng tôi đề xuất các khách sạn nên sử dụng mô hình XGBoost làm công cụ dự đoán chính cho việc quản lý đặt phòng. Đồng thời, cần tiếp tục theo dõi và tối ưu hóa các mô hình này để duy trì và cải thiện hiệu suất trong thời gian dài. Điều này sẽ giúp khách sạn tối ưu hóa lợi ích từ việc quản lý phòng trống, tăng cường sự hài lòng của khách hàng và đạt được sự cạnh tranh cao trong ngành du lịch đang phát triển.