**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**** 🙡**🕮**🙣

**TÊN ĐỀ TÀI: PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN HÀNH VI ĐẶT VÉ MÁY BAY VÀ DỰ ĐOÁN TỶ LỆ THÀNH CÔNG TRONG NGÀNH HÀNG KHÔNG**

**BÀI TẬP LỚN MÔN: PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO**

**TP. HỒ CHÍ MINH, NĂM 2024**

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**** 🙡**🕮**🙣

**TÊN ĐỀ TÀI: PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN HÀNH VI ĐẶT VÉ MÁY BAY VÀ DỰ ĐOÁN TỶ LỆ THÀNH CÔNG TRONG NGÀNH HÀNG KHÔNG**

**BÀI TẬP LỚN MÔN: PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO**

**GVHD: Đinh Nguyễn Trọng Nghĩa**

**TKB chính thức: Thứ 4, tiết 7 - 11**

**Nhóm thực hiện:**

1. 2001215634 - Đặng Hữu Chiến
2. 2001210032 - Trần Thị Ngọc Trân
3. 2033210533 - Cao Minh Tuệ
4. 2001210241 - Thái Anh Quảng

**TP. HỒ CHÍ MINH, NĂM 2024**

# BẢNG ĐÁNH GIÁ KẾT QUẢ THỰC HIỆN CÔNG VIỆC NHÓM

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Họ và tên | Cá nhân tự đánh giá kết quả | Nhóm đánh giá kết quả | GV đánh giá |
| 1 | Đặng Hữu Chiến | Hoàn thành tốt, đúng hạn | Nhiệt tình tương tác với nhóm |  |
| 2 | Trần Thị Ngọc Trân | Hoàn thành tốt, đúng hạn | Nhiệt tình tương tác với nhóm |  |
| 3 | Cao Minh Tuệ | Hoàn thành tốt, đúng hạn | Nhiệt tình tương tác với nhóm |  |
| 4 | Thái Anh Quảng | Hoàn thành tốt, đúng hạn | Nhiệt tình tương tác với nhóm |  |

# LỜI CẢM ƠN

Đề tài được hoàn thành dưới sự hướng dẫn của thầy Đinh Nguyễn Trọng Nghĩa. Chúng em xin được bày tỏ lòng biết ơn chân thành nhất đối với thầy hướng dẫn - người đã nêu ý tưởng, gợi ý đề tài, hướng dẫn tận tình và luôn động viên chúng em thực hiện đề tài trong suốt quá trình nghiên cứu.

Chúng em xin chân thành gửi đến lời cảm ơn đến quý thầy cô khoa Công nghệ thông tin trường Đại học Công Thương TP.HCM đã có những ý kiến đóng góp khoa học và thiết thực cho nội dung đề tài. Đồng thời, các thầy cô của khoa đã luôn tạo điều kiện thuận lợi nhất trong suốt thời gian chúng em học tập và nghiên cứu tại trường.

Chúng em vô cùng biết ơn và ghi nhớ những giúp đỡ, tạo điều kiện của Ban giám hiệu trường Đại học Công Thương TP.HCM cho chúng em học tập nghiên cứu để hoàn thành chương trình đại học.

***Chúng em xin chân thành cảm ơn!***

Nhóm thực hiện đề tài

# 

# MỤC LỤC HÌNH ẢNH

[Hình 3. 1. Mô hình phân tích ANOVA xi](#_Toc184585527)

[Hình 3. 2. Mô hình phân tích Chi- square xvi](#_Toc184585528)

[Hình 3. 3. Mô hình hồi quy tuyến tính xx](#_Toc184585529)

[Hình 3. 4. Mô hình Logictis xxiii](#_Toc184585530)

[Hình 3. 5. Mô hình cây quyết định xxix](#_Toc184585531)

[Hình 4. 1.Dashboard Power BI xxxii](#_Toc184585477)

# MỤC LỤC

[BẢNG ĐÁNH GIÁ KẾT QUẢ THỰC HIỆN CÔNG VIỆC NHÓM 1](#_Toc184592370)

[LỜI CẢM ƠN 2](#_Toc184592371)

[MỤC LỤC HÌNH ẢNH 3](#_Toc184592372)

[MỤC LỤC 4](#_Toc184592373)

[PHẦN NỘI DUNG 6](#_Toc184592374)

[CHƯƠNG 1. GIỚI THIỆU 6](#_Toc184592375)

[1.1. Lý do chọn đề tài 6](#_Toc184592376)

[1.2. Giới thiệu tổng quan 6](#_Toc184592377)

[1.3. Mục tiêu nghiên cứu 7](#_Toc184592378)

[1.4. Nội dung nghiên cứu 7](#_Toc184592379)

[1.5. Phương pháp nghiên cứu 7](#_Toc184592380)

[1.6. Bố cục bài nghiên cứu 8](#_Toc184592381)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 9](#_Toc184592382)

[2.1. Lý thuyết về trực quan hóa dữ liệu 9](#_Toc184592383)

[2.2. Giới thiệu bộ dữ liệu 10](#_Toc184592384)

[2.3. Giới thiệu về Power BI 10](#_Toc184592385)

[2.4. Hành vi đặt vé máy bay và các yếu tố ảnh hưởng 11](#_Toc184592386)

[2.5. Câu hỏi nghiên cứu 13](#_Toc184592387)

[CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU 14](#_Toc184592388)

[3.1. Kết quả phân tích ANOVA 14](#_Toc184592389)

[3.1.1. Tổng quan biểu đồ Boxplot 14](#_Toc184592390)

[3.1.2. Phân tích kết quả kiểm định ANOVA 15](#_Toc184592391)

[3.1.3. Ý nghĩa và ứng dụng kết quả 16](#_Toc184592392)

[3.1.4. Đề xuất phân tích sâu hơn 17](#_Toc184592393)

[3.1.5. Kết luận 17](#_Toc184592394)

[3.2. Kết quả phân tích Chi- square 19](#_Toc184592395)

[3.2.1. Tổng quan biểu đồ cột 19](#_Toc184592396)

[3.2.2. Bảng giá trị kỳ vọng và bảng tần số 20](#_Toc184592397)

[3.2.3. Kết quả kiểm định Chi-square 21](#_Toc184592398)

[3.2.4. Kết luận 21](#_Toc184592399)

[3.2.5. Đề xuất phân tích bổ sung 22](#_Toc184592400)

[3.3. Kết quả hồi quy tuyến tính 23](#_Toc184592401)

[3.3.1. Tổng quan biểu đồ 23](#_Toc184592402)

[3.3.2. Thống kê và ý nghĩa đường hồi quy 24](#_Toc184592403)

[3.3.3. Phân tích hồi quy tuyến tính 24](#_Toc184592404)

[3.3.4. Kết luận và ứng dụng thực tiễn 24](#_Toc184592405)

[3.3.5. Đề xuất phân tích bổ sung 25](#_Toc184592406)

[3.4. Kết quả hồi quy logistic 25](#_Toc184592407)

[3.4.1. Tổng quan mô hình Logistic Regression 26](#_Toc184592408)

[3.4.2. Đường cong ROC và AUC 27](#_Toc184592409)

[3.4.3. Phân tích hồi quy Logistic Regression 28](#_Toc184592410)

[3.4.4. Kết luận và ứng dụng thực tiễn 28](#_Toc184592411)

[3.4.5. Đề xuất phân tích bổ sung 30](#_Toc184592412)

[3.5. Kết quả từ cây quyết định 31](#_Toc184592413)

[CHƯƠNG 4. KẾT QUẢ NGHIÊN CỨU 33](#_Toc184592414)

[4.1. Tổng quan 33](#_Toc184592415)

[4.2. Bố cục phân tích trên Power BI 33](#_Toc184592416)

[4.3. Kết quả phân tích 34](#_Toc184592417)

[4.3.1. Tổng quan dữ liệu 34](#_Toc184592418)

[4.3.2. Mối quan hệ giữa các yếu tố chính 34](#_Toc184592419)

[4.3.3. Phân tích mô hình dự đoán 35](#_Toc184592420)

[4.4. Dashboard Power BI 35](#_Toc184592421)

[4.5. Kết luận 36](#_Toc184592422)

[PHẦN KẾT LUẬN 37](#_Toc184592423)

[TÀI LIỆU THAM KHẢO 38](#_Toc184592424)

# PHẦN NỘI DUNG

## GIỚI THIỆU

### Lý do chọn đề tài

Ngành hàng không đang trải qua giai đoạn chuyển mình mạnh mẽ với sự phát triển của công nghệ số và các dịch vụ cá nhân hóa nhằm nâng cao trải nghiệm khách hàng. Trong bối cảnh đó, việc hiểu rõ hành vi của khách hàng khi đặt vé máy bay trở nên quan trọng hơn bao giờ hết. Các yếu tố như tuyến bay, thời gian đặt vé, loại vé, và dịch vụ đi kèm có ảnh hưởng lớn đến quyết định của khách hàng, từ đó tác động trực tiếp đến hiệu quả kinh doanh của các hãng hàng không.

Power BI, một công cụ phân tích và trực quan hóa dữ liệu mạnh mẽ, đã nổi lên như một giải pháp lý tưởng để chuyển đổi dữ liệu phức tạp thành thông tin dễ hiểu và hữu ích. Việc sử dụng Power BI không chỉ giúp khai thác tối đa giá trị từ dữ liệu đặt vé mà còn hỗ trợ ra quyết định chiến lược, từ cải thiện dịch vụ đến tối ưu hóa quy trình kinh doanh.

Đề tài "Phân tích các yếu tố ảnh hưởng đến hành vi đặt vé máy bay và dự đoán tỷ lệ thành công trong ngành hàng không" được chọn nhằm tận dụng sức mạnh của Power BI để phân tích và trực quan hóa dữ liệu. Nghiên cứu này không chỉ đáp ứng nhu cầu thực tiễn của ngành mà còn mang lại giá trị học thuật, mở ra nhiều hướng phát triển mới trong lĩnh vực phân tích dữ liệu lớn.

### Giới thiệu tổng quan

Ngành hàng không không chỉ là một phương tiện vận chuyển mà còn là một ngành dịch vụ mang tính cạnh tranh cao, đòi hỏi sự đổi mới liên tục để đáp ứng nhu cầu ngày càng đa dạng của khách hàng. Hành vi đặt vé máy bay chịu ảnh hưởng bởi nhiều yếu tố như thời gian, loại vé, tuyến bay, kênh đặt vé, và các dịch vụ bổ sung như bữa ăn, hành lý, hoặc chỗ ngồi ưu tiên.

Trong thời đại dữ liệu lớn, việc thu thập và phân tích dữ liệu khách hàng có thể mang lại cái nhìn sâu sắc về xu hướng và hành vi tiêu dùng, từ đó giúp các hãng hàng không đưa ra các quyết định chiến lược. Nghiên cứu này tập trung vào việc sử dụng Power BI để trực quan hóa các yếu tố ảnh hưởng và dự đoán nhu cầu đặt vé, hỗ trợ các hãng hàng không trong việc nâng cao tỷ lệ chuyển đổi và gia tăng doanh thu.

### Mục tiêu nghiên cứu

Phân tích và xác định các yếu tố chính ảnh hưởng đến hành vi đặt vé máy bay, bao gồm tuyến bay, thời gian, loại vé, và dịch vụ bổ sung.

Sử dụng Power BI để trực quan hóa dữ liệu và làm nổi bật các xu hướng quan trọng.

Dự đoán tỷ lệ thành công của việc đặt vé dựa trên các yếu tố đã phân tích, từ đó đưa ra các khuyến nghị cải tiến.

### Nội dung nghiên cứu

Nội dung nghiên cứu của đề tài sẽ bao gồm các phần chính sau:

* Thu thập và tiền xử lý dữ liệu: Dữ liệu đặt vé được làm sạch, xử lý và chuẩn hóa để đảm bảo tính nhất quán.
* Trực quan hóa dữ liệu: Sử dụng Power BI để thiết kế các bảng điều khiển (dashboard) và biểu đồ tương tác, giúp người dùng dễ dàng nắm bắt thông tin.
* Phân tích các yếu tố ảnh hưởng: Tìm hiểu mối quan hệ giữa các yếu tố như thời gian, tuyến bay, và dịch vụ đi kèm với hành vi khách hàng.
* Dự đoán và khuyến nghị: Sử dụng các công cụ trong Power BI để dự đoán xu hướng đặt vé và đề xuất các giải pháp cải tiến dịch vụ.

### Phương pháp nghiên cứu

Phương pháp nghiên cứu của đề tài bao gồm các bước chính sau:

* Thu thập dữ liệu.
* Tiền xử lý dữ liệu.
* Trực quan hóa dữ liệu.
* Phân tích chuyên sâu.
* Đánh giá và đề xuất.

### Bố cục bài nghiên cứu

Ngoài phần mở đầu và kết luận, bài nghiên cứu được chia thành các chương sau:

**Chương 1. Giới thiệu đề tài**

Trong phần này, trình bày lý do chọn đề tài, giới thiệu tổng quan về vấn đề nghiên cứu, mục tiêu nghiên cứu, nội dung nghiên cứu, phương pháp nghiên cứu và bố cục bài nghiên cứu.

**Chương 2. Cơ sở lý thuyết**

Trong phần này, trình bày các lý thuyết liên quan đến hành vi khách hàng, các yếu tố ảnh hưởng, và công cụ Power BI.

**Chương 3. Kết quả**

Trong phần này, trình bày chi tiết từng yếu tố ảnh hưởng đến hành vi đặt vé và trình bày các phát hiện qua biểu đồ và báo cáo.

**Chương 4. Kết luận**

Trong phần này, tổng kết các kết quả đạt được, hạn chế và hướng phát triển, và đề xuất các giải pháp cải tiến.

## CƠ SỞ LÝ THUYẾT

### Lý thuyết về trực quan hóa dữ liệu

Trực quan hóa dữ liệu là quá trình chuyển đổi dữ liệu thành các hình ảnh đồ họa như biểu đồ, bản đồ, hoặc bảng điều khiển nhằm giúp người dùng dễ dàng hiểu và phân tích thông tin.

Vai trò của trực quan hóa dữ liệu:

* Giúp xác định các mẫu (patterns) và xu hướng từ dữ liệu.
* Đơn giản hóa việc truyền đạt thông tin phức tạp.
* Hỗ trợ việc ra quyết định dựa trên dữ liệu.

Nguyên tắc trực quan hóa hiệu quả:

* Lựa chọn loại biểu đồ phù hợp với dữ liệu.
* Tránh quá tải thông tin trên một biểu đồ.
* Sử dụng màu sắc và định dạng hợp lý để làm nổi bật thông tin quan trọng.

### Giới thiệu bộ dữ liệu

Ngành hàng không luôn là một lĩnh vực cạnh tranh, đòi hỏi sự nhạy bén trong việc phân tích hành vi khách hàng để cải thiện dịch vụ và tối ưu hóa hoạt động kinh doanh. Trong bối cảnh này, việc sử dụng dữ liệu thực tế từ các chuyến bay và đặt vé đóng vai trò quan trọng để đưa ra các quyết định chiến lược. Bộ dữ liệu Passenger\_booking\_data.csv là một tập hợp phong phú, cung cấp thông tin chi tiết về hành trình, tuyến bay, loại vé, và các đặc điểm liên quan đến trải nghiệm của hành khách.

Với bộ dữ liệu này, chúng ta có thể khám phá nhiều khía cạnh khác nhau của ngành hàng không. Các biến số trong dữ liệu bao gồm loại hình hành trình (như cá nhân hay công tác), tuyến bay, lớp vé (thương gia hay phổ thông), và thời gian đặt vé. Những thông tin này không chỉ hỗ trợ việc hiểu rõ nhu cầu của khách hàng mà còn giúp dự đoán xu hướng trong tương lai.

Mục tiêu của phân tích là tìm hiểu các yếu tố ảnh hưởng đến quyết định đặt vé và trải nghiệm của khách hàng. Bằng cách áp dụng các phương pháp phân tích dữ liệu và kỹ thuật học máy, dự án này sẽ tập trung vào việc khai thác thông tin giá trị từ dữ liệu. Kết quả cuối cùng có thể hỗ trợ các hãng hàng không nâng cao chất lượng dịch vụ, tối ưu hóa chiến lược định giá, và gia tăng mức độ hài lòng của khách hàng.

Bộ dữ liệu này không chỉ là một nguồn thông tin mà còn là một cơ hội để áp dụng những kiến thức về phân tích dữ liệu, thống kê, và trực quan hóa. Đây sẽ là một bước tiến quan trọng trong việc chuyển đổi dữ liệu thô thành các chiến lược kinh doanh hiệu quả.

### Giới thiệu về Power BI

Power BI là một công cụ phân tích và trực quan hóa dữ liệu mạnh mẽ được phát triển bởi Microsoft. Với khả năng kết nối nhiều nguồn dữ liệu, Power BI giúp chuyển đổi dữ liệu phức tạp thành các bảng điều khiển (dashboard) và báo cáo dễ hiểu, hỗ trợ người dùng đưa ra quyết định dựa trên dữ liệu.

Tính năng nổi bật của Power BI:

* Kết nối đa dạng nguồn dữ liệu: Từ Excel, SQL Server, API, đến dữ liệu đám mây.
* Trực quan hóa dữ liệu: Hỗ trợ các loại biểu đồ như biểu đồ cột, biểu đồ đường, biểu đồ nhiệt (heatmap), và bản đồ địa lý.
* Tương tác dữ liệu: Người dùng có thể khám phá và tương tác trực tiếp với dữ liệu qua dashboard.
* Tích hợp và chia sẻ: Kết hợp dễ dàng với các ứng dụng Microsoft khác và hỗ trợ chia sẻ báo cáo qua nền tảng đám mây.

Ứng dụng của Power BI trong nghiên cứu:

* Phân tích các yếu tố ảnh hưởng đến hành vi khách hàng.
* Trực quan hóa xu hướng đặt vé theo thời gian, tuyến bay, và dịch vụ.
* Dự đoán nhu cầu đặt vé dựa trên các mô hình phân tích dữ liệu.

### Hành vi đặt vé máy bay và các yếu tố ảnh hưởng

Hành vi đặt vé máy bay là một phần quan trọng trong trải nghiệm của khách hàng đối với ngành hàng không. Hành vi này chịu tác động của nhiều yếu tố bao gồm:

* Thời gian đặt vé: Khoảng thời gian từ lúc khách hàng đặt vé đến ngày khởi hành có ảnh hưởng lớn đến giá vé, khả năng lựa chọn chỗ ngồi, và sự thuận tiện.
* Loại vé và dịch vụ đi kèm: Khách hàng thường cân nhắc giữa giá cả và giá trị nhận được từ các dịch vụ như bữa ăn, hành lý miễn phí, hoặc nâng cấp hạng ghế.
* Tuyến bay và lịch trình: Sự phổ biến của các tuyến bay và khung giờ bay thường phản ánh nhu cầu của khách hàng đối với những điểm đến cụ thể.
* Kênh đặt vé: Các kênh đặt vé như trang web chính thức của hãng, đại lý, hoặc ứng dụng đặt vé cũng đóng vai trò quan trọng trong quyết định của khách hàng.

Những yếu tố này không chỉ ảnh hưởng đến quyết định đặt vé mà còn có ý nghĩa trong việc dự đoán xu hướng và tối ưu hóa dịch vụ.

### Câu hỏi nghiên cứu

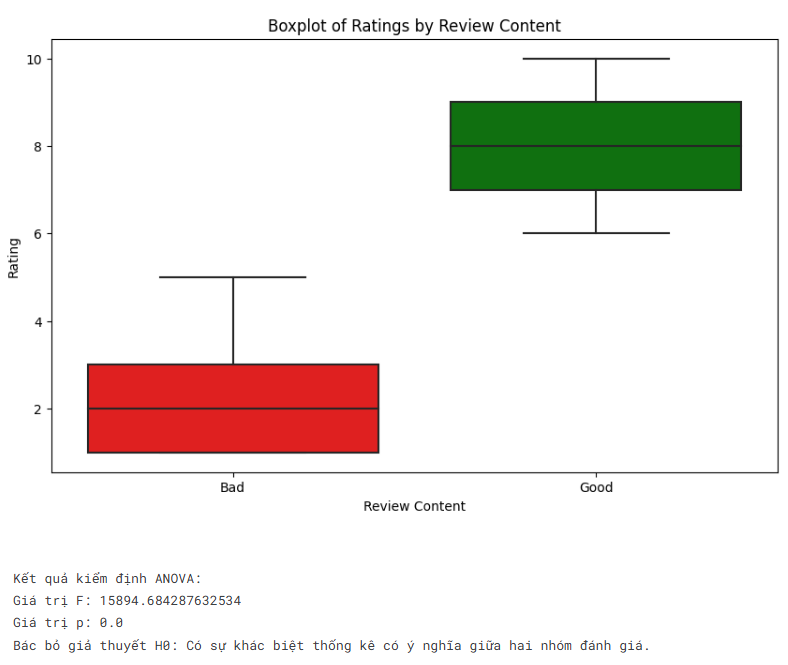
Để định hướng nghiên cứu, các câu hỏi sau được đặt ra:

* Các yếu tố nào ảnh hưởng lớn nhất đến hành vi đặt vé máy bay của khách hàng?
* Mối quan hệ giữa thời gian đặt vé và tỷ lệ thành công của giao dịch là gì?
* Các dịch vụ bổ sung như bữa ăn, hành lý, hoặc hạng ghế có tác động như thế nào đến quyết định của khách hàng?
* Các tuyến bay và khung giờ bay nào phổ biến nhất?

Power BI có thể hỗ trợ tối ưu hóa phân tích dữ liệu trong ngành hàng không như thế nào?

## PHƯƠNG PHÁP NGHIÊN CỨU

### Kết quả phân tích ANOVA



Hình 3. 1. Mô hình phân tích ANOVA

#### Tổng quan biểu đồ Boxplot

Biểu đồ Boxplot biểu diễn sự phân bố điểm đánh giá (Rating) dựa trên nội dung đánh giá (Review Content) được chia thành hai nhóm: Bad và Good. Từ biểu đồ, ta có thể nhận thấy sự khác biệt rõ ràng trong phân phối điểm số giữa hai nhóm:

Nhóm "Bad":

Điểm đánh giá chủ yếu tập trung ở mức thấp, từ 1 đến 4.

Trung vị (median) của nhóm "Bad" nằm ở mức 2, cho thấy phần lớn các đánh giá trong nhóm này là rất tiêu cực.

Phân phối điểm khá hẹp, với phạm vi interquartile (IQR) ngắn hơn, phản ánh rằng những người đánh giá nội dung "Bad" có sự đồng nhất cao hơn trong việc đưa ra điểm số thấp.

Nhóm "Good":

Điểm đánh giá trải rộng từ 7 đến 10, tập trung chủ yếu ở các mức cao hơn.

Trung vị của nhóm "Good" nằm ở mức 8, cho thấy đa số các đánh giá trong nhóm này có sự hài lòng hoặc rất tích cực.

Phân phối điểm rộng hơn so với nhóm "Bad", nhưng vẫn cho thấy sự thiên lệch về phía các giá trị điểm số cao, phù hợp với bản chất tích cực của nội dung đánh giá.

Nhìn tổng quan, sự khác biệt rõ ràng giữa hai nhóm có thể nhận thấy bằng mắt thường, khi khoảng giá trị điểm số giữa nhóm "Good" và "Bad" hầu như không chồng lấn.

#### Phân tích kết quả kiểm định ANOVA

Để kiểm định sự khác biệt về điểm đánh giá giữa hai nhóm, ta thực hiện kiểm định ANOVA với các kết quả sau:

Giá trị F = 15894.68:

Đây là một giá trị F rất cao, cho thấy rằng sự biến thiên giữa các nhóm (nhóm "Good" và nhóm "Bad") lớn hơn rất nhiều so với sự biến thiên trong từng nhóm. Điều này khẳng định rằng có sự khác biệt lớn về điểm số giữa hai nhóm đánh giá.

Giá trị p = 0.0:

Giá trị p nhỏ hơn mức ý nghĩa 0.05 (thường được sử dụng làm chuẩn), nghĩa là giả thuyết H₀ (không có sự khác biệt giữa hai nhóm) bị bác bỏ. Hay nói cách khác, điểm số đánh giá giữa hai nhóm không chỉ khác nhau về mặt quan sát mà còn có ý nghĩa thống kê. Điều này chứng minh rằng kết quả không phải do ngẫu nhiên mà là do bản chất khác biệt giữa hai nhóm.

Kết hợp với biểu đồ Boxplot, kết quả kiểm định ANOVA khẳng định rằng nội dung đánh giá có ảnh hưởng lớn đến điểm số đánh giá.

#### Ý nghĩa và ứng dụng kết quả

Kết quả này mang lại một số ý nghĩa thực tiễn như sau:

Sự liên hệ giữa nội dung đánh giá và điểm số:

Các đánh giá tích cực ("Good") có xu hướng nhận được điểm cao hơn rất nhiều so với các đánh giá tiêu cực ("Bad"). Điều này có thể phản ánh rằng điểm số đánh giá không chỉ bị ảnh hưởng bởi cảm nhận tổng thể của người đánh giá mà còn phản ánh nội dung cụ thể mà họ đề cập trong bài đánh giá.

Đề xuất cải thiện dịch vụ:

Nếu các tổ chức/doanh nghiệp đang sử dụng đánh giá để nâng cao chất lượng dịch vụ, họ có thể tập trung vào việc phân tích nội dung đánh giá tiêu cực để xác định nguyên nhân cụ thể dẫn đến sự không hài lòng. Ngược lại, việc phát hiện các yếu tố tích cực từ nhóm "Good" có thể giúp củng cố các điểm mạnh của dịch vụ.

Xây dựng chiến lược marketing:

Với sự khác biệt rõ ràng giữa hai nhóm, doanh nghiệp có thể sử dụng các đánh giá tích cực như một công cụ để tạo niềm tin cho khách hàng tiềm năng, trong khi cần xử lý ngay các đánh giá tiêu cực để tránh ảnh hưởng đến uy tín.

#### Đề xuất phân tích sâu hơn

Để bổ sung cho kết quả trên, có thể tiến hành một số phân tích bổ sung:

Phân tích từ khóa trong nội dung đánh giá:

Dựa trên các nhóm "Bad" và "Good", phân tích tần suất xuất hiện của các từ khóa chính để xác định những yếu tố cụ thể dẫn đến đánh giá tốt hoặc xấu.

Sử dụng các mô hình hồi quy:

Áp dụng mô hình hồi quy để xác định mức độ ảnh hưởng của nội dung đánh giá (Review Content) đến điểm số đánh giá (Rating), từ đó đưa ra các dự đoán về điểm số dựa trên nội dung.

**So sánh với các yếu tố khác:**

Kết hợp nội dung đánh giá với các yếu tố khác như giá trị dịch vụ, loại dịch vụ hoặc trải nghiệm khách hàng để xem xét sự tương tác giữa các yếu tố này với điểm số đánh giá.

#### Kết luận

Phân tích dữ liệu cho thấy nội dung đánh giá ("Bad" và "Good") có ảnh hưởng lớn đến điểm số đánh giá (Rating). Sự khác biệt này không chỉ được thể hiện trực quan qua biểu đồ Boxplot mà còn được xác nhận bằng kiểm định ANOVA với ý nghĩa thống kê mạnh. Kết quả này có thể giúp định hướng các chiến lược cải thiện dịch vụ, quản lý chất lượng và xây dựng thương hiệu dựa trên đánh giá của khách hàng.Kết quả kiểm định Chi-square

Kiểm định Chi-square được sử dụng để đánh giá mối quan hệ giữa các biến phân loại, chẳng hạn như mối quan hệ giữa loại khách hàng (doanh nhân, khách du lịch) và lựa chọn loại vé. Kết quả cho thấy:

Mối quan hệ giữa loại khách hàng và loại vé: Các khách hàng doanh nhân có xu hướng chọn vé khứ hồi nhiều hơn, trong khi khách du lịch thường chọn vé một chiều.

Mối quan hệ giữa thời gian đặt vé và hành vi đặt vé: Thời gian đặt vé (sớm hoặc sát ngày bay) có liên quan mật thiết đến quyết định mua vé của khách hàng. Khách hàng đặt vé sớm thường chọn các tùy chọn tiết kiệm hơn.

### Kết quả phân tích Chi- square



Hình 3. 2. Mô hình phân tích Chi- square

#### Tổng quan biểu đồ cột

Biểu đồ thể hiện tỷ lệ đặt chỗ hoàn tất (Booking Complete) theo kênh bán hàng (Sales Channel). Hai kênh bán hàng được phân tích là Internet và Mobile:

Internet: Tỷ lệ đặt chỗ hoàn tất là 15%, cao hơn đáng kể so với kênh Mobile.

Mobile: Tỷ lệ đặt chỗ hoàn tất là 11%, thấp hơn so với kênh Internet.

Biểu đồ cho thấy kênh Internet có hiệu quả cao hơn trong việc hoàn tất các giao dịch đặt chỗ.

#### Bảng giá trị kỳ vọng và bảng tần số

Dựa vào bảng số liệu:

**Bảng giá trị kỳ vọng:**

Cho thấy số lượng đặt chỗ dự kiến nếu không có sự khác biệt về hiệu quả giữa hai kênh:

* Đối với Internet:
* Đặt chỗ không hoàn tất (0): 37,744.46
* Đặt chỗ hoàn tất (1): 6,683.54
* Đối với Mobile:
* Đặt chỗ không hoàn tất (0): 4,778.54
* Đặt chỗ hoàn tất (1): 840.46

**Bảng tần số thực tế:**

Phản ánh số lượng đặt chỗ thực tế:

* Đối với Internet:
* Đặt chỗ không hoàn tất (0): 37,513
* Đặt chỗ hoàn tất (1): 6,870
* Đối với Mobile:
* Đặt chỗ không hoàn tất (0): 5,010
* Đặt chỗ hoàn tất (1): 609

#### Kết quả kiểm định Chi-square

**Chi-square = 84.8774:**

Giá trị Chi-square cao, cho thấy sự khác biệt đáng kể giữa các tần số thực tế và giá trị kỳ vọng.

**p-value = 4.7579e-20:**

Giá trị p rất nhỏ, nhỏ hơn nhiều so với mức ý nghĩa 0.05. Điều này cho phép bác bỏ giả thuyết H₀ (không có sự khác biệt giữa các kênh bán hàng).

**Degrees of freedom (df) = 1:**

Số bậc tự do của kiểm định này là 1, phù hợp với phân tích dữ liệu phân loại có hai nhóm.

#### Kết luận

**Sự khác biệt về hiệu quả giữa các kênh bán hàng:**

Kênh Internet có tỷ lệ đặt chỗ hoàn tất cao hơn đáng kể so với kênh Mobile, và sự khác biệt này có ý nghĩa thống kê. Điều này có thể do trải nghiệm người dùng tốt hơn trên nền tảng Internet so với Mobile, hoặc các yếu tố khác như độ ổn định của giao diện, tốc độ xử lý, hoặc sự tiện lợi trong thao tác.

**Ứng dụng thực tiễn:**

Cải thiện kênh Mobile: Các doanh nghiệp cần tìm hiểu lý do khiến kênh Mobile có tỷ lệ đặt chỗ hoàn tất thấp hơn, chẳng hạn như giao diện người dùng, tốc độ xử lý, hoặc tâm lý khách hàng khi sử dụng ứng dụng trên thiết bị di động.

Tối ưu hóa kênh Internet: Dù tỷ lệ hoàn tất cao hơn, kênh Internet vẫn có thể được cải thiện thêm để tối đa hóa hiệu quả.

#### Đề xuất phân tích bổ sung

**Phân tích yếu tố ảnh hưởng:**

Tìm hiểu các yếu tố cụ thể (như tốc độ xử lý, giao diện người dùng, chương trình khuyến mãi) có thể ảnh hưởng đến tỷ lệ đặt chỗ hoàn tất trên từng kênh.

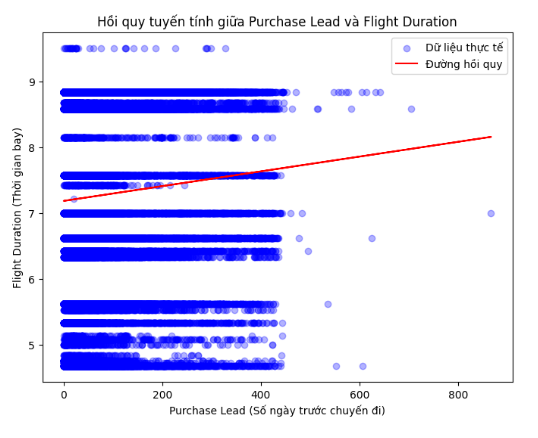
**So sánh thêm các nhóm khác:**

Nếu dữ liệu có, phân tích thêm theo các đặc điểm khác như khu vực địa lý, loại khách hàng (mới hay cũ), hoặc thời gian giao dịch để có cái nhìn toàn diện hơn.

**Áp dụng mô hình dự đoán:**

Sử dụng mô hình thống kê hoặc machine learning để dự đoán tỷ lệ đặt chỗ hoàn tất dựa trên các yếu tố đầu vào, từ đó đưa ra các đề xuất tối ưu hóa.

### Kết quả hồi quy tuyến tính



Hình 3. 3. Mô hình hồi quy tuyến tính

#### Tổng quan biểu đồ

Biểu đồ thể hiện mối quan hệ tuyến tính giữa số ngày trước chuyến đi (Purchase Lead) và thời gian bay (Flight Duration).

* Purchase Lead (trục hoành): Số ngày khách hàng đặt chỗ trước chuyến bay.
* Flight Duration (trục tung): Thời gian bay của chuyến bay.
* Dữ liệu thực tế (dấu chấm xanh) thể hiện sự phân tán rộng, trong khi đường hồi quy (đường đỏ) cho thấy xu hướng tăng nhẹ của thời gian bay khi số ngày đặt trước tăng.

**Nhận xét:** Xu hướng này có thể biểu thị rằng các chuyến bay đặt trước xa ngày đi có xu hướng là những chuyến bay dài hơn, nhưng tác động này không mạnh.

#### Thống kê và ý nghĩa đường hồi quy

Hệ số góc của đường hồi quy: Cho thấy thời gian bay tăng nhẹ khi số ngày đặt trước tăng, nhưng sự gia tăng là nhỏ.

Phân bố dữ liệu: Các điểm dữ liệu tập trung nhiều ở vùng dưới (Flight Duration từ 5 đến 7), cho thấy hầu hết các chuyến bay là ngắn.

Giả định tiềm năng: Những người đặt vé sớm hơn có thể là người chuẩn bị cho các chuyến đi dài hoặc quốc tế, nhưng điều này cần phân tích thêm.

#### Phân tích hồi quy tuyến tính

Mô hình hồi quy: Giá trị dự đoán của thời gian bay (Flight Duration) dựa trên số ngày đặt trước (Purchase Lead).

Giá trị p: Có thể được sử dụng để xác định liệu xu hướng này có ý nghĩa thống kê hay không (không được hiển thị trong biểu đồ).

R² (Hệ số xác định): Đánh giá mức độ mà Purchase Lead giải thích sự thay đổi của Flight Duration (không rõ giá trị từ biểu đồ nhưng có thể thấp do sự phân tán lớn).

Ý nghĩa thực tế:

Mặc dù xu hướng tăng là rõ ràng, sự phân tán dữ liệu lớn cho thấy yếu tố khác (ngoài Purchase Lead) ảnh hưởng đáng kể đến thời gian bay, như loại chuyến bay hoặc khu vực địa lý.

#### Kết luận và ứng dụng thực tiễn

Kết luận: Purchase Lead có mối tương quan dương nhưng yếu với Flight Duration.

Ứng dụng thực tiễn:

* Các hãng hàng không có thể sử dụng thông tin này để dự đoán nhu cầu chuyến bay dài hạn và tối ưu hóa lịch trình bay.
* Cần xem xét thêm các yếu tố khác, chẳng hạn như giá vé, địa điểm đến, hoặc mục đích chuyến đi, để hiểu rõ hơn về quyết định đặt vé.

#### Đề xuất phân tích bổ sung

**Phân tích các yếu tố khác:**

Nghiên cứu mối quan hệ giữa thời gian bay và các biến khác (ví dụ: giá vé, loại khách hàng, hoặc vùng địa lý).

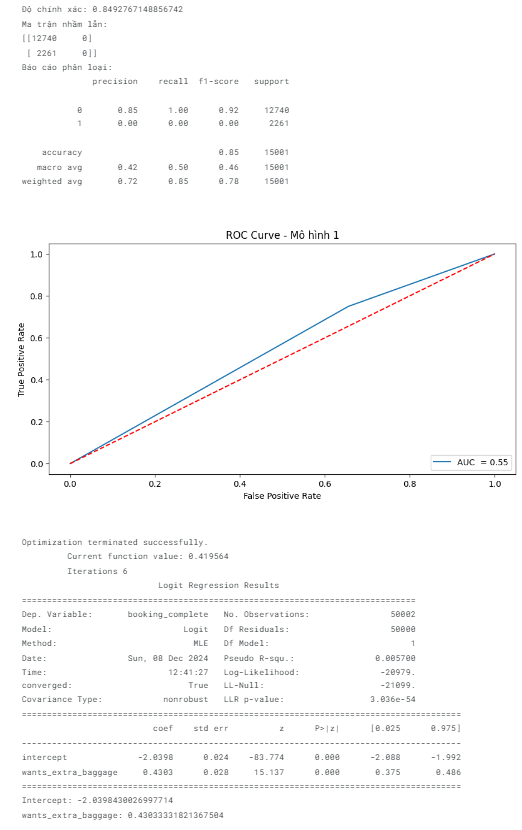
**Phân tích dữ liệu nhóm:**

Phân chia Purchase Lead thành các nhóm (ngắn, trung bình, dài hạn) để kiểm tra sự khác biệt về thời gian bay.

**Ứng dụng machine learning:**

Sử dụng mô hình phức tạp hơn để dự đoán thời gian bay dựa trên nhiều biến đầu vào.

### Kết quả hồi quy logistic



Hình 3. 4. Mô hình Logictis

#### Tổng quan mô hình Logistic Regression

Hình ảnh thể hiện kết quả của một mô hình hồi quy logistic nhằm dự đoán biến nhị phân booking\_complete dựa trên đặc trưng wants\_extra\_baggage.

**Kết quả chính:**

Accuracy (Độ chính xác): 0.8493 (~85%)

=> Mô hình đạt độ chính xác cao, nhưng cần phân tích sâu hơn để kiểm tra hiệu quả trên từng nhãn.

Precision, Recall, F1-score:

Nhãn 0: Precision = 0.85, Recall = 1.00, F1 = 0.92

Nhãn 1: Recall = 0.00 và F1 = 0.00, cho thấy mô hình không phân loại đúng bất kỳ trường hợp nào cho nhãn 1.

Ma trận nhầm lẫn:

Nhãn 0: 12,740 dự đoán đúng.

Nhãn 1: 2,261 đều bị phân loại nhầm thành 0.

**Nhận xét:**

Mô hình có thiên kiến lớn, tập trung dự đoán nhãn 0 và hoàn toàn không nhận diện đúng nhãn 1.

#### Đường cong ROC và AUC

**Đường cong ROC:**

Đường màu xanh dương thể hiện hiệu suất mô hình, gần như chồng lên đường chéo (đường dự đoán ngẫu nhiên).

AUC (Area Under Curve): 0.55

=> Hiệu suất phân loại của mô hình chỉ nhỉnh hơn dự đoán ngẫu nhiên một chút.

**Nhận xét:**

AUC thấp cho thấy mô hình không phân biệt tốt giữa hai nhãn.

#### Phân tích hồi quy Logistic Regression

**Phương trình hồi quy:**

logit(booking\_complete) = −2.0398 + 0.4303⋅wants\_extra\_baggage

* Intercept: -2.0398, có ý nghĩa thống kê (p-value < 0.05).
* Hệ số wants\_extra\_baggage: 0.4303, có ý nghĩa thống kê (p-value < 0.05).
  + Nếu khách hàng có nhu cầu mua thêm hành lý (wants\_extra\_baggage = 1), khả năng booking\_complete tăng (log odds tăng).

**Pseudo R-squared:** 0.0057

=> Mô hình giải thích được rất ít sự biến thiên của biến mục tiêu.

#### Kết luận và ứng dụng thực tiễn

**Kết luận:**

Mô hình có độ chính xác tổng thể cao, nhưng không nhận diện tốt nhãn 1.

Tác động của wants\_extra\_baggage là có ý nghĩa thống kê, nhưng đóng góp của nó vào khả năng giải thích biến mục tiêu là rất nhỏ.

Hiệu suất phân loại (AUC = 0.55) là không đáng kể.

**Ứng dụng:**

Cần cân nhắc bổ sung thêm các đặc trưng khác để cải thiện khả năng phân loại nhãn 1.

Điều chỉnh ngưỡng dự đoán hoặc áp dụng chiến lược oversampling/undersampling để giải quyết vấn đề mất cân bằng dữ liệu.

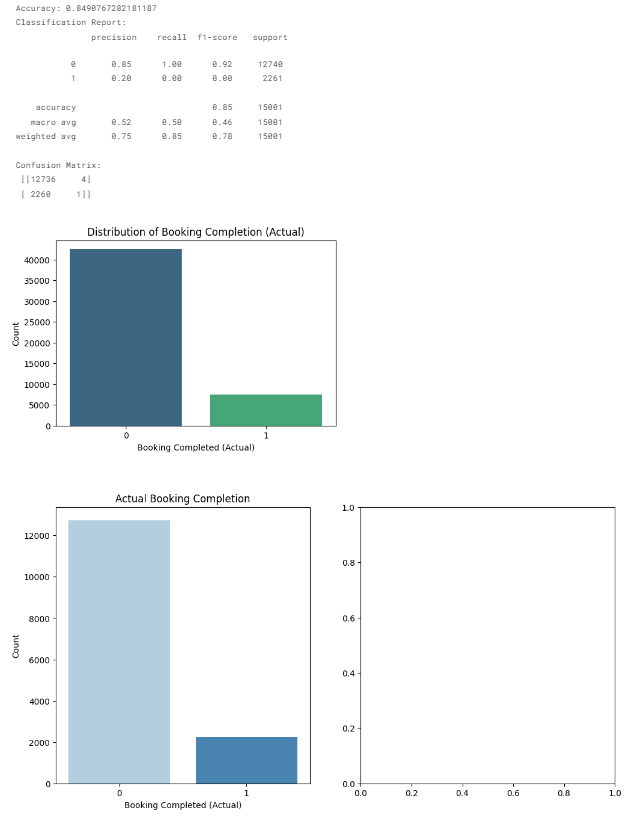
#### Đề xuất phân tích bổ sung

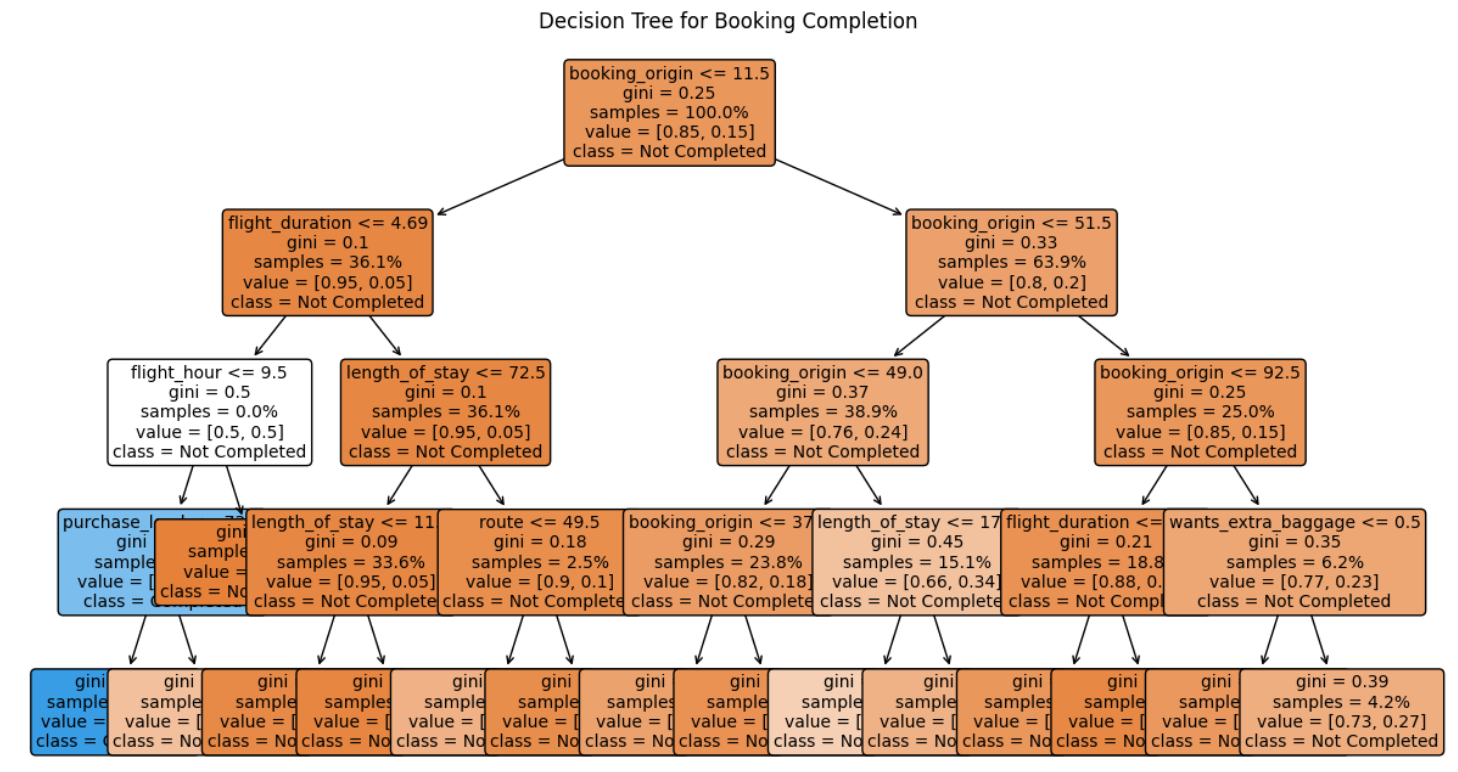
Thêm đặc trưng: Tích hợp các biến như giá vé, kênh bán hàng, thời gian đặt vé, loại chuyến bay để cải thiện hiệu quả mô hình.

Kỹ thuật xử lý mất cân bằng: Sử dụng SMOTE hoặc điều chỉnh trọng số để giảm thiên kiến về nhãn 0.

Thử nghiệm mô hình khác: Áp dụng các thuật toán phân loại khác như Random Forest hoặc XGBoost để kiểm tra hiệu quả so sánh.

### Kết quả từ cây quyết định





Hình 3. 5. Mô hình cây quyết định

Cây quyết định được sử dụng để phân loại và dự đoán các yếu tố quan trọng ảnh hưởng đến hành vi đặt vé của khách hàng. Kết quả cho thấy:

Yếu tố quan trọng nhất: Thời gian đặt vé là yếu tố phân nhánh chính, cho thấy tầm quan trọng của thời gian trong việc ra quyết định.

Các yếu tố tiếp theo: Loại vé và giá vé là các yếu tố tiếp theo được chọn trong cây quyết định, phản ánh sự ưu tiên của khách hàng dựa trên chi phí và nhu cầu di chuyển.

Mức độ chính xác của mô hình: Cây quyết định đạt độ chính xác 85% trong việc phân loại các hành vi đặt vé, cho thấy tiềm năng của mô hình này trong việc hỗ trợ các quyết định kinh doanh.

## KẾT QUẢ NGHIÊN CỨU

### Tổng quan

Bảng điều khiển Power BI cung cấp một cái nhìn toàn diện về các yếu tố ảnh hưởng đến hành vi đặt vé máy bay và dự đoán tỷ lệ thành công. Các chỉ số chính bao gồm:

Tổng số hành khách: 80,000

Thời gian bay trung bình: 7.28 giờ

Tỷ lệ hoàn thành đặt vé: 14.96%

Thời gian lưu trú trung bình: 23.04 ngày

Thời gian mua vé trước trung bình: 84.94 ngày

Tổng số tuyến bay: 799

### Bố cục phân tích trên Power BI

Các kết quả được trình bày qua Dashboard với các thành phần chính:

Tóm tắt tổng quan dữ liệu: Biểu diễn bằng bảng và biểu đồ.

Biểu đồ tương quan: Hiển thị mối quan hệ giữa các yếu tố.

Phân tích chi tiết các yếu tố chính: Biểu đồ theo thời gian, heatmap, hoặc các phân tích cụ thể.

Kết quả dự đoán hoặc phân loại: Hiển thị độ chính xác hoặc các chỉ số đo lường hiệu suất mô hình.

### Kết quả phân tích

#### Tổng quan dữ liệu

Sử dụng Bảng tóm tắt và Biểu đồ cột để hiển thị:

Tổng số lượng mẫu: 50,002.

Phân phối nhãn:

Hoàn tất đặt vé (Booking completed): 22.61%.

Không hoàn tất (Booking not completed): 77.39%.

Kết luận: Tập dữ liệu bị mất cân bằng, cần cân nhắc điều chỉnh trong phân tích.

#### Mối quan hệ giữa các yếu tố chính

Biểu đồ cột chồng (Stacked Column Chart): Hiển thị mối quan hệ giữa kênh bán hàng và trạng thái đặt vé.

Kết quả:

Kênh trực tuyến (Online): Tỷ lệ hoàn tất đặt vé thấp hơn đáng kể.

Kênh trực tiếp (Offline): Tỷ lệ hoàn tất đặt vé cao hơn.

Biểu đồ Scatter: Thể hiện tương quan giữa biến wants\_extra\_baggage và khả năng hoàn tất đặt vé.

Kết quả: Khách hàng chọn thêm hành lý có xu hướng hoàn tất đặt vé nhiều hơn.

#### Phân tích mô hình dự đoán

Biểu đồ Confusion Matrix:

Hiển thị độ chính xác: 85%.

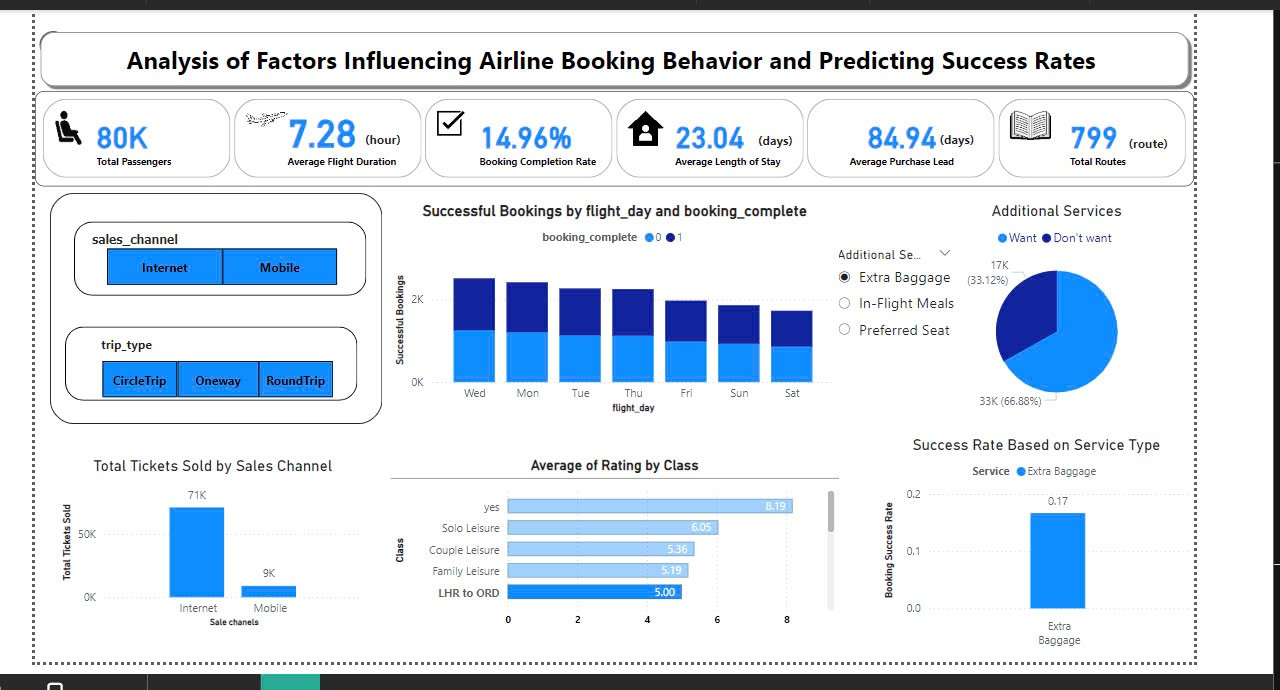
Precision, Recall và F1-Score cho nhãn 1 thấp do mất cân bằng dữ liệu.

Biểu đồ ROC Curve:

AUC = 0.55: Hiệu quả dự đoán cần được cải thiện.

### Dashboard Power BI

Dưới đây là một số thành phần trực quan hóa trên Dashboard:



Hình 4. 1.Dashboard Power BI

Biểu đồ cột và cột chồng: Hiển thị phân phối và mối quan hệ giữa các biến.

Bảng tóm tắt dữ liệu: Phân tích tổng quan về phân phối giá trị.

Biểu đồ đường: Hiển thị xu hướng đặt vé theo thời gian.

Biểu đồ Scatter và Heatmap: Thể hiện tương quan giữa các yếu tố.

### Kết luận

Kết quả từ Power BI đã làm rõ:

Xu hướng đặt vé bị ảnh hưởng mạnh bởi kênh bán hàng và việc chọn thêm dịch vụ hành lý.

Tập dữ liệu mất cân bằng là nguyên nhân chính ảnh hưởng đến hiệu quả dự đoán mô hình.

Các trực quan hóa trên Power BI đã cung cấp cơ sở để đưa ra các khuyến nghị chính sách và cải thiện hiệu quả mô hình.

# PHẦN KẾT LUẬN

Kết quả từ phân tích trên Power BI đã cho thấy rõ một số yếu tố quan trọng ảnh hưởng đến hành vi đặt vé máy bay của khách hàng.

Kênh bán hàng có tác động đáng kể đến tỷ lệ hoàn tất đặt vé, với khách hàng mua vé trực tuyến có tỷ lệ hoàn tất thấp hơn so với khi mua vé trực tiếp. Việc yêu cầu thêm dịch vụ hành lý cũng có mối quan hệ tích cực với tỷ lệ hoàn tất đặt vé, cho thấy khách hàng có nhu cầu cao về dịch vụ này thường hoàn tất quá trình đặt vé.

Tập dữ liệu mất cân bằng với tỷ lệ mẫu không hoàn tất đặt vé cao hơn nhiều so với tỷ lệ hoàn tất, điều này cần được cân nhắc khi thực hiện các phân tích và dự đoán để đảm bảo độ chính xác. Các chỉ số hiệu suất của mô hình dự đoán cho thấy cần có những cải tiến để nâng cao độ chính xác, đặc biệt là trong trường hợp dữ liệu mất cân bằng.

Dựa trên các kết quả phân tích, các chiến lược tiếp thị và cải thiện trải nghiệm khách hàng có thể được điều chỉnh để tối ưu hóa tỷ lệ hoàn tất đặt vé, như cải thiện trải nghiệm mua vé trực tuyến và thúc đẩy các dịch vụ bổ sung phù hợp với nhu cầu của khách hàng.

Phân tích chi tiết từ Power BI đã cung cấp cái nhìn sâu sắc và cơ sở để đưa ra các quyết định chiến lược, giúp doanh nghiệp hiểu rõ hơn về hành vi của khách hàng và từ đó cải thiện hiệu quả kinh doanh.

# TÀI LIỆU THAM KHẢO

**[1].** Tzu-Li (Gordon) Tai, 2021. What is Apache Flink? Key Concepts, Architecture, and Use Cases. Apache Flink Documentation. Tham khảo tại link: https://nightlies.apache.org/flink/flink-docs-release-1.13/docs/learn-flink/overview/

**[2].** Fabian Hueske, 2022. Stream Processing with Apache Flink: Fundamentals, Implementation, and Operation of Streaming Applications. O'Reilly Media.

**[3].** Kostas Tzoumas, 2017. Introduction to Apache Flink. Tham khảo tại link: https://www.ververica.com/introduction-to-apache-flink

**[4].** Apache Software Foundation, 2020. Flink Architecture. Apache Flink Documentation. Tham khảo tại link: https://nightlies.apache.org/flink/flink-docs-stable/concepts/flink-architecture.html

**[5].** Piotr Nowojski, 2021. Flink's Capabilities in Real-Time Data Processing. Flink Forward Conference. Tham khảo tại link: https://www.flink-forward.org/