TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: TÀI DỰ ĐOÁN XU HƯỚNG ĐẶT PHÒNG ĐẶT PHÒNG KHÁCH SẠN**

**Giảng viên: TS Trần Quý Nam || Th.S Lê Thị Thu Trang**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | | **Mã sv** | **Họ và Tên** | **Ngày Sinh** | **Lớp** |
| 1 | 1671020028 | | Nguyễn Thị Lan Anh | 24/02/2004 | CNTT 16-03 |
| 2 | 1671020220 | | Đỗ Ngọc Nghĩa | 02/06/2004 | CNTT 16-03 |
| 3 | 1671020248 | | Ngô Ngọc Phú | 28/10/2004 | CNTT 16-03 |

**Hà Nội, năm 2025**

TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: TÀI DỰ ĐOÁN XU HƯỚNG ĐẶT PHÒNG ĐẶT PHÒNG KHÁCH SẠN**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Mã sv** | **Họ và Tên** | **Ngày Sinh** | **Điểm** | |
| **Bằng Số** | **Bằng Chữ** |
| 1 | 1671020028 | Nguyễn Thị Lan Anh | 24/02/2004 |  |  |
| 2 | 1671020220 | Đỗ Ngọc Nghĩa | 02/06/2004 |  |  |
| 3 | 1671020248 | Ngô Ngọc Phú | 28/10/2004 |  |  |

|  |  |
| --- | --- |
| **CÁN BỘ CHẤM THI 1**  **Trần Quý Nam** | **CÁN BỘ CHẤM THI 2**  **Lê Thị Thu Trang** |

### 

**Hà Nội, năm 2025**

# LỜI NÓI ĐẦU

Trong thời đại công nghệ số, dữ liệu lớn (Big Data) đã trở thành một yếu tố cốt lõi trong hầu hết các lĩnh vực, từ kinh doanh, tài chính, y tế đến du lịch và khách sạn. Với khả năng lưu trữ và phân tích khối lượng dữ liệu khổng lồ, Big Data không chỉ giúp các doanh nghiệp nâng cao hiệu quả vận hành mà còn hỗ trợ ra quyết định dựa trên dữ liệu thực tế.

Môn học Dữ liệu lớn cung cấp những kiến thức quan trọng về các phương pháp thu thập, xử lý, phân tích và khai thác dữ liệu ở quy mô lớn. Sinh viên được tiếp cận với các công nghệ tiên tiến như Apache Spark, Hadoop, SQL phân tán và các thư viện phân tích dữ liệu trong R và Python. Thông qua môn học, sinh viên không chỉ hiểu rõ về các nguyên tắc cơ bản mà còn có cơ hội áp dụng vào các bài toán thực tế, từ đó nâng cao kỹ năng xử lý và phân tích dữ liệu lớn.

Trong bối cảnh ngành du lịch và khách sạn ngày càng phát triển mạnh mẽ, nhu cầu dự đoán xu hướng đặt phòng trở nên cấp thiết đối với các doanh nghiệp khách sạn. Việc phân tích dữ liệu đặt phòng giúp các khách sạn đưa ra chiến lược tối ưu về giá cả, quản lý phòng trống và giảm tỷ lệ hủy phòng. Xuất phát từ thực tiễn này, nhóm nghiên cứu thực hiện đề tài "Dự đoán xu hướng đặt phòng khách sạn" với mục tiêu:

* Phân tích dữ liệu đặt phòng để xác định các yếu tố ảnh hưởng đến xu hướng đặt phòng và hủy phòng.
* Ứng dụng công nghệ dữ liệu lớn (Big Data) và ngôn ngữ R kết hợp với Spark để xử lý và phân tích dữ liệu quy mô lớn.
* Xây dựng mô hình học máy (Random Forest) để dự đoán xu hướng đặt phòng trong tương lai.
* Trực quan hóa dữ liệu bằng thư viện ggplot2 để cung cấp cái nhìn toàn diện về tình hình đặt phòng theo thời gian, loại phòng, mức giá và tỷ lệ hủy.

Quá trình thực hiện đề tài giúp nhóm nghiên cứu có cơ hội tiếp cận với những thách thức thực tế trong việc xử lý dữ liệu lớn, từ việc làm sạch, tiền xử lý dữ liệu đến áp dụng các mô hình phân tích và dự đoán. Ngoài ra, đề tài cũng mang lại giá trị ứng dụng cao, góp phần giúp các doanh nghiệp khách sạn có thể cải thiện hoạt động kinh doanh và nâng cao trải nghiệm khách hàng.

# lời cảm ơn

Nhóm chúng em xin gửi lời cảm ơn chân thành đến giảng viên hướng dẫn, người đã tận tình chỉ bảo, cung cấp kiến thức và định hướng quan trọng trong suốt quá trình thực hiện đề tài. Sự hướng dẫn của thầy/cô không chỉ giúp chúng em hiểu rõ hơn về các khái niệm liên quan đến dữ liệu lớn mà còn giúp chúng em có cái nhìn thực tế về cách ứng dụng các mô hình phân tích và dự đoán trong lĩnh vực khách sạn.

Chúng em cũng xin trân trọng cảm ơn Khoa Công nghệ Thông tin đã tạo điều kiện thuận lợi để nhóm có thể tiếp cận tài liệu, công cụ và môi trường nghiên cứu phù hợp. Nhờ đó, nhóm có thể hoàn thành đề tài này một cách hiệu quả và có hệ thống.

Mặc dù đã cố gắng hoàn thiện đề tài với tất cả sự nỗ lực, nhưng chắc chắn không thể tránh khỏi những thiếu sót. Nhóm rất mong nhận được những góp ý từ thầy/cô và các bạn để có thể tiếp tục cải thiện và phát triển hơn trong tương lai.

Nhóm thực hiện:

Nguyễn Thị Lan Anh

Đỗ Ngọc Nghĩa

Ngô Ngọc Phú

**MỤC LỤC**

[**LỜI NÓI ĐẦU 3**](#_Toc192971474)

[**LỜI CẢM ƠN 4**](#_Toc192971475)

[**CHƯƠNG I: TỔNG QUAN VỀ DỮ LIỆU LỚN TRONG NGÀNH KHÁCH SẠN 10**](#_Toc192971476)

[1.1. Giới thiệu về dữ liệu lớn và ứng dụng trong ngành du lịch – khách sạn. 10](#_Toc192971477)

[1.1.1. Ứng dụng của Big Data trong ngành du lịch – khách sạn 11](#_Toc192971478)

[1.1.2. Cá nhân hóa trải nghiệm khách hàng 11](#_Toc192971479)

[1.1.3. Dự báo nhu cầu và tối ưu giá phòng (Dynamic Pricing) 11](#_Toc192971480)

[1.1.4. Cải thiện hoạt động vận hành khách sạn 11](#_Toc192971481)

[1.1.5. Phân tích đánh giá của khách hàng 11](#_Toc192971482)

[1.1.6. Dự báo xu hướng du lịch 12](#_Toc192971483)

[1.1.7. Cải thiện hệ thống đặt phòng và chăm sóc khách hàng 12](#_Toc192971484)

[1.2. Tầm quan trọng của dự đoán xu hướng đặt phòng khách sạn. 12](#_Toc192971485)

[1.2.1. Tối ưu hóa quản lý doanh thu 12](#_Toc192971486)

[1.2.2. Cải thiện trải nghiệm khách hàng. 12](#_Toc192971487)

[1.2.3. Tăng cường an toàn và bảo mật 13](#_Toc192971488)

[1.2.4. Tiếp thị đích đáng 13](#_Toc192971489)

[1.2.5. Các lợi ích bổ sung của phân tích dữ liệu lớn trong ngành khách sạn 13](#_Toc192971490)

[1.3. Các nguồn dữ liệu liên quan đến đặt phòng khách sạn 14](#_Toc192971491)

[1.3.1. Dữ liệu lịch sử đặt phòng 14](#_Toc192971492)

[1.3.2. Phản hồi và đánh giá của khách hàng 14](#_Toc192971493)

[1.3.3. Dữ liệu thời tiết 14](#_Toc192971494)

[1.3.4. Dữ liệu sự kiện & lễ hội 15](#_Toc192971495)

[1.3.5. Dữ liệu hành vi khách hàng & mạng xã hội 15](#_Toc192971496)

[1.3.6. Dữ liệu kinh tế - xã hội 15](#_Toc192971497)

[1.3.7. Dữ liệu từ đối tác và OTA (Online Travel Agency) 15](#_Toc192971498)

[1.4. Các công nghệ và phương pháp phân tích dữ liệu lớn trong ngành khách sạn. 16](#_Toc192971499)

[1.4.1. Trí tuệ nhân tạo (AI) và Học máy (Machine Learning - ML) 16](#_Toc192971500)

[1.4.2. Internet vạn vật (IoT) & Cảm biến thông minh 16](#_Toc192971501)

[1.4.3. Điện toán đám mây (Cloud Computing) 16](#_Toc192971502)

[1.4.4. Phân tích dữ liệu lớn (Big Data Analytics) 17](#_Toc192971503)

[1.4.5. Blockchain trong quản lý khách sạn 17](#_Toc192971504)

[**CHƯƠNG 2: PHƯƠNG PHÁP VÀ MÔ HÌNH DỰ ĐOÁN XU HƯỚNG ĐẶT PHÒNG 18**](#_Toc192971505)

[2.1. Quy trình thu thập và xử lí dữ liệu đặt phòng 18](#_Toc192971506)

[2.2. Các Phương Pháp Khai Thác Dữ Liệu và Machine Learning Dùng Để Dự Đoán Xu Hướng Đặt Phòng 18](#_Toc192971507)

[2.2.1.Kiểm Định Giả Thuyết: 19](#_Toc192971508)

[2.2.2. Học Máy (Machine Learning) 19](#_Toc192971509)

[2.3. Ứng Dụng Các Thuật Toán Dự Đoán Trong Dự Đoán Xu Hướng Đặt Phòng 20](#_Toc192971510)

[2.3.1. Hồi Quy Tuyến Tính (Linear Regression) 20](#_Toc192971511)

[2.3.2. Cây Quyết Định (Decision Trees) 21](#_Toc192971512)

[2.3.3. Mạng Nơ-ron Nhân Tạo (Artificial Neural Networks - ANN) 22](#_Toc192971513)

[2.3.4. Mô Hình Gradient Boosting (như XGBoost) 23](#_Toc192971514)

[2.4. Đánh Giá Hiệu Quả Mô Hình Dự Đoán 23](#_Toc192971515)

[2.4.1. Chia Dữ Liệu 24](#_Toc192971516)

[2.4.2. Các Chỉ Số Đánh Giá Mô Hình 24](#_Toc192971517)

[2.4.3. Kỹ Thuật Cross-Validation 25](#_Toc192971518)

[2.4.4. Kiểm Tra Trên Dữ Liệu Mới 25](#_Toc192971519)

[2.4.5. Phân Tích Kết Quả 25](#_Toc192971520)

[**CHƯƠNG 3. THỰC HIỆN CHƯƠNG TRÌNH VÀ KẾT QUẢ 27**](#_Toc192971521)

[3.1. Thực hiện chương trình 27](#_Toc192971522)

[3.1.1. Kết nối với Spark 27](#_Toc192971523)

[3.1.2. Đọc và kiểm tra dữ liệu từ CSV 27](#_Toc192971524)

[3.1.3. Xử lý dữ liệu 28](#_Toc192971525)

[3.1.4. Lưu dữ liệu đã xử lý 29](#_Toc192971526)

[3.1.5. Trực quan hóa dữ liệu 29](#_Toc192971527)

[3.1.6. Phân tích mô hình Random Forest trên dữ liệu đặt phòng khách sạn 31](#_Toc192971528)

[3.2. Kết quả 35](#_Toc192971529)

[**KÉT LUẬN 40**](#_Toc192971530)

[**TÀI LIỆU THAM KHẢO 41**](#_Toc192971531)

MỤC LỤC HÌNH ẢNH

[Hình 1. BIGDATA. 9](#_Toc192975961)

[Hình 2. Kết nối với Spark. 26](#_Toc192975962)

[Hình 3. Đọc và kiểm tra dữ liệu từ CSV. 27](#_Toc192975963)

[Hình 4. Xử lí dữ liệu. 27](#_Toc192975964)

[Hình 5. Lưu dữ liệu đã xử lý. 28](#_Toc192975965)

[Hình 6. Trực quan hóa dữ liệu . 28](#_Toc192975966)

[Hình 7. Tỷ lệ hủy theo tháng. 29](#_Toc192975967)

[Hình 8. Số lượng đặt phòng theo loại khách sạn. 29](#_Toc192975968)

[Hình 9. Biểu đồ phân phối ADR. 30](#_Toc192975969)

[Hình 10. Số lượng đặt phòng theo loại phòng. 30](#_Toc192975970)

[Hình 11. Chia dữ liệu thành tập huấn luyện. 31](#_Toc192975971)

[Hình 12. Huấn luyện mô hình Ramdom Forest. 31](#_Toc192975972)

[Hình 13. Dự đoán trên tập kiểm tra. 32](#_Toc192975973)

[Hình 14. Đánh giá mô hình bằng AUC-ROC. 32](#_Toc192975974)

[Hình 15. Xem dữ liệu Spark. 33](#_Toc192975975)

[Hình 16. Ngắt kết nối Spark. 34](#_Toc192975976)

[Hình 17. Kết quả biểu đồ số lượng đặt phong theo tháng. 34](#_Toc192975977)

[Hình 18.kết quả biểu đồ tỷ lệ hủy đặt phòng theo tháng. 35](#_Toc192975978)

[Hình 19. Kết quả biểu đồ số lượng đặt phòng theo loại khách sạn. 35](#_Toc192975979)

[Hình 20. Kết quả biểu đồ phân phối ADR. 36](#_Toc192975980)

[Hình 21. Biểu đồ số lượng đặt phòng theo loại phòng. 37](#_Toc192975981)

[Hình 22. Hiệu suất mô hình 37](#_Toc192975982)

[Hình 23. Xem dữ liệu từ Spark. 37](#_Toc192975983)

[Hình 24. Xem bảng dữ liệu. 38](#_Toc192975984)

# CHƯƠNG I: TỔNG QUAN VỀ DỮ LIỆU LỚN TRONG NGÀNH KHÁCH SẠN

## Giới thiệu về dữ liệu lớn và ứng dụng trong ngành du lịch – khách sạn.

Dữ liệu lớn(Big data) là nền tảng để khách sạn thấu hiểu thị trường, nhu cầu khách hàng nhằm tăng doanh thu với chất lượng trải nghiệm

Dữ liệu lớn ngày càng quan trọng và phổ biến trong số hóa ngành du lịch và khách sạn. Với sự phát triển nhanh chóng của công nghệ và internet, ngành này đã tiếp nhận và tận dụng dữ liệu lớn để cải thiện trải nghiệm của du khách. Đồng thời, tối ưu hóa hoạt động kinh doanh và dự đoán xu hướng tương lai. Bài viết dưới đây sẽ giúp bạn khám phá chi tiết về dữ liệu lớn, tác động đến ngành du lịch và khách sạn. Và lý giải lý do việc nắm vững dữ liệu lớn trở thành chìa khóa thành công trong thời đại số hóa hiện nay.



#### Hình 1. BIGDATA.

Dữ liệu lớn (Big Data) là tập hợp dữ liệu có khối lượng khổng lồ, tốc độ tạo ra nhanh và đa dạng về định dạng. Nó thường được mô tả qua mô hình **5V:**

* **Volume (Khối lượng):** Dữ liệu có dung lượng lớn, thường tính bằng terabyte hoặc petabyte.
* **Velocity (Tốc độ)**: Dữ liệu được tạo ra và xử lý với tốc độ cao.
* **Variety (Đa dạng):** Dữ liệu đến từ nhiều nguồn khác nhau, có thể ở dạng văn bản, hình ảnh, video, cảm biến, mạng xã hội...
* **Veracity (Tính xác thực):** Dữ liệu cần được lọc và xử lý để đảm bảo độ tin cậy.
* **Value (Giá trị):** Dữ liệu mang lại giá trị lớn khi được phân tích và ứng dụng hợp lý.

Big Data giúp các tổ chức và doanh nghiệp có thể khai thác thông tin quan trọng từ lượng dữ liệu khổng lồ, từ đó đưa ra các quyết định chính xác và hiệu quả hơn.

### **Ứng dụng của Big Data trong ngành du lịch – khách sạn**

Ngành du lịch và khách sạn là một trong những lĩnh vực hưởng lợi nhiều nhất từ Big Data, nhờ khả năng phân tích hành vi khách hàng, tối ưu hóa hoạt động kinh doanh và nâng cao trải nghiệm dịch vụ.

### **Cá nhân hóa trải nghiệm khách hàng**

* Big Data giúp thu thập thông tin về sở thích, thói quen và lịch sử đặt phòng của khách hàng.
* Các khách sạn có thể sử dụng dữ liệu này để cung cấp dịch vụ cá nhân hóa, như gợi ý phòng phù hợp, ưu đãi riêng biệt hoặc dịch vụ đặc biệt theo nhu cầu.
* Ví dụ: Marriott và Hilton sử dụng AI và Big Data để cung cấp trải nghiệm đặt phòng và dịch vụ theo sở thích của từng khách hàng.

### **Dự báo nhu cầu và tối ưu giá phòng (Dynamic Pricing)**

* Big Data giúp khách sạn dự đoán nhu cầu theo mùa, xu hướng du lịch và các sự kiện lớn.
* Thuật toán phân tích dữ liệu giúp điều chỉnh giá phòng theo thời gian thực dựa trên cung – cầu.
* Ví dụ: Hệ thống đặt phòng của Booking.com và Airbnb sử dụng AI để thay đổi giá dựa trên xu hướng thị trường.

### **Cải thiện hoạt động vận hành khách sạn**

* Hệ thống quản lý khách sạn có thể theo dõi mức tiêu thụ năng lượng, điều chỉnh điều hòa, đèn điện nhằm tiết kiệm chi phí.
* Dữ liệu từ cảm biến IoT giúp tối ưu hóa quản lý phòng, bảo trì thiết bị và giảm thiểu sự cố.
* Ví dụ: Một số khách sạn lớn như InterContinental Hotels sử dụng dữ liệu IoT để tự động điều chỉnh nhiệt độ phòng theo thời tiết và mức tiêu thụ năng lượng.

### **Phân tích đánh giá của khách hàng**

* Big Data giúp thu thập và phân tích hàng triệu đánh giá từ các nền tảng như TripAdvisor, Google Reviews và mạng xã hội.
* Các khách sạn có thể sử dụng AI để xác định xu hướng, phản hồi nhanh chóng và cải thiện chất lượng dịch vụ dựa trên phản hồi thực tế.

### **Dự báo xu hướng du lịch**

* Phân tích dữ liệu từ mạng xã hội, tìm kiếm Google, và hành vi đặt phòng giúp dự đoán xu hướng du lịch mới.
* Các công ty lữ hành có thể xây dựng chiến lược marketing nhắm đến khách hàng tiềm năng theo thời gian thực.

### **Cải thiện hệ thống đặt phòng và chăm sóc khách hàng**

* Chatbot AI và trợ lý ảo giúp trả lời câu hỏi, hỗ trợ đặt phòng 24/7.
* Big Data kết hợp với AI giúp khách sạn hiểu rõ hơn về nhu cầu khách hàng, từ đó đề xuất dịch vụ phù hợp.

## Tầm quan trọng của dự đoán xu hướng đặt phòng khách sạn.

* Tối ưu hóa quản lý doanh thu
* Quản lý chiến lược
* Cải thiện trải nghiệm khách hàng
* Tăng cường an toàn và bảo mật
* Cải thiện hiệu suất của tiếp thị đích đáng
* Quản lý uy tín và nghiên cứu tiếp thị

### Tối ưu hóa quản lý doanh thu

Phân tích dữ liệu lớn trong ngành du lịch liên quan đến nhiều yếu tố. Bao gồm: thu thập, phân tích cách đặt phòng, tỷ lệ thành viên, sở thích thị trường và nhu cầu của khách hàng. Các công ty triển khai phương pháp này thường thành công dự đoán nhu cầu tương lai.

Ngoài ra, bạn có thể tận dụng sự tương tác giữa dữ liệu lớn và ngành du lịch để mở rộng mô hình giá động. Từ đó, dễ dàng đáp ứng các thay đổi về nhu cầu, kỳ vọng theo mùa hoặc sự kiện đặc biệt. Điều này giúp khách sạn thành công thu hút khách hàng và tối đa hóa lợi nhuận.

Bằng việc phân tích quản lý doanh thu khách sạn, bạn có thể duy trì tối ưu mức tồn kho. Sau đó, có thể đưa ra quyết định mua hàng thông minh và tối ưu hóa quy trình cung ứng. Tất cả điều này giúp tăng cường hiệu suất hoạt động và tiết kiệm tối đa chi phí.

### Cải thiện trải nghiệm khách hàng.

Các khách sạn không ngừng đầu tư phân tích dữ liệu lớn để thấu hiểu sở thích và hành vi của khách hàng. Họ có thể thu thập và khám phá thông qua chatbot, lịch sử đặt hàng. Ngoài ra, còn dựa trên mức độ quan tâm ở các phương tiện truyền thông xã hội. Từ đó, có thể khám thác dễ dàng các mô hình và xu hướng ẩn. Việc hiểu biết sâu hơn về khách hàng giúp mang đến trải nghiệm chất lượng nhất.

Điều quan trọng hơn cả là tùy chỉnh thông điệp và cung cấp nội dung nhắm đến từng nhóm khách hàng. Để đạt được điều đó, cần triển khai các chiến lược tiếp thị cá nhân hóa và quảng cáo có chọn lọc. Khách hàng sẽ cảm thấy ấn tượng bởi cách tiếp cận đặc biệt. Điều này dẫn đến sự đồng cảm và kết nối cảm xúc mạnh mẽ với thương hiệu. Nhờ đó, dữ liệu lớn trong ngành du lịch giúp tăng cường tỷ lệ chuyển đổi và sự trung thành.

Với các phân tích dự đoán trong kinh doanh lưu trú, bạn có thể dễ dàng đáp ứng nhu cầu của khách hàng. Trước khi nó trở thành “nỗi đau” của họ. Điều này có thể thực hiện thông qua nghiên cứu dữ liệu, lịch sử khách hàng. Đồng thời, quan sát các mô hình hành vi và theo dõi xu hướng thị trường hiện tại. Nhờ đó, bạn có thể đề xuất các gợi ý, trải nghiệm và dịch vụ phù hợp với sở thích cá nhân của khách hàng.

### Tăng cường an toàn và bảo mật

Phân tích dữ liệu lớn trong ngành du lịch là công cụ hoàn hảo để tăng cường bảo mật. Cụ thể là xác định các nguy cơ tiềm năng và vi phạm. Bao gồm hình ảnh từ camera quan sát, nhật ký truy cập và thông tin khách hàng. Điều này giúp khách sạn có thể triển khai kịp thời các biện pháp. Đồng thời, tăng cường bảo mật, đảm bảo an toàn cho khách hàng và nhân viên. Công nghệ phân tích dữ liệu còn giúp xác định các hoạt động đáng ngờ. Đó có thể là các giao dịch bất thường, hành vi khách hàng kỳ lạ và các biểu hiện gian lận. Do đó, ngành du lịch sử dụng dữ liệu lớn để giảm thiểu tổn thất tài chính và bảo vệ hoạt động toàn vẹn.

### Tiếp thị đích đáng

Bạn có muốn cải thiện hiệu suất của chiến dịch tiếp thị không? Tận dụng lượng lớn dữ liệu khách hàng và áp dụng phân tích dự đoán cho các khách sạn. Đây là hai giải pháp hàng đầu giúp cải thiện.

Dữ liệu lớn trong du lịch được thu thập giúp khách sạn phân tích lịch sử đặt phòng, sở thích, dân số học và mức độ tương tác với thương hiệu. Nhờ đó, họ có thể tạo ra hồ sơ khách hàng chi tiết và tùy chỉnh thông điệp tiếp thị phù hợp. Sau đó, bạn sẽ đề xuất các sản phẩm, dịch vụ và ưu đãi liên quan đến từng cá nhân khách hàng.

Còn dữ liệu lớn trong ngành khách sạn giúp khách sạn khám phá được sở thích của khách hàng về loại phòng hoặc dịch vụ tiện nghi. Nắm được các thông tin trên, khách sạn có thể triển khai các chương trình ưu đãi hiệu quả. Chẳng hạn như cung cấp gói dịch vụ spa độc quyền hoặc giảm giá, gửi email tiếp thị để thể hiện sự quan tâm. Cuối cùng, có thể gửi lời chúc mừng sinh nhật và đề xuất cá nhân về các điểm tham quan.

### Các lợi ích bổ sung của phân tích dữ liệu lớn trong ngành khách sạn

Phân tích dữ liệu lớn trong kinh doanh khách sạn mang lại lợi thế cạnh tranh ấn tượng. Ví dụ, bằng cách phân tích dữ liệu khách hàng, khách sạn có thể xác định xu hướng quan tâm đến môi trường của du khách. Khách sạn có thể thực hiện các chiến lược hướng tới đảm bảo phát triển bền vững. Từ đó, xây dựng thương hiệu tích cực và nổi bật so với các đối thủ khác.

Hoạt động phân tích còn hữu ích trong việc hợp tác và thu hút đối tác. Cụ thể là hợp tác với ngành giao thông, giải trí hoặc du lịch địa phương. Khách sạn có thể chia sẻ dữ liệu, ý tưởng để cung cấp gói dịch vụ đính kèm, quảng cáo chéo hoặc chương trình tích hợp. Như vậy, bạn có thể mở rộng khả năng tiếp cận và cung cấp thêm giá trị cho khách hàng.

## Các nguồn dữ liệu liên quan đến đặt phòng khách sạn

Dự đoán xu hướng đặt phòng khách sạn yêu cầu một lượng lớn dữ liệu từ nhiều nguồn khác nhau. Những nguồn dữ liệu này giúp khách sạn hiểu rõ hơn về nhu cầu khách hàng, tối ưu hóa giá cả, nâng cao chất lượng dịch vụ và đưa ra quyết định kinh doanh hiệu quả.

### **Dữ liệu lịch sử đặt phòng**

* Chứa thông tin về các đặt phòng trước đây, bao gồm:
  + Ngày đặt phòng, ngày nhận/trả phòng.
  + Loại phòng được đặt.
  + Số lượng khách.
  + Giá phòng tại thời điểm đặt.
  + Thời gian lưu trú trung bình.
  + Tỷ lệ hủy đặt phòng.
* **Ứng dụng:** 
  + Phân tích xu hướng theo mùa, ngày lễ, sự kiện đặc biệt.
  + Dự báo công suất phòng trong tương lai.
  + Xây dựng mô hình Dynamic Pricing để tối ưu hóa giá phòng.

### **Phản hồi và đánh giá của khách hàng**

* Dữ liệu từ các nền tảng đánh giá như **TripAdvisor, Google Reviews, Booking.com, Agoda, Expedia.**
* Chứa thông tin về:
  + Điểm số đánh giá (rating).
  + Nhận xét tích cực và tiêu cực.
  + Phản hồi về dịch vụ, tiện nghi, thái độ nhân viên.
* **Ứng dụng:** 
  + Cải thiện chất lượng dịch vụ dựa trên phản hồi thực tế.
  + Xây dựng hệ thống đề xuất cá nhân hóa dựa trên sở thích khách hàng.
  + Phân tích sentiment (cảm xúc) của khách hàng để đánh giá mức độ hài lòng.

### **Dữ liệu thời tiết**

* Dữ liệu từ **các trạm khí tượng, API thời tiết** như OpenWeatherMap, Weather.com.
* Chứa thông tin về:
  + Nhiệt độ, độ ẩm, lượng mưa.
  + Dự báo thời tiết trong tương lai.
  + Số giờ nắng trong ngày.
* **Ứng dụng:** 
  + Dự đoán lượng khách du lịch theo mùa.
  + Điều chỉnh giá phòng dựa trên điều kiện thời tiết (ví dụ: khách sạn gần biển tăng giá vào mùa hè).
  + Đề xuất các dịch vụ phù hợp như tour du lịch, hồ bơi, spa.

### **Dữ liệu sự kiện & lễ hội**

* Thu thập từ **lịch sự kiện địa phương, trang web du lịch, chính quyền thành phố.**
* Chứa thông tin về:
  + Lễ hội văn hóa, sự kiện thể thao, hội nghị, triển lãm.
  + Ngày nghỉ lễ quốc gia và quốc tế.
* **Ứng dụng:** 
  + Dự đoán nhu cầu đặt phòng trong các sự kiện lớn.
  + Điều chỉnh giá phòng dựa trên độ hot của sự kiện.
  + Triển khai chiến lược marketing nhắm đến du khách tham dự sự kiện.

### **Dữ liệu hành vi khách hàng & mạng xã hội**

* Thu thập từ **Facebook, Instagram, Twitter, TikTok, Google Trends.**
* Chứa thông tin về:
  + Tìm kiếm du lịch phổ biến.
  + Xu hướng du lịch theo thời gian thực.
  + Các điểm đến đang được quan tâm.
* **Ứng dụng:** 
  + Dự báo xu hướng đặt phòng dựa trên hành vi tìm kiếm.
  + Nhắm mục tiêu quảng cáo đến khách hàng có nhu cầu du lịch.
  + Đánh giá chiến dịch quảng bá khách sạn có hiệu quả hay không.

### **Dữ liệu kinh tế - xã hội**

* Thu thập từ **chỉ số kinh tế, báo cáo thị trường du lịch**.
* Chứa thông tin về:
  + Tỷ lệ thất nghiệp, lạm phát, mức sống.
  + Xu hướng chi tiêu của khách du lịch.
* **Ứng dụng:** 
  + Dự đoán khả năng chi tiêu của khách du lịch.
  + Xây dựng chiến lược giá phù hợp với từng đối tượng khách hàng.

### **Dữ liệu từ đối tác và OTA (Online Travel Agency)**

* Thu thập từ **Booking.com, Agoda, Expedia, Airbnb**.
* Chứa thông tin về:
  + Lượng đặt phòng trên các nền tảng này.
  + Xu hướng đặt phòng theo khu vực.
* **Ứng dụng:** 
  + So sánh hiệu suất đặt phòng giữa các kênh.
  + Điều chỉnh chiến lược marketing theo nền tảng OTA hiệu quả nhất.

## Các công nghệ và phương pháp phân tích dữ liệu lớn trong ngành khách sạn.

Ngành khách sạn ngày càng ứng dụng **Big Data** để phân tích xu hướng, tối ưu hóa dịch vụ và nâng cao trải nghiệm khách hàng. Để khai thác tối đa giá trị từ dữ liệu, các khách sạn sử dụng nhiều công nghệ và phương pháp phân tích hiện đại như **AI, Machine Learning, IoT, Cloud Computing** và các công cụ phân tích dữ liệu lớn.

### **Trí tuệ nhân tạo (AI) và Học máy (Machine Learning - ML)**

* **Ứng dụng:** 
  + **Dự đoán xu hướng đặt phòng:** Sử dụng ML để dự báo nhu cầu theo mùa, sự kiện và điều chỉnh giá phòng.
  + **Cá nhân hóa dịch vụ:** Phân tích hành vi khách hàng để đề xuất ưu đãi phù hợp.
  + **Chatbot & Trợ lý ảo:** AI hỗ trợ đặt phòng, tư vấn dịch vụ 24/7.
* **Công cụ phổ biến:** TensorFlow, Scikit-Learn, PyTorch.

### **Internet vạn vật (IoT) & Cảm biến thông minh**

* **Ứng dụng:** 
  + **Quản lý phòng thông minh:** Cảm biến IoT giúp điều chỉnh nhiệt độ, ánh sáng, tiêu thụ năng lượng.
  + **Tối ưu hóa bảo trì:** Phát hiện sự cố kỹ thuật trong hệ thống điều hòa, đèn, cửa tự động.
  + **Cải thiện trải nghiệm khách hàng:** Khách có thể điều khiển thiết bị trong phòng qua smartphone.
* **Công nghệ:** ESP32, Arduino.

### **Điện toán đám mây (Cloud Computing)**

* **Ứng dụng:** 
  + **Lưu trữ dữ liệu khách hàng và đặt phòng trên cloud** để dễ dàng truy xuất từ mọi nơi.
  + **Phân tích dữ liệu thời gian thực** để tối ưu hóa chiến lược kinh doanh.
  + **Tích hợp với các nền tảng OTA** (Booking.com, Agoda, Expedia).
* **Dịch vụ phổ biến:** AWS, Google Cloud, Microsoft Azure.

### **Phân tích dữ liệu lớn (Big Data Analytics)**

* **Ứng dụng:** 
  + **Phân tích xu hướng thị trường** dựa trên dữ liệu từ mạng xã hội, Google Trends.
  + **Đo lường hiệu quả marketing** bằng cách theo dõi hành vi khách hàng trên website và OTA.
  + **Cải thiện dịch vụ** bằng cách phân tích phản hồi từ TripAdvisor, Google Reviews.
* **Công nghệ phổ biến:** Apache Hadoop, Spark, Tableau, Power BI.

### **Blockchain trong quản lý khách sạn**

* **Ứng dụng:** 
  + **Bảo mật dữ liệu khách hang:** Lưu trữ thông tin đặt phòng, thanh toán an toàn hơn.
  + **Chương trình khách hàng thân thiết:** Tích hợp blockchain để quản lý điểm thưởng, giảm gian lận.
  + **Hợp đồng thông minh (Smart Contract):** Tự động hóa quá trình thanh toán đặt phòng.
* **Công nghệ:** Ethereum, Hyperledger, Solana.

# CHƯƠNG 2: PHƯƠNG PHÁP VÀ MÔ HÌNH DỰ ĐOÁN XU HƯỚNG ĐẶT PHÒNG

## 2.1. Quy trình thu thập và xử lí dữ liệu đặt phòng

Quy trình thu thập và xử lý dữ liệu đặt phòng là bước quan trọng trong việc xây dựng mô hình dự đoán. Dưới đây là các bước chính trong quy trình này:

.**Bước 1: Xác định nguồn dữ liệu**

* **Dữ liệu lịch sử đặt phòng**: Thu thập từ hệ thống quản lý khách sạn , bao gồm thông tin về thời gian đặt phòng, loại phòng, giá cả, và thông tin khách hàng.
* **Dữ liệu bên ngoài**: Thông tin từ các nguồn như thời tiết, sự kiện địa phương, và các yếu tố kinh tế có thể ảnh hưởng đến xu hướng đặt phòng.

**Bước 2: Thu thập dữ liệu**

* Sử dụng API hoặc các công cụ ETL (Extract, Transform, Load) để thu thập dữ liệu từ các nguồn xác định.
* Đảm bảo dữ liệu được lưu trữ một cách có tổ chức, thường trong cơ sở dữ liệu SQL hoặc NoSQL.

.**Bước 3: Tiền xử lý dữ liệu**

* **Làm sạch dữ liệu**: Loại bỏ các bản ghi trùng lặp, xử lý các giá trị thiếu và sửa chữa các lỗi dữ liệu.
* **Chuyển đổi dữ liệu**: Chuyển đổi dữ liệu thành định dạng phù hợp cho việc phân tích, bao gồm việc mã hóa các biến phân loại và chuẩn hóa các biến số.

**Bước 4: Phân tích dữ liệu**

* Sử dụng các phương pháp thống kê để khám phá dữ liệu và tìm ra các mẫu xu hướng trong dữ liệu lịch sử đặt phòng.
* Sử dụng trực quan hóa dữ liệu để giúp hiểu rõ hơn về các yếu tố ảnh hưởng đến xu hướng đặt phòng.

## 2.2. Các Phương Pháp Khai Thác Dữ Liệu và Machine Learning Dùng Để Dự Đoán Xu Hướng Đặt Phòng

* Sử dụng biểu đồ thời gian để trực quan hóa thay đổi trong đặt phòng, từ đó nhận diện các mùa cao điểm và thấp điểm.

### 2.2.1.Kiểm Định Giả Thuyết:

* Sử dụng kiểm định t-test hoặc ANOVA để so sánh sự khác biệt giữa các nhóm dữ liệu (ví dụ: đặt phòng vào cuối tuần so với giữa tuần).
* Giúp xác định liệu một yếu tố cụ thể (như giá phòng) có ảnh hưởng rõ rệt đến quyết định đặt phòng hay không.

### 2.2.2. Học Máy (Machine Learning)

* Hồi Quy (Regression):
  + Hồi Quy Tuyến Tính:
    - Mô hình đơn giản để dự đoán số lượng đặt phòng dựa trên các yếu tố như giá phòng, thời tiết, và sự kiện địa phương.
    - Ví dụ, nếu giá phòng tăng lên 20%, mô hình có thể dự đoán rằng số lượng đặt phòng sẽ giảm 10%.
  + Hồi Quy Logistic:
    - Dùng để dự đoán xác suất khách hàng sẽ đặt phòng. Kết quả được diễn giải dưới dạng xác suất (0-1).
    - Ví dụ, mô hình có thể dự đoán rằng 70% khách hàng sẽ đặt phòng nếu họ nhận được một mức giá ưu đãi.
* Cây Quyết Định (Decision Trees):
  + Tạo mô hình cây, trong đó mỗi nút đại diện cho một quyết định dựa trên một biến (như giá phòng hoặc loại phòng).
  + Dễ hiểu và có thể trực quan hóa, giúp nhận diện rõ ràng các yếu tố ảnh hưởng đến quyết định đặt phòng.
* Random Forest (RF):
  + Sử dụng nhiều cây quyết định để cải thiện độ chính xác. Mỗi cây sẽ đưa ra một dự đoán, và kết quả cuối cùng là trung bình của tất cả các dự đoán.
  + Giúp giảm thiểu overfitting, làm cho mô hình mạnh mẽ hơn với dữ liệu mới.
* Mô Hình Gradient Boosting:
  + XGBoost:
    - Một thuật toán mạnh mẽ nổi tiếng trong các cuộc thi dữ liệu, giúp cải thiện độ chính xác bằng cách xây dựng các cây quyết định từng bước.
    - Tối ưu hóa hiệu suất thông qua việc điều chỉnh các tham số như học suất và chiều cao của cây.
* Mạng Nơ-ron (Neural Networks):
  + Mạng nơ-ron sâu có thể học các mẫu phức tạp trong dữ liệu, rất hữu ích khi làm việc với dữ liệu lớn và phi tuyến tính.
  + Sử dụng các lớp ẩn để trích xuất đặc trưng từ dữ liệu, cải thiện khả năng dự đoán.
  + Phân Tích Dữ Liệu Thời Gian (Time Series Analysis)
* ARIMA (AutoRegressive Integrated Moving Average):
  + Kỹ thuật phổ biến cho dữ liệu chuỗi thời gian, phù hợp cho việc dự đoán số lượng đặt phòng theo thời gian.
  + Giúp nhận diện các mẫu theo mùa và xu hướng dài hạn trong dữ liệu.
* Holt-Winters:
  + Phương pháp này cho phép dự đoán dữ liệu chuỗi thời gian có tính mùa vụ, bằng cách áp dụng trọng số khác nhau cho dữ liệu quá khứ.
  + Phân Tích Cảm Xúc (Sentiment Analysis)
* Phân tích dữ liệu từ các đánh giá trực tuyến để xác định cảm xúc của khách hàng đối với dịch vụ.
* Sử dụng các kỹ thuật NLP (Natural Language Processing) để trích xuất cảm xúc từ văn bản, giúp hiểu rõ hơn về yếu tố tác động đến quyết định đặt phòng.
  + Học Tăng Cường (Reinforcement Learning)
* Mô hình học tăng cường có thể được sử dụng để tối ưu hóa các quyết định về giá phòng và các chương trình khuyến mãi.
* Mô hình sẽ học từ các phản hồi của thị trường để cải thiện chiến lược kinh doanh theo thời gian.

## 2.3. Ứng Dụng Các Thuật Toán Dự Đoán Trong Dự Đoán Xu Hướng Đặt Phòng

Dưới đây là một cái nhìn chi tiết về các thuật toán dự đoán như hồi quy tuyến tính, cây quyết định, mạng nơ-ron nhân tạo, và mô hình gradient boosting, cùng với các ứng dụng cụ thể trong dự đoán xu hướng đặt phòng.

### 2.3.1. Hồi Quy Tuyến Tính (Linear Regression)

* **Mô tả**:
  + Hồi quy tuyến tính là một phương pháp thống kê tìm kiếm mối liên hệ tuyến tính giữa một biến phụ thuộc (số lượng đặt phòng) và một hoặc nhiều biến độc lập (giá phòng, thời tiết, sự kiện, v.v.).
* **Công thức**:
* **Ứng dụng**:
  + Dự đoán số lượng đặt phòng theo tháng dựa trên biến độc lập như giá phòng trung bình, thời tiết (nhiệt độ, độ ẩm), và các sự kiện địa phương (như lễ hội hay hội nghị).
  + Ví dụ: Nếu giá phòng của khách sạn giảm 10% trong mùa thấp điểm, mô hình có thể dự đoán rằng số lượng đặt phòng sẽ tăng 15%.
* **Ưu điểm**:
  + Đơn giản, dễ hiểu, và dễ triển khai.
  + Hiệu quả với dữ liệu nhỏ và mối quan hệ tuyến tính.
* **Nhược điểm**:
  + Hạn chế trong việc mô hình hóa các mối quan hệ phi tuyến tính.
  + Nhạy cảm với các giá trị ngoại lai (outliers).

### 2.3.2. Cây Quyết Định (Decision Trees)

* **Mô tả**:
  + Cây quyết định là mô hình dự đoán phân loại hoặc hồi quy, trong đó mỗi nút trong cây đại diện cho một quyết định dựa trên giá trị của một thuộc tính (như giá phòng hay loại phòng).
* **Quy trình xây dựng**:
  + **Chia dữ liệu**: Dữ liệu được chia thành các nhóm dựa trên các thuộc tính cụ thể.
  + **Tạo nút**: Mỗi nút trong cây đại diện cho một câu hỏi (ví dụ: "Giá phòng > 100$?").
  + **Duy trì tiếp tục**: Quá trình được lặp lại cho đến khi đạt đến một tiêu chí dừng (như không còn thuộc tính nào để chia hoặc độ sâu tối đa).
* **Ứng dụng**:
  + Dự đoán xem một khách hàng có khả năng đặt phòng hay không dựa trên các yếu tố như độ tuổi, địa điểm, và đánh giá.
  + Phân loại khách hàng thành các nhóm để cá nhân hóa các chương trình khuyến mãi.
* **Ưu điểm**:
  + Dễ hiểu và trực quan hóa, dễ dàng diễn giải các quyết định.
  + Có khả năng xử lý dữ liệu phi tuyến tính và không yêu cầu tiền xử lý dữ liệu phức tạp.
* **Nhược điểm**:
  + Dễ bị overfitting nếu cây quá sâu.
  + Khả năng tổng quát kém trên dữ liệu mới.

### 2.3.3. Mạng Nơ-ron Nhân Tạo (Artificial Neural Networks - ANN)

* **Mô tả**:
  + Mạng nơ-ron nhân tạo bao gồm nhiều lớp nơ-ron, giúp mô hình hóa các mối quan hệ phi tuyến tính phức tạp trong dữ liệu.
* **Cấu trúc**:
  + **Lớp đầu vào**: Nhận dữ liệu đầu vào (giá phòng, thời tiết, sự kiện).
  + **Lớp ẩn**: Chứa nhiều nơ-ron, xử lý và học các mẫu từ dữ liệu.
  + **Lớp đầu ra**: Dự đoán số lượng đặt phòng hoặc xác suất khách hàng sẽ đặt phòng.
* **Ứng dụng**:
  + Dự đoán xu hướng đặt phòng bằng cách phân tích dữ liệu từ nhiều nguồn (đánh giá khách hàng, giá phòng, sự kiện).
  + Dự đoán dựa trên hành vi khách hàng, chẳng hạn như thời gian lưu trú và loại phòng thường đặt.
* **Ưu điểm**:
  + Khả năng học sâu từ dữ liệu lớn và phức tạp.
  + Có thể áp dụng cho nhiều loại bài toán khác nhau, từ phân loại đến hồi quy.
* **Nhược điểm**:
  + Cần tập dữ liệu lớn để đào tạo hiệu quả.
  + Khó khăn trong việc giải thích kết quả do tính chất "hộp đen".

### 2.3.4. Mô Hình Gradient Boosting (như XGBoost)

* **Mô tả**:
  + Mô hình gradient boosting kết hợp nhiều cây quyết định yếu để tạo ra một mô hình mạnh hơn thông qua quá trình học từ sai số của các cây trước đó.
* **Quy trình hoạt động**:
  + **Tạo cây đầu tiên**: Mô hình bắt đầu với một cây quyết định đơn giản.
  + **Học từ sai số**: Mỗi cây tiếp theo học từ sai số (residuals) của cây trước đó, nhằm cải thiện độ chính xác.
  + **Kết hợp**: Cuối cùng, các cây được kết hợp để đưa ra dự đoán cuối cùng.
* **Ứng dụng**:
  + Dự đoán số lượng đặt phòng với độ chính xác cao bằng cách kết hợp nhiều yếu tố.
  + Sử dụng trong các cuộc thi dữ liệu để tối ưu hóa mô hình dự đoán nhờ vào khả năng tổng quát tốt.
* **Ưu điểm**:
  + Thường cho kết quả tốt hơn so với cây quyết định đơn giản hoặc hồi quy tuyến tính.
  + Có khả năng xử lý các biến không có cấu trúc và dữ liệu thiếu.
* **Nhược điểm**:
  + Cần thời gian đào tạo lâu hơn so với hồi quy tuyến tính hoặc cây quyết định đơn giản.
  + Có thể dễ bị overfitting nếu không điều chỉnh tham số hợp lý.

## 2.4. Đánh Giá Hiệu Quả Mô Hình Dự Đoán

Đánh giá hiệu quả mô hình dự đoán là một bước thiết yếu trong quá trình phát triển và triển khai mô hình. Điều này không chỉ giúp xác định độ chính xác của mô hình mà còn đảm bảo rằng mô hình có thể tổng quát tốt trên dữ liệu chưa thấy. Dưới đây là các bước chi tiết để đánh giá mô hình dự đoán trong lĩnh vực dự đoán xu hướng đặt phòng.

### 2.4.1. Chia Dữ Liệu

Việc chia dữ liệu là bước đầu tiên và quan trọng trong quá trình xây dựng mô hình.

* **Chia thành Tập Huấn Luyện và Tập Kiểm Tra**:
  + **Tập huấn luyện**: Thường chiếm 70-80% tổng dữ liệu, được sử dụng để xây dựng và huấn luyện mô hình.
  + **Tập kiểm tra**: Chiếm 20-30% còn lại, được sử dụng để đánh giá hiệu suất của mô hình trên dữ liệu chưa thấy.
* **Ví dụ**: Nếu bạn có 1000 bản ghi dữ liệu, bạn có thể chia thành 800 bản ghi cho tập huấn luyện và 200 bản ghi cho tập kiểm tra.

### 2.4.2. Các Chỉ Số Đánh Giá Mô Hình

Các chỉ số này giúp đánh giá hiệu suất của mô hình dự đoán.

* **Đối với Mô Hình Hồi Quy**:
  + **Mean Absolute Error (MAE)**:
    - Đo lường độ chính xác trung bình của dự đoán, với độ lệch tuyệt đối giữa giá trị thực và giá trị dự đoán.
  + **Mean Squared Error (MSE)**:
    - Nhấn mạnh các sai số lớn hơn, do đó, các giá trị ngoại lai sẽ ảnh hưởng nhiều hơn đến kết quả.
  + **R-squared (R²)**:
    - Mô tả tỷ lệ biến thiên trong dữ liệu mà mô hình giải thích. Giá trị R² nằm trong khoảng từ 0 đến 1, với giá trị càng gần 1 thì mô hình càng tốt.
* **Đối với Mô Hình Phân Loại**:
  + **Accuracy**:
    - Tỷ lệ dự đoán đúng so với tổng số dự đoán.
  + **Precision**:
    - Tỷ lệ giữa số dự đoán dương tính đúng và tổng số dự đoán dương tính.
  + **Recall (Sensitivity)**:
    - Tỷ lệ giữa số dự đoán dương tính đúng và tổng số trường hợp thực sự dương tính.
  + **F1 Score**:
    - Chỉ số tổng hợp giữa Precision và Recall, thường được sử dụng trong các bài toán phân loại không cân bằng.
  + **ROC Curve và AUC**:
    - ROC Curve là đồ thị biểu diễn tỷ lệ True Positive Rate (TPR) và False Positive Rate (FPR) tại các ngưỡng khác nhau. AUC (Area Under the Curve) cho biết khả năng phân loại đúng của mô hình.

### 2.4.3. Kỹ Thuật Cross-Validation

Cross-validation là phương pháp đánh giá mô hình mạnh mẽ giúp giảm thiểu thiên lệch trong đánh giá hiệu suất.

* **K-fold Cross-Validation**:
  + Dữ liệu được chia thành k phần (folds). Mô hình được huấn luyện trên k-1 phần và kiểm tra trên phần còn lại. Quá trình này lặp lại k lần.
* **Lợi ích**:
  + Cung cấp ước lượng chính xác hơn về độ chính xác của mô hình.
  + Giúp phát hiện vấn đề overfitting và underfitting.

### 2.4.4. Kiểm Tra Trên Dữ Liệu Mới

Sau khi mô hình đã được huấn luyện và đánh giá bằng dữ liệu kiểm tra, cần kiểm tra mô hình trên một tập dữ liệu hoàn toàn mới.

* **Tại sao quan trọng**:
  + Đảm bảo rằng mô hình không chỉ hoạt động tốt trên dữ liệu đã thấy mà còn có khả năng tổng quát trên dữ liệu thực tế.

### 2.4.5. Phân Tích Kết Quả

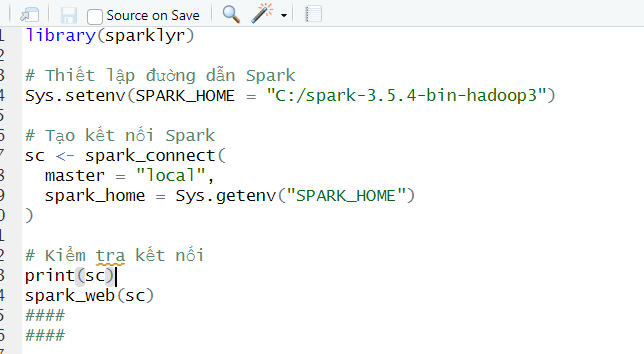
Sau khi có các chỉ số đánh giá, bước tiếp theo là phân tích kết quả.

* **So sánh các mô hình**:
  + So sánh hiệu suất của nhiều mô hình khác nhau để xác định mô hình nào hoạt động tốt nhất cho dữ liệu cụ thể.
* **Điều chỉnh mô hình**:
  + Dựa trên kết quả đánh giá, có thể cần điều chỉnh mô hình (ví dụ: thay đổi tham số, thêm biến số) để cải thiện hiệu suất.

# CHƯƠNG 3. THỰC HIỆN CHƯƠNG TRÌNH VÀ KẾT QUẢ

## 3.1. Thực hiện chương trình

### 3.1.1. Kết nối với Spark



#### Hình 2. Kết nối với Spark.

Giải thích:

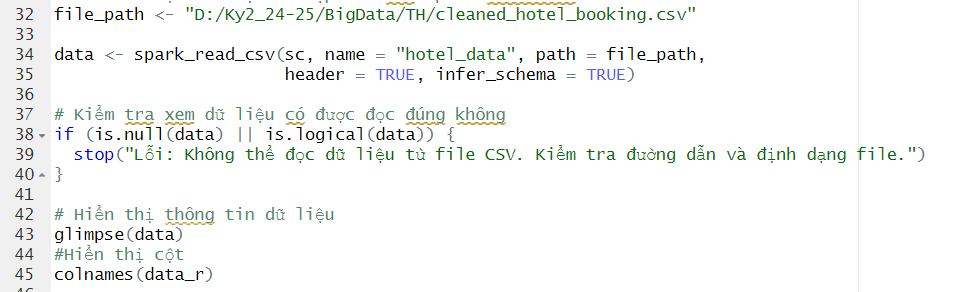
 **sparklyr** là thư viện hỗ trợ tích hợp Spark với R.

 **Sys.setenv(SPARK\_HOME)** thiết lập đường dẫn cài đặt Spark.

 **spark\_connect()** tạo kết nối với Spark trên máy cục bộ.

 **spark\_web(sc)** mở giao diện web của Spark để kiểm tra thông tin tiến trình.

### 3.1.2. Đọc và kiểm tra dữ liệu từ CSV



#### Hình 3. Đọc và kiểm tra dữ liệu từ CSV.

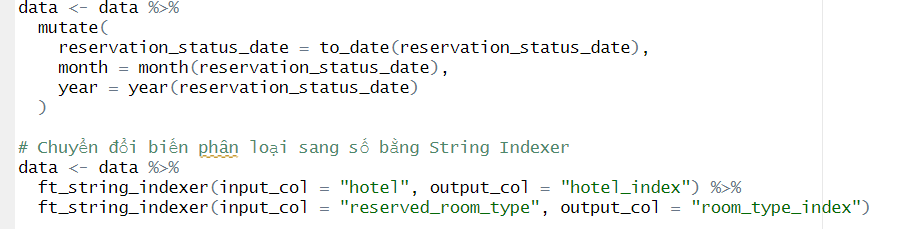
Giải thích:

 **spark\_read\_csv()** đọc dữ liệu CSV vào Spark DataFrame.

 **glimpse(data)** hiển thị thông tin tổng quan về dữ liệu.

 **colnames(data\_r)** kiểm tra danh sách các cột.

### 3.1.3. Xử lý dữ liệu



#### Hình 4. Xử lí dữ liệu.

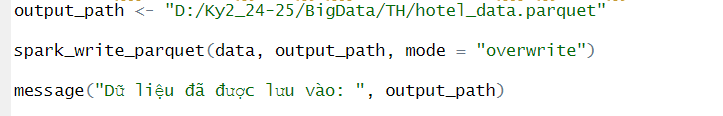
Giải thích:

 Chuyển đổi cột **reservation\_status\_date** thành kiểu ngày.

 Tạo thêm hai cột **month** và **year** để phân tích theo tháng/năm.

 **ft\_string\_indexer()** chuyển đổi dữ liệu dạng chuỗi (hotel, room\_type) thành số để sử dụng cho mô hình Machine Learning.

### 3.1.4. Lưu dữ liệu đã xử lý



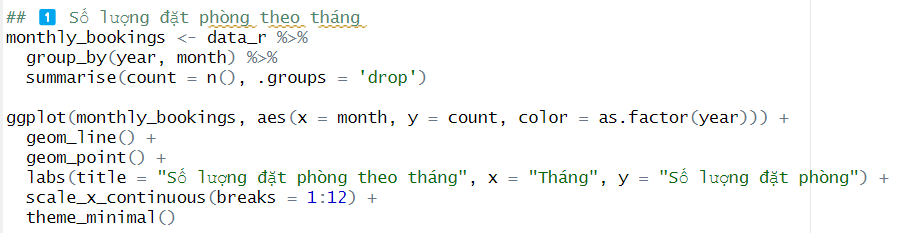
#### Hình 5. Lưu dữ liệu đã xử lý.

Giải thích:

Dữ liệu đã xử lý được lưu dưới định dạng **Parquet** (tối ưu hơn CSV về tốc độ truy xuất).

### 3.1.5. Trực quan hóa dữ liệu

- Biểu đồ 1: Số lượng đặt phòng theo tháng



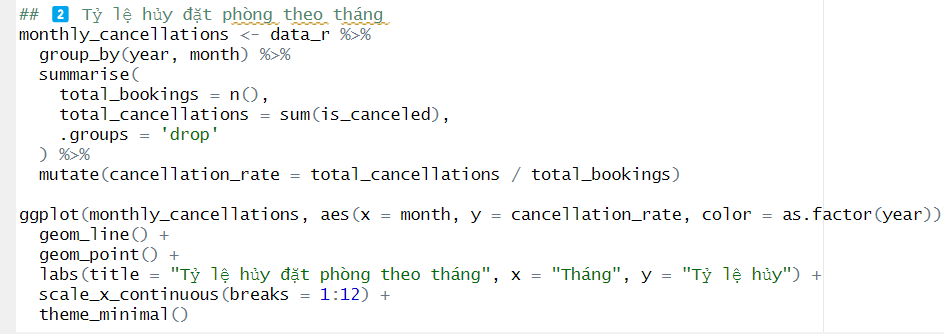
#### Hình 6. Trực quan hóa dữ liệu .

Giải thích:

 Nhóm dữ liệu theo **tháng** và **năm** để đếm số lượng đặt phòng.

 **ggplot()** tạo biểu đồ **đường** hiển thị xu hướng đặt phòng theo tháng.

* Biểu đồ 2: Tỷ lệ hủy đặt phòng theo tháng



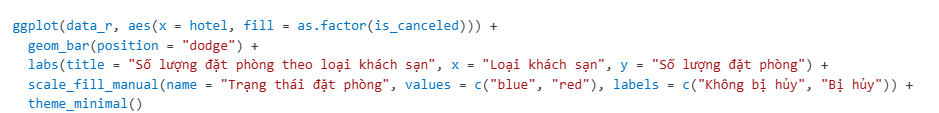
#### Hình 7. Tỷ lệ hủy theo tháng.

Giải thích:

 Tính toán tỷ lệ hủy đặt phòng hàng tháng.

 **geom\_line()** vẽ xu hướng hủy đặt phòng qua thời gian.

* Biểu đồ 3: Số lượng đặt phòng theo loại khách sạn

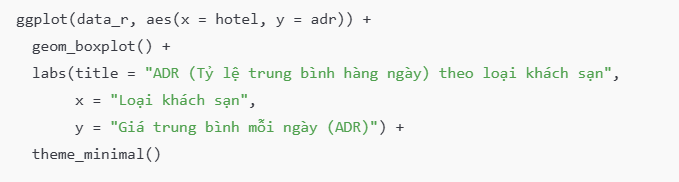


#### Hình 8. Số lượng đặt phòng theo loại khách sạn.

Giải thích:

Biểu đồ **cột** thể hiện số lượng đặt phòng của từng loại khách sạn, phân biệt theo trạng thái hủy hay không.

* Biểu đồ 4: Phân phối ADR (Giá trung bình mỗi ngày) theo loại khách sạn



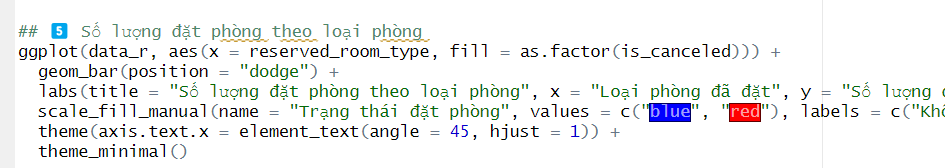
#### Hình 9. Biểu đồ phân phối ADR.

Giải thích:

 **ADR (Average Daily Rate)** là giá trung bình mỗi ngày cho một phòng khách sạn.

 **Biểu đồ Boxplot** giúp quan sát phân phối ADR của từng loại khách sạn.

* Biểu đồ 5: Số lượng đặt phòng theo loại phòng



#### Hình 10. Số lượng đặt phòng theo loại phòng.

Giải thích:

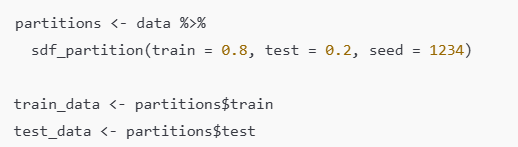
 **Sử dụng biểu đồ cột** để so sánh số lượng đặt phòng theo từng loại phòng.

 Phân biệt trạng thái đặt phòng (bị hủy hay không) bằng màu sắc.

 **Dodge position** giúp hiển thị dữ liệu rõ ràng hơn.

### 3.1.6. Phân tích mô hình Random Forest trên dữ liệu đặt phòng khách sạn

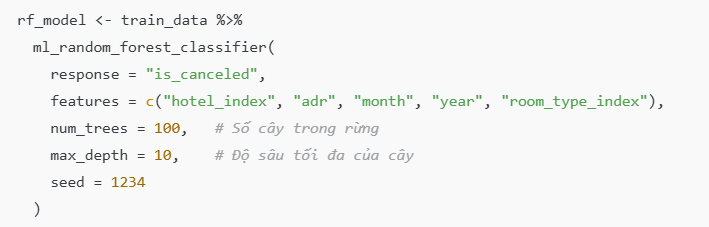
- Chia dữ liệu thành tập huấn luyện (80%) và kiểm tra (20%)



#### Hình 11. Chia dữ liệu thành tập huấn luyện.

Giải thích:

* Chia tập dữ liệu thành **80% huấn luyện** và **20% kiểm tra** để đảm bảo mô hình có dữ liệu để học và đánh giá.
* **seed = 1234** giúp tái lập kết quả nhất quán.
* Huấn luyện mô hình Random Forest



#### Hình 12. Huấn luyện mô hình Ramdom Forest.

Giải thích:

 **Mô hình Random Forest** được huấn luyện để **dự đoán khả năng hủy đặt phòng**.

 Các tham số:

* **num\_trees = 100**: Sử dụng 100 cây để tăng độ chính xác.
* **max\_depth = 10**: Độ sâu tối đa của cây giúp kiểm soát overfitting.
* **features** gồm:
  + - **hotel\_index**: Loại khách sạn.
    - **adr**: Giá trung bình mỗi ngày.
    - **month, year**: Thời gian đặt phòng.
    - **room\_type\_index**: Loại phòng đặt.
* Dự đoán trên tập kiểm tra



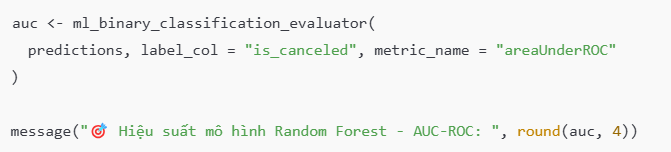
#### Hình 13. Dự đoán trên tập kiểm tra.

Giải thích:

 Mô hình **dự đoán xem đặt phòng có bị hủy hay không** dựa trên dữ liệu kiểm tra.

 Kết quả dự đoán sẽ bao gồm:

* **is\_canceled** (giá trị thực tế)
* **prediction** (giá trị dự đoán)
* **probability** (xác suất dự đoán)
* Đánh giá mô hình bằng AUC-ROC



#### Hình 14. Đánh giá mô hình bằng AUC-ROC.

Giải thích:

 **AUC-ROC (Area Under Curve - Receiver Operating Characteristic)** đo độ chính xác của mô hình.

 **Giá trị AUC**:

* **0.5 - 0.6**: Mô hình yếu.
* **0.6 - 0.7**: Mô hình trung bình.
* **0.7 - 0.8**: Mô hình khá tốt.
* **0.8 - 0.9**: Mô hình tốt.
* **> 0.9**: Mô hình xuất sắc.
* Xem dữ liệu Spark



#### Hình 15. Xem dữ liệu Spark.

Giải thích:

 **Xem dữ liệu trong Spark** để đảm bảo dữ liệu đã được tải và xử lý đúng.

 **collect()** chuyển dữ liệu từ Spark về R để dễ dàng trực quan hóa.

* Ngắt kết nối Spark



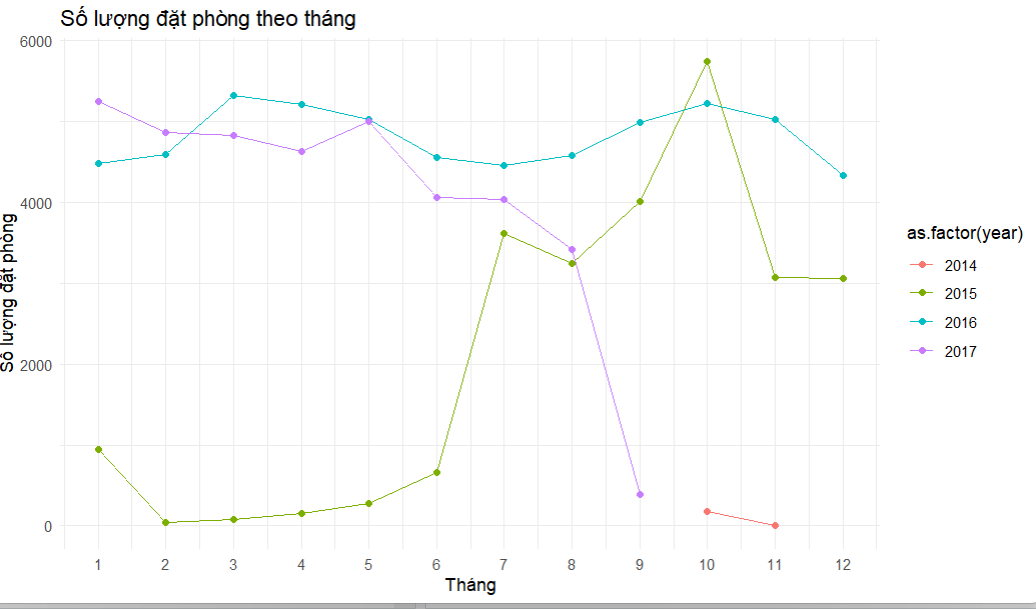
#### Hình 16. Ngắt kết nối Spark.

Giải thích:

* Giải phóng tài nguyên Spark sau khi hoàn thành phân tích.

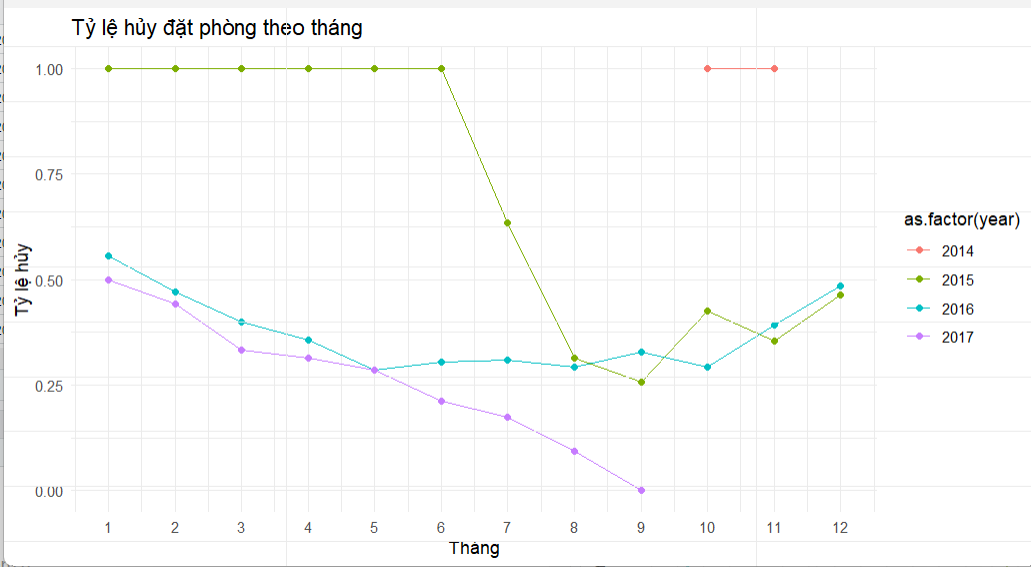
## 3.2. Kết quả

- Biểu đồ 1: Số lượng đặt phòng theo tháng



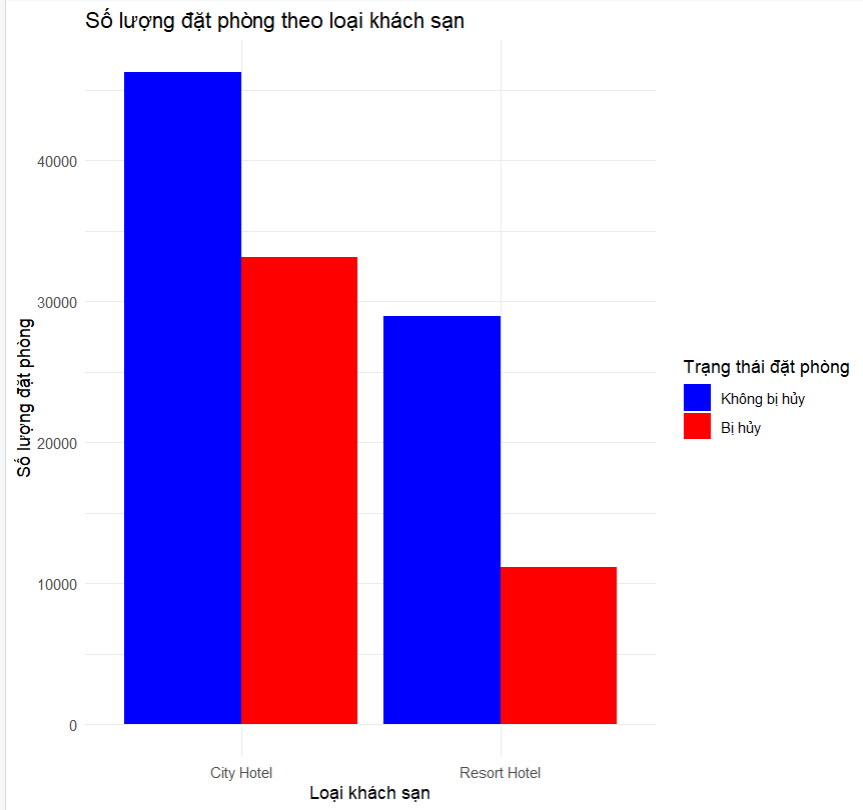
#### Hình 17. Kết quả biểu đồ số lượng đặt phong theo tháng.

- Biểu đồ 2: Tỷ lệ hủy đặt phòng theo tháng



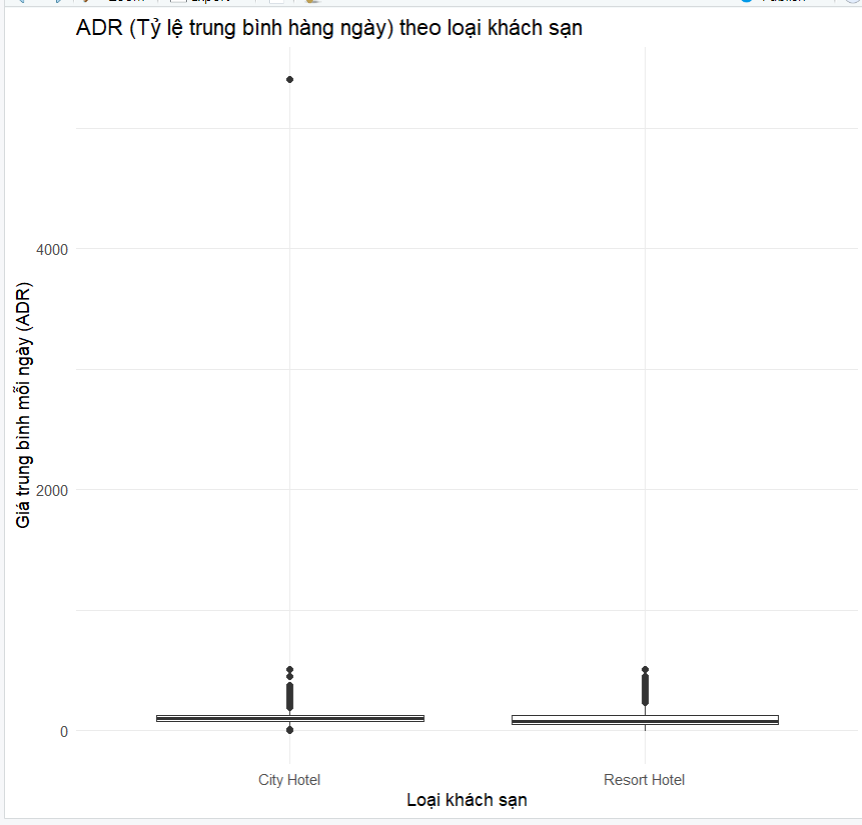
#### Hình 18.kết quả biểu đồ tỷ lệ hủy đặt phòng theo tháng.

- Biểu đồ 3: Số lượng đặt phòng theo loại khách sạn



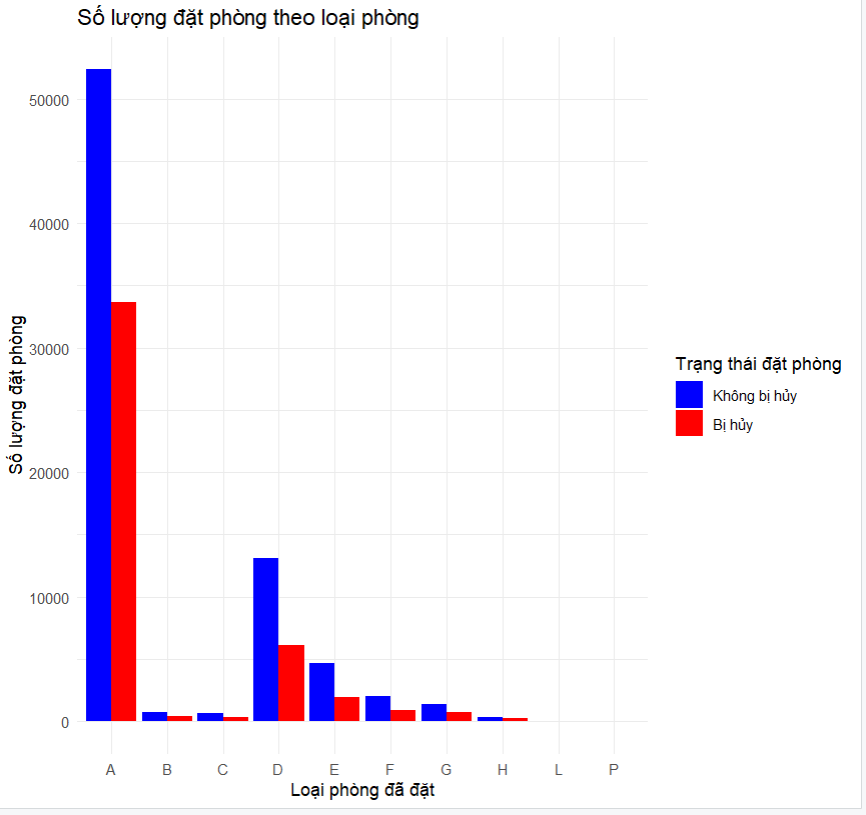
#### Hình 19. Kết quả biểu đồ số lượng đặt phòng theo loại khách sạn.

- Biểu đồ 4: Phân phối ADR (Giá trung bình mỗi ngày) theo loại khách sạn



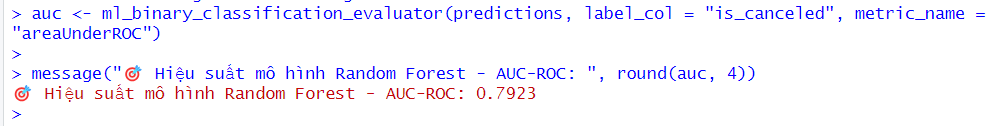
#### Hình 20. Kết quả biểu đồ phân phối ADR.

- Biểu đồ 5: Số lượng đặt phòng theo loại phòng



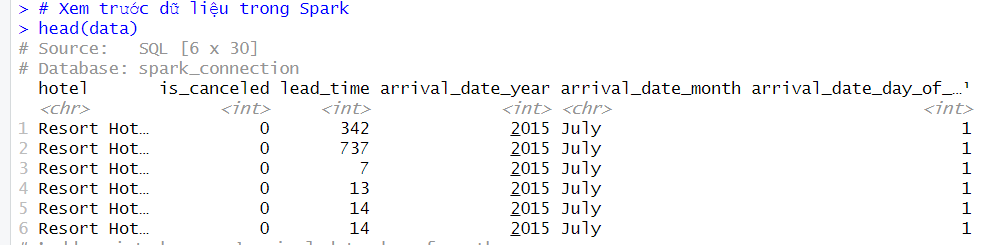
#### Hình 21. Biểu đồ số lượng đặt phòng theo loại phòng.

* Hiệu xuất mô hình



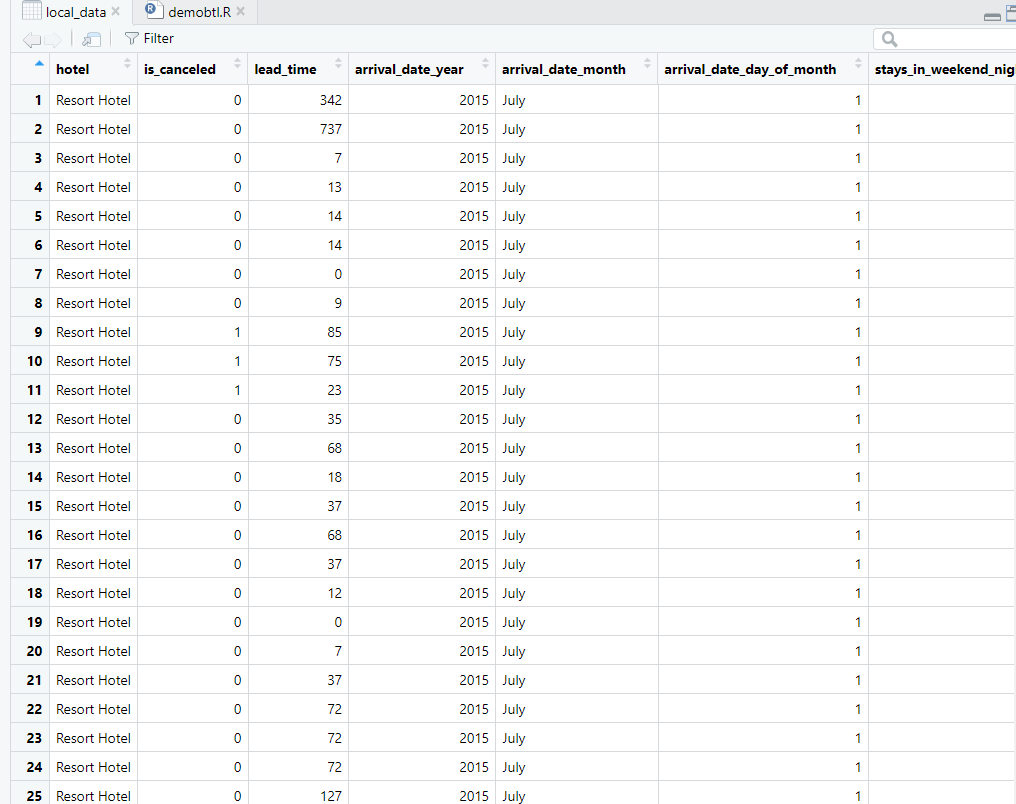
#### Hình 22. Hiệu suất mô hình

* Xem dữ liệu từ spark



#### Hình 23. Xem dữ liệu từ Spark.

* Xem bảng dữ liệu



#### Hình 24. Xem bảng dữ liệu.

# KẾT LUẬN

Việc áp dụng phân tích dữ liệu lớn vào lĩnh vực khách sạn giúp tối ưu hóa chiến lược kinh doanh và nâng cao trải nghiệm khách hàng. Bằng cách khai thác dữ liệu đặt phòng, khách sạn có thể dự đoán xu hướng đặt phòng theo mùa, tối ưu giá cả dựa trên nhu cầu thực tế và triển khai các chiến dịch khuyến mãi hiệu quả. Ngoài ra, việc phân tích tỷ lệ hủy phòng giúp khách sạn điều chỉnh chính sách đặt phòng và giảm thiểu tổn thất doanh thu. Hơn nữa, công nghệ Big Data kết hợp với mô hình AI như Random Forest giúp nâng cao độ chính xác trong dự báo hành vi khách hàng, hỗ trợ ra quyết định chiến lược.

Mặc dù phân tích dữ liệu lớn mang lại nhiều lợi ích, nhưng việc triển khai cũng gặp không ít thách thức. Thứ nhất, yêu cầu về cơ sở hạ tầng mạnh mẽ, bao gồm phần cứng và phần mềm, có thể tốn kém đối với các khách sạn nhỏ. Thứ hai, dữ liệu khách sạn thường rất đa dạng và không đồng nhất, dẫn đến khó khăn trong việc làm sạch và chuẩn hóa dữ liệu trước khi phân tích. Thứ ba, vấn đề bảo mật dữ liệu và quyền riêng tư của khách hàng là một mối lo ngại lớn, đặc biệt khi xử lý dữ liệu cá nhân. Ngoài ra, việc áp dụng mô hình AI đòi hỏi đội ngũ nhân sự có chuyên môn cao, điều này có thể là một rào cản đối với các doanh nghiệp chưa quen thuộc với công nghệ này.

Trong tương lai, hệ thống phân tích dữ liệu lớn trong khách sạn có thể được mở rộng bằng cách tích hợp thêm các nguồn dữ liệu bên ngoài như phản hồi khách hàng, dữ liệu mạng xã hội và dự báo thời tiết để đưa ra dự đoán chính xác hơn. Bên cạnh đó, việc ứng dụng các mô hình AI tiên tiến như deep learning có thể giúp nâng cao khả năng cá nhân hóa trải nghiệm khách hàng, gợi ý dịch vụ phù hợp với từng cá nhân. Ngoài ra, hệ thống có thể được triển khai trên nền tảng đám mây để tăng khả năng mở rộng và tối ưu hóa chi phí. Cuối cùng, việc phát triển dashboard trực quan giúp quản lý khách sạn dễ dàng theo dõi và đưa ra quyết định dựa trên dữ liệu theo thời gian thực.

# TÀI LIỆU THAM KHẢO

[1] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[2] Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.

[3] Moro, S., Rita, P., & Vala, B. (2016). Predicting hotel booking cancellations using data mining techniques. International Journal of Hospitality Management, 51, 89-96.

[4] McKinsey & Company. (2021). The future of hotel revenue management. Retrieved from <https://www.mckinsey.com>

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.