TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI

ĐỒ ÁN TỐT NGHIỆP

ỨNG DỤNG MÔ HÌNH ARIMA VÀ PROPHET TRONG VIỆC DỰ BÁO NHU CẦU ĐẶT PHÒNG KHÁCH SẠN

Giảng viên hướng dẫn: PGS. TS. Trần Văn Long

Thực hiện: Nguyễn Đức Anh

Mã sinh viên: 213010736

Lớp: Toán Ứng Dụng 62

HÀ NỘI, 2025

Nhận xét của giảng viên hướng dẫn

1. Mục tiêu và nội dung của đồ án:

. . . . . .

. . . . . .

. . . . . .

. . . . . .

2. Kết quả đạt được:

. . . . . .

. . . . . .

. . . . . .

. . . . . .

3. Ý thức làm việc của sinh viên:

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Hà Nội, ngày ... tháng ... năm 2025

Giảng viên hướng dẫn

Lời cảm ơn

Trong thời gian bốn năm theo học chương trình Toán ứng dụng tại Đại học Giao thông Vận tải - Hà Nội, đây là quá trình đầy nỗ lực và cố gắng tìm kiếm hướng đi đúng cho bản thân. Sau một thời gian dài tích lũy những kiến thức chuyên ngành và kinh nghiệm thực tế từ thầy cô thì kết quả được đúc kết trong đề tài này. Em xin chân thành cảm ơn PGS.TS. Trần Văn Long đã tận tình hướng dẫn, chỉ bảo và tạo điều kiện thuận lợi trong suốt quá trình học tập và nghiên cứu. Sự hỗ trợ và những lời khuyên quý báu của thầy đã giúp cho em nâng cao kiến thức và hoàn thành bài báo cáo này. Em cũng xin gửi lời cảm ơn đến Khoa Khoa học Cơ bản, Trường Đại học Giao thông Vận tải, vì đã cung cấp môi trường học tập, nghiên cứu thuận lợi và tạo mọi điều kiện tốt nhất để chúng em phát triển và hoàn thiện bản thân.

Do kinh nghiệm và kiến thức còn hạn chế, đồ án này không tránh khỏi những thiếu sót. Em mong có thể nhận được sự góp ý từ thầy cô để em có thể hoàn thành tốt hơn những nhiệm vụ sau này. Em xin chân thành cảm ơn!

Hà Nội, ngày 20 tháng 5 năm 2025

Nguyễn Đức Anh

Tóm tắt

Đồ án tốt nghiệp này tập trung vào việc áp dụng các mô hình chuỗi thời gian gồm mô hình Prophet và mô hình ARIMA để dự báo nhu cầu đặt phòng khách sạn, một chủ đề có tính ứng dụng cao trong lĩnh vực du lịch lưu trú. Đồ án được chia thành 3 chương chính.

Trong chương 1: Chương này cung cấp một cái nhìn tổng quan về thị trường khách sạn và tầm quan trọng của dự báo nhu cầu đặt phòng. Tiếp theo, chương giới thiệu về chuỗi thời gian, một dạng dữ liệu quan trọng trong dự báo.

Trong chương 2: Chương này đi chi tiết về hai mô hình chuỗi thời gian Prophet và ARIMA. Trình bày những định nghĩa và cơ sở toán học của cả hai mô hình.

Trong chương 3: Chương này trình bày chi tiết về dữ liệu của giao dịch đặt phòng khách sạn đầu vào được sử dụng trong đồ án.Phân tích dữ liệu đầu vào và sử dụng dữ liệu để huấn luyện hai mô hình Prophet và ARIMA để dự báo số người đặt phòng, bao gồm đánh giá về độ chính xác của dự báo.

Mục tiêu chính của đồ án: Xây dựng một mô hình dự báo nhu cầu đặt phòng khách sạn chính xác, giúp người kinh doanh và các bên liên quan đưa ra quyết định thông minh. Đánh giá hiệu quả của mô hình Prophet và mô hình ARIMA trong việc dự báo nhu cầu đặt phòng. Đóng góp vào nghiên cứu ứng dụng các mô hình dự báo chuỗi thời gian trong lĩnh vực kinh doanh lưu trú.

Mục lục

Danh sách hình vẽ

Danh sách bảng

Danh Mục Các Từ Viết Tắt

Danh sách các từ viết tắt và ý nghĩa tương ứng

Viết tắt	**Ý nghĩa**
ML	Machine Learning – Học máy
AI	Artificial Intelligence – Trí tuệ nhân tạo
AR	Autoregressive – Tự hồi quy
I	Integrated – Tích phân (sai phân để làm dừng chuỗi)
MA	Moving Average – Trung bình trượt
ARMA	Autoregressive Moving Average – Mô hình tự hồi quy kết hợp trung bình trượt
ARIMA	Autoregressive Integrated Moving Average – Mô hình tự hồi quy kết hợp trung bình trượt và sai phân
MAE	Mean Absolute Error – Sai số tuyệt đối trung bình
MAPE	Mean Absolute Percentage Error – Sai số phần trăm tuyệt đối trung bình
ACF	Autocorrelation Function – Hàm tự tương quan
PACF	Partial Autocorrelation Function – Hàm tự tương quan riêng phần
AIC	Akaike Information Criterion – Tiêu chí thông tin Akaike
BIC	Bayesian Information Criterion – Tiêu chí thông tin Bayes
ADF	Augmented Dickey-Fuller – Kiểm định Dickey-Fuller mở rộng (dùng để kiểm tra tính dừng)

CHƯƠNG 1

Cơ sở lý thuyết

Sự phát triển mạnh mẽ của công nghệ thông tin trong những thập kỷ gần đây, đặc biệt là sự bùng nổ của dữ liệu lớn (Big Data) và các tiến bộ vượt bậc trong lĩnh vực trí tuệ nhân tạo (AI) và học máy (ML), đã tạo ra những thay đổi sâu rộng trong cách thức các tổ chức và doanh nghiệp tiếp cận, khai thác và sử dụng dữ liệu nhằm tối ưu hóa hoạt động kinh doanh và nâng cao hiệu quả quản lý. Trong bối cảnh đó, ngành dịch vụ khách sạn, với đặc thù là lĩnh vực chịu ảnh hưởng mạnh mẽ từ các yếu tố mang tính mùa vụ, xu hướng du lịch và hành vi tiêu dùng, đang đứng trước nhiều cơ hội để áp dụng các phương pháp phân tích và dự báo hiện đại nhằm nâng cao khả năng dự báo nhu cầu và cải thiện chất lượng dịch vụ.

Trong lĩnh vực dự báo chuỗi thời gian, việc ứng dụng các mô hình thống kê và thuật toán học máy đang dần trở nên phổ biến, nhờ khả năng xử lý dữ liệu lớn và khai thác hiệu quả các đặc điểm cấu trúc của chuỗi thời gian, bao gồm xu hướng dài hạn, tính chu kỳ và tính mùa vụ. Trong số các công cụ và phương pháp được sử dụng phổ biến hiện nay, mô hình Prophet – do Facebook phát triển – và mô hình ARIMA – một mô hình kinh điển trong phân tích chuỗi thời gian, cả hai được đánh giá là hai công cụ có độ linh hoạt và hiệu quả cao trong việc dự báo dữ liệu có yếu tố mùa vụ và xu hướng biến động theo thời gian.

Nghiên cứu này tập trung vào việc phân tích, so sánh và đánh giá hiệu quả của hai mô hình Prophet và ARIMA trong bối cảnh dự báo nhu cầu đặt phòng trong ngành khách sạn. Dữ liệu đầu vào được sử dụng trong nghiên cứu là tập dữ liệu thực tế ghi nhận lịch sử đặt phòng tại các cơ sở lưu trú, từ đó tiến hành xử lý, phân tích và xây dựng mô hình dự báo. Quá trình thực nghiệm bao gồm các bước như tiền xử lý dữ liệu, huấn luyện mô hình, dự báo và đánh giá hiệu suất dự báo thông qua các chỉ số định lượng như MAE, MAPE,... nhằm đảm bảo độ khách quan và chính xác trong việc so sánh các mô hình.

Kết quả nghiên cứu không chỉ giúp xác định mô hình phù hợp hơn cho bài toán dự báo nhu cầu đặt phòng trong ngành lưu trú mà còn cung cấp cơ sở khoa học và thực tiễn cho việc ứng dụng các phương pháp dự báo hiện đại vào hoạt động quản trị và lập kế hoạch kinh doanh. Việc áp dụng thành công các mô hình chuỗi thời gian không những góp phần nâng cao hiệu quả khai thác nguồn lực, mà còn giúp các doanh nghiệp trong ngành khách sạn tăng cường năng lực cạnh tranh trong môi trường kinh doanh ngày càng biến động.

1.1 Giới thiệu về thị trường kinh doanh lưu trú

1.1.1 Định nghĩa

Thị trường kinh doanh lưu trú là một bộ phận quan trọng của ngành dịch vụ, tập trung vào việc cung cấp dịch vụ chỗ ở có trả phí cho khách hàng trong một khoảng thời gian nhất định. Các cơ sở kinh doanh lưu trú rất đa dạng về quy mô, hình thức và dịch vụ cung cấp, bao gồm khách sạn, khu nghỉ dưỡng, nhà nghỉ, căn hộ dịch vụ, homestay và các loại hình khác.
Các đặc điểm của thị trường kinh doanh lưu trú bao gồm:

- **Tính vô hình:** Dịch vụ lưu trú là vô hình, khách hàng không thể kiểm tra trước khi tiêu dùng. Trải nghiệm của khách hàng phụ thuộc vào nhiều yếu tố như chất lượng phòng, thái độ phục vụ, tiện nghi và không gian.
- **Tính không tách rời:** Quá trình cung cấp và tiêu thụ dịch vụ thường diễn ra đồng thời. Sự tương tác giữa nhân viên và khách hàng đóng vai trò quan trọng trong việc tạo ra giá trị.
- **Tính không đồng nhất:** Chất lượng dịch vụ có thể khác nhau tùy thuộc vào người cung cấp, thời điểm cung cấp và khách hàng nhận dịch vụ. Việc tiêu chuẩn hóa chất lượng là một thách thức lớn.
- **Tính dễ hư hỏng:** Phòng trống là một nguồn lực không thể lưu trữ. Nếu một phòng không được bán trong một đêm, doanh thu tiềm năng từ đêm đó sẽ mất đi vĩnh viễn.

1.1.2 Tổng quan về ngành khách sạn và vai trò của dự báo nhu cầu đặt phòng

Ngành khách sạn đóng vai trò then chốt trong sự phát triển của ngành du lịch và đóng góp đáng kể vào nền kinh tế quốc gia thông qua việc tạo ra việc làm, thu hút đầu tư và thúc đẩy các ngành kinh tế liên quan. Sự cạnh tranh trong ngành ngày càng gay gắt, đòi hỏi các nhà quản lý khách sạn phải đưa ra các quyết định kinh doanh thông minh và kịp thời để tối ưu hóa hiệu quả hoạt động và tăng cường lợi nhuận.Dự báo nhu cầu đặt phòng là một công cụ quản lý chiến lược thiết yếu trong ngành khách sạn. Việc dự đoán chính xác số lượng phòng cần thiết trong tương lai giúp khách sạn có thể quản lí được doanh thu, giúp điều chỉnh được giá phòng linh hoạt và hợp lí để tối ưu hóa doanh thu. Lập kế hoạch nhân sự đảm bảo có đủ nhân sự cần thiết để phục vụ khách hàng trong thời gian cao điểm và tránh lãng phí nguồn lực trong thời gian thấp điểm. Hơn nữa, dự báo được nhu cầu đặt phòng của khách sạn có thể giúp lập được kế hoạch marketing và quảng bả nhằm mục tiêu các chiến dịch marketing đúng thời điểm và đối tượng khách hàng phù hợp.

1.2 Chuỗi thời gian

1.2.1 Giới thiệu về mô hình chuỗi thời gian

Một chuỗi thời gian là một tập các quan sát về giá trị của một biến nhận được tại những thời điểm liên tiếp(cách đều) nhau. Dự báo chuỗi thời gian là một lớp mô hình quan trọng trong thống kê, kinh tế lượng và machine learning. Sở dĩ gọi lớp mô hình này là chuỗi thời gian là vì mô hình được áp dụng trên các chuỗi đặc thù có yếu tố thời gian. Một mô hình chuỗi thời gian thường dự báo dựa trên giả định rằng các quy luật trong quá khứ sẽ lặp lại ở tương lai. Do đó, xây dựng mô hình chuỗi thời gian là chúng ta đang mô hình hóa mối quan hệ trong quá khứ giữa biến độc lập (biến đầu vào) và biến phụ thuộc (biến mục tiêu). Dựa vào mối quan hệ này để dự đoán giá trị trong tương lai của biến phụ thuộc.

Do là dữ liệu chịu ảnh hưởng bởi tính chất thời gian nên chuỗi thời gian thường xuất hiện những quy luật đặc trưng như : yếu tố chu kỳ, mùa vụ và yếu tố xu hướng. Đây là những đặc trưng thường thấy và xuất hiện ở hầu hết các chuỗi thời gian.Bên cạnh đó, chuỗi thời gian còn có những thuộc tính cơ bản như thuộc tính ngẫu nhiên, tự tương quan, dừng hay không dừng.

Yếu tố chu kỳ và mùa vụ

Yếu tố chu kỳ, mùa vụ là những đặc tính lặp lại theo chu kỳ. Ví dụ như nhiệt độ trung bình các tháng trong năm sẽ chịu ảnh hưởng bởi các mùa xuân, hạ, thu, đông. Hay xuất nhập khẩu của một quốc gia thường có chu kỳ theo các quý.

Yếu tố xu hướng

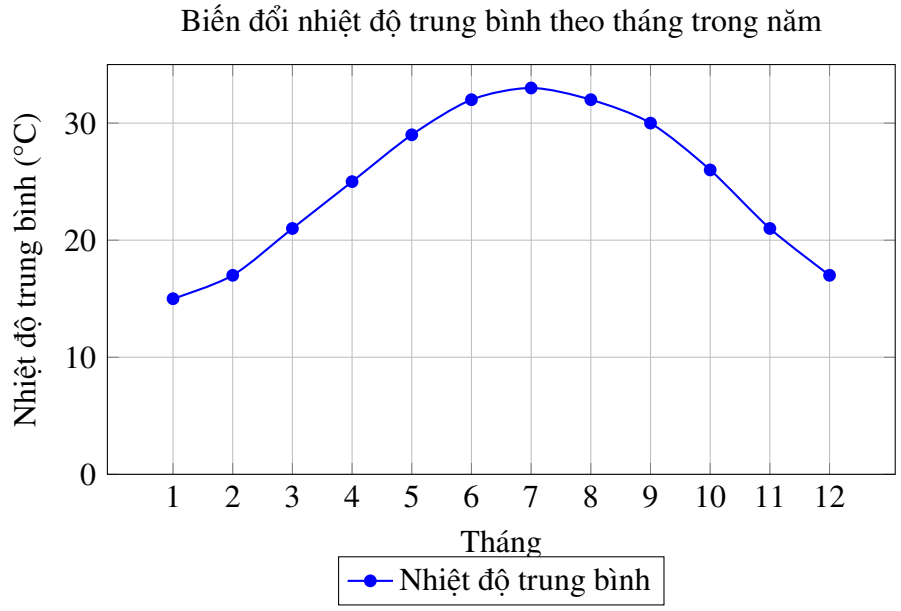
Yếu tố xu hướng thể hiện đà tăng hoặc giảm của chuỗi trong tương lai. Chẳng hạn như lạm phát là xu hướng chung của các nền kinh tế, do đó giá cả trung bình của giỏ hàng hóa cơ sở hay còn gọi là chỉ số CPI luôn có xu hướng tăng và xu hướng tăng này đại diện cho sự mất giá của đồng tiền.

Thuộc tính ngẫu nhiên

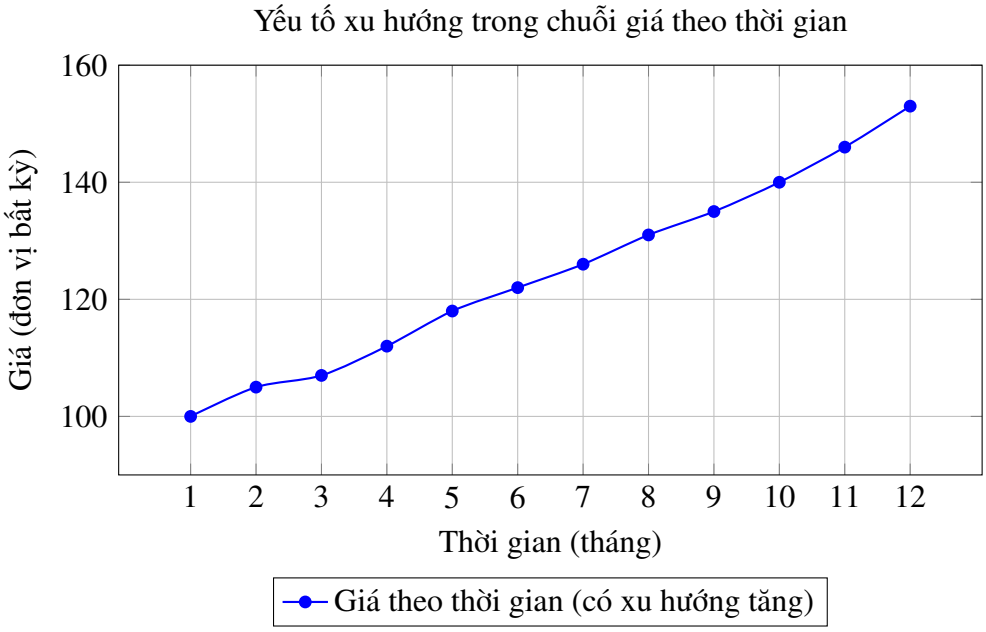
Thuộc tính ngẫu nhiên của dữ liệu có thể được thấy một cách rất dễ dàng, nó xuất hiện một cách ngẫu nhiên không theo quy luật nào rõ ràng.

Thuộc tính tự tương quan

Thuộc tính tự tương quan của dữ liệu trong chuỗi thời gian nếu một giá trị của chuỗi có tương quan hay có xu hướng biến thiên theo các giá trị khác của chuỗi.



Hình 1.2.1: Đồ thị minh họa yếu tố mùa vụ: nhiệt độ trung bình trong năm



Hình 1.2.2: Đồ thị minh họa yếu tố xu hướng: Chuỗi giá tăng theo thời gian

1.2.2 Các cơ sở toán học trong chuỗi thời gian

Cơ sở toán học đóng vai trò nền tảng và thiết yếu trong toàn bộ quá trình xử lý, phân tích và mô hình hóa chuỗi thời gian – một lĩnh vực quan trọng trong thống kê ứng dụng và khoa học dữ liệu. Việc dự báo các hiện tượng biến thiên theo thời gian đòi hỏi sự hiểu biết sâu sắc về các khái niệm toán học cốt lõi, bao gồm xác suất thống kê, giải tích toán học và đại số tuyến tính. Những nguyên lý này không chỉ cung cấp công cụ lý thuyết để xây dựng và hiệu chỉnh mô hình, mà còn đóng vai trò quan trọng trong việc giải thích và đánh giá tính hợp lý của kết quả dự báo.

Phân tích chuỗi thời gian là một nhánh đặc biệt của thống kê, tập trung vào việc mô tả, trích xuất đặc điểm và dự báo các chuỗi dữ liệu được thu thập theo thứ tự thời gian. Để thực hiện được các phân tích này một cách khoa học, cần thiết phải làm rõ các thành phần cấu thành nên một chuỗi thời gian, chẳng hạn như xu hướng, tính mùa vụ, chu kỳ và phần dư. Việc nhận diện và tách biệt các thành phần này đòi hỏi việc áp dụng các phép biến đổi toán học thích hợp như sai phân, lấy logarit, hoặc chuẩn hóa dữ liệu, nhằm ổn định chuỗi và làm nổi bật các đặc tính cần thiết cho mô hình hóa.

Bên cạnh đó, việc thiết lập và lựa chọn mô hình dự báo phù hợp cần được hỗ trợ bởi các tiêu chí định lượng rõ ràng và khách quan. Các chỉ số đánh giá mô hình như sai số tuyệt đối trung bình (MAE), và phần trăm sai số tuyệt đối trung bình (MAPE) được sử dụng phổ biến nhằm so sánh hiệu quả dự báo của các mô hình khác nhau. Các chỉ số này dựa trên nguyên lý của lý thuyết xác suất và thống kê, cho phép đo lường mức độ chênh lệch giữa giá trị thực tế và giá trị dự báo, từ đó cung cấp căn cứ định lượng để tối ưu hóa mô hình.

Chương này tập trung trình bày một cách hệ thống các khái niệm toán học nền tảng có liên quan đến phân tích chuỗi thời gian, bao gồm các khái niệm thống kê mô tả như tự tương quan và hàm tự tương quan; các kỹ thuật biến đổi chuỗi nhằm xử lý tính dừng; phân loại và đặc trưng hóa các thành phần chính trong chuỗi thời gian; cũng như giới thiệu các tiêu chí đánh giá hiệu quả mô hình dự báo. Việc nghiên cứu và nắm vững những nội dung này có ý nghĩa quan trọng, không chỉ giúp đảm bảo tính đúng đắn về mặt lý thuyết cho quá trình mô hình hóa, mà còn tạo nền tảng vững chắc cho việc áp dụng hiệu quả các phương pháp dự báo hiện đại trong thực tiễn.

Làm trơn chuỗi thời gian và Trung bình trượt

Trong thực tế, dữ liệu chuỗi thời gian thường chứa nhiều "nhiễu", làm cản trở việc nhận diện xu hướng tổng thể hoặc cấu trúc lặp lại. Do đó, một trong những bước tiền xử lý quan trọng là làm trơn dữ liệu, nhằm loại bỏ biến động ngẫu nhiên ngắn hạn, giữ lại các đặc điểm tổng quát quan trọng.Phương pháp làm trơn phổ biến nhất là trung bình trượt (MA), trong đó giá trị tại một thời điểm được tính bằng trung bình cộng của một số quan sát gần nhất.

Định nghĩa:

Giả sử chuỗi thời gian được biểu diễn bởi $\{y_t\}_{t=1}^{T}$, trung bình trượt tại thời điểm t với độ dài số lượng phần tử trung bình n được tính bằng công thức:

$$MA_t = \frac{1}{n}\sum_{i=0}^{n-1} y_{t-i}$$

Trong phân tích chuỗi thời gian, việc biến đổi chuỗi để đạt được các tính chất thuận lợi hơn cho mô hình hóa là vô cùng quan trọng. Một trong những tính chất quan trọng cần đạt được đó là tính dừng của chuỗi thời gian. Một chuỗi dừng có các đặc trưng thống kê không thay đổi theo thời gian, điều này giúp các mô hình dự báo hoạt động hiệu quả và ổn định hơn.

Sai phân

Phép sai phân là một trong những kỹ thuật phổ biến được sử dụng nhằm làm cho chuỗi thời gian trở nên dừng. Khi một chuỗi dữ liệu không dừng, giá trị trung bình hoặc phương sai của chuỗi có thể thay đổi theo thời gian, gây khó khăn trong việc áp dụng các mô hình thống kê. Thông qua việc áp dụng phép sai phân, xu hướng tăng hoặc giảm trong chuỗi dữ liệu có thể được loại bỏ, giúp duy trì tính ổn định của giá trị trung bình theo thời gian. Việc loại bỏ xu hướng thông qua sai phân không chỉ làm ổn định chuỗi mà còn góp phần giảm hiện tượng tự tương quan giữa các điểm dữ liệu gần nhau trong chuỗi thời gian. Giảm tự tương quan là một yếu tố quan trọng trong việc nâng cao độ chính xác của mô hình, vì sự phụ thuộc nội tại giữa các giá trị liên tiếp có thể làm sai lệch các ước lượng thống kê. Kỹ thuật này đặc biệt quan trọng trong quá trình xây dựng và hiệu chỉnh các mô hình chuỗi thời gian như ARIMA, vốn yêu cầu dữ liệu đầu vào phải đạt được tính dừng và không có tự tương quan rõ rệt. Việc đảm bảo điều kiện này góp phần cải thiện hiệu quả mô hình hóa và độ tin cậy của dự báo trong các ứng dụng thực tiễn.
Công thức tính sai phân bậc nhất:

$$\Delta y_t = y_t - y_{t-1}$$

- Δy_t: Giá trị sai phân bậc nhất.
- y_t: Giá trị tại thời điểm hiện tại.
- y_{t-1}: Giá trị tại thời điểm liền trước.

Việc tính toán sai phân bậc nhất cho phép xác định sự thay đổi của giá trị giữa các điểm dữ liệu liên tiếp trong chuỗi dữ liệu. Nếu giá trị của sai phân bậc nhất là dương, điều đó cho thấy chuỗi dữ liệu đang có xu hướng tăng; nếu giá trị là âm, chuỗi dữ liệu đang có xu hướng giảm. Việc tính sai phân này giúp loại bỏ xu hướng tuyến tính trong chuỗi. Khi sử dụng sai phân bậc nhất để chuẩn bị dữ liệu cho mô hình, cần kiểm tra xem chuỗi dữ liệu đã đạt được tính dừng hay chưa. Nếu chuỗi dữ liệu chưa dừng, cần tiếp tục thực hiện sai phân cho đến khi chuỗi đạt được tính dừng.
Công thức tính sai phân bậc d:

$$\Delta^d y_t = \Delta(\Delta^{d-1} y_t) = \Delta^{d-1} y_t - \Delta^{d-1} y_{t-1}$$

- $\Delta^d y_t$: Giá trị sai phân bậc d.
- $\Delta^{d-1} y_t$: Giá trị tại thời điểm hiện tại.
- $\Delta^{d-1} y_{t-1}$: Giá trị tại thời điểm liền trước.

Trong nhiều trường hợp, ngoài xu hướng tổng thể, chuỗi thời gian còn thể hiện các mẫu lặp lại theo chu kỳ, thường gọi là yếu tố mùa vụ. Ví dụ, doanh số bán hàng thường tăng vào dịp lễ Tết, hoặc nhiệt độ trung bình thay đổi theo mùa trong năm. Những yếu tố này làm cho chuỗi không dừng và gây khó khăn trong mô hình hóa và dự báo.Để loại bỏ ảnh hưởng của mùa vụ, sẽ áp dụng kỹ thuật **sai phân theo mùa vụ**, nhằm trừ đi giá trị hiện tại với giá trị trong cùng kỳ của chu kỳ trước. Kỹ thuật này rất hiệu quả trong việc loại bỏ cấu trúc lặp lại có chu kỳ cố định trong chuỗi.
Công thức tính sai phân theo mùa vụ:

$$\Delta_s y_t = y_t - y_{t-s}$$

- $y'_s(i)$: giá trị sai phân mùa vụ tại thời điểm i
- $y(i)$: giá trị tại thời điểm hiện tại i
- $y(i-s)$: giá trị tại thời điểm cách trước s chu kỳ

Sau khi áp dụng các phép biến đổi như sai phân hoặc làm trơn, bước tiếp theo là kiểm tra xem chuỗi có đạt được tính dừng hay chưa.

Tự tương quan

Tự tương quan là khái niệm chỉ mối liên hệ giữa một giá trị trong chuỗi thời gian với các giá trị trước đó của chính nó. Nói cách khác, nó cho biết liệu giá trị hiện tại có bị ảnh hưởng bởi các giá trị trong quá khứ hay không. Khi phân tích chuỗi thời gian, người ta thường quan tâm đến độ trễ tức là khoảng cách thời gian giữa hai quan sát. Ví dụ, độ trễ bằng 1 nghĩa là so sánh giá trị hiện tại với giá trị liền trước đó; độ trễ bằng 2 là so sánh với giá trị cách đó hai thời điểm, và cứ như vậy. Để đo lường mức độ tự tương quan tại các độ trễ khác nhau, người ta sử dụng hàm tự tương quan (ACF). Hàm này giúp xác định xem chuỗi thời gian có tính phụ thuộc theo thời gian hay không – nghĩa là liệu các giá trị có đang lặp lại theo một khuôn mẫu nào đó. Trong thực tế, biểu đồ ACF thường được sử dụng để trực quan hóa các hệ số tự tương quan. Những độ trễ nào có hệ số lớn và vượt ngưỡng thống kê thường sẽ gợi ý rằng chuỗi có cấu trúc đáng kể ở những khoảng thời gian đó.

Tự tương quan một phần

Tự tương quan một phần tương tự như hàm tự tương quan (ACF) ngoại trừ nó chỉ hiển thị mối tương quan giữa hai quan sát mà độ trễ ngắn hơn giữa các quan sát đó không giải thích được. Ví dụ, tự tương quan một phần cho độ trễ 3 chỉ là tương quan mà độ trễ 1 và 2 không giải thích được. Nói cách khác, mối tương quan từng phần cho mỗi độ trễ là mối tương quan duy nhất giữa hai quan sát đó sau khi loại bỏ các mối tương quan xen kẽ.

Đánh giá mô hình dự báo

Đánh giá mô hình dự báo dựa trên sự phù hợp của mô hình với dữ liệu quá khứ. Có nhiều kĩ thuật thống kê giúp mô tả sự phù hợp của mô hình với dữ liệu mẫu. Chú ý rằng sự phù hợp này dựa trên đánh giá phần dư của mô hình, nên không thực sự đảm bảo mô hình sẽ dự báo thành công trong tương lai. Người ta quan tâm đến sự chính xác của dự báo tương lai, chứ không phải sự phù hợp của mô hình với dữ liệu quá khứ. Do đó, việc đánh giá sai số dự báo ngoài mẫu (out-of-sample) là rất quan trọng. **Sai số dự báo** Hiệu quả/ chất lượng của một mô hình dự báo có thể được đánh giá dựa trên sai số dự báo một bước:

$$e_t(1) = y_{t+1} - \hat{y}_t(1)$$

trong đó $\hat{y}_t(1) = E(y_{t+1}|y_1, \ldots, y_t)$ là giá trị dự báo của y_{t+1} thực hiện tại thời điểm t. Giả sử có n quan sát được dự báo và n sai số dự báo một bước, $e_t(1)$, $t = 1, 2, \ldots, n$. Các đại lượng đo độ chính xác của mô hình ước lượng bao gồm:

- **Sai số tuyệt đối trung bình (mean absolute error):**

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t(1)|$$

- **Sai số bình phương trung bình (mean squared error):**

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2(1)$$

- **Sai số phần trăm tuyệt đối trung bình (mean absolute percentage error):**

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{y_{t+1} - \hat{y}_t(1)}{y_{t+1}}\right| \times 100\%$$

Có nhiều bằng chứng thực nghiệm cho thấy sai số dự báo có phân phối xấp xỉ phân phối chuẩn. Có thể sử dụng biểu đồ xác suất chuẩn (normal probability plot) của sai số dự báo để xác nhận điều này. Mô hình dự báo được kỳ vọng sẽ diễn giải đầy đủ các cấu trúc của dữ liệu (như xu thế, mùa, ...) và các sai số dự báo sẽ chỉ mang tính ngẫu nhiên. Khi đó, ACF mẫu của sai số dự báo sẽ giống như ACF của dữ liệu hoàn toàn ngẫu nhiên, tức là ACF mẫu khác 0 không đáng kể. Trái lại, ACF mẫu có giá trị khác 0 đáng kể với một độ trễ nào đó; chứng tỏ rằng các sai số dự báo không phải hoàn toàn ngẫu nhiên. Do đó, mô hình dự báo chưa diễn giải tốt các cấu trúc của dữ liệu và có thể cải tiến hơn nữa.

Nhiễu trắng

Một quá trình hoàn toàn ngẫu nhiên, như được đề cập đến trước đó, được gọi là một nhiễu trắng (white noise). Cụ thể, một chuỗi thời gian a_t bao gồm các quan sát không tương quan, có kỳ vọng bằng 0 và phương sai hằng số thì được gọi là một nhiễu trắng. Do đó, nhiễu trắng a_t thỏa mãn các tính chất:

$$E(a_t) = 0$$

$$E(a_t^2) = \sigma^2$$

$$E(a_t a_{t+k}) = 0 \quad \text{với } k \neq 0$$

Thông thường, nhiễu trắng a_t được ký hiệu là $a_t \sim WN(0, \sigma^2)$. Ngoài ra, nếu các quan sát của chuỗi thời gian có phân phối chuẩn, được gọi là nhiễu trắng Gauss. Trong trường hợp lý tưởng, sai số dự báo là nhiễu trắng Gauss. Trong phân tích chuỗi thời gian, có thể kiểm tra chuỗi phần dư của mô hình có sinh ra từ quá trình nhiễu trắng Gauss hay không bằng cách sử dụng lược đồ tự tương quan (ACF), histogram và biểu đồ xác suất chuẩn.

CHƯƠNG 2

Mô hình Arima và mô hình Prophet

Chương trước đã trình bày các khái niệm nền tảng trong phân tích chuỗi thời gian, bao gồm các yếu tố cấu thành như xu hướng, mùa vụ, chu kỳ, tính dừng và tự tương quan. Trên cơ sở những kiến thức này, chương tiếp theo sẽ giới thiệu hai mô hình phổ biến được ứng dụng rộng rãi trong thực tiễn để phân tích và dự báo chuỗi thời gian, đó là mô hình ARIMA và mô hình Prophet.

Mô hình ARIMA là một mô hình thống kê truyền thống, được phát triển dựa trên giả định rằng chuỗi dữ liệu đầu vào phải đạt được tính dừng thông qua phép sai phân. Mô hình này bao gồm ba thành phần chính: tự hồi quy (AR), sai phân (I), và trung bình trượt (MA). Ngoài ra, mô hình có thể được mở rộng để xử lý yếu tố mùa vụ bằng cách sử dụng phiên bản SARIMA, cho phép phân tích dữ liệu có chu kỳ rõ rệt theo mùa.

Ngược lại, Prophet là một mô hình hiện đại hơn, do Facebook (nay là Meta) phát triển, được xây dựng dựa trên phương pháp phân rã thành phần. Prophet được thiết kế với mục tiêu hỗ trợ việc xây dựng mô hình dự báo dễ dàng, đặc biệt phù hợp trong các trường hợp chuỗi dữ liệu có xu hướng rõ rệt và yếu tố mùa vụ mạnh. Mô hình này thể hiện sự linh hoạt trong việc xử lý dữ liệu thiếu, dữ liệu dị thường, và có khả năng tự động điều chỉnh theo các ngày lễ hoặc sự kiện đặc biệt gây ảnh hưởng đến chuỗi dữ liệu.

Chương này sẽ trình bày chi tiết về cấu trúc lý thuyết, công thức mô hình, quy trình xây dựng, phương pháp ước lượng, tiêu chí lựa chọn và kiểm định đối với hai mô hình ARIMA và Prophet. Thông qua việc phân tích các yếu tố kỹ thuật và phương pháp luận của từng mô hình, chương này cung cấp cơ sở để đánh giá và so sánh hiệu quả ứng dụng của hai phương pháp trong các bối cảnh dự báo chuỗi thời gian thực tế.

2.1 Mô hình ARIMA

Như đã nói ở trên, mô hình ARIMA được xây dựng dựa trên giả định rằng chuỗi đầu vào cần được làm dừng thông qua sai phân, chính vì vậy nó là một trong những mô hình phổ biến nhất trong phân tích chuỗi thời gian, đặc biệt khi dữ liệu có tính không dừng. Mô hình ARIMA được ký hiệu là ARIMA(p,d,q) nó được kết hợp từ 3 thành phần chính: p là bậc của phần tự hồi quy(AR), d là số lần sai phân(I) cần thiết để chuỗi thời gian dừng, q là bậc của phần trung bình trượt (MA). Thành phần tự hồi quy (AR) biểu thị mối quan hệ giữa giá trị hiện tại của chuỗi thời gian và các giá trị trong quá khứ của chính chuỗi đó. Thành phần sai phân thể hiện việc sai phân chuỗi thời gian để loại bỏ xu hướng và biến chuỗi không dừng thành chuỗi dừng. Việc áp dụng phép sai phân giúp ổn định giá trị trung bình và phương sai của chuỗi theo thời gian.Thành phần trung bình trượt mô hình hóa mối quan hệ giữa giá trị hiện tại của chuỗi và các nhiễu loạn ngẫu nhiên (sai số) trong quá khứ. Chính vì vậy, trong mô hình chuỗi thời gian, mô hình ARIMA khái niệm

mang tính nền tảng là tính dừng của chuỗi. Vì vậy, trước khi đi vào mô hình ARIMA cụ thể, đầu tiên cần làm rõ: chuỗi dừng là gì, và tại sao phải thực hiện phép sai phân trước khi áp dụng mô hình. Những nội dung này sẽ được trình bày chi tiết trong phần tiếp theo.

2.1.1 Tính dừng

Trong phân tích chuỗi thời gian, một chuỗi được gọi là dừng nếu các đặc tính thống kê cơ bản của nó không thay đổi theo thời gian. Tính dừng của chuỗi thời gian được chia ra làm hai loại đó là dừng mạnh và dừng yếu.

Dừng mạnh

Một chuỗi thời gian $\{X_t\}$ được gọi là dừng mạnh hay còn được gọi là dừng theo phân phối xác suất nghĩa là nếu phân phối xác suất chung của bất kỳ tập hợp hữu hạn các biến ngẫu nhiên trong chuỗi không thay đổi khi dịch chuyển thời gian. Cụ thể, với mọi $k \in \mathbb{N}$, mọi tập chỉ số $t_1, t_2, \ldots, t_k \in \mathbb{Z}$ và mọi dịch chuyển $h \in \mathbb{Z}$, công thức là:

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_k}) \overset{d}{=} (X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$$

Ký hiệu $\overset{d}{=}$ biểu thị rằng hai vector ngẫu nhiên có cùng phân phối xác suất.Tức là các đặc trưng thống kê của chuỗi (bao gồm phân phối kết hợp, không chỉ trung bình hay phương sai) đều không thay đổi theo thời gian.
Ví dụ trực quan: Nếu lấy ba giá trị liên tiếp của chuỗi tại thời điểm $t = 5, 6, 7$ và so sánh với ba giá trị tại $t = 100, 101, 102$, thì vector phân phối của hai bộ ba giá trị đó phải giống hệt nhau nếu chuỗi là dừng mạnh.

Dừng yếu

Trong thực tế, đặc biệt khi xây dựng các mô hình như ARIMA, người ta thường sử dụng khái niệm dừng yếu, vì nó cung cấp điều kiện đủ để mô hình hoạt động mà không đòi hỏi phân phối đầy đủ của chuỗi.Một chuỗi thời gian $\{X_t\}$ được gọi là dừng yếu nếu thỏa mãn đồng thời các đặc tính thống kê cơ bản của nó không thay đổi theo thời gian. Cụ thể, một chuỗi dừng thỏa mãn ba điều kiện chính gồm kỳ vọng, phương sai và tự tương quan. Kỳ vọng không đổi theo thời gian tức là trung bình của chuỗi tại mọi thời điểm đều bằng nhau.

$$\mathbb{E}[X_t] = \mu, \quad \forall t$$

Ví dụ trực quan: Trung bình của chuỗi trong tháng 1 năm 2020 phải xấp xỉ bằng trung bình trong tháng 1 năm 2022.
Phương sai không đổi theo thời gian nghĩa là độ biến động của chuỗi không thay đổi theo thời gian. Nếu phương sai thay đổi, chuỗi sẽ có độ "nhiễu" khác nhau ở các giai đoạn.

$$\mathrm{Var}(X_t) = \mathbb{E}[(X_t - \mu)^2] = \sigma^2, \quad \forall t$$

Tự tương quan không phụ thuộc vào thời điểm nghĩa là mối quan hệ giữa các giá trị tại hai thời điểm chỉ phụ thuộc vào khoảng cách giữa chúng (độ trễ), chứ không phụ thuộc

vào thời gian cụ thể.

$$\gamma(h) = \text{Cov}(X_t, X_{t+h}) = \mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)], \quad \forall t$$

Trên thực tế, nhiều chuỗi thời gian không phải là chuỗi dừng. Các chuỗi dữ liệu như giá cổ phiếu, doanh số bán hàng, hoặc các chỉ số kinh tế thường thể hiện xu hướng tăng hoặc giảm theo thời gian, đồng thời có thể chịu ảnh hưởng của các yếu tố mùa vụ. Những đặc điểm này khiến kỳ vọng và phương sai của chuỗi thay đổi theo thời gian, từ đó xác định chuỗi là chuỗi không dừng. Trong trường hợp này, việc mô hình hóa chuỗi bằng các phương pháp thống kê như ARIMA yêu cầu dữ liệu đầu vào phải có tính dừng. Do đó, cần thiết phải chuyển đổi chuỗi không dừng về chuỗi dừng trước khi tiến hành xây dựng mô hình. Phương pháp được sử dụng phổ biến nhất để đạt được điều này là phép sai phân, như đã trình bày trong chương trước về chuỗi thời gian. Phép biến đổi này giúp loại bỏ xu hướng trong dữ liệu, ổn định kỳ vọng và phương sai, tạo điều kiện cho việc áp dụng hiệu quả các mô hình dự báo tuyến tính như ARIMA.

2.1.2 Công thức của mô hình ARIMA

Như đã nêu ở đầu chương, mô hình ARIMA là sự kết hợp của ba thành phần chính gồm tự hồi quy (AR), tích hợp (I – thực hiện phép sai phân để chuỗi trở thành dừng), và trung bình trượt (MA). Công thức tổng quát của mô hình ARIMA(p,d,q) được biểu diễn như sau:

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t$$

- $\phi(B)$: biểu diễn phần tự hồi quy (AR) bậc p.
- $(1-B)^d$: biểu diễn phép sai phân bậc d để làm chuỗi trở thành dừng.
- $\theta(B)$: biểu diễn phần trung bình trượt (MA) bậc q.
- Z_t: là nhiễu trắng tại thời điểm t.

Toán tử trễ B

Toán tử trễ, hay còn gọi là toán tử backshift ký hiệu là B, là một phép toán quan trọng trong phân tích chuỗi thời gian. Toán tử này được dùng để dịch chuyển các giá trị của chuỗi thời gian lùi về phía trước theo số bước thời gian nhất định. Cụ thể, khi áp dụng toán tử trễ B lên giá trị chuỗi thời gian tại thời điểm t, kết quả sẽ thu được giá trị của chuỗi tại thời điểm $t-1$:

$$BX_t = X_{t-1}$$

Tương tự, khi áp dụng toán tử trễ B lặp lại k lần, biểu thức thu được là:

$$B^k X_t = X_{t-k}$$

Toán tử trễ giúp biểu diễn các mối quan hệ giữa các giá trị trong chuỗi thời gian một cách cô đọng và hiệu quả. Ví dụ, trong mô hình ARIMA, các thành phần tự hồi quy (AR) và trung bình trượt (MA) được thể hiện thông qua các đa thức của toán tử trễ B, giúp viết

công thức mô hình một cách ngắn gọn. Việc sử dụng toán tử trễ giúp phân tích và xây dựng mô hình dễ dàng hơn, bởi nó biểu diễn các giá trị quá khứ của chuỗi thông qua một ký hiệu duy nhất thay vì viết dài dòng từng giá trị.

2.1.3 Các bước xây dựng Mô hình ARIMA

Trước tiên, để xây dựng mô hình ARIMA phù hợp, cần xác định thứ tự của các thành phần tự hồi quy (p), sai phân (d) và trung bình trượt (q). Việc xác định này được thực hiện thông qua các công cụ phân tích chuỗi thời gian.

Xác định bậc kết hợp của một chuỗi thời gian

Trong phân tích chuỗi thời gian, việc xác định bậc kết hợp là một bước quan trọng nhằm lựa chọn mô hình phù hợp, đặc biệt là trong mô hình ARIMA. Một chuỗi thời gian được gọi là kết hợp bậc d (ký hiệu là $I(d)$) nếu sau khi thực hiện phép sai phân d lần, chuỗi trở thành dừng. Việc xác định giá trị d chính là xác định bậc kết hợp của chuỗi. Để xác định bậc kết hợp, cần sử dụng các kiểm định nghiệm đơn vị, phổ biến nhất là kiểm định ADF.

Kiểm định ADF

Giả sử chuỗi thời gian y_t được mô hình hóa thông qua phương trình hồi quy trong kiểm định ADF như sau:

$$\Delta y_t = \alpha + \phi y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \varepsilon_t$$

- $\Delta y_t = y_t - y_{t-1}$ là sai phân bậc nhất của chuỗi.
- α là hằng số..
- ϕ là hệ số kiểm định chính, dùng để kiểm tra nghiệm đơn vị.
- β_i là hệ số của các độ trễ sai phân.
- ε_t là nhiễu trắng (sai số ngẫu nhiên).
- p là số độ trễ được chọn.

Với kiểm định ADF, kiểm định có giả thuyết gốc $H_0 : \phi = 0$ là chuỗi có nghiệm đơn vị (nghĩa là không dừng), còn $H_1 : \phi < 0$ nghĩa là chuỗi dừng. Thống kê kiểm định được xác định theo công thức như sau

$$\tau = \frac{\hat{\phi}}{\mathrm{SE}(\hat{\phi})}$$

- $\hat{\phi}$ là ước lượng của ϕ.
- $\mathrm{SE}(\hat{\phi})$ là sai số chuẩn của ước lượng ϕ.

Sau khi áp dụng công thức kiểm định, để bác bỏ huyết H_0 nếu $\tau < u_\alpha$, trong đó u_α là giá trị tới hạn tại mức ý nghĩa α. Các giá trị tới hạn thường dùng được trình bày trong bảng sau:

Mức ý nghĩa α	Giá trị tới hạn u_α
1.0%	-3.43
2.5%	-3.12
5.0%	-2.86

Nếu giá trị thống kê kiểm định τ nhỏ hơn giá trị tới hạn tương ứng, giả thuyết gốc H_0 bị bác bỏ và có thể kết luận rằng chuỗi y_t là chuỗi dừng. Khi đó, bậc sai phân $d = 0$, mô hình có thể được xây dựng dưới dạng ARIMA$(p,0,q)$. Trong trường hợp kiểm định ADF không bác bỏ giả thuyết $H_0 : \phi = 0$, điều này cho thấy chuỗi thời gian y_t có nghiệm đơn vị và không ổn định theo thời gian, tức là chuỗi không dừng. Lúc này, cần thực hiện phép sai phân để chuyển chuỗi không dừng thành chuỗi dừng. Cụ thể, phép sai phân bậc nhất được tính theo công thức:

$$\Delta y_t = y_t - y_{t-1}$$

Sau đó, kiểm định tính dừng được áp dụng cho chuỗi sai phân Δy_t bằng kiểm định ADF. Nếu chuỗi sai phân trở nên dừng, có thể kết luận rằng chuỗi ban đầu là kết hợp bậc 1, do đó bậc sai phân $d = 1$ được sử dụng trong mô hình ARIMA.Nếu chuỗi Δy_t vẫn chưa dừng, phép sai phân tiếp theo sẽ được thực hiện lần thứ hai:

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = y_t - 2y_{t-1} + y_{t-2}$$

Quy trình này được lặp lại cho đến khi chuỗi đạt được tính dừng. Số lần sai phân cần thiết để chuỗi trở nên dừng chính là bậc sai phân d trong mô hình ARIMA(p,d,q). Việc sai phân giúp loại bỏ xu hướng trong chuỗi và đảm bảo tính dừng — một điều kiện cần thiết để áp dụng hiệu quả các mô hình chuỗi thời gian như ARIMA. Tuy nhiên, cần lưu ý tránh sai phân quá mức, vì điều này có thể làm mất đi các đặc điểm quan trọng của dữ liệu và ảnh hưởng tiêu cực đến hiệu quả dự báo của mô hình.

Xác định bậc của thành phần tự hồi quy p

Thành phần tự hồi quy (AR) trong mô hình ARIMA được ký hiệu là $AR(p)$, trong đó p là số bậc trễ. Thành phần này biểu thị mối quan hệ giữa giá trị hiện tại của chuỗi thời gian và các giá trị trong quá khứ của chính chuỗi đó. Việc xác định đúng bậc p là rất quan trọng để đảm bảo mô hình ARIMA phản ánh đúng cấu trúc phụ thuộc trong dữ liệu. Sau khi đã thực hiện làm chuỗi thời gian trở thành chuỗi dừng(thông qua phép sai phân ở trên), thì sẽ sử dụng hàm tự tương quan một phần(PACF) để xác định được giá trị p bậc trễ của mô hình. Hàm tự tương quan một phần tại độ trễ. Hàm tự tương quan riêng phần tại độ trễ h, ký hiệu là $\phi(h)$, đo lường mối tương quan giữa X_t và X_{t-h} sau khi đã loại bỏ ảnh hưởng của tất cả các độ trễ trung gian từ 1 đến $h-1$.Một cách tính phổ biến là sử dụng mô hình hồi quy tuyến tính đa biến sau, với Z_t là phần dư trắng:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_h X_{t-h} + Z_t$$

Nói cách khác, PACF chỉ đo lường ảnh hưởng trực tiếp tại độ trễ h. Lúc đó sẽ có hai trường hợp, nếu đồ thị PACF **cắt cụt** sau độ trễ p, tức là các giá trị $\phi(h)$ gần bằng 0 với $h > p$, thì có thể suy luận rằng mô hình AR bậc p là phù hợp.Nếu PACF giảm dần từ từ và không cắt cụt rõ ràng, có thể cần dùng mô hình kết hợp (ARMA) hoặc điều chỉnh

lại sai phân.Đồ thị PACF mẫu thường được vẽ kèm theo các khoảng tin cậy (confidence intervals), thông thường là 95%, tương ứng với giới hạn $\pm\frac{2}{\sqrt{n}}$ trong đó n là số lượng quan sát. Nếu một hệ số PACF nằm ngoài khoảng này, nó được xem là có ý nghĩa thống kê và góp phần đáng kể vào mô hình. Trong thực hành, việc xác định p không chỉ dựa hoàn toàn vào đồ thị PACF, mà còn cần kết hợp với các tiêu chí chọn mô hình như AIC hoặc BIC để xác nhận mô hình tối ưu.

Xác định bậc của thành phần trung bình trượt q

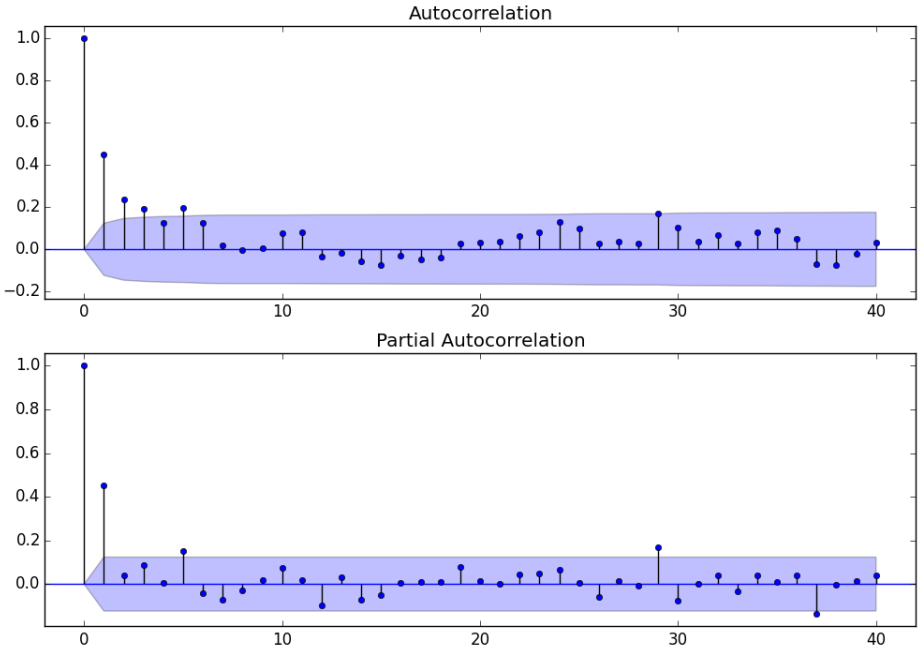
Trong mô hình ARIMA, thành phần trung bình trượt (MA) được dùng để mô hình hóa ảnh hưởng của các nhiễu loạn ngẫu nhiên trong quá khứ đến giá trị hiện tại của chuỗi thời gian. Bậc của thành phần MA, ký hiệu là q, đại diện cho số lượng độ trễ của các nhiễu loạn ngẫu nhiên $Z_{t-1}, Z_{t-2}, \ldots, Z_{t-q}$ có ảnh hưởng trực tiếp đến giá trị X_t. Để xác định giá trị phù hợp của q, sẽ sử dụng biểu đồ hàm tự tương quan (Autocorrelation Function - ACF). Trong một chuỗi thời gian đã được làm dừng (thường sau khi sai phân), nếu biểu đồ ACF cắt cụt rõ ràng sau độ trễ q, tức là giá trị ACF giảm nhanh và gần như bằng 0 từ độ trễ $q+1$ trở đi, thì đây là dấu hiệu cho thấy mô hình MA(q) là phù hợp. Ví dụ, nếu biểu đồ ACF cho thấy các giá trị tương quan tại độ trễ 1 và 2 là lớn, trong khi từ độ trễ 3 trở đi đều gần bằng 0, thì có thể mô hình MA(2) là thích hợp. Công thức tổng quát của mô hình MA(q) được viết như sau:

$$X_t = \mu + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}$$

- X_t: giá trị tại thời điểm t.
- μ: kỳ vọng của chuỗi.
- Z_t: nhiễu trắng.
- $\theta_1, \theta_2, \ldots, \theta_q$: các hệ số MA.

Việc đánh giá biểu đồ ACF để xác định q cần kết hợp cùng với việc kiểm định phần dư và các tiêu chí lựa chọn mô hình như AIC hoặc BIC để đảm bảo rằng mô hình không bị quá khớp và phù hợp với dữ liệu.

Ví dụ:



Hình 2.1.1: Đồ thị ACF và PACF

Giải thích:
Phân tích biểu đồ ACF cho thấy biểu đồ giảm dần dần không cắt cụt ngay và có ý nghĩa tại một vài độ trễ đầu (lag 1, 2, 3), sau đó dao động quanh 0 và dần nằm trong vùng tin cậy màu xanh.Điều này là đặc trưng của một thành phần AR (tự hồi quy), biểu đồ ACF như thế này cho thấy không nên chọn MA (vì MA thường khiến ACF cắt cụt). Vậy p ở đây bằng 0. Phân tích biểu đồ PACF cho thấy PACF cắt cụt sau lag 2: các giá trị tại lag 1 và lag 2 là đáng kể, còn lại hầu hết nằm trong vùng không ý nghĩa. Vậy nên q ở đây bằng 2. Mô hình phù hợp ở đây sẽ là ARIMA(2,d,0).

Lựa chọn mô hình bằng tiêu chí AIC và BIC

Sau khi đã xác định sơ bộ các giá trị p, d, và q thông qua phân tích đồ thị ACF, PACF và kiểm định tính dừng, bước tiếp theo trong quá trình xây dựng mô hình ARIMA là lựa chọn mô hình tối ưu. Trong thực tế, có nhiều tổ hợp (p,d,q) khác nhau có thể phù hợp với dữ liệu. Do đó, cần có một tiêu chí khách quan để so sánh và lựa chọn mô hình tốt nhất. Hai tiêu chí phổ biến thường được sử dụng trong việc lựa chọn mô hình là AIC và BIC. Cả hai tiêu chí đều đánh giá sự phù hợp của mô hình dựa trên độ khớp với dữ liệu và mức độ đơn giản của mô hình (số lượng tham số). Mô hình với quá nhiều tham số sẽ bị phạt để tránh hiện tượng quá khớp. Mô hình AIC và BIC là:

$$\text{AIC} = -2\ln(L) + 2k$$

và

$$\text{BIC} = -2\ln(L) + k\ln(n)$$

- L là giá trị của hàm hợp lý cực đại của mô hình.
- k là số lượng tham số trong mô hình.
- n là số lượng quan sát .

Mô hình có giá trị AIC hoặc BIC thấp hơn được xem là tốt hơn. Tuy nhiên, AIC thiên về khả năng dự đoán và có thể chọn mô hình phức tạp hơn, trong khi BIC nghiêng về mô hình đơn giản hơn với khả năng tổng quát hóa tốt hơn, đặc biệt khi kích thước mẫu lớn.

Dự báo bằng mô hình ARIMA

Sau khi lựa chọn được mô hình ARIMA phù hợp, bước tiếp theo là sử dụng mô hình này để dự báo các giá trị tương lai của chuỗi thời gian. Dự báo là một ứng dụng quan trọng, giúp đưa ra các quyết định liên quan đến quản lý, tài chính hoặc hoạch định chính sách. Quá trình dự báo dựa trên việc sử dụng thông tin từ quá khứ để ước lượng giá trị tại các thời điểm tương lai.Công thức tổng quát của mô hình ARIMA (p,d,q) sau khi được xác định có dạng:

$$\phi(B)(1-B)^d X_t = \theta(B)Z_t$$

Trong đó, X_t là chuỗi thời gian gốc, B là toán tử trễ, d là số lần sai phân để chuỗi trở nên dừng, $\phi(B)$ và $\theta(B)$ lần lượt là đa thức của các thành phần tự hồi quy (AR) và trung bình trượt (MA), còn Z_t là nhiễu trắng. Dự báo giá trị tại thời điểm tương lai $t+h$, ký hiệu là $\hat{X}_{t+h}$, được thực hiện dựa trên các giá trị quá khứ và sai số đã mô hình hóa. Để tăng độ tin cậy của dự báo, bên cạnh việc đưa ra giá trị trung bình kỳ vọng $\hat{X}_{t+h}$, người ta còn tính khoảng tin cậy $(1-\alpha)\%$ (thường là 95%) theo công thức:

$$\hat{X}_{t+h} \pm z_{\alpha/2} \cdot \sigma_h$$

Trong đó, $z_{\alpha/2}$ là điểm tới hạn của phân phối chuẩn, thường lấy $z_{0.025} \approx 1.96$ với mức ý nghĩa 5%, còn σ_h là độ lệch chuẩn của sai số dự báo tại bước h. Cần lưu ý rằng khoảng tin cậy của dự báo sẽ tăng dần theo thời gian do sai số dự báo tích lũy. Điều này có nghĩa là các dự báo càng xa thì càng có xu hướng kém chính xác hơn. Vì vậy, trong thực tế, người phân tích thường kết hợp biểu đồ và khoảng tin cậy để đưa ra quyết định phù hợp.

Sai số dự báo trong mô hình ARIMA

Sai số dự báo là sự chênh lệch giữa giá trị thực tế và giá trị dự báo được mô hình đưa ra tại một thời điểm tương lai. Ký hiệu sai số dự báo tại bước dự báo h là e_{t+h}, được xác định theo công thức:

$$e_{t+h} = X_{t+h} - \hat{X}_{t+h}$$

Trong đó, X_{t+h} là giá trị thực tế tại thời điểm $t+h$ và $\hat{X}_{t+h}$ là giá trị dự báo tương ứng. Độ lớn của sai số dự báo phụ thuộc vào cấu trúc của mô hình và mức độ biến động của chuỗi thời gian. Khi dự báo các bước xa hơn trong tương lai, sai số dự báo thường tăng do mô hình phải sử dụng các giá trị dự báo trước đó thay vì giá trị thực tế. Để đánh giá sai số dự báo, người ta thường sử dụng các chỉ số thống kê như:

- **MAE (Mean Absolute Error)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|X_i - \hat{X}_i|$$

- **MAPE (Mean Absolute Percentage Error)**:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{X_i - \hat{X}_i}{X_i}\right|$$

Trong quá trình đánh giá mô hình ARIMA, sai số dự báo không chỉ là công cụ đo lường hiệu suất của mô hình mà còn đóng vai trò trong việc điều chỉnh và cải thiện cấu trúc mô hình nhằm tăng độ chính xác trong dự báo thực tế.

2.2 Mô hình Prophet

Prophet là một mô hình dự báo chuỗi thời gian được phát triển bởi Facebook, đặc biệt hiệu quả trong việc xử lý các chuỗi có đặc điểm gồm xu hướng, tính mùa vụ, và các ngày lễ. Mô hình này được thiết kế để dễ sử dụng, linh hoạt và có khả năng tự động điều chỉnh với những thay đổi trong cấu trúc dữ liệu, từ đó rất phù hợp cho cả các nhà phân tích dữ liệu không chuyên về thống kê.

2.2.1 Đặc điểm của Mô hình Prophet

Mô hình Prophet sở hữu nhiều đặc điểm nổi bật khiến nó trở thành một lựa chọn lý tưởng cho việc dự báo chuỗi thời gian, đặc biệt trong các trường hợp dữ liệu có tính phức tạp cao và đòi hỏi sự linh hoạt. Trước hết, Prophet hoạt động dựa trên mô hình cộng tính, trong đó chuỗi thời gian được phân tách thành ba thành phần chính gồm xu hướng, mùa vụ và ảnh hưởng của các ngày lễ. Việc phân tách này giúp tăng khả năng diễn giải và phân tích các yếu tố tác động đến dữ liệu, đồng thời hỗ trợ nâng cao hiệu quả trong dự báo.

Bên cạnh đó, Prophet đặc biệt hiệu quả với những chuỗi thời gian có tính mùa vụ rõ rệt, kể cả khi các chu kỳ mùa vụ có thể phức tạp hoặc thay đổi theo thời gian. Mô hình sử dụng chuỗi Fourier để mô hình hóa thành phần mùa vụ, nhờ đó cho phép điều chỉnh linh hoạt mức độ phức tạp của mô hình theo nhu cầu cụ thể của người dùng.

Một ưu điểm lớn khác là khả năng xử lý tốt các tập dữ liệu thiếu hoặc chứa giá trị ngoại lệ. Prophet không yêu cầu tiền xử lý dữ liệu quá phức tạp, giúp giảm thiểu công sức chuẩn bị dữ liệu mà vẫn duy trì được độ chính xác trong mô hình hóa.

Thêm vào đó, mô hình cũng hỗ trợ mô hình hóa các xu hướng phi tuyến và cho phép người dùng chỉ định các điểm thay đổi xu hướng. Nhờ đó, Prophet có thể phản ánh sát thực hơn những biến động đột ngột hoặc xu hướng thay đổi trong dữ liệu theo thời gian.

Prophet cũng rất linh hoạt trong việc tích hợp các yếu tố đặc biệt như ngày lễ hoặc sự kiện cụ thể có thể ảnh hưởng mạnh đến chuỗi thời gian. Người dùng có thể dễ dàng đưa các sự kiện này vào mô hình, giúp cải thiện đáng kể chất lượng dự báo, đặc biệt trong các

lĩnh vực như bán lẻ, tài chính, hoặc du lịch.
Cuối cùng, một trong những đặc điểm khiến Prophet được ưa chuộng rộng rãi là tính dễ sử dụng. Mô hình được thiết kế với giao diện lập trình ứng dụng (API) đơn giản, thân thiện với người dùng, kể cả những người không chuyên về thống kê. Đồng thời, Prophet còn hỗ trợ tự động hóa nhiều bước trong quá trình xây dựng mô hình, từ việc xác định điểm thay đổi đến việc điều chỉnh tham số, giúp tiết kiệm đáng kể thời gian và công sức trong quá trình triển khai mô hình dự báo.

2.2.2 Mô hình cộng tính của mô hình Prophet

Mô hình cộng tính là nền tảng cốt lõi trong Prophet, cho phép phân tách và phân tích các yếu tố khác nhau ảnh hưởng đến chuỗi thời gian. Mô hình này giả định rằng chuỗi thời gian được tạo thành từ tổng của các thành phần độc lập là xu hướng, mùa vụ , và ảnh hưởng của các ngày lễ. Cụ thể, mô hình được biểu diễn bằng phương trình:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Trong đó, xu hướng dài hạn của chuỗi được kí hiệu là $g(t)$, thành phân chu kì là $s(t)$, ảnh hưởng của các sự kiện đặc biệt như ngày lễ $h(t)$ và ε_t là nhiễu.

Thành phần xu hướng

Xu hướng $g(t)$ mô hình hóa sự thay đổi dài hạn của chuỗi thời gian, có thể là tuyến tính hoặc logistic. Mô hình Prophet cho phép tồn tại nhiều điểm thay đổi xu hướng tại các thời điểm s_j.Trong trường hợp xu hướng được giả định là tuyến tính, hàm $g(t)$ được định nghĩa như sau:

$$g(t) = \left(k + \sum_{j=1}^{S} \delta_j \mathbf{1}_{(t \geq s_j)}\right)(t - t_0) + \left(m + \sum_{j=1}^{S} \delta_j (s_j - t_0)\mathbf{1}_{(t \geq s_j)}\right)$$

Trong đó, k là hệ số góc ban đầu thể hiện tốc độ thay đổi của xu hướng, m là hệ số chặn ban đầu xác định vị trí của đường xu hướng tại thời điểm t_0. Tại mỗi điểm thay đổi s_j, xu hướng sẽ điều chỉnh hệ số góc thêm một lượng δ_j, với $\mathbf{1}_{(t \geq s_j)}$ là hàm chỉ báo kích hoạt sự thay đổi khi t vượt qua s_j. Đồng thời, thành phần chặn cũng được điều chỉnh tương ứng để đảm bảo hàm $g(t)$ duy trì tính liên tục tại các điểm thay đổi. Mô hình này phù hợp với những chuỗi có sự thay đổi tuyến tính về tốc độ tăng hoặc giảm theo thời gian. Tuy nhiên, trong một số trường hợp dữ liệu có xu hướng tiến dần đến một ngưỡng giới hạn tăng trưởng (carrying capacity), chẳng hạn như thị trường đạt mức bão hòa, mô hình logistic được sử dụng thay cho mô hình tuyến tính. Khi đó, hàm xu hướng $g(t)$ được mô tả bởi:

$$g(t) = \frac{C(t)}{1 + \exp\left(-\left(k + \sum_{j=1}^{S} \delta_j \mathbf{1}_{(t \geq s_j)}\right)\left(t - \left(m + \sum_{j=1}^{S} \gamma_j \mathbf{1}_{(t \geq s_j)}\right)\right)\right)}$$

Trong công thức này, $C(t)$ là ngưỡng mang tải – giá trị cực đại mà chuỗi có thể tiến gần đến nhưng không vượt qua. Các tham số k và m lần lượt là hệ số góc và hệ số chặn ban

đầu như trong mô hình tuyến tính. Các điểm thay đổi s_j tiếp tục điều chỉnh tốc độ tăng trưởng thông qua δ_j, và để đảm bảo tính liên tục của hàm số tại các điểm này, mô hình bổ sung thêm hệ số dịch γ_j, được xác định theo công thức $\gamma_j = -s_j\delta_j$. Tóm lại, thành phần xu hướng $g(t)$ trong Prophet là một thành phần linh hoạt, cho phép mô hình thích ứng tốt với các thay đổi dài hạn của chuỗi thời gian, kể cả trong các trường hợp tăng trưởng tuyến tính hay có giới hạn.

Thành phần thời vụ

Thành phần thời vụ $s(t)$ trong mô hình Prophet được sử dụng để mô tả các mẫu lặp lại theo chu kỳ trong chuỗi thời gian. Để mô hình hóa các hiệu ứng định kỳ này, Prophet sử dụng chuỗi Fourier, một phương pháp mạnh mẽ cho phép biểu diễn các hàm tuần hoàn dưới dạng tổng của các hàm sin và cos. Cụ thể, một hàm tuần hoàn với chu kỳ P có thể được biểu diễn bằng công thức:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

Trong đó, P là chu kỳ của tính thời vụ, chẳng hạn 365,25 ngày cho mùa vụ hàng năm hoặc 7 ngày cho chu kỳ hàng tuần. Số lượng hạng tử Fourier, ký hiệu là N, quyết định mức độ linh hoạt của mô hình trong việc bắt các dao động lặp lại. Các hệ số a_n và b_n là các trọng số được ước lượng từ dữ liệu thông qua hồi quy tuyến tính. Ví dụ, đối với tính thời vụ hàng năm, có chu kỳ $P = 365{,}25$, và mô hình thời vụ hàng năm được viết dưới dạng:

$$s_{annual}(t) = \sum_{n=1}^{N_{annual}} \left(a_n \cos\left(\frac{2\pi nt}{365.25}\right) + b_n \sin\left(\frac{2\pi nt}{365.25}\right) \right)$$

Tương tự, với tính thời vụ hàng tuần, chu kỳ là $P = 7$, và hàm mô tả như sau:

$$s_{weekly}(t) = \sum_{n=1}^{N_{weekly}} \left(c_n \cos\left(\frac{2\pi nt}{7}\right) + d_n \sin\left(\frac{2\pi nt}{7}\right) \right)$$

Nếu dữ liệu có tần suất cao, ví dụ dữ liệu theo giờ, cũng có thể mô hình hóa tính thời vụ hàng ngày với chu kỳ $P = 1$ theo cách tương tự. Tất cả các hệ số trong các chuỗi Fourier (a_n, b_n, c_n, d_n) đều được học từ dữ liệu thông qua hồi quy tuyến tính, nhờ đó giúp mô hình xác định các mô hình tuần hoàn chính xác hơn.

Để minh họa cụ thể cho vai trò của thành phần $s(t)$, giả sử xét bài toán đang phân tích số lượng khách đặt phòng hàng ngày tại một khách sạn trong vòng 5 năm. Dữ liệu cho thấy vào các tháng mùa du lịch cao điểm như tháng 6 đến tháng 8, và trong các dịp lễ lớn như Tết Nguyên Đán hay Giáng Sinh, lượng đặt phòng có xu hướng tăng cao rõ rệt. Đây là một biểu hiện điển hình của tính thời vụ hàng năm. Ngoài ra, số lượng đặt phòng cũng biến động theo ngày trong tuần: thường giảm vào các ngày đầu tuần (thứ Hai đến thứ Năm) và tăng vào cuối tuần (thứ Sáu đến Chủ Nhật), phản ánh tính thời vụ hàng tuần. Trong trường hợp này, Prophet sẽ kết hợp cả hai thành phần thời vụ để mô hình hóa như sau:

$$s(t) = s_{weekly}(t) + s_{annual}(t)$$

Trong đó, $s_{annual}(t)$ phản ánh các quy luật lặp lại theo năm do mùa du lịch và các sự kiện lễ hội gây ra, còn $s_{weekly}(t)$ giúp mô hình nắm bắt các dao động mang tính quy luật trong tuần do thói quen hành vi của khách hàng. Việc kết hợp các chuỗi Fourier vào mô hình Prophet giúp mô tả chính xác hơn tính chất lặp lại của dữ liệu theo thời gian, làm tăng độ chính xác của các dự báo.

Thành phần hiệu ứng ngày lễ

Một trong những tính năng mạnh mẽ của Prophet là khả năng mô hình hóa hiệu ứng của các ngày lễ hoặc sự kiện đặc biệt có thể ảnh hưởng đến chuỗi thời gian. Người dùng có thể chỉ định danh sách các ngày lễ cùng với khoảng thời gian tác động (trước, trong và sau ngày lễ), từ đó mô hình có thể tự động học được ảnh hưởng của những ngày này đến dữ liệu. Cụ thể, giả sử có L ngày lễ hoặc sự kiện đặc biệt, mỗi ngày lễ thứ i xảy ra tại các thời điểm thuộc tập D_i, bao gồm cả ngày chính thức và các ngày trước hoặc sau tùy theo thiết lập. Khi đó, Prophet định nghĩa hàm chỉ báo $Z_i(t)$ cho ngày lễ i như sau:

$$Z_i(t) = \begin{cases} 1 & \text{nếu } t \in D_i \\ 0 & \text{ngược lại} \end{cases}$$

Tập D_i có thể gồm một hoặc nhiều ngày liên tiếp, không chỉ giới hạn ở ngày lễ chính. Sau đó, hiệu ứng tổng hợp từ tất cả các ngày lễ được mô hình hóa thông qua hàm $h(t)$, là tổng trọng số tuyến tính của các hàm chỉ báo:

$$h(t) = \sum_{i=1}^{L} \kappa_i Z_i(t)$$

Trong đó, κ_i là hệ số mô tả cường độ và chiều hướng ảnh hưởng của ngày lễ thứ i đối với chuỗi thời gian. Các tham số κ_i này được ước lượng đồng thời với các thành phần xu hướng $g(t)$ và thời vụ $s(t)$ thông qua hồi quy tuyến tính, thường bằng phương pháp bình phương tối thiểu. Để minh họa cụ thể, giả sử xét bài toán đang phân tích chuỗi thời gian liên quan đến số lượng khách đặt phòng khách sạn. Trong trường hợp này, hai ngày lễ quan trọng có thể ảnh hưởng mạnh đến số lượng đặt phòng là Tết Nguyên Đán và Giáng Sinh. Ví dụ, định nghĩa:

- Tết Nguyên Đán năm 2023 và 2024 lần lượt rơi vào ngày `2023-01-22` và `2024-02-10`, với khoảng ảnh hưởng từ ngày trước đến ngày sau, tức là từ `2023-01-21` đến `2023-01-23`, và từ `2024-02-09` đến `2024-02-11`.
- Giáng Sinh rơi vào ngày `2023-12-25` và `2024-12-25`, cũng với khoảng ảnh hưởng tương tự, từ `2023-12-24` đến `2023-12-26`.

Tập các ngày ảnh hưởng từ các ngày lễ có thể được mô hình hóa bằng các hàm chỉ báo tương ứng $Z_{\text{Tết}}(t)$ và $Z_{\text{GiángSinh}}(t)$. Khi đó, hàm $h(t)$ được viết thành:

$$h(t) = \kappa_1 Z_{\text{Tết}}(t) + \kappa_2 Z_{\text{GiángSinh}}(t)$$

Trong biểu thức này, các giá trị của $Z_i(t)$ là nhị phân: nhận giá trị 1 nếu thời điểm t thuộc vào khoảng ảnh hưởng của ngày lễ i, và 0 nếu không. Các hệ số κ_i mang ý nghĩa thực tế

quan trọng: nếu $\kappa_i > 0$, ngày lễ i có xu hướng làm tăng giá trị của chuỗi (ví dụ: số lượng đặt phòng); ngược lại, nếu $\kappa_i < 0$, sự kiện đó có thể làm giảm hoạt động — mặc dù điều này hiếm xảy ra với các ngày lễ phổ biến nhưng có thể xuất hiện với các sự kiện đặc biệt khác. Tóm lại, thông qua việc đưa vào hàm $h(t)$, Prophet cung cấp một cơ chế linh hoạt để mô hình hóa các ảnh hưởng không định kỳ nhưng có tính lặp lại theo lịch, giúp tăng độ chính xác của dự báo trong các bối cảnh nhạy cảm với lịch sự kiện.

Sai số và ước lượng khoảng tin cậy trong dự báo

Trong mô hình Prophet, dự báo không chỉ đơn thuần là một giá trị cố định cho mỗi thời điểm tương lai mà còn bao gồm ước lượng về sự không chắc chắn của dự báo đó, được biểu diễn qua khoảng tin cậy. Sự không chắc chắn này bao gồm hai thành phần chính: sai số ngẫu nhiên và sai số do ước lượng các tham số trong mô hình, như các thành phần xu hướng, tính thời vụ và hiệu ứng ngày lễ. Sai số ngẫu nhiên được giả định là phần dư (residual) hay nhiễu ε_t tuân theo phân phối chuẩn với trung bình bằng 0 và phương sai σ^2:

$$\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

Đây chính là sai số giữa giá trị quan sát thực tế và giá trị dự báo của mô hình tại thời điểm t. Để phản ánh sự không chắc chắn này trong dự báo, Prophet cung cấp khoảng tin cậy (prediction interval), thể hiện vùng giá trị mà dự báo tương lai có khả năng rơi vào với một xác suất nhất định, ví dụ 80% hoặc 95%, tùy cấu hình. Khoảng tin cậy này được xây dựng dựa trên việc kết hợp phân phối sai số ngẫu nhiên với sự không chắc chắn trong ước lượng các tham số của mô hình. Khi sử dụng phương pháp lấy mẫu Markov Chain Monte Carlo (MCMC), Prophet mô phỏng nhiều kịch bản dự báo bằng cách lấy mẫu từ phân phối hậu nghiệm của các tham số. Qua đó, tại mỗi thời điểm dự báo, sẽ thu được một tập hợp các giá trị dự báo mẫu:

$$\{y_1(t^*), y_2(t^*), \ldots, y_N(t^*)\}$$

với N là số lượng mẫu (ví dụ 1000). Giá trị dự báo trung bình được tính bằng trung bình cộng các mẫu này:

$$\hat{y}(t^*) = \frac{1}{N} \sum_{i=1}^{N} y_i(t^*)$$

và khoảng tin cậy 80% được xác định bằng đoạn giữa phân vị thứ 10% và phân vị thứ 90% của các mẫu dự báo:

$$[\text{quantile}_{10\%}(y), \text{quantile}_{90\%}(y)]$$

Khoảng tin cậy này giúp người dùng hiểu rõ hơn về độ biến động và rủi ro tiềm tàng trong dự báo, từ đó có thể đưa ra quyết định phù hợp hơn dựa trên mức độ chắc chắn của các dự báo.

CHƯƠNG 3

Dữ liệu và kết quả thực nghiệm

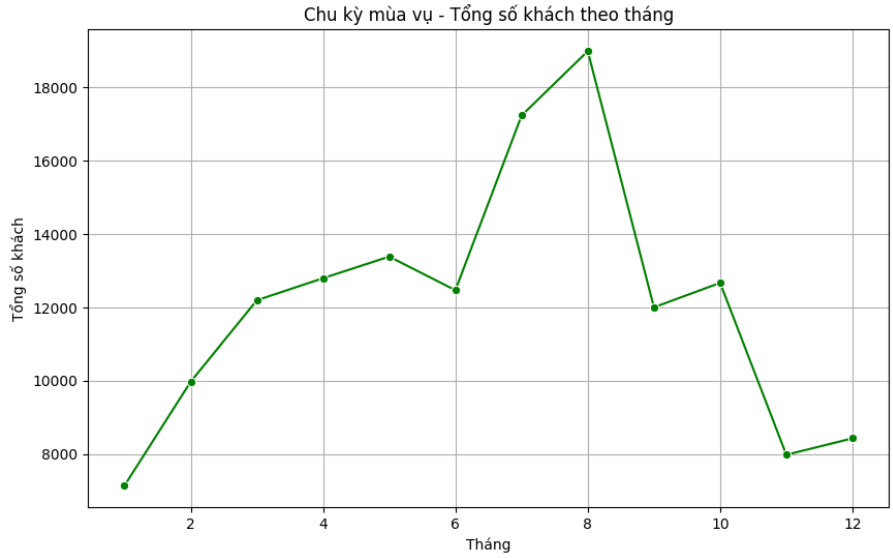
3.1 Dữ liệu đầu vào

Để kiểm nghiệm khả năng dự báo của 2 mô hình gồm mô hình ARIMA và mô hình Prophet trong bối cảnh thực tế, đồ án này đã sử dụng một tập dữ liệu công khai về đặt phòng khách sạn được cung cấp trên nền tảng Kaggle tại địa chỉ: https://www.kaggle.com/datasets/abhi97/hotel-bookings. Dữ liệu này được thu thập từ hai khách sạn ở Bồ Đào Nha, bao gồm một khách sạn nghỉ dưỡng (resort hotel) và một khách sạn thành phố (city hotel), với hơn 119.000 bản ghi đặt phòng trong khoảng thời gian từ tháng 7 năm 2015 đến tháng 8 năm 2017. Mỗi bản ghi trong tập dữ liệu tương ứng với một đơn đặt phòng và chứa thông tin chi tiết như ngày đặt phòng, thời gian lưu trú, loại phòng, số lượng người lớn và trẻ em, thông tin hủy đặt phòng, nguồn đặt phòng, quốc tịch khách hàng, cũng như các đặc điểm liên quan đến giá cả và kênh phân phối. Trong đồ án này, dữ liệu được tiền xử lý và trích chọn những thuộc tính phù hợp để phục vụ cho bài toán dự báo chuỗi thời gian. Cụ thể, trọng tâm được đặt vào việc tổng hợp số lượng đặt phòng theo ngày nhằm tạo ra chuỗi thời gian đầu vào cho mô hình ARIMA và mô hình Prophet. Sau khi xử lý dữ liệu bằng cách loại bỏ khuyết thiếu, xử lý các đơn hàng bị hủy thì dữ liệu có khoảng 75.000 bản ghi gồm các trường dữ liệu là ngày nhận phòng, tổng số khách bao gồm cả người lớn và trẻ con. Sau đó sẽ tổng hợp các bảng ghi thành các ngày sẽ có 793 ngày.Điều này giúp đơn giản hóa bài toán và tập trung đánh giá hiệu quả mô hình trong việc dự báo xu hướng và tính thời vụ của nhu cầu đặt phòng theo thời gian. Việc lựa chọn tập dữ liệu này mang lại nhiều lợi thế, bao gồm độ tin cậy cao, dữ liệu có tính thực tế và đa dạng về đặc điểm, cho phép kiểm chứng khả năng học xu hướng, tính thời vụ và hiệu ứng ngày lễ của mô hình dự báo trong môi trường kinh doanh khách sạn thực tế.

3.2 Phân tích dữ liệu cơ bản để hiểu nhu cầu đặt phòng khách sạn

Việc phân tích dữ liệu là bước đầu tiên và quan trọng trong quá trình xây dựng chiến lược kinh doanh dịch vụ lưu trú. Dựa vào các biểu đồ trực quan hóa dữ liệu, có thể nhận định và dự đoán xu hướng đặt phòng khách sạn theo thời gian.

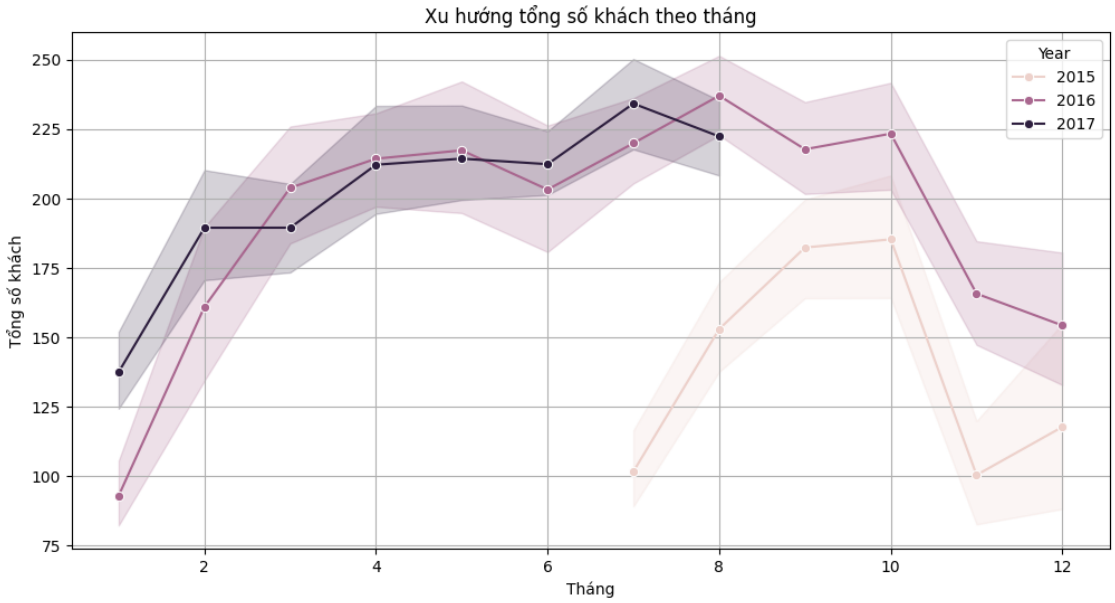
Tổng số khách theo tháng



Hình 3.2.1: Chu kỳ mùa vụ - Tổng số khách theo tháng

Dữ liệu cho thấy được xu hướng rõ rệt về mùa vụ. Tăng trưởng mạnh vào giữa năm, từ tháng 1 đến tháng 8, số lượng khách có xu hướng tăng liên tục, đặc biệt tăng mạnh từ tháng 6 đến tháng 8 – đây có thể là mùa cao điểm du lịch hè. Giảm mạnh sau tháng 8, sau tháng cao điểm là sự giảm sút đáng kể từ tháng 9 đến tháng 11. Tháng thấp điểm nhất là các tháng 1, 11, và 12 thường là thời kỳ thấp điểm trong năm.

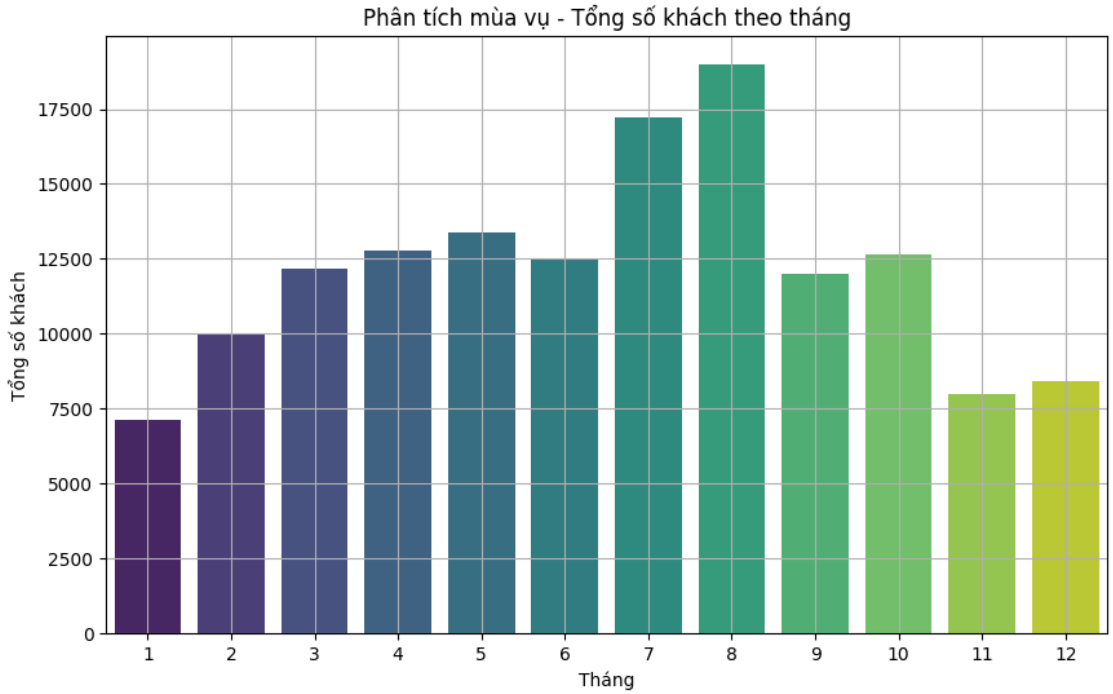
Xu hướng tổng số khách theo tháng qua các năm



Hình 3.2.2: Xu hướng tổng số khách theo tháng (2015–2017)

Qua biểu đồ minh họa xu hướng tổng số khách theo tháng qua ba năm liên tiếp từ 2015 đến 2017. Dù có sự khác biệt về mức độ nhưng mô hình mùa vụ vẫn được giữ nguyên qua từng năm. Điều này cho thấy các yếu tố ảnh hưởng đến hành vi du lịch – chẳng hạn như kỳ nghỉ hè, thời tiết thuận lợi – diễn ra ổn định theo chu kỳ. Đáng chú ý là năm 2016 có sự tăng trưởng vượt trội ở giai đoạn giữa năm, đặc biệt là quý III. Biểu đồ cũng sử dụng vùng bóng thể hiện độ lệch chuẩn, cho thấy mức biến động không lớn trong từng tháng giữa các năm, từ đó khẳng định tính ổn định của xu hướng.

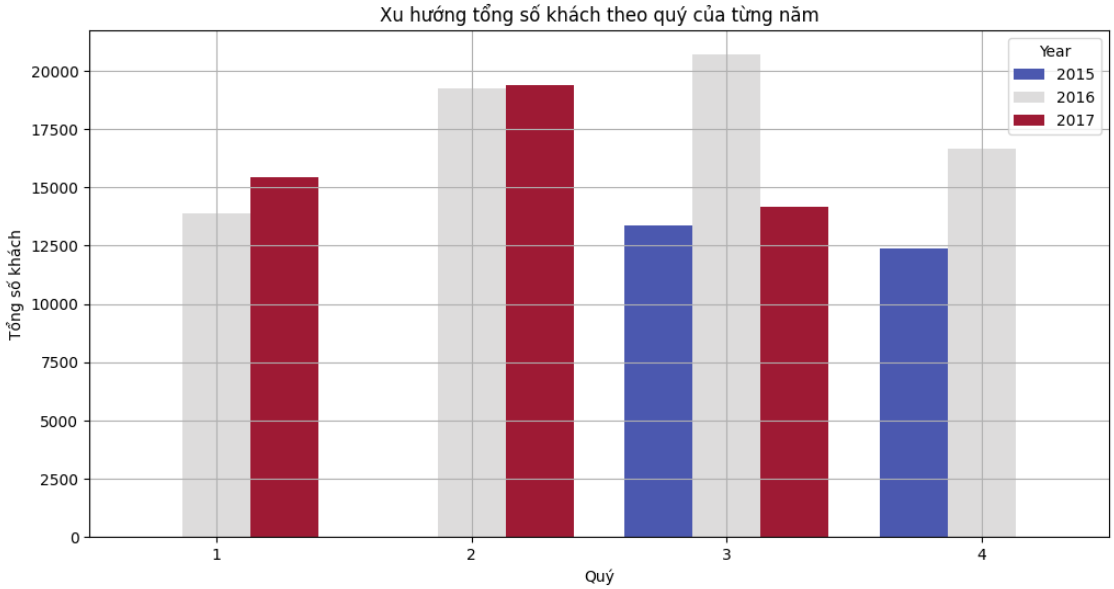
Phân tích mùa vụ - Tổng số khách theo tháng (biểu đồ cột)



Hình 3.2.3: Phân tích mùa vụ - Tổng số khách theo tháng (biểu đồ cột)

Điều này càng nhấn mạnh sự phân hóa rõ ràng giữa các tháng trong năm. Tháng 7 và tháng 8 có số lượng khách vượt trội, trong khi các tháng đầu và cuối năm thường ghi nhận lượng khách thấp hơn nhiều. Phân tích dạng này đặc biệt hữu ích khi lên kế hoạch nguồn lực vận hành khách sạn: vào mùa cao điểm, cần tăng cường nhân sự, mở rộng dịch vụ, và có thể điều chỉnh giá bán phòng; ngược lại, vào mùa thấp điểm, nên đẩy mạnh khuyến mãi hoặc hợp tác với các công ty lữ hành để kích cầu.

Phân tích theo quý



Hình 3.2.4: Xu hướng tổng số khách theo quý của từng năm

Biểu đồ thể hiện tổng số khách theo từng quý của các năm, cho thấy rõ ràng rằng quý II và quý III là hai giai đoạn kinh doanh sôi động nhất. Đặc biệt, quý III năm 2016 chứng kiến lượng khách tăng đột biến, vượt xa các quý còn lại trong cùng năm. Trong khi đó, quý IV luôn là thời điểm có ít khách nhất. Điều này cho thấy các khách sạn cần chuẩn bị kỹ cho hai quý giữa năm – từ sản phẩm, nhân lực đến chương trình bán hàng – và đồng thời cần các chiến lược riêng để giữ công suất hoạt động ổn định trong quý đầu và cuối năm.

Tổng hợp các kết quả phân tích trên, có thể kết luận rằng hành vi đặt phòng khách sạn có tính mùa vụ rất rõ ràng, tập trung cao vào giữa năm – đặc biệt là tháng 7 và tháng 8. Từ đó, các nhà quản lý khách sạn cần có những chiến lược linh hoạt tùy theo mùa: tăng giá và đẩy mạnh marketing vào mùa cao điểm, đồng thời áp dụng chính sách khuyến mãi, gói dịch vụ đi kèm hoặc hợp tác đa kênh để thu hút khách trong mùa thấp điểm. Những phân tích chuyên sâu hơn về các yếu tố như loại phòng, quốc tịch khách, kênh đặt phòng hoặc các dịp lễ đặc biệt sẽ càng giúp tối ưu hiệu quả kinh doanh.

3.3 Kết quả thực nghiệm

Trong phần này, sẽ đi vào xây dựng mô hình ARIMA và mô hình Prophet và đánh giá mô hình với việc dự báo số người đặt phòng khách sạn.

3.3.1 Xây dựng mô hình ARIMA

Đối với mô hình ARIMA cần làm đầy đủ các bước từ kiểm tra tính dừng của chuỗi, sau đó xác định các tham số (p,d,q). Từ đó, xác định được các giá trị (p,d,q) tối nhất dựa trên các chỉ số đánh giá AIC và BIC. Mô hình có giá trị AIC,BIC nhỏ nhất được lựa chọn để huấn luyện trên tập dữ liệu chính thức.

Kiểm định tính dừng của dữ liệu

Để kiểm tra tính dừng của dữ liệu, cần sử dụng phương pháp ADF test để kiểm tra xem chuỗi có dừng không. Sau khi kiểm tra, kiểm định được đầu ra như sau:

Bảng 3.1: Kết quả kiểm định ADF ban đầu

Tham số	**Giá trị**
Test Statistic	-2.603322
p-value	0.092313
Số lượng độ trễ sử dụng	20
Số quan sát sử dụng	772
Giá trị tới hạn ở mức 1%	-3.438849
Giá trị tới hạn ở mức 5%	-2.865291
Giá trị tới hạn ở mức 10%	-2.568767

Dựa vào kết quả kiểm định, giá trị thống kê kiểm định là -2.603322, lớn hơn giá trị tới hạn ở mức 5% (-2.865291) và mức 1% (-3.438849). Đồng thời, giá trị p-value là 0.092313, lớn hơn mức ý nghĩa 0.05. Do đó, sẽ không bác bỏ giả thuyết không, tức là chuỗi dữ liệu không có tính dừng tại thời điểm hiện tại. Vì vậy, cần thực hiện sai phân để làm dừng chuỗi trước khi tiến hành xây dựng mô hình ARIMA. Việc chuỗi không dừng sẽ gây sai lệch trong quá trình dự báo và vi phạm giả định của mô hình ARIMA.

Sử dụng phép sai phân để làm dừng dữ liệu

Sau khi kiểm định ADF với dữ liệu thì nhận thấy dữ liệu chưa dừng, do đó cần áp dụng phép sai phân đối với dữ liệu để làm cho chuỗi dừng. Sau khi áp dụng các phép sai phân và sử dụng kiểm định ADF để kiểm định lại, kiểm định được đầu ra như sau:

Bảng 3.2: Kết quả kiểm định ADF theo các bậc sai phân

Bậc sai phân (d)	**Giá trị ADF**	**p-value**	**Chuỗi dừng**
0	-2.603322	0.092313	Không
1	-10.274579	3.92×10^{-18}	Có
2	-13.093078	1.77×10^{-24}	Có
3	-14.450682	7.11×10^{-27}	Có

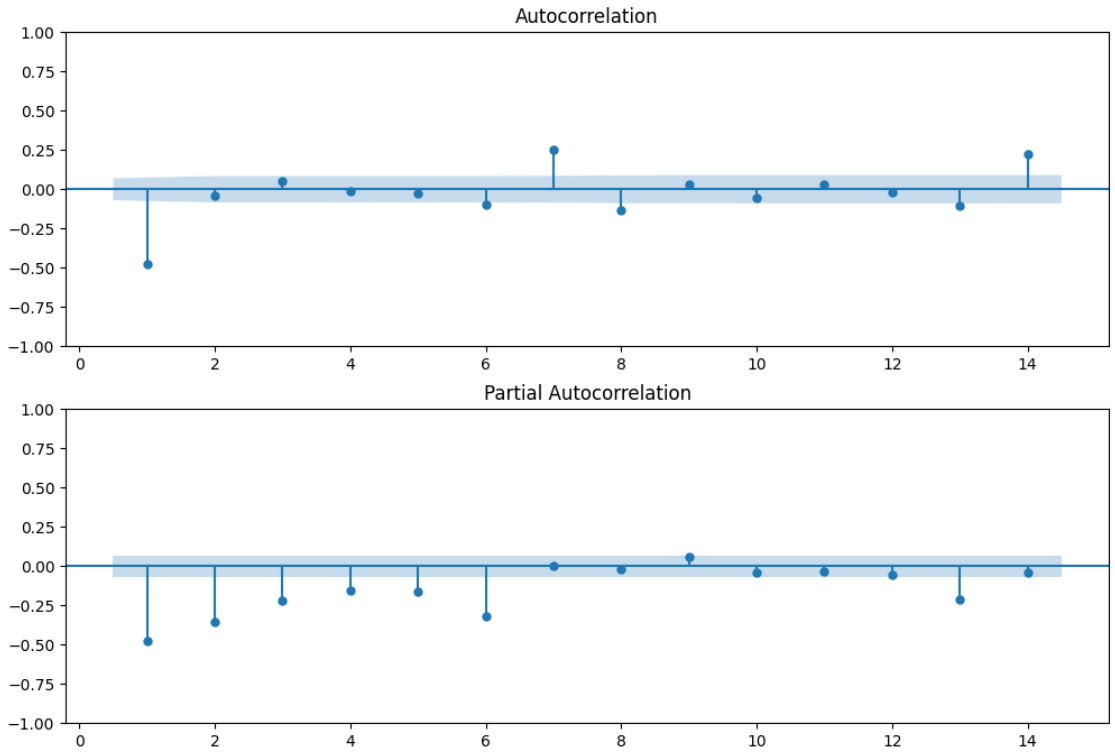
Dựa trên bảng kết quả, tại bậc $d = 0$, chuỗi chưa đạt tính dừng (p-value = 0.092313 $>$ 0.05). Tuy nhiên, ngay tại bậc $d = 1$, giá trị p giảm xuống rất nhỏ (< 0.05), chứng tỏ chuỗi đã đạt tính dừng. Các bậc sai phân cao hơn ($d = 2$ và $d = 3$) cũng cho thấy chuỗi dừng, nhưng không cần thiết phải áp dụng sai phân quá nhiều vì có thể làm mất thông tin trong chuỗi và ảnh hưởng đến chất lượng mô hình. Do đó, bậc sai phân tối ưu được lựa chọn là:

d = 1

Việc lựa chọn bậc sai phân thấp nhất mà chuỗi trở nên dừng là một bước quan trọng giúp đảm bảo mô hình ARIMA không bị "quá sai phân", từ đó duy trì tính hiệu quả và ổn định của mô hình.

Lựa chọn tham số p và q cho mô hình ARIMA

Sau khi xác định được bậc sai phân $d = 1$, bước tiếp theo là lựa chọn các tham số p (số bậc tự hồi quy - AR) và q (số bậc trung bình trượt - MA) phù hợp. Để xác định giá trị khởi đầu cho p và q, cần phân tích hai biểu đồ hàm tự tương quan (ACF) và tự tương quan riêng phần (PACF) như dưới đây.



Hình 3.3.1: Biểu đồ ACF và PACF sau khi sai phân bậc 1

Từ biểu đồ trên cho thấy được PACF có ý nghĩa đến độ trễ khoảng 6 nên giá trị p có thể là $p = 1,2,3,4,5,6$ Cũng tương tự như vậy, ACF có ý nghĩa đến độ trễ khoảng 6 nên giá trị q có thể là $q = 1,2,3,4,5,6$. Sau khi xác định phạm vi tiềm năng cho p và q, mô hình được huấn luyện với nhiều tổ hợp khác nhau của các tham số này.Nên đồ án sẽ sử dụng các tiêu chí AIC và BIC để tìm ra được mô hình tốt nhất.

Bảng 3.3: Một số tổ hợp (p, q) có AIC thấp nhất

p	q	AIC	BIC
6	6	**8609.657**	8670.426
5	6	8611.236	8667.331
6	5	8618.624	8674.719
3	6	8628.279	8675.025
6	0	8633.308	8666.030

Kết quả được so sánh dựa trên hai tiêu chí AIC và BIC để tìm ra mô hình tốt nhất. Các mô hình có AIC thấp nhất sẽ được ưu tiên lựa chọn, vì AIC phản ánh độ phù hợp của mô hình với dữ liệu trong khi vẫn phạt độ phức tạp của mô hình. Kết quả so sánh cho thấy mô hình ARIMA(6, 1, 6) có giá trị AIC thấp nhất (8609.657), đồng thời cũng nằm trong nhóm các mô hình có BIC cạnh tranh. Do đó, mô hình ARIMA(6, 1, 6) được lựa chọn là mô hình tối ưu để sử dụng trong quá trình dự báo chuỗi thời gian.

Huấn luyện mô hình ARIMA(6,1,6)

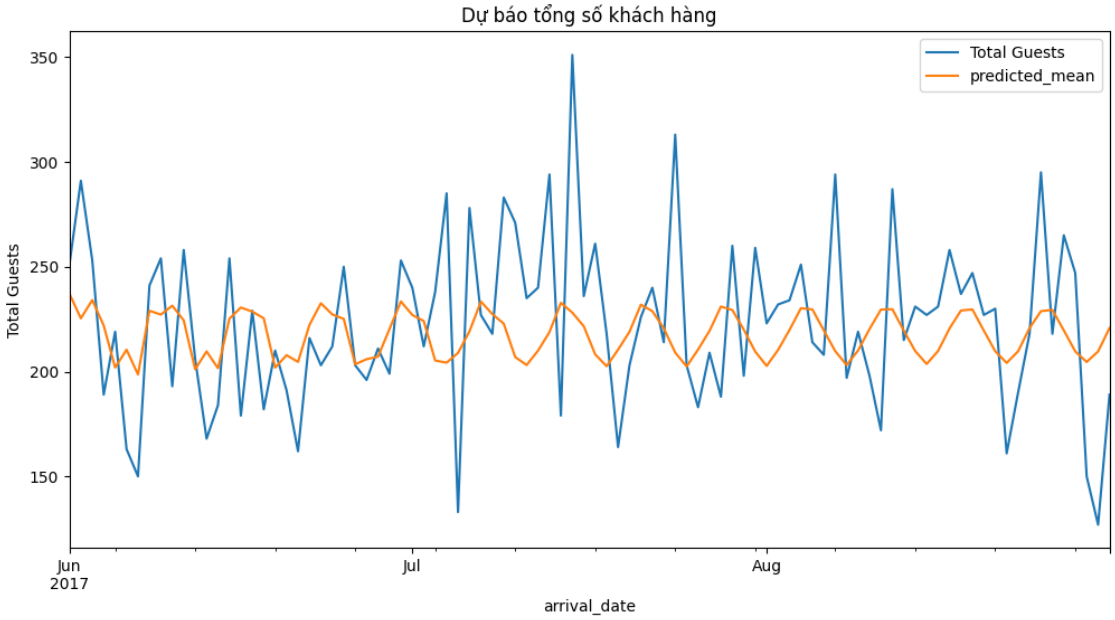
Để huấn luyện mô hình thì sẽ phải chia dữ liệu thành 2 phần huấn luyện và kiểm tra, tập dữ liệu sẽ gồm từ 701 ngày đầu và tập dữ liệu kiểm tra sẽ từ ngày 701 đổ đi. Sau khi huấn luyện mô hình sẽ đánh giá mô hình bằng sai số MAE và MAPE như sau:

Bảng 3.4: Các chỉ số đánh giá hiệu quả mô hình dự báo mô hình ARIMA

Chỉ số đánh giá	**Giá trị**
MAE	30.671
MAPE	0.141

MAE có giá trị 30.671 cho thấy trung bình sai số tuyệt đối giữa giá trị dự báo và giá trị thực tế là khoảng 30.671 đơn vị, đây là một mức sai số khá chấp nhận được. Trong khi đó, MAPE bằng 0.141 (tương đương 14.1%) thể hiện rằng sai số trung bình chiếm khoảng 14.1% so với giá trị thực tế, mức độ lỗi này được xem là hợp lý trong các bài toán dự báo chuỗi thời gian và cho thấy mô hình có khả năng dự báo khá chính xác.

Dự báo



Hình 3.3.2: Dự báo tổng số khách hàng bằng mô hình ARIMA

Hình trên mô tả kết quả dự báo tổng số khách hàng trong tương lai gần sử dụng mô hình ARIMA. Đường màu xanh thể hiện dữ liệu thực tế về số lượng khách hàng, trong khi đường màu cam biểu diễn giá trị dự báo từ mô hình. Nhìn vào biểu đồ, có thể thấy mô hình ARIMA thể hiện xu hướng trung bình khá tốt, tuy nhiên mô hình có xu hướng làm trơn dữ liệu, dẫn đến việc không thể phản ánh được các dao động mạnh trong thực tế. Điều này là đặc trưng của ARIMA khi không kết hợp với thành phần ngoại sinh hoặc yếu tố mùa vụ phức tạp. Tuy nhiên, đường dự báo vẫn bám sát theo xu hướng chung, chứng tỏ mô hình có khả năng dự báo mức tổng thể khá ổn định.

3.3.2 Xây dựng mô hình Prophet

Bên cạnh mô hình ARIMA truyền thống với yêu cầu nghiêm ngặt về tính dừng của chuỗi và quá trình lựa chọn tham số p, d, q thông qua các biểu đồ ACF/PACF và tiêu chí AIC/BIC, mô hình Prophet do Facebook phát triển đã nổi lên như một phương pháp tiếp cận linh hoạt và thân thiện hơn với người dùng trong bài toán dự báo chuỗi thời gian. Khác với ARIMA, Prophet không yêu cầu chuỗi dữ liệu phải dừng, mà thay vào đó, nó tự động phân tách và mô hình hóa các thành phần xu hướng, mùa vụ và ngày đặc biệt, ngày lễ một cách riêng biệt. Điều này đặc biệt phù hợp với dữ liệu thực tế như nhu cầu đặt phòng khách sạn – vốn thường chứa xu hướng dài hạn và chu kỳ rõ rệt theo tuần, tháng, và các dịp lễ. Hơn nữa, Prophet không yêu cầu xác định thủ công các tham số mô hình, giúp tiết kiệm thời gian xử lý và tránh sai sót trong lựa chọn cấu hình. Việc đánh giá mô hình cũng trở nên trực quan hơn khi Prophet hỗ trợ trực tiếp các biểu đồ dự báo

và đo lường sai số như MAE, MAPE mà không cần can thiệp sâu vào cấu trúc thống kê của mô hình.

Huấn luyện mô hình Prophet

Để huấn luyện mô hình Prophet, ngoài bước chia dữ liệu huấn luyện và kiểm tra thì em sẽ tạo thêm một dataframe các ngày lễ của Bồ Đào Nha có ảnh hưởng đến chuỗi thời gian để huấn luyện mô hình. Và chọn các tùy chọn như sau

- **Thành phần ngày lễ (holidays)** được đưa vào nhằm mô hình hóa các dịp lễ có ảnh hưởng đến số lượng khách.
- **Yếu tố mùa vụ**:
  - `yearly_seasonality=True`: cho phép mô hình nhận diện chu kỳ theo năm.
  - `weekly_seasonality=True`: phản ánh các dao động theo từng ngày trong tuần.
  - `daily_seasonality=False`: không xét đến yếu tố theo ngày vì không cần thiết.
- **Chế độ mùa vụ**: `seasonality_mode='multiplicative'`, thích hợp khi hiệu ứng mùa vụ biến đổi tỷ lệ với mức độ tổng thể của chuỗi.
- **Khoảng tin cậy**: `interval_width=0.95`, tương ứng với khoảng dự báo 95%.

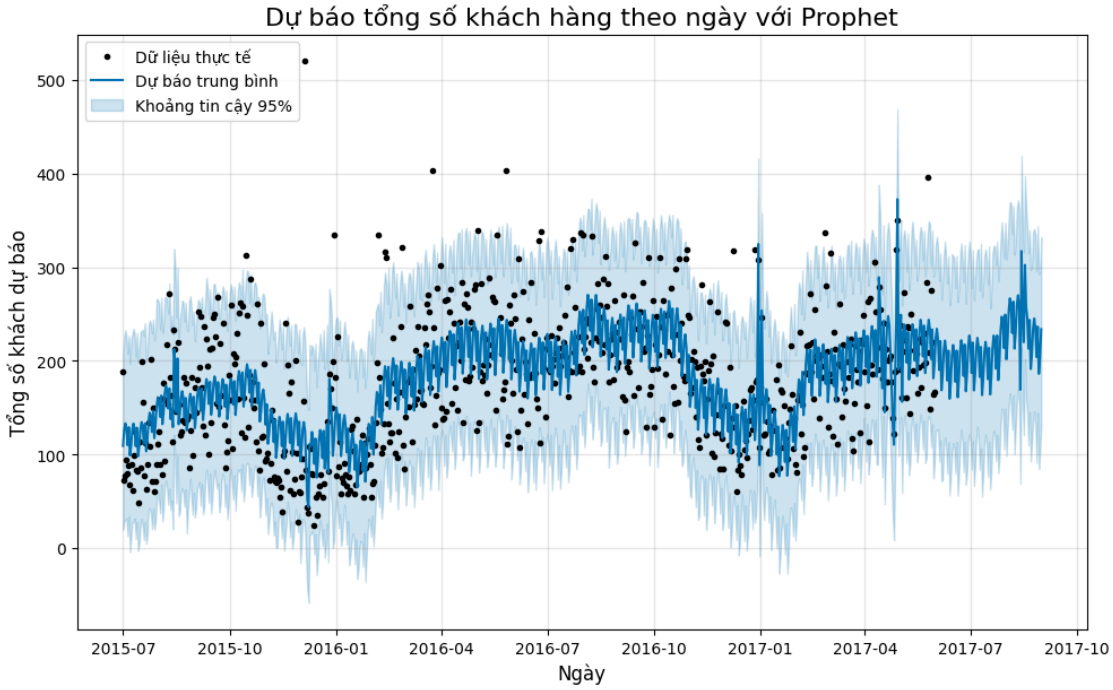
Sau khi huấn luyện mô hình, đồ án sẽ đánh giá mô hình bằng sai số MAE và MAPE như sau:

Bảng 3.5: Các chỉ số đánh giá hiệu quả mô hình dự báo Prophet

Chỉ số đánh giá	Giá trị
MAE	33.096
MAPE	0.148

Kết quả đánh giá mô hình Prophet cho thấy sai số trung bình tuyệt đối (MAE) đạt 33.096, và sai số phần trăm trung bình tuyệt đối (MAPE) là 0.148, tương ứng với 14.8%. Đây là mức sai số tương đối thấp, cho thấy mô hình Prophet có khả năng dự báo tổng số lượng khách ở mức khá chính xác, đặc biệt phù hợp trong bối cảnh dữ liệu có yếu tố mùa vụ và dịp lễ.

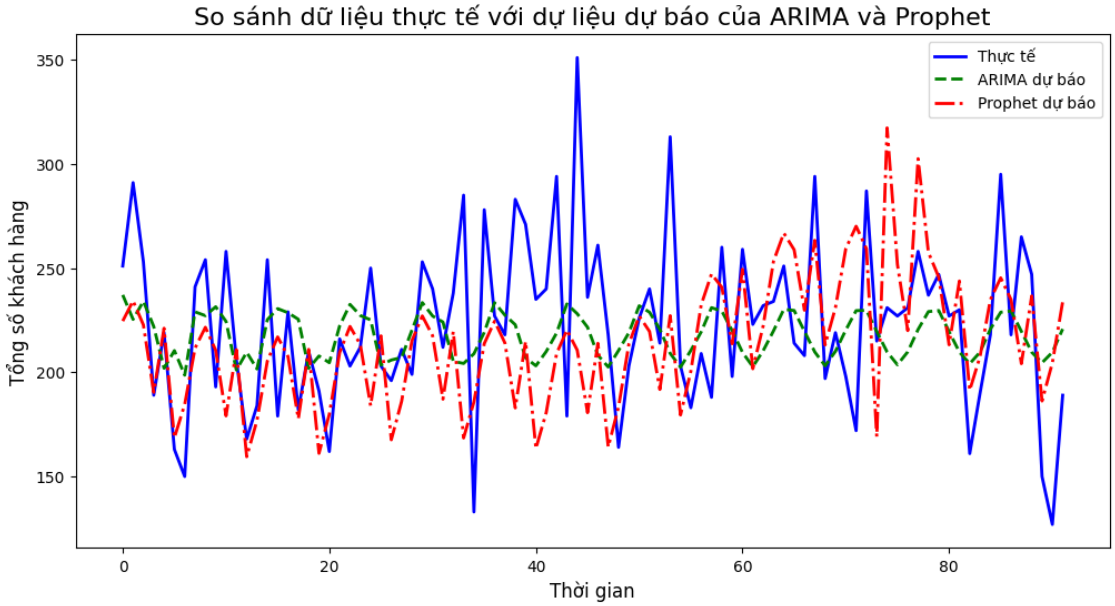
Dự báo



Hình 3.3.3: Dự báo tổng số khách hàng bằng mô hình Prophet

Biểu đồ dự báo tổng số khách hàng theo ngày với mô hình Prophet cho thấy mô hình đã bám sát tốt xu hướng thực tế của dữ liệu, đặc biệt thể hiện rõ các chu kỳ mùa vụ và dao động định kỳ theo tuần và năm. Đường dự báo trung bình phù hợp với các điểm dữ liệu thực tế, chứng tỏ Prophet xử lý hiệu quả các đặc trưng thời gian như xu hướng và mùa vụ. Khoảng tin cậy 95% bao phủ phần lớn các điểm dữ liệu, phản ánh độ tin cậy tương đối cao của mô hình. Tuy nhiên, vẫn có một số điểm thực tế nằm ngoài khoảng tin cậy này, cho thấy sự tồn tại của các giá trị ngoại lệ hoặc biến động bất thường mà mô hình chưa dự báo chính xác. Mặc dù Prophet đã sử dụng thông tin về ngày lễ để cải thiện dự báo, một số đột biến về số lượng khách vẫn chưa được phản ánh đầy đủ, có thể do tính phức tạp của các sự kiện đặc biệt hoặc biến động ngắn hạn. Ngoài ra, mô hình có xu hướng làm mượt dữ liệu nên chưa phản ứng nhạy bén với các biến động cực mạnh trong ngắn hạn, dẫn đến một số sai lệch so với dữ liệu thực tế. Tổng thể, mô hình Prophet phù hợp với dữ liệu có tính mùa vụ rõ rệt như chuỗi khách hàng này, nhưng có thể cải thiện hơn nữa bằng cách bổ sung các biến ngoại sinh hoặc điều chỉnh các tham số để tăng độ nhạy trong dự báo.

3.3.3 So sánh 2 mô hình



Hình 3.3.4: So sánh dữ liệu thực tế với dự liệu dự báo của ARIMA và Prophet

Hai mô hình ARIMA(6,1,6) và Prophet đều được sử dụng để dự báo tổng số khách hàng theo chuỗi thời gian, nhưng có những điểm khác biệt rõ rệt về cách tiếp cận và hiệu quả dự báo. Mô hình ARIMA dựa trên cấu trúc tự hồi quy và trung bình trượt, cho phép phân tích sâu về các thành phần tự tương quan trong dữ liệu thông qua các chỉ số như ACF và PACF. ARIMA cho kết quả tương đối chính xác với MAE là 30.671 và MAPE là 0.141, đồng thời phần dư không có tự tương quan còn lại, thể hiện mô hình phù hợp với dữ liệu. Tuy nhiên, phần dư có đặc điểm không chuẩn và phương sai không đồng nhất, làm hạn chế khả năng dự báo trong một số trường hợp biến động mạnh hoặc không tuyến tính. Ngược lại, mô hình Prophet không cần phải xác định các tham số p, d, q như ARIMA, mà tự động xử lý các thành phần xu hướng, mùa vụ theo năm và tuần cùng với các ngày lễ. Prophet có MAE là 33.096 và MAPE là 0.148, chỉ hơi kém hơn ARIMA về mặt độ chính xác, nhưng bù lại có ưu điểm trong việc nắm bắt các chu kỳ phức tạp và đặc biệt hiệu quả với dữ liệu có mùa vụ rõ ràng và các biến động theo kỳ. Khoảng tin cậy dự báo của Prophet cũng cung cấp thông tin về độ tin cậy của dự báo, trong khi ARIMA tập trung vào mô hình hóa các cấu trúc thời gian. Tóm lại, nếu dữ liệu có cấu trúc tương quan phức tạp và cần phân tích chi tiết các tham số, ARIMA là lựa chọn phù hợp hơn. Còn nếu dữ liệu có tính mùa vụ rõ ràng, có nhiều biến động định kỳ theo ngày, tuần, năm hoặc có các ngày lễ ảnh hưởng, Prophet là mô hình linh hoạt, dễ sử dụng và hiệu quả trong dự báo tổng quát. Việc lựa chọn mô hình cuối cùng cần cân nhắc dựa trên mục tiêu cụ thể và đặc điểm dữ liệu thực tế.

CHƯƠNG 4

Kết luận

Báo cáo đã tiến hành nghiên cứu và xây dựng hai mô hình dự báo chuỗi thời gian phổ biến là ARIMA và Prophet để dự báo tổng số khách hàng trong ngành kinh doanh lưu trú dựa trên dữ liệu thực tế từ năm 2015 đến 2017.

Chương 1 trình bày cơ sở lý thuyết, cung cấp cái nhìn tổng quan về thị trường kinh doanh lưu trú và các kiến thức toán học cơ bản trong chuỗi thời gian, làm nền tảng cho việc xây dựng mô hình dự báo.

Chương 2 tập trung vào giới thiệu và phân tích chi tiết hai mô hình ARIMA và Prophet. Mô hình ARIMA thể hiện khả năng khai thác cấu trúc tự hồi quy và tính dừng của dữ liệu, trong khi Prophet mang lại sự linh hoạt trong xử lý yếu tố mùa vụ và các biến ngoại sinh như ngày lễ, cùng khả năng dự báo với phương pháp đơn giản và hiệu quả.

Chương 3 trình bày dữ liệu và kết quả thực nghiệm. Qua đó, mô hình ARIMA(6,1,6) đạt được kết quả khá tốt với các chỉ số MAE và MAPE lần lượt là 30.671 và 0.141, trong khi mô hình Prophet với các ưu điểm về khai thác mùa vụ và yếu tố lễ tết cũng thể hiện kết quả gần tương đương với MAE = 33.096 và MAPE = 0.148. Kết quả so sánh cho thấy cả hai mô hình đều có khả năng dự báo tương đối chính xác, tuy nhiên mỗi mô hình có những ưu điểm và hạn chế riêng, phụ thuộc vào đặc điểm của dữ liệu và yêu cầu ứng dụng thực tế.

Tổng kết lại, nghiên cứu đã làm rõ các bước xây dựng, đánh giá mô hình dự báo và đưa ra các nhận xét giúp người dùng lựa chọn mô hình phù hợp. Mô hình ARIMA phù hợp với dữ liệu có tính tự tương quan mạnh, còn Prophet thích hợp cho dữ liệu có tính mùa vụ phức tạp và yếu tố ngoại lai. Đây cũng là tiền đề cho các nghiên cứu tiếp theo như kết hợp mô hình, xử lý dữ liệu phi tuyến hoặc thêm các biến ngoại sinh nhằm nâng cao độ chính xác dự báo trong ngành lưu trú.

CHƯƠNG 5

Tài liệu tham khảo

1. Nguyễn Mạnh Hùng (Chủ biên) và Vũ Thị Hương, *Thống kê các quá trình ngẫu nhiên*, Nhà xuất bản Trường Đại học Giao thông Vận tải, Hà Nội, 2023.

2. G. E. P. Box, G. M. Jenkins, G. C. Reinsel và G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed., Wiley, 2015.

3. Taylor, S. J., và Letham, B. (2017, February). *Prophet: Forecasting at scale.* Facebook Research Blog. https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale/

4. Modeh, S. (n.d.). *Hotel booking EDA and demand forecasting*.* Kaggle. Retrieved May 20, 2025, from https://www.kaggle.com/code/sabrimodeh/hotel-booking-eda-demand-forecasting