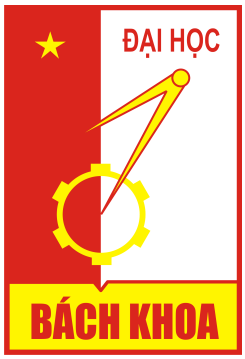
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC



BÀI TOÁN DỰ BÁO PHÂN BỔ TÀI NGUYÊN TRONG MẠNG WIFI MARKETING

ĐỒ ÁN TỐT NGHIỆP

Chuyên ngành: TOÁN TIN

Giảng viên hướng dẫn: TS. TẠ ANH SƠN

Sinh viên thực hiện: PHẠM NGỌC BÁCH

Lớp: CTTN Toán Tin - K63

HÀ NỘI – 2022

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

(a) Mục tiêu:

(b) Nội dung:

2. Kết quả đạt được

(a)

3. Ý thức làm việc của sinh viên:

(a)

Hà Nội, ngày 00 tháng 00 năm 2022

Giảng viên hướng dẫn

TS. Tạ Anh Sơn

Lời cảm ơn

Lời đầu tiên, tôi xin gửi lời cảm ơn sâu sắc nhất tới TS. Tạ Anh Sơn, Bộ môn Toán ứng dụng, Viện Toán ứng dụng và Tin học, Trường Đại học Bách Khoa Hà Nội. Trong thời gian thực hiện đồ án, thầy đã trực tiếp dành nhiều thời gian tận tình hướng dẫn cho tôi những định hướng khoa học, chỉ dẫn sát sao trong thời gian nghiên cứu và thực hiện đồ án này. Thầy cũng dành nhiều thời gian nhận xét về các ý tưởng, kết quả của tôi.

Tôi cũng xin cảm ơn các thầy cô Viện Toán ứng dụng và Tin học, Trường Đại học Bách khoa Hà Nội đã dạy dỗ tôi trong suốt những năm vừa qua, giúp tôi có nền tảng kiến thức vững chắc để hoàn thành đồ án này.

Tôi xin cảm ơn công ty cổ phần Công nghệ và Truyền thông AWING Việt Nam đã cho phép tôi sử dụng bộ dữ liệu thực nghiệm để nghiên cứu và sử dụng kết quả trong đồ án này.

Tôi xin chân thành cảm ơn!

Hà Nội, ngày 00 tháng 00 năm 2022

Tác giả đồ án

Phạm Ngọc Bách

Tóm tắt nội dung Đồ án

1. Trình bày khái niệm, các thuật ngữ trong WiFi marketing và bài toán dự báo phân bổ tài nguyên, ý nghĩa của bài toán
2. Mô tả lý thuyết về phương pháp phát hiện điểm bất thường SR, lý thuyết về mô hình NeuralProphet. Trình bày cách chọn các hệ số cho mô hình học sâu, lý thuyết các phương pháp đánh giá mô hình.
3. Xây dựng mô hình dự báo cho bài toán dự báo
4. Thực nghiệm với bộ dữ liệu và đánh giá kết quả
5. Kết luận và định hướng tiếp theo của bài toán

Hà Nội, ngày 00 tháng 00 năm 2022

Tác giả đồ án

Phạm Ngọc Bách

Mục lục

Bảng ký hiệu và chữ viết tắt

SR	Spectral Residual
AR	Auto-regression
WiFi	Wireless Fidelity
NP	Neural Prophet
NN	Neural Netwwork
MSE	Mean Square Error
ReLU	Rectified Linear Unit
MAE	Mean Absolute Error
sMAPE	Symmetric Mean Absolute Percentage Error
MAPE	Mean Absolute Percentage Error

Danh sách bảng

Danh sách hình vẽ

Mở đầu

Với sự phát triển mạnh mẽ của công nghệ trên thế giới nhự hiện nay, Việt Nam cũng đang trong quá trình hội nhập để vào trong cuộc chạy đua này.Nắm bắt được nhu cầu đó, các công ty khởi nghiệp về lĩnh vực marketing trên các thiết bị thông minh đang đầu tư nhiều. Mobile marketing đi cùng công cuộc đó trở thành một lĩnh vực phát triển chóng mặt hiện nay. Với thói quen của người dùng ở Việt Nam, khi đi đến các địa điểm công cộng thường có Wifi miễn phí để người dùng truy cập sử dụng. Đây là một kênh đang phát triển và có tiềm năng về marketing rất lớn. Khái niệm WiFi marketing từ đó được quan tâm.

WiFi marketing là hình thức quảng cáo thông qua lượt truy cập của người dùng vào mạng WiFi công cộng . Những điểm truy cập này được cài đặt cấu hình những bước yêu cầu đăng nhập của đơn vị quảng cáo nhằm giới thiệu tới người sử dụng WiFi sản phẩm, dịch vụ và thương hiệu từ các doanh nghiệp.

WiFi Marketing phù hợp với tất cả các ngành nghề:

- Dịch vụ tài chính, ngân hàng
- Chuỗi cửa hàng dịch vụ, trung tâm vui chơi giải trí
- Dịch vụ y tế chăm sóc sức khỏe
- Khách sạn - nhà hàng - quán cafe
- Khu vực chăm sóc khách hàng của các doanh nghiệp
- Trường đại học, các trung tâm về giáo dục

Để có một kế hoạch kinh doanh hiệu quả, những người quản lý cần quan tâm đến lượt truy cập vào mạng WiFi nhằm lên kế hoạch phân chia các đợt quảng cáo cho các chiến dịch một cách tối ưu nhất. Từ nhu cầu đó, bài toán dự báo lượt truy cập vào mạng WiFi được đặt ra.

Tại mỗi một địa điểm, người quản lý sẽ thống kê lượt truy cập theo từng ngày. Khi sắp xếp các số lượng người truy cập này theo thời gian, ta được một chuỗi thời gian với các điểm giá trị tương ứng là số lượng người truy cập từng ngày. Thực tế cho thấy, lượng người truy cập là khác nhau tại các thời điểm. Ví dụ vào cuối tuần, lượng người truy cập sẽ tăng hơn các ngày tong tuần với các địa điểm vui chơi; hoặc với kì lễ dài như 30/4-1/5 lượt truy cập này tăng đột biến so với ngày bình thường.

Từ những gì bài toán được đặt ra, đồ án "Dự báo phân bổ tài nguyên trong mạng WiFi marketing" trình bày một quy trình để xử lý dữ liệu và dự báo số lượt người truy cập WiFi của các điểm truy cập trong 30 ngày tiếp theo.

Chương 1

Khái quát mô hình WiFi marketing và bài toán dự báo tài nguyên

Chương này trình bày khái niệm, tình hình kinh tế tại Việt Nam về WiFi marketing, bài toán dự báo phân bổ tài nguyên và ý nghĩa thực tiễn.

1.1 Quảng cáo trực tuyến

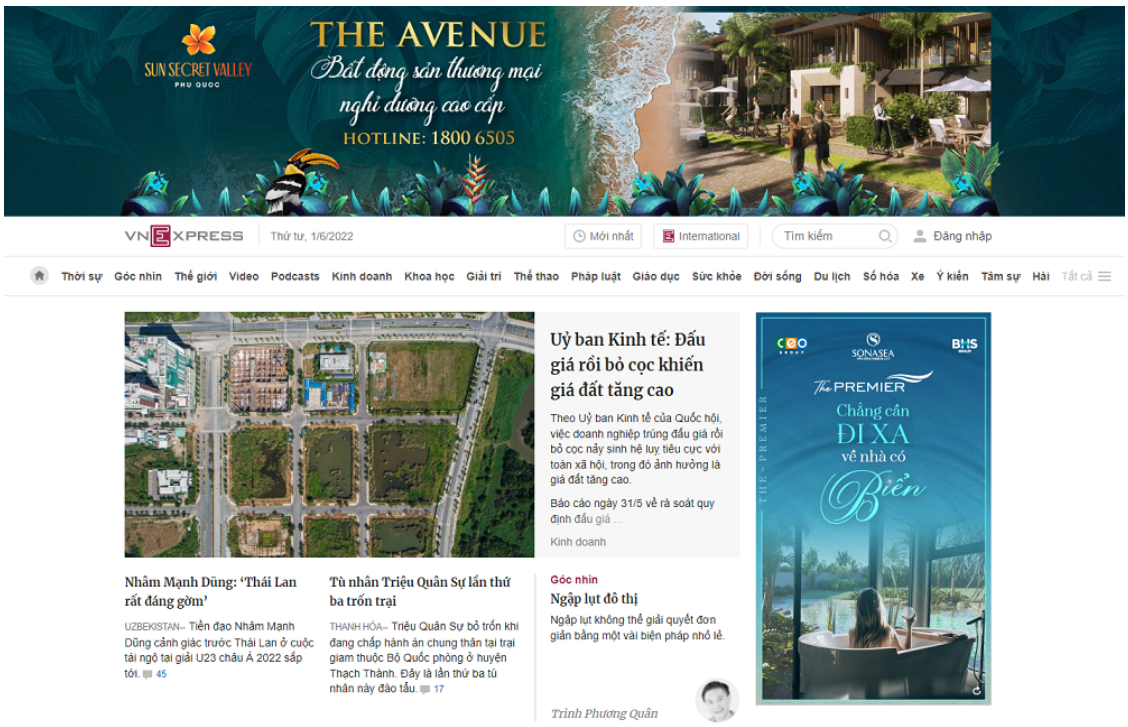
Trong những năm gần đây, cùng với sự phát triển của Internet, nhu cầu và hành vi mua sắm của người tiêu dùng có sự thay đổi rõ rệt. Đặc biệt, với sự ra đời của smartphone thì hiện nay, các doanh nghiệp vừa phải tìm cách đáp ứng được nhu cầu mua sắm và thông tin của người dùng trên nền tảng mới, lại vừa phải tìm cách phát triển mạng lưới khách hàng mới.

Vậy nên, song song với mạng xã hội thì quảng cáo trực tuyến là một sự lựa chọn không thể thiếu trong quá trình xây dựng kế hoạch marketing. Với sự bùng nổ công nghệ như hiện nay, cơ hội tiếp cận với Internet đối với nhiều người ngày càng dễ dàng. Quảng cáo trực tuyến đang ngày càng phát triển với tốc độ nhanh. Năm 2016, doanh thu quảng cáo trực tuyến ở Mỹ đã vượt qua doanh thu quảng cáo qua truyền hình, đạt tổng doanh thu 72.5 tỷ đô-la. Năm 2021, tổng doanh thu của quảng cáo trực tuyến tại Mỹ đạt 190 tỷ đô-la, tăng trưởng lên đến 262% [1].

[1] https://www.statista.com/

Có nhiều loại hình quảng cáo trực tuyến. Tiêu biểu có thể kể đến:

- Quảng cáo hiển thị
- Quảng cáo Retargeting
- Quảng cáo tìm kiếm
- Quảng cáo video
- Quảng cáo thông qua các mạng xã hội



Hình 1.1: Quảng cáo trực tuyến hiển thị trên trang báo điện tử VNEXPRESS

Với hỗ trợ của Internet và thiết bị, quảng cáo trực tuyến có thể hiển thị đa dạng loại nội dung như văn bản, hình ảnh, video, thậm chí quảng cáo tương tác.

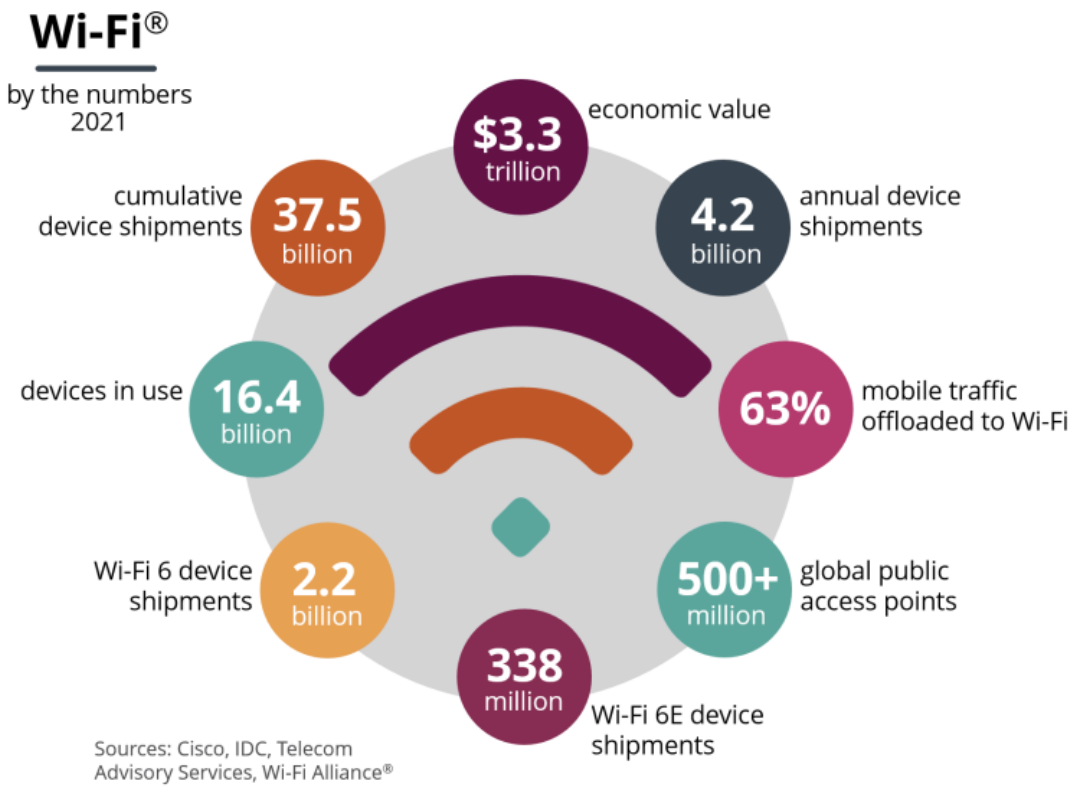
Một trong nhiều ưu điểm của quảng cáo trực tuyến là cách tính toán chi phí quảng cáo. Có nhiều hình thức thanh toán khác nhau, dựa vào tương tác của người tiêu dùng với quảng cáo, một số cách thức phổ biến như:

- CPC (cost per click - giá mỗi lần click)

- CPM (cost per mille - giá cho mỗi một nghìn lượt hiển thị)
- CPE (cost per engagement - giá trên mỗi lượt tương tác)
- CPV (cost per view - giá cho mỗi lần xem)
- CPI (cost per install - giá cho mỗi lần cài đặt)

1.2 Mô hình WiFi marketing

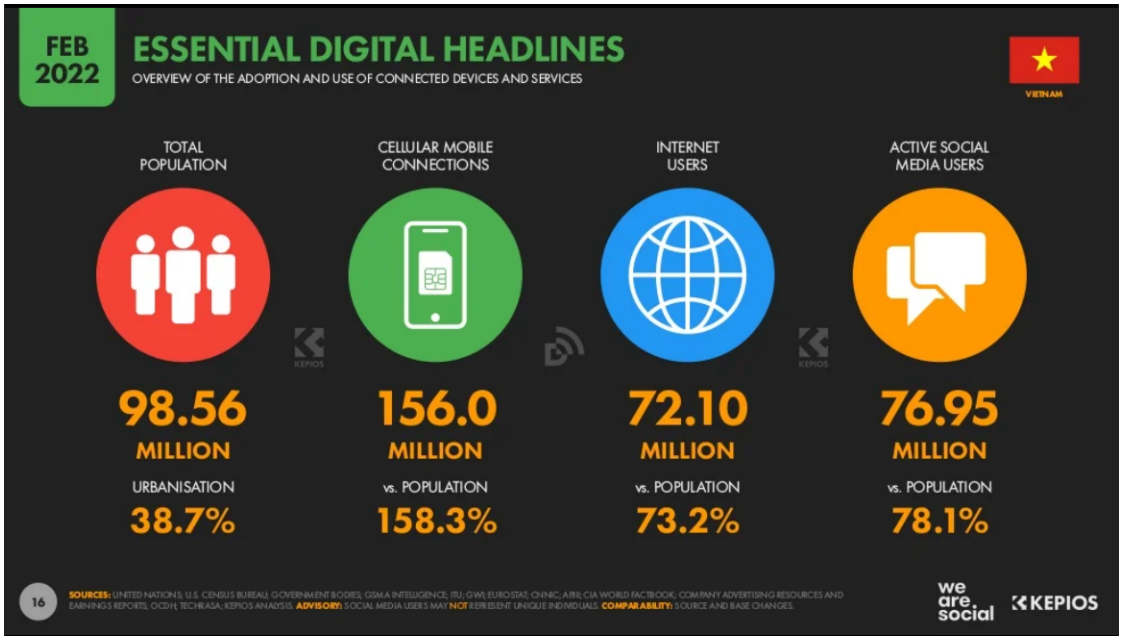
Theo thống kê của Statista, tính đến hết năm 2021 trên toàn thế giới có 22,1 nghìn tỉ thiết bị truy cập mạng Wifi. Ở các nước phát triển, mạng Wifi nắm vai trò quan trọng trong nền kinh tế. Giá trị kinh tế của Wi-Fi ở Australia là 34,7 tỷ USD vào năm 2021, dự kiến sẽ tăng lên 41,7 tỷ USD vào năm 2025. Theo Opensignal, người truy cập Wifi tại Australia dành hơn 52% thời gian của họ để kết nối với Wi-Fi.



Hình 1.2: Thống kê về mạng Wifi trên toàn thế giới

Hoa Kỳ là một trong những quốc gia có lượng người sử dụng Wi-Fi nhiều nhất trên thế giới. Cisco ước tính trên cả Hoa Kỳ có 33,5 triệu điểm truy cập Wi-Fi trả phí, trong đó ước tính số điểm truy cập Wi-Fi công cộng miễn phí vào khoảng 18,6 triệu. 85% thuê bao băng thông rộng của Hoa Kỳ có khả năng truy cập Wi-Fi tại nhà và người dùng di động kết nối Internet thông qua Wi-Fi qua mạng di động với hơn 55% thời gian của họ. Hoa Kỳ cũng có một hệ sinh thái sản xuất mạnh mẽ và doanh nghiệp sử dụng ngày càng nhiều, điều này đã hỗ trợ sự gia tăng giá trị của Wi-Fi. Tổng giá trị kinh tế của Wi-Fi tại Hoa Kỳ vào năm 2021 là 995 tỷ USD.

Theo We Are Social[2], tại Việt Nam lượng người truy cập Internet lên tới 72,10 triệu người, tương ứng với 73,2% dân số. Dữ liệu từ GSMA Intelligence cho thấy Việt Nam hiện có 156 triệu thuê bao di động. Thời gian online trung bình mỗi ngày của người Việt Nam khoảng 6 tiếng 38 phút, trong đó thời gian truy cập bằng thiết bị di động chiếm tới 53,2%. Tất cả những con số trên cho ta thấy được tiềm năng marketing giữa việc kết hợp giữa điện thoại, WiFi và mạng xã hội là rất lớn, đây cũng chính là cở sở nền tảng để WiFi marketing ra đời.



Hình 1.3: Thống kê người sử dụng Internet tại Việt Nam tính tới tháng 2 năm 2022

[2]https://www.wearesocial.com/

1.3 Bài toán dự báo phân bổ tài nguyên trong WiFi marketing

1.3.1 Một vài khái niệm

Trước khi đi vào bài toán dự báo, có các khái niệm cần nắm được:

- Điểm truy cập: là một địa điểm công cộng được lắp đặt WiFi miễn phí.
- Một lượt truy cập: là một lượt người dùng truy cập vào mạng WiFi của điểm truy cập. Một lượt truy cập tương đương với một lượt xem quảng cáo.
- Số lượt truy cập hàng ngày: là tổng lượt truy cập theo ngày tính từ 0h ngày hôm nay đến 0h ngày hôm sau (theo múi giờ Việt Nam)
- Nhà cung cấp dịch vụ WiFi marketing: là nhà cung cấp dịch vụ WiFi marketing cho các địa điểm công cộng và nhận đặt quảng cáo từ các nhãn hàng, các doanh nghiệp.
- Nhà quảng cáo: là khách hàng sử dụng dịch vụ WiFi marketing, đây có thể các doanh nghiệp, các cửa hàng, ...
- Người dùng: là người truy cập vào mạng WiFi marketing.
- Chuỗi dữ liệu thời gian: là chuỗi được tạo thành khi số lượt truy cập của mỗi điểm WiFi marketing được sắp xếp theo thứ tự thời gian, mỗi điểm dữ liệu tương ứng với một ngày.

1.3.2 Phát biểu bài toán

Bài toán dự báo tài nguyên trong mạng WiFi marketing là bài toán dự báo lượt truy cập người dùng tại mỗi điểm truy cập.

Đầu vào:

- Danh sách ngày lễ, ngày có sự kiện lớn tại địa điểm dự đoán

- Dữ liệu gồm số lượt truy cập theo ngày của các địa điểm truy cập vào mạng WiFi marketing theo dạng chuỗi thời gian

Đầu ra:

- Kết quả sau khi dự báo lượt truy cập vào mạng WiFi marketing vào từng địa điểm. Thời gian dự báo trong tương lai là 30 ngày.

1.4 Ý nghĩa bài toán

Việc dự báo bài lượt truy cập WiFi để thuận tiện cho việc lên kế hoạch trong công việc kinh doanh như nhận các hợp đồng để cho doanh thu tốt nhất. Đầu ra của bài toán cũng được sử dụng cho bài toán tối ưu việc phân bổ tài nguyên giữa các địa điểm khi chạy các chiến dịch quảng cáo.

1.5 Tóm tắt chương

Trong chương 1 đã trình bày các khái niệm về WiFi marketing và bài toán dự báo phân bổ tài nguyên trong mạng WiFi này, ý nghĩa bài toán với thực tế.

Chương 2

Lý thuyết các phương pháp

2.1 Phát hiện bất thường trên chuỗi thời gian bằng phương pháp Spectral Residual

Khi chúng ta xây dựng mô hình học máy, việc cần làm trước đó, và hầu như chiếm nhiều thời gian nhất, là tiền xử lý dữ liệu. Tiền xử lý dữ liệu (Data Pre-Processing) là một kỹ thuật được sử dụng để chuyển đổi dữ liệu thô thành một định dạng dễ hiểu. Dữ liệu trong thế giới thực (dữ liệu thô) luôn không đầy đủ và dữ liệu đó không thể được gửi qua các mô hình vì nó sẽ gây ra một số lỗi nhất định. Bởi lẽ các bộ dữ liệu ứng với các bài toán trong thực tế rất khác nhau và mỗi bài toàn thì đối mặt với những thách thức khác nhau về mặt dữ liệu. Và trong đồ án này tôi đề xuất sử dụng phương pháp Spectral Residual (SR) để phát hiện các điểm bất thường trên chuỗi dữ liệu, sau đó có phương án đề xuất xử lý với những điểm đó.

Được nghiên cứu và phát triển từ năm 2019 bởi đội ngũ nghiên cứu của Microsoft, thuật toán Spectral Residual (SR) [1] cho ta một phương pháp phát hiện điểm bất thường bằng sơ đồ saliency .Thuật toán SR bao gồm 3 bước chính:

1. Biến đổi chuỗi Fourier để có được phổ biên độ log
2. Tính toán spectral residual
3. Biến đổi ngược chuỗi Fourier thành miền không gian

Với chuỗi x là chuỗi đầu vào một chiều với độ dài bằng, các bước trên có thể được thể hiện qua một chuỗi công thức biến đổi như sau:

$$\begin{aligned} A(f) &= Amplitude(\mathcal{F}(x)) \\ P(f) &= Phrase(\mathcal{F}(x)) \\ L(f) &= log(A(f)) \\ AL(f) &= h_q(f).L(f) \\ R(f) &= L(f) - AL(f) \\ S(x) &= \|\mathcal{F}^{-1}(exp(R(f) + iP(f)))\| \end{aligned}$$

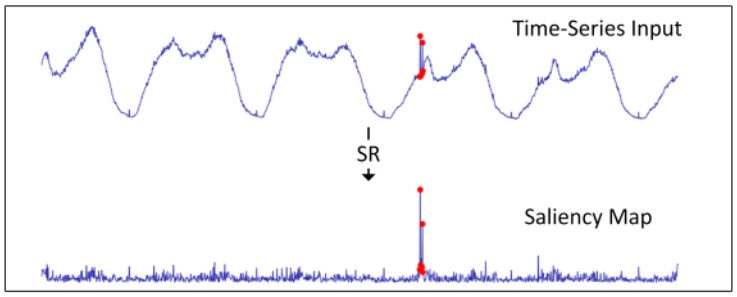
trong đó :

- $\mathcal{F}$ và $\mathcal{F}^{-1}$ là phép biến đổi Fourier và phép biến đổi Fourier ngược
- $A(f)$ là phổ biên độ của chuỗi x
- $P(f)$ là phổ pha của chuỗi x
- $L(f)$ là log của $A(f)$
- $AL(f)$ là phổ trung bình của $L(f)$, được tính bằng tích chập với $h_q(f)$
- $h_q(f)$ là một ma trận $q \times q$:

$$h_q(f) = \frac{1}{q^2} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}$$

- $R(f)$ là spectral residual
- $S(x)$ là kết quả của phép biến đổi, gọi là saliency map

Như mô tả trên, đầu vào của thuật toán SR là chuỗi thời gian thực, kết quả là một saliency map. Hình 2.1 biểu diễn cho ta thấy một ví dụ về quá trình xử lý SR:



Hình 2.1: Minh họa ví dụ kết quả của thuật toán SR

Quan sát hình 2.1 có thể thấy những điểm màu đỏ trong saliency map có ý nghĩa hơn những điểm còn lại. Những điểm bất thường có thể dùng quy tắc để chú thích. Bằng cách áp dụng ngưỡng giá trị τ, chuỗi đầu ra $O(x)$ được tính bằng:

$$O(x) = \begin{cases} 1, & \text{nếu } \frac{S(x_i) - \overline{S(x_i)}}{\overline{S(x_i)}} > \tau \\ 0, & \text{các trường hợp còn lại} \end{cases} \tag{2.1}$$

trong đó

- x_i là một điểm thuộc chuỗi x
- $S(x_i)$ là điểm tương ứng với x_i trong chuỗi saliency map
- $\overline{S(x_i)}$ là các giá trị trung bình các điểm z trước $S(x_i)$

2.2 Dự báo chuỗi thời gian bằng NeuralProphet

Được phát triển dựa trên kế thừa kết quả của Facebook Prophet, Neural Prophet được cải tiến cùng với những phương pháp học sâu để tăng độ chính xác cho việc dự báo.

Về cơ bản, NeuralProphet vẫn giữ nguyên nguyên lý thiết kế của FacebookProphet, đều cho các mô hình thành phần cơ bản giống nhau nhưng NP được kết hợp với thư viện Pytorch để cho ra kết quả tốt hơn [5]. Sau đây ta đi vào tìm hiểu từng phần của mô hình dự báo NeuralProphet

2.2.1 Thành phần mô hình

Với mỗi một bộ dữ liệu ta cần phân tích chúng thành các thành phần để hiểu rõ hơn về đối tượng mà ta đang nghiên cứu biến động ra sao theo thời gian, xu hướng tăng giảm,... sẽ hỗ trợ rất tốt trong việc đưa ra dự báo.

Điểm mạnh của mô hình NeuralProphet là khả năng kết hợp các mô đun thành phần. Mô hình bao gồm nhiều mô đun thành phần và mỗi mô đun đó đều là một thành phần cho việc dự báo. Các mô đun thành phần đều có thể thay đổi tỷ lệ theo xu hướng để tạo ra hiệu ứng nhân. Mỗi một mô đun này đều có đầu vào và quy trình mô hình hóa riêng. Tuy nhiên, tất cả các các mô đun đều cho đầu ra h, trong đó h xác định số bước được dự báo trong tương lai. Những giá trị này được cộng lại dưới dạng các giá trị dự đoán $\hat{y}_t, \ldots, \hat{y}_{t+h-1}$ứng với các giá trị $y_t, \ldots, y_{t+h-1}$. Ta có thành phần dự báo một bước h=1:

$$\hat{y}_t = T(t) + S(t) + E(t) + F(t) + A(t) + L(t) \tag{2.2}$$

trong đó

- $T(t)$: là xu hướng tại thời điểm t
- $S(t)$: Hiệu ứng theo mùa tại thời điểm t
- $E(t)$: Hiệu ứng ngày lễ và sự kiện tại thời điểm t
- $F(t)$: Những tác động hồi quy tại thời điểm t đối với các biến ngoại sinh được biết đến trong tương lai
- $A(t)$:Hiệu ứng tự động hồi quy tại thời điểm t dựa trên các quan sát trong quá khứ

- $L(t)$: Hiệu ứng hồi quy tại thời điểm t đối với các quan sát trễ của các biến ngoại sinh

Sau đây ta sẽ thảo luận chi tiết hơn về từng thành phần này

Thành phần xu hướng

Chúng ta có thể tổng quát hóa xu hướng bằng cách xác định tốc độ tăng trưởng phụ thuộc thời gian $\delta(t)$ và độ lệch phụ thuộc thời gian $\rho(t)$.

$$T(t) = \delta(t).t + \rho(t)$$

Xu hướng tuyến tính của từng đoạn chỉ thay đổi tốc độ tăng trưởng tại một số điểm hữu hạn. Ta gọi tập C gồm n_c điểm thay đổi tại các thời điểm khác nhau có $C = (c_1, c_2, \ldots, c_{n_c})$. Giữa các điểm thay đổi, tốc độ tăng trưởng xu hướng được giữ không đổi. Tốc độ tăng trưởng và điều chỉnh bù của những đoạn dữ liệu đầu tiên được định nghĩa tương ứng là δ_0 và ρ_0. Điều chỉnh tốc độ tại mỗi điểm thay đổi có thể được định nghĩa là một vectơ $\delta \in \mathbb{R}^{n_c}$, trong đó δ_j là thay đổi tốc độ tại điểm thay đổi thứ j. Tốc độ tăng trưởng tại thời điểm t được xác định bằng cách cộng tốc độ tăng trưởng ban đầu δ_0 với tổng các điều chỉnh tốc độ tại tất cả các điểm thay đổi cho đến bước thời gian t. Mỗi lần thay đổi tốc độ tăng trưởng δ_j là một tham số được đưa vào dữ liệu. Một vectơ điều chỉnh bù tương ứng có thể được định nghĩa là ρ thuộc $\mathbb{R}^{n_c}$. Tương tự, độ lệch tại thời điểm t được cho bởi độ lệch ban đầu ρ_0 và tổng các điều chỉnh độ lệch tại mỗi điểm thay đổi cho đến thời điểm t. Khác so với tốc độ tăng trưởng, độ bù cho các điểm thay đổi được tính bằng $\rho_j = -c_j.\delta_j$. Ta định nghĩa một vectơ nhị phân $\Gamma(t) \in \mathbb{R}^{n_c}$ biểu thị với thời gian t có qua mỗi điểm thay đổi hay không. Từ đó ta có vector xu hướng $T(t)$ tại thời điểm t:

$$T(t) = [\delta_0 + \Gamma(t)^T \delta].t + [\rho_0 + \Gamma(t)^T \rho] \tag{2.3}$$

trong đó:

$$\delta = (\delta_0, \delta_1, \ldots, \delta_{n_C})$$
$$\rho = (\rho, \rho_1, \ldots, \rho_{n_C})$$
$$\Gamma(t) = (\Gamma_1(t), \Gamma_2(t), \ldots, \Gamma_{n_c}(t))$$
$$\Gamma_j(t) = \begin{cases} 1, & \text{nếu } t \geq c_j \\ 0, & \text{với các trường hợp khác} \end{cases}$$

Thành phần mùa

Tính mùa trong NeuralProphet được định nghĩa bằng chuỗi Fourier. Các thuật ngữ Fourier được định nghĩa là các cặp sin, cos và cho phép lập mô hình nhiều mùa cũng như các mùa có chu kỳ không phải số nguyên, chẳng hạn như thời vụ hàng năm với dữ liệu hàng ngày ($p = 365.25$) hoặc với dữ liệu hàng tuần ($p = 52.18$). Trong một dữ liệu có nhiều tính mùa vụ, các giá trị khác nhau của n có thể được xác định cho mỗi chu kỳ:

$$S_p(t) = \sum_{j=1}^{k} (a_j . \cos(\frac{2\pi jt}{p}) + b_j . \sin(\frac{2\pi jt}{p})) \tag{2.4}$$

Đối với bước thời gian t, ảnh hưởng từ tất cả các mùa được xem xét trong mô hình có thể được biểu thị bằng $S(t)$ trong công thức dưới đây, trong đó $\mathbb{P}$ là tập hợp tất cả các chu kỳ

$$S(t) = \sum_{p \in \mathbb{P}} S_p^*(t) \tag{2.5}$$

Cả hai mẫu theo mùa cộng thêm và số nhân đều được hỗ trợ. Mỗi chu kỳ theo mùa S_p^* có thể được định cấu hình riêng lẻ để trở thành đa nhân, trong trường hợp đó thành phần mùa được nhân với xu hướng.

$$S_p^*(t) = \begin{cases} S_p^\dagger = T(t).S_p(t) & \text{nếu } S_P \text{ có tính nhân} \\ S_p^*(t) & \text{với các trường hợp còn lại} \end{cases}$$

Auto-Regression

Mô đun AR trong NP được cải tiến từ mô hình AR-net [3]. AR-Net có thể tính toán được tất cả các dự báo h với một mô hình, có thể tuyến tính hoặc phi tuyến tính. Trong bất kỳ cấu hình nào, p lần quan sát cuối cùng của biến mục tiêu $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, còn được gọi là độ trễ, là các đầu vào cho mô-đun. Kết quả đầu ra là các giá trị h tương ứng với hiệu ứng AR cho mỗi bước dự báo $A^t(t), A^t(t+1), \ldots, A^t(t+h-1)$.

$$A^t(t), A^t(t+1), \ldots, A^t(t+h-1) = AR - Net(y_{t-1}, y_{t-2}, \ldots, y_{t-p}) \quad (2.6)$$

Điều quan trọng cần lưu ý, với mỗi lần dự báo tại một điểm gốc cụ thể, ta thu được h dự đoán. Do đó, tại một thời điểm nhất định, ta có tối đa h dự đoán khác nhau, mỗi dự đoán bắt nguồn từ một dự báo khác nhau được thực hiện trong quá khứ. Chúng khác nhau dựa trên dữ liệu có sẵn cho mô hình tại thời điểm dự báo.

AR order Tham số quan trọng nhất đối với mô-đun này là số lượng các giá trị trong quá khứ được hồi quy, còn được gọi là bậc p của mô hình $AR(p)$. Tham số này nên được chọn dựa trên độ dài gần đúng của ngữ cảnh có liên quan trong các quan sát trước đây.

Linear AR Cấu hình AR-Net mặc định không chứa các lớp ẩn. Trong thực tế, nó là một NN lớp đơn với đầu vào p, đầu ra h, không có biases và không có chức năng kích hoạt. Trọng số của mỗi lớp đơn lẻ hồi quy một độ trễ cụ thể vào một bước dự báo cụ thể. Do đó, mỗi trọng số có thể được so khớp với một hệ số tương ứng của tập hợp h các mô hình $AR(p)$ cổ điển, làm cho việc diễn giải mô hình trở nên đơn giản.

$$y = Wx$$

với

$$x = (y_{t-1}, y_{t-2}, \ldots, y_{t-p})$$
$$y = (A^t(t), A^t(t+1), \ldots, A^t(t+h-1))$$

Deep AR Mô-đun đào tạo một NN được kết nối đầy đủ với số lượng lớp và kích thước ẩn được chỉ định. Ta có p quan sát cuối cùng của chuỗi thời gian là đầu vào của lớp đầu tiên. Sau mỗi lớp ẩn, các bản ghi được chuyển qua một chức năng kích hoạt. Lớp cuối cùng xuất ra h logits, không bị biến đổi bởi một hàm kích hoạt và không có bias. Đối với l lớp ẩn có kích thước lớp ẩn là d, ta có:

$$\begin{aligned} a_1 &= f_a(W_1 x + b_1) \\ a_i &= f_a(W_i a_{i-1} + b_i) \text{với } i \in [2, \ldots, l] \\ y &= W_{l+1} a_l \end{aligned}$$

trong đó $f_a(x)$ là một hàm phi tuyến, có thể là Sigmoid, ReLU, $max(0, z), \ldots$

Theo đó, độ lệch của lớp đều có cùng kích thước $b \in \mathbb{R}^d$, trong khi trọng số của lớp là $W \in \mathbb{R}^{d \times p}$, ngoại trừ trọng lượng lớp(layer weights) $W_1 \in \mathbb{R}^{d \times p}$ đầu tiên và $W_{l+1} \in \mathbb{R}^{h \times d}$ cuối cùng.

Sparse AR AR-Net đã chứng minh rằng thứ tự chính xác có thể được tính gần đúng bằng cách đặt thứ tự thành một giá trị lớn hơn một chút so với giá trị mong đợi khi sự chính quy hóa được sử dụng để chia nhỏ trọng số mô hình. Chức năng chính quy ban đầu được đề xuất trong AR-Net được đưa ra :

$$\Lambda_{AR-Net}(\theta, c_1, c_2) = \frac{1}{p} \sum_{i=1}^{p} 2.(1 + \exp(-c_1.|\theta|_i^{\frac{1}{c_2}})^{-1} - 1) \tag{2.7}$$

với được đề xuất từ tác giả giá trị $c_1 \approx 3, c_2 \approx 3$

Lagged Regressors

Hồi quy độ trễ được sử dụng để tương quan các biến quan sát khác với chuỗi thời gian mục tiêu. Chúng thường được gọi là đồng biến. Không giống như các bộ hồi quy trong tương lai, ta chưa biết được tương lai của các bộ hồi quy có độ trễ.

Tại thời điểm dự báo t, ta chỉ có quyền truy cập vào các giá trị đã quan sát, trong quá khứ của chúng cho đến thời điểm $t-1$.

$$L(t) = \sum_{x \in \mathbb{X}} L_x(x_{t-1}, x_{t-2}, \ldots, x_{t-p}) \tag{2.8}$$

Cho một tập các đồng biến $\mathbb{X} \in \mathbb{R}^{T \times n_l}$, ta có một mô-đun hồi quy có độ trễ riêng biệt cho mỗi m trong số m đồng biến biến của chiều dài T. Điều này cho phép xác định riêng ảnh hưởng của từng hiệp biến đối với các dự đoán. Mỗi mô-đun hồi quy có độ trễ về mặt chức năng giống hệt với mô-đun AR, với sự khác biệt duy nhất là đầu vào. Ở đây, p quan sát cuối cùng của đồng biến biến x là đầu vào cho mô-đun. Các đầu ra có dạng giống hệt nhau, mỗi mô-đun tạo ra h các thành phần:

$$L_x^t(t), L_x^t(t-1), \ldots, L_x^t(t+h-1) = AR - Net(x_{t-1}, x_{t-1}, \ldots, x_{t-p}) \tag{2.9}$$

Future Regressors

Để lập mô hình cho các hàm hồi quy đặc trưng, cả giá trị trong quá khứ và tương lai của các hàm hồi quy này phải được biết đến. Với tập hợp các bộ hồi quy đặc trưng là $\mathbb{F} \in \mathbb{R}^{T \times n_f}$, trong đó n_f là số bộ hồi quy, ảnh hưởng từ tất cả các bộ hồi quy đặc trưng tại thời điểm bước t có thể được ký hiệu là $F(t)$ như trong công thức (2.10) , trong đó d_f là hệ số của mô hình cho bộ hồi quy đặc trưng. Theo mặc định, các bộ hồi quy đặc trưng có hiệu ứng cộng, có thể được định cấu hình thành phép nhân thay thế.

$$F(t) = \sum_{f \in \mathbb{F}} F_f^*(t) \tag{2.10}$$

trong đó:

$$F_f^t(t) = d_f.f(t)$$

$$F_f^*(t) = \begin{cases} F_f^\dagger(t) = T(t).F_f(t) & \text{nếu } f \text{ có tính nhân} \\ F_f(t), & \text{các trường hợp còn lại} \end{cases}$$

Sự kiện và ngày lễ

Với dữ liệu thực tế, luôn bị biến động vào những ngày lễ hoặc sự kiện. Tuy nhiên ảnh hưởng từ các sự kiện đặc biệt hoặc ngày lễ là việc xảy ra không thường xuyên. Các sự kiện như vậy được mô hình hóa tương tự với các bộ hồi quy đặc trưng, với mỗi sự kiện e là một biến nhị phân, $e \in [0; 1]$, báo hiệu sự kiện có xảy ra hay không. Đối với một tập hợp các sự kiện $\mathbb{E} \in \mathbb{R}^{T \times n_e}$ với n_e là số lượng sự kiện và độ dài của chuỗi là T, ảnh hưởng từ tất cả các sự kiện tại bước thời gian t có thể được ký hiệu là $E(t)$ trong công thức (2.11) , trong đó z_e biểu thị hệ số của mô hình tương ứng với sự kiện $e \in \mathbb{E}$

$$E(t) = \sum_{e \in \mathbb{E}} E_e^*(t) \tag{2.11}$$

trong đó

$$E_e(t) = z_e e(t)$$

$$E_e^*(t) = \begin{cases} E_e^\dagger(t) = T(t).E_e(t), & \text{với } e \text{ có tính nhân} \\ E_e(t) & \text{với trường hợp còn lại} \end{cases}$$

Tại Việt Nam, với những ngày nghỉ lễ, sự kiện trùng cuối tuần thì sẽ được nghỉ bù dẫn đến có những ngày nghỉ lễ đặc trưng riêng của từng năm. Khi đó, đối với một sự kiện đã cho tại thời điểm t_e, một cửa sổ $[t_e - i, t_e + j]$ của $i + j$ ngày có thể được cấu hình để được coi là sự kiện đặc biệt của riêng chúng. Bằng cách này, một biến mới được tạo cho mỗi ngày trong cửa sổ xung quanh sự kiện và được thêm vào tập hợp các sự kiện E.

2.2.2 Quá trình Training

Có sự cải tiến so với tiền thân của NP là Prophet, NP cho phép ta điều chỉnh các tham số của mô hình với dữ liệu mà ta đang xét. Được tích hợp thêm PyTorch làm cho mô hình linh hoạt và dễ sử dụng. NP được sử dụng phương pháp stochastic gradient descent(SGD). Đây là phương pháp phù hợp với hầu hết các mô hình deep learning. Về mô hình học sâu ta đi chi tiết các thành phần để xây dựng mô hình

Hàm mất mát

Hàm mất mát mặc định của thư viện NP là hàm Hubber, còn được gọi là hàm $L1-loss$ trơn, được thể hiện qua công thức (2.12) . Quan sát công thức, ta thấy hàm mất mát sẽ tương tự với sai số bình phương trung bình (MSE) nếu $|y - \hat{y}| < \beta$ và hàm mất mát sẽ giống sai số tuyệt đối trung bình (MAE) trong trường hợp còn lại. Giá trị β được mặc định bằng 1. Người dùng có thể điều chỉnh được tham số này để phù hợp với các bộ dữ liệu riêng.

$$L_{hubber(y,\hat{y})} = \begin{cases} \frac{1}{2\beta}(y-\hat{y})^2, & \text{với} |y-\hat{y}| < \beta \\ |y-\hat{y}| - \frac{\beta}{2}, & \text{với trường hợp còn lại} \end{cases} \tag{2.12}$$

Regularization

NP sử dụng hàm Regularization với trọng số mô hình θ được tham số hóa :

$$\Lambda(\theta, \epsilon, \alpha) = \frac{1}{n}\sum_{i=1}^{n} \log(\frac{1}{\epsilon.e} + \alpha.|\theta_i|) + \log(\epsilon) + 1 \tag{2.13}$$

với $\epsilon \in (0, \infty)$ và $\alpha \in (0, \infty)$
Thay đổi giá trị của ϵ và α theo độ lớn của trọng số θ. NP đề xuất người dùng với giá trị $\epsilon = 1$ và $\alpha = 1$

$$\Lambda(\theta, \epsilon = 1, \alpha = 1) = \frac{1}{n}\sum_{i=1}^{n} \log(\frac{1}{e} + |\theta_i|) + 1 \tag{2.14}$$

Optimizer

NP đề xuất người dùng sử dụng thuật toán tối ưu AdamW [6]. Thuật toán được tác giá nghiên cứu cho kết quả phù hợp với nhiều bộ dữ liệu. Giá trị khởi tạo của AdamW với learning rate (được giới thiệu phần tiếp theo), $\beta = (0.9, 0.999), \epsilon = 1e{-}8$ và độ lớn phân rã là $1e{-}4$

Còn sự đề xuất của các nhà nghiên cứu sử dụng SGD thay thế AdamW với $\beta = 0.9$ và độ lớn phân rã là $1e{-}4$. Kết quả cho thấy SGD tốt hơn trong nhiều trường hợp so với AdamW, vẫn có 1 vài trường hợp phân kì. Tuy nhiên khi sử dụng SGD cần tinh chỉnh các hyperparameter, yêu cầu người dùng có kinh nghiệm nhất định.

Learning rate

Kiểm tra phạm vi tốc độ học tập được thực hiện cho $100 + \log_{10}(10 + T) * 50$ lần lặp, bắt đầu từ $\eta = 1e - 7$, kết thúc ở $\eta = 1e + 2$ với kích thước batch-size đã được cấu hình [7]. Learning rate cao nhất chính là kết quả learning rate cần tìm . Độ dốc learning rate cao nhất được tìm thấy bằng cách chọn learning rate tại điểm lớn nhất tại giá trị âm gradient của hàm mất mát (maximizes the negative gradient of the losses.)

Để có được giá trị tốt nhất, ta thực hiện kiểm tra ba lần và lấy trung bình $\log_{10}$ của ba giá trị η_1, η_2, η_3 được learing rate cần tìm là η:

$$\eta^\star = \frac{1}{3}(\log_{10}(\eta_1) + \log_{10}(\eta_2) + \log_{10}(\eta_3))$$
$$\eta = 10^{\eta^\star}$$

Batch Size

Batch size B là một tham số tùy chọn. Nếu không phải do người dùng chỉ định, phương pháp phỏng đoán sau sẽ xác định batch size dựa trên độ dài T của tập dữ liệu:

$$B^\star = 2^{2+\log_{10} T}$$
$$B = \min(T, \max(16, \min(256, B^*)))$$

Training Epochs

Tương tự như batch size, tham số N_{epoch} cũng có thể tùy chọn của người dùng. Nếu người dùng không thiết lập ban đầu, N_{epoch} được xác định:

$$N^{\star}_{epoch} = \frac{1000.2^{\frac{5}{2}.\log_{10}(T)}}{T}$$
$$N_{epoch} = \min(500, \max(500, N^{\star}_{epoch}))$$

Scheduler

Vì tất cả các hyperparameter liên quan đến mô hình đều được tính gần đúng tự động, khi đó ta không biết đâu là bộ hệ số tối ưu và do đó có thể dẫn đến các vấn đề khi train các mô hình. Để khắc phục vấn đề này, NP có chính sách '1cycle' [8], cho phép đào tạo ra mô hình "siêu hội tụ". Theo đó, tỷ lệ học ban đầu $\frac{\eta}{100}$ được tăng dần lên đến tỷ lệ học tập cao nhất η, đạt 30% quá trình đào tạo. Sau đó, learning rate được chạy dọc theo đường cong cosin xuống $\frac{\eta}{5000}$ khi kết thúc quá trình đào tạo.

2.2.3 Các phương pháp đánh giá

Trong đồ án này tôi sử dụng các phương pháp đánh giá được gọi là sai số phần trăm tuyệt đối trung bình đối xứng (sMAPE) và sai số tuyệt đối trung bình (MAE)

sMAPE

Ta thường thấy phương pháp đánh giá sai số tương đối:

$$s(y_i, \hat{y}_i) = \frac{|y_i - \hat{y}_i|}{|y_i|}$$

trong đó y_i là giá trị thực tế, $\hat{y}_i$ là giá trị dự đoán
Phương pháp này sẽ gặp vấn đề khi giá trị của $x = 0$ dẫn đến kết quả bị sai. Khắc phục vấn đề đó, phương pháp sai số phần trăm tuyệt đối trung bình đối xứng được nghiên cứu.

sMAPE có thể thể hiện qua công thức:

$$s(y_i, \hat{y}_i) = \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{2.15}$$

MAPE

Sai số tương đối mà một dự báo mắc phải có thể được đo lường bằng phần trăm sai số tuyệt đối trung bình (MAPE). MAPE được tính theo công thức sau:

$$MAPE = \frac{100}{T} \sum_{i=1}^{T} \frac{|y_i - \hat{y}_i|}{y_i} \tag{2.16}$$

với T là độ dài của dữ liệu

2.3 Tóm tắt chương

Chương 2 trình này cho ta thấy được lý thuyết của phương pháp điểm bất thường SR, mô hình NP, các thành phần mô hình, cách chọn các tham số của mô hình học sâu và các phương pháp đánh giá. Đây là những phương pháp trình nằm trong quy trình đề xuất mà đồ án sẽ trình bay trong phần tiếp theo.

Chương 3

Xây dựng mô hình dự đoán cho bài toán dự báo phân bổ tài nguyên

Trên cở sở lý thuyết đã nêu ở chương trước,ta xây dựng mô hình dự báo

3.1 Mô tả bài toán dự báo tài nguyên trong mạng WiFi marketing

Bài toán chi tiết đã nêu ở chương 1, ta có thể mô tả ngắn gọn:

Đầu vào:

Tập dữ liệu D có danh sách N địa điểm, mỗi địa điểm có một mảng số nguyên v độ dài l_i là mảng chứa số lượt truy cập theo từng ngày của địa điểm đó. Thêm vào đó là danh sách các ngày lễ hoặc ngày xảy ra sự kiện tương ứng với địa điểm đó. (Ví dụ: danh sách các ngày nghỉ lễ, các ngày diễn ra sự kiện giảm giá hoặc sự kiện âm nhạc ở địa điểm đó)

Đầu ra

Một tập F: gồm N mảng số nguyên độ dài k, mỗi mảng chứa số lượt truy cập được dự báo trong tương lai của N địa điểm đầu vào.

3.2 Quy trình đề xuất

3.2.1 Tiền xử lý dữ liệu

Xử lý dữ liệu trước khi huấn luyện mô hình là một việc vô cùng quan trọng trong bất kì một bài toán học máy nào. Tiền xử lý giúp việc huấn luyện tăng độ chính xác, cải thiện thời gian huấn luyện,...

Dữ liệu bị thiếu

Trong thời kì covid vừa qua việc thu thập dữ liệu gặp nhiều khó khăn. Điều đó dẫn đến tình trạng thiết bị thực tế có thể bị hỏng, dẫn đến không thể thu thập dữ liệu. Hoặc trường hợp lỗi hệ thống cũng sẽ gặp tình trạng tương tự. Vậy việc ta cần làm xử lý dữ liệu đó.
Với dữ liệu bị thiếu vào các ngày lễ, sự kiện: các sự kiện được coi là không xảy ra. Sự kiện bị mất được điền bằng các số không, cho biết sự vắng mặt của chúng.
Với dữ liệu là ngày bình thường, ta có 3 bước xử lý:

1. Các giá trị bị thiếu được xấp xỉ bằng phương pháp nội suy tuyến tính 2 chiều. Ta sẽ thực hiện xấp xỉ với tối đa 5 giá trị bị thiếu tại mỗi chiều. Nếu thiếu hơn 10 giá trị, các giá trị đó vẫn là NaN

2. Các giá trị còn thiếu còn lại được tính bằng giá trị trung bình xoay ở giữa(centred rolling average.). Giá trị trung bình được tính trong khoảng thời gian tối đa là 30 và điền vào nhiều nhất 20 giá trị bị thiếu liên tiếp.

3. Nếu có hơn 30 giá trị bị thiếu liên tiếp, thuật toán nhập sẽ hủy bỏ và thay vào đó sẽ loại bỏ tất cả các điểm dữ liệu bị thiếu.

Chuẩn hóa dữ liệu

Với NP, ta có thể tùy chọn nhiều cách chuẩn hóa như trong bảng . Khi người dùng không tùy chọn, thư viện sẽ trả về giá trị auto

Tên	Quy trình chuẩn hóa
'auto'	'minmax' nếu dữ liệu là nhị phân, còn lại là 'soft'
'off'	Bỏ qua việc chuẩn hóa dữ liệu
'minmax'	Quy chuẩn giá trị nhỏ nhất =0, lớn nhất bằng 1
'standardize'	Tính toán độ lệch trung bình sau đó chia cho độ lệch chuẩn
'soft'	Quy chuẩn giá trị nhỏ nhất là 0, giá trị quantile thứ 95 là 1.0
'soft1'	Quy chuẩn giá trị nhỏ nhất là 0.1, giá trị quantile thứ 90 là 0.9

Bảng 3.1: Các tùy chọn chuẩn hóa dữ liệu có sẵn trong NP

3.2.2 Xử lý dữ liệu

Dữ liệu sau khi tiền xử lý với những trường hợp ở trên, ta sử dụng thuật toán SR để phát hiện ra các điểm bất thường. Các điểm này thường là những ngày có giá trị cao hơn hoặc thấp hơn nhiều so với những điểm lân cận của nó. Sau khi đã có được những điểm bất thường, ta xử lý chúng bằng cách cho chúng bằng trung bình cộng của số lượt truy cập hai ngày cách đó trước và sau 7 ngày. Ta có thể quan sát qua công thức:

$$v_i = \frac{v_{i-7} + v_{i+7}}{2}$$

trong đó

- v_i là số lượt truy cập ngày thứ i được thuật toán SR phát hiện
- v_{i+7}, v_{i-7} là số lượt truy cập ngày thứ $i-7$ và $i+7$ trong dư liệu là chuỗi thời gian

3.2.3 Dự báo bằng NeuralProphet

Sau khi đã xử lý dữ liệu với phương pháp SR, ta cho dữ liệu đó vào mô hình NeuralProphet đã được tạo để huấn luyện mô hình dự báo. Với số ngày dự báo là 30, ta sẽ chia tập dữ liệu ra làm 2 phần train và test, độ dài tập test bằng 30. Có điểm cần lưu ý đầu vào của NeuralProphet là dữ liệu gồm 2 cột, cột thứ nhất là ngày, cần đổi tên cột thành 'ds', cột thứ 2 là số lượt truy cập sẽ đổi tên thành 'y'. Cùng với dữ liệu sau khi xử lý, ta sẽ tạo 1 dataframe chứa những ngày lễ, sự kiện. Dataframe này gồm 2 cột: cột thứ nhất là tên của những sự kiện đó, cột thứ 2 là ngày diễn ra dự kiện đó. NeuralProphet cần tham số về ngày lễ để tính toán độ ảnh hưởng của chúng tới việc dự báo.

3.2.4 Đánh giá kết quả

Sau khi đã lượt dự báo của 30 ngày tiếp theo, ta tính toán sai số bằng 2 phương pháp đánh giá đã nêu ở phần lý thuyết : phương pháp MAE và phương pháp sMAPE.

3.3 Tóm tắt chương

Chương này đề xuất quy trình cho bài toán dự báo phân bổ tài nguyên trong mạng WiFi marketing. Quy trình bao gồm các bước chính: xử lý dữ liệu với các điểm bất thường, dự báo và đánh giá kết quả sau khi dự báo. Quy trình được đề xuất sử sụng phương pháp SR, mô hình NeuralProphet và phương pháp đánh giá sMAPE, MAPE

Chương tiếp theo, đồ án trình bày về phần thực nghiệm với bộ dữ liệu thực tế cùng đó đannhs giá kết quả và nhận xét.

Chương 4

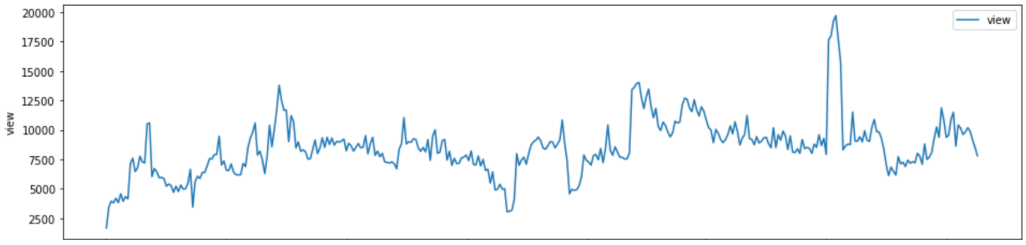
Kết quả thực nghiệm

Trong chương này sẽ trình bày thực nghiệm theo quy trình đề xuất ở chương 3. Sau khi có kết quả, ta so sánh với phương pháp đang được sử dụng tại công ty và một vài phương pháp phổ biến khác cũng đó sẽ đưa ra nhưng định hướng tiếp theo

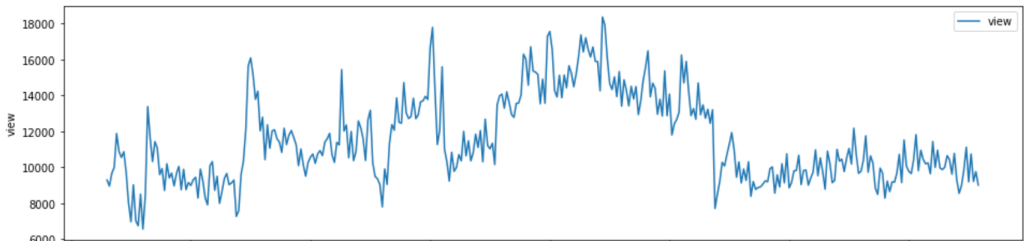
4.1 Dữ liệu thực nghiệm

Đồ án tiến hành thực nghiệm để đưa ra dự báo lượt người truy cập tại 20 địa điểm thuộc các địa điểm có nhu cầu sử dụng khác nhau và các vị trí địa lý khác nhau cả 3 miền để tăng sự đa dạng về dữ liệu cho mô hình. 20 địa điểm gồm những thương hiệu: trung tâm thương mại Vincom, chuỗi cafe HighLand Coffee và sân bay của Gold Sun. Các dữ liệu được thu thập từ 01/01/2018 - 01/05/2022, được lưu dưới dạng file csv có dạng gồm 2 cột: cột 'date' và cột 'view'. Tổng cả bộ dữ liệu có 31620 điểm dữ liệu.

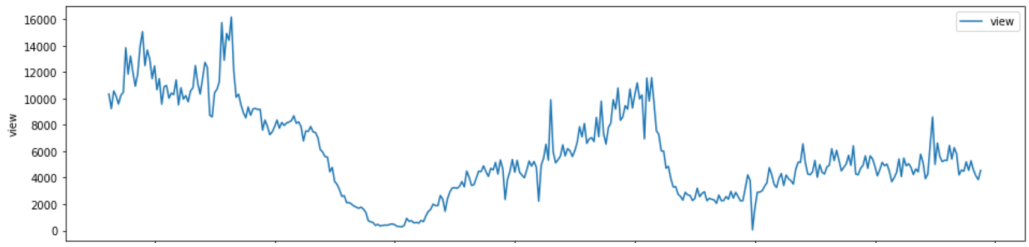
Ta có thể mô tả một



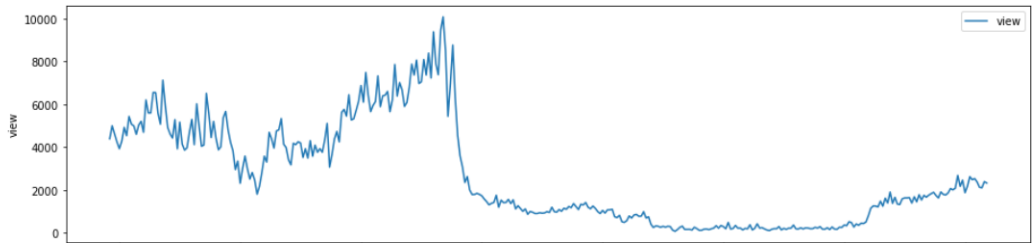
Hình 4.1: Dữ liệu thực tế năm 2018



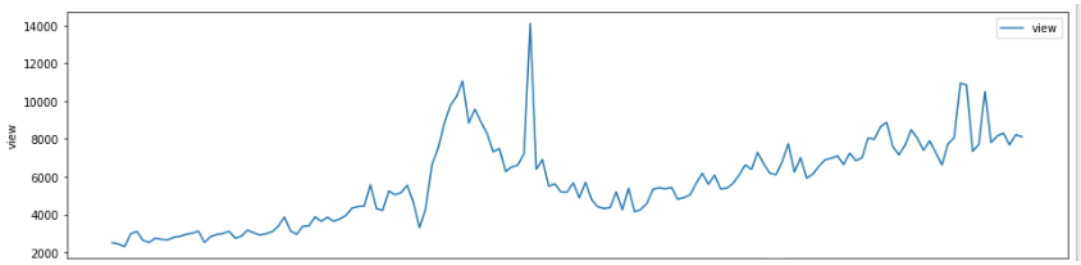
Hình 4.2: Dữ liệu thực tế năm 2019



Hình 4.3: Dữ liệu thực tế năm 2020



Hình 4.4: Dữ liệu thực tế năm 2021



Hình 4.5: Dữ liệu thực tế năm 2022

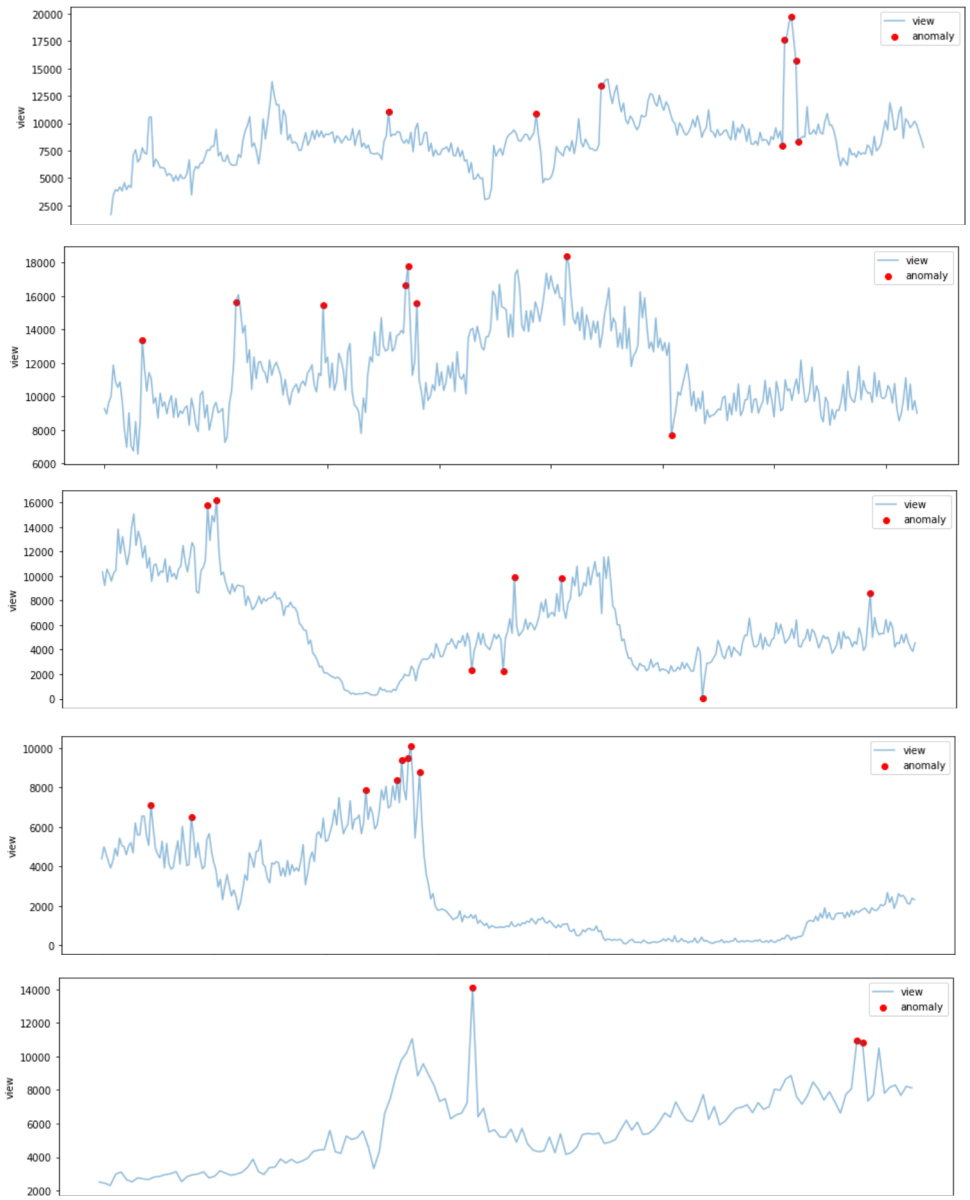
Hình trên được trực quan hóa dữ liệu tại một địa điểm. Ta phân chia chúng thành các năm để tiện cho việc quan sát

4.2 Thực nghiệm

Thực nghiệm bên dưới mô tả lại quá trình dự báo số lượt truy cập tại một điểm WiFi marketing. Khoảng thời gian dự báo là từ ngày 01/04/2022 - 01/05/2022

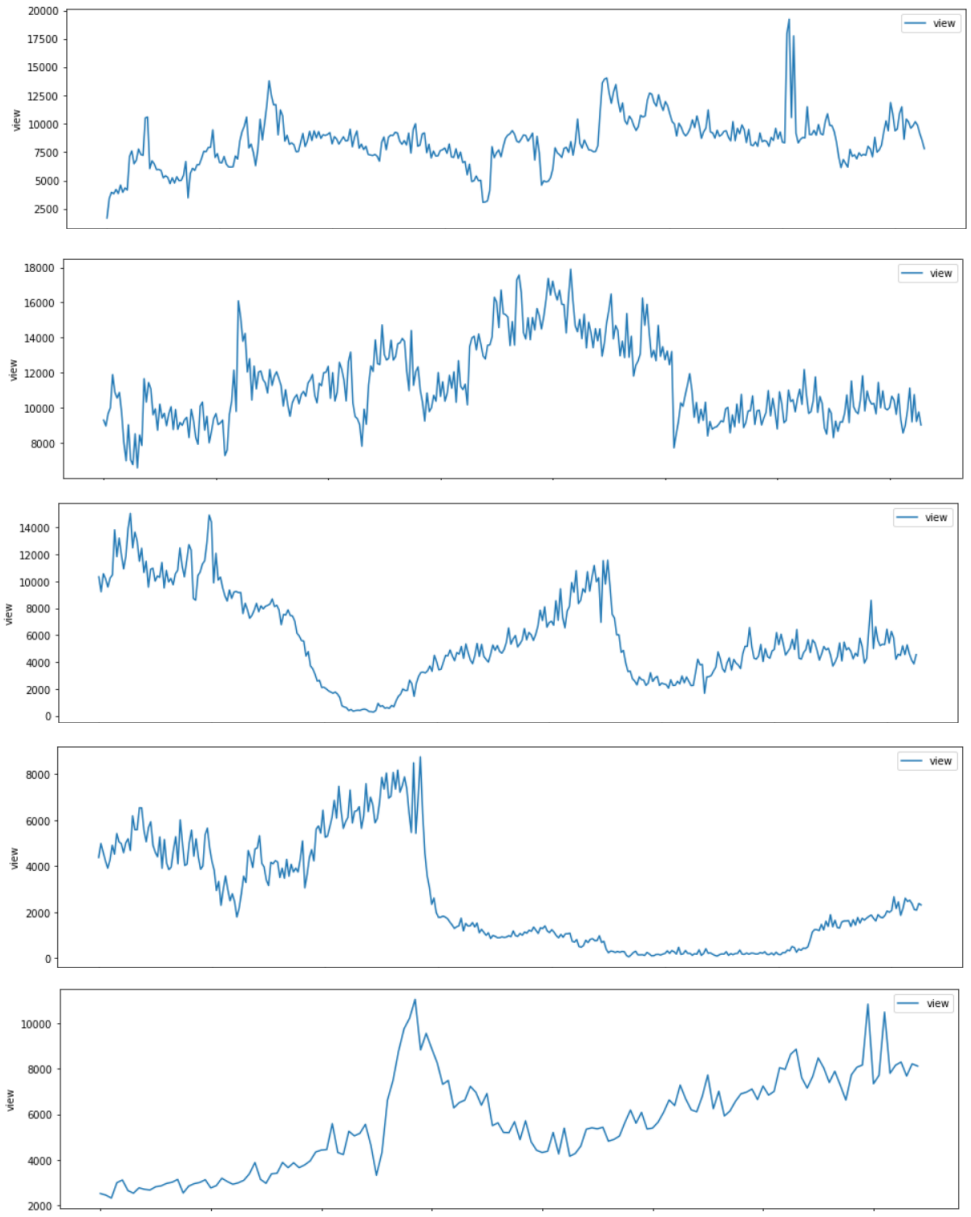
4.2.1 Xử lý dữ liệu

Như đã nêu ở phần lý thuyết, ta sẽ đưa dữ liệu vào thuật toán SR để phát hiện ra điểm bất thường sau đó xử lý chúng.



Hình 4.6: Phát hiện bất thường bằng phương pháp SR

Sau khi đã phát hiện điểm bất thường này, ta sẽ xử lý chúng theo cách đã nêu. Kết quả sau khi xử lý:



Hình 4.7: Dữ liệu sau khi xử lý

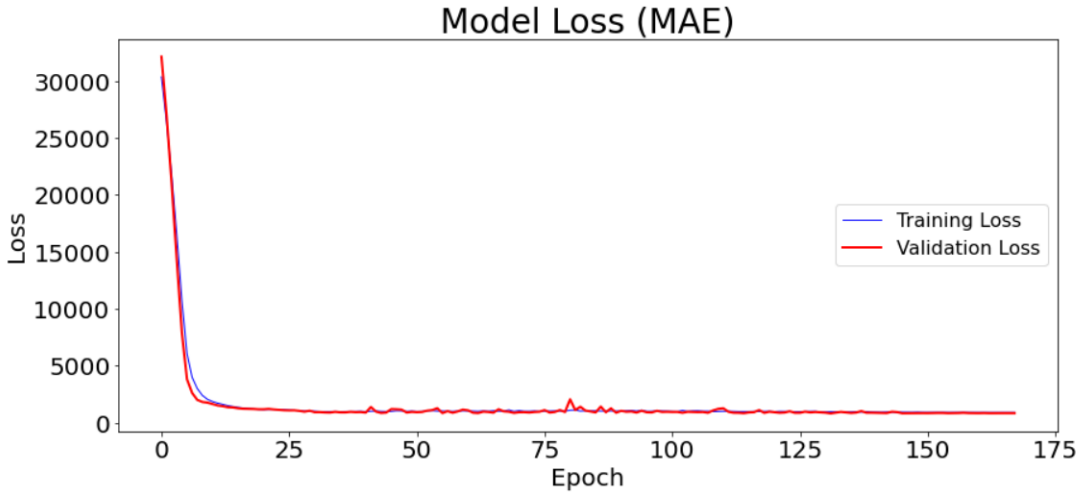
4.2.2 Dự báo bằng mô hình NeuralProphet

Sau khi dữ liệu đã được xử lý những điểm bất thường, ta đưa chúng vào model NeuralProphet. Mô tả quy trình chạy của mô hình:

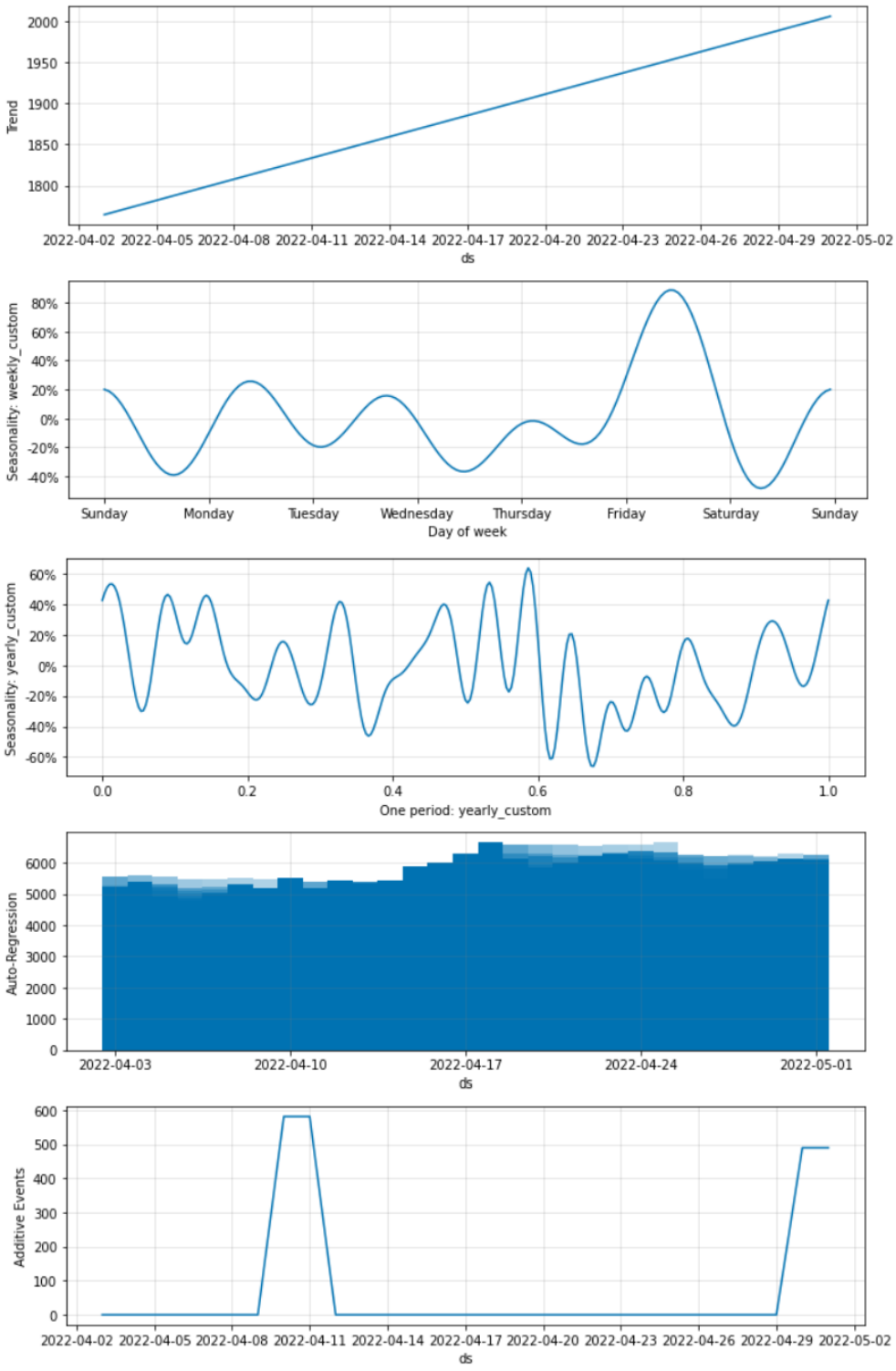
```
INFO - (NP.df_utils._infer_frequency) - Major frequency D corresponds to 99.939% of the data.
INFO - (NP.df_utils._infer_frequency) - Defined frequency is equal to major frequency - D
INFO - (NP.config.init_data_params) - Setting normalization to global as only one dataframe provided for training.
INFO - (NP.config.set_auto_batch_epoch) - Auto-set batch_size to 32
INFO - (NP.config.set_auto_batch_epoch) - Auto-set epochs to 168
92% 120/130 [00:00<00:00, 219.82it/s]
INFO - (NP.utils_torch.lr_range_test) - lr-range-test results: steep: 6.18E-02, min: 2.12E+00
92% 119/130 [00:00<00:00, 216.94it/s]
INFO - (NP.utils_torch.lr_range_test) - lr-range-test results: steep: 6.18E-02, min: 5.85E-01
INFO - (NP.forecaster._init_train_loader) - lr-range-test selected learning rate: 7.25E-02
Epoch[168/168]: 100%|██████████| 168/168 [00:20<00:00,  8.10it/s, SmoothL1Loss=0.00362, MAE=805, RMSE=1.16e+3, RegLoss=0]
```

Hình 4.8: Quá trình chạy của mô hình

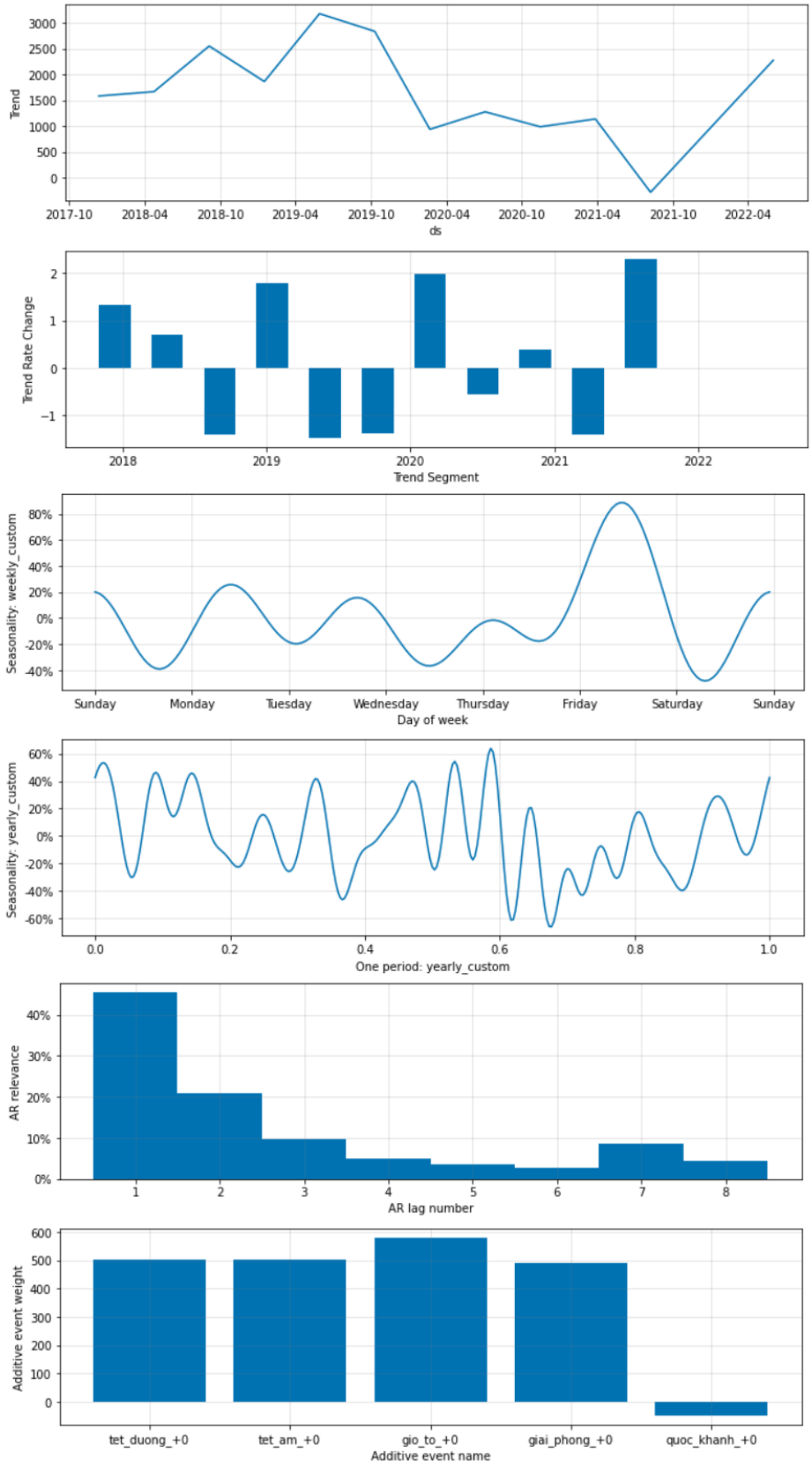
Quan sát trong hình 4.8, ta thấy mô hình tự động chọn các tham số như: learning rate, batch size, epochs, định dạng chuẩn hóa của dữ liệu. Việc tự động chọn này có thể chưa phải tốt nhất với bộ dữ liệu mà ta đang xét, tuy nhiên với mô hình mà ta xây dựng chạy cho rất nhiều bộ dữ liệu, việc ta định nghĩa sẵn các tham số kể trên sẽ gây ra những kết quả không tốt với một số địa điểm nhất định. Quan sát kết quả trong hình 4.10, ta thấy dự đoán xu hướng và tính mùa của địa điểm mà ta đang xét. Các thành phần này được phân rã theo lý thuyết đã trình bày ở chương 2.



Hình 4.9: Optimization Learning Curves

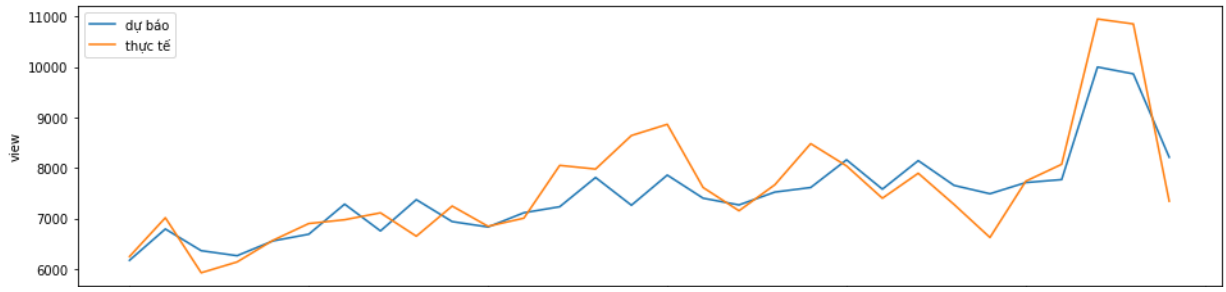


Hình 4.10: Dự đoán các thành phần của mô hình



Hình 4.11: Các tham số của mô hình

Sau khi tổng hợp lại các kết quả, ta thu được kết quả dự báo cuối cùng:



Hình 4.12: Kết quả dự báo cuối cùng

4.3 Đánh giá kết quả

Kết quả của quy trình đề xuất được so sánh với một số phương pháp phổ biến như AR-LSTM [11], Prophet [10].

- Prophet là một phương pháp dự báo chuỗi thời gian được Facebook giới thiệu. Quy trình này dựa trên mô hình cộng trong đó các xu hướng phi tuyến tính phù hợp với thời vụ hàng năm, hàng tuần và hàng ngày, cộng với các hiệu ứng ngày lễ. Nó hoạt động tốt nhất với chuỗi thời gian có hiệu ứng theo mùa và một số mùa dữ liệu lịch sử.
- Phương pháp AR-LSTM là phương pháp lai giữa AR(tự hồi quy) và mạng học sâu LSTM được phân rã các thành phần theo STL

Sử dụng phương pháp đánh giá sMAPE và MAPE để so sánh 3 phương pháp này. Giá trị sMAPE và MAPE được tính trung bình cho 20 địa điểm trong bộ dữ liệu. Bảng sau đây so sánh kết quả dự báo của tháng 5 năm 2022:

Phương pháp	NP	AR-LSTM	Prophet
MAPE	8.52%	12.34%	15%
sMAPE	4.26%	7.12%	9.48%

Bảng 4.1: So sánh giữa các phương pháp

Nhận xét:

Quan sát trong bảng 4.1, ta thấy kết quả dự báo bằng NP tốt hơn so với 2 phương pháp AR-LSTM và Prophet. Tuy chưa thể kết luận phương pháp NP tốt hơn hẳn so với AR-LSTM và Prophet do kết quả dự báo chỉ bao gồm 1 tháng nhưng bước đầu cho thấy phương pháp NP xử lý ngày lễ khả quan hơn.

4.4 Tóm tắt chương

Trong chương này, khóa luận đã trình bày thực nghiệm quy trình mà khóa luận đã đề xuất, thực hiện dự báo trên dữ liệu số lượt người truy cập vào hệ thống WiFi marketing 20 địa điểm tại trung tâm thương mại Vincom, chuỗi cafe HighLand Coffee và các sân bay thuộc Gold Sun. Kết quả được so sánh với một số phương pháp phổ biến là Prophet và AR-LSTM, cho thấy tính hợp lý của quy trình đề xuất.

Chương 5

Kết luận

5.1 Kết luận

Đồ án đã đạt được mục tiêu đề ra

" Tìm hiểu phương pháp để giải quyết bài toán dự báo phân bổ tài nguyên, đưa ra bài toán, lý thuyết phương pháp và đạt được những kết quả nhất định".

Kết quả của đồ án

Đồ án đã trình bày một phương pháp dự báo khá mới tại thời điểm đồ án này được viết và một ngành kinh doanh có phát triển mạnh trong thời gian sắp tới. Cụ thể:

1. Giới thiệu được bài toán dự báo phân bổ tài nguyên trong mạng WiFi marketing, định nghĩa, khái niệm về ngành marketing này.
2. Trình bày lý thuyết của phương pháp, tính mới cải tiến ơn so với bản tiền nhiệm của nó là Prophet
3. Chạy thử nghiệm với 20 bộ dữ liệu thực tế được viết bằng ngôn ngữ PYTHON.

Kỹ năng đạt được

1. Bước đầu biết tìm kiếm, đọc, dịch tài liệu chuyên ngành liên quan đến nội dung đồ án.
2. Biết tổng hợp các kiến thức đã học và kiến thức trong tài liệu tham khảo để viết báo cáo đồ án.
3. Chế bản đồ án bằng LaTeX, viết chương trình tính toán cho ví dụ minh họa bằng sử dụng ngôn ngữ PYTHON.
4. Biết tóm tắt nội dung đồ án và biết trình bày một báo cáo khoa học.

5.2 Hướng phát triển của đồ án trong tương lai

1. Tiếp tục tìm hiểu sâu hơn để cải thiện độ chính xác của mô hình.
2. Xây dựng quy trình phù hợp hơn với từng loại địa điểm.
3. Kết hợp với dữ liệu về thời tiết tại các tỉnh thành để dự báo có kết quả chính xác hơn.

Tài liệu tham khảo

[1] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, Qi Zhang *Time-series anomaly detection service at microsoft*, Microsoft, Bejing, China.

[2] Oskar Triebe , Hansika Hewamalage, Polina Pilyugina, Nikolay Laptev, Christoph Bergmeir, Ram Rajagopal (2021)*NeuralProphet: Explainable Forecasting at Scale*, arXiv.

[3] Triebe, O., Laptev, N., & Rajagopal, R. (2019). *Ar-net: A simple auto-regressive neural network for time-series*,arXiv.

[4] Girshick, R. (2015). *Fast r-cnn.*

[5] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). *Pytorch:An imperative style, high-performance deep learning library.* In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024-8035). Curran Associates, Inc

[6] Loshchilov, I., & Hutter, F. (2019) *Decoupled weight decay regularization.*

[7] Smith, L. N. (2017).*Cyclical learning rates for training neural networks*

[8] Smith, L. N., & Topin, N. (2018). *Super-convergence: Very fast training of neural networks using large learning* rates.

[9] Harvey, A. C., & Shephard, N. (1993). *Structural time series models. In Handbook of Statistics* (pp. 261-302).
Amsterdam: North Holland volume Vol. 11:Econometrics. ((edited by g.s. maddala, c.r. rao and h.d. vinod) ed.).

[10] Taylor, S. J., & Letham, B. (2017).*Forecasting at scale. PeerJ, .*

[11] Ta Anh Son, Nguyen Thi Thuy Linh & Nguyen Ngoc Dang (2021) *Solving Resource Forecasting in Wifi Networks by Hybrid AR-LSTM Model*, Springer

Phụ lục