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

ỨNG DỤNG CHUỖI THỜI GIAN VÀO DỰ ĐOÁN DOANH THU BÁN HÀNG

HỆ HỖ TRỢ QUYẾT ĐỊNH

Chuyên ngành: HỆ THỐNG THÔNG TIN QUẢN LÝ

Giảng viên:	TS. Lê Hải Hà
Sinh viên thực hiện:	Đặng Minh Anh
Lớp:	Hệ thống thông tin quản lý – K64
Mã số sinh viên:	20195947

HÀ NỘI – 2022

NHẬN XÉT CỦA GIẢNG VIÊN

1. Mục tiêu

(a)

(b)

(c)

2. Nội dung

(a)

(b)

(c)

3. Đánh giá kết quả đạt được

(a)

(b)

(c)

Hà Nội, ngày 01 tháng 08 năm 2022

Giảng viên

TS. Lê Hải Hà

Lời cảm ơn

Để hoàn thành báo cáo này, em xin gửi lời cảm ơn chân thành đến:

Ban Lãnh đạo trường Đại Học Bách khoa Hà Nội vì đã tạo điều kiện về cơ sở vật chất với hệ thống thư viện hiện đại, đa dạng các loại sách, tài liệu thuận lợi cho việc tìm kiếm, nghiên cứu thông tin.

Xin gửi lời cảm ơn chân thành đến Viện Toán ứng dụng và Tin học đã đưa thiết kế học phần Hệ hỗ trợ quyết định vào chương trình giảng dạy.

Xin cảm ơn giảng viên hướng dẫn - Thầy Lê Hải Hà đã giảng dạy tận tình, chi tiết để em có đủ kiến thức và vận dụng chúng vào bài báo cáo này.

Rất mong nhận được sự nhận xét, ý kiến đóng góp, phê bình từ phía Thầy để đồ án được hoàn thiện hơn.

Lời cuối cùng, em xin kính chúc thầy nhiều sức khỏe, thành công và hạnh phúc.

Hà Nội, tháng 08 năm 2022

Sinh viên

Đặng Minh Anh

Tóm tắt nội dung Báo cáo

1. Cơ sở lý thuyết

   Một số lý thuyết như Ra quyết định, Hệ hỗ trợ quyết định, Chuỗi thời gian.

2. Ứng dụng chuỗi thời gian vào dự báo doanh thu

   Chương này phân tích các yếu tố ảnh hưởng đến doanh thu để dự báo doanh thu trong tương lai, sử dụng Python trên Jupyter Notebook để thể hiện dữ liệu.

3. Kết luận

   Chương này đưa ra một số nhận định, kết luận sau khi phân tích.

Mục lục

Mở đầu **2**

Chương 1 Cơ sở lý thuyết **3**

1.1 Ra quyết định 3

1.1.1 Khái niệm 3

1.1.2 Vai trò 3

1.2 Hệ hỗ trợ quyết định 4

1.2.1 Khái niệm 4

1.2.2 Vai trò 5

1.3 Dự báo (Forecasting) 5

1.4 Chuỗi thời gian (Time Series) 5

Chương 2 Ứng dụng chuỗi thời gian vào dự báo doanh thu **7**

2.1 Giới thiệu bộ dữ liệu 7

2.2 Xử lý, quan sát dữ liệu 9

2.3 Dự báo doanh thu sử dụng chuỗi thời gian 9

Kết luận **12**

Tài liệu tham khảo **13**

Mở đầu

Lí do chọn đề tài

Dự báo không chỉ dành cho các nhà khí tượng học. Các chính phủ cần dự báo tăng trưởng kinh tế. Các nhà khoa học cố gắng dự đoán dân số trong tương lai. Và các doanh nghiệp dự báo nhu cầu sản phẩm – một nhiệm vụ chung của các nhà khoa học dữ liệu chuyên nghiệp. Dự báo đặc biệt liên quan đến các cửa hàng tạp hóa truyền thống, những cửa hàng này phải cân nhắc khéo léo với lượng hàng tồn kho cần mua. Dự đoán cao hơn với thực tế, các cửa hàng tạp hóa sẽ mắc kẹt với hàng hóa quá tải, dễ hư hỏng. Dự đoán thấp hơn thực tế, các mặt hàng phổ biến sẽ nhanh chóng bán hết, dẫn đến doanh thu bị giảm và khách hàng không hài lòng. Dự báo chính xác hơn, nhờ máy học, có thể giúp đảm bảo các nhà bán lẻ làm hài lòng khách hàng bằng cách có đủ sản phẩm phù hợp vào đúng thời điểm.

Các phương pháp dự báo chủ quan hiện tại cho bán lẻ có rất ít dữ liệu để sao lưu chúng và không có khả năng được tự động hóa. Vấn đề thậm chí còn trở nên phức tạp hơn khi các nhà bán lẻ thêm các địa điểm mới với nhu cầu riêng biệt, sản phẩm mới, thị hiếu theo mùa luôn thay đổi và tiếp thị sản phẩm không thể đoán trước.

Mục tiêu báo cáo

- Giúp sinh viên có cái nhìn tổng quát về chuỗi thời gian và ứng dụng của nó.
- Có kỹ năng cơ bản sử dụng chuỗi thời gian để dự đoán vấn đề quan tâm.
- Dự đoán được doanh số bán hàng cho từng nhóm sản phẩm và các cửa hàng.

- Nhìn nhận chi tiết, đưa ra nhận xét về vấn đề quan tâm được nhắc đến trong báo cáo.

Mã nguồn

Toàn bộ code, báo cáo được lưu thành một thư mục lớn và được tải lên Github: https://github.com/minhanh-dang/store-sales-ts-forecasting.git

Chương 1

Cơ sở lý thuyết

1.1 Ra quyết định

1.1.1 Khái niệm

"Ra quyết định là một quá trình lựa chọn có ý thức giữa hai hoặc nhiều phản ánh để chọn ra một phản ánh và phản ánh này sẽ tạo ra được một kết quả mong muốn trong các điều kiện ràng buộc đã biết".

- Quyết định có thể là nhận thức ở dạng mô tả;
- Quyết định có thể là nhận thức ở dạng quá trình;
- Quyết định có thể là một hoạt động giàu ý thức;
- Quyết định có thể là những thay đổi trạng thái kiến thức.

1.1.2 Vai trò

Tại sao phải hỗ trợ ra quyết định?

1. Nhu cầu hỗ trợ ra quyết định
    - Ra quyết định luôn cần xử lý kiến thức;
    - Kiến thức là nguyên liệu và thành phẩm của ra quyết định, cần được sở hữu hoặc tích lũy bởi người ra quyết định.

2. Giới hạn về nhận thức;
3. Giới hạn về kinh tế;
4. Giới hạn về thời gian;
5. Áp lực cạnh tranh.
6. Bản chất của hỗ trợ ra quyết định:
   - Quyết định có cấu trúc đến phi cấu trúc;
   - Cung cấp thông tin, tri thức;
   - Thể hiện qua tương tác người – máy;
   - Thể hiện qua mô phỏng.

1.2 Hệ hỗ trợ quyết định

1.2.1 Khái niệm

- Hệ hỗ trợ quyết định (Decision Support System – DSS) là hệ thống dựa trên máy tính, có tính tương tác, giúp các nhà ra quyết định dùng dữ liệu và mô hình để giải quyết các bài toán phi cấu trúc (Scott Morton, 1971).
- DSS kết hợp trí lực của con người với năng lực của máy tính để cải tiến chất lượng của quyết định. Đây là các hệ dựa vào máy tính hỗ trợ cho người ra quyết định giải các bài toán nửa cấu trúc (Keen and Scott Morton, 1978).
- DSS là tập các thủ tục dựa trên mô hình nhằm xử lý dữ liệu và phán đoán của con người để giúp nhà quản lý ra quyết định (Little, 1970).

Nhu cầu về DSS: Những năm 1980, 1990 điều tra các công ty lớn cho thấy:

- Kinh tế thiếu ổn định;
- Khó theo dõi vận hành của doanh nghiệp;
- Cạnh tranh gay gắt;

- Bộ phận IT quá bận, không giải quyết được các yêu cầu quản lý;
- Cần phân tích lợi nhuận, hiệu quả và thông tin chính xác, mới, kịp thời;
- Giảm giá phí hoạt động;
- Xu hướng tính toán của người dùng.

1.2.2 Vai trò

- Cải thiện tốc độ tính toán;
- Tăng năng suất của cá nhân liên đới;
- Cải tiến kỹ thuật trong việc lưu trữ, tìm kiếm, trao đổi dữ liệu trong và ngoài tổ chức theo hướng nhanh và kinh tế;
- Nâng cao chất lượng của các quyết định đưa ra;
- Tăng cường năng lực cạnh tranh của tổ chức;
- Khắc phục khả năng hạn chế của con người trong việc xử lý và lưu chứa thông tin.

1.3 Dự báo (Forecasting)

Dự báo có lẽ là ứng dụng phổ biến nhất của học máy trong thế giới thực. Các doanh nghiệp dự báo nhu cầu sản phẩm, các chính phủ dự báo tăng trưởng kinh tế và dân số, các nhà khí tượng học dự báo thời tiết. Sự hiểu biết về những điều sắp xảy ra là một nhu cầu cấp thiết trong khoa học, chính phủ và ngành công nghiệp (không đề cập đến cuộc sống cá nhân của chúng ta) và những người thực hành trong các lĩnh vực này đang ngày càng áp dụng máy học để giải quyết nhu cầu này.

1.4 Chuỗi thời gian (Time Series)

Dự báo chuỗi thời gian là một lĩnh vực rộng lớn có lịch sử lâu đời. Đối tượng cơ bản của dự báo là chuỗi thời gian, là một tập hợp các quan sát được ghi lại theo

thời gian. Trong các ứng dụng dự báo, các quan sát thường được ghi lại với tần suất thường xuyên, như hàng ngày hoặc hàng tháng.

Chương 2

Ứng dụng chuỗi thời gian vào dự báo doanh thu

2.1 Giới thiệu bộ dữ liệu

Dữ liệu đến từ một công ty Ecuador có tên là Corporación Favorita và đây là một nhà bán lẻ tạp hóa lớn. Ngoài ra, công ty hoạt động ở các quốc gia khác ở Nam Mỹ.

Bộ dữ liệu được đăng tải bởi Kaggle:

https://www.kaggle.com/competitions/store-sales-time-series-forecasting

Bộ dữ liệu gồm có các tập tin sau:

1. **train.csv**

   - Dữ liệu được luyện, bao gồm chuỗi thời gian của các tính năng **store_nbr**, family và onpromotion cũng như doanh số mục tiêu.
   - **store_nbr** xác định cửa hàng nào bán mà sản phẩm nhắc đến được bán.
   - **family** xác định loại sản phẩm được bán.
   - **sales** cho biết tổng doanh số bán hàng cho một nhóm sản phẩm tại một cửa hàng cụ thể vào một ngày nhất định. Giá trị phân số là có thể vì sản phẩm có thể được bán theo đơn vị phân số (ví dụ: 1,5 kg pho mát, trái ngược với 1 túi khoai tây chiên).

- **onpromotion** cho biết tổng số mặt hàng trong một nhóm sản phẩm đang được khuyến mại tại một cửa hàng vào một ngày nhất định.

2. **test.csv**

   - Dữ liệu kiểm tra, có các tính năng tương tự như dữ liệu đào tạo. Bạn sẽ dự đoán doanh số mục tiêu cho các ngày trong tệp này.
   - Ngày trong dữ liệu kiểm tra là 15 ngày sau ngày cuối cùng trong dữ liệu đào tạo.

3. **sample_submission.csv**

   - File mẫu để trình bày theo đúng định dạng.

4. **store.csv**

   - Lưu trữ siêu dữ liệu, bao gồm thành phố, tiểu bang, loại và cụm.
   - **cluster** là một cụm, một nhóm các cửa hàng tương tự.

5. **oil.csv**

   - Giá dầu hàng ngày. Bao gồm các giá trị trong cả khung thời gian dữ liệu thử nghiệm và đào tạo. (Ecuador là một quốc gia phụ thuộc vào dầu mỏ và nền kinh tế của nước này rất dễ bị ảnh hưởng bởi những cú sốc về giá dầu).

6. **holiday_events.csv**

   - Ngày lễ và sự kiện, với siêu dữ liệu
   - *Lưu ý*: Đặc biệt chú ý đến cột **transferred**. Một ngày lễ được chuyển chính thức rơi vào ngày dương lịch đó, nhưng đã được chính phủ dời sang một ngày khác. Một ngày được chuyển giao giống như một ngày bình thường hơn là một ngày lễ. Để tìm ngày mà nó thực sự được tổ chức, hãy tìm hàng tương ứng có loại là Đã được chuyển (Tranferred). Ví dụ, ngày lễ Independencia de Guayaquil được chuyển từ 2012-10-09 sang 2012-10-12,

có nghĩa là nó được tổ chức vào 2012-10-12. Những ngày thuộc loại Cầu nối là những ngày bổ sung được thêm vào một kỳ nghỉ (ví dụ: để kéo dài thời gian nghỉ trong một ngày cuối tuần dài). Những điều này thường được tạo thành bởi loại Ngày làm việc là một ngày thường không được lên lịch làm việc (ví dụ: Thứ bảy) nhằm mục đích hoàn vốn cho Cầu.

- Các ngày lễ bổ sung là những ngày được thêm vào một ngày lễ theo lịch thông thường, chẳng hạn như thường xảy ra vào khoảng lễ Giáng sinh (làm cho đêm Giáng sinh trở thành một ngày lễ).

7. **Ghi chú bổ sung**

- Tiền lương trong khu vực công được trả hai tuần một lần vào ngày 15 và ngày cuối cùng của tháng. Doanh số siêu thị có thể bị ảnh hưởng bởi điều này.
- Một trận động đất mạnh 7,8 độ Richter đã xảy ra ở Ecuador vào ngày 16 tháng 4 năm 2016. Mọi người tập hợp trong các nỗ lực cứu trợ quyên góp nước và các sản phẩm cần thiết khác, điều này đã ảnh hưởng lớn đến doanh số bán hàng của siêu thị trong vài tuần sau trận động đất.

2.2 Xử lý, quan sát dữ liệu

Phần này nằm trong phần thực hành ở dưới hoặc tại https://github.com/minhanh-dang/store-sales-ts-forecasting.git

2.3 Dự báo doanh thu sử dụng chuỗi thời gian

Phần này nằm trong phần thực hành ở dưới hoặc tại https://github.com/minhanh-dang/store-sales-ts-forecasting.git

Kết luận

2.4 Kết quả đạt được

Việc xử lý dữ liệu đòi hỏi rất nhiều kỹ năng. Sau khi hoàn thiện báo cáo, em nhận thấy mình đã đạt được những kết quả sau:

- Cải thiện kỹ năng xử lý dữ liệu
- Nâng cao kỹ năng biểu diễn dữ liệu
- Học được kiến thức cơ bản về chuỗi thời gian
- Doanh thu phụ thuộc vào rất nhiều thuộc tính, như thời tiết, thời điểm trong năm, trong tuần, thậm chí là cả giá dầu, tuỳ thuộc vào nền kinh tế của quốc gia được xét.

2.5 Hạn chế

Ngoài việc đã hoàn thiện báo cáo, em cảm thấy báo cáo còn tồn tại những mặt sau:

- Chưa phân tích được một số họ sản phẩm còn hoạt động/ bán được hay không do doanh thu các sản phẩm này quá ít, mang tính thời vụ
- Chưa đính kèm được phần thực hành vào báo cáo vì các bước phức tạp

2.6 Hướng phát triển

- Học chuỗi thời gian mức nâng cao hơn để phân tích được sâu hơn

- Phân tích các họ sản phẩm có doanh thu hiếm trong vòng 60 ngày
- Đính kèm phần thực hành vào trong báo cáo

Tài liệu tham khảo

Tiếng Anh

[1] Corporación Favorita, https://www.corporacionfavorita.com/en/

[2] ÌM Country Focus (2019), Ecuador's New Economic Plan Explained, https://www.imf.org/en/News/Articles/2019/03/20/NA032119-Ecuador-New-Economic-Plan-Explained

[3] Kashish Rastogi (2021), Store Sales Analysis – Time Series, https://www.kaggle.com/kashishrastogi/store-sales-forecasting

[4] Gianluca Malato (2020), An algorithm to find the best moving average for stock trading, https://towardsdatascience.com/an-algorithm-to-find-the-best-moving-average-for-stock-trading-1b024672299c