**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN CÔNG NGHỆ KHOA HỌC**

**PHÂN TÍCH VÀ DỰ ĐOÁN SỐ LƯỢNG HÀNG HÓA**

**Giảng viên giảng dạy : ThS. Sử Nhật Hạ**

**Sinh viên thực hiện : Trần Trọng Khang**

**MSSV : 2100009261**

**Chuyên ngành : KHOA HỌC DỮ LIỆU**

**Môn học : Công nghệ khoa học**

**Khóa : 2021**

**Tp.HCM, tháng 09 Năm 2024**

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN CÔNG NGHỆ KHOA HỌC**

**PHÂN TÍCH VÀ DỰ ĐOÁN SỐ LƯỢNG HÀNG HÓA**

**Giảng viên giảng dạy : ThS. Sử Nhật Hạ**

**Sinh viên thực hiện : Trần Trọng Khang**

**MSSV : 2100009261**

**Chuyên ngành : KHOA HỌC DỮ LIỆU**

**Môn học : Công nghệ khoa học**

**Khóa : 2021**

**Tp.HCM, tháng 09 Năm 2024**

# LỜI MỞ ĐẦU

Trong những năm gần đây, với sự hội nhập kinh tế ngày càng sâu rộng, cạnh tranh trên thị trường trở nên khốc liệt hơn bao giờ hết. Để tồn tại và phát triển, các doanh nghiệp phải đối mặt với nhiều thách thức, đòi hỏi họ không ngừng cải thiện hoạt động kinh doanh và phát triển các chiến lược linh hoạt để thích ứng với môi trường thị trường luôn biến động.

Phân tích và dự đoán số lượng hàng hóa là một bước quan trọng giúp các nhà quản trị nắm bắt được tình hình thị trường cũng như nhu cầu của khách hàng. Thông qua việc phân tích dữ liệu, doanh nghiệp có thể đánh giá chính xác về tình trạng hàng hóa hiện tại, từ đó đưa ra các quyết định hợp lý về việc nhập hàng, điều chỉnh nguồn cung, và tối ưu hóa quy trình kinh doanh.

Trong bối cảnh kinh tế thế giới và trong nước đang phát triển nhanh chóng, việc dự đoán và quản lý số lượng hàng hóa trở thành một yếu tố then chốt giúp doanh nghiệp tồn tại và thành công. Điều này đặt ra thách thức lớn đối với các nhà quản trị, nhưng đồng thời cũng mở ra cơ hội để họ khai thác tối đa tiềm năng và nguồn lực của doanh nghiệp, từ đó nâng cao hiệu quả kinh doanh và khả năng cạnh tranh trên thị trường.

Phân tích và dự đoán số lượng hàng hóa không chỉ giúp doanh nghiệp đảm bảo nguồn cung ổn định, mà còn hạn chế rủi ro liên quan đến hàng tồn kho và chi phí lưu trữ. Việc này đòi hỏi sự cẩn thận và chính xác trong quá trình đánh giá dữ liệu, nhằm đảm bảo rằng các quyết định được đưa ra phù hợp với xu hướng thị trường và định hướng phát triển của doanh nghiệp.

# LỜI CẢM ƠN

Với lòng biết ơn sâu sắc ,em chân thành gửi lời biết ơn tới quý Thầy Cô ở Khoa Công Nghệ Thông Tin, Trường Đại Học Nguyễn Tất Thành, vì đã chia sẻ với chúng em những kiến thức quý báu trong suốt thời gian học tại trường. Nhờ những chỉ dẫn và sự hỗ trợ của quý Thầy Cô, đề tài dự đoán thời tiết với các mô hình hồi quy của em đã thuận lợi và thành công.

Em xin bày tỏ lòng biết ơn đặc biệt đến cô ThS Sử Nhật Hạ, người đã tận tâm hướng dẫn, giúp đỡ và quan tâm để em có thể hoàn thành sản phẩm và báo cáo này một cách xuất sắc.

Bài báo cáo của em còn hạn chế và có một số thiếu sót, em rất mong nhận được sự đóng góp quý báu từ phía quý Thầy Cô. Những ý kiến này sẽ giúp em cải thiện kiến thức và kỹ năng trong lĩnh vực này, đồng thời nâng cao ý thức và trình độ cá nhân.

Em chân thành cảm ơn sự hỗ trợ và đóng góp quý báu của quý Thầy Cô. Rất mong nhận được sự chỉ bảo và hướng dẫn để phát triển bản thân mình.

Sinh viên thực hiện

TRẦN TRỌNG KHANG

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ 3 NĂM HỌC 2023 - 2024** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: CÔNG NGHỆ KHOA HỌC DỮ LIỆU Lớp học phần: 2324HK3-21DTH2D

Nhóm sinh viên thực hiện :NHÓM 14

1.TRẦN TRỌNG KHANG Tham gia đóng góp:100%

Đề tài tiểu luận/báo cáo của sinh viên : **PHÂN TÍCH VÀ DỰ ĐOÁN SỐ LƯỢNG HÀNG HÓA**

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  |  |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  |  |  |
| * Lập luận |  |  |  |
| * Kết luận |  |  |  |
| Trình bày |  |  |  |
| **TỔNG ĐIỂM** |  |  |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

**Ths.Sử Nhật Hạ**

**MỤC LỤC**

[LỜI MỞ ĐẦU 1](#_Toc174277007)

[LỜI CẢM ƠN 2](#_Toc174277008)

[DANH MỤC BẢNG 6](#_Toc174277009)

[DANH MỤC HÌNH 7](#_Toc174277010)

[CHƯƠNG 1: Cơ sở lý thuyết 1](#_Toc174277011)

[1.1 Tổng quan về phân tích dữ liệu. 1](#_Toc174277012)

[1.1.1 Phân tích dữ liệu là gì? 1](#_Toc174277013)

[1.1.2 Quá trình phân tích dữ liệu 2](#_Toc174277014)

[1.1.3 Có những kỹ thuật phân tích dữ liêu nào? 3](#_Toc174277015)

[1.1.4 Tại sao phân tích dữ liệu lại quan trọng ? 4](#_Toc174277016)

[1.1.5 Một số ứng dụng tiêu biểu của Khoa học dữ liệu 5](#_Toc174277017)

[1.2. Giới thiệu đề tài 5](#_Toc174277018)

[1.2.1. Mục tiêu nghiên cứu 6](#_Toc174277019)

[1.2.2. Mô tả dữ liệu 6](#_Toc174277020)

[CHƯƠNG 2: Phân tích yêu cầu 7](#_Toc174277021)

[2.1. Tổng quan về Google Colab 7](#_Toc174277022)

[2.1.1 Mô tả 7](#_Toc174277023)

[2.1.2 Tính năng 7](#_Toc174277024)

[2.2 Quy trình phân lớp dữ liệu 8](#_Toc174277025)

[2.2.1 Phương pháp hồi quy tuyến tính 8](#_Toc174277026)

[2.2.2 Phương pháp phân lớp dữ liệu 8](#_Toc174277027)

[2.3 Ứng dụng khoa học dữ liệu 9](#_Toc174277028)

[2.3.1. Dự đoán vào doanh số bàn hàng thương mại điện tử 9](#_Toc174277029)

[2.3.2 Phân tích ảnh hưởng của thuộc tính 10](#_Toc174277030)

[2.3.3 Tối ưu hóa dữ liệu 10](#_Toc174277031)

[CHƯƠNG 3: Xây dựng mô hình 11](#_Toc174277032)

[3.1 Tổng quan về dữ liệu 11](#_Toc174277033)

[3.2 Tiền xử lý dữ liệu 12](#_Toc174277034)

[3.2.1 Truy cập thư viện và tải dữ liệu 12](#_Toc174277035)

[3.2.2 Làm sạch và tiền xử lý dữ liệu 15](#_Toc174277036)

[CHƯƠNG 4: Thực nghiệm mô hình 18](#_Toc174277037)

[4.1 Phân tích và trực quan hóa dữ liệu 18](#_Toc174277038)

[4.1.1 Tháng nào có doanh số tốt nhất? Doanh số tháng đó là bao nhiêu? 18](#_Toc174277039)

[4.1.2 Thành phố nào có doanh số cao nhất? 19](#_Toc174277040)

[4.1.3 Doanh nghiệp cần chiếu quảng cáo vào khung thời gian nào để tăng khả năng mua hàng của khách hàng? 20](#_Toc174277041)

[4.1.4 Những sản phẩm nào thường được bán cùng nhau? 21](#_Toc174277042)

[4.1.5 Sản phẩm nào được bán nhiều nhất? Giả thiết của bạn về lý do sản phẩm này được bán nhiều nhất là gì? 22](#_Toc174277043)

[4.2 Dự đoán (prediction) 22](#_Toc174277044)

[4.2.1 Lấy mẫu dữ liệu 23](#_Toc174277045)

[4.2.2 Lựa chon mô hình phù hợp để dự đoán: 23](#_Toc174277046)

[4.2.3 Tiến hành dự báo: 27](#_Toc174277047)

[CHƯƠNG 5: Kết luận và hướng phát triển 29](#_Toc174277048)

[5.1. Kết quả đạt được 29](#_Toc174277049)

[5.2. Hạn chế của đề tài 29](#_Toc174277050)

[5.3. Hướng phát triển 29](#_Toc174277051)

# DANH MỤC BẢNG

[Bảng 1-1: Mô tả thuộc tính 6](#_Toc174276130)

[Bảng 3-3: Mô tả thuộc tính của bộ dữ liệu qua xử lý 12](#_Toc174276131)

# DANH MỤC HÌNH

[Hình 1‑1: Các phân tích trong dữ liệu 2](#_Toc174277142)

[Hình 1‑2: Quá trình phân tích 3](#_Toc174277143)

[Hình 2-1: Dự báo doanh thu ngành công nghiệp sản phẩm thương mại điện tử 9](#_Toc174277144)

[Hình 3‑1: Bộ dữ liệu sau khi qua xử lý để phân tích 11](#_Toc174277145)

[Hình 3‑2: Các thuộc tính của bộ dữ liệu sau xử lý 12](#_Toc174277146)

[Hình 3‑4: Bộ dữ liệu cần được xử lý 13](#_Toc174277147)

[Hình 3‑5: Gọi thư viện 13](#_Toc174277148)

[Hình 3‑6: Phương thức gộp dữ liệu 14](#_Toc174277149)

[Hình 3‑7: Bộ dữ liệu sau khi gộp 15](#_Toc174277150)

[Hình 3‑8: Thêm cột Month 15](#_Toc174277151)

[Hình 3‑9: Kiểm tra giá trị null trong dữ liệu 16](#_Toc174277152)

[Hình 3‑10: Loại bỏ các dữ liệu thiếu 16](#_Toc174277153)

[Hình 3‑11: Các thuộc tính bổ sung 17](#_Toc174277154)

[Hình 4‑1:Biểu đồ cột doanh thu theo Tháng 18](#_Toc174277155)

[Hình 4‑2: Biểu đồ cột doanh thu theo Thành phố 19](#_Toc174277156)

[Hình 4‑3: Biểu đồ đường về thời gian mua sắm của khách hàng 20](#_Toc174277157)

[Hình 4‑4: Các sản phẩm thường được bán cùng nhau 21](#_Toc174277158)

[Hình 4‑5: Biểu đồ cột và đường thể hiện số lượng mua và giá sản phẩm 22](#_Toc174277159)

[Hình 4‑6: Bộ dữ liệu qua xử lý và đã được rút ngắn các thuộc tính cần thiết 23](#_Toc174277160)

[Hình 4‑7: Lựa chọn thuộc tính 23](#_Toc174277161)

[Hình 4‑8: Chia tập dữ liệu 23](#_Toc174277162)

[Hình 4‑9: Mô hình Linear Regression 24](#_Toc174277163)

[Hình 4‑10: Mô hình Ridge Regression 25](#_Toc174277164)

[Hình 4‑11: Mô hình Lasso 25](#_Toc174277165)

[Hình 4‑12: Mô hình Catboost 26](#_Toc174277166)

[Hình 4‑13: Mô hình Randoom Forest 27](#_Toc174277167)

[Hình 4‑14: Biểu đồ trực quan 27](#_Toc174277168)

[Hình 4‑15: Kết quả dự đoán 28](#_Toc174277169)

# CHƯƠNG 1: Cơ sở lý thuyết

**1.1 Tổng quan về phân tích dữ liệu.**

**1.1.1 Phân tích dữ liệu là gì?**

Phân tích dữ liệu là quá trình thu thập, xử lý, làm sạch, và mô hình hóa dữ liệu , nhằm rút ra thông tin hữu ích, hỗ trợ quyết định và dự đoán xu hướng. Quá trình này có thể được thực hiện bằng nhiều phương pháp khác nhau, từ các phương pháp thống kê truyền thống đến các kỹ thuật học máy hiện đại.

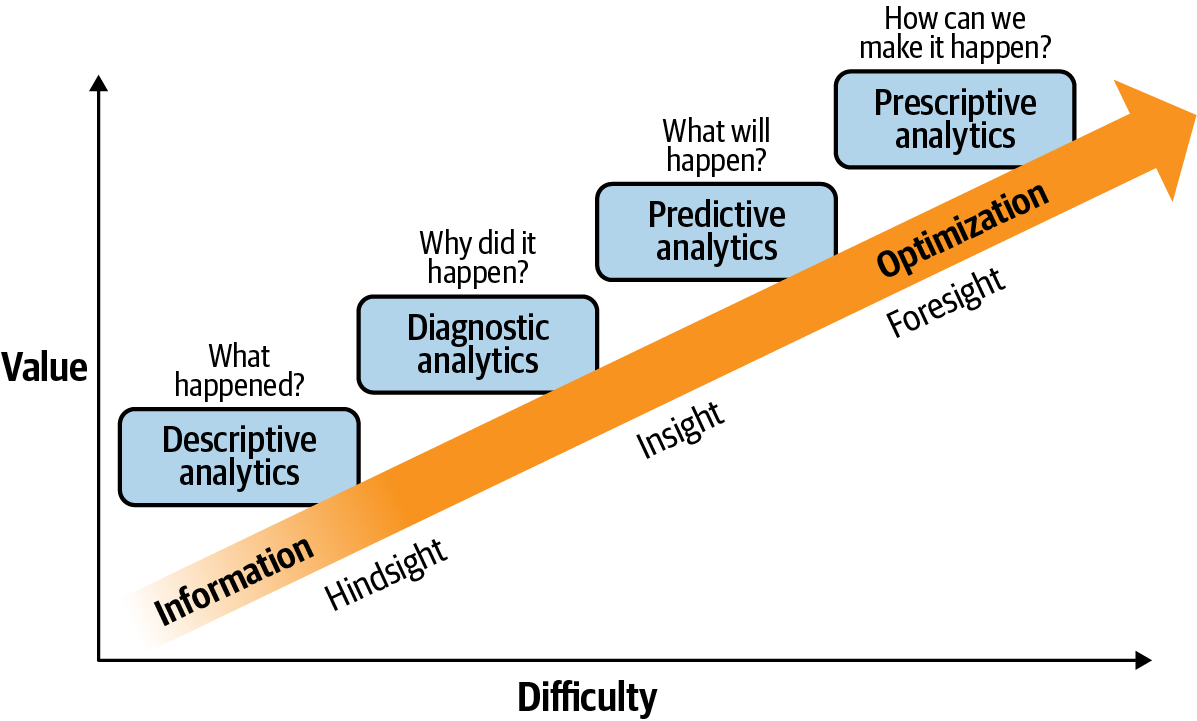
Các loại phân tích dữ liệu?

**Phân tích mô tả** : Chuyện gì đã xảy ra? Cung cấp thông tin chi tiết về các sự kiện trong quá khứ

**Phân tích chuẩn đoán** : Tại sao điều đó xảy ra? Sử dụng thông tin chi tiết từ phân tích mô tả, đào sâu hơn nhằm tìm ra nguyên nhân của kết quả

**Phân tích dự đoán :** Chuyện gì sẽ xảy ra tiếp theo? Tận dụng dữ liệulịch sử và xu hướng để dự đoán tương lai

**Phân tích theo quy định :**Những gì nên được thực hiện về nó ? Phân tích các quyết định và sự kiện trong quá khứ để ước tính khả năng xảy ra các kết quả khác nhau



Hình 1‑1: Các phân tích trong dữ liệu

**1.1.2 Quá trình phân tích dữ liệu**

* **Hiểu vấn đề và kết quả mong muốn:** Xác định bạn đang ở đâu và bạn muốn ở đâu
* **Đặt thước đo rõ ràng :** Quyết định cái gì sẽ được đo lường và nó được đo lường như thế nào
* **Thu thập dữ liệu ràng:** Các định dữ liệu bạn yêu cầu, cácnguồn dữ liệu mà bạn sẽ truy cập và công cụ thu thập dữ liệu tốt nhất
* **Làm sạch dữ liệu :** Khắc phục các vấn đề về chất lượng trong dữ liệu và chuẩn hóa dữ liệu đến từ nhiều nguồn
* **Phân tích và khai thác dữ liệu:** Trích xuất, phân tích và thao tác dữ liệu từ các quan điểm khác nhau để hiểu xu hướng, xác địnhmối tương quan, tìm các mẫu và biến thể
* **Giải thích kết quả:** Diễn giải kết quả, đánh giá khả năng bảo vệ của phân tích trước các trường hợp mà phân tích có thể không đúng
* **Trình bày những phát hiện :** Giao tiếp và trình bày  
  những phát hiện của bạn theo những cách rõ ràng, có tác động và thuyết phục



Hình 1‑2: Quá trình phân tích

**1.1.3 Có những kỹ thuật phân tích dữ liêu nào?**

• **Phân Tích Mô Tả (Descriptive Analytics):**

Tóm tắt và mô tả dữ liệu hiện tại để hiểu rõ tình hình.

• **Phân Tích Hồi Quy (Regression Analysis):**

Xác định mối quan hệ giữa các biến và dự đoán một biến phụ thuộc dựa trên các biến độc lập.

**• Phân tích nhân tố (Factor Analysis):**

Mục tiêu: Giảm số lượng biến số quan sát và tìm ra các biến ẩn (nhân tố) thể hiện dữ liệu.

**• Phân tích phân tán (Dispersion Analysis):**

Mục tiêu: Đo lường sự phân tán và biến thiên trong dữ liệu.

**• Phân tích biệt thức (Discriminant Analysis):**

Mục tiêu: Phân loại các đối tượng vào các nhóm dựa trên các biến độc lập.

**• Logic mờ (Fuzzy Logic):**

Mục tiêu: Xử lý các thông tin không chắc chắn hoặc mờ nhạt, không rõ ràng.

**• Lập trình tiến hóa (Evolutionary Programming):**

Mục tiêu: Tối ưu hóa các giải pháp dựa trên nguyên lý tiến hóa và chọn lọc tự nhiên.

**• Cây quyết định (Decision Trees):**

Mục tiêu: Phân loại hoặc dự đoán kết quả dựa trên các quyết định được biểu diễn dưới dạng cây.

**• Mạng thần kinh nhân tạo (Artificial Neural Networks):**

Mục tiêu: Mô phỏng hoạt động của bộ não con người để nhận diện mẫu và dự đoán kết quả.

Công cụ: Perceptron, Multi-layer Perceptron (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN).

**• Phân Tích Chuỗi Thời Gian (Time Series Analysis):**

Phân tích dữ liệu theo thời gian để dự đoán xu hướng tương lai.

Công cụ: ARIMA, SARIMA, Holt-Winters.

**1.1.4 Tại sao phân tích dữ liệu lại quan trọng ?**

Tại sao phân tích dữ liệu lại quan trọng?

- Hỗ trợ quyết định thông minh:

Giúp đưa ra quyết định dựa trên dữ liệu thực tế, giảm thiểu rủi ro.

- Tối ưu hóa quy trình:

Nâng cao hiệu suất hoạt động và tiết kiệm chi phí.

- Dự đoán và dự báo:

Nhận diện xu hướng và dự đoán hành vi trong tương lai.

- Phát hiện và ngăn chặn gian lận:

Phát hiện các mẫu gian lận, bảo vệ tài sản của tổ chức.

- Cải thiện trải nghiệm khách hàng:

Hiểu rõ nhu cầu khách hàng, nâng cao sự hài lòng và giữ chân khách hàng.

- Khám phá cơ hội kinh doanh mới:

Tìm kiếm thị trường và sản phẩm tiềm năng, mở rộng kinh doanh.

- Hỗ trợ nghiên cứu và phát triển:

Đóng góp vào tiến bộ khoa học, kỹ thuật và y tế.

- Đánh giá và điều chỉnh chiến lược:

Đo lường hiệu quả của các chiến lược kinh doanh, marketing, và hoạt động để điều chỉnh và cải thiện.

Những điểm này làm nổi bật vai trò quan trọng của phân tích dữ liệu trong việc giúp các tổ chức và doanh nghiệp hoạt động hiệu quả, cải thiện trải nghiệm khách hàng, và mở rộng cơ hội kinh doanh.

**1.1.5 Một số ứng dụng tiêu biểu của Khoa học dữ liệu**

Các công ty phụ thuộc vào nền tảng dữ liệu để cấu trúc, phát triển và cải tiến doanh nghiệp. Các Data Scientist làm việc với các con số, phân tích một khối lượng lớn Data để xuất ra những Insight ý nghĩa. Những insight này rất hữu ích khi phân tích công ty và cách hoạt động của công ty trên thị trường từ đó đưa ra các quyết định đúng đắn. Những ứng dụng tiêu biểu của phân tích dữ liệu trong thời kỳ chuyển đổi số mạnh mẽ 4.0:

* Tối ưu hóa Marketing
* Phân tích dữ liệu các trang web
* Phân tích dữ liệu rủi ro
* Phân tích dữ liệu an ninh
* Phân tích dữ liệu phần mềm
* Đặc biệt được quan tâm nhất: Phân tích dữ liệu nhân sự

Trong bối cảnh cuộc cách mạng công nghiệp 4.0 đang phát triển mạnh mẽ, việc các doanh nghiệp sản xuất phương tiện đi dại, các nhà kinh tế, các kỹ thuật sư,…ứng dụng công nghệ mới sẽ mang lại hiệu quả trong hoạt động marketing, quản lý tài chính, phương thức thanh toán, hỗ trợ khách hàng.

**1.2. Giới thiệu đề tài**

Sau khi tìm hiểu và thảo luận, em đã chọn đề tài: “­­PHÂN TÍCH VÀ DỰ ĐOÁN SỐ LƯỢNG HÀNG HÓA” làm đồ án. Đề tài này tập trung vào việc phân tích một số cơ sở dữ liệu của “Sales-Reporting”. Cơ sở dữ liệu này sẽ bao gồm các thông tin về nhiều thuộc tính có thể kể đến như: Trong nặm đó mà một nhà sản xuất có thể thu được bao nhiêu, số lượng doanh thu bán ra được, tổng doanh thu trên khu vực đó,…

Trong đề tài này, chúng ta sẽ sử dụng ứng dụng Google Colab để khám phá và phân tích cơ sở dữ liệu. Bằng cách trực quan hóa dữ liệu và áp dụng các kỹ thuật phân tích, chúng ta có thể hiểu rõ hơn về mối quan hệ giữa các thuộc tính và hiệu suất nhiên liệu của các thuộc tính. Với phương pháp phân lớp dữ Logistic Regression, Hồi quy sẽ được em áp dụng trong đề tài này.

Qua việc phân tích cơ sở dữ liệu bằng ứng dụng Google Colab, ta có thể dự đoán mức tiêu thụ của một thị trường cho một tựa trò chơi điện tử hoặc nhiều hơn thế dựa trên các thuộc tính kỹ thuật của nó. Điều này có thể giúp hỗ trợ các nhà sản xuất có thể hiểu được người tiêu dùng cần những gì và theo đó sẽ thận trọng hơn và người tiêu dùng có thể hõa mãn được những nhu cầu họ cần.

**1.2.1. Mục tiêu nghiên cứu**

* Khám phá dữ liệu.
* Dự báo doanh thu thông qua dữ liệu chứa doanh số hiện tại.
* Nêu ra hướng phát triển và hạn chế của mô hình

**1.2.2. Mô tả dữ liệu**

Dữ liệu gốc bao gồm thuộc tính 6 và 186850 kiểu dữ liệu.

Tiếp theo, em sẽ tiến hành mô tả dữ liệu và giải thích các thuộc tính của bộ dữ liệu “Sales-Reporting” và kiểu dữ liệu.

|  |  |  |
| --- | --- | --- |
| **Thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** |
| Order ID | Thông tin mua hàng | object |
| Product | Sản phẩm | object |
| Quantity Ordered | Số lượng mua hàng | int8 |
| Price Each | Giá mỗi sản phẩm | float32 |
| Order Date | Ngày mua | datetime64 |
| Purchase Address | Địa chỉ mua hàng | object |

Bảng 1-1: Mô tả thuộc tính

# CHƯƠNG 2: Phân tích yêu cầu

**2.1. Tổng quan về Google Colab**

**2.1.1 Mô tả**

Khai phá dữ liệu (Data Mining) và Học máy (Machine Learning) là hai lĩnh vực rộng lớn và vô cùng phức tạp, đòi hỏi cần có sự hiểu biết sâu sắc và nghiên cứu kỹ lưỡng. Để hỗ trợ và giúp đỡ người dùng trong việc khám phá và nghiên cứu các bài toán trong lĩnh vực này, đã có nhiều phần mềm ra đời và phát triển. Một trong số đó không thể nhắc đến chính là phần mềm Google Colab.

Google Colab là một sản phẩm giống Jupyter Notebook của Google Research. Google đã lập trình đặc biệt công cụ mã hóa Python dựa trên đám mây này để ghi nhớ nhu cầu của các lập trình viên máy học, nhà phân tích dữ liệu lớn, nhà khoa học dữ liệu, nhà nghiên cứu AI và người học Python.

Phần tốt nhất là một sổ ghi chép mã cho tất cả các thành phần cần thiết để trình bày một dự án khoa học dữ liệu hoặc máy học hoàn chỉnh cho người giám sát hoặc nhà tài trợ chương trình. Ví dụ: sổ ghi chép Colab của bạn có thể chứa mã thực thi, mã Python trực tiếp, văn bản có định dạng, HTML, LaTeX, hình ảnh, hình ảnh hóa dữ liệu, biểu đồ, đồ thị, bảng, v.v.

**2.1.2 Tính năng**

Một trong những ưu điểm nổi bật của Google Colab là việc sử dụng ngôn ngữ lập trình Python. Python là một ngôn ngữ lập trình phổ biến và mạnh mẽ, có cộng đồng hỗ trợ lớn và nhiều thư viện phong phú cho khai phá dữ liệu và học máy. Việc sử dụng Python làm ngôn ngữ chính của Google Colab mang lại sự linh hoạt và khả năng mở rộng cao, cho phép người dung tận dụng toàn bộ sức mạnh của Python trong quá trình nghiên cứu và phân tích dữ liệu.

**2.2 Quy trình phân lớp dữ liệu**

**2.2.1 Phương pháp hồi quy tuyến tính**

Mô hình hồi quy tuyến tính (Linear Regression Model) là một mô hình dự đoán được sử dụng để xác định mối quan hệ tuyến tính giữa một biến phụ thuộc (đầu ra) và một hoặc nhiều biến độc lập (đầu vào). Mô hình này dự đoán giá trị đầu ra dựa trên các hệ số tương ứng với các biến đầu vào.

Phương trình tổng quát của mô hình hồi quy tuyến tính như sau:

y= β₀+β₁x₁+β₂x₂+...+ βkxk

Trong đó:

y là biến phụ thuộc cần dự đoán,

x, x₂, ..., xk là các biến độc lập,

β, β₁, β₂, ..., βk là các hệ số tương ứng với các biến độc lập,

β là hệ số chặn (intercept).

**2.2.2 Phương pháp phân lớp dữ liệu**

Phân lớp dữ liệu là quá trình phân chia các mẫu dữ liệu vào các lớp hoặc nhãn khác nhau dựa trên các đặc trưng và thuộc tính của chúng. Đây là một nhiệm vụ quan trọng trong Machine Learning và Data Mining, giúp xây dựng mô hình dự đoán hoặc phân loại dữ liệu mới dựa trên các mẫu đã được gán nhãn.

Có nhiều phương pháp phân lớp dữ liệu khác nhau, tùy thuộc vào đặc điểm của dữ liệu và mục tiêu của bài toán. Dưới đây là một số phương pháp phân lớp phổ biến:

●Random Forest: Random Forest sử dụng nhiều cây quyết định ngẫu nhiên để phân lớp dữ liệu. Kết quả cuối cùng là sự kết hợp của các dự đoán từ các cây quyết định riêng lẻ, giúp cải thiện độ chính xác và khả năng tổng quát hóa.

●Ridge Regression :Ridge Regression thường áp dụng nó trong bối cảnh của Logistic Regression thay vì Linear Regression. Điều này có nghĩa là mô hình dự đoán xác suất của một điểm dữ liệu thuộc về một lớp nhất định và sau đó sử dụng Ridge để điều chỉnh các hệ số của mô hình nhằm tránh overfitting.

●Lasso: Lasso cũng có thể được áp dụng trong bối cảnh của Logistic Regression. Lasso sẽ phạt các hệ số của mô hình bằng cách sử dụng chuẩn L1L1L1, điều này không chỉ giúp ngăn overfitting mà còn có thể loại bỏ hoàn toàn một số biến không cần thiết, làm cho mô hình đơn giản hơn.

●Catboost: CatBoost xây dựng các cây quyết định để dự đoán xác suất của các lớp, sau đó kết hợp các cây này để cải thiện độ chính xác của mô hình tổng thể. CatBoost tự động xử lý các biến phân loại mà không cần chuyển đổi phức tạp, giúp đơn giản hóa quy trình làm việc.

**2.3 Ứng dụng khoa học dữ liệu**

**2.3.1. Dự đoán vào doanh số bàn hàng thương mại điện tử**

Dựa vào khoa học dữ liệu, ta có thể sử dụng rất nhiều phương pháp để xây dựng mô hình dự đoán doanh thu đồ điện tử dựa trên các thông số doanh thu về thị trường mỗi khu vực.



Hình 2-1: Dự báo doanh thu ngành công nghiệp  
 sản phẩm thương mại điện tử

**2.3.2 Phân tích ảnh hưởng của thuộc tính**

Sử dụng phương pháp thống kê và khai phá dữ liệu, chúng ta có thể phân tích và đánh giá tác động của các thuộc tính đến doanh số thị trường khu vực. Xác định những thuộc tính quan trọng nhất và tìm hiểu chúng có thể giúp chúng ta rất nhiều điều:

● Hiểu rõ mức độ ảnh hưởng: Phân tích giúp ta xác định mức độ ảnh hưởng của từng thuộc tính đến doanh thu thị trường đó. Điều này giúp chúng ta biết được thuộc tính nào có tác động mạnh và thuộc tính nào có tác động yếu hơn. Điều này rất hữu ích để tập trung vào các thuộc tính quan trọng nhất khi nghiên cứu và cảit hiện hiệu suất.

● Xác định yếu tố quyết định: Chúng ta có thể xác định những yếu tố quyết định chính trong việc ảnh hưởng đến thị trường trò chơi điện tử. Nhờ đó, ta có thể tập trung vào việc cải thiện những yếu tố này.

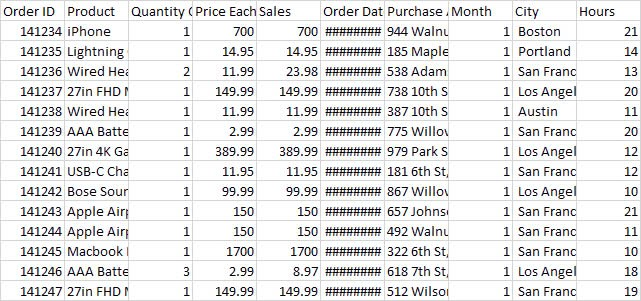
**2.3.3 Tối ưu hóa dữ liệu**

Khoa học dữ liệu có thể giúp ta tối ưu hóa các lựa chọn, từ đó có thể xác định các thông tin về thị trường biến động như thế nào để nắm bắt xu thế và từ đó phát triển. Việc áp dụng thuật toán tối ưu hóa và mô phỏng có thể giúp đưa ra các giải pháp cải tiến như xu hướng của người tiêu dung cần những gì hay những thể loại trò chơi điện tử đang có xu thế gia tăng và mang lại một lợi nhuận khổng lồ.

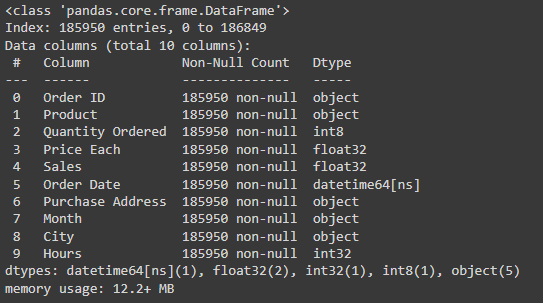
# CHƯƠNG 3: Xây dựng mô hình

**3.1 Tổng quan về dữ liệu**

Sử dụng bộ dữ liệu “[**Sales-Reporting**](https://github.com/boringPpl/Sales-Reporting)” trên github từ đường dẫn. Bộ dữ liệu gồm 12 tập dữ liệu csv với gần 200000 mẫu dữ liệu mục tiêu của bộ dữ liệu là giúp kiểm toán viên xây dựng mô hình có thể phân tích và dự đoán doanh số của các thị trường thương mại điện tử trên các khu vực trong các năm tới.



Hình 3‑1: Bộ dữ liệu sau khi qua xử lý để phân tích



Hình 3‑2: Các thuộc tính của bộ dữ liệu sau xử lý

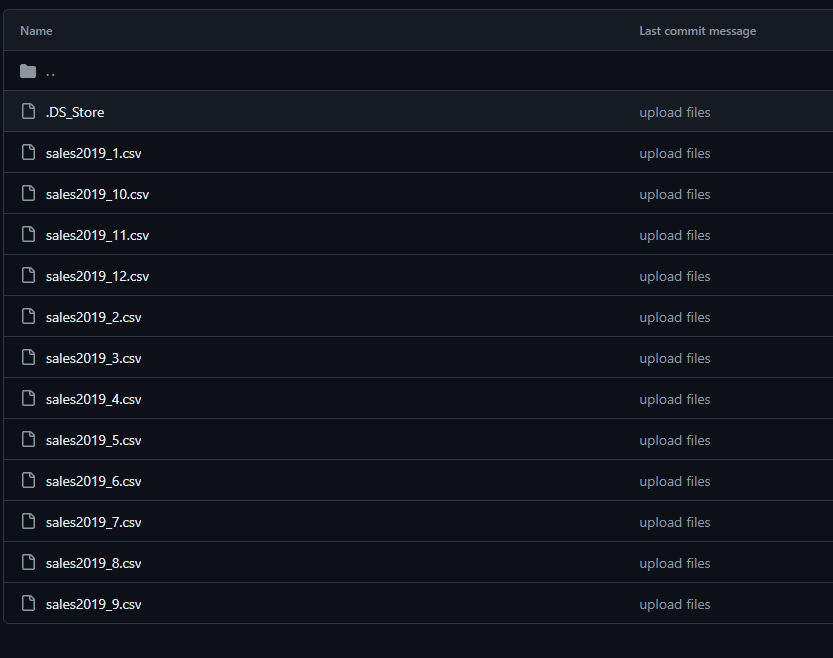
|  |  |  |
| --- | --- | --- |
| **Thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** |
| Order ID | Thông tin mua hàng | object |
| Product | Sản phẩm | object |
| Quantity Ordered | Số lượng mua hàng | int8 |
| Price Each | Giá mỗi sản phẩm | float32 |
| Sales | Doanh thu bán hàng | float32 |
| Order Date | Ngày mua | datetime64 |
| Purchase Address | Địa chỉ mua hàng | object |
| Month | Tháng mua hàng | object |
| City | Thành phố mua hàng | object |
| Hours | Thời gian mua | int32 |

Bảng 3-3: Mô tả thuộc tính của bộ dữ liệu qua xử lý

**3.2 Tiền xử lý dữ liệu**

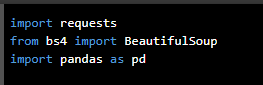
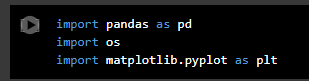
**3.2.1 Truy cập thư viện và tải dữ liệu**

Vì bộ dữ liệu được chia thành 12 file nên chúng ta cần phải làm hàm gộp các bộ dữ liệu lại để tiến hành phân tích và trực quan.



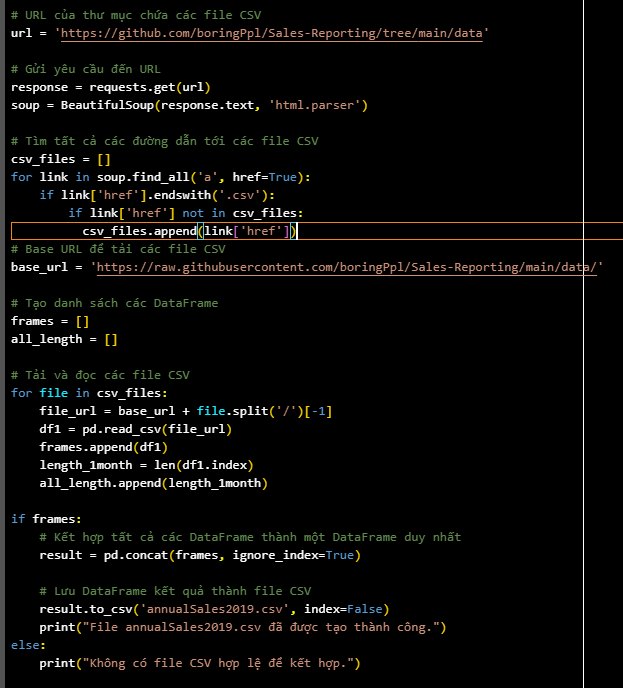
Hình 3‑4: Bộ dữ liệu cần được xử lý

Đầu tiên gọi thư viện



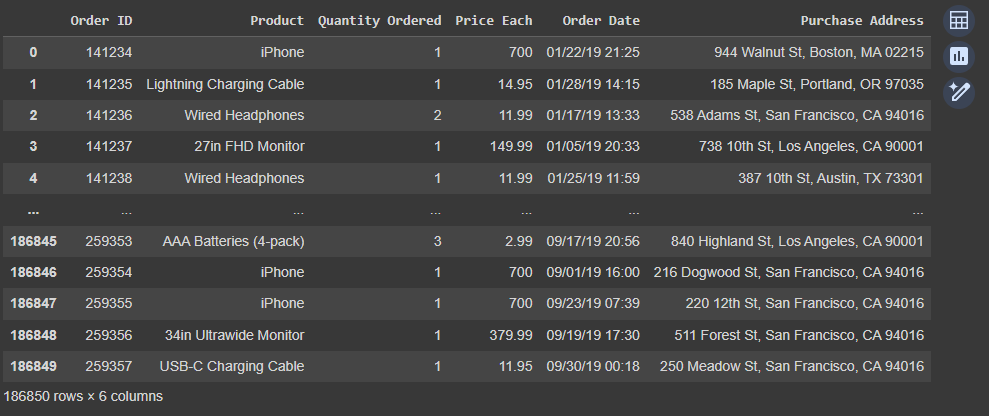
Hình 3‑5: Gọi thư viện

Tiếp theo là tạo một phương thức để gộp dữ liêu thành 1 file csv hoàn chỉnh.



Hình 3‑6: Phương thức gộp dữ liệu

Sau khi gộp thành công các file ta sẽ có file hoàn chỉnh như bên dưới:

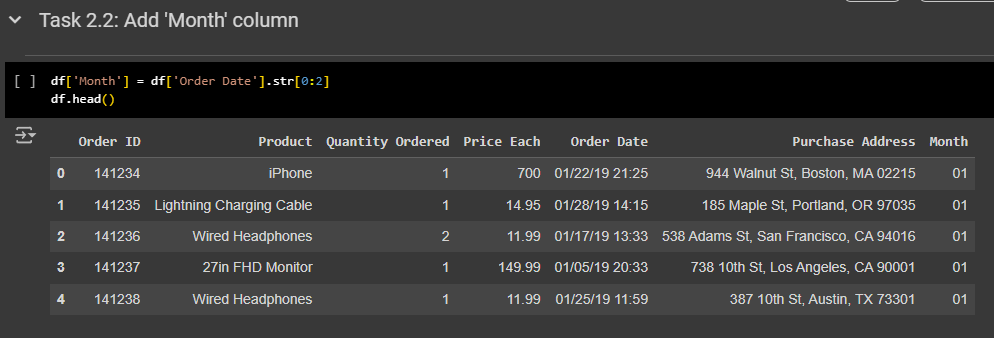


Hình 3‑7: Bộ dữ liệu sau khi gộp

**3.2.2 Làm sạch và tiền xử lý dữ liệu**

1. **Khởi tạo cột Month(Tháng)**

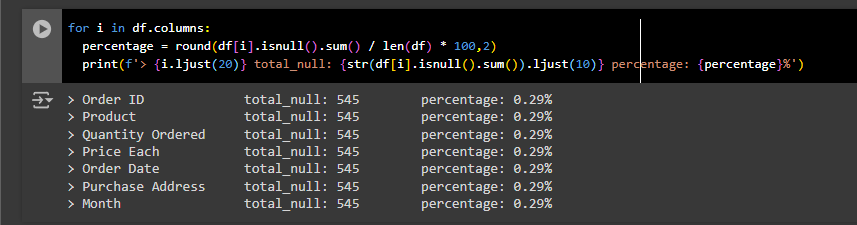
Khởi tạo cột Month(Tháng) để việc phân tích dữ liệu trở nên hiểu quả hơn.



Hình 3‑8: Thêm cột Month

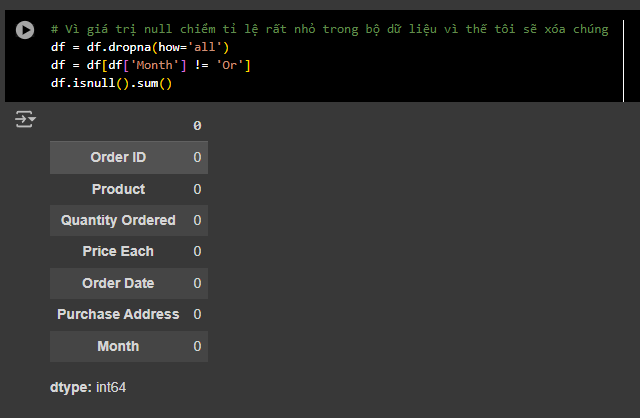
1. **Loại bỏ các giá trị “NaN” và “Or”**

Kiểm tra dữ liệu có tồn tại dữ liệu có bị thiếu hay không qua phương thức trả về hàm isnull() giúp kiểm tra các giá trị trong cột có bị lỗi hay thiếu.



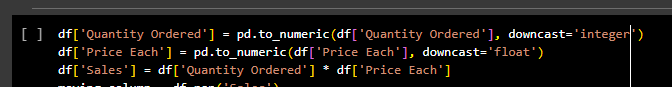
Hình 3‑9: Kiểm tra giá trị null trong dữ liệu

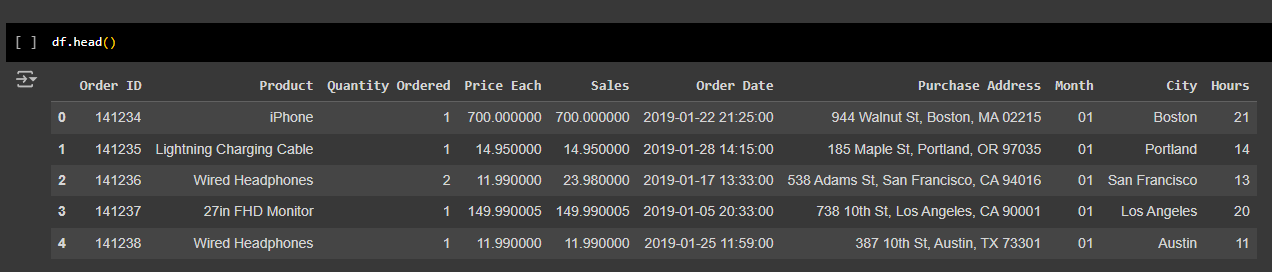
Sau khi kiểm tra dữ liệu thì có thể thấy dữ liệu mỗi cột đều thiếu rất ít dữ liệu khoảng 0.29% nên ta sẽ loại bỏ các dữ liệu bị thiếu đấy bằng hàm dropna().



Hình 3‑10: Loại bỏ các dữ liệu thiếu

1. **Khởi tạo các thuộc tính bổ sung để trực quan**





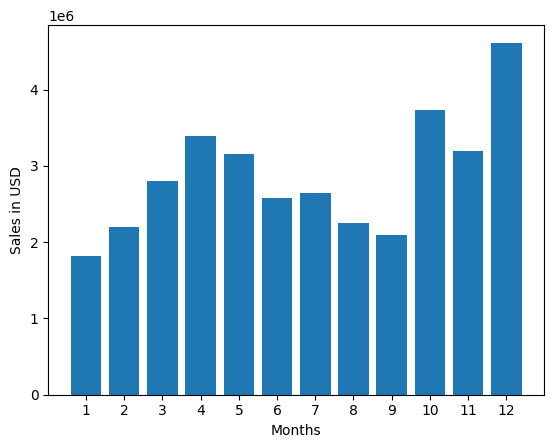
Hình 3‑11: Các thuộc tính bổ sung

Sau khi bổ sung các cột cần thiết cho bộ dữ liệu tiến hành trực quan hóa dữ liệu theo xu hướng để hiểu rõ hơn về bộ dữ liệu mà các doanh nghiệp cần hiểu để dựa vào đó mà phát triển.

# CHƯƠNG 4: Thực nghiệm mô hình

**4.1 Phân tích và trực quan hóa dữ liệu**

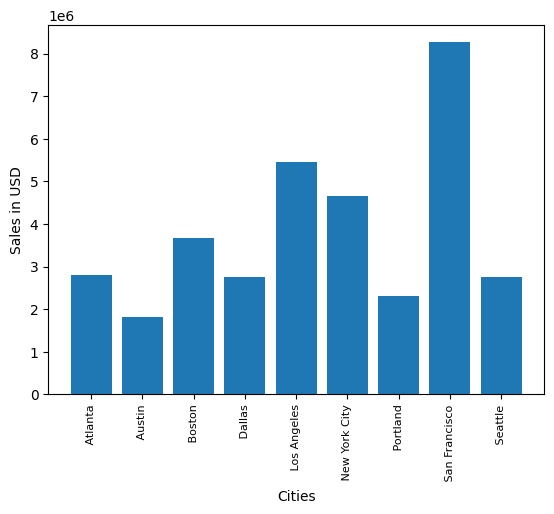
**4.1.1 Tháng nào có doanh số tốt nhất? Doanh số tháng đó là bao nhiêu?**



Hình 4‑1: Biểu đồ cột doanh thu theo Tháng

Theo như biểu cột trên ta có thể thấy doanh trăng trưởng cao từ tháng 10 đến tháng 12 trở đi và đặc biệt tháng 12 doanh thu cao nhất với 4.6 triệu USD.

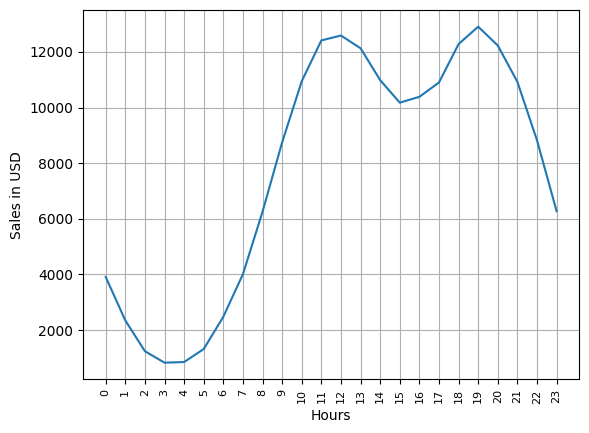
**4.1.2 Thành phố nào có doanh số cao nhất?**



Hình 4‑2: Biểu đồ cột doanh thu theo Thành phố

Theo như biểu đồ trên cho thấy rằng Thành phố San Francisco chiếm doanh thu lớn nhất và gần gấp đối so với Thành phố đứng thứ hai Los Angeles với tổng doanh thu về thị trường thương mai điện tử là hơn 8 triệu USD.

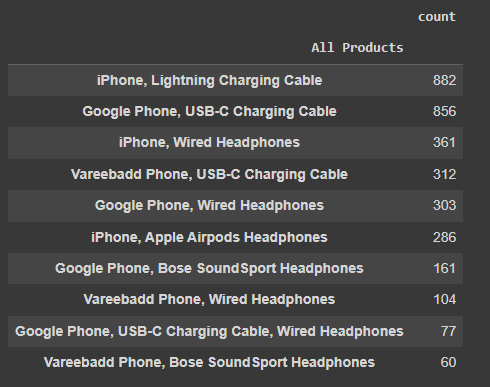
**4.1.3 Doanh nghiệp cần chiếu quảng cáo vào khung thời gian nào để tăng khả năng mua hàng của khách hàng?**



Hình 4‑3: Biểu đồ đường về thời gian mua sắm của khách hàng

Dựa trên biểu đồ khoảng thời gian ít mua sắm của khách hàng vào tối khuya là dĩ nhiên vì khoảng thời gian đa số mọi người trong giấc ngủ, về khoảng thời gian mua sắm nhiều nhất của khách hàng vào khung giờ từ 11h-12h và từ 18h-20h là vào giờ nghỉ trưa và kết thúc giờ hành chính cho thấy rằng ta nên tận dụng khoảng thời gian trên để phát quảng cao nhằm tăng cao khách hàng và doanh thu.

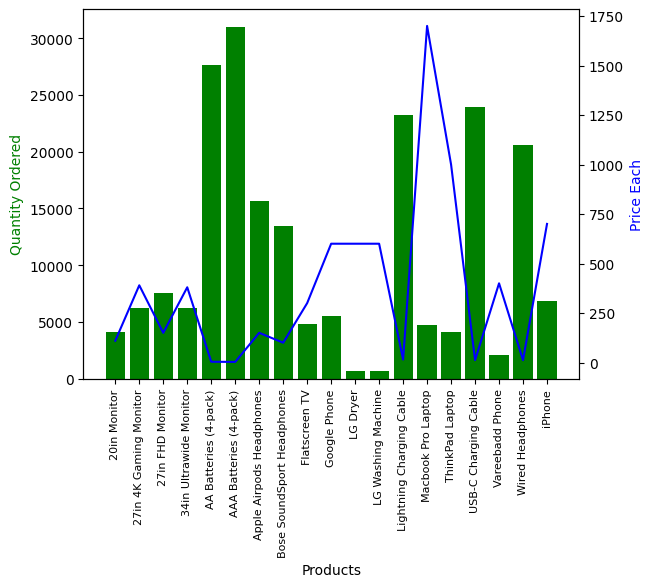
**4.1.4 Những sản phẩm nào thường được bán cùng nhau?**



Hình 4‑4: Các sản phẩm thường được bán cùng nhau

Có thể thấy trên dữ liệu có hơn 1500 đơn hàng mua Iphone thì ước tính sẽ có hơn 50% trong đơn hàng có cáp sặc và 25% có tai nghe không dây và 25% còn lại là tai nghe Airpod . Có thể thấy tổng quan không chỉ khách hàng mua Iphone thì sẽ có 50% mua cáp sặc mà đa số các điện thoại khác cũng chiếm đa số khi mua điện thoại thì khách hàng sẽ mua kèm cáp sặc. Dựa trên yếu tố đó mà doanh nghiệp có thể biết được những gì cần tư vấn cho khách hàng và nhằm tăng lợi nhuận từ yếu tố đó.

**4.1.5** **Sản phẩm nào được bán nhiều nhất? Giả thiết của bạn về lý do sản phẩm này được bán nhiều nhất là gì?**



Hình 4‑5: Biểu đồ cột và đường thể hiện số lượng mua và giá sản phẩm

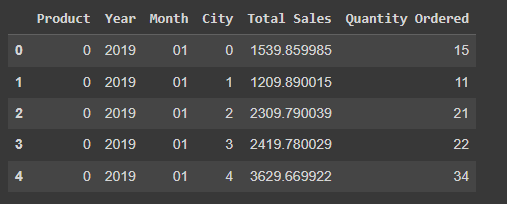
Dựa trên biểu đồ trên thấy rằng số lượng mua pin AA và AAA chiếm nhất nhiều do có thể thấy được là về giá thành khá rẻ khiến khách hàng sẽ ưu chuộng các sản phẩm này và các sẩn phẩm như cáp sạc và tai nghe cũng được mua rất với giá thành tương tư.

## **4.2 Dự đoán (prediction)**

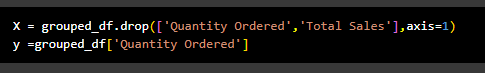
Lựa chọn phương pháp Hồi quy tuyến tính và Rừng ngẫu nhiên, Ridge Regression, Lasso , Catboost dùng kiểm tra độ phù hợp mô hình và để dự đoán bộ dữ liệu doanh số thương mại điện tử.

**4.2.1 Lấy mẫu dữ liệu**

Đưa bộ dữ liệu ban đầu đã qua tiền xử lý và rút ngắn các thuộc tính vào mô hình dự đoán.

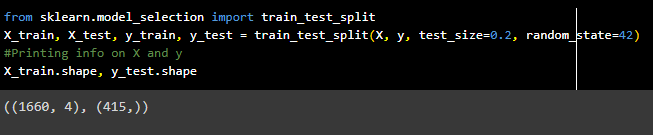


Hình 4‑6: Bộ dữ liệu qua xử lý và đã được rút ngắn các thuộc tính cần thiết



Hình 4‑7: Lựa chọn thuộc tính

Lựa chọn các thuộc tính của X và y trong tập dữ liệu. Với X là các thuộc tính Product(Sản phẩm), Year(Năm mua sản phẩm), Month(Tháng mua sản phẩm), City(Thành phố) và với y giá trị dự đoán là Quantity Ordered (Số lượng sản phẩm)



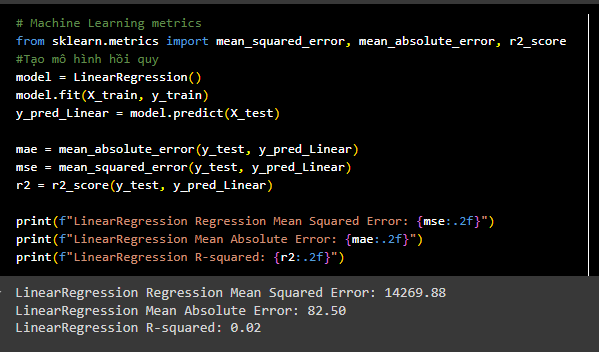
Hình 4‑8: Chia tập dữ liệu

Chia dữ liệu theo tỷ lệ train và test là 80% và 20%

**4.2.2 Lựa chon mô hình phù hợp để dự đoán:**

Tiến hành khởi tạo các mô hình và kiểm tra độ phù hợp bộ dữ liệu với mô hình

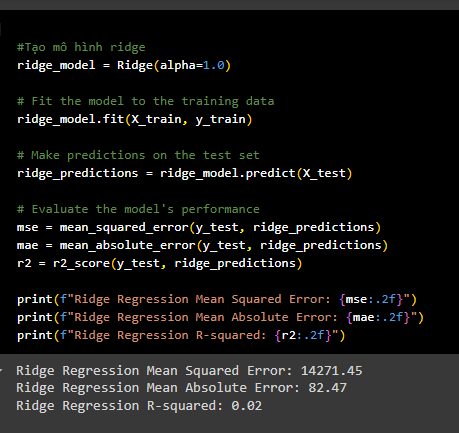
1. **Mô hình Linear Regression**



Hình 4‑9: Mô hình Linear Regression

Thông qua đánh giá các chỉ số của mô hình thì về chỉ số R-squared rất thấp cho thấy bộ dữ liệu không phù với mô hình nên ta sẽ không sử dụng mô hình này trong bài toán dự đoán.

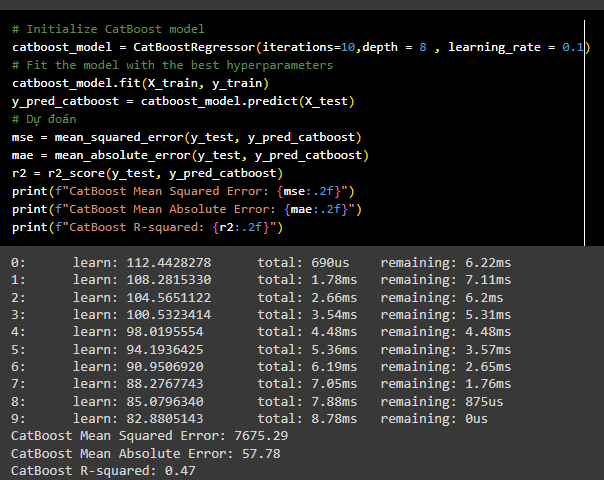
1. **Mô hình Ridge Regression**



Hình 4‑10: Mô hình Ridge Regression

Cũng giống như mô hình Linear thì về chỉ số R-squared rất thấp cho thấy bộ dữ liệu không phù với mô hình nên ta sẽ không sử dụng mô hình này trong bài toán dự đoán.

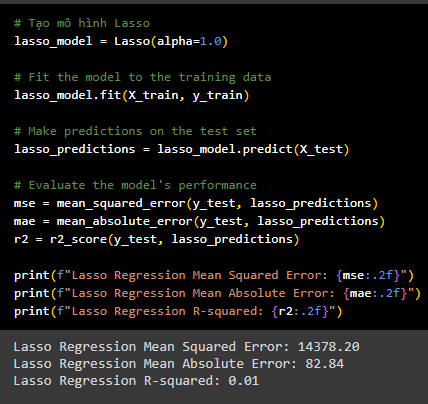
1. **Mô hình Lasso**



Hình 4‑11: Mô hình Lasso

Với mô hình Lasso thì về độ đánh giá có thể cái thiện hơn nhưng vẫn chưa đủ tốt để thực hiện với bài toán dự đoán .

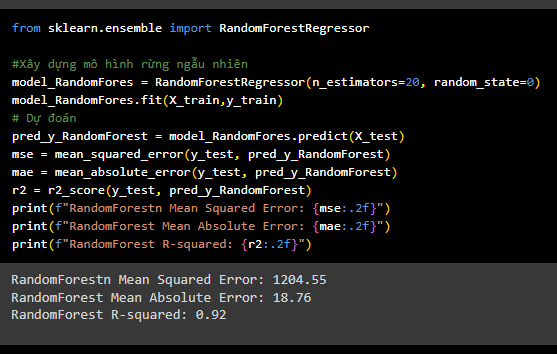
1. **Mô hình Catboost**



Hình 4‑12: Mô hình Catboost

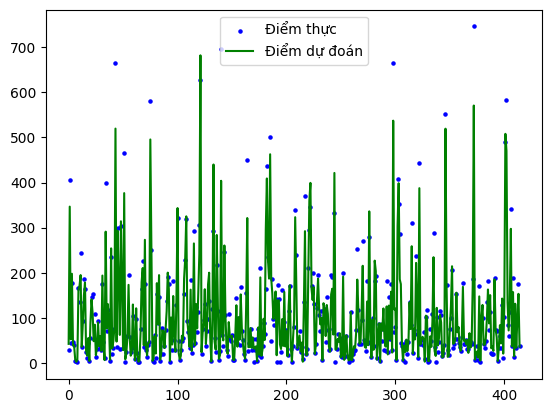
Cũng giống như với đánh giá 2 mô hình đầu tiên ta sẽ không sử dụng mô hình này cho bài toán toán dự đoán.

1. **Mô hình Randoom Forest**



Hình 4‑13: Mô hình Randoom Forest

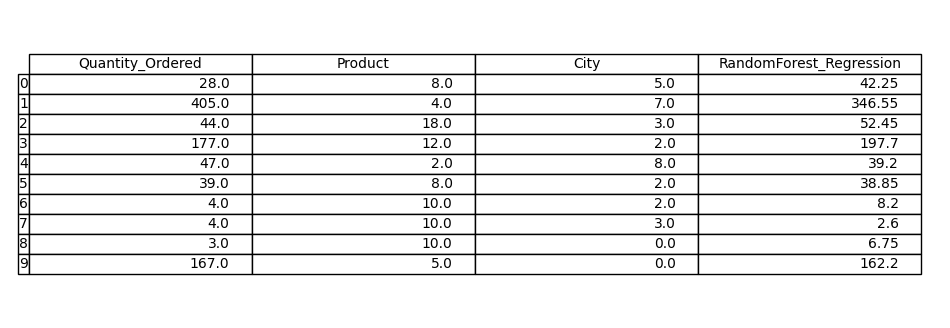
Với mô hình Randoom Forest thì lại mang kết quả tốt nhất trong các số mô hình trên với chỉ số R-squared rất cao tới 0.92 cho thấy độ phù hợp với dữ liệu với mô hình là vô cùng tốt. Nên ta sẽ sử dựng mô hình này cho bài toán dự đoán.



Hình 4‑14: Biểu đồ trực quan

**4.2.3 Tiến hành dự báo:**

Tiến hành áp dụng Mô hình Randoom Forest cho tập dữ liêu. Lựa chọn 10 biến dữ liệu ngẫu nhiên trong tập test để dự đoán



Hình 4‑15: Kết quả dự đoán

Kết quả của mô hình dự đoán cho thấy các giá trị dự đoán không quá chênh lệch và có độ chính xác cao.

# CHƯƠNG 5: Kết luận và hướng phát triển

**5.1. Kết quả đạt được**

* **Mô hình dự đoán hiệu quả**: Đã phát triển thành công một mô hình dự đoán số lượng sản phẩm trên thị trường thương mại điện tử với độ chính xác cao. Các thuật toán như CatBoost, Ridge Regression, và Lasso Regression đã được áp dụng và so sánh.
* **Phân tích yếu tố ảnh hưởng**: Mô hình đã giúp xác định được các yếu tố quan trọng nhất ảnh hưởng đến số lượng sản phẩm, chẳng hạn như giá bán, số lượng đánh giá, đánh giá trung bình của khách hàng, và ngày phát hành sản phẩm.
* **Hiệu năng mô hình**: Mô hình Randoom Forest đã cho thấy hiệu năng vượt trội với khả năng xử lý tốt dữ liệu không cân bằng và các biến phân loại. Mô hình đã đạt được các chỉ số đánh giá (như RMSE, MAE) thỏa đáng trên tập dữ liệu kiểm tra.
* **Ứng dụng thực tế**: Kết quả của mô hình có thể được ứng dụng trong việc tối ưu hóa chiến lược bán hàng, quản lý kho hàng, và dự báo nhu cầu thị trường, giúp các nhà bán lẻ và doanh nghiệp đưa ra quyết định dựa trên dữ liệu một cách chính xác hơn.

**5.2. Hạn chế của đề tài**

* **Chất lượng dữ liệu**: Mặc dù mô hình đã đạt được kết quả tốt, chất lượng dữ liệu đầu vào vẫn còn nhiều hạn chế. Các dữ liệu thiếu sót, không đầy đủ hoặc không đồng nhất có thể ảnh hưởng đến độ chính xác của mô hình.
* **Mô hình với dữ liệu**: Do chưa tương thích được nhiều mô hình nên bộ dữ liệu khó có thể có thể gặp khó khăn trong việc áp dụng trên các tập dữ liệu mới hoặc các thị trường khác nhau mà có những đặc thù riêng

**5.3. Hướng phát triển**

* **Cải thiện chất lượng dữ liệu**: Tìm kiếm các phương pháp làm sạch và làm giàu dữ liệu, hoặc thu thập thêm dữ liệu từ nhiều nguồn khác nhau để tăng cường chất lượng đầu vào cho mô hình.
* **Mô hình hóa động**: Phát triển mô hình có khả năng thích nghi với sự thay đổi của thị trường theo thời gian, có thể thông qua việc triển khai các thuật toán học trực tuyến (online learning) hoặc cập nhật định kỳ mô hình.
* **Phát triển mô hình đa chiều**: Nghiên cứu và phát triển các mô hình có thể không chỉ dự đoán số lượng sản phẩm mà còn dự đoán các yếu tố khác như xu hướng giá, mức độ phổ biến của sản phẩm, hoặc các yếu tố rủi ro thị trường.
* **Mở rộng phạm vi nghiên cứu**: Khám phá và áp dụng mô hình vào các ngành hàng hoặc thị trường khác, từ đó kiểm tra tính khả thi và hiệu quả của mô hình trong các bối cảnh khác nhau.

**DANH MỤC TÀI LIỆU THAM KHẢO**

1. Nguyễn Thị Thùy Linh (2005), bài luận tốt nghiệp “Nghiên cứu các thuật toán phân lớp dựa trên cây quyết định”, Trường Đại học Công nghệ, ĐHQGHN.
2. Vi Văn Sơn (2016), luận văn thạc sĩ “Phân cụm thô của dữ liệu tuần tự”, Trường Đại học Công nghệ, ĐHQGHN.
3. García, S., Luengo, J., & Herrera, F. (2015), “Data Preprocessing in Data Mining”, NewYork: Springer.
4. Foster Provos, Tom Fawcett (2013), Data Science for Business, “What you need to knowabout Data Mining and Data-Analytic Thinking”.
5. ElCOM. (2022, October 17). Máy học (Machine Learning) là gì? Ứng dụng công nghệ máy học. https://www.elcom.com.vn/may-hoc-machine-learning-la-gi-ung-dung-cong-nghe-may-hoc-trong-thuc-tien-1666003970
6. TOPONSEEK (2023, May 24) Google Colab là gì? <https://www.toponseek.com/blogs/google-colab/>
7. Bộ dữ liệu Sales-Reporting <https://github.com/boringPpl/Sales-Reporting/tree/main/data>