****

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO CUỐI KỲ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH YẾU TỐ ẢNH HƯỞNG ĐẾN**

**BÁN HÀNG TRONG BIGMART**

**GVHD: ThS. Nguyễn Văn Thành**

**Lớp: DAAN436277\_23\_2\_01**

**Sinh viên thực hiện:** Nhóm 18

**Họ và tên MSSV**

Nguyễn Trương Thành Long 21133053

Nguyễn Duy Hải 21133101

Dương Hoàng Khang 21133046

Nguyễn Thị Thanh Hiền 21133032

*Thành phố Hồ Chí Minh, tháng 05 năm 2024*

**PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

1. **Môn học:** Phân tích dữ liệu
2. **Giảng viên hướng dẫn:** ThS. Nguyễn Văn Thành
3. **Đề tài:** Phân tích yếu tố ảnh hưởng đến bán hàng trong bigmart
4. **Danh sách nhóm 18:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** | **Kí tên** |
| 1 | Nguyễn Trương Thành Long | 21133053 |  |
| 2 | Nguyễn Duy Hải | 21133101 |  |
| 3 | Dương Hoàng Khang | 21133046 |  |
| 4 | Nguyễn Thị Thanh Hiền | 21133032 |  |

**Nhận xét của giảng viên:**

**Điểm:**

**Thành phố Hồ Chí Minh, tháng 05 năm 2024**

*Giáo viên hướng dẫn ký tên*

*(Ký & ghi rõ họ tên)*

**BẢNG PHÂN CÔNG NHIỆM VỤ**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên** | **Nhiệm vụ** | **Tỉ lệ hoàn thành (%)** |
| 1 | Nguyễn Trương Thành Long | Phân tích dữ liệu, áp dụng mô hình Decision tree | 100 |
| 2 | Nguyễn Duy Hải | Kiểm tra tập dữ liệu và tiền xử lý dữ liệu, tối ưu mô hình áp dụng Grid search | 100 |
| 3 | Dương Hoàng Khang | Phân tích dữ liệu, áp dụng mô hình Random forest | 100 |
| 4 | Nguyễn Thị Thanh Hiền | Phân tích dữ liệu, áp dụng mô hình Linear Regression | 100 |

**LỜI CẢM ƠN**

Đề tài này là sản phẩm của quá trình học tập, làm việc nhóm. Để có thể hoàn thành đề tài này chúng em đã nhận được rất nhiều sự hỗ trợ từ thầy và các bạn. Do đó nhóm chúng em xin được phép gửi lời cảm ơn chân thành đến:

1. Thầy Nguyễn Văn Thành, giảng viên Khoa Công nghệ thông tin Đại học Sư Phạm Kỹ Thuật thành phố Hồ Chí Minh đã tận tình hỗ trợ nhóm chúng em trong việc thực hiện đề tài này.

2. Các bạn học cùng lớp đã có những nhận xét, đóng góp về mặc kiến thức lẫn tinh thần cho nhóm.

Một lần nữa, nhóm chúng em xin chân thành cảm ơn!

MỤC LỤC

[CHƯƠNG 1: TỔNG QUAN VỀ DỰ ÁN 1](#_Toc166743889)

[1. Giới thiệu vấn đề 1](#_Toc166743890)

[2. Giới thiệu về tập dữ liệu 1](#_Toc166743891)

[CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU 3](#_Toc166743892)

[1. Kiểm tra dữ liệu 3](#_Toc166743893)

[2. Tiền xử lý dữ liệu 3](#_Toc166743894)

[3. Phân tích dữ liệu 4](#_Toc166743895)

[3.1. Biểu đồ thể hiện mối quan hệ khả năng nhìn thấy của sản phẩm và doanh số bán hàng 5](#_Toc166743896)

[3.2. Biểu đồ thể hiện mối quan hệ giữa giá bán lẻ của sản phẩm và doanh số bán hàng 6](#_Toc166743897)

[3.3. Biểu đồ thể hiện phân phối giá bán đề xuất của sản phẩm 7](#_Toc166743898)

[3.4. Biểu đồ tỉ lệ hàm lượng chất béo của tất cả sản phẩm 8](#_Toc166743899)

[3.5. Biểu đồ doanh số, số lượng theo loại cửa hàng 9](#_Toc166743900)

[3.6. Biểu đồ doanh số trung bình của từng cửa hàng 10](#_Toc166743901)

[4. Kết luận 11](#_Toc166743902)

[CHƯƠNG 3: ÁP DỤNG MÔ HÌNH MACHINE LEARNING VÀ DEEP LEARNING 13](#_Toc166743903)

[1. Chỉ số hồi quy 13](#_Toc166743904)

[1.1. Mean Squared Error 13](#_Toc166743905)

[1.2. Mean Absolute Error 13](#_Toc166743906)

[1.3. Root Mean Square Error 13](#_Toc166743907)

[1.4. Hệ số xác định (R2) 14](#_Toc166743908)

[2. Áp dụng Machine learning vào dự đoán doanh số của BigMart 14](#_Toc166743909)

[2.1. Mã hóa dữ liệu và Correclation 14](#_Toc166743910)

[2.2. Áp dụng machine learning trong trường hợp lựa chọn các biến có ảnh hưởng nhiều đến cột “Item\_Outlet\_Sales” 15](#_Toc166743911)

[2.2.1. Linear Regression 16](#_Toc166743912)

[2.2.2. Decision Tree Regressor 16](#_Toc166743913)

[2.2.3. Random Forest Regressor 16](#_Toc166743914)

[2.3. Áp dụng machine learning trong trường hợp chọn tất cả các biến trong tập dữ liệu 16](#_Toc166743915)

[2.3.1. Linear Regression 17](#_Toc166743916)

[2.3.2. Decision Tree Regressor 17](#_Toc166743917)

[2.3.3. Random Forest Regressor 17](#_Toc166743918)

[2.4. Nhận xét 17](#_Toc166743919)

[2.4.1. Linear Regression 17](#_Toc166743920)

[2.4.2. Decision Tree Regressor 18](#_Toc166743921)

[2.4.3. Random Forest Regressor 18](#_Toc166743922)

[2.5. Tối ưu mô hình áp dụng Grid search 18](#_Toc166743923)

[2.5.1. Linear Regression 18](#_Toc166743924)

[2.5.2. Decision Tree Regressor 18](#_Toc166743925)

[2.5.3. Random Forest Regressor 19](#_Toc166743926)

[2.6. Kết luận 19](#_Toc166743927)

# **CHƯƠNG 1: TỔNG QUAN VỀ DỰ ÁN**

1. **Giới thiệu vấn đề**

Trên con đường khốc liệt của thị trường bán lẻ, việc hiểu rõ những yếu tố nào thực sự thúc đẩy doanh số bán hàng của sản phẩm có thể là yếu tố quyết định giữa sự thành công và thất bại. BigMart, một trong những nhà bán lẻ lớn với mạng lưới cửa hàng rộng khắp, đã thu thập một bộ dữ liệu vô cùng quý giá trong năm 2013. Bộ dữ liệu này không chỉ cung cấp thông tin về các yếu tố sản phẩm như trọng lượng, hàm lượng chất béo, và giá bán lẻ, mà còn về các đặc điểm của cửa hàng như kích thước, địa điểm và loại hình cửa hàng.

Đứng trước một cơ hội để khai thác sâu vào dữ liệu này và tìm hiểu sâu hơn về cách mỗi yếu tố ảnh hưởng đến doanh số bán hàng. Từ sự nổi bật của sản phẩm trên kệ đến chiến lược giá cả và quảng cáo, bộ dữ liệu này mở ra cánh cửa để hiểu rõ hơn về sự liên kết giữa các yếu tố này và hiệu suất của cửa hàng. Phân tích sâu hơn có thể giúp chúng ta dự đoán được không chỉ doanh số bán hàng của mỗi sản phẩm mà còn cơ hội thị trường và giá trị giỏ hàng trung bình.

Bằng cách này, không chỉ có thể nắm bắt được bức tranh tổng quan về hiệu suất của các sản phẩm tại BigMart mà còn có thể đưa ra những chiến lược cụ thể hơn để tối ưu hóa doanh số bán hàng và nắm bắt cơ hội thị trường một cách chính xác.

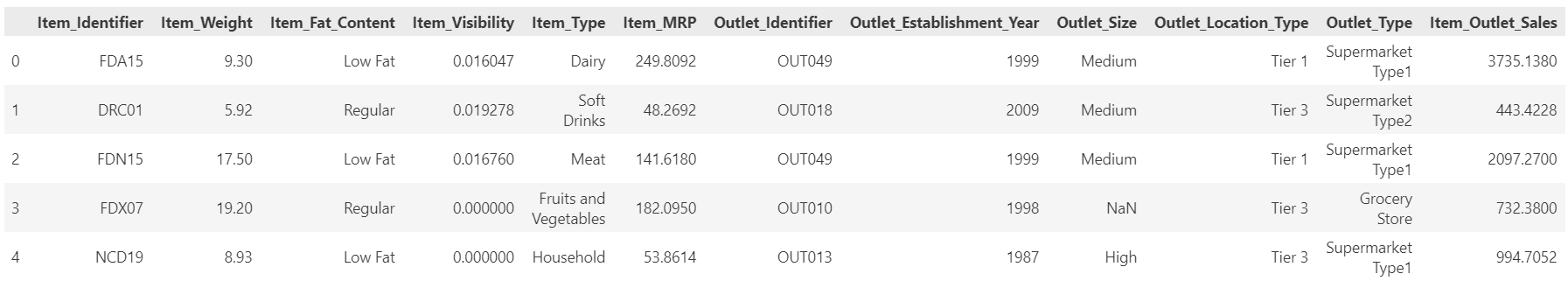
1. **Giới thiệu về tập dữ liệu**

Tập dữ liệu: **data.csv**

Gồm: 12 cột và 14204 dòng

Tập dữ liệu này cung cấp thông tin về các sản phẩm và cửa hàng bán lẻ, cung cấp một cơ sở dữ liệu phong phú để nghiên cứu về yếu tố ảnh hưởng đến doanh số bán hàng

Mục tiêu của việc sử dụng tập dữ liệu này là phân tích và đánh giá các yếu tố như trọng lượng sản phẩm, giá bán lẻ, vị trí cửa hàng, loại cửa hàng, và năm thành lập cửa hàng để hiểu rõ hơn về mối quan hệ giữa các yếu tố này và doanh số bán hàng. Bằng cách phân tích tập dữ liệu này, chúng ta có thể xây dựng các mô hình dự đoán để dự báo doanh số bán hàng của các sản phẩm trong các cửa hàng bán lẻ khác nhau, từ đó giúp cải thiện hiệu suất kinh doanh và chiến lược bán hàng.



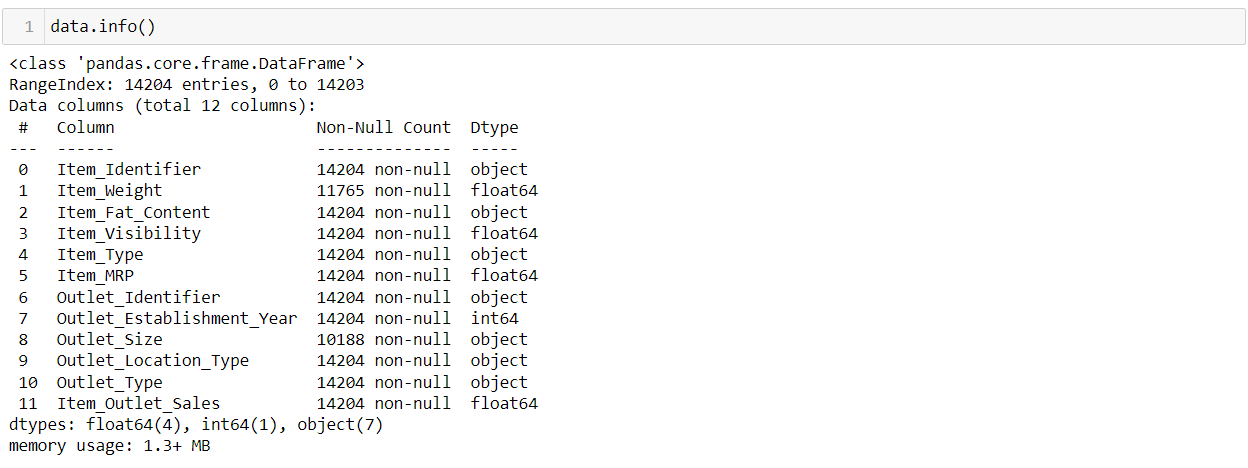
Ý nghĩa các cột trong tập dữ liệu

|  |  |
| --- | --- |
| **Column** | **Description** |
| Item\_Identifier | ID sản phẩm duy nhất |
| Item\_Weight | Trọng lượng của sản phẩm tính bằng kilogam. |
| Item\_Fat\_Content | Chỉ định sản phẩm có ít chất báo hay không |
| Item\_Visibility | Khả năng hiển thị của sản phẩm trong cửa hàng hoặc trực tuyến |
| Item\_Type | Danh mục sản phẩm |
| Item\_MRP | Giá bán lẻ tối đa của sản phẩm |
| Outlet\_identifier | ID của cửa hàng duy nhất |
| Outlet\_Establishment\_Year | Năm cửa hàng được thành lập |
| Outlet\_Size | Quy mô cửa hàng theo diện tích mặt bằng |
| Outlet\_Location\_Type | Loại thành phố nơi của cửa hàng tọa lạc |
| Outlet\_Type | Loại cửa hàng: cửa hàng bách hóa, siêu thị |
| Item\_Outlet\_Sales | Doanh số bán sản phẩm tại cửa hàng |

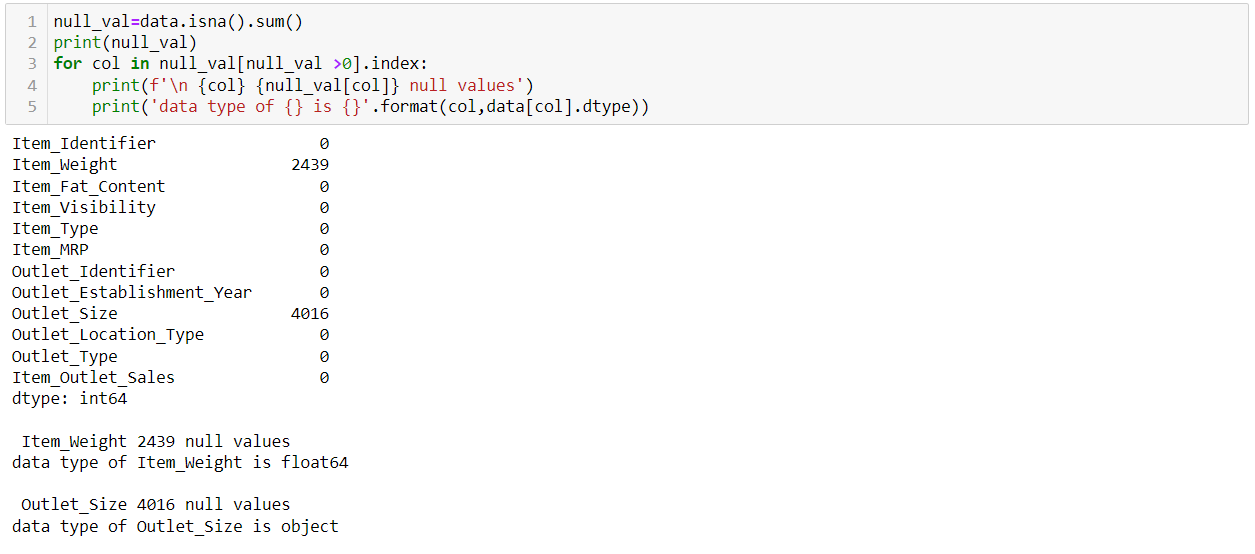
# **CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU**

1. **Kiểm tra dữ liệu**

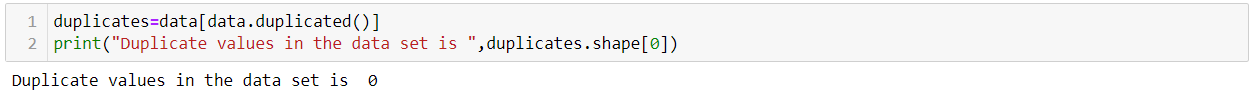
Kiểm tra kiểu giá trị của dữ liệu



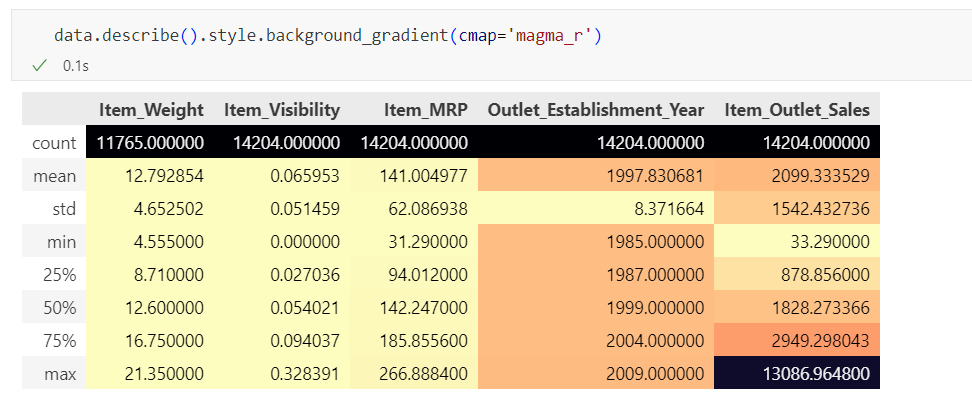
Kiểm tra các dữ liệu Null



Kiểm tra các dữ liệu bị trùng

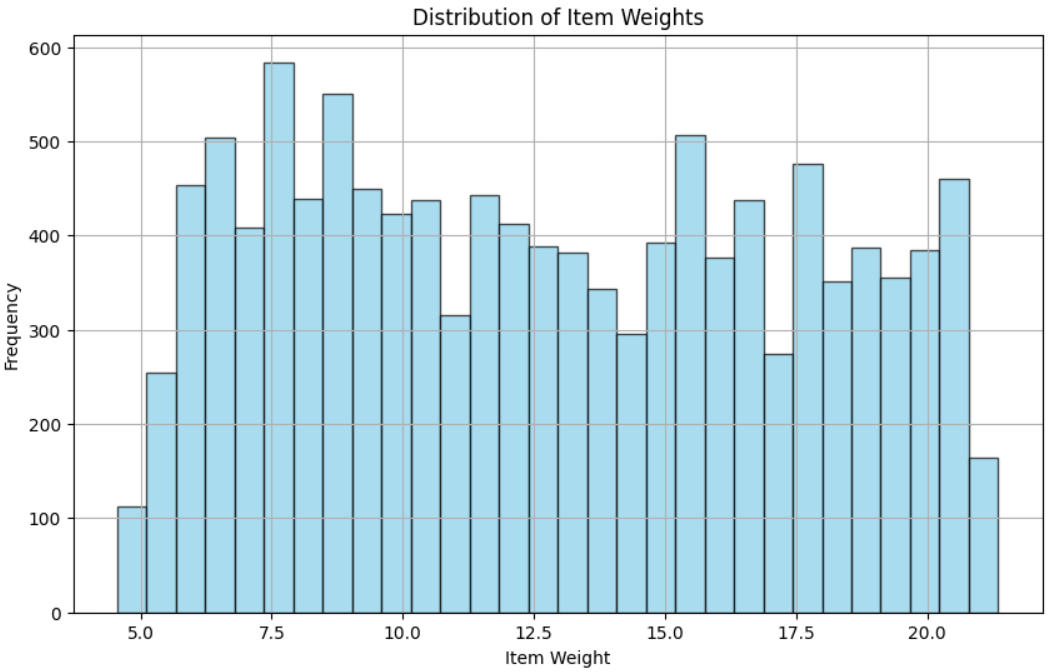


Kiểm tra describe các biến có giá tri là số

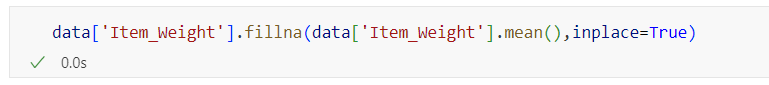


1. **Tiền xử lý dữ liệu**

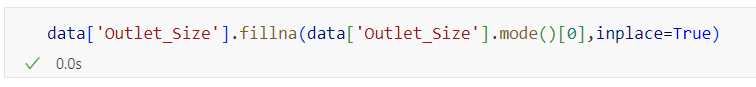
Đối với thuộc tính item weight (kiểu dữ liệu numeric), vẽ biểu đồ phối để biết sự phân bổ các giá trị trong thuộc tính



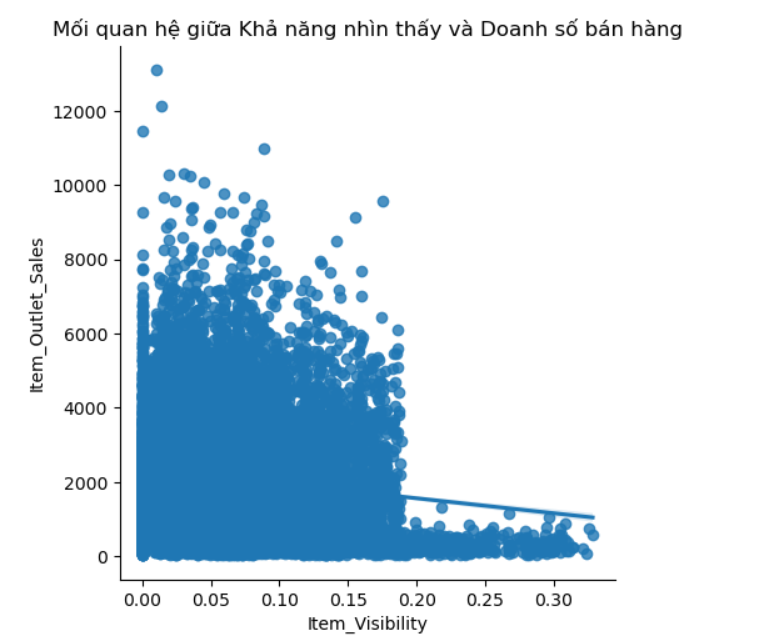
Dựa vào biểu đồ trên ta thấy các thuộc tính phân bổ khá đồng đều, vì thế ta chọn thay thế các giá trị null trong cột bằng giá trị trung bình



Đối với thuộc tính Outlet\_Size có kiểu giá trị là Object ta sẽ thay thế các giá trị null tron cột bằng giá trị phổ biến

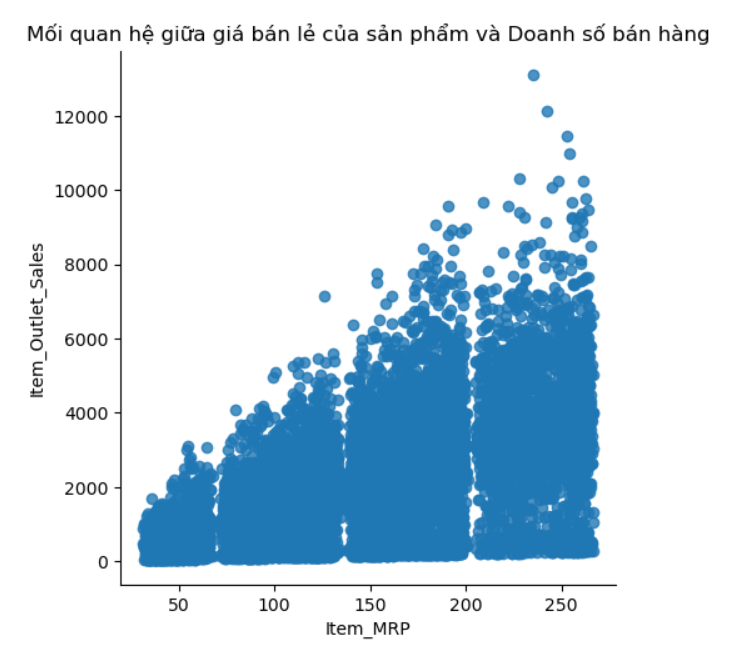


1. **Phân tích dữ liệu**
   1. **Biểu đồ thể hiện mối quan hệ khả năng nhìn thấy của sản phẩm và doanh số bán hàng**



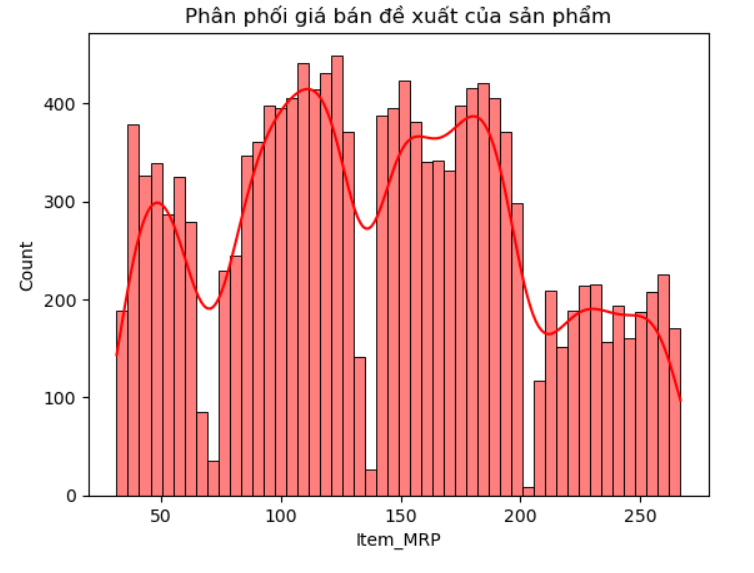
Ta có thể thấy mối quan hệ tuyến tính (nghịch biến) giữa khả năng nhìn thấy của sản phẩm với doanh số. Nhiều sản phẩm có khả năng nhìn thấy là không, nhưng lại có doanh số lớn. Có vẻ như khả năng nhìn thấy không thực sự quan trọng, vì vậy không nên dành quá nhiều công sức vào nó.

* 1. **Biểu đồ thể hiện mối quan hệ giữa giá bán lẻ của sản phẩm và doanh số bán hàng**



Ta có thể thấy mối quan hệ tuyến tính (đồng biến) giữa giá bán đề xuất (Item\_MRP) và doanh số bán hàng (Khi giá bán của sản phẩm tăng thì doanh số cũng tăng) là một phản ánh quan trọng về tầm quan trọng của giá cả đối với hành vi mua hàng của khách hàng. Khi giá sản phẩm tăng, doanh số cũng tăng theo, cho thấy sự đáp ứng tích cực từ phía khách hàng đối với các sản phẩm có giá cao.

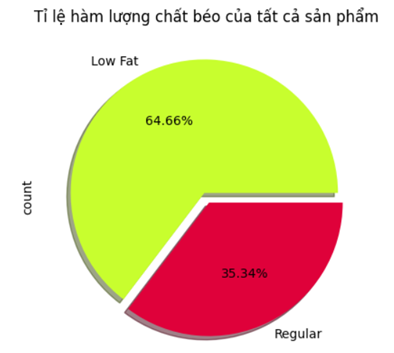
* 1. **Biểu đồ thể hiện phân phối giá bán đề xuất của sản phẩm**



Biểu đồ histogram cho thấy rằng các giá trị Item\_MRP phân bố không đều. Có những khoảng giá trị mà số lượng sản phẩm nhiều hơn so với các khoảng khác. Điều này có thể cho thấy rằng các sản phẩm thường được định giá theo các mức giá nhất định. Có các đỉnh (peaks) trong histogram tại các khoảng giá trị cụ thể như quanh 50, 100, 150, và 200, cho thấy rằng nhiều sản phẩm có giá bán lẻ tối đa rơi vào các khoảng này.

Đường KDE (đường cong màu đỏ): mượt mà hơn và giúp làm nổi bật các đỉnh và đáy trong phân phối Item\_MRP. Đường này cũng cho thấy rằng có nhiều đỉnh và đáy trong phân phối giá bán lẻ tối đa, phản ánh sự biến động trong các mức giá của sản phẩm.

* 1. **Biểu đồ tỉ lệ hàm lượng chất béo của tất cả sản phẩm**



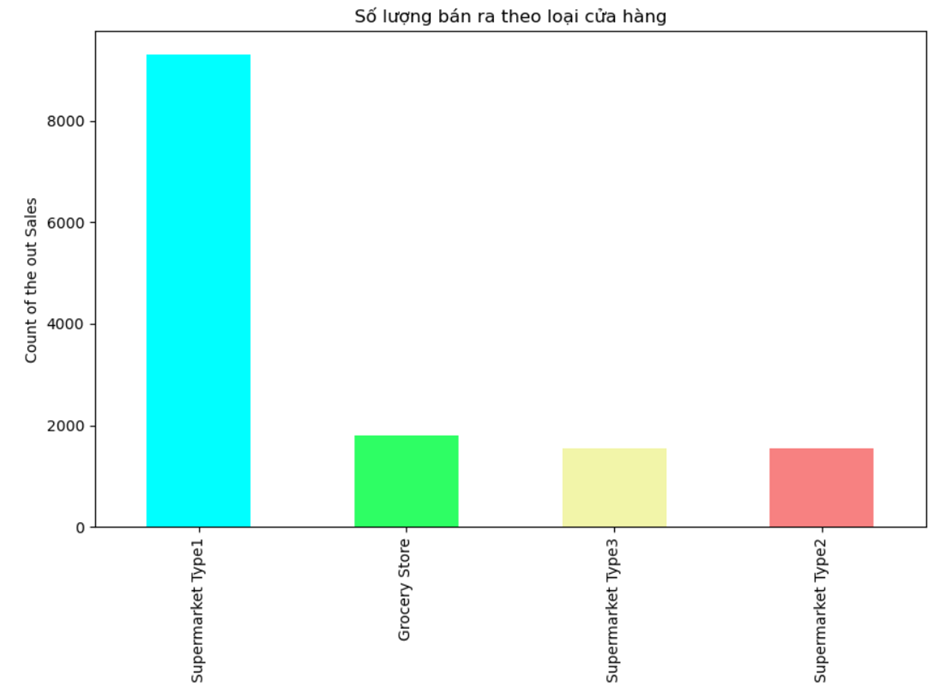
Biểu đồ tròn trên minh họa phân phối phần trăm của hai loại chất béo trong tập dữ liệu: "Low Fat" (64.66%) và "Regular" (35.34%).

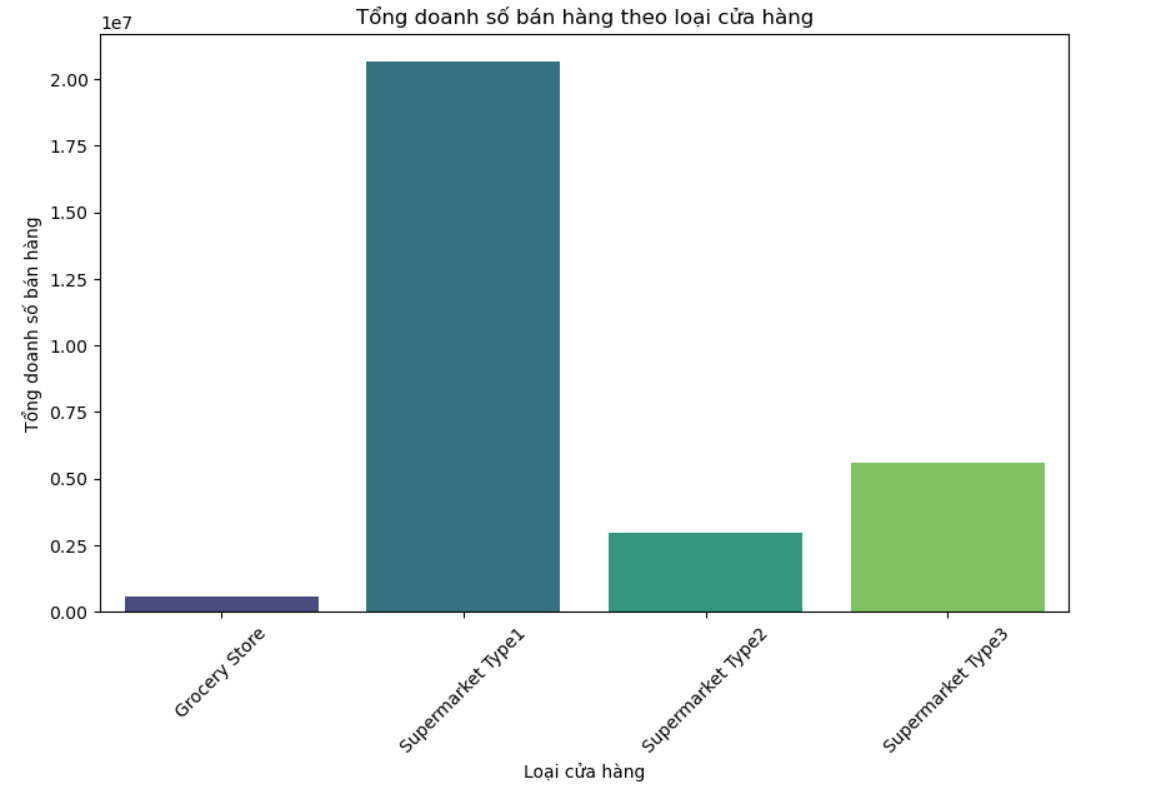
Hiểu về sự ưu tiên của khách hàng: Biểu đồ tròn cho thấy rằng phần lớn sản phẩm trong cửa hàng thuộc loại "Low Fat". Nếu "Low Fat" chiếm tỷ lệ lớn trong số lượng sản phẩm bán ra, điều này có thể cho thấy rằng khách hàng của bạn có xu hướng ưu tiên sản phẩm "Low Fat". Do đó, để tăng doanh thu, bạn có thể tập trung vào việc cung cấp và quảng cáo sản phẩm "Low Fat" để đáp ứng nhu cầu của khách hàng.

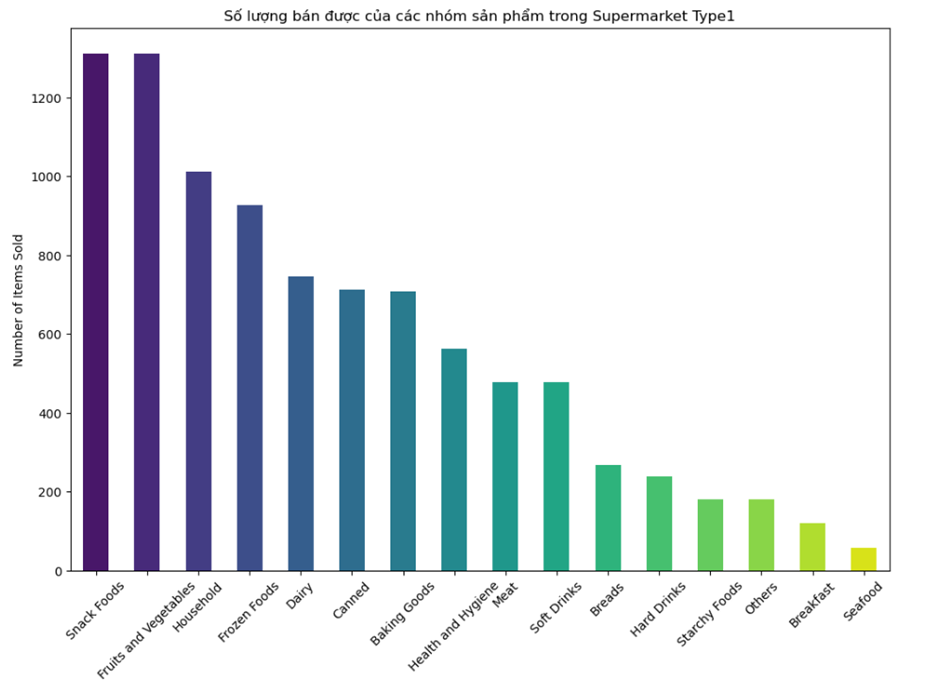
Chiến lược quảng cáo và tiếp thị: Nắm bắt được tỷ lệ phần trăm của các loại sản phẩm có thể giúp bạn tối ưu hóa chiến lược quảng cáo và tiếp thị. Bạn có thể chọn định hình chiến lược tiếp thị của mình để tập trung vào sản phẩm "Low Fat" hoặc phát triển các chương trình khuyến mãi và giảm giá đặc biệt cho sản phẩm này để thu hút khách hàng.

Tối ưu hóa sản phẩm và giá cả: Hiểu rõ phân phối của sản phẩm giữa các nhóm có thể giúp bạn điều chỉnh dịch vụ và sản phẩm của mình. Nếu "Regular" chiếm một phần nhỏ trong tập dữ liệu nhưng có doanh thu cao, bạn có thể xem xét việc tăng giá hoặc cải thiện chất lượng để tối ưu hóa doanh thu từ loại sản phẩm này.

* 1. **Biểu đồ doanh số, số lượng theo loại cửa hàng**

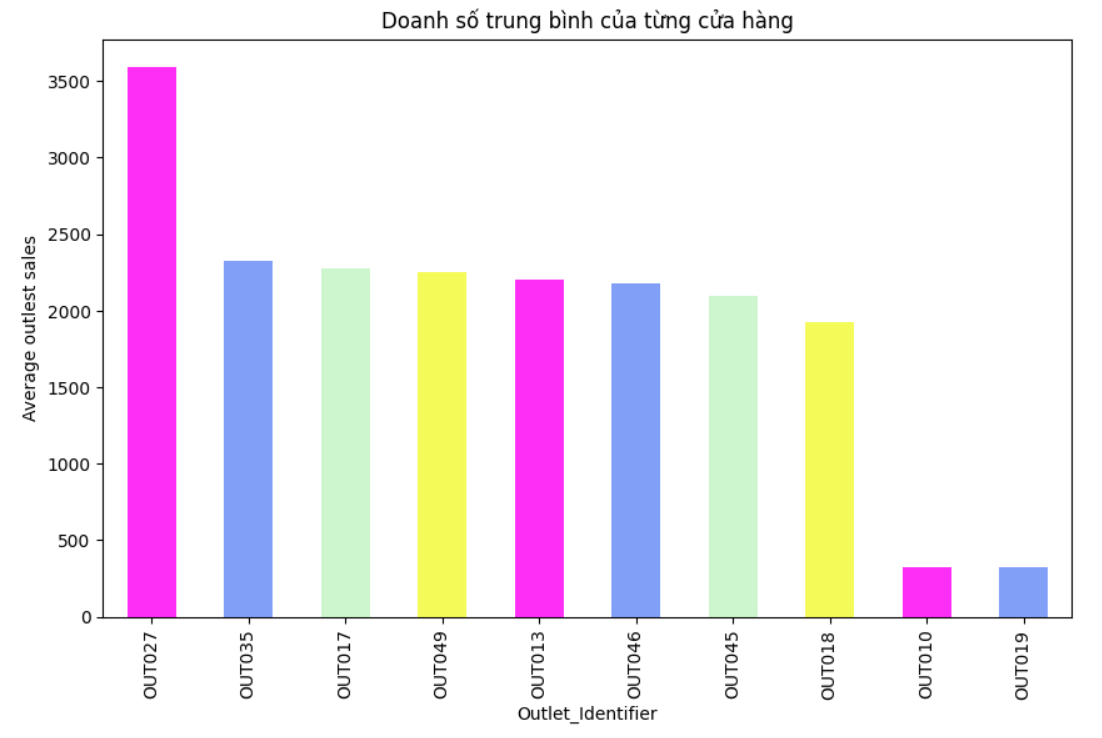






So sánh doanh số bán hàng giữa các loại cửa hàng: Biểu đồ này giúp so sánh doanh số bán hàng giữa các loại cửa hàng khác nhau. Trong đó, có thể thấy là Supermarket Type1 có số lượng bán và doanh thu nhiều nhất, Supermarket Type3 có số lượng bán ra thấp nhất và thấy được số lượng bán ra của các nhóm sản phẩm trong Supermarket Type 1.

* 1. **Biểu đồ doanh số trung bình của từng cửa hàng**

****

Biểu đồ thanh trên hiển thị doanh số trung bình của mỗi nhà bán lẻ trong năm qua. Trục tung biểu thị doanh số trung bình, trục hoành là mã cửa hàng.Với doanh số trung bình là khoảng 2.000, với một số cửa hàng có lượng cửa hàng cao hơn đáng kể, như OUT027 với doanh số trung bình hơn 3500. Tuy nhiên lại có 2 cửa hàng có doanh số trung bình khá thấp đó là OUT010 và OUT019, vì vậy cần có những chiến lược để có thể cải thiện doanh số ở các cửa hàng này.

1. **Kết luận**

Từ việc phân tích dữ liệu, chúng ta đã có cái nhìn tổng quan về các yếu tố ảnh hưởng đến doanh số bán hàng trong cửa hàng. Một số điểm chính có thể được rút ra từ các phân tích là:

Khả năng nhìn thấy của sản phẩm không đóng vai trò quan trọng trong quyết định doanh số bán hàng. Biểu đồ thể hiện mối quan hệ âm giữa khả năng nhìn thấy của sản phẩm và doanh số bán hàng, ngụ ý rằng việc tập trung vào các biện pháp để tăng khả năng nhìn thấy có thể không hiệu quả cho việc tăng doanh số.

Giá bán lẻ của sản phẩm có mối quan hệ tích cực với doanh số bán hàng, nhưng không hoàn toàn tuyến tính. Điều này cho thấy cần phải xem xét nhiều yếu tố khác ngoài giá cả khi xây dựng chiến lược tiếp thị và giá cả. Phân phối số lượng bán hàng theo các mức giá và tỉ lệ hàm lượng chất béo của sản phẩm cũng cung cấp thông tin quan trọng về sở thích và ưu tiên của khách hàng.

Điều này có thể được sử dụng để tối ưu hóa sự đa dạng sản phẩm và chiến lược tiếp thị. Sự khác biệt giữa các loại cửa hàng cũng ảnh hưởng đến doanh số bán hàng. Việc hiểu rõ về hiệu suất của từng loại cửa hàng có thể giúp tập trung nguồn lực vào những phần của kinh doanh có hiệu suất cao nhất.

Tóm lại, việc hiểu rõ các yếu tố này và áp dụng chúng vào chiến lược kinh doanh có thể giúp tối ưu hóa doanh số bán hàng và tăng cường hiệu suất kinh doanh.

# **CHƯƠNG 3: ÁP DỤNG MÔ HÌNH MACHINE LEARNING VÀ DEEP LEARNING**

1. **Chỉ số hồi quy**
   1. **Mean Squared Error**

Mean Squared Error (MSE) có lẽ là số liệu phổ biến nhất được sử dụng cho các bài toán hồi quy. Về cơ bản, nó tìm thấy sai số bình phương trung bình giữa các giá trị được dự đoán và thực tế. MSE là thước đo chất lượng của một công cụ ước tính - nó luôn không âm và các giá trị càng gần 0 càng tốt.

trong đó n là số điểm dữ liệu, yᵢ là giá trị quan sát và ŷ ᵢ là giá trị dự đoán.

Trong phân tích hồi quy, vẽ biểu đồ là một cách tự nhiên hơn để xem xu hướng chung của toàn bộ dữ liệu. Đơn giản MSE cho bạn biết mức độ gần của đường hồi quy với một tập hợp các điểm. Nó thực hiện điều này bằng cách lấy khoảng cách từ các điểm đến đường hồi quy (những khoảng cách này là “sai số”) và bình phương chúng. Bình phương là rất quan trọng để giảm độ phức tạp với các dấu hiệu tiêu cực. Nó cũng tạo ra nhiều trọng lượng hơn cho sự khác biệt lớn hơn.

Để giảm thiểu MSE, mô hình có thể chính xác hơn, có nghĩa là mô hình gần với dữ liệu thực tế hơn. Một ví dụ về hồi quy tuyến tính sử dụng phương pháp này là - phương pháp bình phương nhỏ nhất đánh giá sự phù hợp của mô hình hồi quy tuyến tính với tập dữ liệu hai biến, nhưng giới hạn của nó liên quan đến phân phối dữ liệu đã biết.

MSE càng thấp thì dự báo càng tốt.

* 1. **Mean Absolute Error**

Mean Absolute Error (MAE) đo độ lớn trung bình của các lỗi trong một tập hợp các dự đoán mà không cần xem xét hướng của chúng. Đó là giá trị trung bình trên mẫu thử nghiệm về sự khác biệt tuyệt đối giữa dự đoán và quan sát thực tế, trong đó tất cả các khác biệt riêng lẻ có trọng số bằng nhau.

trong đó n là số điểm dữ liệu, xᵢ là giá trị thực và yᵢ là giá trị dự đoán.

Có thể diễn đạt MAE là tổng hòa của hai thành phần: Bất đồng về số lượng và Bất đồng về phân bổ.

MAE được biết đến là mạnh mẽ hơn đối với các yếu tố ngoại lai so với MSE. Lý do chính là trong MSE bằng cách bình phương các sai số, các giá trị ngoại lai (thường có sai số cao hơn các mẫu khác) được chú ý nhiều hơn và chiếm ưu thế trong sai số cuối cùng và tác động đến các tham số của mô hình.

* 1. **Root Mean Square Error**

Root Mean Square Error (RMSE) hoặc Root Mean Square Deviation (RMSD) là căn bậc hai của mức trung bình của các sai số bình phương. RMSE là độ lệch chuẩn của các phần dư (sai số dự đoán).

Phần dư là thước đo khoảng cách từ các điểm dữ liệu đường hồi quy; RMSE là thước đo mức độ dàn trải của những phần dư này, nói cách khác, nó cho bạn biết mức độ tập trung của dữ liệu xung quanh đường phù hợp nhất.

i: variable i

N: number of non-missing data points

: actual observations time series

: estimated time series

Ảnh hưởng của mỗi lỗi đối với RMSE tỷ lệ với kích thước của lỗi bình phương; do đó các sai số lớn hơn có ảnh hưởng lớn đến RMSE một cách không cân xứng. Do đó, RMSE nhạy cảm với các yếu tố ngoại lai. Sai số bình phương trung bình gốc thường được sử dụng trong khí hậu học, dự báo và phân tích hồi quy để xác minh kết quả thực nghiệm.

Khi các quan sát và dự báo chuẩn hóa được sử dụng làm đầu vào RMSE, có mối quan hệ trực tiếp với hệ số tương quan . Ví dụ, nếu hệ số tương quan là 1, RMSE sẽ bằng 0, bởi vì tất cả các điểm nằm trên đường hồi quy (và do đó không có sai số).

RMSE luôn không âm và giá trị 0 (hầu như không bao giờ đạt được trong thực tế) sẽ chỉ ra sự phù hợp hoàn hảo với dữ liệu. Nói chung, RMSE thấp hơn sẽ tốt hơn RMSE cao hơn.

* 1. **Hệ số xác định (R2)**

Hệ số xác định (coefficient of determination) là một đại lượng trong thống kê được sử dụng để đánh giá mức độ phù hợp của một mô hình hồi quy tuyến tính với dữ liệu. Hệ số này thường được ký hiệu là R².

Hệ số xác định cho biết tỉ lệ phương sai của biến mục tiêu (outcome variable) được giải thích bởi các biến độc lập (independent variables) trong mô hình hồi quy tuyến tính. Nó thường được tính bằng cách so sánh phương sai giữa mô hình hồi quy và phương sai của giá trị trung bình của biến mục tiêu.

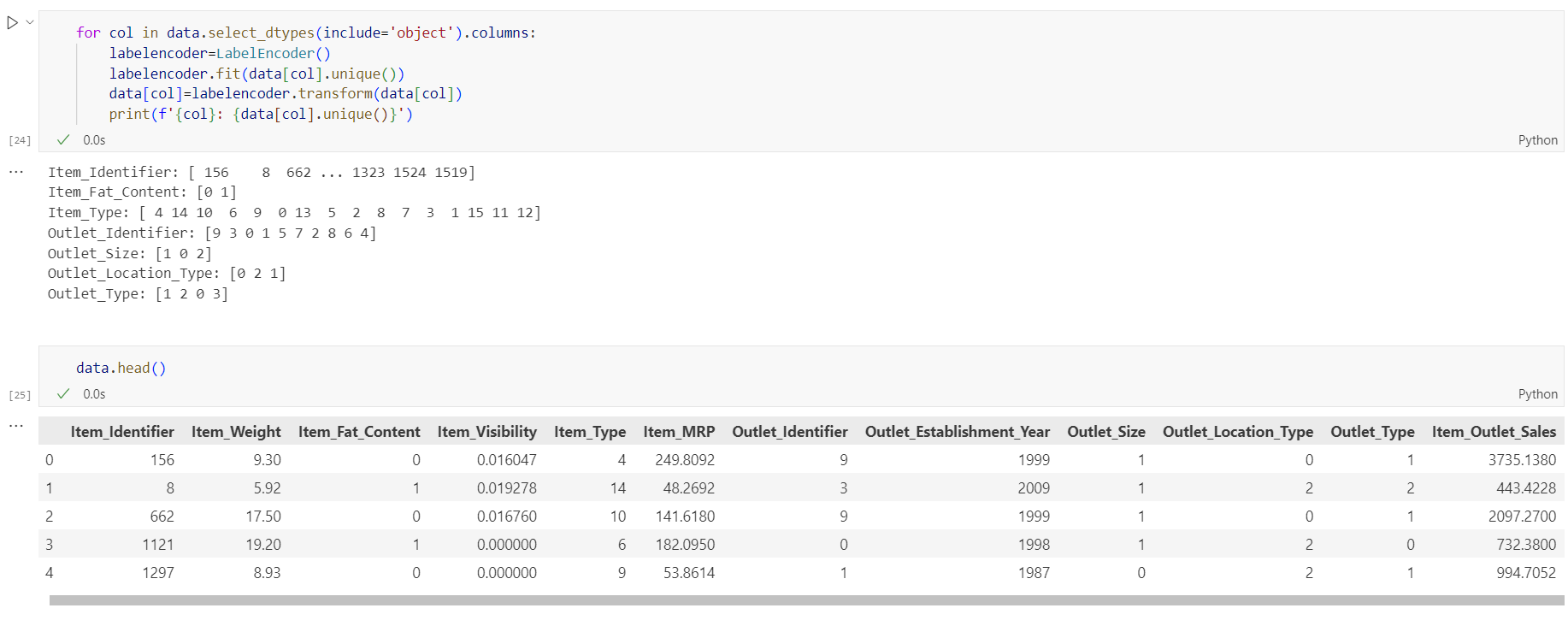
: predicted value of y

: mean value of y

Giá trị của hệ số xác định nằm trong khoảng từ 0 đến 1. Nếu giá trị của hệ số xác định gần bằng 1, tức là mô hình hồi quy tuyến tính giải thích được một phần lớn sự biến thiên của biến mục tiêu. Trong trường hợp giá trị của hệ số xác định gần bằng 0, mô hình hồi quy tuyến tính không giải thích được sự biến thiên của biến mục tiêu và cho thấy mô hình không phù hợp.

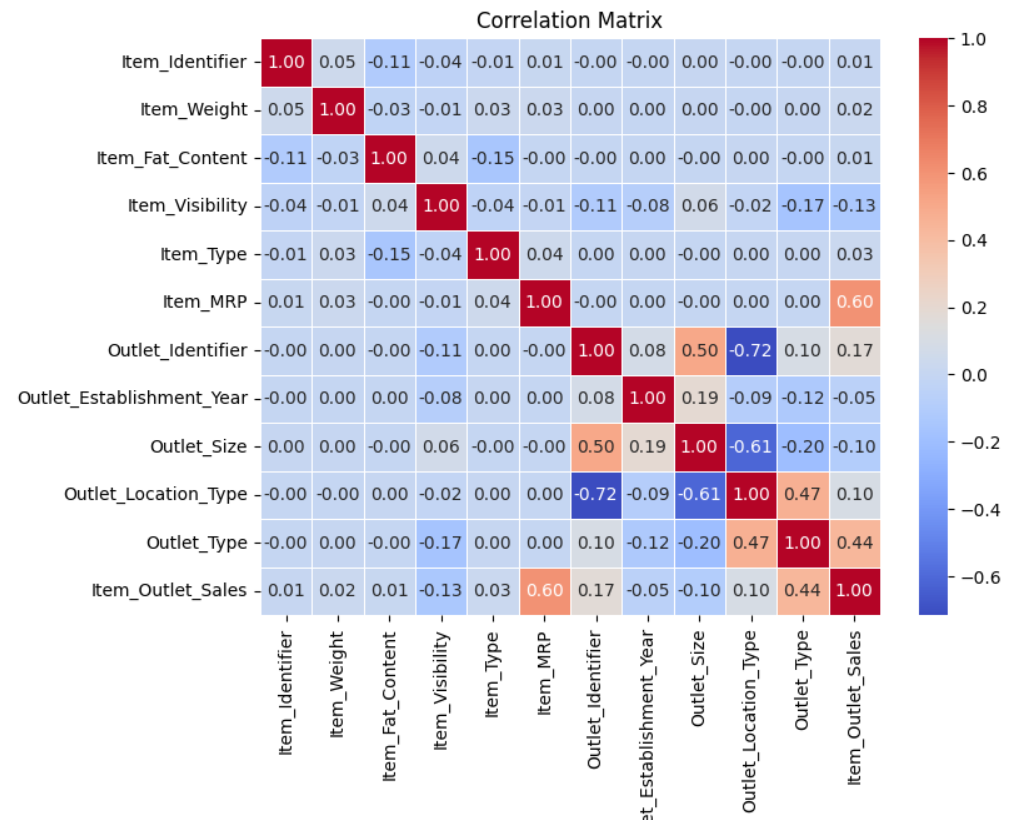
1. **Áp dụng Machine learning vào dự đoán doanh số của BigMart**
   1. **Mã hóa dữ liệu và Correclation**

Mã hóa các dữ liệu dạng chữ thành số



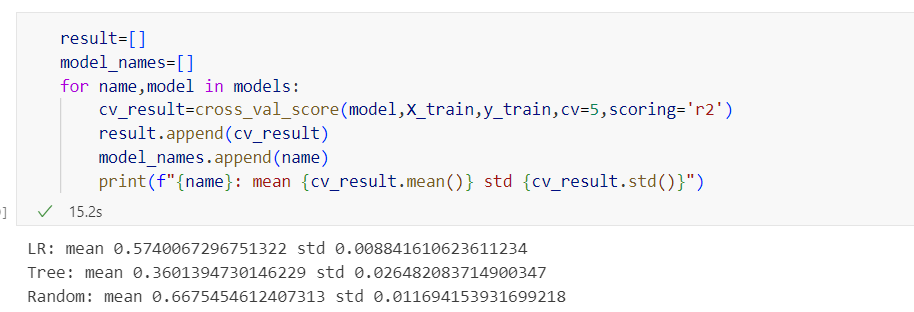
Mối quan hệ tương quan của các biến



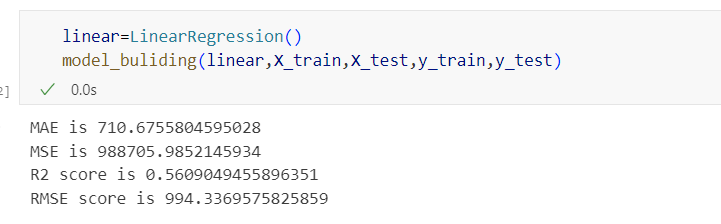
****

* 1. **Áp dụng machine learning trong trường hợp lựa chọn các biến có ảnh hưởng nhiều đến cột “Item\_Outlet\_Sales”**

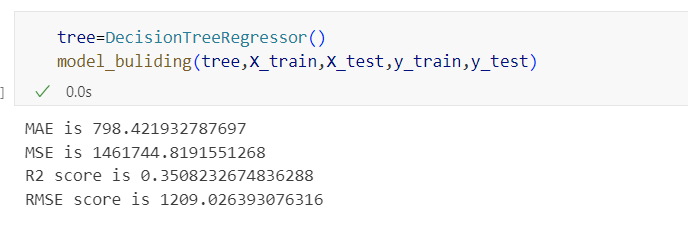
Lựa chọn các biến có ảnh hưởng nhiều đến cột “Item\_Outlet\_Sales” (Item\_MRP, Outlet\_Type, Outlet\_Identifier, Outlet\_Location\_Type, Item\_Type, Item\_Weight, Outlet\_Establishment\_Year, Item\_Visibility).



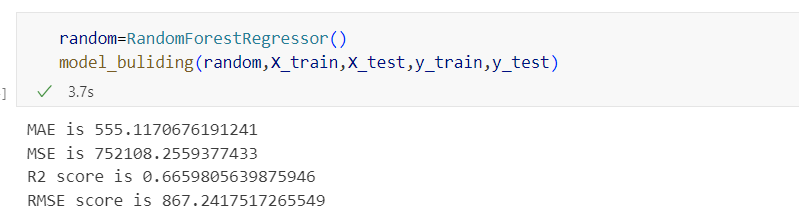
* + 1. **Linear Regression**

****

* + 1. **Decision Tree Regressor**

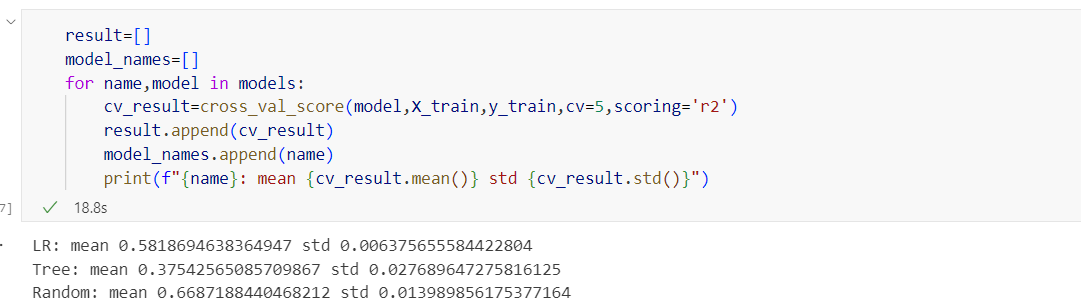
****

* + 1. **Random Forest Regressor**

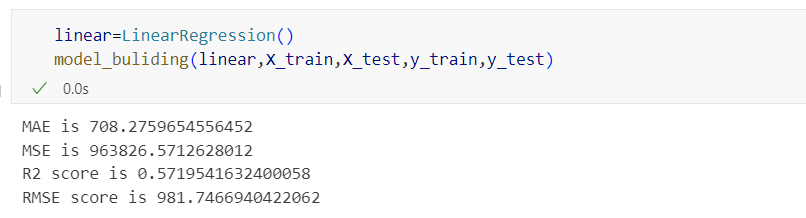
****

* 1. **Áp dụng machine learning trong trường hợp chọn tất cả các biến trong tập dữ liệu**

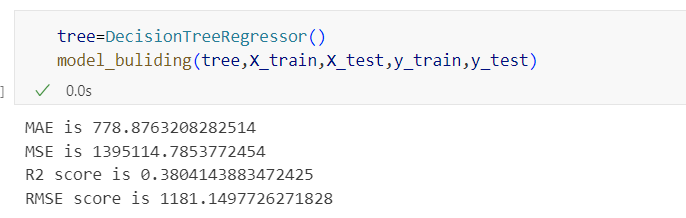
Lựa chọn tất cả các biến trong tập dữ liệu



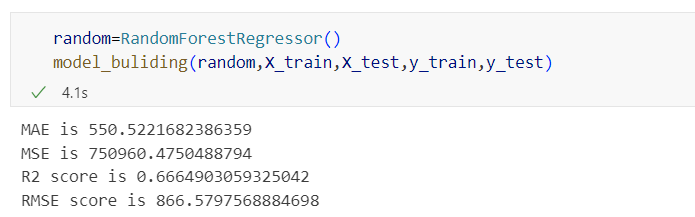
* + 1. **Linear Regression**

****

* + 1. **Decision Tree Regressor**

****

* + 1. **Random Forest Regressor**

****

* 1. **Nhận xét**
     1. **Linear Regression**

Trường hợp chọn tất cả các biến cho mean cao hơn (0.5818 so với 0.5740 của trường hợp chỉ lựa chọn các biến có ảnh hưởng nhiều đến cột “Item\_Outlet\_Sales”) và std thấp hơn (0.0063 so với 0.0088 của trường hợp chỉ lựa chọn các biến có ảnh hưởng nhiều đến cột “Item\_Outlet\_Sales”), tức là hiệu quả mô hình tốt hơn và ổn định hơn khi sử dụng tất cả các biến.

Trường hợp chọn tất cả các biến cho kết quả tốt hơn với MAE (708.2759 so với 710.6756), MSE (963826.5712 so với 988705.9852), R2 (0.5719 so với 0.5609), và RMSE (981.7466 so với 994.3369) đều cải thiện.

* + 1. **Decision Tree Regressor**

Mean và std của trường hợp chọn tất cả các biến (0.3758 và 0.0352) cao hơn so với trường hợp chỉ lựa chọn các biến có ảnh hưởng nhiều đến cột “Item\_Outlet\_Sales (0.3723 và 0.0388), cho thấy rằng việc sử dụng tất cả các biến cải thiện hiệu quả nhưng lại kém ổn định hơn.

Trường hợp chọn tất cả các biến cũng tốt hơn với MAE(778.8763 so với 798.4219), MSE (1395117.7853 so với 1461744.8191), R2 (0.3804 so với 0.3508), và RMSE (1181.1497 so với 1209.0263) đều cải thiện.

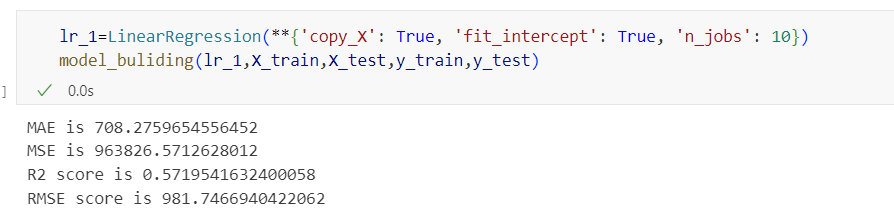
* + 1. **Random Forest Regressor**

Mean của trường hợp chọn tất cả các biến cao hơn (0.6684 so với 0.6682) và std cũng cao hơn (0.0136 so với 0.0130), cho thấy mô hình tốt hơn khi sử dụng tất cả các biến, nhưng sự thay đổi giữa các lần chạy cũng nhiều hơn một chút.

Trường hợp chọn tất cả các biến tốt hơn với MAE (550.5221 so với 555.1170), MSE (750960.4750 so với 752108.2559), và R2 (0.6664 so với 0.6659), và RMSE (866.5797 so với 867.2417).

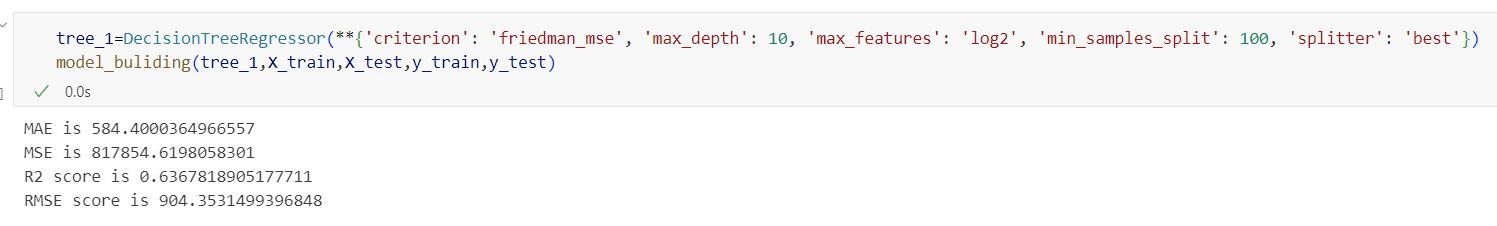
* 1. **Tối ưu mô hình áp dụng Grid search**
     1. **Linear Regression**

****

****

* + 1. **Decision Tree Regressor**

****

****

* + 1. **Random Forest Regressor**

****

****

**Kết luận chung:** Sau khi tối ưu ta thấy các mô hình được cải thiện hiệu suất, các chỉ số đều tốt hơn, đặc biệt là random forest

* 1. **Kết luận**

Việc sử dụng các mô hình như Linear Regression, Decision Tree Regressor và Random Forest Regressor đem lại những kết quả quan trọng về dự đoán và hiểu biết về hiệu suất kinh doanh trong các cửa hàng bán lẻ. Các mô hình này đã chứng minh rằng việc sử dụng tất cả các biến có sẵn trong tập dữ liệu giúp cải thiện hiệu quả của mô hình và giảm sự biến động, từ đó tăng tính ổn định của dự đoán doanh số bán hàng. Qua đó có thể thấy được đối với tập dữ liệu này thì mô hình Random Forest có hiệu quả cao nhất.

Kết hợp với EDA cung cấp cái nhìn tổng quan và sâu rộng về dữ liệu, giúp xác định các yếu tố quan trọng và mối quan hệ giữa chúng. Bằng cách này, chúng ta có thể điều chỉnh mô hình và chọn lọc các biến quan trọng nhất để tối ưu hóa hiệu suất dự đoán. Đồng thời, EDA cũng giúp định hình chiến lược và phương pháp tiếp cận phù hợp với bộ dữ liệu cụ thể, từ đó tăng cơ hội thành công trong việc áp dụng mô hình vào thực tiễn kinh doanh.

Dựa vào những phân tích và kết quả từ EDA và mô hình hóa, có một số giải pháp có thể được đề xuất để tăng doanh thu trong cửa hàng bán lẻ:

Tối ưu hóa chiến lược giá cả: Với việc giá bán lẻ có mối quan hệ tích cực với doanh số bán hàng, cân nhắc việc điều chỉnh giá cả để thu hút khách hàng mà vẫn đảm bảo lợi nhuận.

Bên cạnh đó cần phải có những chiến lược cụ thể để duy trì những mặt hàng đang có doanh số bán cao như là Snack Foods, Fruits and Vegetable và cải thiện những mặt hàng đang có danh số thấp như Breakfast hay Seafood

Đẩy mạnh hơn việc đầu tư và mở thêm cửa hàng loại Supermarket Type1 vì dựa vào những phân tích ở trên có thể thấy sự hiệu quả trong doanh số ở loại Supermarket Type1 vượt lên hẵng các loại cửa hàng khác. Đồng thơi, càng phải chú trọng hơn về những loại cửa hàng có doanh số thấp hơn đặc biệt là Grocery Store loại cửa hàng mà có hiệu quả về doanh số là thấp nhất

Kết hợp EDA và các mô hình dự đoán không chỉ giúp chúng ta hiểu sâu hơn về dữ liệu mà còn tạo ra những giải pháp có thể thực thi và có hiệu quả trong thực tế. Điều này làm nổi bật vai trò quan trọng của việc kết hợp cả hai phương pháp để tối ưu hóa hiệu suất kinh doanh trong các cửa hàng bán lẻ.