**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

****



**đỒ ÁN KẾT THÚC HỌC PHẦN**

**Môn: PHÂN TÍCH DỮ LIỆU**

Mã môn học: 841432

Khoa: Công nghệ thông tin

Chuyên ngành: Hệ thống thông tin

Nhóm nghiên cứu:

Trần Nguyên Lộc – 3120410297

Võ Đăng Quang – 3120410429

**Giảng viên phụ trách:**

**PHAN THÀNH HUẤN**

**Thành phố Hồ Chí Minh, tháng 12 năm 2023**

# **Lời cảm ơn**

Trước hết em xin gửi đến lời cảm ơn chân thành và sâu sắc nhất đến thầy TS. Phan Thành Huấn người trực tiếp hướng dẫn và tận tình chỉ bảo cho nhóm chúng em cho tới khi em hoàn thành đồ án của mình.

Tiếp đến em xin giành lời cảm ơn đến quý thầy cô Trường Đại học Sài Gòn – khoa Công nghệ thông tin đã truyền đạt cho em những kiến thức vô cùng quý báu và bổ ích trong suốt quá trình nghiên cứu và học tập tại trường.

Xin chân thành cảm ơn tới những người bạn đã luôn sát cánh cùng em, những lời động viên, những lần hỗ trợ những lúc cần thiết đã phần nào giúp em hoàn thành đồ án này.

Cuối cùng, em xin cảm ơn đến ba mẹ và người thân trong gia đình đã hỗ trợ và tạo điều kiện thuận lợi cho em trong suốt thời gian học tập và nghiên cứu tại Đại học Sài Gòn.

# **BẢNG ĐÁNH GIÁ MỨC ĐỘ THAM GIA**

|  |  |  |
| --- | --- | --- |
| Danh sách thành viên | MSSV | Mức độ tham gia |
| Trần Nguyên Lộc | 3120410297 | 100% |
| Võ Đăng Quang | 3120410429 | 100% |

Đồ án này trình bày quá trình tìm kiếm, khám phá và phân tích thông tin từ tập dữ liệu tìm được. Trong đó có sử dụng các các thuật toán và phương pháp khai thác dữ liệu. Cuối cùng, quá trình khai thác dữ liệu dẫn dến việc trích xuất tri thức từ dữ liệu và đưa ra quyết định thông minh có lợi cho việc marketing.

# **Mục lục**

[Lời cảm ơn i](#_Toc152512445)

[BẢNG ĐÁNH GIÁ MỨC ĐỘ THAM GIA ii](#_Toc152512446)

[Mục lục iii](#_Toc152512447)

[Danh mục hình ảnh v](#_Toc152512448)

[Chương 1: KHÁI QUÁT ĐỒ ÁN 1](#_Toc152512449)

[1.1. Lí do chọn đề tài 1](#_Toc152512450)

[1.2. Mô tả dữ liệu và cấu trúc dữ liệu 2](#_Toc152512451)

[Chương 2: QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ 4](#_Toc152512452)

[2.1. Các phương pháp phục vụ cho việc phân tích 4](#_Toc152512453)

[2.1.1. Phân tích hồi quy 4](#_Toc152512454)

[2.1.1.1. Hồi quy tuyến tính 4](#_Toc152512455)

[2.1.1.2. Hồi quy phi tuyến 4](#_Toc152512456)

[2.1.2. Support Vector Regression 4](#_Toc152512457)

[2.2. Phân tích bộ dự liệu 5](#_Toc152512458)

[2.3. Tiền xử lí dữ liệu 6](#_Toc152512459)

[2.3.1. Kiểm tra bộ dữ liệu 6](#_Toc152512460)

[2.3.2. Biến đổi và thêm mới dữ liệu 9](#_Toc152512461)

[2.3.3. Bộ dữ liệu tổng thể 11](#_Toc152512462)

[2.4. Thăm dò dữ liệu sau tiền xử lý 12](#_Toc152512463)

[2.4.1. Khái niệm về EDA 13](#_Toc152512464)

[2.4.2. Mục đích sử dụng EDA 13](#_Toc152512465)

[2.4.3. Thực nghiệm EDA đối với dữ liệu 14](#_Toc152512466)

[2.5. Kết quả 18](#_Toc152512467)

[Chương 3: THỰC NGHIỆM VÀ KẾT LUẬN 19](#_Toc152512468)

[3.1. Huấn luyện mô hình 19](#_Toc152512469)

[3.1.1. Linear Regression 2 biến 19](#_Toc152512470)

[3.1.2. Linear Regression 1 biến 23](#_Toc152512471)

[3.1.3. Support Vector Regression 25](#_Toc152512472)

[3.2. Bài toán 1: Ngày giao hàng có thay đổi không khi trở thành khách hàng hạng Bạch kim? 26](#_Toc152512473)

[3.3. Bài toán 2: Dự đoán giá bán lẻ tốt nhất 27](#_Toc152512474)

[3.4. Kết quả 27](#_Toc152512475)

[Tài liệu tham khảo 28](#_Toc152512476)

# **Danh mục hình ảnh**

[Hình 1. 1. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu orders.csv 4](#_Toc150285481)

[Hình 1. 2. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv 4](#_Toc150285482)

[Hình 1. 3. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv 5](#_Toc150285483)

[Hình 1. 4. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu product-supplier.csv 5](#_Toc150285484)

[Hình 1. 5. Kết quả trả về từ câu lệnh isna().sum() đối với bộ dữ liệu product-supplier.csv 5](#_Toc150285485)

[Hình 1. 6. Kết quả trả về từ câu lệnh isnull().sum() đối với bộ dữ liệu product-supplier.csv 6](#_Toc150285486)

[Hình 1. 7. Kiểu dữ liệu ngày và tháng ban đầu 6](#_Toc150285487)

[Hình 1. 8. Kiểu dữ liệu ngày tháng sau khi được biến đổi để phù hợp cho việc phân tích 6](#_Toc150285488)

[Hình 1. 9. Lỗi không thống nhất dữ liệu trên cột Customer Status 7](#_Toc150285489)

[Hình 1. 10. Thông nhất kiểu dữ liệu trên cột Customer Status 7](#_Toc150285490)

[Hình 1. 11. Thêm một cột dữ liệu mới tên là Item Retail Value 7](#_Toc150285491)

[Hình 1. 12. Tạo dataframe mới nhằm thống kê số lượng sản phẩm được bán cũng như giá trị vốn - lãi trung bình của sản phẩm 8](#_Toc150285492)

[Hình 1. 13. Thông tin của bảng dữ liệu được merge lại với nhau 8](#_Toc150285493)

[Hình 1. 14. Dữ liệu của bảng dữ liệu được merge lại với nhau 9](#_Toc150285494)

[Hình 1. 15. Bộ dữ liệu cuối cùng trong quá trình tiền xử lý 9](#_Toc150285495)

# **Chương 1: KHÁI QUÁT ĐỒ ÁN**

## **Lí do chọn đề tài**

Ngày nay, khi cuộc sống chúng ta ngày một phát triển thì việc mua bán, trao đổi hàng hoá ngày càng được chú trọng và quan tâm. Khách hàng ngày càng có nhiều sự lựa chọn, nhiều cơ hội để tìm cho mình một loại sản phẩm phù hợp nhất và giá cả lại phải chăng. Sự hiểu biết về hành vi tiêu dùng là chìa khóa cho một chiến lược marketing thành công cả trong nước và quốc tế. Những hoạt động marketing chủ yếu này sẽ hiệu quả hơn khi được đặt trên cơ sở một sự hiểu biết về hành vi tiêu dùng.

Xuất phát từ nhận thức trên, nhận thấy được tầm quan trọng của sự thay đổi hành vi của người tiêu dùng của trong thời điểm hiện tại, nên nhóm chúng em quyết định lựa chọn đề tài: ”Phân tích các mẫu mua hàng trực tuyến để tìm hiểu các luật kết hợp giữa các mặt hàng” để nhận diện ra mặt hàng nào là tiềm năng, phát hiện những yếu tố quan trọng tác động đến hành vi mua sắm từ đó có những chiến lược Marketing phù hợp.

Mục tiêu của đồ án này sẽ:

Sử dụng dữ liệu về lịch sử mua hàng để phân tích và xác định:

* Mặt hàng nào có tỉ suất lợi nhuận cao nhất?
* Mặt hàng nào bán chạy nhất, mặt hàng nào không bán chạy.
* Mối tương quan giữa giá bán lẻ?

Sử dụng dữ liệu về thời gian để phân tích và xác định:

* Tần suất mua hàng của các thời điểm trong năm.
* Ngày giao hàng có thay đổi không khi trở thành khách hàng hạng Bạch kim?

Sử dụng dữ liệu về lịch sử thanh toán để phân tích và xác định giá bản sỉ và giá bán lẻ để rồi từ đó ta có thể dự đoán giá bán lẻ khi đã biết giá bán sỉ.

## **Mô tả dữ liệu và cấu trúc dữ liệu**

Bộ dữ liệu được tổng hợp từ các nguồn sau Wholesale & Retail Orders Dataset (kaggle.com):

<https://www.kaggle.com/datasets/gabrielsantello/wholesale-and-retail-orders-dataset?select=orders.csv>

<https://www.kaggle.com/datasets/gabrielsantello/wholesale-and-retail-orders-dataset?select=product-supplier.csv>

Đây là những nguồn được đánh giá là có những bộ dữ liệu đáng tin cậy để phục vụ cho việc nghiên cứu hành vi tiêu dùng.

*Bảng 1.4a. Bảng mô tả cấu trúc của bộ dữ liệu orders.csv*

|  |  |
| --- | --- |
| Thuộc tính | Ý nghĩa |
| Customer ID | Mã khách hàng |
| Customer Status | Trạng thái khách hàng thân thiết (Ví dụ: Gold, Slive, Platinum) |
| Date Order was placed | Ngày đặt hàng |
| Delivery Date | Ngày giao hàng |
| Order ID | Mã đặt hàng |
| Product ID | Mã sản phẩm |
| Quantity Ordered | Số lượng sản phẩm đặt |
| Total Retail Price for This Order | Tổng giá bán lẻ cho đơn hàng này |
| Cost Price Per Unit | Giá vốn của mỗi món hàng |

*Bảng 1.4b. Bảng mô tả cấu trúc của bộ dữ liệu product-supplier.csv*

|  |  |
| --- | --- |
| Thuộc tính | Ý nghĩa |
| Product ID | Mã sản phẩm |
| Product Line | Dòng sản phẩm |
| Product Category | Danh mục sản phẩm |
| Product Group | Nhóm sản phẩm |
| Product Name | Tên sản phẩm |
| Supplier Country | Nước phân phối |
| Supplier Name | Tên nhà phân phối |
| Supplier ID | Mã nhà phân phối |

# **Chương 2: QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ**

## **Các phương pháp phục vụ cho việc phân tích**

### **Phân tích hồi quy**

Phân tích hồi quy là phương pháp thống kê được phát triển từ thế kỉ 19 và trải qua một quá trình lịch sử dài. Lịch sử đã chứng kiến sự phát triển liên tục của công nghệ và phần mềm thống kê. Phân tích hồi quy đã trở thành một công cụ mạnh mẽ trong nghiên cứu và ứng dụng thống kê trong nhiều lĩnh vực khác nhau. Hiện tại phân tích hồi quy vẫn là một kĩ thuật quan trọng được sử dụng để khám phá mối quan hệ giữa một biến phụ thuộc và một hay nhiều biến độc lập. Phân tích hồi quy thương được sử dụng để dự đoán, hiểu sự mạnh yếu và hướng của mối quan hệ, và xác định các biến quan trọng trong một tập dữ liệu.

#### **Hồi quy tuyến tính**

Hồi quy tuyến tính là một kĩ thuật phân tích dự đoán giá trị của dữ liệu không xác định bằng cách sử dụng một giá trị dữ liệu liên quan và đã biết khác. Về bản chất, một kỹ thuật hồi quy tuyến tính đơn giản cố gắng vẽ một đồ thị đường giữa hai biến dữ liệu, x và y. Là biến độc lập, x được vẽ dọc theo trục hoành. Các biến độc lập còn được gọi là biến giải thích hoặc biến dự báo. Biến phụ thuộc, y, được vẽ trên trục tung. Bạn cũng có thể tham chiếu các giá trị y như các biến phản hồi hoặc các biến dự báo.

#### **Hồi quy phi tuyến**

Hồi quy phi tuyến là một kỹ thuật phân tích dữ liệu dự đoán giá trị của biến phụ thuộc bằng cách sử dụng biến độc lập

### **Support Vector Regression**

## **Phân tích bộ dự liệu**

Trong đồ án môn học này, chúng em sử dụng bộ dữ liệu trong hoạt động mua bán sỉ lẻ, bộ dữ liệu phân tích gồm 2 file csv tương ứng:

* *Bộ dữ liệu orders.csv* *(gồm 185004 bản ghi và 9 trường thuộc tính):*

Đây là bộ dữ liệu chứa thông tin về lịch sử giao dịch về mua bán sản phẩm đối với khách hàng và người bán. Trong đó, mỗi bản ghi (record) trong bộ dữ liệu tượng trưng cho việc thông tin sản phẩm trong quá trình mua bán như mã sản phẩm, mã đặt hàng, mã khách hàng, mức giá, ngày đặt, ngày giao, số lượng giao dịch, tổng vốn thu được, giá trị mỗi sản phẩm.

Các thuộc tính trong bộ dữ liệu được mô tả bao như sau:

* **Customer ID:** Mã khách hàng.
* **Customer Status:** Trạng thái khách hàng thân thiết.
* **Date Order was placed:** Ngày đặt hàng.
* **Delivery Date:** Ngày giao hàng.
* **Order ID:** Mã đặt hàng.
* **Product ID:** Mã sản phẩm.
* **Quantity Ordered:** Số lượng sản phẩm được đặt hàng.
* **Total Retail Price for This Order:** Tổng giá bán lẻ cho cả đơn hàng.
* **Cost Price Per Unit:** Giá vốn thu được của mỗi món hàng.
* Qua bộ dữ liệu trên, nhóm em có một vài nhận định như sau: Đối với mỗi bản ghi trong bộ dữ liệu chỉ đánh giá cụ thể đối với một sản phẩm thông qua mã sản phẩm và bộ dữ liệu này cũng là một bộ dữ liệu đầy đủ không thiếu hay NULL dòng/trường nào.
* *Bộ dữ liệu product-supplier.csv (gồm 5505 bản ghi và 8 trường thuộc tính):*

Đây là bộ dữ liệu mô tả thông tin chi tiết về các sản phẩm được mua bán.Trong đó, mỗi bản ghi (record) thể hiện thông tin chi tiết về từng thuộc tính của sản phẩm như mã sản phẩm, tên sản phẩm, nhóm sản phẩm, phân loại, dòng sản phẩm, nước phân phối, tên nhà phân phối, mã nhà phân phối.

Các thuộc tính trong bộ dữ liệu được mô tả như sau:

* **Product ID:** Mã sản phẩm
* **Product Line:** Dòng sản phẩm
* **Product Category:** Danh mục sản phẩm
* **Product Group:** Nhóm sản phẩm
* **Product Name:** Tên sản phẩm
* **Supplier Country:** Nước phân phối
* **Supplier Name:** Tên nhà phân phối
* **Supplier ID:** Mã nhà phân phối
* Qua bộ dữ liệu trên, nhóm em có một vài nhận định như sau: Bộ dữ liệu trên mô tả rất đầy đủ thông tin chi tiết của từng đối tượng sản phẩm.

Đánh giá tổng thể cho thấy đây là hai bộ dữ liệu khá đầy đủ cho việc phân tích dữ liệu. Ngoài ra, hướng đi chính cho quá trình phân trình phân tích dữ liệu trong đồ án môn học này sẽ là tìm ra các bộ luật nhằm xác định các mối liên hệ giữa sản phẩm và người dùng.

## **Tiền xử lí dữ liệu**

### **Kiểm tra bộ dữ liệu**

Trước khi tiền xử lý dữ liệu, nhóm chúng em tiến hành kiểm tra bộ dữ liệu một cách tổng quan nhất. Bắt đầu với bộ dữ liệu *orders.csv* trước tiên ta đọc bộ dữ liệu và gán vào biến dataset bằng lệnh ***info()***. Kết quả trả về thu được sẽ là:



Hình 1. 1. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu orders.csv

Điều đó cho thấy kiểu dữ liệu tổng thể của toàn bộ bộ dữ liệu trên. Tiếp sau đó là kiểm tra xem bộ dữ liệu trên có chứa tập Null và NaN hay không bằng cách đếm số lượng của chúng chứa trong bộ dữ liệu bằng lệnh *isna().sum()* và *isnull().sum()*, sau đó kết quả trả về sẽ là:



Hình 1. 2. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv



Hình 1. 3. Kết quả trả về từ câu lệnh isnal().sum() đối với bộ dữ liệu orders.csv

Tiếp sau đó ta vẫn tiến hành kiểm tra bộ dữ liệu như trên đối với bộ dữ liệu *product-supplier.csv* và kết quả thu được như sau:



Hình 1. 4. Kết quả trả về từ câu lệnh info() đối với bộ dữ liệu product-supplier.csv



Hình 1. 5. Kết quả trả về từ câu lệnh isna().sum() đối với bộ dữ liệu product-supplier.csv



Hình 1. 6. Kết quả trả về từ câu lệnh isnull().sum() đối với bộ dữ liệu product-supplier.csv

Kết quả trả về từ việc kiểm tra 2 bộ dữ liệu trên cho thấy không hề có một bản ghi nào bị trống dữ liệu.

### **Biến đổi và thêm mới dữ liệu**

Để thuận tiện hơn trong quá trình phân tích dữ liệu, chúng em tiến hành biến đổi dữ liệu bằng cách thay đổi một số trường dữ liệu, thay đổi kiểu dữ liệu, tạo dữ liệu mới và kết hợp dữ liệu với nhau tạo ra một dataframe phù hợp cho quá trình phân tích.

Đầu tiên, đối với bộ dữ liệu *orders.csv*, ta chuyển kiểu dữ liệu của cột *Date Order was placed* và *Delivery Date* về lại dạng ngày tháng bằng số hoàn chỉnh, thay vì kiểu dữ liệu viết tắt tháng như ban đầu.



Hình 1. 7. Kiểu dữ liệu ngày và tháng ban đầu



Hình 1. 8. Kiểu dữ liệu ngày tháng sau khi được biến đổi để phù hợp cho việc phân tích

Trong quá trình kiểm tra lại bộ dữ liệu thì nhóm chúng em có nhận thấy một thiếu sót trong quá trình tổng hợp dữ liệu của bên cung cấp bộ dữ liệu, bằng cách sử dụng câu lệnh *unique()* đối với mỗi cột trong bộ dữ liệu, chúng em nhận thấy đối với cột *Customer Status* bị lỗi lặp dữ liệu trùng trong một số bản ghi. Ví dụ: Gold và GOLD, Silver và SILVER, một số lỗi viết hoa và viết thường ko thông nhất dẫn đến lỗi không khớp dữ liệu với nhau:



Hình 1. 9. Lỗi không thống nhất dữ liệu trên cột Customer Status

Để xử lý vấn đề này, chúng em xử dụng câu lệnh *str.lower()* và *str.capitalize()* để thống nhất lại kiểu dữ liệu chung trong cột này. Ngoài ra chúng em còn chuyển kiểu dữ liệu của cột từ *object* sang *string* để thuận tiện cho việc đọc – ghi dữ liệu:



Hình 1. 10. Thông nhất kiểu dữ liệu trên cột Customer Status

Ngoài ra, để thuận tiện hơn cho việc phân tích nhóm chúng em quyết định thêm mới dữ liệu bằng cách tạo một cột mới tên là *Item Retail Value* (giá bán lẻ trên từng sản phẩm)dựa trên dữ liệu đến từ cột *Total Retail Price for This Order* (Tổng giá bán lẻ trên order này)và *Quanlity Order* (Số lượng sản phẩm) theo công thức *Item Retail Value = Total Retail Price for This Order / Quanlity Order*.



Hình 1. 11. Thêm một cột dữ liệu mới tên là Item Retail Value

Bên cạnh đó, để có thể biết được số lượng được bán ra của từng sản phẩm cũng như là giá trị vốn lẫn lãi trung bình thu được của sản phẩm đó, nhóm chúng em tiến hành tạo mới một dataframe dựa trên bộ dữ liệu *orders.csv* đã được tiền xử lý trước như sau: Đầu tiên nhóm (group) các mã sản phẩm với nhau (nhằm tạo ra một bản ghi về sản phẩm là duy nhất trong dataframe), song đếm số lượng sản phẩm được group lại, bên cạnh đó tính toán lại giá trị trung bình trên từng sản phẩm của các cột như *Cost Price Per Unit* (Giá vốn trên từng sản phẩm) và *Item Retail Value* (giá bán lẻ trên từng sản phẩm) sao cho phù hợp với lại từng sản phẩm đã được nhóm (group) lại với nhau.



Hình 1. 12. Tạo dataframe mới nhằm thống kê số lượng sản phẩm được bán cũng như giá trị vốn - lãi trung bình của sản phẩm

### **Bộ dữ liệu tổng thể**

Sau quá trình tiền xử lý ở phần trên, dữ liệu chúng ta thu được ở bước này là bộ dữ liệu dataframe về số lượng, giá vốn – lãi trung bình trên từng mặt hàng, do đó cái mà nhóm chúng em thiếu để có thể phân tích đặc tính của mặt hàng đó chính là thông tin tổng thể của từng mặt hàng. Thông tin tổng thể của từng mặt hàng được chứa trong bộ dữ liệu *product-supplier.csv*, cho nên nhóm chúng em quyết định trộn (merge) dữ liệu từ 2 dataframe đó lại với nhau nhằm mục đích tạo ra một bộ dữ liệu hoàn chỉnh hơn.

Kết hợp dữ liệu tư 2 bộ dữ liệu với nhau khác giống với việc join 2 table lại với nhau trong SQL, bằng cách sử dụng câu lệnh *merge()*với tham số điều kiện kết hợp bảng là *Product ID* thì ta có thể trộn dữ liệu từ 2 bảng lại với nhau. Dữ liệu được kết hợp lại sẽ có mô tả như sau:



Hình 1. 13. Thông tin của bảng dữ liệu được merge lại với nhau

Và bộ dữ liệu sẽ có dạng như sau:



Hình 1. 14. Dữ liệu của bảng dữ liệu được merge lại với nhau

Tuy vậy, dữ liệu bên trong khá là nhiều và có thể gây tốn kém tài nguyên trong quá trình phân tích. Nhóm chúng em quyết định lược bỏ đi những cột không cần thiết như *Product Line, Product Group, Supplier Country, Supplier Name* và *Supplier ID*. Bên cạnh đó, đổi tên các cột sao cho dễ hiễu và phù hợp hơn như *N Rows* đổi thành *Total Sold* (Tổng sản phẩm bán được), *Cost Price Per Unit* đổi thành *Wholesale Price* (Giá sỉ), *Item Retail Value* đổi thành *Retail Price* (Giá lẻ).



Hình 1. 15. Bộ dữ liệu cuối cùng trong quá trình tiền xử lý

Tóm lại, qua một chuỗi tiền xử lý trên, từ 2 bộ dữ liệu về lịch sử mua bán và thông tin mặt hàng, nhóm chúng em đã tóm gọn và xử lý thành một bộ dữ liệu hoàn chỉnh và đầy đủ về thông tin được bán ra như số lượng, giá sỉ-lẻ của từng mặt hàng. Bộ liệu thu được qua quá trình tiền xử lý trên sẽ chính là tiền đề để nhóm chúng em phân tích sâu hơn trong những phần sau.

## **Thăm dò dữ liệu sau tiền xử lý**

Đối với một người làm về phân tích dữ liệu, không dễ để nhìn vào một cột số cụ thể hay toàn bộ bảng dữ liệu để xác định các đặc điểm quan trọng của dữ liệu, nếu thực hiện bằng cách thức thủ công, sẽ mất rất nhiều thời gian và mức độ hiệu quả không được đảm bảo. Vì vậy, *EDA –Phân tích khám phá dữ liệu (Exploratory Data Analyst)* sẽ là một giải pháp phù hợp dành cho các nhà phân tích dữ liệu. Vậy EDA là gì? Mục đích của việc sử dụng EDA như thế nào?

### **Khái niệm về EDA**

Phân tích dữ liệu thăm dò là quá trình mô tả dữ liệu bằng các kỹ thuật thống kê và trực quan hoá nhằm tập trung vào các khía cạnh quan trọng của dữ liệu để tiếp tục phân tích. Điều này bao gồm cả việc kiểm tra tập dữ liệu từ nhiều góc độ, mô tả và tóm tắt nó mà không đưa ra bất kỳ giả định nào khác về nội dung của nó. Trong đề tài của nhóm chúng em sẽ tập trung hiển thị các đặc trưng của dữ liệu trong bộ dữ liệu đã tiền xử lý ở phần trước.

### **Mục đích sử dụng EDA**

Một số mục đích của việc sử dụng EDA vào các dự án phân tích dữ liệu như:

* Tìm hiểu về cấu trúc dữ liệu: EDA là phương pháp giúp xác định cấu trúc dữ liệu bao gồm số lượng, kiểu dữ liệu, trường dữ liệu, sự liên kết giữa các trường dữ liệu,... Khi xác định được cấu trúc dữ liệu, các nhà phân tích dữ liệu có thể hiểu được mối quan hệ giữa các dữ liệu trong tệp.
* Điều chỉnh và thay đổi: EDA giúp giải quyết các trường hợp thiếu giá trị, dữ liệu lỗi, các ngoại lệ trong dữ liệu. Điều này giúp các nhà phân tích dữ liệu điều chỉnh các phương án khắc phục kịp thời, tránh những ảnh hưởng nghiêm trọng đến dự án.
* Xác định mối tương quan giữa các biến: Các biến đều chứa các giá trị riêng, EDA có khả năng phát hiện các liên hệ tiềm ẩn và sự ảnh hưởng giữa các biến với nhau, tạo sự liên kết giữa các thông tin dữ liệu nhằm xây dựng một quy trình phân tích tổng thể, rõ ràng.
* Xây dựng cơ sở dữ liệu quan hệ: Các đối tượng dữ liệu quan trọng được phát triển mối quan hệ nhằm cấu trúc hóa dữ liệu theo sơ đồ, tiết kiệm thời gian xử lý những thông tin thừa, hạn chế sự sai sót của kết quả phân tích.
* Chuẩn bị cho bước phân tích tiếp theo: Áp dụng EDA giúp loại bỏ các dữ liệu không cần thiết, dữ liệu thiếu giá trị và chuẩn hóa dữ liệu. Đây là yếu tố nền tảng để chuẩn bị cho các bước phân tích bằng thuật toán học máy.

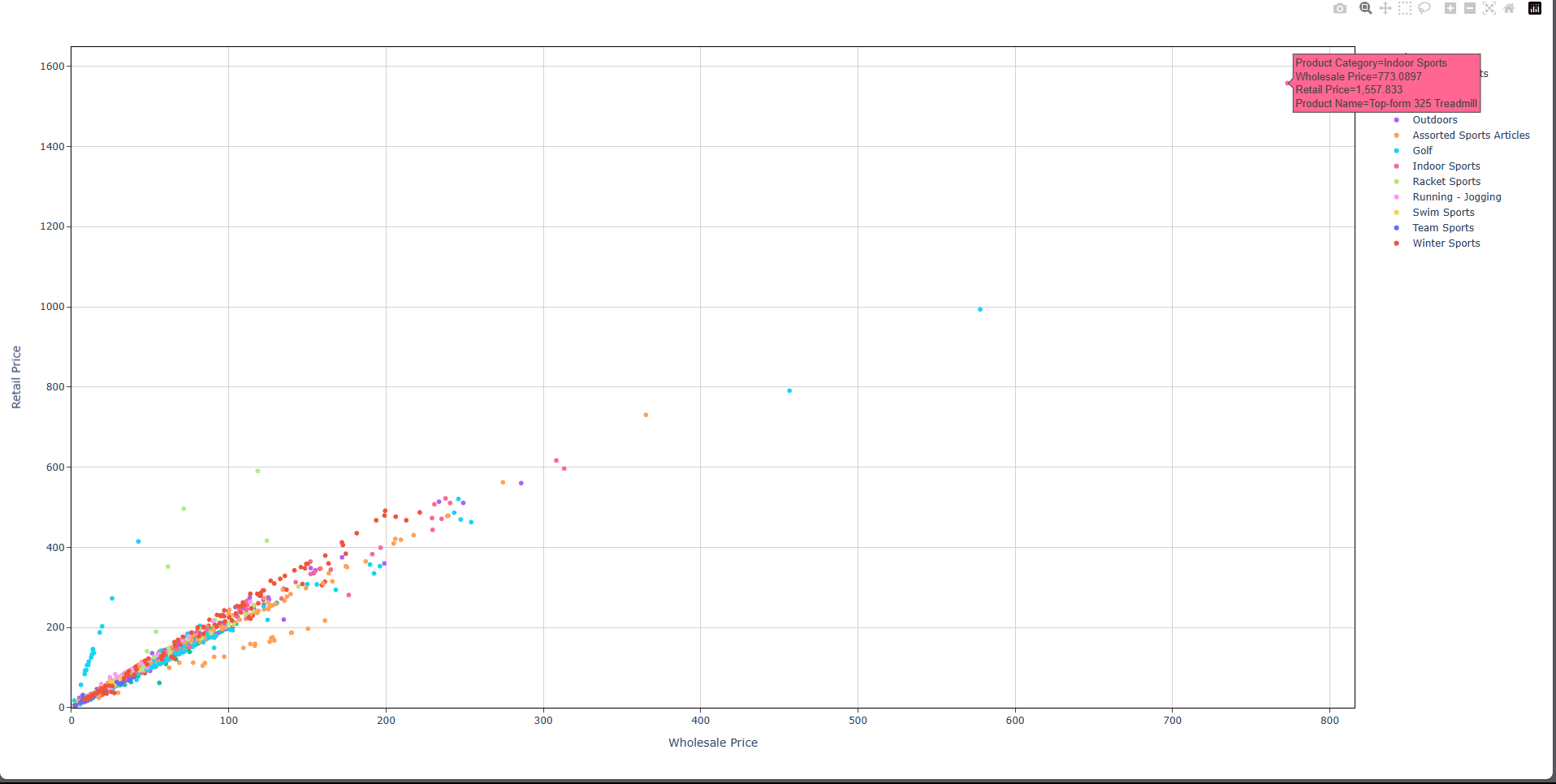
### **Thực nghiệm EDA đối với dữ liệu**

Để có một cái nhìn bao quát hơn về dữ liệu mà nhóm chúng em đang thực nghiện phân tích, nhóm chúng em tiến hành phân tích và tổng hợp dữ liệu một lần dưới dạng mô hình nữa nhằm đưa ra cái nhìn trực quan hơn về dữ liệu mà chúng em đang làm việc.

* *Hiển thị danh mục mặt hàng với giá bán sỉ - lẻ (1)*



Như hình vẽ trên, thì các danh mục phân loại mặt hàng đã được sắp xếp và hiển thị trên biểu đồ theo mức giá của các mặt hàng. Do đó, theo mô hình hiển thị thì mặt hàng “Top-form 325 Treadmill” đang có mức giá dao động là cao nhất so với toàn bộ mặt hàng.



* *Tổng tiền thu được từ các danh mục mặt hàng (2)*



Biểu đồ trên cho biết tổng tiền thu được từ các danh mục mặt hàng. Trong đó cho biết danh mục quần áo (Clothes) là danh mục chiếm tỷ trọng tổng thu nhập là lớn nhất và danh mục quần áo thể thao trong nhà (Indoor Sports) chiếm tỷ trọng bé nhất trong tất cả danh mục sản phẩm.

* Tổng tiền thu được từ danh mục quần áo (Clothes) là 38.953k
* Tổng tiền thu được từ danh mục quần áo thể thao trong nhà là 2995
* *Số lượng sản phẩm bán được dao động trong các khoảng mức giá khác nhau*
* *Bán sỉ (3)*



Biểu đồ trên cho biết tổng tiền thu được từ các sản phẩm có các khoảng mức giá bán sỉ khác nhau. Ở biểu đồ trên thì các sản phẩm có mức giá từ 5-10 có tổng số lượng bán được 23 ngàn sản phẩm, chiếm tỷ trọng lớn nhất trong toàn bộ sản bộ sản phẩm. Do đó, ta có thể ngầm kết luận rằng ở các khoảng mức giá bản sỉ từ 5-10 chiếm phần lớn người mua nhất và khoảng mức giá từ 770-775 thì chiếm số lượng người mua là ít nhất.

Ngoài ra, bảng nhỏ phía trên cũng mô tả thông tin các sản phẩm nằm trong khoảng mức giá ấy. Bên cạnh đó, ta cũng có kết luận từ một số thống kê cơ bản của các cột như sau:



* Tổng các cột (count): 3124
* Giá trị trung bình (mean): 34.1039767
* Độ lệch chuẩn (std): 40.766120
* Giá trị nhỏ nhất (min): 0.2
* Giá trị phân vị thứ 25 (25%): 11.996711
* Giá trị phân vị thứ 50 (50%): 22.36443
* Giá trị phân vị thứ 75 (75%): 40.180476
* Giá trị lớn nhất (max): 773.089744
* *Bán lẻ (4)*



Biểu đồ trên cho biết tổng tiền thu được từ các sản phẩm có các khoảng mức giá bán lẻ khác nhau. Ở biểu đồ trên thì các sản phẩm có mức giá từ 20-29.99 có tổng số lượng bán được 21.387 sản phẩm, chiếm tỷ trọng lớn nhất trong toàn bộ sản bộ sản phẩm. Do đó, ta có thể ngầm kết luận rằng ở các khoảng mức giá bán lẻ từ 20-29.99 chiếm phần lớn người mua nhất và khoảng mức giá từ 1555-1559.99 thì chiếm số lượng người mua là ít nhất.

Ngoài ra, bảng nhỏ phía trên cũng mô tả thông tin các sản phẩm nằm trong khoảng mức giá ấy. Bên cạnh đó, ta cũng có kết luận từ một số thống kê cơ bản của các cột như sau:



* Tổng các cột (count): 3124
* Giá trị trung bình (mean): 73.575585
* Độ lệch chuẩn (std): 84.644969
* Giá trị nhỏ nhất (min): 0.625000
* Giá trị phân vị thứ 25 (25%): 26.658506
* Giá trị phân vị thứ 50 (50%): 48.963529
* Giá trị phân vị thứ 75 (75%): 86.635305
* Giá trị lớn nhất (max): 1557.833333

## **Kết quả**

Kết quả thu được sau khi áp dụng phương pháp EDA như sau:

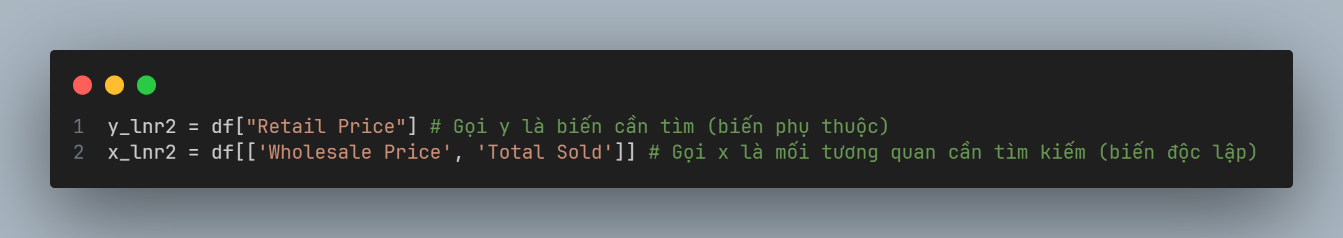
* Danh mục Golf có sản phẩm có tỷ suất lợi nhuận cao nhất (Kết quả từ hình 1)
* Danh mục quần áo (Clothes) là danh mục bán chạy nhất (Kết quả từ hình 2)
* Danh mục thể thao trong nhà (Indoor Sport) là danh mục bán ít nhất. (Kết quả từ hình 2)
* Hầu hết giá các mặt hàng đều dưới 50, dao động trên cả 2 loại mặt hàng bán lẻ và bán sỉ (Kết quả từ hình 3 và 4)
* Giá bán lẻ thường cao hơn nhiều so với giá bán sỉ

# **Chương 3: THỰC NGHIỆM VÀ KẾT LUẬN**

## **Huấn luyện mô hình**

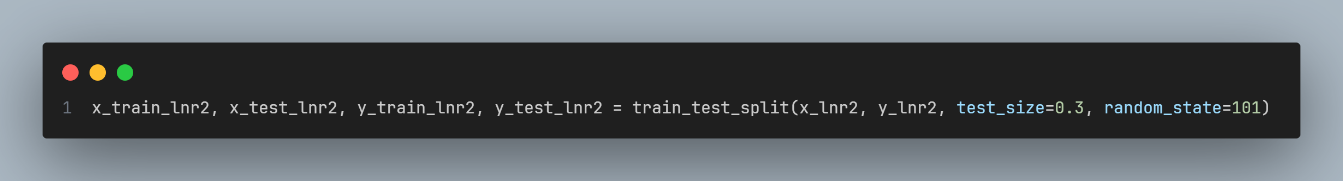
### **Linear Regression 2 biến**

Đầu tiên, chúng ta chọn *Retail Price* làm biến phụ thuộc và *Wholesale Price*, *Total Sold* làm biến độc lập trong tập dữ liệu đã được tiền xử lý trước đó, trong code này thì nó được gọi là *df* .



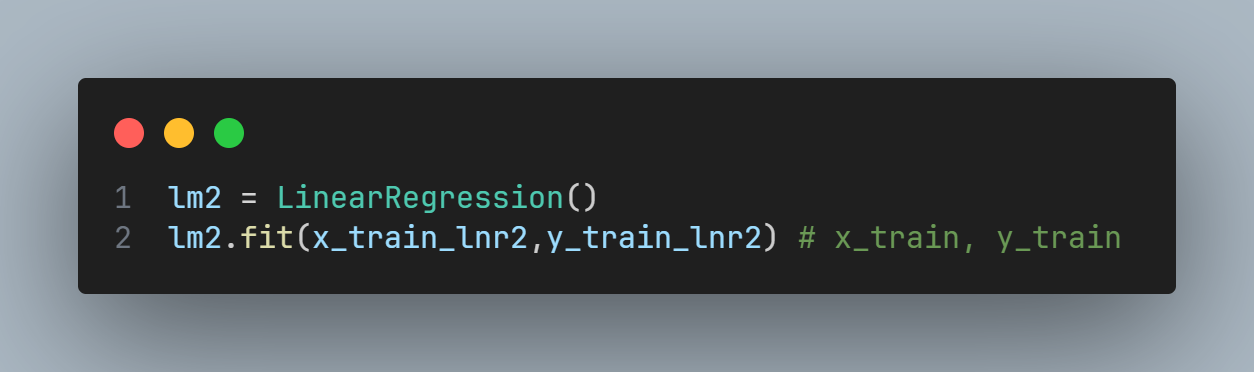
Hình 3. 1. Chọn biến độc lập và biến phụ thuộc LNR-2

Tiếp sau đó ta sử dụng hàm *train\_test\_split* trong thư viện *sklearn* với tham số *text\_size* (Kích thước của tập kiểm tra) là 0.3 và *random\_state* (số lần tái tạo trong việc kiểm soát việc xáo trộn dữ liệu trước khi áp dụng phân chia) là 101 lần. Hàm này sẽ trả về các tham số là *x\_train, y\_train, x\_test và y\_test*.



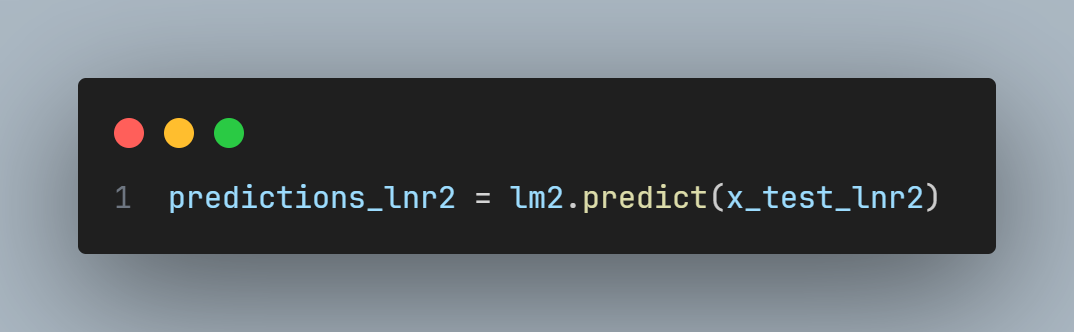
Hình 3. 2. Sử dụng hàm train\_test\_split để sinh các x\_train, y\_train, x\_test và y\_test trong LNR-2

Giải thích cho các tham số trong hàm *train\_test\_split* như sau: *test\_size* là đại diện cho tỷ lệ của tập dữ liệu để chia kiểm tra, *random\_state* là kiểm soát cách dữ liệu được xáo trộn trước khi thực hiện phân chia.



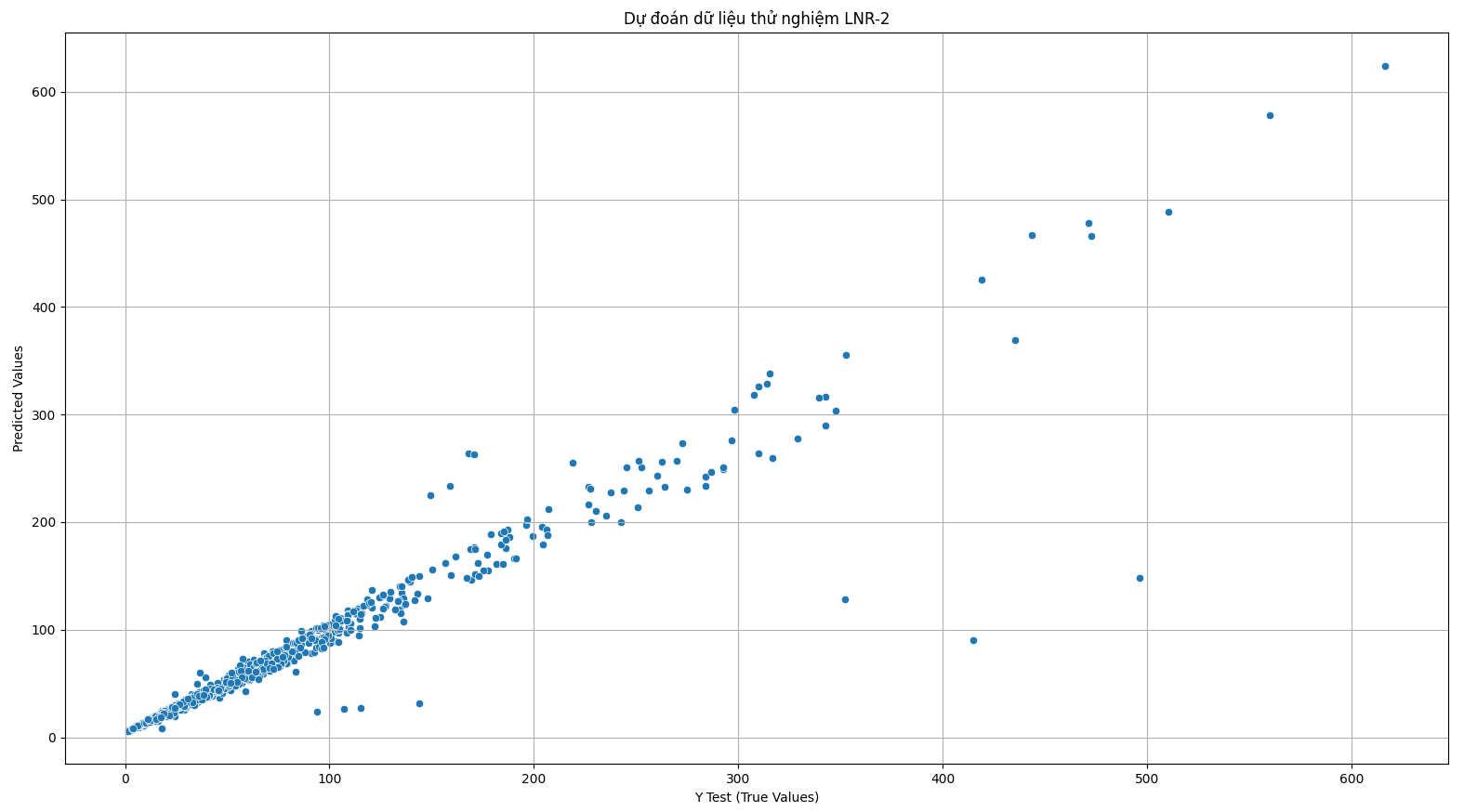
Hình 3. 3. Bắt đầu quá trình huấn luyện mô hình LNR-2

Sau khi có được *x\_train ­*và *y\_train* từ bước trên, ta tiến hành huấn luyện mô hình LNR. Sau khi chạy xong huấn luyện mô hình hồi quy, ta tiếp tục chạy thử nghiệm dự đoán *predictions\_lnr2* với *x\_test* và *y\_test*:



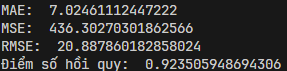
Hình 3. 4. Dự đoán kết quả LNR-2

Sau khi thực nghiệm dự đoán, ta có mô hình trực quan hóa như sau:



Hình 3. 5. Trực quan hóa kết quả dự đoán LNR-2

Sau đó, ta đánh giá mô hình huấn luyện bằng các chỉ tiêu đánh giá MAE, MSE, RMSE và đánh giá hồi quy bằng điểm số hồi quy (điểm số hồi quy đánh giá theo *y\_test* và kết quả đánh giá *predictions\_lnr2* mà ta đã chạy trên):

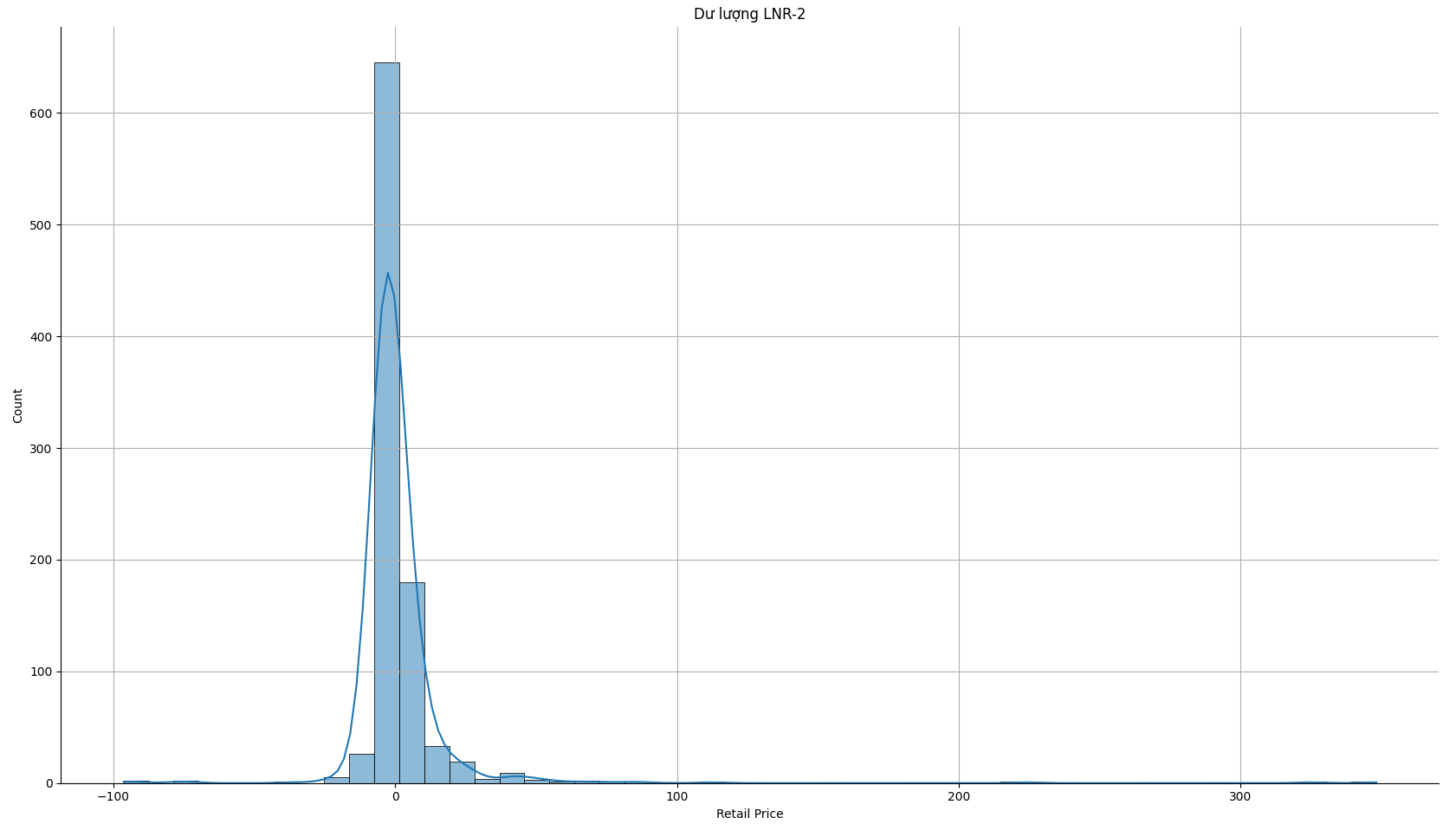


Hình 3. 6. Đánh giá chỉ tiêu và điểm số hồi quy LNR-2

MAE là 7.024611124472225 có nghĩa là trung bình mô hình dự đoán sai lệch khoảng 7.02 đơn vị so với giá trị thực tế, MSE là 436.3027030186259 có nghĩa là bình phương sai số dự đoán của mô hình là 436.30 và RMSE là 20.88786018285803 có nghĩa là mô hình dự đoán sai lệch khoảng 20.89 đơn vị so với giá trị thực tế. Với một dataframe có 3124 bản ghi thì mức độ sai số này là có thể chấp nhận được, kết luận rằng mô hình rất tốt và phù hợp.

Bên cạnh đó điểm số hồi quy cũng sấp sỉ 0.9235, cho thấy rằng phương pháp hồi quy LNR này phù hợp với bộ dữ liệu trong việc huấn luyện mô hình.

Tiếp sau đó là trực quan hóa dữ liệu của một số mô hình:



Hình 3. 7. Trực quan hóa phân bổ dư lượng LNR-2

Trực quan hóa mức độ phân bổ dư lượng (Residuals) là sự khác biệt giữa giá trị thực tế và giá trị dự đoán bởi mô hình đã được phân tích ở trên. Ở hình trên ta có thể thấy mức độ phân bổ dư lượng là đồng đều và không bị tỏa ra quá nhiều. Kết luận rằng phần dư lượng này phân bổ bình thường và nó không phải là vấn đề trong việc phân tích này.

Cuối cùng, ta kiểm tra hệ số (Coefficients) giữa các biến độc lập bằng hàm *coef\_* cho ra kết quả như sau:

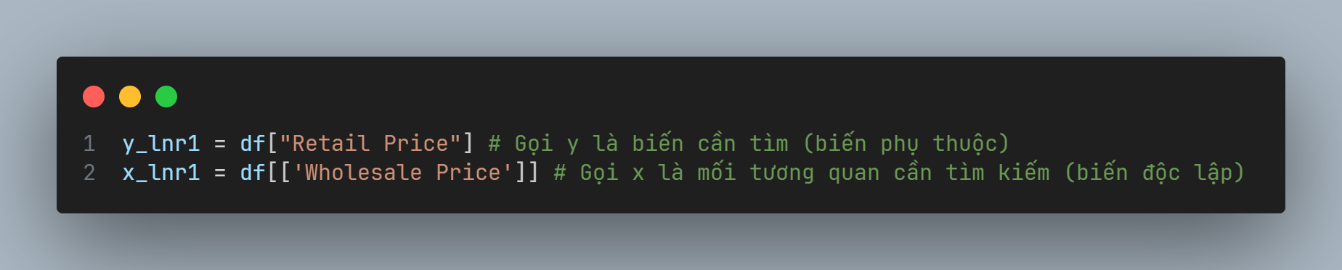


Hình 3. 8. Kết quả hệ số của LNR-2

Kết quả từ hình trên cho thấy hệ số của *Wholesale Price* là 2.01 và *Total Sold* là -0.0039, do đó có thể kết luận là không có mối tương quan đối với *Total Sold* , điều này có nghĩa là có một mối quan hệ nghịch đảo giữa biến độc lập và biến phụ thuộc. Do đó ta sẽ thử lại phương pháp LNR một lần nữa nhưng không có biến *Total Sold*, điều này sẽ được trình bày trong phần kế tiếp.

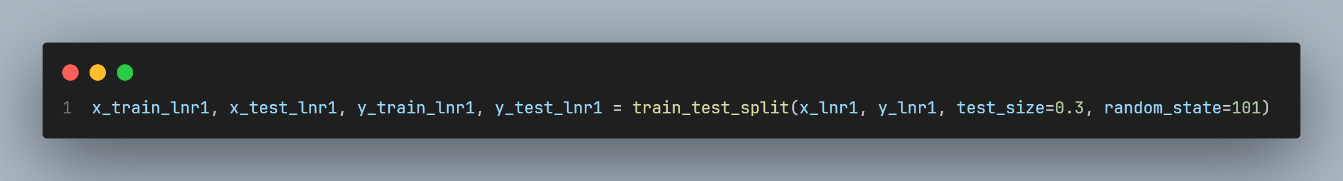
### **Linear Regression 1 biến**

Cũng giống như LNR 2 biến nhưng lần này ta bắt đầu lại chỉ với 1 biến độc lập là *Wholesale Price* thay vì 2 như trước là *Whosale Price* và *Retail Price*, ta gọi phương pháp sử dụng lần này là LNR-1:



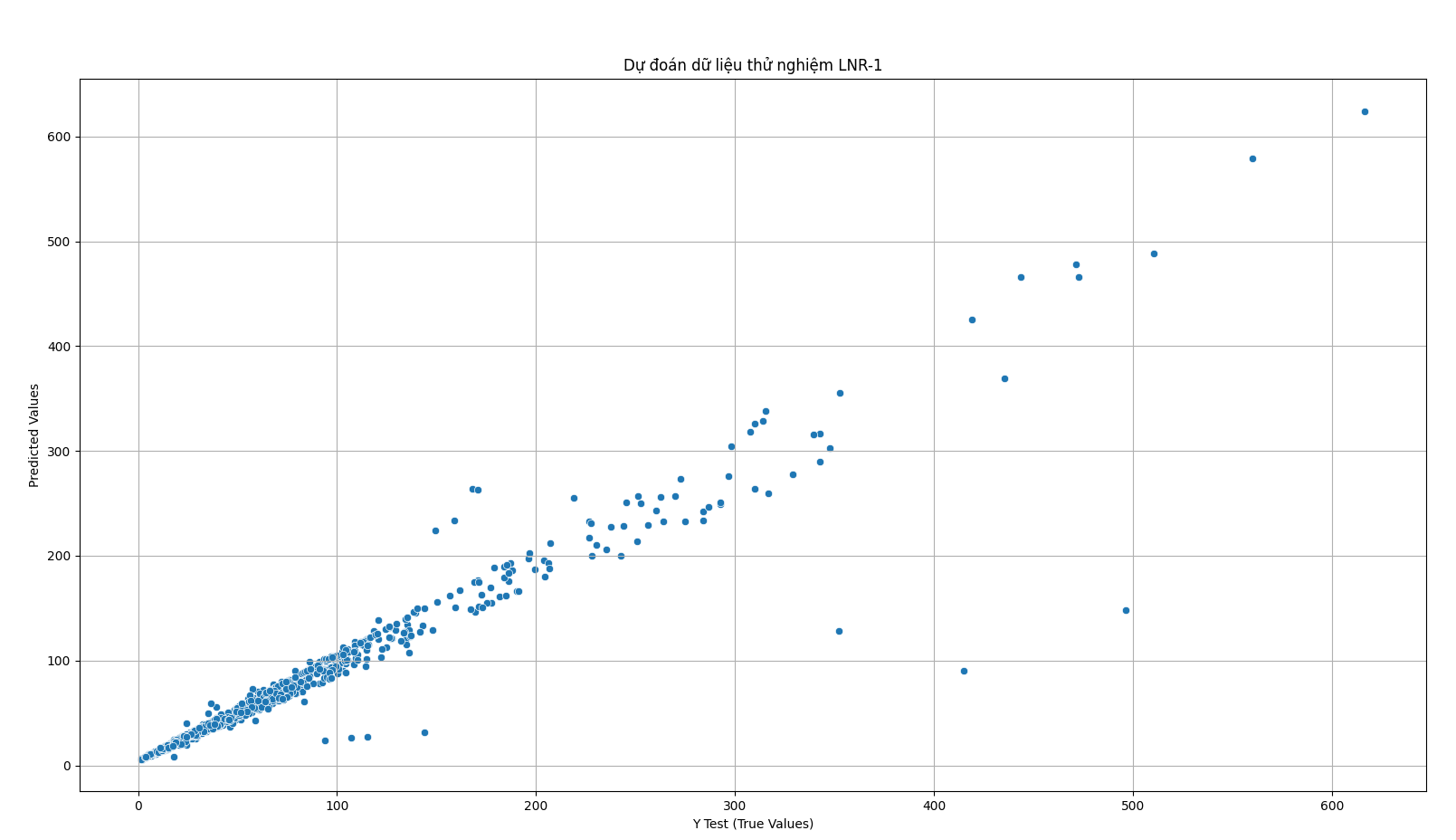
Hình 3. 9. Chọn biến độc lập và biến phụ thuộc LNR-1

Ta cũng sử dụng lại hàm *train\_test\_split* giống như lần trước và các tham số *text\_size* là 0.3 và *random\_state* là 101 lần.



Hình 3. 10. Sử dụng hàm train\_test\_split để sinh các x\_train, y\_train, x\_test và y\_test trong LNR-1

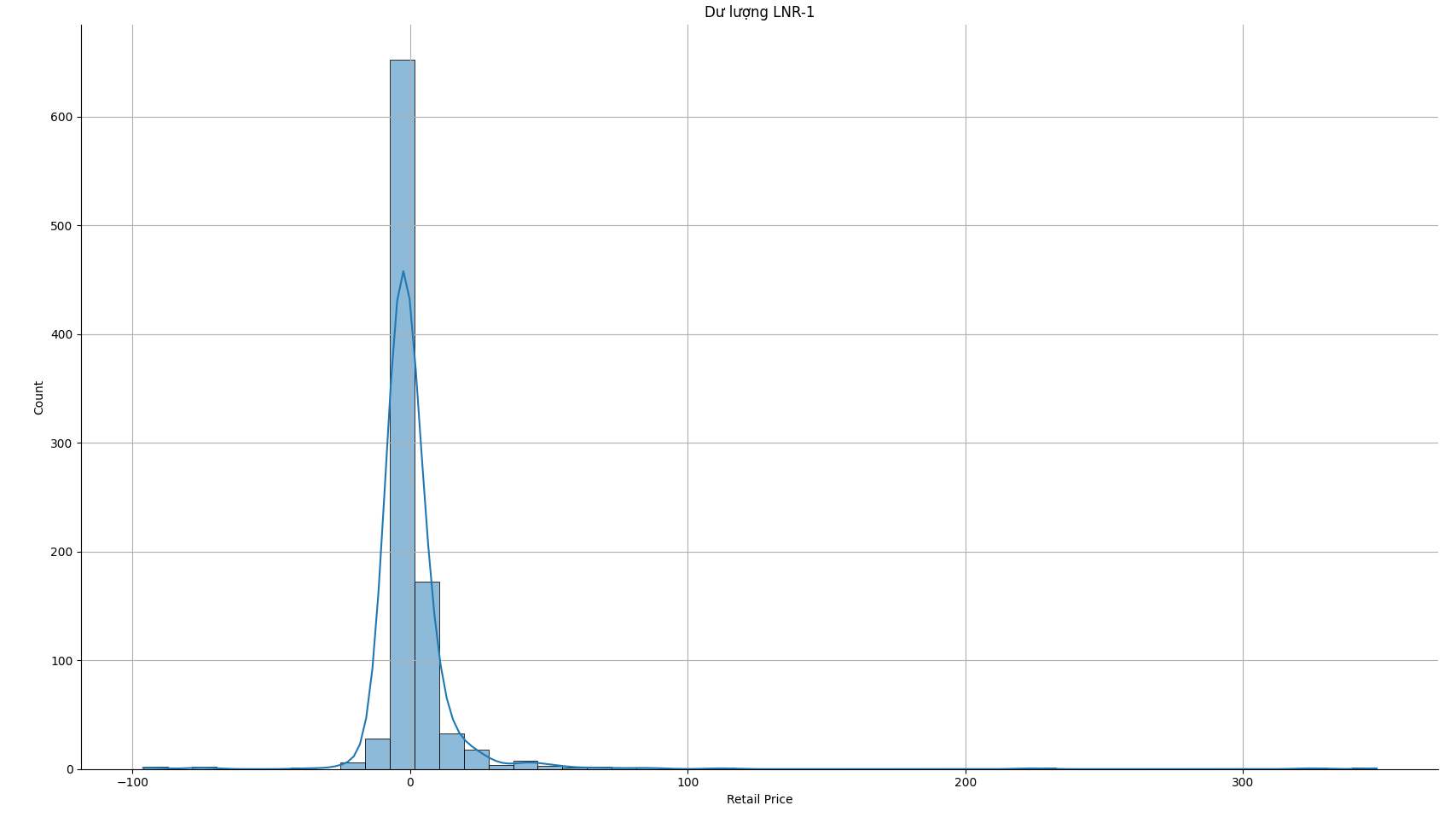
Các bước huấn luyện mô hình hồi quy và dự doán kết quả ta cũng thực hiện tương tự như LNR-2. Các kết quả thu được được trực quan hóa như sau:  
 *Mô hình trực quan hóa cho việc dự đoán:*



Hình 3. 11. Trực quan hóa kết quả dự đoán LNR-1

Có thể thấy rằng kết quả dự đoán của LNR-1 và LNR-2 là không có sự khác biệt nào cả, giống nhau đến 99%.

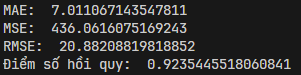
*Mô hình trực quan hóa phân bổ dư lượng*



Hình 3. 12. Trực quan hóa phân bổ dư lượng LNR-1

Có thể thấy rằng sự phân bố dư lượng của LNR-1 không có quá nhiều sự khác biệt với LNR-2. Kết luận rằng nó giống với LNR-2, không phải là vấn đề quá lớn trong việc phân tích dữ liệu.

*Đánh giá chỉ tiêu phương sai và điểm số hồi quy LNR-1*



Hình 3. 13. Đánh giá chỉ tiêu và điểm số hồi quy LNR-1

Điểm số hồi quy và đánh giá chỉ tiêu phương sai MAE, MSE và RMSE không có quá nhiều sự khác biệt so với LNR-2. Có thể đánh giá rằng LNR-1 cũng rất phù hợp cho việc phân tích dữ liệu.

*Kết quả hệ số LNR-1*

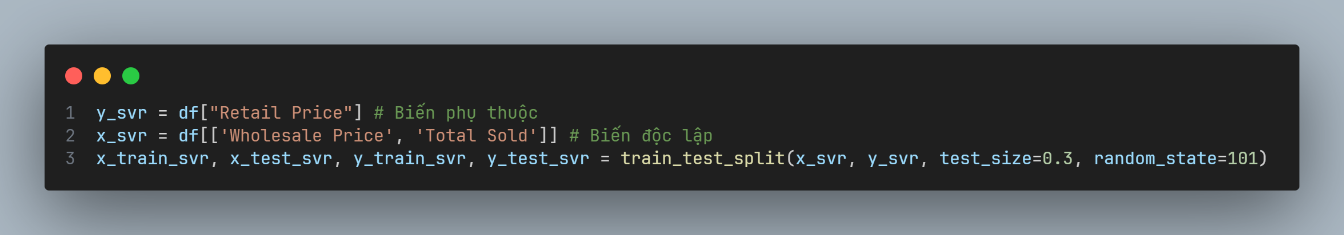
**

Hình 3. 14. Kết quả hệ số LNR-1

Kết quả hệ số của LNR-1 và LNR-2 là tương đồng nhau, không có quá nhiều sự khác biệt. Kết luận đối với LNR-1 rằng ta có thể giữ phương pháp này cùng với LNR-2 cho những giai đoạn phân tích dữ liệu sau.

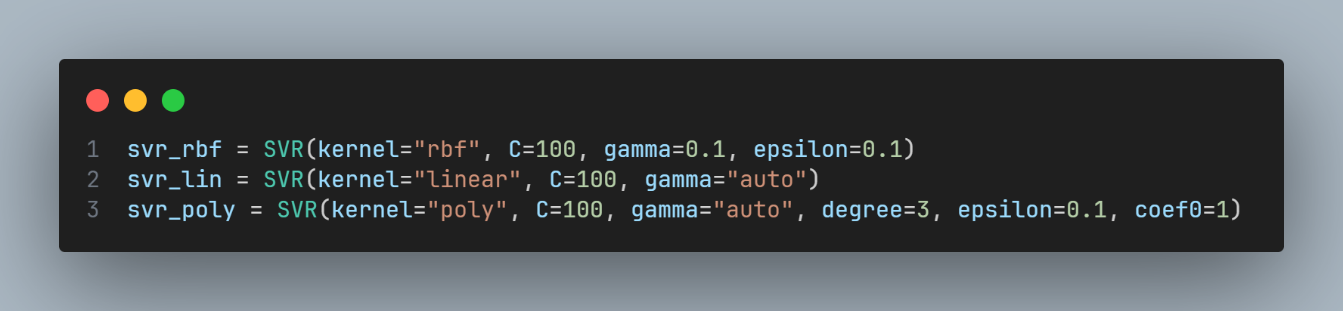
### **Support Vector Regression**

Cũng giống như LNR-2, chúng ta chọn *Retail Price* làm biến phụ thuộc và *Wholesale Price*, *Total Sold* làm biến độc lập. Các tham số của *train\_test\_split* ta cũng giữ nguyên tương tự, gồm *text\_size* là 0.3 và *random\_state* là 101 lần.



Hình 3. 15. Biến độc lập, biến phụ thuộc và tham số train\_test\_split của SVR

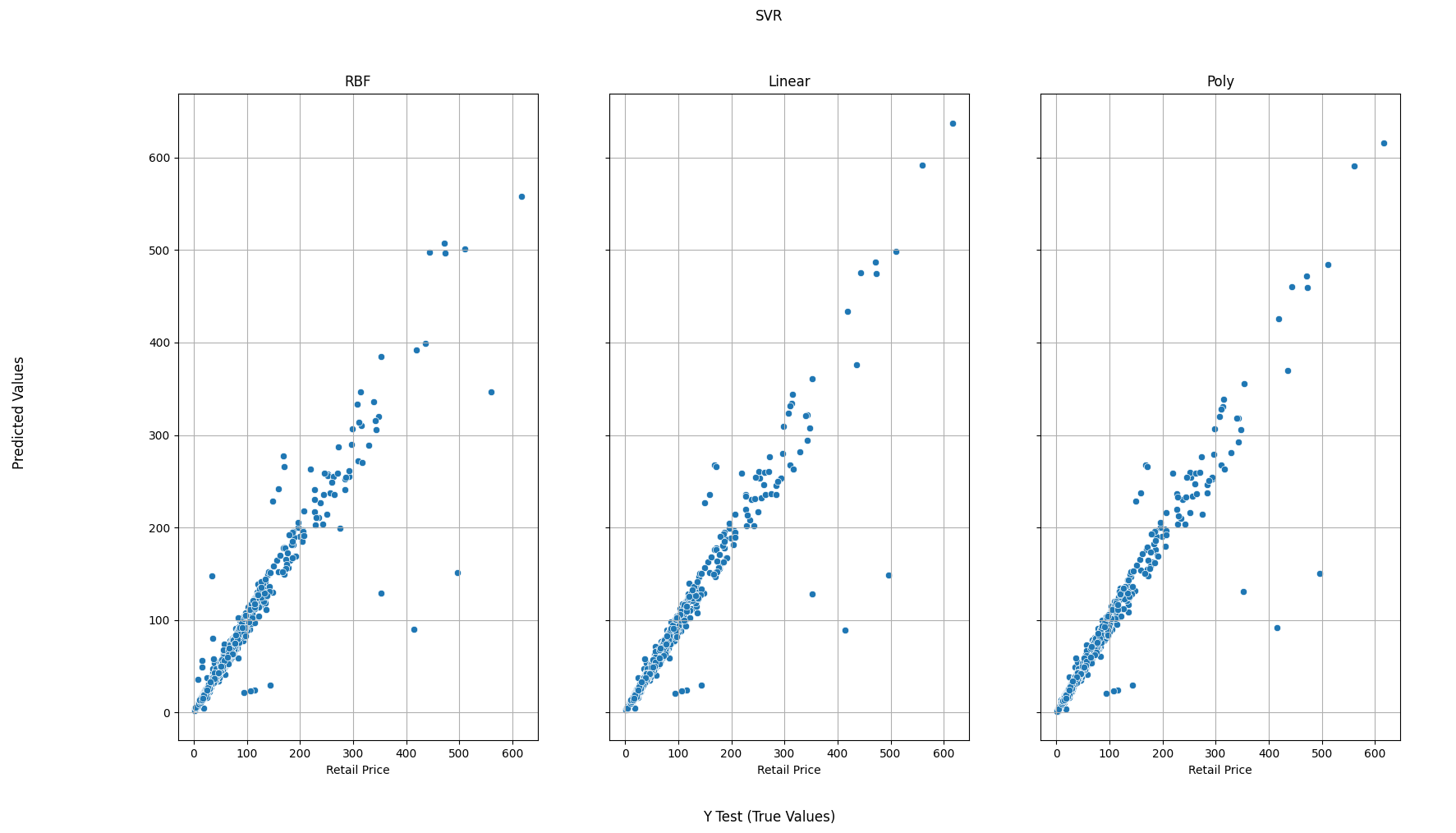
Tiếp sau đó, ta xác định các tham số của từng kernel trong SVR mà ta chọn, trong đây ta chọn 3 kernel để huấn luyện mô hình gồm *RBF, Linear* và *Poly:*



Hình 3. 16. Xác định các tham số của từng kernel trong SVR

Trong đó, đối với từng kernel gồm có các tham số *C, gamma, epsilon, degree* và *coef0.* Trong đó *C* là tham số chuẩn hóa, *gamma* là Hệ số kernel, *epsilon* là Epsilon trong mô hình epsilon-SVR (Nó chỉ định ống epsilon mà trong đó không có hình phạt nào được liên kết trong hàm mất mát huấn luyện với các điểm được dự đoán trong khoảng cách epsilon từ giá trị thực tế), *degree*là Bậc của hàm kernel đa thức (‘poly’), *coef0* là hạng tử độc lập trong hàm kernel (được sử dụng trong kernel *poly* và *sigmoid*).

Sau khi huấn luyện mô hình, ta mô hình trực quan hóa dự đoán như sau:

****

Hình 3. 17. Trực quan hóa kết quả dự đoán của SVR

**<**Đánh giá kết quả mô hình**>**

****

Hình 3. 18. Điểm số hồi quy của SVR

Đánh giá điểm số hồi quy của SVR: Điểm số hồi quy của SVR trên cả 3 kernel đều cho mức tốt và chấp nhận được, cho nên phương pháp này cũng được đưa vào để dự đoán cho những phần phân tích sau.

## **Bài toán 1: Ngày giao hàng có thay đổi không khi trở thành khách hàng hạng Bạch kim?**

## **Bài toán 2: Dự đoán giá bán lẻ tốt nhất**

## **Kết quả**

# **Tài liệu tham khảo**