TRƯỜNG CAO ĐẲNG CÔNG NGHỆ   
THỦ ĐỨC

----- □ & □ -----



**ĐỒ ÁN LẬP TRÌNH PYTHON NÂNG CAO**

Giảng viên: Phan Thị Thể

Sinh viên thực hiện:

|  |  |  |
| --- | --- | --- |
| STT | Họ và tên | MSSV |
| 1 | Huỳnh Thanh Trọng | 21211TT0497 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Sài Gòn 2023**

Mục Lục

[I. Mở Đầu 3](#_Toc3192)

[II. Phân Tích Dữ Liệu](#_Toc7914) **[Error! Bookmark not defined.](#_Toc7914)**3

[III. Tiền Xử lý dữ liệu 3](#_Toc26952)

[IV. Xây Dựng Mô Hình](#_Toc24449) **[Error! Bookmark not defined.](#_Toc24449)**6

[V. Đánh Giá Mô Hình 7](#_Toc11657)

[VI. Phân Tích Kết Quả 8](#_Toc24391)

[VII. Giải Pháp 8](#_Toc11446)

[VIII. Demo Chương Trình 8](#_Toc14816)

1. **) Mở Đầu**

Một trong những cửa hàng bán lẻ hàng đầu ở Mỹ, Walmart, muốn dự đoán chính xác doanh số và nhu cầu. Có một số sự kiện và ngày lễ nhất định ảnh hưởng đến doanh số bán hàng mỗi ngày. Có dữ liệu bán hàng có sẵn cho 45 cửa hàng của Walmart. Doanh nghiệp đang phải đối mặt với thách thức do nhu cầu không lường trước được và đôi khi hết hàng do thuật toán học máy không phù hợp. Một thuật toán lý tưởng sẽ dự đoán chính xác nhu cầu và sử dụng các yếu tố như điều kiện kinh tế bao gồm CPI, Chỉ số thất nghiệp, nhiệt độ và giá nguyên liệu. Dữ liệu bán hàng lịch sử của 45 cửa hàng Walmart ở các khu vực khác nhau đều có sẵn.

1. **Phân Tích Dữ Liệu**

Trong file Walmart.csv  
- Store: Mã số cửa hàng

- Date: Ngày nhập dữ liệu

- Weekly\_Sales: Doanh số bán hàng hàng tuần của cửa hàng

- Holiday\_Flag: Chỉ mục cho biết ngày liên quan có phải là ngày lễ hay không(1 là ngày lễ, 0 là không phải)

- Temperature: Nhiệt độ trong khu vực của cửa hàng.

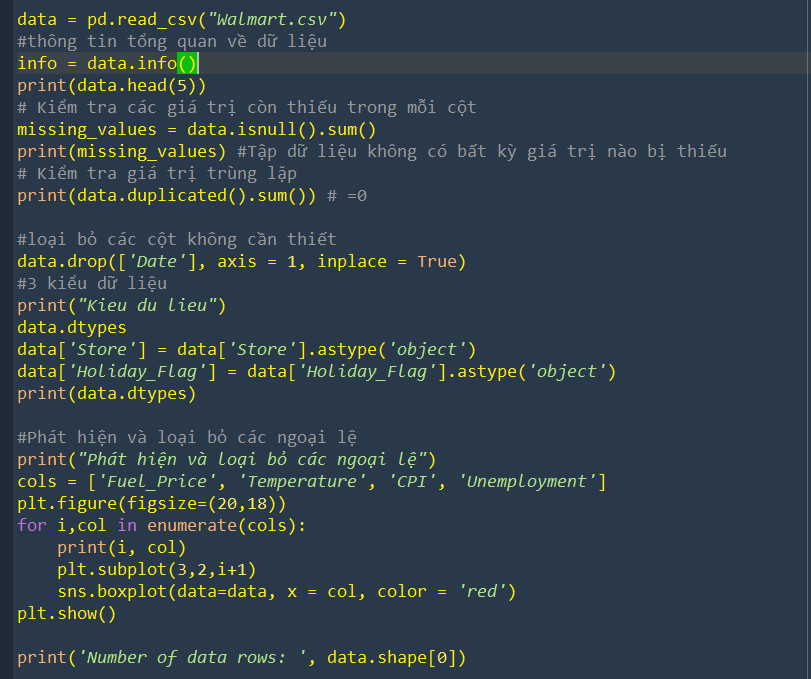
- Fuel\_Price: Giá nhiên liệu tại khu vực cửa hàng.

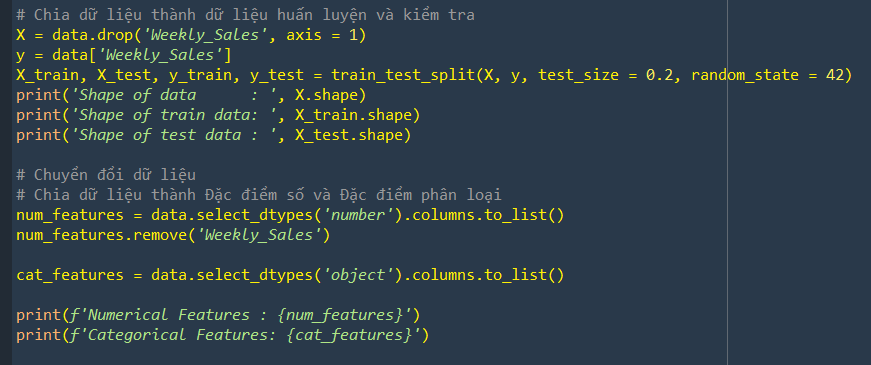
- CPI: Chỉ số giá tiêu dùng, đo lường sự thay đổi trong giá cả hàng hóa và dịch vụ tiêu dùng

- Unemployment: Tỉ lệ thất nghiệp trong khu vực cửa hàng

1. **Xử lý dữ liệu**

Xử lý dữ liệu bằng cách loại bỏ các giá trị thiếu, giá trị trùng lặp, cột không cần thiết, chuẩn hóa dữ liệu để đảm bảo mô hình không chứa ngoại lệ





1. **Tương Quan Dữ Liệu**

**Tương quan Pearson**

Tương quan Pearson đo lường sự phụ thuộc tuyến tính giữa hai biến X và Y.

Hệ số kết quả là giá trị bao gồm từ -1 đến 1, trong đó:

**1** : Tương quan tuyến tính dương hoàn hảo.

**0** : Không có tương quan tuyến tính, hai biến nhiều khả năng không ảnh hưởng lẫn nhau.

**-1** : Tương quan tuyến tính âm hoàn hảo.

**giá trị P**

Giá trị P là giá trị xác suất mà mối tương quan giữa hai biến này có ý nghĩa thống kê. Tôi chọn mức ý nghĩa là 0,05, có nghĩa là tôi tin cậy 95% rằng mối tương quan giữa các biến là đáng kể.

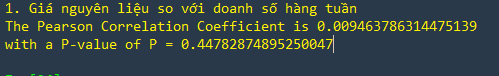
giá trị p là < 0,001: chúng tôi nói rằng có bằng chứng chắc chắn rằng mối tương quan là đáng kể.

giá trị p là < 0,05: có bằng chứng vừa phải cho thấy mối tương quan là đáng kể.

giá trị p < 0,1: có bằng chứng yếu cho thấy mối tương quan là đáng kể.

giá trị p > 0,1: không có bằng chứng nào cho thấy mối tương quan là đáng kể.

### **Giá nhiên liệu so với doanh số hàng tuần**



Vì giá trị p > 0,1 nên mối tương quan giữa giá nhiên liệu và doanh thu hàng tuần không có ý nghĩa thống kê.

Giá nhiên liệu dường như không phải là một yếu tố dự báo tốt cho doanh số bán hàng hàng tuần vì đường hồi quy gần như nằm ngang. Vì vậy, nó không phải là một biến đáng tin cậy.

### **Thất nghiệp so với doanh số hàng tuần**



Vì giá trị p < 0,001 nên mối tương quan giữa thất nghiệp và doanh số hàng tuần là bằng chứng mạnh mẽ cho thấy mối tương quan này rất đáng kể.

Thất nghiệp dường như là một yếu tố dự báo tốt về doanh số bán hàng hàng tuần. Tỷ lệ thất nghiệp càng cao thì doanh số bán hàng hàng tuần càng thấp.

### **CPI so với doanh thu hàng tuần**



Vì giá trị p < 0,001 nên mối tương quan giữa CPI và doanh số bán hàng hàng tuần là bằng chứng mạnh mẽ cho thấy mối tương quan này rất đáng kể.

### **Nhiệt độ so với doanh số hàng tuần**



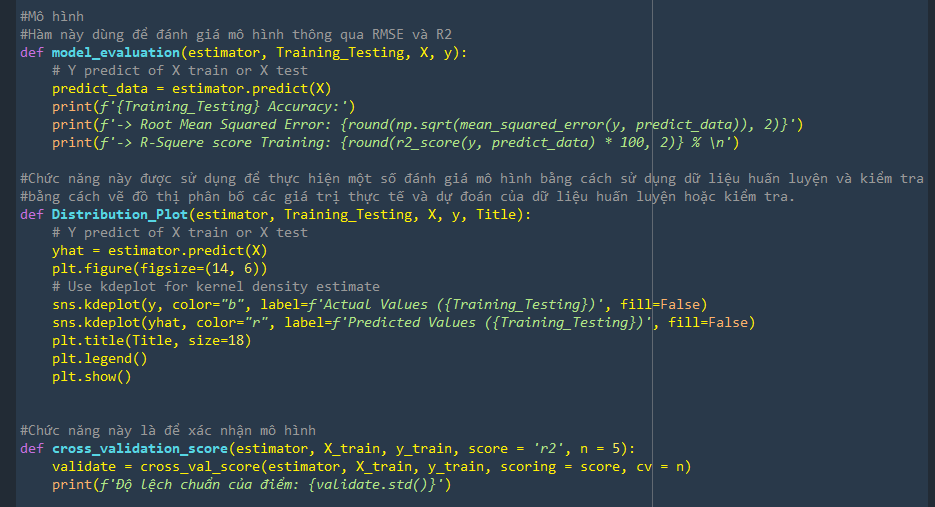
Vì giá trị p < 0,05 nên mối tương quan giữa nhiệt độ và doanh số bán hàng hàng tuần là bằng chứng vừa phải cho thấy mối tương quan này rất đáng kể.

Nhiệt độ có vẻ như là một yếu tố dự báo tốt về doanh số bán hàng hàng tuần. Tỷ lệ Nhiệt độ càng cao thì doanh số hàng tuần càng thấp.

1. **Xây Dựng Mô Hình**

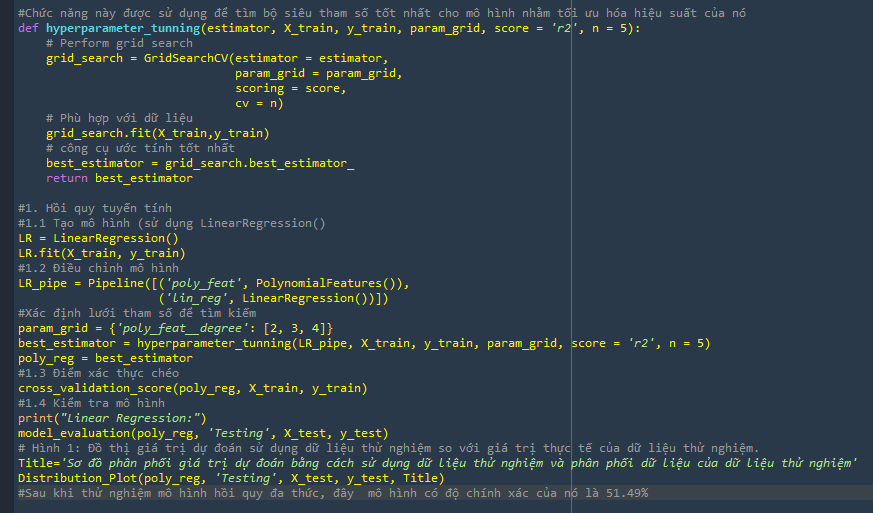
Tôi sử dụng Hồi Quy Tuyến Tính để mô tả mối quan hệ tuyến tính giữa các biến đọc lập và biến phụ thuộc, qua đó mô hình được huấn luyện trên tập dữ liệu.

Thuật toán hồi quy tuyến tính là một phương pháp trong thống kê để mô hình hóa mối quan hệ tuyến tính giữa một biến độc lập và một biến phụ thuộc. Mục tiêu của thuật toán là tìm ra đường hồi quy tuyến tính sao cho tổng bình phương của sự chênh lệch giữa các giá trị dự đoán và giá trị thực tế của biến phụ thuộc là nhỏ nhất.



Hàm cross\_validation\_score:

Dùng để đánh giá mô hình sử dụng phương pháp cross-validation và in ra độ lệch chuẩn của các điểm số (scores)



Hàm hyperparameter\_tunning:

Tìm kiếm bộ siêu tham số tốt nhất cho mô hình nhằm tối ưu hóa hiệu suất

1. **Đánh Giá Mô Hình**

Kết quả của mô hình Được đánh giá bởi MSE và R^2, từ đó rút ra được một số thông tin quan trọng về hiệu suất của mô hình trên tập dữ liệu kiểm thử.

Ngoài ra tôi còn sử dụng một số mô hình như Công Cụ Hồi Quy KNN và Cây Quyết Định Hồi Quy để so sánh:

Giá trị bình phương R tốt là gì?

- Khi so sánh các mô hình, mô hình có giá trị bình phương R cao hơn sẽ phù hợp hơn với dữ liệu.

Một MSE tốt là gì?

- Khi so sánh các mô hình, mô hình có giá trị MSE nhỏ nhất sẽ phù hợp hơn với dữ liệu.

**Hồi quy tuyến tính:**

Bình phương R: 51.49 %

RMSE: 395316.04

**Công cụ hồi quy KNN:**

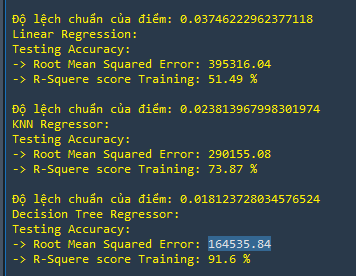
Bình phương R: 73.87 %

RMSE: 290155.08

**Cây quyết định hồi quy:**

Bình phương R:91.6 %

RMSE: 164535.84



1. **Phân Tích Kết Quả**

So sánh ba mô hình này, kết luận rằng Mô hình hồi quy tuyến tính là mô hình không tốt với độ chính xác 51.49% , mô hình cây quyết định hồi quy là mô hình tốt có thể dự đoán doanh số bán hàng hàng tuần từ tập dữ liệu của tôi với độ chính xác 91.6%

1. **Giải Pháp**

Cân nhắc sử dụng mô hình khác phù hợp với dữ liệu và điều chỉnh siêu tham số của các mô hình để xem xét xem có cải thiện hiệu xuât không, Để cửa hàng kịp thời đưa ra giải pháp hợp lý.

1. **Demo Chương Trình**