**TRƯỜNG CAO ĐẲNG CÔNG NGHỆ THỦ ĐỨC**

**KHOA CÔNG NGHỆ THÔNG TIN**

---🙡🙣---



**BÁO CÁO ĐỒ ÁN**

**Đề tài:** Báo cáo Hồi quy Tuyến tính và Ứng dụng trong Dự đoán Phân Phối Mẫu Xe Của Các Hãng Xe

***Nhóm gồm các thành viên:***

Lê Hoàng Nam

***Giáo viên hướng dẫn:***

Phan Thị Thể

**TP.HỒ CHÍ MINH - Năm 2023**

MỤC LỤC

[**I.** **Lời Mở Đầu** 4](#_Toc153263684)

[**II.** **Phân Tích Dữ Liệu** 4](#_Toc153263685)

[1. Model(Mẫu xe): Chỉ các mẫu xe của hãng xe 4](#_Toc153263686)

[2. Years(Năm sản xuất): Năm sản xuất ra các laoij xe của từng hãng 4](#_Toc153263687)

[3. Prices(Giá): Giá cả các loại xe 4](#_Toc153263688)

[4. Transmission(Hộp số): Hộp số của xe 4](#_Toc153263689)

[5. Mieage(Quãng đường): Quãng đường xe đi 4](#_Toc153263690)

[6. FuelType(Loại nhiên liệu): các loại nhiên liệu cho xe 4](#_Toc153263691)

[7. Tax(Thuế): thuế xe 4](#_Toc153263692)

[8. Mpg: Do lường nhiên liệu xe 4](#_Toc153263693)

[9. EngineSize(Kích thước động cơ): kích thước động cơ của mỗi xe mỗi hãng 4](#_Toc153263694)

[10. Brands(Hãng Xe): các hãng xe có trên thị trường 5](#_Toc153263695)

[**III.** **Tiền Xử lý Dữ liệu:** 5](#_Toc153263696)

[**1.** **Chuyển đổi dữ liệu phân loại thành biến giả:** 5](#_Toc153263697)

[**2.** **Chuẩn hóa dữ liệu:** 5](#_Toc153263698)

[**3.** **Đổi giá trị dữ liệu cột trong file dữ liệu:** 6](#_Toc153263699)

[**IV.** **Xây dựng mô hình:** 6](#_Toc153263700)

[**V.** **Đánh giá mô hình** 7](#_Toc153263701)

[**1.** **Set cứng 1 dữ liệu để đánh giá mô hình:** 7](#_Toc153263702)

[**2.** **Kết Quả:** 7](#_Toc153263703)

[**3.** **Phân Tích Kết Quả** 8](#_Toc153263704)

[**VI.** **Đề Xuất Ứng Dụng Và Giải Pháp:** 8](#_Toc153263705)

[**VII.** **Tài liệu Kham khảo:** 8](#_Toc153263706)

[**1.** **Sách kham khảo** 8](#_Toc153263707)

[**2.** **Các nguồn video kham khảo** 8](#_Toc153263711)

1. **Lời Mở Đầu**

Trong thời đại hiện nay, ngành công nghiệp ô tô đang phát triển rất nhanh chóng và đang trở thành một lĩnh vực đầy tiềm năng. Để tồn tại và phát triển trong thị trường cạnh tranh này, các công ty sản xuất ô tô phải không ngừng cải tiến và nâng cao chất lượng sản phẩm, đồng thời tìm kiếm những cách thức để tang lợi nhuân. Sau đây tôi xin hân hạnh được giới thiệu dự án “Brands Cars Price”. Đây là một dự án giá xe ô tô của các hãng ô tô lớn trên thế giới. Dự án này nhắm đến việc phát triển quy mô về doanh thu và dự đoán chính xác giá xe cho các hãng xe như Audi, Huynhdai, Ford, BMW,…

Khi nói đến những mô hình và dự đoán trên thì hiện nay ngoài các phần mềm phân tích như Execl, Data Studio ra thì hiện nay đã có một ngôn ngữ lập trình đã cho phép ta phân tích các số liệu hay là vẽ những mô hình trực quan các số liệu giống như Execl hay là Data Studio. Bằng việc sử dụng Python và các thư viện mạnh mẽ như Pandas, NumPy, và Scikit-learn thì ta sẽ có thể đi sâu hơn vào các dữ liệu liên quan đến giá cả, các mẫu xe hay là các hãng xe hay là số lượng phân phối xe của các năm. Đây là những thứ thú vị mà ngôn ngữ Python đem lại cho chúng ta.

Giới thiệu thuật toán LinearRegression:

Thuật toán Linear Regression là một mô hình học máy thuộc dạng học máy giám sát và được sử dụng trong các tác vụ dự đoán giá trị số (regression). Mục tiêu chính của Linear Regression là xây dựng một đường tuyến tính (linear function) sao cho tổng bình phương sai số (sự chênh lệch giữa giá trị dự đoán và giá trị thực tế) là nhỏ nhất. Trong ngữ cảnh này, "sai số" đại diện cho sự chênh lệch giữa dự đoán và thực tế. Bằng cách sử dụng phương pháp least squares, Linear Regression tối ưu hóa các hệ số của đường tuyến tính để giảm thiểu tổng sai số bình phương.

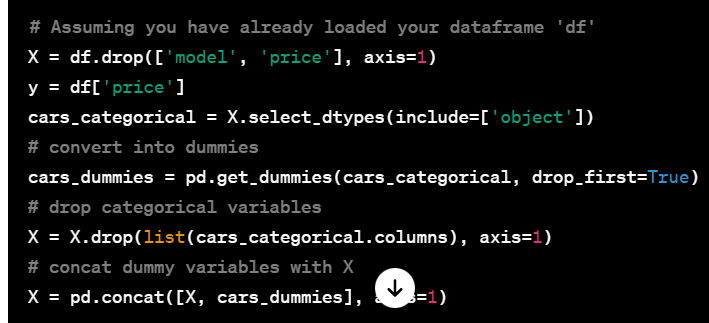
1. **Phân Tích Dữ Liệu**

Dữ liệu được thu thập từ doanh nghiệp phân phối xe:

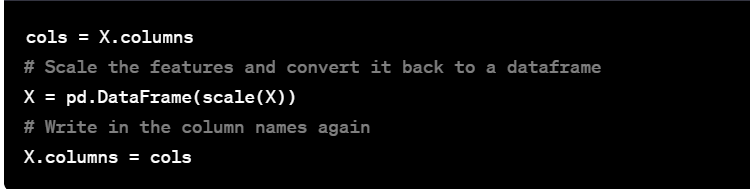
1. Model(Mẫu xe): Chỉ các mẫu xe của hãng xe
2. Years(Năm sản xuất): Năm sản xuất ra các laoij xe của từng hãng
3. Prices(Giá): Giá cả các loại xe
4. Transmission(Hộp số): Hộp số của xe
5. Mieage(Quãng đường): Quãng đường xe đi
6. FuelType(Loại nhiên liệu): các loại nhiên liệu cho xe
7. Tax(Thuế): thuế xe
8. Mpg: Do lường nhiên liệu xe
9. EngineSize(Kích thước động cơ): kích thước động cơ của mỗi xe mỗi hãng
10. Brands(Hãng Xe): các hãng xe có trên thị trường
11. **Tiền Xử lý Dữ liệu:**

Trước khi xây dụng một mô hình thì chúng ta nên xử lý dữ liệu, thay đổi dữ liệu, chuẩn hóa dữ liệu không chứa các ngoại lệ hay độ lệch lớn như là:

1. **Chuyển đổi dữ liệu phân loại thành biến giả:**

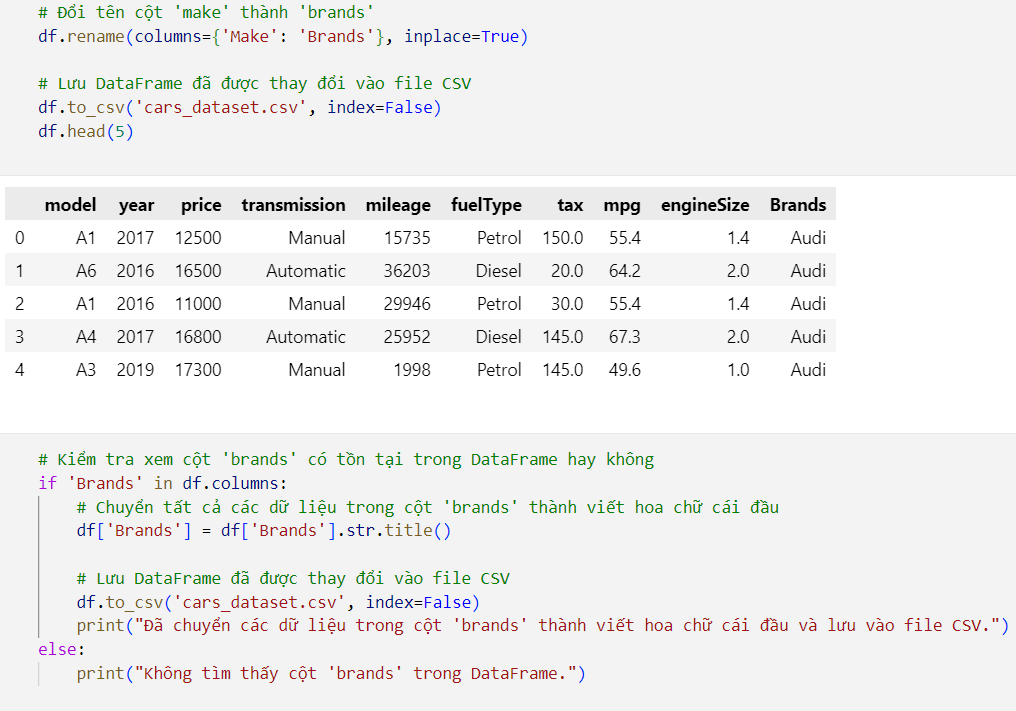


1. **Chuẩn hóa dữ liệu:**



Việc này giúp đưa các đặc trưng về cùng một phạm vi giá trị và cải thiện hiệu suất của mô hình học.

1. **Đổi giá trị dữ liệu cột trong file dữ liệu:**

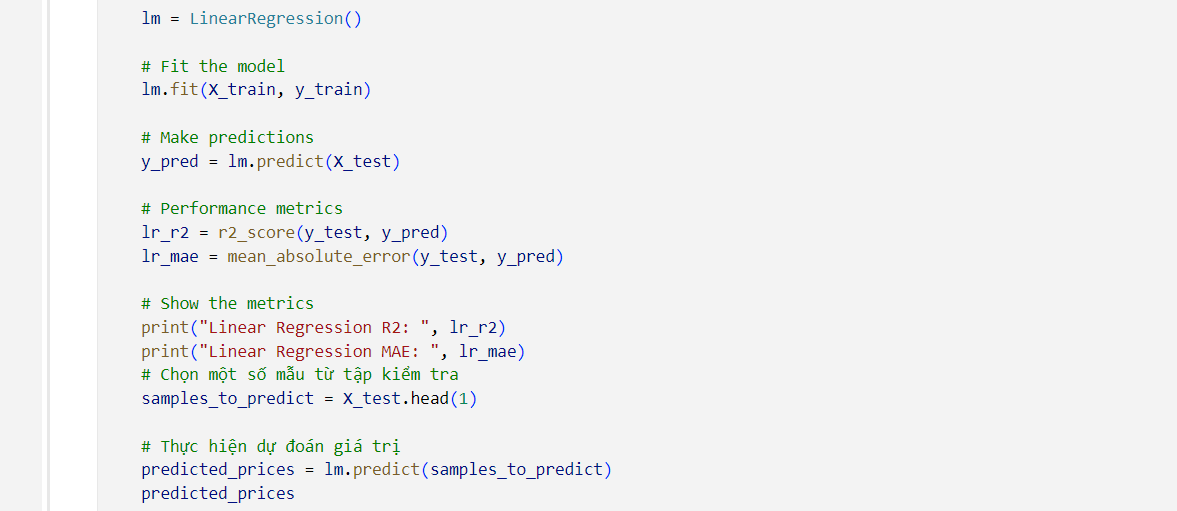


Đây là một trong những tiền xử lý mà chúng ta cần làm đi bắt đầu vào mô hình dữ liệu hay là vẽ biểu đồ phân tích

1. **Xây dựng mô hình:**

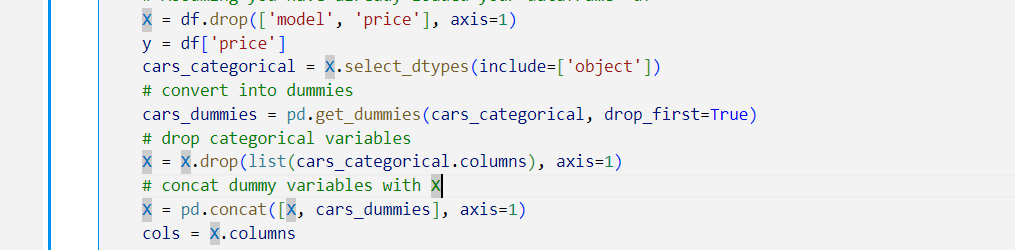
Ta sử dụng Hồi Quy Tuyến Tính để xác định mối quan hệ giữa các biến độc lập với biến mục tiêu. Qua đó, mô hình được huấn luyện trên tập dữ liệu huấn luyện và được đánh giá trên tập dữ liệu kiểm thử.

Hồi quy tuyến tính là một phương pháp mạnh mẽ trong thống kê và học máy, chủ yếu được sử dụng để xác định mối quan hệ giữa biến phụ thuộc và một hoặc nhiều biến độc lập. Đây là một cách hiệu quả để mô hình hóa dữ liệu và dự đoán giá trị dựa trên mối quan hệ tuyến tính giữa các biến.

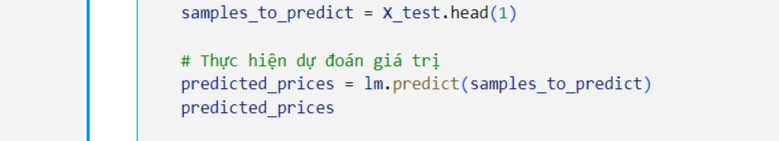


Feature: year, transmission, mileage, fuelType, tax, mpg, engineSize, Brands

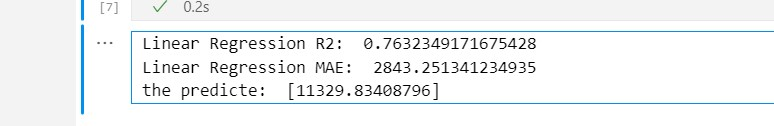
Target: việc chọn "price" và “model” làm biến target giúp bạn tập trung vào việc tạo ra một mô hình có khả năng dự đoán giá trị của xe dựa trên các đặc tính khác nhau.



1. **Đánh giá mô hình**
2. **Set cứng 1 dữ liệu để đánh giá mô hình:**



1. **Kết Quả:**



Lưu ý: Những mô hình có giá trị tốt nghĩa là “Mean Squared Error” và “RMSE” càng thấp thì mô hình đó càng phù hợp với tập dữ liệu của chúng ta.

1. **Phân Tích Kết Quả**

Dựa trên kết quả price có kết quả chính xác là 12000, mô hình Linear Regression cho kết quả khoảng [11329.83408796]

Phân tích kết quả cho thấy mô hình học máy của ta có khả năng dự đoán price và model có độ chính xác dựa trên Mean Squared Error và R2

1. **Đề Xuất Ứng Dụng Và Giải Pháp:**

Dựa vào kết quả phân tích ở trên tôi đề xuất cần tối ưu hóa quản lý tài chính để không có sự sai lệch về giá cả và phải cập nhật dự đoán cho các nhà đầu tư và doanh nghiệp để có thể quản lý chiến lược đầu tư vào thị trường ô tô.

1. **Tài liệu Kham khảo:**
2. **Sách kham khảo**

“Data Science for Business” của Forster Provost và Tom Fawcett

“Python for Data Analysis” của Wes McKinney

“Tidy Data” của Hadley Wickham

1. **Các nguồn video kham khảo**

Data school

StarQuest with Josh Starmer

Corey Schafer