ĐỒ ÁN CUỐI KÌ KHOA HỌC DỮ LIỆU

NỘI DUNG: DỰ ĐOÁN GIÁ Ô TÔ

NHÓM 9

HỨA PHÚ THÀNH - 18120563

TRẦN LUẬT VY - 18120656

GDVH: TRẦN TRUNG KIÊN

NỘI DUNG

1. CÂU HỎI ĐẶT RA
2. THU THẬP DỮ LIỆU
3. KHÁM PHÁ DỮ LIỆU
4. TIỀN XỬ LÝ
5. MÔ HÌNH HÓA
6. NHÌN LẠI QUÁ TRÌNH

1. CÂU HỎI ĐẶT RA

- Vấn đề:
  - Dự đoán giá xe ô tô từ các thuộc tính như **hãng sản xuất, số chỗ ngồi, công suất động cơ**,…
- Tác dụng:
  - Giúp người dùng có thể ước lượng được chi phí mua xe
  - Đưa ra những lựa chọn tốt nhất phù hợp với mục đích sử dụng
  - Giúp doanh nghiệp cạnh tranh trên thị trường

2. THU THẬP DỮ LIỆU

- **Nguồn dữ liệu:** www.cars-data.com
- **Cách thức:** parse HTML
- **Thư viện dùng:** HTMLSession và JSON
- **Nội dung trang web:**
  - Hình ảnh, mẫu mã ô tô
  - Giá cả
  - Thông số kĩ thuật

2. THU THẬP DỮ LIỆU

- ✓ Check file robots.txt
- ✓ Vào link www.cars-data.com/en/all-cars.html lấy url các page chứa thông tin xe
- ✓ Vào page nói trên lấy url của các xe cụ thể
- ✓ Vào url các xe cụ thể chọn dòng xe (lấy …)
- ✓ Lấy thông tin chi tiết

2. THU THẬP DỮ LIỆU

- Dữ liệu sau khi thu thập còn thô cần được xử lý

	url	name	model	brand	eLabel	bodyType	length	height	width	weight	...	fuelConsumption
0	http://www.cars-data.com//en/audi-rs4-avant-2....	Audi RS4 Avant 2.9 TFSI quattro	Audi RS4 Avant	Audi	G	stationwagon	4781	1404 mm	1866 mm	1730 kg	...	9,2 l/100km
1	http://www.cars-data.com//en/audi-rs5-sportbac...	Audi RS5 Sportback 2.9 TFSI quattro	Audi RS5 Sportback	Audi	G	hatchback	4783	1399 mm	1866 mm	1695 kg	...	9,1 l/100km
2	http://www.cars-data.com//en/audi-tt-coupe-40-...	Audi TT Coupe 40 TFSI	Audi TT Coupe	Audi	D	coupe	4191	1376 mm	1832 mm	1245 kg	...	6,0 l/100km
3	http://www.cars-data.com//en/audi-tt-coupe-45-...	Audi TT Coupe 45 TFSI	Audi TT Coupe	Audi	E	coupe	4191	1376 mm	1832 mm	1225 kg	...	6,5 l/100km
4	http://www.cars-data.com//en/audi-tt-coupe-45-...	Audi TT Coupe 45 TFSI	Audi TT Coupe	Audi	E	coupe	4191	1376 mm	1832 mm	1260 kg	...	6,4 l/100km

3. KHÁM PHÁ DỮ LIỆU

- Dữ liệu gồm: 36 thuộc tính, 45452 đối tượng
- Output missing: 143

	Name_col	Description
0	url	Link xem thông tin chi tiết của xe
1	name	Tên xe
2	model	Dòng xe
3	brand	Thương hiệu
4	eLabel	Nhãn năng lượng, cho biết hiệu suất năng lượng...
5	bodyType	Loại thân xe (station wagon, hatch back, coupe...
6	length	Chiều dài xe (mm)
7	height	Chiều cao xe (mm)
8	width	Chiều rộng xe (mm)
9	weightTotal	Trọng lượng xe (kg)
10	emissonsCO2	Lượng khí thải CO2 (g/km)
11	modelData	Năm sản xuất của dòng xe
12	fuelType	Loại nhiên liệu sử dụng
13	numbnumberOfAxleser	Số trục
14	numberOfDoors	Số lượng cửa
15	numberOfForwardGears	Cấp của hộp số
16	seatingCapacity	Số ghế ngồi
17	vehicleTransmission	Hệ thống chuyển số của xe
18	cargoVolume	Dung tích xe (l)
19	roofLoad	Tải trọng của đồ tối đa để trên nóc xe
20	accelerationTime	Thời gian tăng tốc (s)
21	driveWheelConfiguration	Cấu hình bánh xe lái
22	fuelCapacity	Dung tích nhiên liệu (l)
23	fuelConsumption	Độ tiêu hao nhiên liệu (l/km)
24	speed	Tốc độ (km/h)
25	payload	Tải trọng xe (kg)
26	trailerWeight	Khối lượng tối đa của xe móc kéo mà xe có thể kéo
27	vEengineType	Loại động cơ của xe
28	vEfuelType	Loại nhiên liệu
29	vEengineDisplacement	Dung tích xi lanh của xe
30	vEenginePower	Công suất
31	torque	Mô men xoắn (Nm)
32	price	Giá xe

4. TIỀN XỬ LÝ

- Tách các tập dữ liệu
  - Tập train: 60%
  - Tập validation: 20%
  - Tập test: 20%
- Xóa các dòng thiếu output
- Xử lý các thuộc tính numeric
  - Xử lý đơn vị đo, chuyển sang dạng số
- Xử lý các thuộc tính dạng categorical

4. TIỀN XỬ LÝ

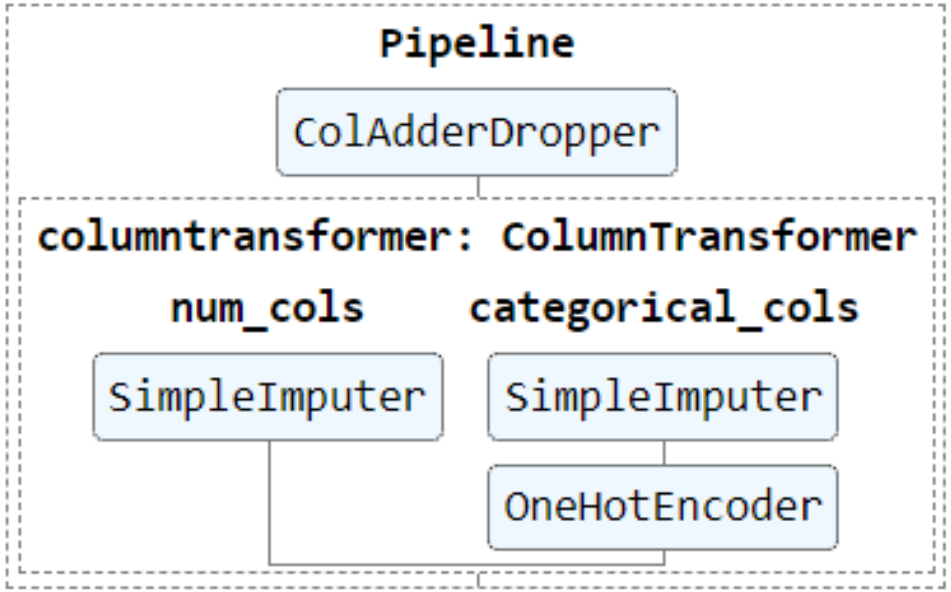
- Loại bỏ các thuộc tính không cần thiết
  1. Các cột **url, name, model** sẽ được loại bỏ vì không cần thiết
  2. Cột **VehicleTransmission** chỉ tồn tại duy nhất 1 giá trị, nên khi khai thác sẽ không có tác dụng
  3. Hai cột **fuelType** và **vEfuelType** giống nhau 99%, xóa 1 cột.
  4. Cột **cargoVolume** có rất nhiều giá trị rác khó xử lí nên loại bỏ.
  5. Cột **modelDate** chỉ gồm năm sản xuất, nên không có ý nghĩa cho dự đoán cần loại bỏ.

4. TIỀN XỬ LÝ

- Điền giá trị thiếu
  - Mean: thuộc tính numeric
  - Mode: thuộc tính categorical
  - OneHotEncoder cho thuộc tính categorical không có thứ tự
  - Không tồn tại thuộc tính categorical có thứ tự
- Dùng StandardScaler scale dữ liệu, mục đích nhằm **hội tụ nhanh**

5. MÔ HÌNH HÓA

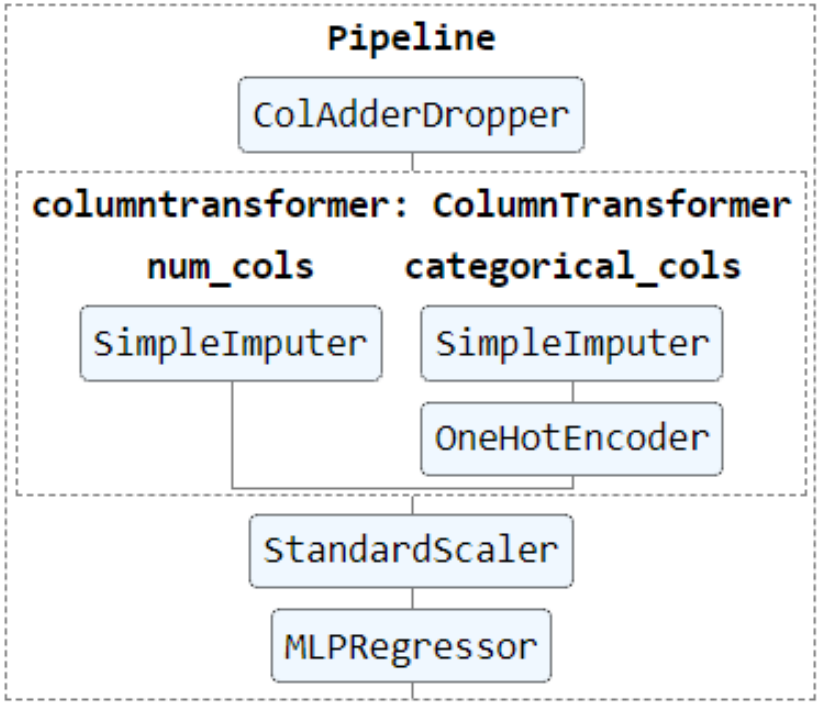
- Tổng quan pipeline tiền xử lý
  - ColAdderDropper
  - SimpleImputer
  - OneHotEncoder
  - StandardScaler



5. MÔ HÌNH HÓA

- **MLPRegressor**

```
MLPRegressor(hidden_layer_sizes=(512, 512, ),
             solver='adam', learning_rate='adaptive'\
             ,random_state=0, max_iter=500, early_stopping=True, verbose=0)
```



5. MÔ HÌNH HÓA

- MLPRegressor chạy trên tập train và validation
  - **hidden_layers** = [(50,), (100,), (150,), (50,50,), (50,50,50,)]
  - **num_top_brands** = [5, 10, 15, 20, 25]

5. MÔ HÌNH HÓA

- Siêu tham số tốt nhất

  best_num_top_brand = 20
  best_hidden_layer = (150,)

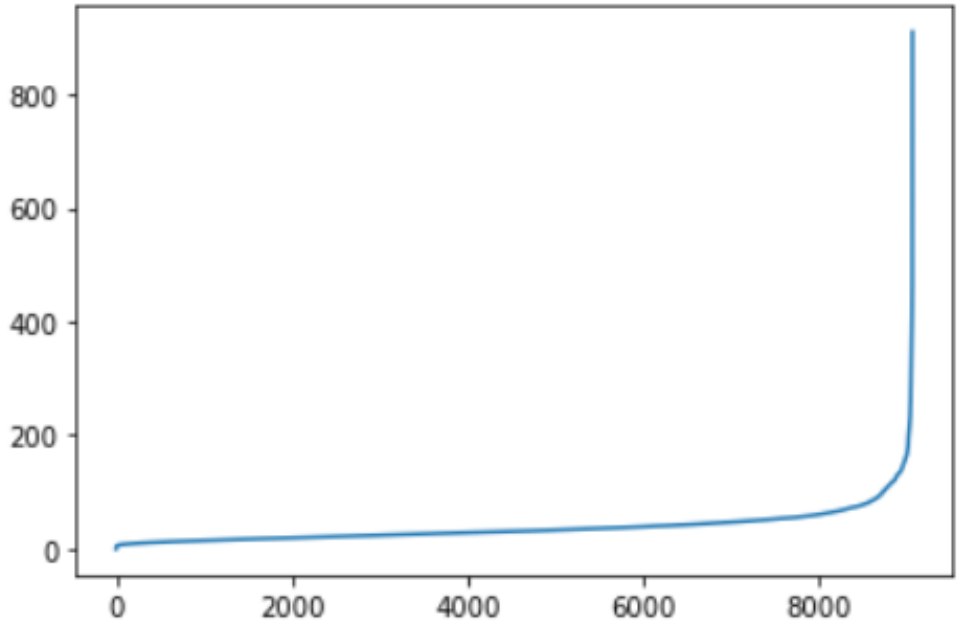
- Tiến hành chạy trên tập test

5. MÔ HÌNH HÓA

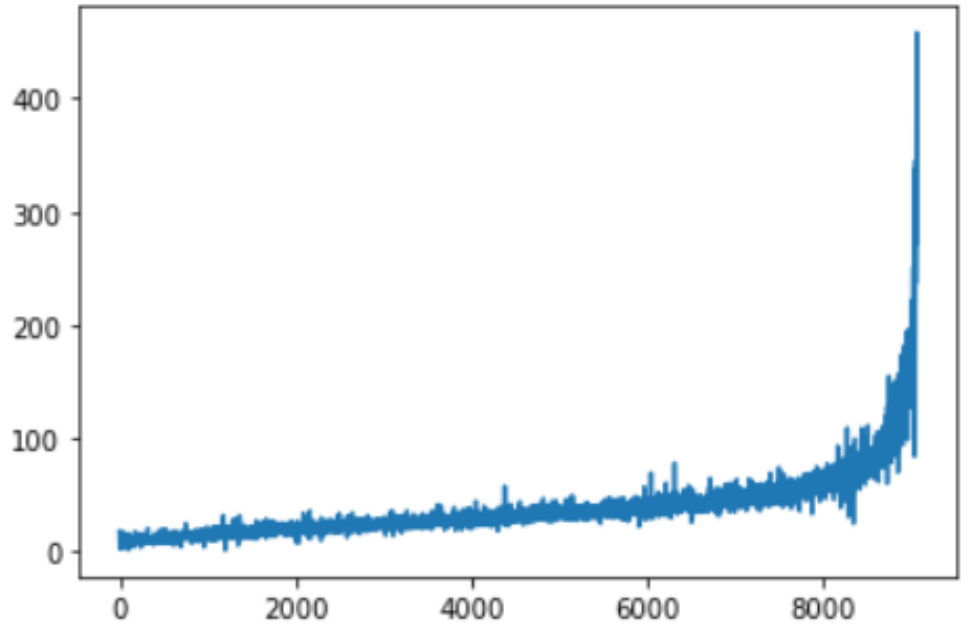
- Kết quả trên tập test:

```
pred_y=full_pipeline.predict(test_X_df)
(1 - full_pipeline.score(test_X_df, test_y_price))*100
```

```
8.487421409857976
```



Actual



Predict

6. NHÌN LẠI QUÁ TRÌNH

- Khó khăn gặp phải
  1. Việc chọn chủ đề (có quá nhiều chủ đề để chọn)
  2. Lấy dữ liệu ở đâu (nguồn dữ liệu khá nhiều, nhưng cái nào tốt)
  3. Dữ liệu khá lớn thời gian parse lâu
  4. Khâu tiền xử lý khá phức tạp
  5. Chọn mô hình và các siêu tham số

6. NHÌN LẠI QUÁ TRÌNH

- Những điều học được
  1. Biết thêm được thư viện HTMLSession
  2. Nâng cao khả năng sử dụng Jupyter notebook
  3. Hiểu rõ hơn về quy trình tìm dữ liệu và huấn luyện mô hình
  4. **Không có việc gì khó, chỉ sợ lòng không bền**

6. NHÌN LẠI QUÁ TRÌNH

- Dự định nếu có thêm thời gian
  - Giá trên do nhà sản xuất đưa ra chưa tính thuế
  - Tính giá xe dựa trên thuế, phí lăn bánh của các quốc gia

Tài liệu tham khảo và link đồ án

- Tài liệu tham khảo
  - https://drive.google.com/drive/folders/1HsO9vSWpbp1xfpOa7zA4MzIZbSRV7jw
  - https://github.com/hmhuan/Data-science-project
- Link đồ án
  - https://github.com/huasen07/Data_Science_Final_Project