ĐỒ ÁN CUỐI KỲ
NHẬP MÔN KHOA HỌC DỮ LIỆU

GV: thầy Trần Trung Kiên

DỰ ĐOÁN GIÁ XE ÔTÔ

NHÓM 17

18120066 – Bùi Đoàn Hữu Nhân

18120097 – Đinh Hữu Phúc Trung

NỘI DUNG

1. Giới thiệu chủ đề
2. Thu thập dữ liệu
3. Khám phá dữ liệu ban đầu
4. Tách tập và khám phá dữ liệu
5. Tiền xử lý và mô hình hóa dữ liệu
6. Đánh giá mô hình
7. Nhìn lại quá trình làm đồ án
8. Tài liệu tham khảo

1. Giới Thiệu Chủ Đề

Câu hỏi: Giá ôtô được tính từ các đặc trưng, bộ phận của ôtô theo công thức nào?

Lợi ích:

- Về phía người bán hoặc nhà phân phối: giúp họ định giá được xe ôtô dựa vào cấu tạo của xe.
- Về phía khách hàng (người mua): giúp người mua nắm được giá xe ôtô đúng với chất lượng thực sự của nó.

1. Giới Thiệu Chủ Đề

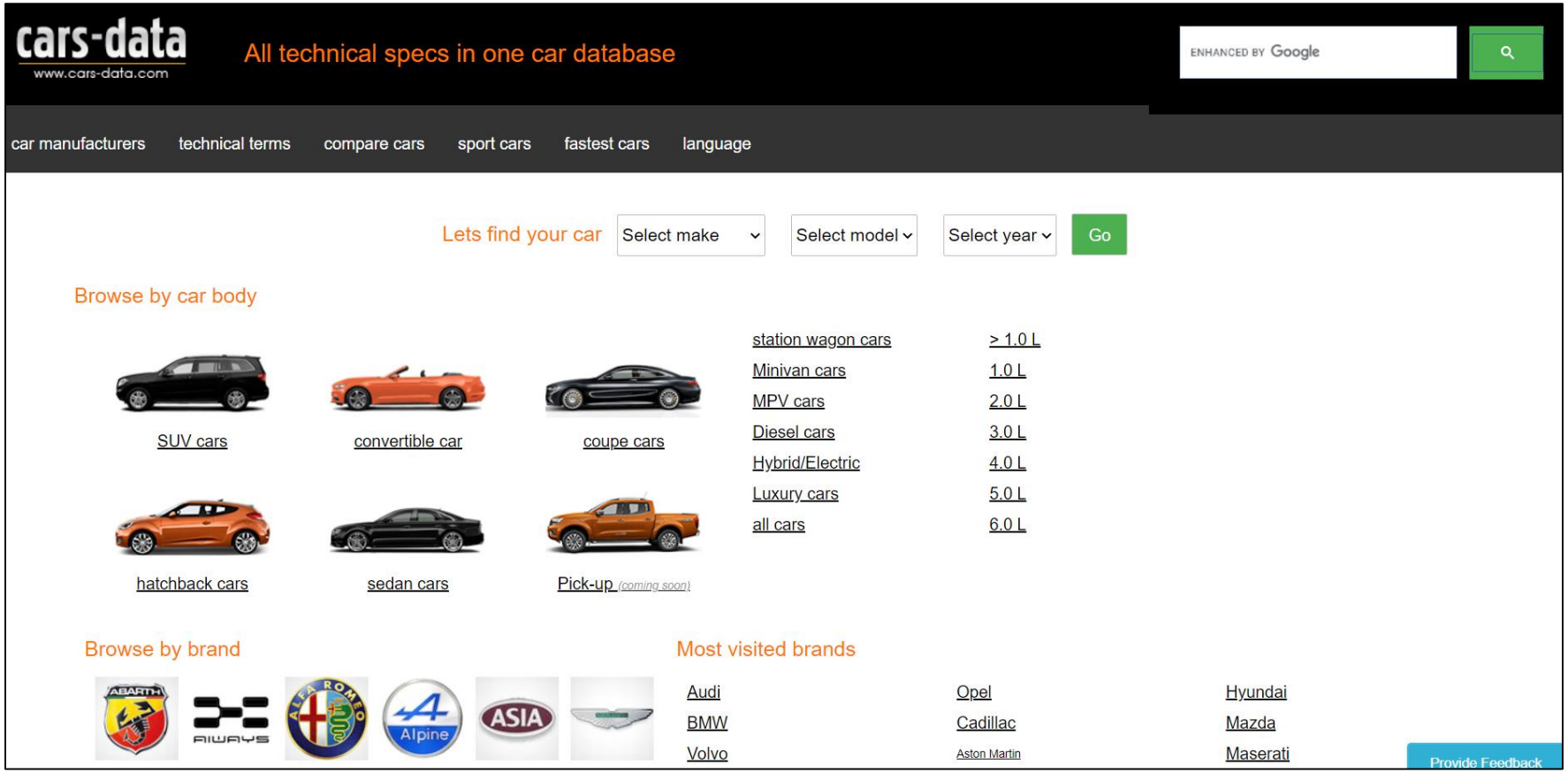
Nguồn gốc chủ đề: Nhóm tự nghĩa ra.

Cảm hứng: Từ việc có ý định mua xe ôtô từ lâu nên đã đề xuất chủ đề này để tận dụng cơ hội tìm hiểu về cấu tạo và giá xe ôtô, và với mong muốn sẽ tìm được câu trả lời chính xác cho câu hỏi trên để sau này mua xe với giá hợp lý nhất.

2. Thu Thập Dữ Liệu

Trang web dùng để crawl dữ liệu:

[Car specs database | cars-data.com (cars-data.com)]



2. Thu Thập Dữ Liệu

- Sử dụng **Selenium** để thu thập dữ liệu từ trang web.
- **Các bước thu thập dữ liệu:**
  - Thu thập link chứa thông tin của ôtô trong các đường link sau:

    https://www.cars-data.com/en/all-cars/page<x>.html
  - Trong các link thu thập ở trên, sẽ có một số link chỉ chứa các option của các xe cùng loại (như hình dưới), tiến hành thu thập tiếp các option này.

2. Thu Thập Dữ Liệu

- **Các bước thu thập dữ liệu:**
  - Với mỗi link thu thập được thêm vào “/tech” thì sẽ thu được link dẫn đến bảng thông số kỹ thuật chi tiết của xe ôtô đó.
  - Cuối cùng lựa chọn các thông tin cần thiết và lưu vào dataframe.
- Dữ liệu thu thập được từ các bước trên ở dạng rất thô, ta sẽ tiến hành xử lý các vấn đề sau trước khi lưu dữ liệu về:
  - Định dạng các cột về đúng kiểu dữ liệu.
  - Tách các cột chứa nhiều thông tin ra thành nhiều cột hoặc chỉ lấy thông tin cần thiết trong cột đó.

3. Khám phá dữ liệu

- Dữ liệu gồm **4529** dòng và **24** cột.
- Mỗi dòng chứa thông tin của một xe ôtô.
- Ý nghĩa của mỗi cột là:

Name	Tên xe
Brand	Hãng xe
Price	Giá xe
Body	Loại thân xe
Transmission	Số cấp của hộp số
Number Of Seats	Số chỗ ngồi
Segment	Loại kích cỡ xe
Introduction	Năm sản xuất
Drive	Hệ thống dẫn động
Drive System	Loại động cơ
Fuel	Loại nhiên liệu
Cylinder Capacity	Dung tích xilanh

Max Power Hp	Công suất tối đa đơn vị là mã lực
Max Torque	Momen xoắn cực đại
Fuel System	Hệ thống nhiên liệu
Valve Actuation	Kiểu kích hoạt van
Turbo	Bộ tăng áp
Fuel Tank	Dung tích bình nhiên liệu
Top Speed	Tốc độ tối đa
Energy Label	Nhãn năng lượng
Front Stabilizer	Bộ ổn định phía trước
Rear Stabilizer	Bộ ổn định phía sau
Num_doors	Số cửa
Dt_Transmission	Loại hộp số

4. Tách Tập Và Khám Phá Dữ liệu

- Cột output không chứa giá trị thiếu, thực hiện tách tập dữ liệu thành 3 tập:
  - Tập huấn luyện (60%) gồm **2717** dòng.
  - Tập validation (20%) gồm **906** dòng.
  - Tập test (20%) gồm **906** dòng.

4. Tách Tập Và Khám Phá Dữ liệu

- Kiểu dữ liệu của các cột:

Name	object
Brand	object
Body	object
Transmission	float64
Number Of Seats	int64
Segment	object
Introduction	int64
Drive	object
Drive System	object
Fuel	object
Cylinder Capacity	float64

Max Power Hp	float64
Max Torque	float64
Fuel System	object
Valve Actuation	object
Turbo	object
Fuel Tank	float64
Top Speed	float64
Energy Label	object
Front Stabilizer	object
Rear Stabilizer	object
Num_doors	int64
Dt_Transmission	object

4. Tách Tập Và Khám Phá Dữ liệu

- Phân bố của các cột dữ liệu dạng số trong tập huấn luyện:

	Transmission	Number Of Seats	Introduction	Cylinder Capacity	Max Power Hp	Max Torque	Fuel Tank	Top Speed	Num_doors
missing_ratio	1.3	0.0	0.0	0.3	0.3	0.4	0.5	1.1	0.0
min	3.0	0.0	1969.0	0.0	29.0	38.0	22.0	110.0	2.0
lower_quartile	5.0	5.0	1994.0	2.0	80.0	120.0	50.0	163.0	3.0
median	5.0	5.0	2004.0	4.0	110.0	165.0	60.0	180.0	4.0
upper_quartile	6.0	5.0	2012.0	4.0	160.0	250.0	70.0	205.0	5.0
max	9.0	9.0	2021.0	5.0	1001.0	1250.0	121.0	407.0	5.0

4. Tách Tập Và Khám Phá Dữ liệu

- Phân bố của các cột dữ liệu không phải dạng số trong tập huấn luyện:

	Front Stabilizer	Fuel	Fuel System	Turbo	Drive	Drive System	Valve Actuation	Brand	Rear Stabilizer	Segment	Body	Name	Dt_Transmission	Energy Label
missing_ratio	0	0	0	0	0	0	0	0	0	12.6	0	0	0	31.3
num_values	2	4	8	3	3	2	3	83	2	14	11	2301	5	7
value_ratios	{'yes': 88.8, 'no': 11.2}	{'gasoline': 94.4, 'diesel': 5.5, 'gasoline / bio ethanol': 0.1, 'lpg / gasoline': 0.0}	{'multipoint injection': 56.6, 'direct injection': 19.0, 'carburettor': 14.1, 'common rail': 5.0, 'singplepoint injection': 4.5, 'injection': 0.3, 'indirect injection': 0.3, 'multijet': 0.2}	{'no': 74.5, 'yes': 24.8, 'mechanical': 0.7}	{'front': 65.0, 'rear': 21.5, 'front+rear': 13.5}	{'fuel engine': 98.8, 'hybrid': 1.2}	{'dohc': 61.8, 'ohc': 33.5, 'ohv': 4.7}	{'TOYOTA': 6.3, 'AUDI': 5.3, 'RENAULT': 4.8, 'MERCEDES': 4.7, 'FORD': 4.6, 'OPEL': 4.5, 'VOLKSWAGEN': 4.3, 'NISSAN': 4.1, 'PEUGEOT': 3.8, 'MAZDA': 3.4, 'BMW': 3.2, 'CITROEN': 3.0, 'HYUNDAI': 3.0, ...	{'yes': 70.6, 'no': 29.4}	{'c': 19.9, 'd': 15.1, 'b': 11.1, 'g': 9.9, 'e': 7.7, 'm': 6.9, 'l': 6.8, 'j': 6.4, 'a': 4.9, 'n': 4.0, 'h': 3.1, 'k': 2.0, 'f': 1.6, 'i': 0.6}	{' hatchback': 27.6, ' sedan': 18.7, ' suv/crossover': 12.4, ' station wagon': 11.7, ' coupé': 9.6, ' convertible': 7.8, ' mpv': 7.2, ' van': 2.9, ' bus': 0.9, ' pick-up': 0.7, ' double cabin': 0.4}	{'ALFA ROMEO 147 1 6 T SPARK 16V IMPRESSION': 0.2, 'PORSCHE 911 CARRERA COUPE': 0.2, 'BMW 520I': 0.2, 'OPEL VECTRA 1 6I GL': 0.2, 'AUDI A3 1 6 ATTRACTION': 0.2, 'TOYOTA AYGO 1 0 12V VVT I ACCESS':...	{'manual transmission': 83.8, 'automatic': 12.2, 'automatic with double clutch': 2.3, 'oze automatic': 1.3, 'semi-automatic': 0.4}	{'g': 55.9, 'e': 9.5, 'f': 8.9, 'd': 8.2, 'c': 7.9, 'b': 7.0, 'a': 2.5}

5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Class **ColAdderDropper** xóa và xử lý một số cột:
  - Bỏ cột "Segment" vì có nhiều giá trị thiếu, nhiều giá trị khác nhau và phần nào cột "Body" cũng đã chứa thông tin loại kích cỡ ở cột này.
  - Bỏ cột "Energy Label" vì có quá nhiều giá trị thiếu.
  - Bỏ cột "Name" vì có rất nhiều giá trị khác nhau, thêm vào đó cột này được suy diễn từ những cột khác.
  - Ở cột "Fuel System" và "Body" ta sẽ lấy top các giá trị xuất hiện nhiều nhất.

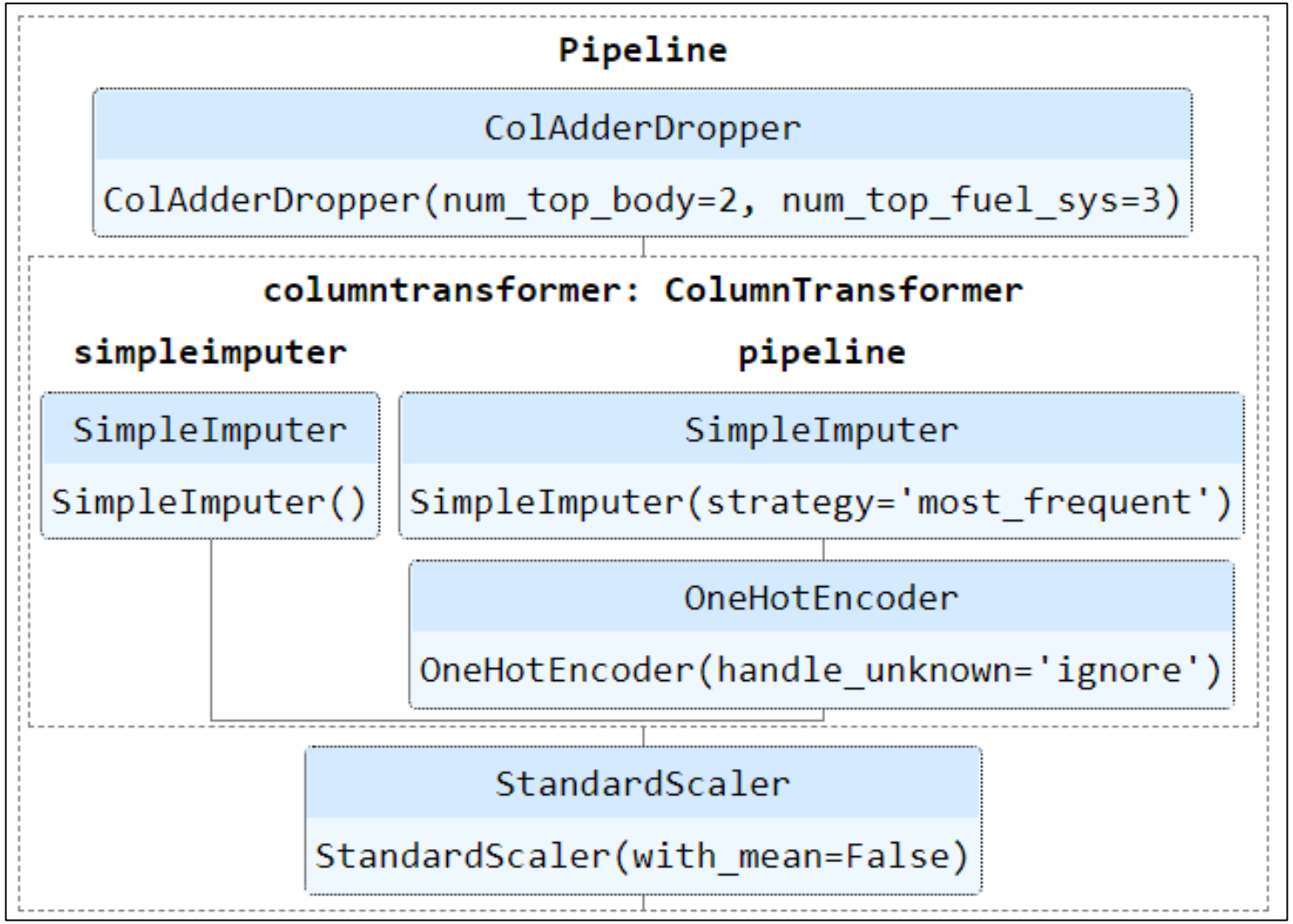
5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Tiếp theo, tạo **preprocess_pipeline** gồm các bước cài trong class **ColAdderDropper** và các bước sau:
  - Với các cột dạng số, điền giá trị thiếu bằng giá trị mean của cột dùng **SimpleImputer**.
  - Với các cột không phải ở dạng số, điền giá trị thiếu bằng giá trị mode (giá trị xuất hiện nhiều nhất) của cột dùng **SimpleImputer**, sau đó chuyển sang dạng số bằng phương pháp mã hóa one-hot (dùng **OneHotEncoder**).
  - Cuối cùng, tiến hành chuẩn hóa bằng cách trừ đi mean và chia cho độ lệch chuẩn của cột dùng **StandardScaler**.

5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Sơ đồ

  preprocess_pipeline

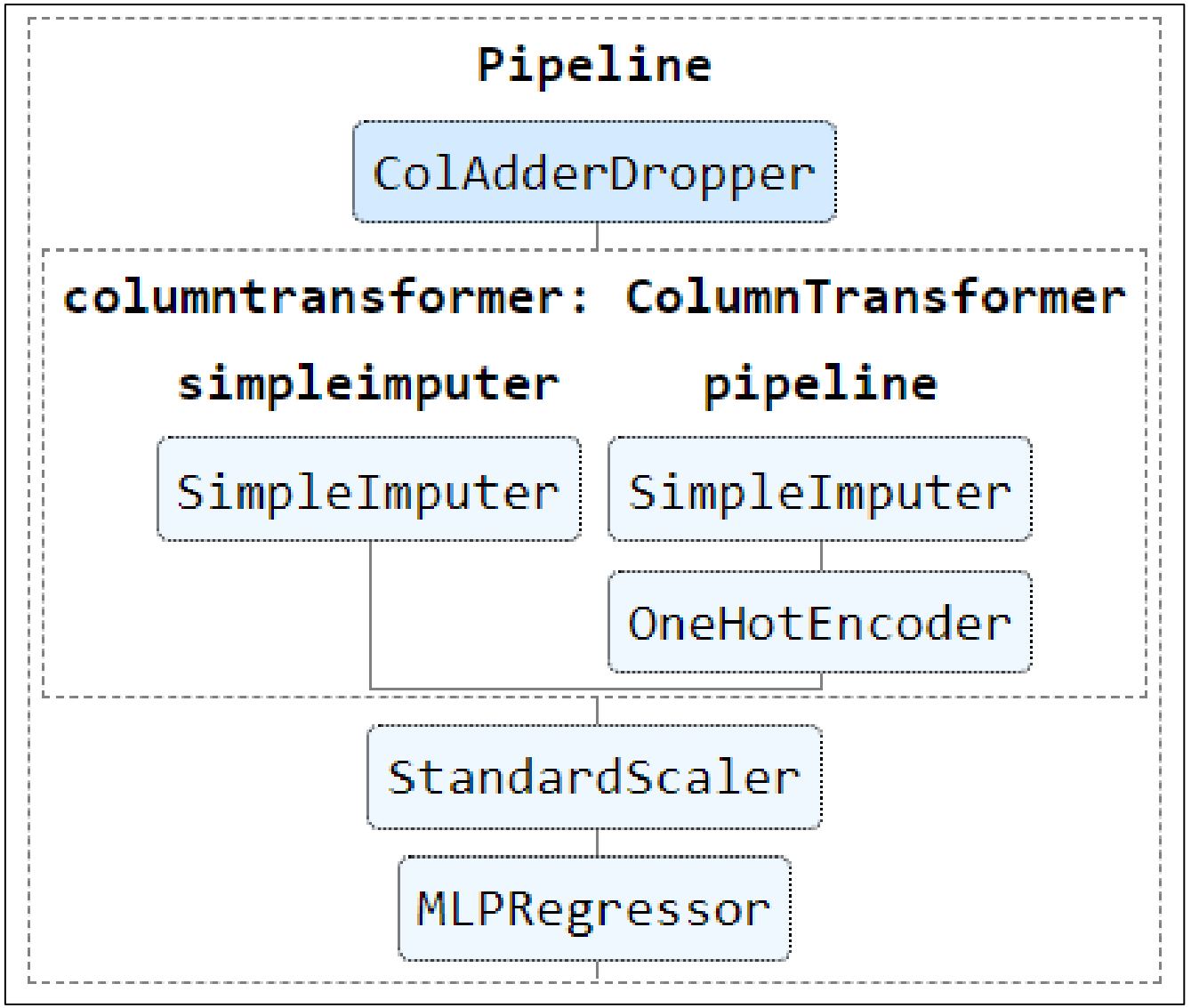


5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Sử dụng mô hình Neural Net để phân lớp, dùng **MLPRegressor** với các siêu tham số:
  - hidden_layer_sizes=(30)
  - activation='relu'
  - solver='adam'
  - max_iter=5000

5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Tạo pipeline **full_pipeline**



5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

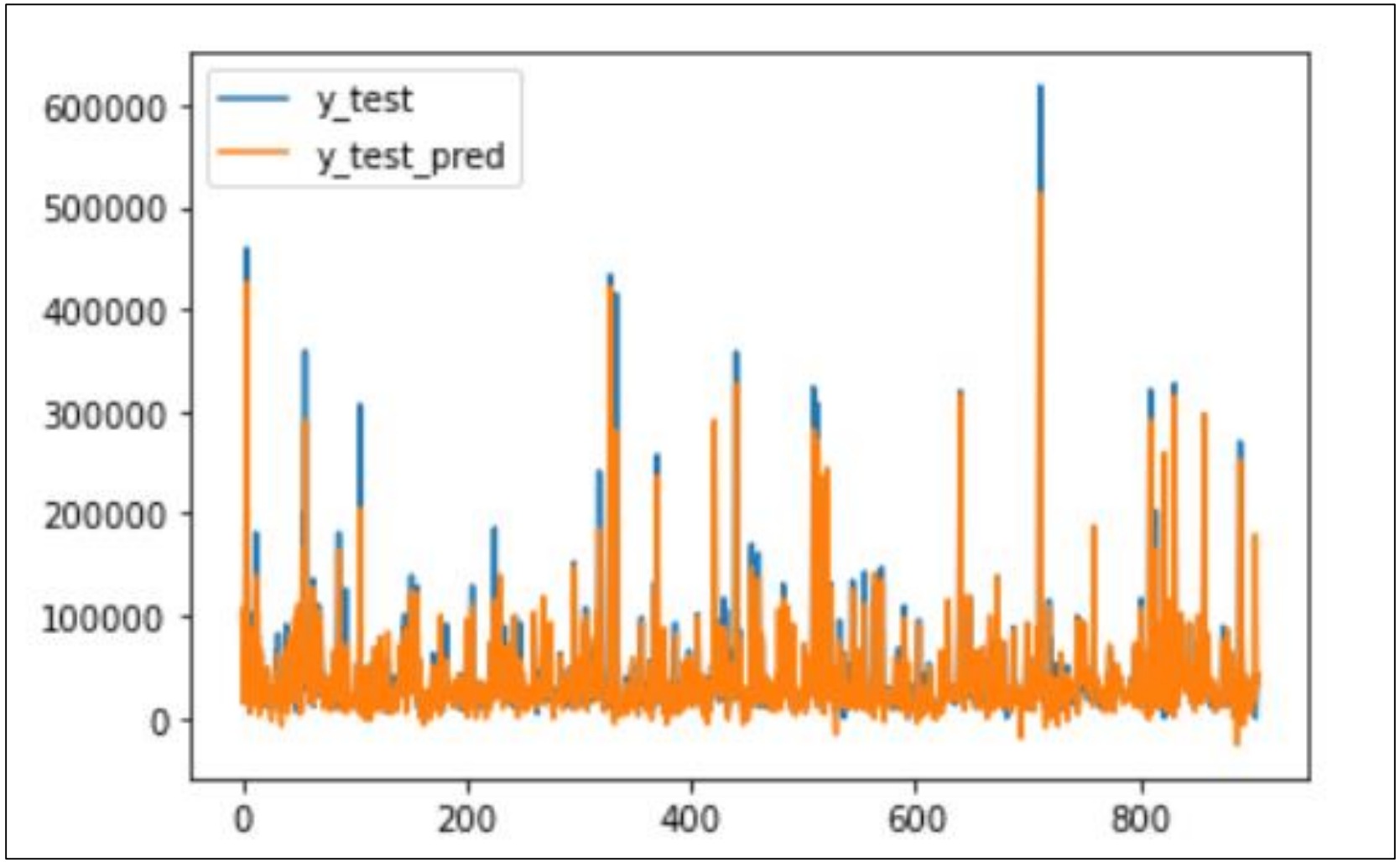
- Thử nghiệm với nhiều giá trị của **alpha**, **num_top_fuel_sys** và **num_top_body**:
  - **alpha**: mức độ L2 regularization hay weight decay của **MLPRegressor**.
  - **num_top_fuel_sys**: số các giá trị xuất hiện nhiều nhất của cột “Fuel System” mà ta muốn lấy.
  - **num_top_body**: số các giá trị xuất hiện nhiều nhất của cột “Body” mà ta muốn lấy.

5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Bộ siêu tham số tốt nhất trong lúc thử nghiệm thu được là:
  - **alpha=0.1**
  - **num_top_fuel_sys=6**
  - **num_top_body=4**
- Huấn luyện lại mô hình với bộ siêu tham số tốt nhất trên tập dữ liệu kết hợp (tập huấn luyện + tập validation) để tìm ra mô hình cuối cùng.

6. Đánh Giá Mô Hình

- Độ chính xác của mô hình thu được trên tập test là: ~89%



7. Nhìn Lại Quá Trình Làm Đồ Án

Đã gặp khó khăn gì?

- **Khó khăn trong việc thu thập dữ liệu:**
  - Những trang web có thông tin chi tiết cụ thể rõ ràng về xe ôtô thì đa số đều không cho crawl dữ liệu nên việc tìm và chọn một trang web ưng ý khá khó.
  - Dữ liệu thu thập được từ trang web lựa chọn thuộc dạng rất thô nên cần phải xử lý nhiều.

7. Nhìn Lại Quá Trình Làm Đồ Án

Đã gặp khó khăn gì?

- **Khó khăn trong việc mô hình hóa dữ liệu:**
  - Vì thời gian train của mô hình tương đối lâu nên việc chạy và tìm các siêu tham số tốt nhất cho mô hình tiêu tốn khá nhiều thời gian.

7. Nhìn Lại Quá Trình Làm Đồ Án

Có học được gì hữu ích?

- Rút ra được để có kết quả mô hình tốt thì cần phải tìm hiểu chủ đề thật kĩ để chọn ra các thông tin cần thiết cho câu trả lời.
- Ngoài việc trao dồi thêm kiến thức về môn học thì còn biết thêm về kiến thức thuộc chủ đề của nhóm (kiến thức về xe ôtô).

7. Nhìn Lại Quá Trình Làm Đồ Án

Nếu có thêm thời gian thì sẽ?

- Cải tiến mô hình để tăng độ chính xác và ổn định độ chính xác trên các tập dữ liệu khác nhau.
- Cải thiện thời gian train mô hình.
- Thử nghiệm nhiều mô hình hơn để tìm ra mô hình tốt nhất.

8. Tài Liệu Tham Khảo

- File “BT03-TienXuLy_MoHinhHoa.ipynb”
- Các file “03-Demo.ipynb”, “04-Demo.ipynb”, “07-Demo.ipynb”
- Trang scikit-learn 0.24.0 documentation
- Trang pandas documentation
- Trang stackoverflow
- Trang webdriver-chromium