DA44 FINAL TEST

BÀI TOÁN ĐỊNH GIÁ ÔTÔ

(CAR PRICE PREDICTION)

NGÀY 10/03/2024

GVHD: TRẦN ĐỨC TRUNG

HV: VÕ HỮU MINH CHÁNH

ĐẶT VẤN ĐỀ

Geely Auto ở Trung Quốc muốn thâm nhập thị trường Mỹ.

Họ cần hiểu các yếu tố ảnh hưởng đến việc định giá ô tô tại Mỹ, vì những yếu tố đó có thể rất khác so với thị trường Trung Quốc.

Cụ thể, họ muốn biết:

- Những yếu tố quan trọng để định giá ôtô?
- Các thông số nào sẽ ảnh hưởng đến giá?

DATABASE

DA_FINALTEST

Nguồn: MindX

Table dbo.CarPrice_Assignment
Kết nối database bằng thư viện pyodbc của Python.

Bộ dữ liệu tương đối sạch:
-205 hàng 26 cột (205 mẫu ôtô unique)
-Không có giá trị null
-Không có giá trị duplicates
-Tồn tại một số outlier do lỗi nhập liệu (kích thước, giá)

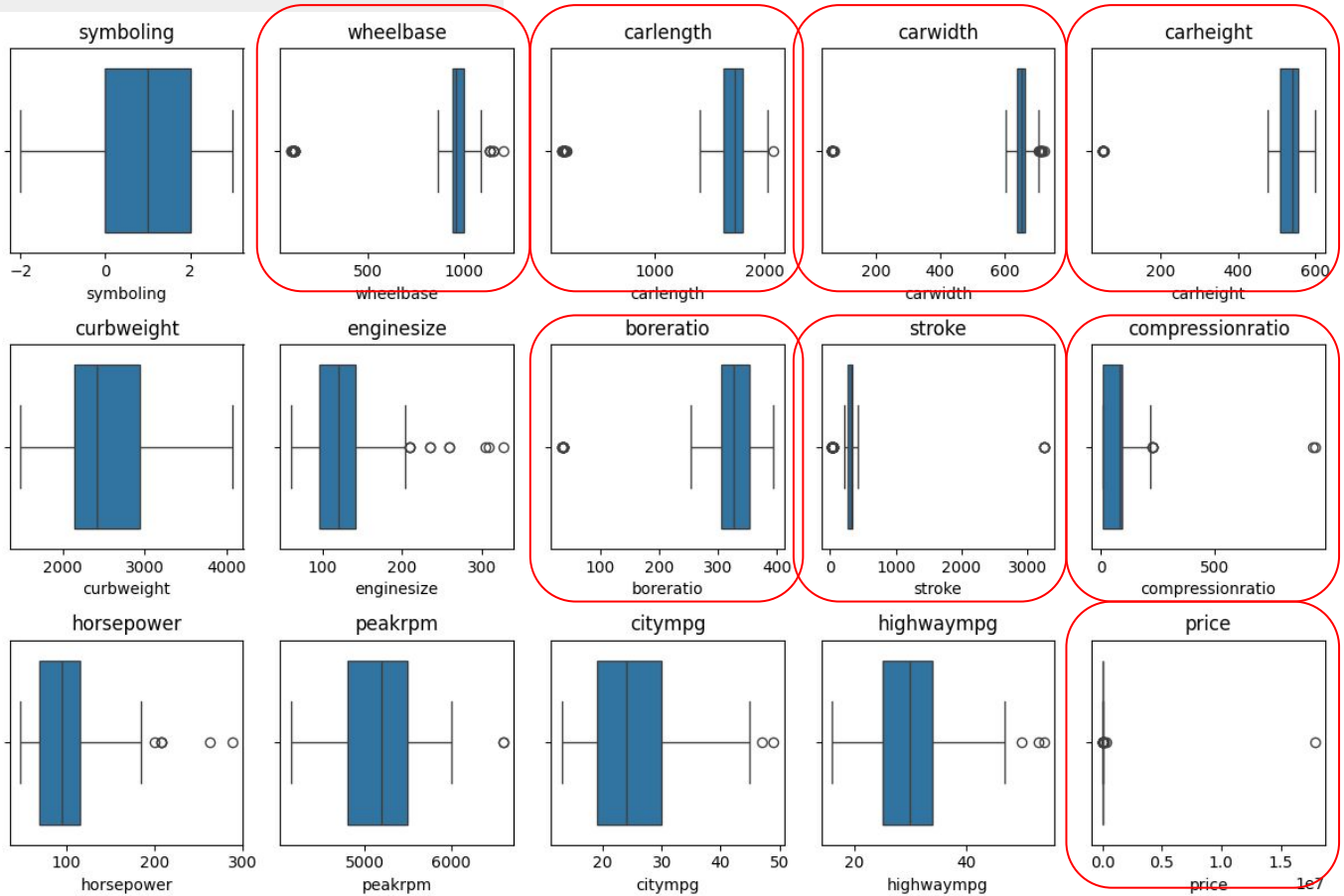
PREPROCESSING
&
EDA

PREPROCESSING

Outliers xuất hiện ở feature:
wheelbase, carlength, carwidth, carheight, boreratio, stroke, compressionratio
và
price

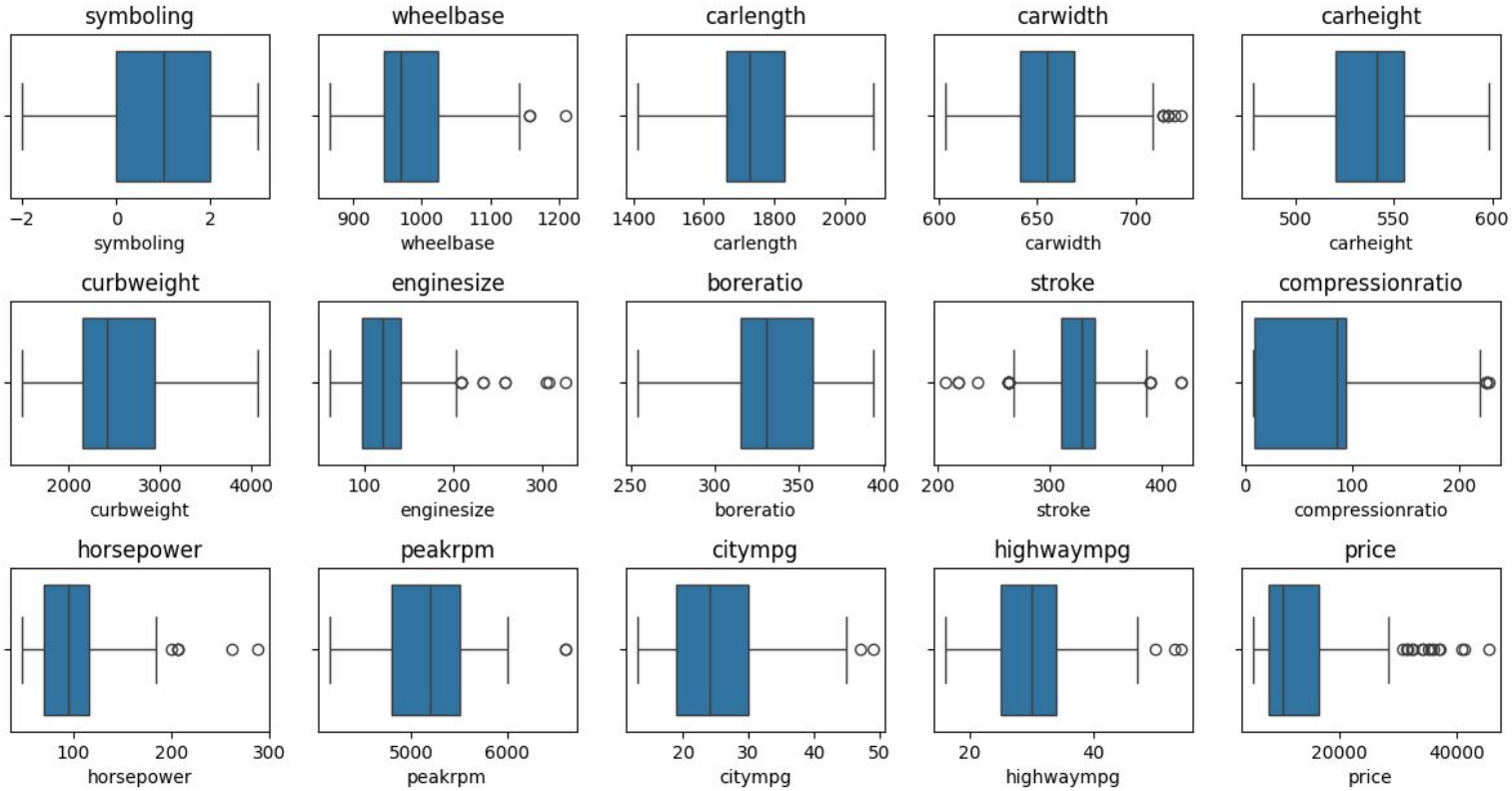
Dựa vào
- Số liệu trên db
- Thông số thực tế của hãng

=> Tất cả là lỗi nhập liệu (lệch 1 hoặc nhiều hàng đơn vị)
=> Replace outliers



PREPROCESSING

Kết quả xử lý



EDA

Loại bỏ các feature không sử dụng

car_ID

highwaympg bằng Correlation

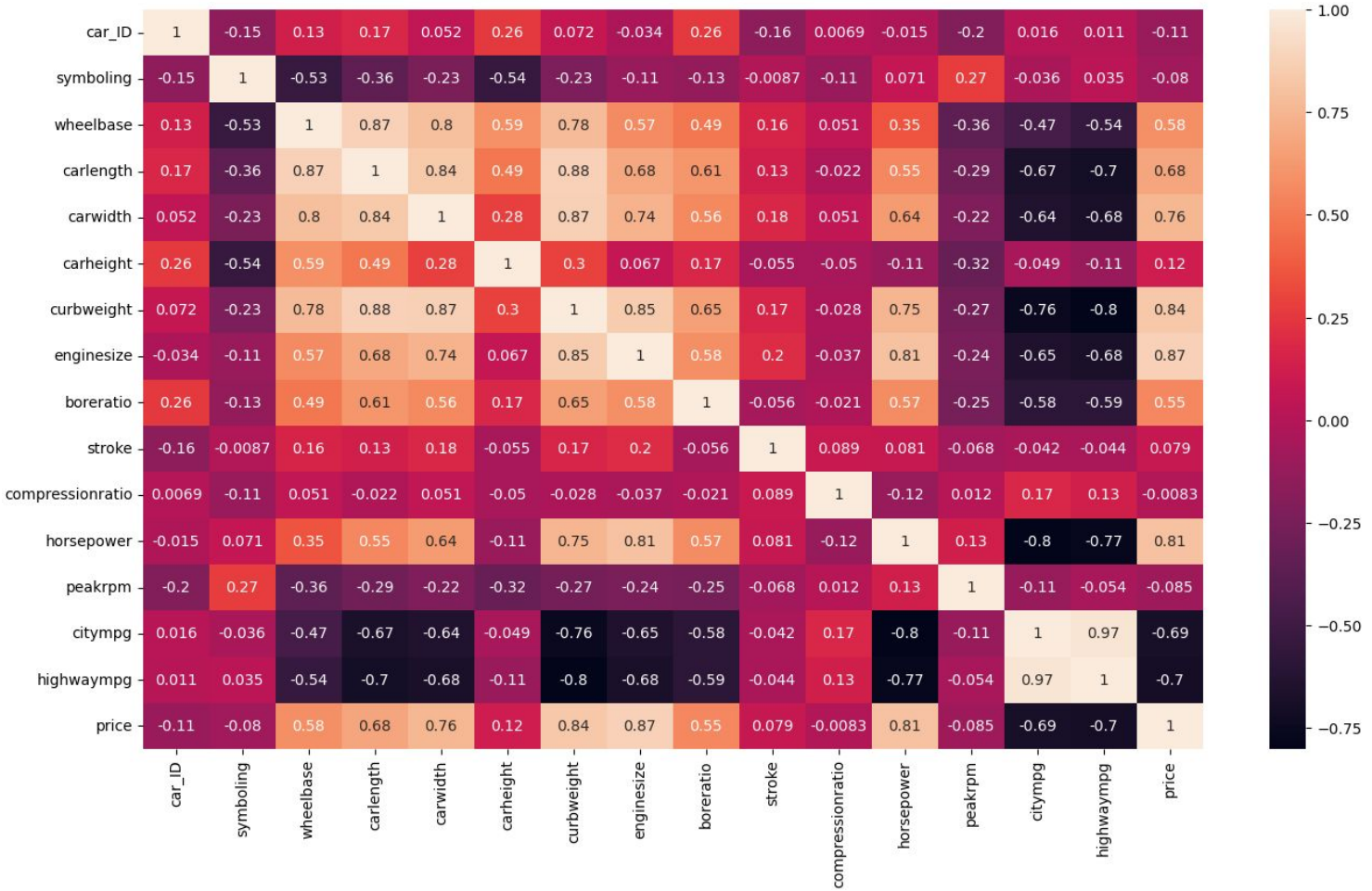
EDA các features rời rạc

Thêm CarBrand

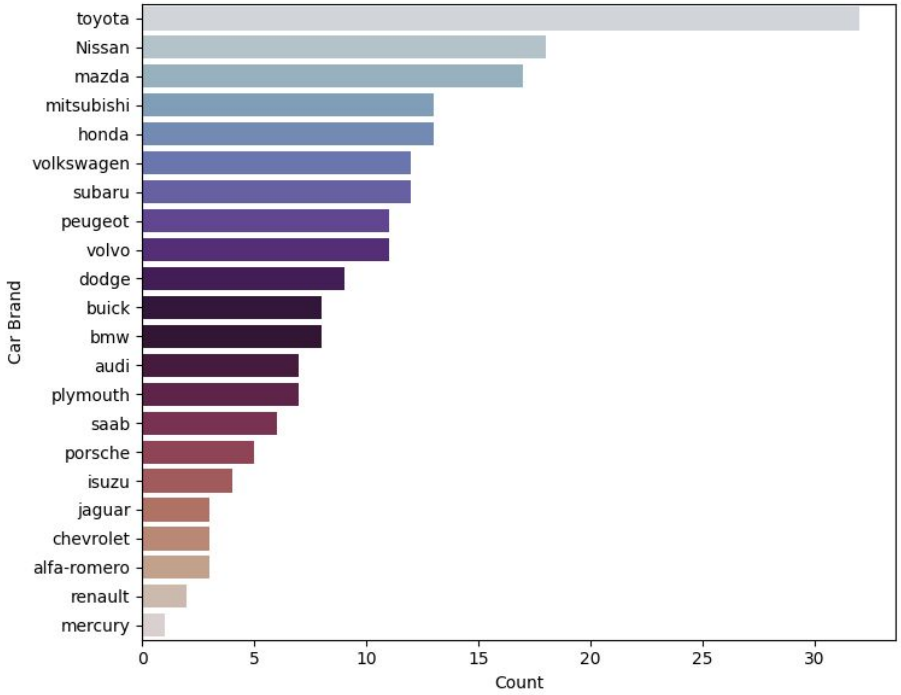
Các features khác

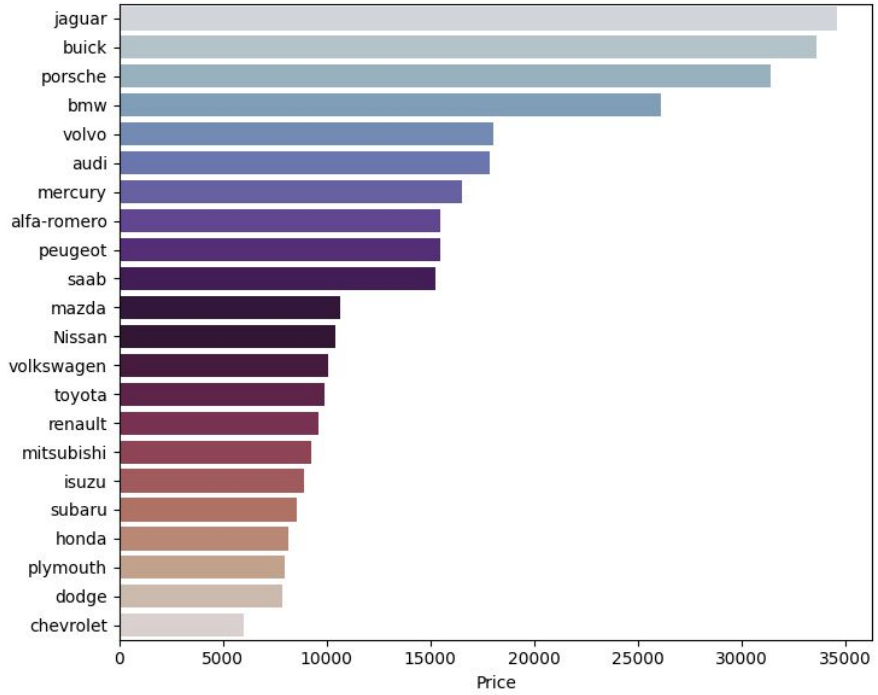
EDA các features liên tục

EDA



highwaypmg là feature có correlation chung khá cao, cần loại bỏ





Nhận xét CarBrand:

- Với số lượng vượt trội, **Toyota** là hãng xe được ưa chuộng nhất, giá ở mức trung bình thấp (top 9 dưới lên).
- **Jaguar** số lượng ít (top 5 dưới) do giá thành cao (#1). Tương tự **Mercury** có số lượng ít nhất, giá cao top 7.
- Đặc biệt **Chevrolet** số lượng ít (top 4 dưới) dù giá thấp nhất. Tương tự với isuzu.

⇒ Hầu hết các hãng xe đều có mức giá đi ngược với độ hiếm (số lượng). Với các hãng xe ít phổ biến như isuzu, chevrolet thì mức giá đi thuận với số lượng.

EDA

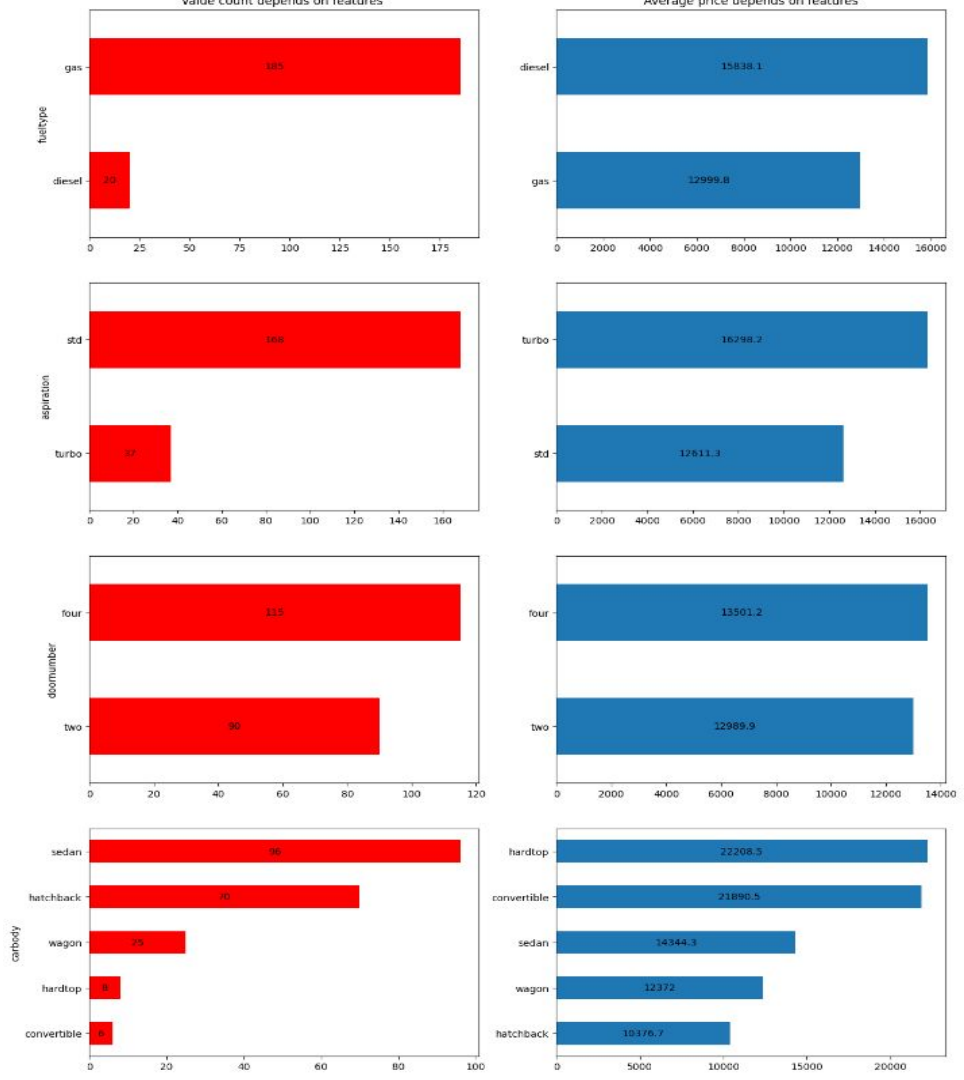
Nhận xét:

-fuel type:
Hầu hết các mẫu xe sử dụng xăng (gas), chiếm khoảng 90% (185/205) số lượng và giá xe xăng thấp hơn 17.9% xe chạy Diesel (12999.8/15838.1)

-aspiration
Xe động cơ tiêu chuẩn (std) chiếm khoảng 82% số lượng, giá hiển nhiên thấp hơn Turbo.

-doornumber
Xe 4 cửa chiếm phần lớn (56%), trung bình giá của 2 loại tương đương nhau (chênh lệch giá khoảng 3.7%)

-carbody
Đa phần người dân sử dụng xe là loại Sedan, còn top trung bình giá cao là Convertible và Hardtop.

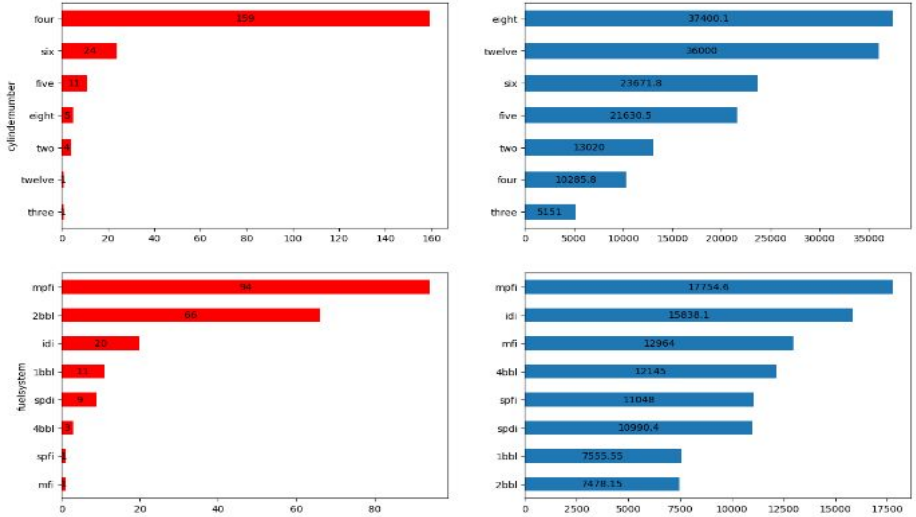


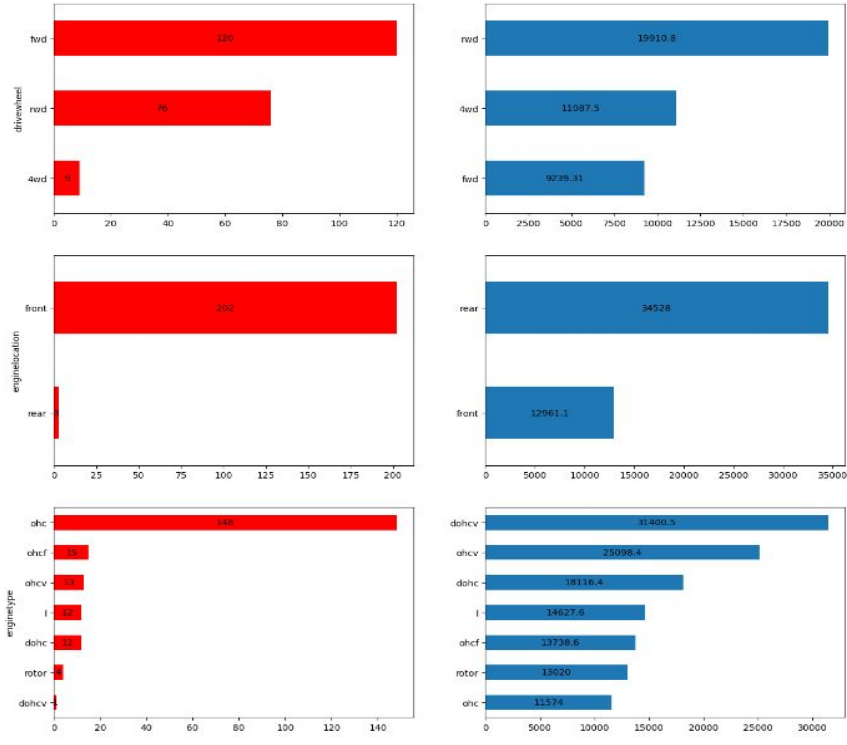
EDA

-drivewheel
Hầu hết đều là dẫn động cầu trước fwd, giá cũng thấp nhất. Tương tự với vị trí động cơ enginelocation và loại động cơ enginetype.

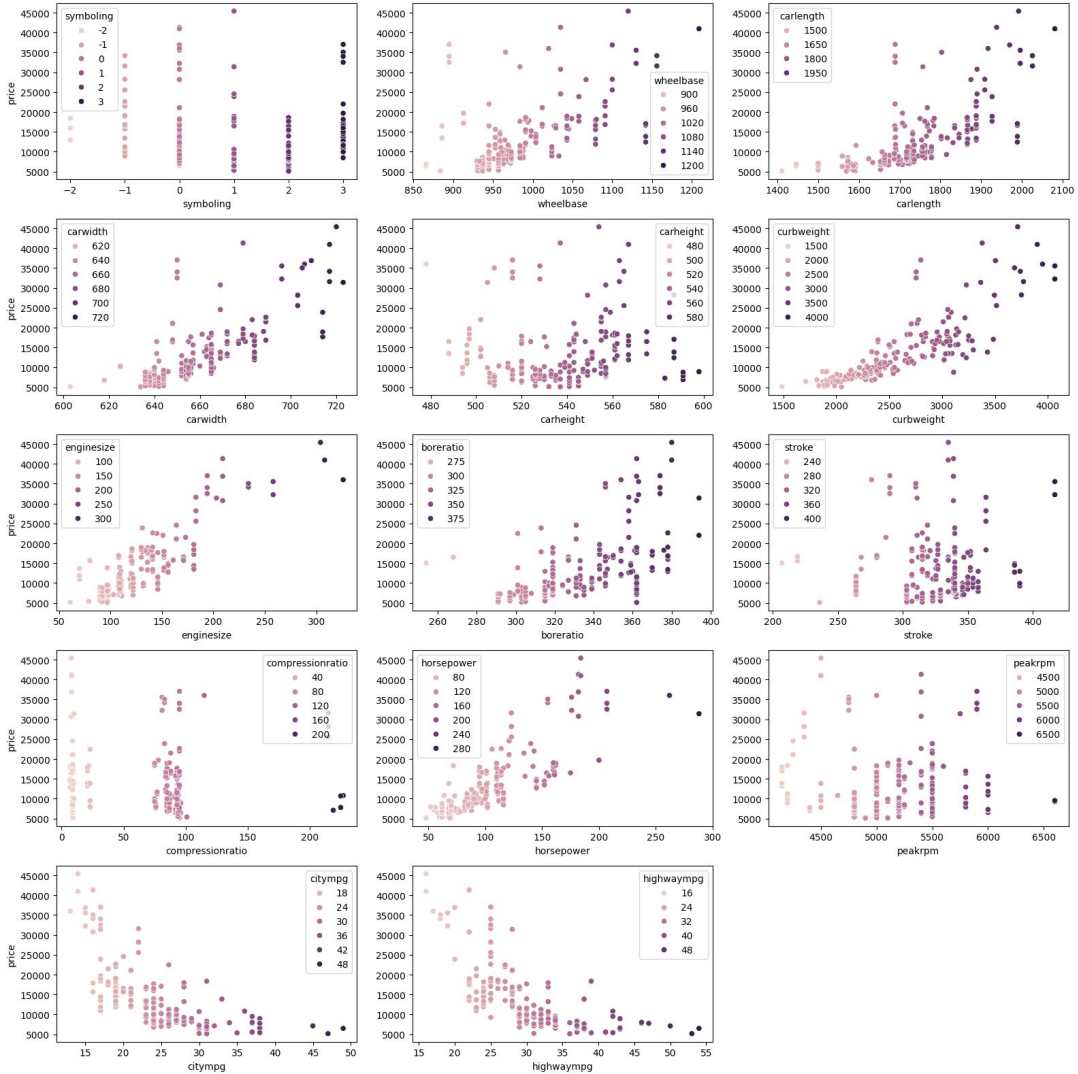
-cylindernumber
Số lượng xi-lanh tăng thì giá cũng tăng,
dòng 4 xi-lanh có số lượng vượt trội nhất.

-fuelsystem
Hệ thống nhiên liệu mpfi dẫn đầu về số lượng lẫn giá cả, còn 2bbl #2 về số lượng là loại có giá rẻ nhất.





EDA



Nhận xét:

-wheelbase, carlength, carwidth, curbweight, enginesize, boreratio, horsepower càng lớn thì giá càng cao.

-citympg càng thấp giá càng cao.

-carheight, stroke, compressionratio, peakrpm không phụ thuộc nhiều vào giá.

EDA

Kết quả phân tích mô tả:

Các feature rời rạc

- Dòng xe có số lượng cao nhất: Toyota.
- Hầu hết là xe xăng (gas), giá thấp hơn xe Diesel.
- Xe có aspiration tiêu chuẩn (std) chiếm phần lớn, nguyên nhân là do giá rẻ. Tương tự với drivewheel, enginelocation và enginetype.
- Xe 4 cửa chiếm ưu thế hơn một chút so với xe thể thao 2 cửa, giá 2 loại tương đương nhau.
- Đa phần là dáng Sedan ở mức giá trung bình, còn top giá là Convertible và Hardtop cũng là số lượng ít nhất.
- Số lượng xi-lanh tăng thì giá cũng tăng, dòng 4 xi-lanh có số lượng vượt trội nhất.
- Hệ thống nhiên liệu (fuelsystem) 'mpfi' dẫn đầu về số lượng lẫn giá cả, còn '2bbl' top 2 về số lượng là loại có giá thành rẻ nhất.

Các feature liên tục:

- Khi tăng enginesize (kích thước động cơ) giá sẽ tăng. wheelbase, carlength, carwidth (trục bánh xe, chiều dài, chiều rộng) có ảnh hưởng đến giá cả. Tương tự với curbweight, boreratio, horsepower, citympg
- carheight (chiều cao xe) không ảnh hưởng gì. Tương tự với stroke, compressionratio, peakrpm

⇒Kết luận:
-Chủ đầu tư có thể dựa vào kết quả của EDA để chọn features khách quan.
-Có thêm sự cân nhắc khi xem xét các features được chọn từ model ML.

Applying machine learning models

Sử dụng 2 cách chọn features:

- RFE (Recursive feature elimination, đệ quy) với Random Forest
- Correlation với price

Sử dụng 4 model Machine Learning:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- K Neighbors Regressor

2 case:

- Trước khi dùng Standard Scaler
- Sau khi dùng Standard Scaler

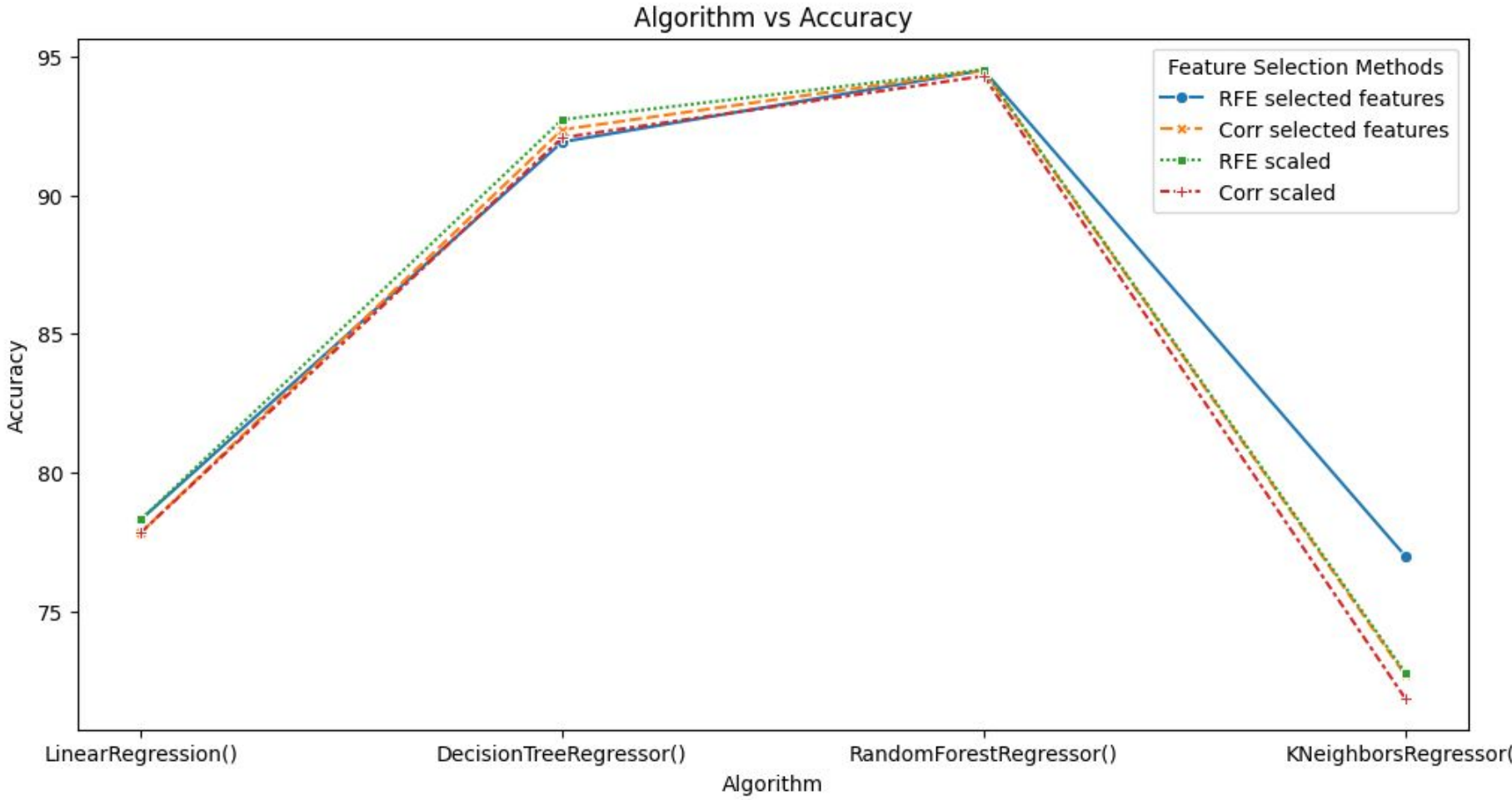
Kết quả tốt nhất là model Random Forest với các features theo RFE và tệp data sau Standard Scaler (~94.52%)

10 features: wheelbase, carlength, carwidth, curbweight, enginesize, stroke, compressionratio, horsepower, peakrpm, citympg

	Trước StandardScaler		Sau StandardScaler	
	RFE selected features	Corr selected features	RFE selected features	Corr selected features
LinearRegression()	78.312537	77.842816	78.312537	77.842816
DecisionTreeRegressor()	91.918353	92.362226	92.725508	92.072954
RandomForestRegressor()	94.506156	94.497210	94.520778	94.293103

ML MODELS

ML MODELS


Algorithm vs Accuracy
Feature Selection Methods
RFE selected features
Corr selected features
RFE scaled
Corr scaled
Accuracy
95
90
85
80
75
LinearRegression()
DecisionTreeRegressor()
RandomForestRegressor()
KNeighborsRegressor(
Algorithm

Các bước xử lý:

Dữ liệu gốc:

- Chọn features theo RFE và Corr
- Chia tệp training và test (70 train 30 test) cho cả 2 option selected features
- Chạy vòng lặp tính toán độ chính xác lần lượt cho 4 model: LR, DTR, RFR, KNN
- Trực quan hoá số liệu (visualize)
- Xem xét các features được chọn theo RFE và Corr

Chạy thử model với số features ít hơn.

Dữ liệu chuẩn hoá:

- Tiến hành chuẩn hoá Standard Scaler
- Chia lại tệp training và test.
- Chạy vòng lặp.
- Visualize sau Standard Scaler.

Visualize before và after để tiện so sánh

Kết luận
-Về mặt model ML: Model có độ chính xác cao nhất là Random Forest.
-Về mặt chọn lọc features: RFE cho ra độ chính xác cao hơn với 3 model Linear Regression, K Neibors, Random Forest (~0.09%). Trong khi Correlation cho kết quả tốt hơn với Decision Tree.
=>Cân nhắc về model Random Forest và chọn lọc feature theo RFE.
carlength, carwidth, citympg, compressionratio, curbweight, enginesize, horsepower, peakrpm, stroke, wheelbase

***Khi giảm số lượng feature, kết quả bị giảm so với 1 trong 2 cách trên.**
-Số features ảnh hưởng nhiều đến giá xe là 10.
-7 features chung vẫn cho ra kết quả tương đối cao, có thể cân nhắc nếu mở rộng quy mô dự đoán (performance tốt hơn)

Sau khi chuẩn hóa (standardization bằng Standard Scaler)
-Random Forest và RFE features được cải thiện (+0.014%), vẫn là kết quả cao nhất.
-Kết quả ở model Random Forest, Decision Tree với features RFE được cải thiện, trong khi với features Corr thì giảm => Dữ liệu càng được chuẩn hóa thì sẽ càng đi thuận với RF do đệ quy theo RF.
-Linear Regression kết quả không đổi. Nguyên nhân có thể do Standard Scaler không ảnh hưởng đến đồ thị tuyến tính của model này.
-K Neighbor giảm mạnh do việc chuẩn hóa ảnh hưởng thuật toán đo khoảng cách. Và K Neighbor model nổi tiếng trong các chủ đề classification hơn.

Đề xuất

- Số lượng features lý tưởng để điều chỉnh giá xe là 10.
- Sử dụng linh hoạt RFE và kết quả từ EDA để chọn lọc features để cải thiện các features quan trọng ảnh hưởng giá xe.
- Chuẩn hoá data bằng Standard Scaler và model RFE để hiểu được cách định giá của một thị trường mới (TQ → Mỹ) và tối ưu hiệu suất khi dự đoán giá xe
  => Tăng tốc độ phản hồi để điều chỉnh các features trên xe cho phù hợp.
- Đưa ra các phương án thiết kế của ô tô, chiến lược kinh doanh, v.v… để đáp ứng với từng phân khúc giá nhất định.

THANK YOU!

Cảm ơn thầy đã hướng dẫn lớp tận tâm và nhiệt tình trong suốt thời gian qua