

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



ĐỒ ÁN
TỐT NGHIỆP ĐẠI HỌC

Đề tài :

XÂY DỰNG MÔ HÌNH HỒI QUI DỰ BÁO GIÁ Ô TÔ

Người hướng dẫn : TS. NGUYỄN HỒNG SƠN

Sinh viên thực hiện : DƯƠNG TẤN PHÁT

Mã số sinh viên : N17DCCN121

Lớp : D17CQCP02-N

Hệ : ĐẠI HỌC CHÍNH QUY

TP. HỒ CHÍ MINH, THÁNG 12 NĂM 2021

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



ĐỒ ÁN
TỐT NGHIỆP ĐẠI HỌC

Đề tài :

XÂY DỰNG MÔ HÌNH HỒI QUI DỰ BÁO GIÁ Ô TÔ

Người hướng dẫn : TS. NGUYỄN HỒNG SƠN

Sinh viên thực hiện : DƯƠNG TẤN PHÁT

Mã số sinh viên : N17DCCN121

Lớp : D17CQCP02-N

Hệ : ĐẠI HỌC CHÍNH QUY

TP. HỒ CHÍ MINH, THÁNG 12 NĂM 2021

ĐỀ TÀI ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

Căn cứ Quyết định số: 402/QĐ-HVCS, ngày 21 tháng 09 năm 2021 của Phó Giám đốc Học viện – Phụ trách Cơ sở tại TP. Hồ Chí Minh về việc “phê duyệt danh sách giáo viên hướng dẫn và giao đề tài đồ án tốt nghiệp cho sinh viên Đại học chính quy Khóa 2017-2022 Ngành Công nghệ thông tin, An toàn thông tin và Công nghệ đa phương tiện”;

Khoa Công nghệ thông tin 2 giao nhiệm vụ thực hiện Đồ án tốt nghiệp cho sinh viên:

1. Họ và tên sv	: Dương Tấn Phát	Mã SV	: N17DCCN121
Lớp	: D17CQCP02-N	Khóa	: 2017-2022
Ngành đào tạo	: Công nghệ thông tin	Hệ đào tạo	: Đại học Chính quy

2. Tên đề tài tốt nghiệp: Xây dựng mô hình hồi qui dự báo giá ô tô

3. Nội dung chính của Đồ án:

Lý thuyết:

- Ý nghĩa của dự báo trong thị trường kinh doanh
- Khái niệm mô hình hồi qui
- Các mô hình hồi qui
- Hồi qui kernel
- Phương pháp xây dựng mô hình dự báo sử dụng hồi qui
- Vai trò của dataset và chất lượng của dataset trong xây dựng mô hình dự báo

Thực hành:

Thiết kế và cài đặt mô hình dự báo:

- Chọn mô hình hồi qui
- Chọn dataset phù hợp
- Huấn luyện mô hình
- Đánh giá mô hình dự báo
- Lập trình cài đặt ứng dụng dự báo giá ô tô

4. Cơ sở dữ liệu ban đầu:

5. Giáo viên hướng dẫn: TS. Nguyễn Hồng Sơn

6. Ngày giao đề tài: 27/09/2021

7. Ngày nộp quyển: 07/12/2021

TRƯỞNG KHOA CNTT2

Nơi nhận:

- Sinh viên có tên tại khoản 1;
- Lưu: VP Khoa.

TS. Nguyễn Hồng Sơn

LỜI CẢM ƠN

Báo cáo đồ án tốt nghiệp chuyên ngành công nghệ phần mềm với Đề tài “Xây dựng mô hình hồi quy dự báo giá ô tô” là kết quả của quá trình cố gắng không ngừng nghỉ của bản thân và được sự giúp đỡ tận tình, động viên khích lệ của thầy cô, bạn bè và người thân. Qua đây, Em xin gửi lời cảm ơn chân thành đến những người đã giúp đỡ em trong thời gian học tập - Nghiên cứu vừa qua.

Em xin trân trọng gửi đến thầy Nguyễn Hồng Sơn - Người đã trực tiếp tận tình hướng dẫn cũng như cung cấp tài liệu, thông tin khoa học cần thiết để em hoàn thiện đề tài này lời cảm ơn chân thành và sâu sắc nhất.

Xin cảm ơn lãnh đạo, ban giám hiệu cùng toàn thể các giảng viên Học viện Công nghệ Bưu chính Viễn thông Cơ sở TPHCM khoa Công nghệ thông tin 2 đã tạo điều kiện và thời gian cho em trong suốt quá trình nghiên cứu.

Cuối cùng, em xin cảm ơn gia đình, người thân, bạn bè đã luôn bên cạnh, ủng hộ, động viên.

Em xin chân thành cảm ơn!

Thành phố Hồ Chí Minh, tháng 11 năm 2021

Sinh viên thực hiện

Dương Tấn Phát

MỤC LỤC

LỜI CẢM ƠN	i
MỤC LỤC	ii
DANH MỤC CÁC BẢNG, SƠ ĐỒ, HÌNH VẼ	iv
DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT	vi
MỞ ĐẦU	1
CHƯƠNG 1: CƠ SỞ LÝ THUYẾT	2
1.1 Ý nghĩa của dự báo trong thị trường kinh doanh	2
1.2 Khái niệm mô hình hồi quy	2
1.2.1 Khái niệm	2
1.2.2 Tại sao chúng ta sử dụng phân tích hồi quy?	2
1.3 Một số mô hình hồi quy	3
1.3.1 Linear Regression – Hồi quy tuyến tính	3
1.3.2 Hồi quy logistic	4
1.3.3 Hồi quy đa thức	5
1.3.4 Hồi quy Stepwise	6
1.3.5 Hồi quy Ridge	7
1.3.6 Hồi quy Lasso	8
1.3.7 Hồi quy ElasticNet	8
1.3.8 Random Forest Regression – Hồi quy rừng ngẫu nhiên	9
1.3.9 Hồi quy Kernel	9
1.4 Một số chỉ số đánh giá	10
1.4.1 R-squared	10
1.4.2 R-squared hiệu chỉnh.....	11
1.4.3 MSE, RMSE.....	11
1.4.4 VIF	12
1.5 Phương pháp xây dựng mô hình dự báo sử dụng hồi quy	12
1.5.1 Chuẩn bị dữ liệu	13
1.5.2 Kiểm tra và loại biến	13
1.5.3 Xử lý các mẫu không đạt chuẩn	13
1.5.4 Xây dựng mô hình dự báo.....	13
1.5.5 Kiểm tra mô hình	13
1.6 Vai trò của dataset và chất lượng của dataset trong xây dựng mô hình dự báo ..	14
CHƯƠNG 2: THIẾT KẾ VÀ CÀI ĐẶT MÔ HÌNH DỰ BÁO.....	15
2.1 Chuẩn bị dữ liệu	15

2.2	Kiểm tra và loại biến.....	15
2.3	Xử lý biến không đạt chuẩn	18
2.4	Tiền xử lý	19
2.5	Huấn luyện mô hình.....	20
2.6	Đánh giá mô hình dự báo.....	23
CHƯƠNG 3: LẬP TRÌNH CÀI ĐẶT ỨNG DỤNG DỰ BÁO GIÁ Ô TÔ.....		24
3.1	Giao diện	24
3.2	Giải thích.....	25
3.3	Chạy thử ứng dụng.....	34
CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN		37
4.1	Kết quả đạt được	37
4.2	Hạn chế	37
4.3	Hướng khắc phục	37
4.4	Hướng mở rộng.....	37
DANH MỤC TÀI LIỆU THAM KHẢO		38

DANH MỤC CÁC BẢNG, SƠ ĐỒ, HÌNH VẼ

Hình 1.1 Ví dụ mô hình tuyến tính.....	3
Hình 1.2 Phương pháp bình phương nhỏ nhất	4
Hình 1.3 Hồi quy logistic	5
Hình 1.4 Hồi quy đa thức	6
Hình 1.5 Overfitting và Underfitting.....	6
Hình 1.6 Phương trình hồi quy Ridge	7
Hình 1.7 Phương trình hồi quy Lasso.....	8
Hình 1.8 Phương trình hồi quy ElasticNet	8
Hình 1.9 Phương trình hồi quy kernel.....	9
Hình 1.10 Mô tả SSE.....	11
Hình 1.11 Mô tả SST.....	11
Hình 1.12 Quy trình xây dựng mô hình dự báo	14
Hình 2.1 Thông tin dữ liệu	15
Hình 2.2 Thống kê dữ liệu.....	15
Hình 2.3 Thống kê dữ liệu theo biểu đồ.....	16
Hình 2.4 Thống kê age_of_car	17
Hình 2.5 Thống kê fuelType và transmission	17
Hình 2.6 Thống kê model.....	18
Hình 2.7 Thống kê số lượng mẫu xe	19
Hình 2.8 Mẫu dữ liệu ban đầu	19
Hình 2.9 Mẫu dữ liệu one_hot_encode	20
Hình 2.10 Thống kê chọn tính năng	21
Hình 2.11 Chỉ số VIF	22
Hình 2.12 Thống kê chọn tính năng cho thuật toán Random Forest.....	23
Hình 2.13 Kiểm định mô hình.....	23
Hình 3.1 Khung giao diện dự đoán	24
Hình 3.2 Giao diện huấn luyện.....	24
Hình 3.3 Popup kết quả dự đoán	25
Hình 3.4 Giao diện sự kiện chọn file dữ liệu	25
Hình 3.5 Popup sự kiện đọc file không thành công	26
Hình 3.6 Giao diện sự kiện thống kê dữ liệu.....	26
Hình 3.7 Giao diện sự kiện chọn mô hình huấn luyện	27
Hình 3.8 Giao diện khung Train khi đang trong quá trình huấn luyện	27
Hình 3.9 Giao diện và popup khi đã hoàn thành huấn luyện	28
Hình 3.10 Popup sự kiện xử lý dữ liệu thất bại.....	28
Hình 3.11 Giao diện combobox chọn hãng xe	28

Hình 3.12 Giao diện combobox chọn mẫu xe	29
Hình 3.13 Giao diện popup sự kiện năm sản xuất cách quá xa.....	29
Hình 3.14 Giao diện popup sự kiện năm sản xuất không nhỏ hơn 2022	30
Hình 3.15 Giao diện popup sự kiện năm dự đoán ở quá khứ.....	30
Hình 3.16 Giao diện combobox chọn loại hộp số	31
Hình 3.17 Giao diện popup sự kiện số dặm không hợp lệ	31
Hình 3.18 Giao diện popup sự kiện kích cỡ động cơ không hợp lệ	32
Hình 3.19 Giao diện popup sự kiện kích cỡ động cơ quá lớn	32
Hình 3.20 Giao diện popup sự kiện dữ liệu mpg không hợp lệ	33
Hình 3.21 Giao diện combobox loại nhiên liệu.....	33
Hình 3.22 Giao diện popup sự kiện dự đoán thất bại do thiếu dữ kiện.....	34
Hình 3.23 Giao diện chạy thử phần huấn luyện	34
Hình 3.24 Giao diện đã hoàn thành huấn luyện	35
Hình 3.25 Ví dụ chạy thử ứng dụng	35
Bảng 3.1 Bảng kết quả chạy thử ứng dụng với các mẫu thử ngẫu nhiên	36

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

AIC: Akaike information criterion	tiêu chí thông tin Akaike
Coef: coefficient	hệ số
Dataset	tập dữ liệu
Mpg: miles per gallon	mức tiết kiệm nhiên liệu của ô tô
MSE: Mean squared error	sai số trung bình bình phương
Regression	hồi quy
RMSE: Root mean squared error	sai số trung bình bình phương gốc
Std: Standard Deviation	độ lệch chuẩn
VIF: Variance inflation factor	hệ số lạm phát phương sai

MỞ ĐẦU

Với việc đất nước Việt Nam đang dần phát triển, các dự án mở rộng cầu, đường đã và đang được thi công, thu nhập của người Việt tăng đi đôi với nhu cầu mua sắm của cải vật chất, thì ô tô là một trong những mặt hàng mà trong tương lai không xa mọi người đều có khả năng mua sắm. Tuy nhiên, ô tô vốn là một mặt hàng xa xỉ, việc mua mới hoàn toàn một chiếc ô tô có thể khó khăn với nhiều người, thì phương án mà họ có thể chọn là mua cũ những chiếc ô tô mà chủ sở hữu ban đầu của nó muốn bán.

Việc xây dựng một ứng dụng có khả năng dự báo giá cả cho một số dòng xe giúp người dùng có nhu cầu mua xe có thể xây dựng kế hoạch tiết kiệm tiền một cách thích hợp nhất, ứng dụng này cũng có thể giúp người dùng có nhu cầu bán xe ô tô đưa ra một mức giá thích hợp cho cả đôi bên.

Bên cạnh đó, đây cũng là một cơ hội để em có thể thử sức nghiên cứu và xây dựng một ứng dụng sử dụng mô hình hồi quy để dự báo giá - một mảng công việc có rất nhiều ứng dụng trong đời sống.

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1 Ý nghĩa của dự báo trong thị trường kinh doanh

- Bằng cách dự báo các mức độ tương lai của hiện tượng, ta có thể giúp các chủ doanh nghiệp chủ động trong việc đề ra các kế hoạch và các quyết định cần thiết phục vụ cho quá trình kinh doanh như sản xuất sản phẩm, đầu tư, quảng bá, phân phối sản phẩm, phân phối nguồn tiền ... và chuẩn bị đầy đủ điều kiện cơ sở vật chất, kỹ thuật cho các kế hoạch trong tương lai được dự báo.
- Trong các doanh nghiệp, nếu công tác dự báo được thực hiện một cách chính xác còn tạo điều kiện nâng cao khả năng cạnh tranh trên thị trường.
- Dự báo chính xác sẽ giảm bớt mức độ rủi ro cho doanh nghiệp nói riêng và toàn bộ nền kinh tế nói chung.
- Dự báo chính xác là căn cứ để các nhà hoạch định các chính sách phát triển kinh tế, văn hoá, xã hội.
- Nhờ có dự báo các chính sách kinh tế, các kế hoạch và chương trình phát triển kinh tế được xây dựng có cơ sở khoa học và mang lại hiệu quả kinh tế cao.
- Nhờ có dự báo thường xuyên và kịp thời, các nhà quản trị doanh nghiệp có khả năng kịp thời đưa ra những biện pháp điều chỉnh các hoạt động kinh tế nhằm thu được hiệu quả kinh doanh cao nhất.

1.2 Khái niệm mô hình hồi quy

1.2.1 Khái niệm

Hồi quy chính là một phương pháp thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập.

Ví dụ: Tuổi = 5 + Chiều cao * 10 + Trọng lượng * 13

Ở đây chúng ta đang thiết lập mối quan hệ giữa Chiều cao và Trọng lượng của một người với Tuổi của anh/cô ta. Đây là một ví dụ rất cơ bản của hồi quy.

1.2.2 Tại sao chúng ta sử dụng phân tích hồi quy?

Như đã đề cập ở trên, phân tích hồi quy ước tính mối quan hệ giữa hai hoặc nhiều biến. Hãy hiểu điều này bằng một ví dụ đơn giản:

Giả sử ta muốn ước tính mức tăng trưởng doanh số bán hàng của một công ty dựa trên điều kiện kinh tế hiện tại. Ta có dữ liệu công ty gần đây cho thấy mức tăng trưởng doanh số bán hàng gấp khoảng 2,5 lần mức tăng trưởng của nền kinh tế. Sử dụng thông tin chi tiết này, chúng ta có thể dự đoán doanh số bán hàng trong tương lai của công ty dựa trên thông tin hiện tại và quá khứ.

Có nhiều lợi ích khi sử dụng phân tích hồi quy. Chúng như sau:

- Nó chỉ ra các mối quan hệ đáng kể giữa biến phụ thuộc và biến độc lập.
- Nó chỉ ra mức độ tác động của nhiều biến độc lập lên một biến phụ thuộc.
- Phân tích hồi quy cũng cho phép chúng ta so sánh tác động của các biến được đo lường trên các quy mô khác nhau, chẳng hạn như ảnh hưởng của sự thay đổi giá và số lượng các hoạt động khuyến mại. Những lợi ích này giúp các nhà nghiên cứu thị trường / nhà phân tích dữ liệu / nhà khoa học dữ liệu loại bỏ và đánh giá tập hợp các biến tốt nhất được sử dụng để xây dựng các mô hình dự báo.

1.3 Một số mô hình hồi quy

Có nhiều loại kỹ thuật hồi quy khác nhau có sẵn để đưa ra dự đoán. Các kỹ thuật này hầu hết được phân biệt bởi ba số liệu (số lượng biến độc lập, loại biến phụ thuộc và hình dạng của đường hồi quy). Trong đó, các phép hồi quy được sử dụng phổ biến nhất là:

1.3.1 Linear Regression – Hồi quy tuyến tính

Nó là một trong những kỹ thuật mô hình hóa được biết đến rộng rãi nhất. Trong kỹ thuật này, biến phụ thuộc là liên tục, (các) biến độc lập có thể liên tục hoặc rời rạc, và bản chất của đường hồi quy là tuyến tính.

Hồi quy tuyến tính thiết lập mối quan hệ giữa biến phụ thuộc (Y) và một hoặc nhiều biến độc lập (X) bằng cách sử dụng một đường thẳng phù hợp nhất (còn được gọi là đường hồi quy).

Nó được biểu diễn bằng phương trình:

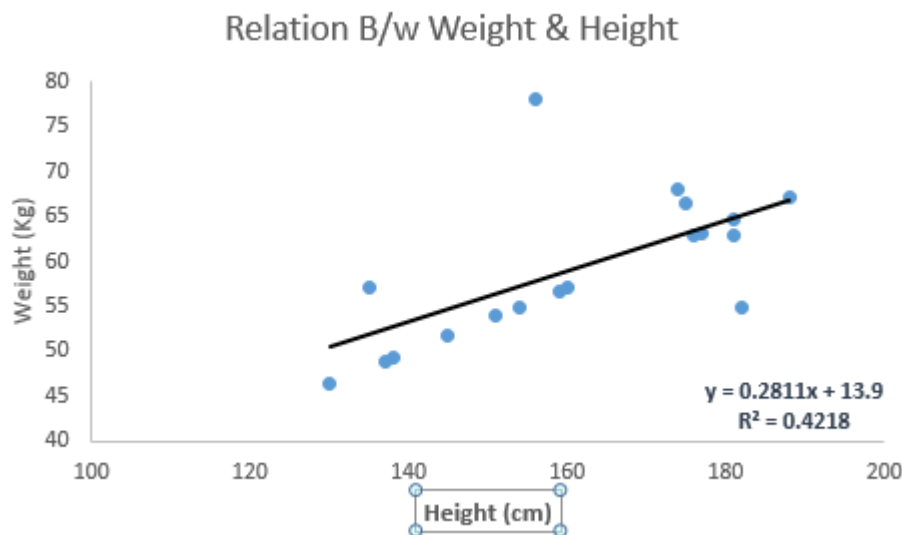
$$Y = a + b * X + e$$

Trong đó a là chặn (intercept),

b là hệ số góc của đường (còn gọi là độ dốc (slope)),

e là sai số (error) (còn gọi là phần dư (residual)).

Phương trình này có thể được sử dụng để dự đoán giá trị của biến mục tiêu dựa trên (các) biến dự báo đã cho.

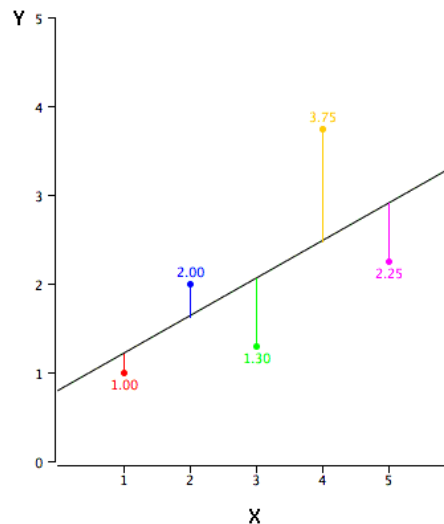


Hình 1.1 Ví dụ mô hình tuyến tính

Sự khác biệt giữa hồi quy tuyến tính đơn và hồi quy tuyến tính bội là hồi quy tuyến tính bội có (> 1) biến độc lập, trong khi hồi quy tuyến tính đơn chỉ có 1 biến độc lập.

Có thể có được đường phù hợp nhất (Giá trị của a và b) dễ dàng bằng Phương pháp bình phương nhỏ nhất. Đây là phương pháp phổ biến nhất được sử dụng để điều chỉnh một đường hồi quy. Nó tính toán đường phù hợp nhất cho dữ liệu quan sát bằng cách giảm thiểu tổng bình phương của độ lệch từ mỗi điểm dữ liệu đến đường. Bởi vì độ lệch được bình phương trước, không có sự triệt tiêu giữa các giá trị âm và dương khi được cộng vào.

$$\min_w ||Xw - y||_2^2$$



Hình 1.2 Phương pháp bình phương nhỏ nhất

Chúng ta có thể đánh giá hiệu suất của mô hình bằng cách sử dụng thước đo R-square. Điểm quan trọng:

- Phải có mối quan hệ tuyến tính giữa các biến độc lập và phụ thuộc
- Hồi quy bội có thể bị
 - đa cộng tuyến (multicollinearity): Đa cộng tuyến là hiện tượng tạo nên từ mối quan hệ tương quan mạnh giữa các biến độc lập với nhau trong mô hình hồi quy tuyến tính.
 - tự tương quan (autocorrelation): Hiện tượng này thường xảy ra trong dữ liệu thời gian (time series) hoặc dữ liệu bảng (panel data). Đây là hiện tượng mà sai số tại thời điểm t có mối quan hệ với sai số tại thời điểm $t-1$ hoặc tại bất kỳ thời điểm nào khác trong quá khứ.
 - phương sai thay đổi (heteroskedasticity): Là tình huống thống kê trong đó có sự thay đổi theo một quy luật nào đó trong phần dư hoặc sai số sau khi phương trình hồi quy được ước lượng từ các kết quả quan sát mẫu của biến độc lập và phụ thuộc.
- Hồi quy tuyến tính rất nhạy cảm với các giá trị ngoại lai. Nó có thể ảnh hưởng mạnh mẽ đến đường hồi quy và các giá trị dự báo.
- Đa cộng tuyến có thể làm tăng phương sai của các hệ số và làm cho các ước lượng rất nhạy cảm với những thay đổi nhỏ trong mô hình. Kết quả là các ước lượng hệ số không ổn định.
- Trong trường hợp có nhiều biến độc lập, chúng ta có thể chọn tiến, loại bỏ lùi và tiếp cận từng bước một cách khôn ngoan để lựa chọn hầu hết các biến độc lập có ý nghĩa.

1.3.2 Hồi quy logistic

Hồi quy logistic được sử dụng để tìm xác suất của sự kiện = Thành công và sự kiện = Thất bại. Chúng ta có thể sử dụng hồi quy logistic khi biến phụ thuộc có bản chất là nhị phân (0/1, True / False, Yes / No). Ở đây giá trị của Y nằm trong khoảng từ 0 đến 1 và nó có thể được biểu diễn bằng phương trình sau.

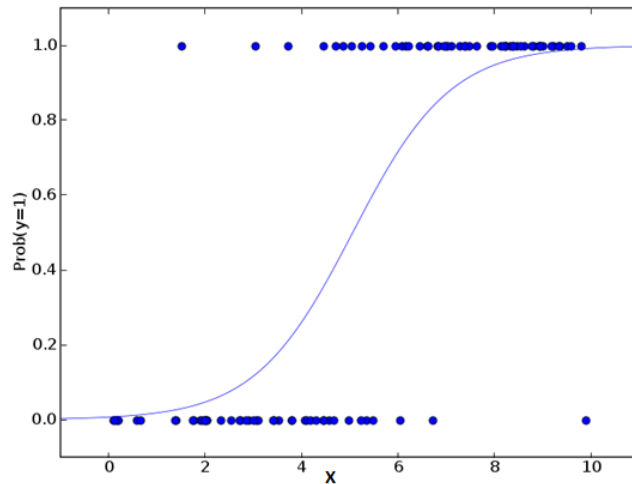
$$\text{tỷ lệ} = p / (1-p) = \text{xác suất xảy ra sự kiện} / \text{xác suất không xảy ra sự kiện}$$

$$\ln(\text{tỷ lệ}) = \ln(p / (1-p))$$

$$\logit(p) = \ln(p / (1-p)) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \dots + b_kX_k$$

Trong đó, p là xác suất hiện diện của đặc tính quan tâm.

Chúng ta sử dụng log trong phương trình vì chúng ta đang làm việc với phân phối nhị thức (biến phụ thuộc), chúng ta cần chọn một hàm liên kết phù hợp nhất với phân phối này. Và, đó là chức năng logit. Trong phương trình trên, các tham số được chọn để tối đa hóa khả năng quan sát các giá trị mẫu hơn là giảm thiểu tổng sai số bình phương (như trong hồi quy thông thường).



Hình 1.3 Hồi quy logistic

Điểm quan trọng:

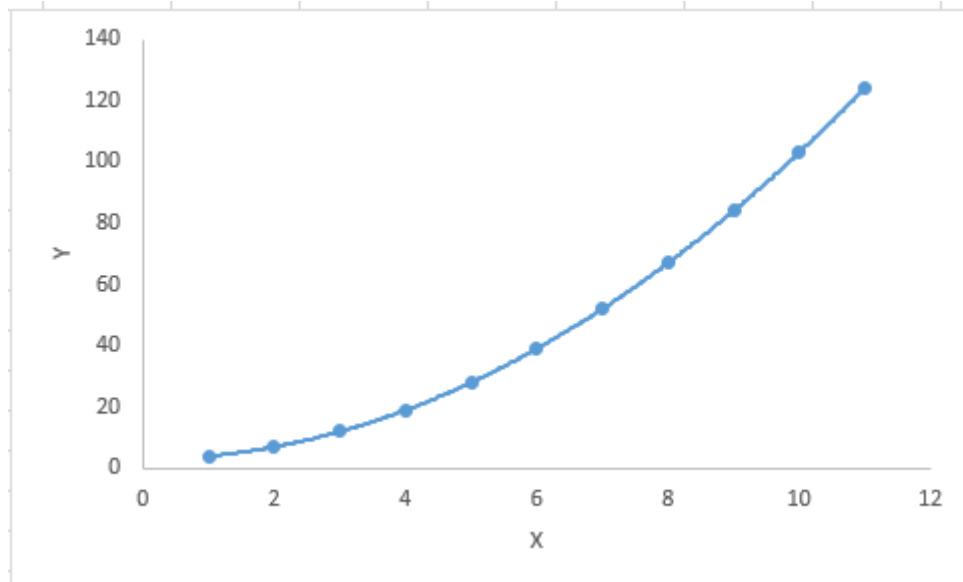
- Hồi quy logistic được sử dụng rộng rãi cho các bài toán phân loại.
- Hồi quy logistic không yêu cầu mối quan hệ tuyến tính giữa các biến phụ thuộc và độc lập. Nó có thể xử lý nhiều loại mối quan hệ khác nhau vì nó áp dụng một phép biến đổi log phi tuyến tính cho tỷ lệ chênh lệch dự đoán.
- Để tránh phù hợp quá mức (overfitting) và phù hợp dưới mức (underfitting), chúng ta nên bao gồm tất cả các biến số quan trọng. Một cách tiếp cận tốt để đảm bảo điều này là sử dụng một phương pháp hồi quy từng bước để ước tính hồi quy logistic.
- Nó yêu cầu kích thước mẫu lớn.
- Các biến độc lập không được tương quan với nhau, tức là không có đa cộng tuyến.
- Nếu các giá trị của biến phụ thuộc dạng thứ tự, thì nó được gọi là hồi quy logistic thứ tự.
- Nếu biến phụ thuộc là biến định tính có nhiều hơn 2 trạng thái thì nó được gọi là hồi quy Logistic đa thức.

1.3.3 Hồi quy đa thức

Phương trình hồi quy là phương trình hồi quy đa thức nếu lũy thừa của biến độc lập lớn hơn 1. Phương trình dưới đây biểu diễn một phương trình đa thức:

$$y = a + b * x^2$$

Trong kỹ thuật hồi quy này, đường phù hợp nhất không phải là đường thẳng. Nó là một đường cong phù hợp với các điểm dữ liệu.

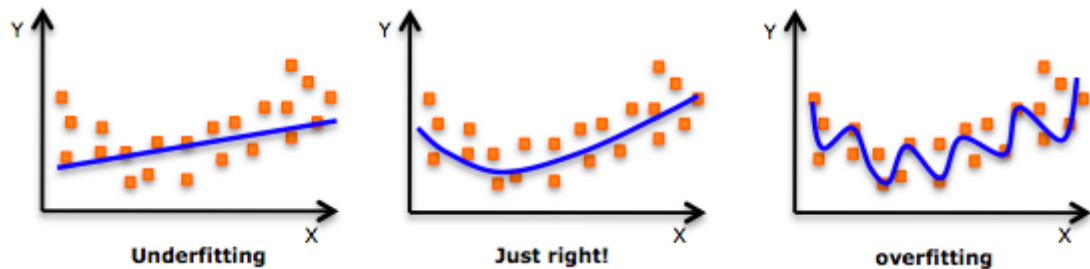


Hình 1.4 Hồi quy đa thức

Điểm quan trọng:

- Mặc dù có thể có sự cám dỗ để điều chỉnh một đa thức bậc cao hơn để nhận được sai số thấp hơn, nhưng điều này có thể dẫn đến việc điều chỉnh quá mức (over-fitting). Luôn vẽ các mối quan hệ để thấy được sự phù hợp và tập trung vào việc đảm bảo rằng đường cong đó phù hợp với bản chất của vấn đề.

Ví dụ:



Hình 1.5 Overfitting và Underfitting

- Đặc biệt là nhìn vào đường cong về sau và xem liệu những hình dạng và hướng đó có hợp lý hay không. Các đa thức cao có thể tạo ra kết quả lạ trên phép ngoại suy.

1.3.4 Hồi quy Stepwise

Dạng hồi quy này được sử dụng khi chúng ta xử lý nhiều biến độc lập. Trong kỹ thuật này, việc lựa chọn các biến độc lập được thực hiện với sự trợ giúp của một quy trình tự động, không có sự can thiệp của con người.

Thành tích này đạt được bằng cách quan sát các giá trị thống kê như R-square, t-stats (t-stats là tỷ số giữa giá trị ước tính của một tham số từ giá trị giả thuyết của nó với sai số chuẩn của nó) và AIC (AIC là viết tắt của Tiêu chí Thông tin Akaike, đây là một kỹ thuật được sử dụng để chọn mô hình tốt nhất từ một nhóm các mô hình) để phân biệt các biến có ý nghĩa. Hồi quy Stepwise về cơ bản phù hợp với mô hình hồi quy bằng cách thêm/ bớt các đồng biến tại một thời điểm dựa trên một tiêu chí được chỉ định.

Một số phương pháp hồi quy Stepwise thường được sử dụng nhiều nhất gồm:

- Hồi quy Stepwise tiêu chuẩn thực hiện hai điều: thêm và loại bỏ các yếu tố dự đoán khi cần thiết cho mỗi bước.

- Lựa chọn chuyển tiếp bắt đầu với dự đoán quan trọng nhất trong mô hình và thêm biến cho mỗi bước.
- Loại bỏ ngược bắt đầu với tất cả các yếu tố dự đoán trong mô hình và loại bỏ biến ít quan trọng nhất cho mỗi bước.

Mục đích của kỹ thuật mô hình hóa này là tối đa hóa khả năng dự đoán với số lượng biến dự báo tối thiểu. Nó là một trong những phương pháp để xử lý số chiều cao của tập dữ liệu.

1.3.5 Hồi quy Ridge

Hồi quy Ridge là một kỹ thuật được sử dụng khi dữ liệu bị đa cộng tuyến (các biến độc lập có tương quan cao). Trong đa cộng tuyến, mặc dù các ước lượng bình phương nhỏ nhất (OLS) là không chệch, nhưng phương sai của chúng rất lớn làm lệch giá trị quan sát ra xa giá trị thực. Bằng cách thêm một mức độ chệch vào các ước tính hồi quy, hồi quy sườn núi làm giảm các sai số tiêu chuẩn.

Nếu phương trình hồi quy tuyến tính được biểu diễn dưới dạng:

$$y = a + b * x$$

Phương trình này cũng có một thuật ngữ lỗi. Phương trình hoàn chỉnh trở thành:

$y = a + b * x + e$ (thuật ngữ lỗi), [thuật ngữ lỗi là giá trị cần thiết để sửa lỗi dự đoán giữa giá trị được quan sát và dự đoán]

$\Rightarrow y = a + b_1x_1 + b_2x_2 + \dots + e$, đối với nhiều biến độc lập.

Trong một phương trình tuyến tính, các lỗi dự đoán có thể được phân tách thành hai thành phần phụ. Đầu tiên là do sự thiên vị và thứ hai là do phương sai. Lỗi dự đoán có thể xảy ra do bất kỳ một trong hai hoặc cả hai thành phần này. Ở đây, chúng ta sẽ thảo luận về lỗi gây ra do phương sai.

Hồi quy Ridge giải quyết vấn đề đa cộng tuyến thông qua tham số co rút λ (lambda). Ta có phương trình:

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

Hình 1.6 Phương trình hồi quy Ridge

Trong đó lambda là một hằng số dương (tham số co rút).

Trong phương trình này, chúng ta có hai thành phần. Số hạng đầu tiên là số hạng bình phương nhỏ nhất và số hạng còn lại là lambda của tổng bình phương của vector hệ số. Nói cách khác, bài toán sườn núi phạt các hệ số hồi quy lớn, và tham số lambda càng lớn thì mức phạt càng lớn.

Điểm quan trọng:

- Các giả định của hồi quy này giống như hồi quy bình phương nhỏ nhất ngoại trừ giả định chuẩn (các phần dư được phân phối chuẩn)
- Hồi quy Ridge thu hẹp giá trị của các hệ số nhưng không đạt đến 0, điều này cho thấy không có tính năng lựa chọn các biến độc lập
- Đây là một phương pháp chính quy hóa và sử dụng chính quy hóa L2.

1.3.6 Hồi quy Lasso

Tương tự như hồi quy Ridge, Lasso (Least Absolute Shrinkage and Selection Operator) cũng phạt giá trị tuyệt đối của các hệ số hồi quy. Ngoài ra, nó có khả năng giảm độ biến thiên và cải thiện độ chính xác của mô hình hồi quy tuyến tính.

Phương trình:

$$= \operatorname{argmin}_{\beta \in \mathbb{R}^p} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Hình 1.7 Phương trình hồi quy Lasso

Hồi quy Lasso khác với hồi quy sườn ở chỗ nó sử dụng các giá trị tuyệt đối trong hàm hình phạt, thay vì bình phương. Điều này dẫn đến việc trừng phạt (hoặc hạn chế một cách tương đương tổng các giá trị tuyệt đối của các ước lượng) các giá trị khiến một số ước lượng tham số trở thành không chính xác. Hình phạt được áp dụng càng lớn thì các ước tính càng bị thu hẹp hoàn toàn về 0. Điều này dẫn đến lựa chọn biến trong số n biến đã cho.

Điểm quan trọng:

- Các giả định của hồi quy lasso cũng giống như hồi quy bình phương nhỏ nhất ngoại trừ giả định chuẩn (các phần dư được phân phối chuẩn).
- Hồi quy Lasso thu nhỏ các hệ số về 0, điều này giúp ích trong việc lựa chọn đối tượng.
- Lasso là một phương pháp chính quy hóa và sử dụng chính quy hóa L1.
- Nếu có một nhóm các yếu tố dự đoán có tương quan cao, lasso chỉ chọn một trong số chúng và thu nhỏ các yếu tố khác xuống 0.

1.3.7 Hồi quy ElasticNet

Elastic-Net là một phương pháp hồi quy kết hợp tuyến tính các hình phạt L1 và L2 của phương pháp LASSO và Ridge tương ứng. Lưới đàn hồi hữu ích khi có nhiều tính năng tương quan với nhau. Lasso chỉ chọn ngẫu nhiên một trong số các tính năng này, trong khi lưới đàn hồi có khả năng chọn một nhóm các tính năng.

$$\hat{\beta} = \operatorname{argmin}_{\beta} (\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1).$$

Hình 1.8 Phương trình hồi quy ElasticNet

Một lợi thế thực tế của việc kết hợp Lasso và Ridge là nó cho phép Elastic-Net kế thừa một số tính ổn định của Ridge.

Điểm quan trọng:

- Nó khuyến khích hiệu ứng nhóm trong trường hợp các biến có tương quan cao
- Không có giới hạn về số lượng biến được chọn
- Nó có thể bị co rút gấp đôi

1.3.8 Random Forest Regression – Hồi quy rừng ngẫu nhiên

Rừng ngẫu nhiên là một thuật toán học có giám sát. Như tên gọi của nó, Rừng ngẫu nhiên sử dụng các cây (tree) để làm nền tảng.

Rừng ngẫu nhiên là một tập hợp của các cây quyết định, mà mỗi cây được chọn theo một thuật toán dựa vào ngẫu nhiên.

- Lấy ngẫu nhiên dữ liệu để xây dựng cây quyết định.
- Lấy ngẫu nhiên các thuộc tính để xây dựng cây quyết định.

Mỗi Node của cây sẽ là các thuộc tính, và các nhánh là giá trị lựa chọn của thuộc tính đó. Bằng cách đi theo các giá trị thuộc tính trên cây, cây quyết định sẽ cho ta biết giá trị dự đoán.

Nhóm thuật toán cây quyết định có một điểm mạnh đó là có thể sử dụng cho cả bài toán Phân loại (Classification) và Hồi quy (Regression).

Khi rừng có nhiều cây hơn, chúng ta có thể tránh được việc Overfitting với tập dữ liệu.

Random Forest hoạt động bằng cách đánh giá nhiều Cây quyết định ngẫu nhiên, và lấy ra kết quả được đánh giá nhiều nhất trong số kết quả trả về đối với bài toán phân loại và lấy kết quả trung bình đối với bài toán hồi quy.

Ví dụ minh họa:

Ta phân vân không biết nên mua sản phẩm máy tính tầm giá bao nhiêu là phù hợp với việc học tập ta sẽ đi hỏi một người bạn để tham khảo ý kiến. Nhưng ý kiến này có thể không khách quan. Ta bắt đầu trưng cầu ý kiến và giá sản phẩm máy tính của nhiều người hơn, sau đó tổng hợp lại và chọn mức giá thích hợp.

Quy trình xây dựng rừng ngẫu nhiên:

- Chọn ngẫu nhiên “k” features từ tập “m” features. ($k \ll m$)
- từ tập “k” features, tính toán ra node “d” là tốt nhất cho Node phân loại.
- Chia các node con theo node tốt nhất vừa tìm được
- Lặp lại bước 1-3 cho đến khi đạt đến k node
- Lặp lại bước 1-4 để tạo ra “n” cây

1.3.9 Hồi quy Kernel

Trong thống kê, hồi quy Kernel là một kỹ thuật phi tham số để ước tính kỳ vọng có điều kiện của một biến ngẫu nhiên. Mục tiêu là để tìm một mối quan hệ phi tuyến tính giữa một cặp biến ngẫu nhiên X và Y.

Trong bất kỳ hồi quy không tham số nào, kỳ vọng có điều kiện của một biến Y liên quan đến một biến X có thể được viết:

$$E(Y|X) = m(X)$$

Hình 1.9 Phương trình hồi quy kernel

ở đây m là một hàm chưa biết.

1.4 Một số chỉ số đánh giá

1.4.1 R-squared

1.4.1.1 Khái niệm

Một thước đo sự phù hợp của mô hình tuyến tính thường dùng là hệ số xác định R bình phương (Coefficient of Determination). Công thức tính R bình phương (R square) xuất phát từ ý tưởng xem toàn bộ biến thiên quan sát được của biến phụ thuộc được chia thành 2 phần: phần biến thiên do Hồi quy (Regression) và phần biến thiên do Phần dư (Residual). Nếu phần biến thiên do Phần dư càng nhỏ, nghĩa là khoảng cách từ các điểm quan sát đến đường ước lượng hồi quy càng nhỏ thì phần biến thiên do Hồi quy sẽ càng cao, khi đó giá trị R bình phương sẽ càng cao.

Hệ số R bình phương là hàm không giảm theo số biến độc lập được đưa vào mô hình, nếu chúng ta càng đưa thêm biến độc lập vào mô hình thì R bình phương càng tăng. Tuy nhiên, điều này cũng được chứng minh rằng không phải phương trình càng có nhiều biến thì càng tốt hơn.

1.4.1.2 Ý nghĩa

R square hay r bình phương được sử dụng nhiều trong kinh tế lượng. R bình phương được sử dụng trong thống kê và được thực hiện bởi phương pháp gọi là hồi quy tuyến tính.

R bình phương cho biết mô hình đó hợp với dữ liệu ở mức bao nhiêu %.

Ví dụ: r bình phương = 0,65. Vậy mô hình hồi quy tuyến tính đang được thống kê sẽ phù hợp với dữ liệu (hoặc biến) ở mức 65%.

R bình phương cũng cho biết độ phù hợp của mô hình, người ta nghiên cứu được rằng, với r bình phương > 50% thì một mô hình được đánh giá là phù hợp.

Tất nhiên, không phải tất cả các mô hình đều phải có r bình phương > 50%, ta có thể loại trừ một số mô hình có sự biến động lớn như giá vàng hay giá cổ phiếu...

Đặc biệt, giá trị r^2 càng cao thì mối quan hệ giữa nhân tố độc lập (biến độc lập) và nhân tố phụ thuộc càng chặt chẽ. Vì thế mà r bình phương còn được biết tới với cái tên hệ số tương quan r bình phương.

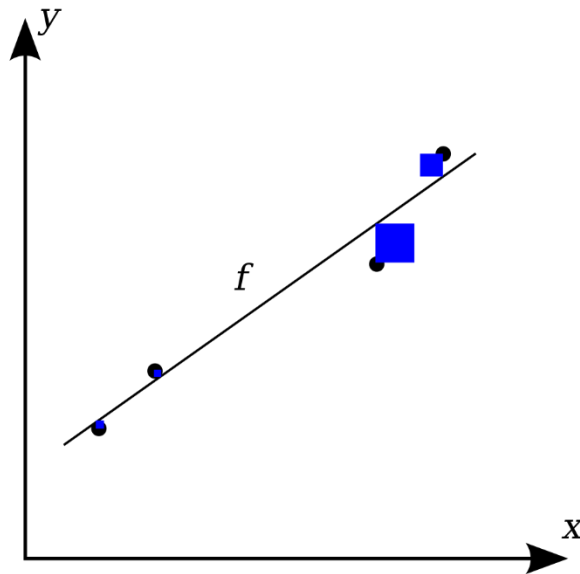
Qua đó có thể thấy ý nghĩa hệ số xác định r^2 là vô cùng quan trọng trong thống kê và nghiên cứu, đặc biệt là trong phương pháp hồi quy tuyến tính.

1.4.1.3 Công thức

$$R^2 = 1 - \frac{SSE}{SST}$$

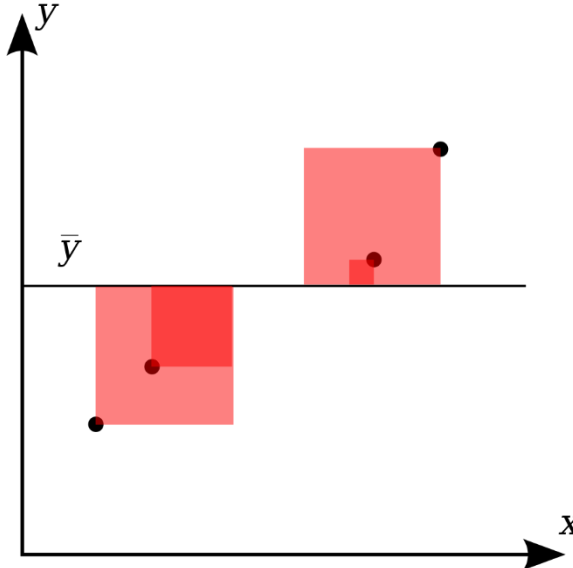
Trong đó,

SSE là tổng bình phương của phần dư (biến thiên của giá trị thực và giá trị dự đoán)



Hình 1.10 Mô tả SSE

SST là tổng bình phương biến thiên của giá trị thực và giá trị trung bình



Hình 1.11 Mô tả SST

1.4.2 R-squared hiệu chỉnh

Hạn chế nổi bật nhất của r square là việc giảm tính chính xác của mô hình khi ta thêm một tham số trong quá trình tính toán. Vì vậy, r bình phương hiệu chỉnh được nghiên cứu giúp khắc phục nhược điểm của r bình phương thông thường. Hệ số này cho phép ta đo độ thích hợp khi ta thêm một tham số nữa. Qua đó giúp giảm sự phức tạp của mô hình.

Công thức:

$$R_{hc}^2 = 1 - \frac{n-1}{n-k} (1 - R^2)$$

Trong đó, n là số mẫu quan sát

k là số biến của mô hình

1.4.3 MSE, RMSE

Trong thống kê học, sai số toàn phương trung bình, viết tắt MSE (Mean squared error) của công cụ ước tính (của thủ tục ước tính số lượng không quan sát được) đo trung bình bình phương của các lỗi – nghĩa là chênh lệch bình phương trung bình giữa các giá trị ước tính và giá trị quan sát được. MSE là một hàm rủi ro, tương ứng với giá trị

dự kiến của mất lỗi bình phương. Việc MSE hầu như luôn luôn tích cực (chứ không phải bằng không) là do tính ngẫu nhiên hoặc do công cụ ước tính không tính đến thông tin có thể tạo ra ước tính chính xác hơn.

MSE được gọi nôm na là giá trị sai số bình phương trung bình hoặc là lỗi bình phương trung bình. Vấn đề khi nói về sai số trung bình của một mô hình thống kê nhất định là rất khó xác định mức độ lỗi là do mô hình và mức độ là do ngẫu nhiên. Lỗi bình phương trung bình (MSE) cung cấp một thống kê cho phép các nhà nghiên cứu đưa ra tuyên bố như vậy. MSE chỉ đơn giản đề cập đến giá trị trung bình của chênh lệch bình phương giữa tham số dự đoán và tham số quan sát được.

Công thức:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Trong đó, Y_i là giá trị quan sát được

\hat{Y}_i là giá trị dự đoán

RMSE (Root Mean Square Error), như tên cho thấy, chỉ là căn bậc hai của MSE.

MSE có giá trị cao hơn vì chúng ta đang bình phương nó. Để mang lại giá trị này trên cùng thang điểm của sai số dự đoán, ta sử dụng RMSE. Điều này làm cho việc giải thích dễ dàng hơn. Khi cả MSE và RMSE bình phương các phần dư, chúng bị ảnh hưởng tương tự bởi các giá trị ngoại lai.

1.4.4 VIF

Hệ số lạm phát phương sai (Variance inflation factor – VIF) là thước đo mức độ đa cộng tuyến trong một tập hợp nhiều biến hồi quy. Về mặt toán học, VIF cho một biến mô hình hồi quy bằng tỷ số giữa phương sai của mô hình tổng thể với phương sai của một mô hình chỉ bao gồm một biến độc lập duy nhất đó. Tỷ lệ này được tính cho từng biến độc lập. VIF cao chỉ ra rằng biến độc lập liên quan có tính tương đồng cao với các biến khác trong mô hình.

Ta có công thức tính VIF như sau:

$$VIF_i = \frac{1}{1 - R_i^2}$$

trong đó R_i^2 là hệ số xác định của phương trình hồi quy bình phương nhỏ nhất thông thường có X_i là một hàm của tất cả các biến độc lập khác (không bao gồm biến phụ thuộc (Y))

- Giá trị VIF bắt đầu từ 1 và không có giới hạn trên.
- Giá trị VIF trong khoảng từ 1-2 chỉ ra rằng không có mối tương quan giữa biến độc lập này và bất kỳ biến nào khác.
- VIF giữa 2 và 5 cho thấy rằng có một mối tương quan vừa phải, nhưng nó không đủ nghiêm trọng để người nghiên cứu phải tìm biện pháp khắc phục.
- VIF lớn hơn 5 đại diện cho mối tương quan cao, hệ số được ước tính kém và các giá trị p - values là đáng nghi ngờ.
- VIF > 10 thì chắc chắn có đa cộng tuyến.

1.5 Phương pháp xây dựng mô hình dự báo sử dụng hồi quy

Theo Wilson và Keating, quy trình dự báo gồm 9 bước:

1. Xác định mục tiêu
2. Xác định đối tượng dự báo
3. Xác định thời đoạn dự báo
4. Thu thập, khảo sát dữ liệu
5. Chọn mô hình
6. Đánh giá mô hình

7. Chuẩn bị dự báo
8. Trình bày dự báo
9. Theo dõi kết quả

Ta có thể tóm gọn làm 5 bước triển khai một mô hình dự báo như sau:

1.5.1 Chuẩn bị dữ liệu

Chất lượng của mô hình phụ thuộc rất nhiều vào chất lượng của dữ liệu. Có thể thu thập dữ liệu từ:

- Nguồn thông tin sơ cấp: từ phỏng vấn trực tiếp, gửi thư, điện thoại ...
- Nguồn thông tin thứ cấp:
 - Bên trong (nội bộ): Các dữ liệu thông tin có sẵn của công ty, cơ quan để phục vụ việc dự báo cho công ty, cơ quan đó
 - Bên ngoài: Các dữ liệu được thu thập, thống kê từ Internet, sách báo, tạp chí, thống kê của cơ quan nhà nước, ...

Yêu cầu:

- Dữ liệu cần phải tin cậy và chính xác
- Dữ liệu cần phải có ý nghĩa
- Dữ liệu cần phải phù hợp
- Dữ liệu cần được thu thập trong một khoảng thời gian nhất định

1.5.2 Kiểm tra và loại biến

Mục tiêu các biến được lựa chọn và đưa vào mô hình phải đáp ứng các tiêu chuẩn như sau:

- Biến phụ thuộc (Y) trong mô hình thường là biến liên tục
- Có mối quan hệ tuyến tính giữa biến phụ thuộc Y với các biến độc lập X
- Dữ liệu không có chứa các điểm dị biệt/ outliers
- Không có sự đa cộng tuyến giữa các biến độc lập X. Điều đó có nghĩa các biến độc lập trong mô hình không có sự tương quan cao với nhau.

1.5.3 Xử lý các mẫu không đạt chuẩn

Có rất nhiều các vấn đề xảy ra với biến và không có chuẩn mực nào cho mọi trường hợp xử lý biến vì vậy trong các trường hợp chúng ta sẽ cần cân nhắc xem vấn đề biến không đạt chuẩn là vì sao:

- Nếu đó là giá trị bất thường, nhưng sau khi xem xét thấy rằng ở thời điểm đó nó bất thường là chính xác thì có thể cân nhắc giữ lại
- Nếu giá trị bất thường đó thực sự có vấn đề và bộ mẫu dữ liệu lớn ta có thể xóa bỏ, thay thế bằng giá trị khác
- Nếu các mối quan hệ không tốt, phân phối dữ liệu không tốt, có thể cân nhắc chia nhỏ các mối quan hệ/ biến hoặc tìm kiếm các mối quan hệ/ biến mới

1.5.4 Xây dựng mô hình dự báo

Việc chọn mô hình tùy thuộc vào các tiêu chí sau:

- Dạng phân bố của dữ liệu
- Số lượng quan sát sẵn có
- Độ dài của thời đoạn dự báo

1.5.5 Kiểm tra mô hình

Một số điều kiện có thể sử dụng để kiểm tra sự tin cậy của mô hình như:

- Đánh giá mức độ phù hợp của mô hình hồi quy tuyến tính thông qua kiểm tra giá trị R Square/ Adjusted R Square

- Có sự độc lập giữa các quan sát trong phần dư, không có sự tương quan giữa các quan sát của phần dư theo thời gian. Sự độc lập giữa các quan sát có thể được kiểm tra qua kiểm định Durbin-Watson.
- Phương sai của phần dư bằng nhau hay có sự đồng nhất về phương sai của phần dư. Phần dư (residual errors) có phân phối chuẩn hoặc xấp xỉ phân phối chuẩn. Có thể kiểm tra bằng các kết quả thống kê giá trị của phần dư
- Đánh giá về mức độ phù hợp của mô hình hồi quy tuyến tính này có suy rộng và áp dụng được cho tổng thể hay không thông qua kiểm định F và kiểm định T. Ý nghĩa của việc này đó là chúng ta dùng 1 tập dữ liệu để làm mẫu xây dựng mô hình và coi như kết quả mô hình với tập dữ liệu mẫu là tin cậy, bây giờ chúng ta kiểm tra xem nó có còn tin cậy với các trường hợp khác ngoài mẫu không.

Thông qua các thông số: R Square/ Adjusted R Square, MSE, RMSE, ...

Nếu không mô hình nào cho ra kết quả có độ chính xác chấp nhận được, ta quay lại bước 4 để chọn mô hình thay thế



Hình 1.12 Quy trình xây dựng mô hình dự báo

1.6 Vai trò của dataset và chất lượng của dataset trong xây dựng mô hình dự báo

Chất lượng của mô hình phụ thuộc rất nhiều vào chất lượng của dữ liệu, vì vậy việc chuẩn bị dữ liệu, kiểm định biến tốt sẽ có tác động lớn đến kết quả, thời gian làm việc này thông thường sẽ chiếm 80% thời lượng xây dựng 1 mô hình.

CHƯƠNG 2: THIẾT KẾ VÀ CÀI ĐẶT MÔ HÌNH DỰ BÁO

2.1 Chuẩn bị dữ liệu

Dữ liệu được lấy từ file “100,000 UK Used Car Data set” trên website kaggle.com

2.2 Kiểm tra và loại biến

- Sau khi xem xét thì thấy dữ liệu có 9 hãng xe (audi, bmw, ford, hyundi, merc, skoda, toyota, vauxhall, vw) và có 9 thuộc tính gồm: model, year, price, transmission, mileage, fuelType, tax, mpg, engineSize.
- Nạp dữ liệu xe audi
- Kiểm tra dữ liệu: data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10666 entries, 0 to 10665
Data columns (total 9 columns):
#   Column          Non-Null Count  Dtype
---  -
0   model           10666 non-null  object
1   year            10666 non-null  int64
2   price           10666 non-null  int64
3   transmission    10666 non-null  object
4   mileage         10666 non-null  int64
5   fuelType       10666 non-null  object
6   tax             10666 non-null  int64
7   mpg            10666 non-null  float64
8   engineSize     10666 non-null  float64
dtypes: float64(2), int64(4), object(3)
memory usage: 750.1+ KB
```

Hình 2.1 Thông tin dữ liệu

Nhận xét: Dữ liệu xe audi có 10666 dòng, không có giá trị nào bị null, các dòng dữ liệu có kiểu thống nhất, có 2 cột dạng float, 4 cột dạng integer và 3 cột còn lại model, transmission và engineSize dạng phân loại

- Thống kê dữ liệu: data.describe(include='all')

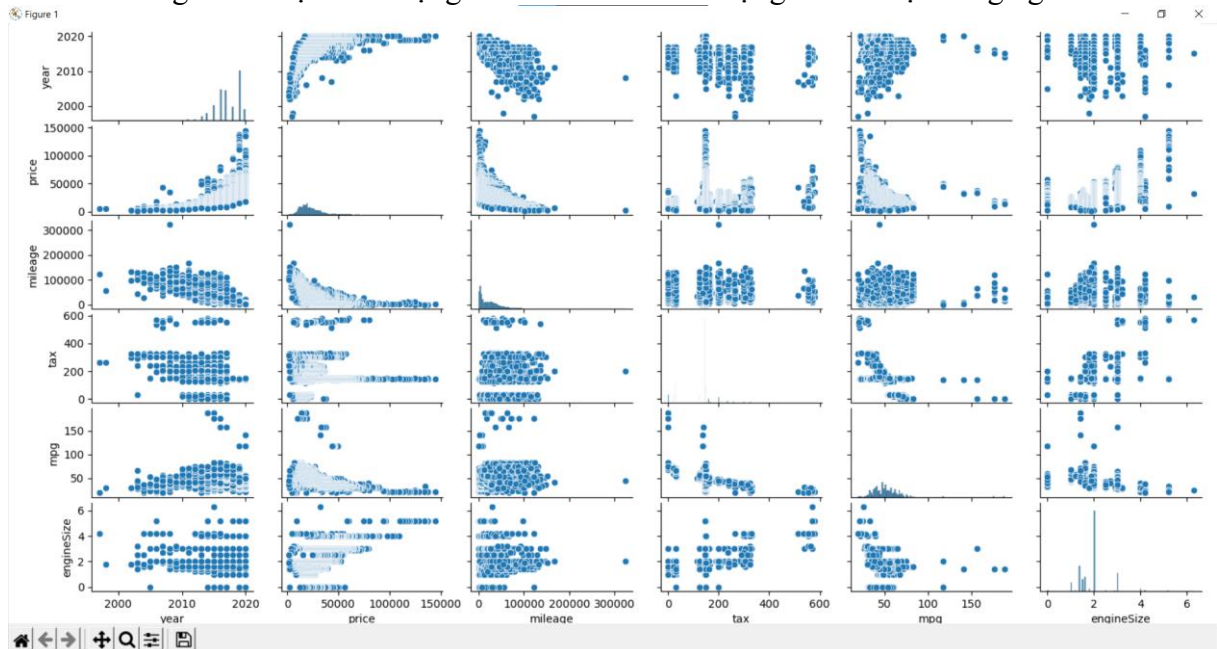
```
model      year      price transmission      mileage \
count  10666  10666.000000  10666.000000      10666  10666.000000
unique      26           NaN           NaN           3           NaN
top         A3           NaN           NaN      Manual           NaN
freq      1929           NaN           NaN      4369           NaN
mean      NaN  2017.100506  22896.834427      NaN  24824.231296
std      NaN    2.167489  11715.308349      NaN  23501.291429
min      NaN  1997.000000  1490.000000      NaN    1.000000
25%      NaN  2016.000000  15145.250000      NaN   5974.000000
50%      NaN  2017.000000  20200.000000      NaN  19000.000000
75%      NaN  2019.000000  27990.000000      NaN  36461.750000
max      NaN  2020.000000  145000.000000      NaN  323000.000000

fuelType      tax      mpg      engineSize
count  10666  10666.000000  10666.000000  10666.000000
unique      3           NaN           NaN           NaN
top      Diesel           NaN           NaN           NaN
freq    5576           NaN           NaN           NaN
mean      NaN   126.018657    50.769651    1.930743
std      NaN    67.169906    12.950426    0.602998
min      NaN    0.000000   18.900000    0.000000
25%      NaN   125.000000   40.900000    1.500000
50%      NaN   145.000000   49.600000    2.000000
75%      NaN   145.000000   58.900000    2.000000
max      NaN   580.000000  188.300000    6.300000
```

Hình 2.2 Thống kê dữ liệu

Nhận xét:

- Dễ nhận thấy dữ liệu engineSize (kích thước động cơ) có giá trị nhỏ nhất là 0 => giá trị sai.
 - Tax có giá trị nhỏ nhất là 0, giá trị này có thể sai, có thể xem xét xử lý hoặc không.
- Thống kê dữ liệu dưới dạng biểu đồ và đếm số lượng các dữ liệu đáng nghi:



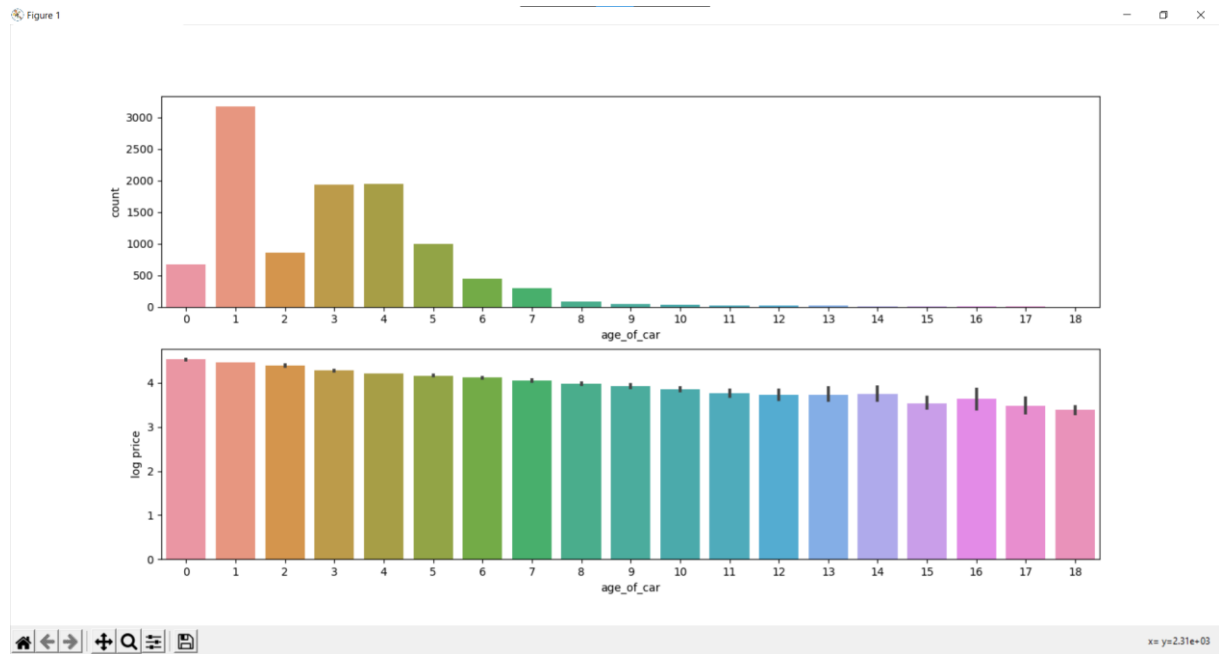
Hình 2.3 Thống kê dữ liệu theo biểu đồ

Nhận xét:

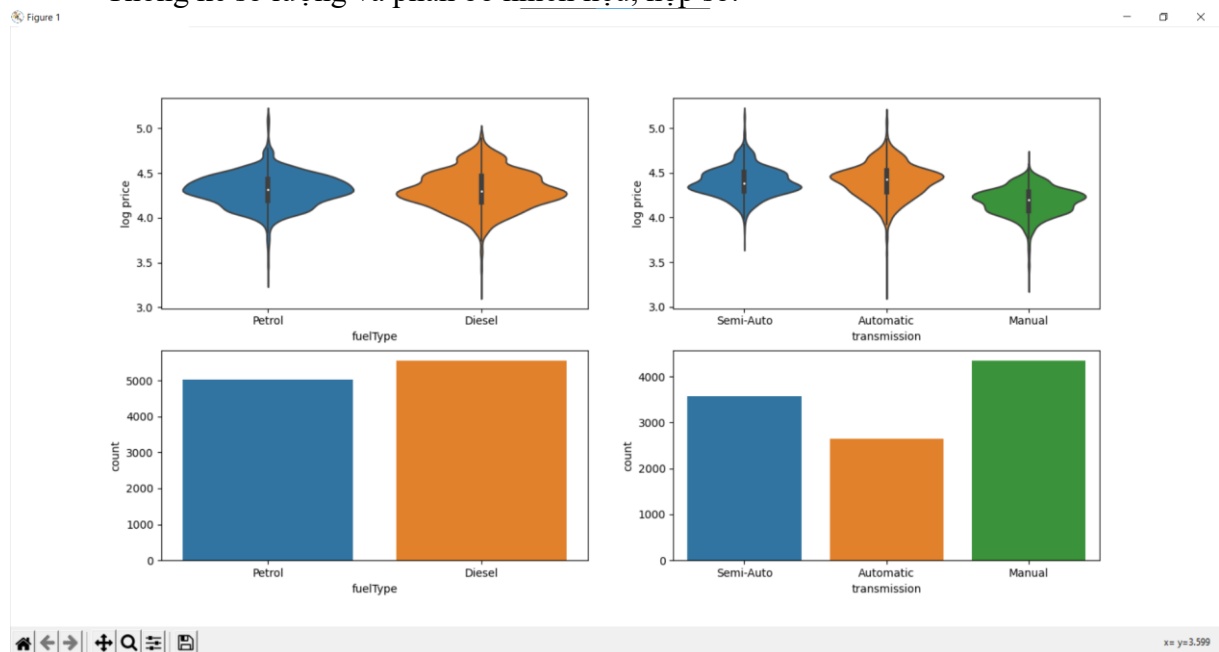
- Giá phụ thuộc vào năm và dặm theo dạng đồ thị hàm log.
- Kích thước động cơ có 57 dòng dữ liệu có giá trị 0.
- Số dặm có 1 dòng dữ liệu bất thường (>300000) nếu so với các giá trị còn lại
- Mpg có 33 giá trị bất thường (>100) nếu so với các giá trị còn lại
- Có 2 chiếc xe trước năm 2000 cũng có vẻ bất thường
- Có 536 xe có tax bằng 0.

Vì số lượng các dữ liệu bất thường ngoại trừ Tax đều rất nhỏ so với tổng 10666 dòng dữ liệu nên có thể bỏ qua các giá trị này ở bước xây dựng mô hình và thống kê tiếp theo

- Thống kê log giá theo số năm:

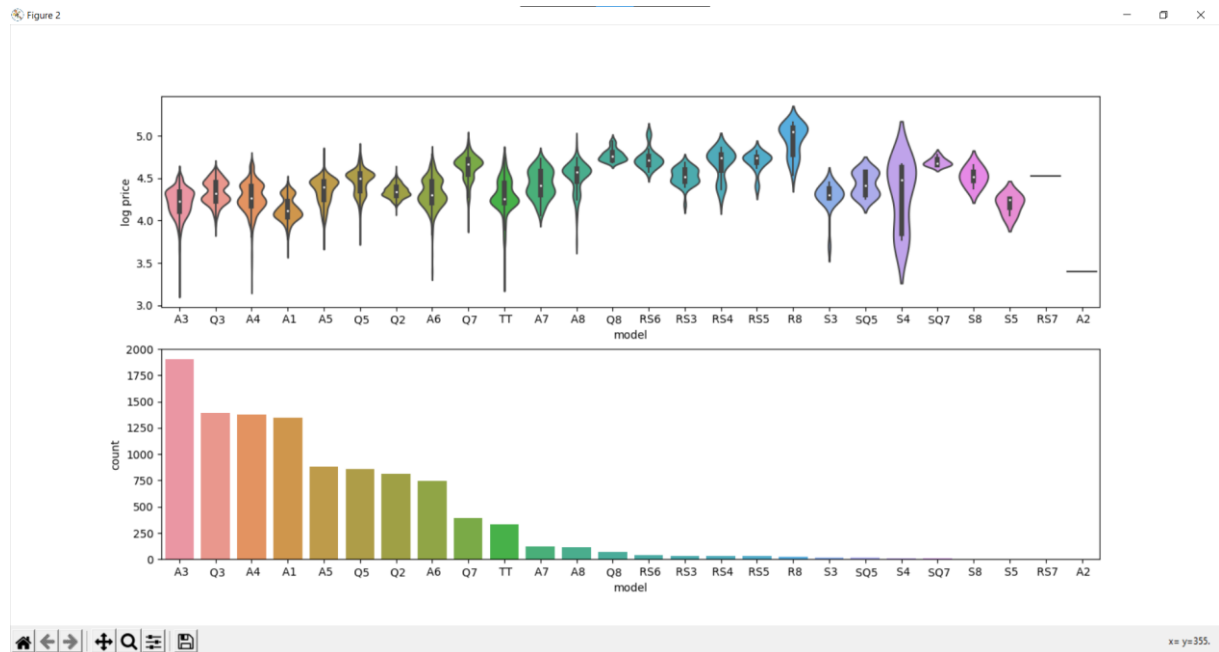
Hình 2.4 Thống kê `age_of_car`

- Thống kê số lượng và phân bố nhiên liệu, hộp số:

Hình 2.5 Thống kê `fuelType` và `transmission`

Nhận xét: Mọi dữ liệu khá tương đồng ngoại trừ giá xe hộp số thủ công rẻ hơn một chút so với 2 loại còn lại.

- Thống kê số lượng và giá cả theo mẫu xe:



Hình 2.6 Thống kê model

Nhận xét:

- Số lượng các mẫu xe chênh lệch khá nhiều, đặc biệt 2 mẫu RS7 và A2 chỉ có một dòng dữ liệu, khác với dữ liệu liên tục, dữ liệu về mẫu xe quá ít rất khó đưa kết quả dự đoán chính xác về mẫu xe này.
- Các mẫu xe khác nhau có khung giá khác nhau, điều này đúng với thực tế.

2.3 Xử lý biến không đạt chuẩn

- Bỏ tất cả các dữ liệu có engineSize=0
- Bỏ tất cả các dòng dữ liệu có mileage > 300000
- Bỏ tất cả các dòng dữ liệu có mpg > 100
- Bỏ tất cả dữ liệu có năm > 2020 và < 2000
- Biến đổi giá trị cột year thành cột chứa giá trị tuổi của xe bằng phép tính 2020 – year

Lý do:

- 2020 là năm dữ liệu được thống kê
- giá trị tuổi của xe là giá trị ảnh hưởng đến giá trực tiếp nhất
- Bỏ tất cả các giá trị thuộc các cột model, transmission và fuelType có số lượng < 20 dòng dữ liệu (20 là số được lấy một cách chủ quan sau khi đếm được số lượng dòng dữ liệu ở các cột này)

model	count	RS6	39
A2	1	Q8	69
RS7	1	A8	117
S5	3	A7	122
S8	4	TT	334
SQ7	8	Q7	394
S4	9	A6	746
SQ5	14	Q2	816
S3	18	Q5	860
R8	28	A5	879
RS5	29	A1	1345
RS4	31	A4	1380
RS3	33	Q3	1390
		A3	1906

Hình 2.7 Thống kê số lượng mẫu xe

2.4 Tiền xử lý

- Tính log cho giá cả và thay thế cho cột price

	model	transmission	mileage	mpg	engineSize	age_of_car
6627	T-Cross	Manual	2694	47.9	1.0	1
4235	Golf	Manual	37958	60.1	1.4	3
2184	Golf	Manual	25867	53.3	1.4	5
13680	Touareg	Automatic	35	34.9	3.0	1
13048	Scirocco	Semi-Auto	19313	57.6	2.0	3

Hình 2.8 Mẫu dữ liệu ban đầu

- Dùng OneHotEncoder() để mã hóa các tính năng model, transmission và fuelType thành các cột nhị phân. Vd: model có các loại A1, A2 sẽ thành -> model_A1, model_A2. Mã hóa thành nhị phân, xe có loại A1 thì cột model_A1 có giá trị 1, còn lại giá trị 0

	mileage	mpg	engineSize	age_of_car	model_Amarok	model_Arteon	\
6627	2694	47.9	1.0	1	0	0	
4235	37958	60.1	1.4	3	0	0	
2184	25867	53.3	1.4	5	0	0	
13680	35	34.9	3.0	1	0	0	
13048	19313	57.6	2.0	3	0	0	
	model_Beetle	model_CC	model_Caddy Maxi	Life	model_Caravelle	\	
6627	0	0		0	0		
4235	0	0		0	0		
2184	0	0		0	0		
13680	0	0		0	0		
13048	0	0		0	0		
	model_Golf	model_Golf SV	model_Jetta	model_Passat	model_Polo	\	
6627	0	0	0	0	0		
4235	1	0	0	0	0		
2184	1	0	0	0	0		
13680	0	0	0	0	0		
13048	0	0	0	0	0		
	model_Scirocco	model_Sharan	model_Shuttle	model_T-Cross	\		
6627	0	0	0	1			
4235	0	0	0	0			
2184	0	0	0	0			
13680	0	0	0	0			
13048	1	0	0	0			
	model_T-Roc	model_Tiguan	model_Tiguan Allspace	model_Touareg	\		
6627	0	0	0	0			
4235	0	0	0	0			
2184	0	0	0	0			
13680	0	0	0	1			
13048	0	0	0	0			
	model_Touran	model_Up	transmission_Automatic	transmission_Manual	\		
6627	0	0	0	1			

Hình 2.9 Mẫu dữ liệu one_hot_encode

- Dùng StandardScaler() để chuẩn hóa dữ liệu số theo cùng một tiêu chuẩn. Vd engineSize từ 1-6 và mileage từ 1-323000 Vì hai cột này có quy mô khác nhau, chúng được Chuẩn hóa để có tỷ lệ chung trong khi xây dựng mô hình học máy.

2.5 Huấn luyện mô hình

- Sử dụng hàm train_test_split() để chia dữ liệu thành phần train và phần test với test_size bằng 20%
- Sau khi đã biến đổi log cột price, nhận thấy mô hình tuyến tính với age_of_car và mileage => sử dụng mô hình Linear
- Thêm lần lượt từng cột chức năng và dựa vào các thông số r-square, std, mse, rmse để chuẩn đoán tầm quan trọng của từng tính năng, xem xét loại bỏ các tính năng ảnh

hưởng xấu đến mô hình thông qua hàm `cross_validate()` và công thức tính `mse`, `rmse`. Trong đó `mse` và `rmse` được tính toán với giá trị ban đầu của `price`

	mean_test_score	mean_train_score	std_test_score	\
Last added feature				
mileage	0.468502	0.470392	0.022205	
age_of_car	0.585736	0.588240	0.031650	
model	0.891470	0.892444	0.005316	
transmission	0.911591	0.912441	0.004757	
engineSize	0.933049	0.933789	0.004430	
mpg	0.939825	0.940557	0.003717	
fuelType	0.940559	0.941296	0.003556	
tax	0.940744	0.941554	0.003594	

	std_train_score	mse	rmse
Last added feature			
mileage	0.005651	1.095691e+08	10467.524770
age_of_car	0.008006	9.697135e+07	9847.403369
model	0.001317	1.928098e+07	4391.011765
transmission	0.001120	1.719396e+07	4146.559913
engineSize	0.001035	1.074739e+07	3278.321586
mpg	0.000867	9.739082e+06	3120.750260
fuelType	0.000824	9.728725e+06	3119.090351
tax	0.000843	9.589680e+06	3096.720812

Hình 2.10 Thống kê chọn tính năng

Nhận xét:

- Tính năng Tax không ảnh hưởng lớn đến điểm số, nhưng lại làm tăng phương sai (std), Tax cũng có nhiều dữ liệu đáng nghi (giá trị 0) => ta lược bỏ tính năng này.
- Tính năng fuelType cũng ít ảnh hưởng đến kết quả dự đoán, có thể lược bỏ dữ liệu này, nhưng vì mục tiêu hướng đến là ứng dụng dự đoán có khả năng dự đoán dù không nhập đầy đủ dữ liệu => ta có thể giữ lại.

- Tính vif để kiểm tra xem mô hình có bị vấn đề đa cộng tuyến không

```

=====vif=====
const                201.487303
onehotencoder__x0_ A3    2.195527
onehotencoder__x0_ A4    2.333757
onehotencoder__x0_ A5    2.028639
onehotencoder__x0_ A6    2.139135
onehotencoder__x0_ A7    1.374900
onehotencoder__x0_ A8    1.388450
onehotencoder__x0_ Q2    1.724204
onehotencoder__x0_ Q3    2.455155
onehotencoder__x0_ Q5    2.632336
onehotencoder__x0_ Q7    2.550172
onehotencoder__x0_ Q8    1.277626
onehotencoder__x0_ R8    1.363348
onehotencoder__x0_ RS3    1.087380
onehotencoder__x0_ RS4    1.110885
onehotencoder__x0_ RS5    1.113084
onehotencoder__x0_ RS6    1.284418
onehotencoder__x0_ TT    1.402895
onehotencoder__x1_Manual  2.192748
onehotencoder__x1_Semi-Auto 1.620632
onehotencoder__x2_Petrol  2.643925
mileage              3.070222
age_of_car           2.932803
engineSize           4.415349
mpg                  3.705735
dtype: float64

```

Hình 2.11 Chỉ số VIF

- Mô hình không có vấn đề đa cộng tuyến
- Và để đảm bảo rằng dữ liệu không bị overfitted (quá vừa vặn), ta thực hiện hồi quy lasso và ridge, cho một loạt giá trị của các alpha tham số của chúng
- Kết quả: Ridge(tol=1e-9, alpha=0.927), Lasso(alpha=0.00003327) và ElasticNet(alpha=0.00005197) thỉnh thoảng (5/10 lần thử) cho MSE và r-squared tốt hơn Linear
- Tiếp tục thử nghiệm với mô hình Random Forest:

	mean_test_score	mean_train_score	std_test_score \
Last added feature			
mileage	0.317314	0.813026	0.041532
age_of_car	0.445223	0.867388	0.047787
model	0.853519	0.974202	0.005569
transmission	0.883286	0.980737	0.005934
engineSize	0.929912	0.988994	0.003857
mpg	0.954671	0.993284	0.002158
fuelType	0.954956	0.993398	0.002177

	std_train_score	mse	rmse
Last added feature			
mileage	0.002412	1.250159e+08	11181.051225
age_of_car	0.003045	1.147429e+08	10711.809178
model	0.000259	2.286450e+07	4781.684221
transmission	0.000385	1.949628e+07	4415.459603
engineSize	0.000255	9.850071e+06	3138.482229
mpg	0.000267	5.931761e+06	2435.520591
fuelType	0.000250	5.869019e+06	2422.605860

Hình 2.12 Thống kê chọn tính năng cho thuật toán Random Forest

Kết quả: Mô hình Random Forest cho các thông số tốt hơn Linear khi có đầy đủ tính năng. Và ngược lại, khi có ít tính năng thì các mô hình tuyến tính có ưu thế hơn.

- Tuy nhiên chênh lệch không lớn lắm, ta có thể tự do chọn bất kì mô hình nào để xây dựng
- Một lần nữa kiểm tra r-squared chéo bằng hàm `r2_score()` từ thư viện `sklearn.metrics` kết quả không khác biệt lắm với điểm nhận được từ hàm `cross_validate()` chứng tỏ mô hình không bị vấn đề overfitted (quá vừa).

2.6 Đánh giá mô hình dự báo

- Sau khi lần lượt kiểm tra, đánh giá mô hình Linear, Ridge, Lasso, ElasticNet, Random Forest thông qua các dữ liệu như r-squared, r-squared adj, MSE, RMSE đều cho kết quả phù hợp, mỗi mô hình đều có ưu nhược điểm riêng.
- Tiếp tục thực hiện kiểm định F và kiểm định t thông qua thống kê OLS từ `statsmodels.api`:

Kết quả Prob (F-statistic) $\sim 0.00 < \alpha = 0.05\% \Rightarrow$ r-squared của tổng thể khác 0, đồng nghĩa với mô hình phù hợp với tổng thể chứ không chỉ đúng với bộ mẫu xây dựng, các biến độc lập có tác động đến biến phụ thuộc

OLS Regression Results						
Dep. Variable:	y	R-squared:		0.942		
Model:	OLS	Adj. R-squared:		0.942		
Method:	Least Squares	F-statistic:		5438.		
Date:	Fri, 19 Nov 2021	Prob (F-statistic):		0.00		
Time:	17:10:21	Log-Likelihood:		13411.		
No. Observations:	8414	AIC:		-2.677e+04		
Df Residuals:	8388	BIC:		-2.659e+04		
Df Model:	25					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	4.3328	0.009	507.618	0.000	4.316	4.350
onehotencoder_x0_A3	0.0456	0.002	22.071	0.000	0.042	0.050

Hình 2.13 Kiểm định mô hình

Ta tiếp tục sử dụng mô hình này với tập dữ liệu đầy đủ và bắt đầu xây dựng ứng dụng

CHƯƠNG 3: LẬP TRÌNH CÀI ĐẶT ỨNG DỤNG DỰ BÁO GIÁ Ô TÔ**3.1 Giao diện**

Dự báo giá oto bán lại

Predict Train

Chọn hãng: Audi

Model:

Year: Predict Year: 2021

Transmission:

Mileage:

Engine size:

Mpg:

Fuel type:

Predict

Hình 3.1 Khung giao diện dự đoán

Dự báo giá oto bán lại

Predict Train

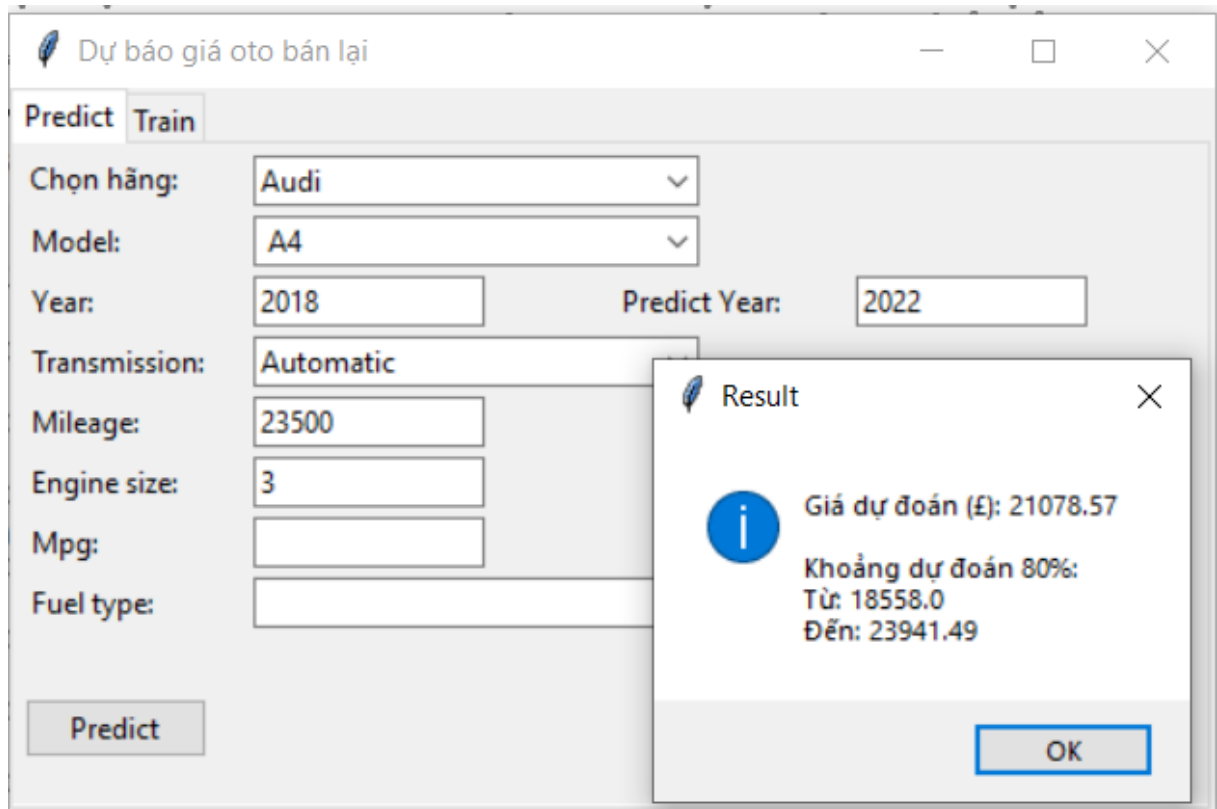
Chọn file train: Select File

Thống kê

Chọn model: LinearRegression

Train

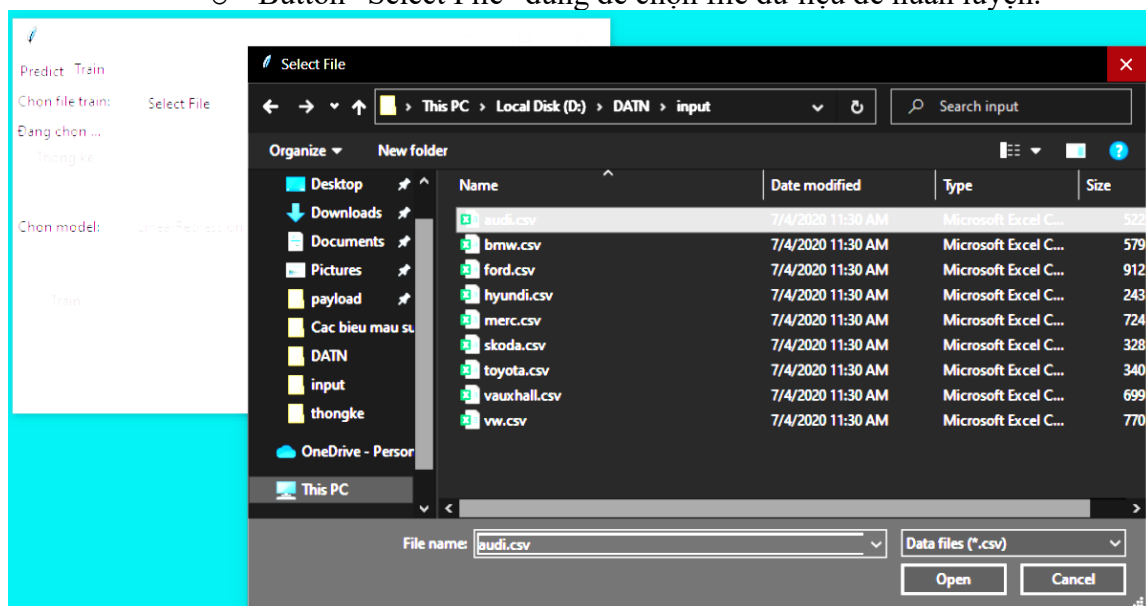
Hình 3.2 Giao diện huấn luyện



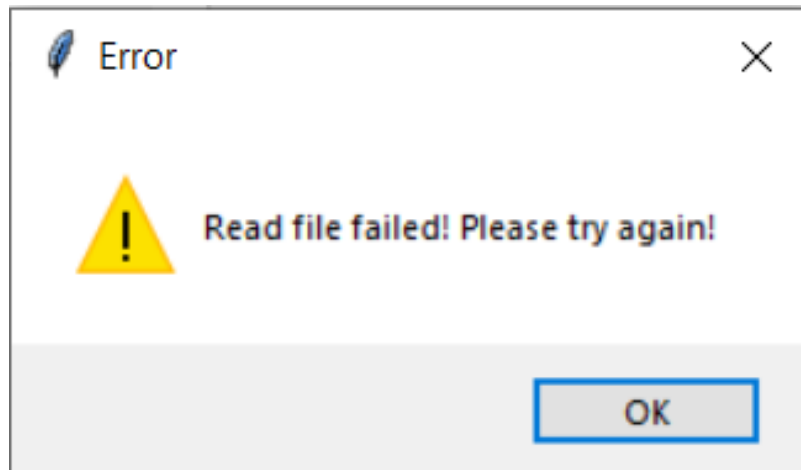
Hình 3.3 Popup kết quả dự đoán

3.2 Giải thích

- Giao diện ứng dụng kích cỡ 500x300 gồm 2 frame “Predict” và “Train”.
 - o Khung Predict dùng để dự đoán giá.
 - o Khung Train dùng để huấn luyện mô hình dùng cho việc dự đoán.
- Khung Train bao gồm các thành phần:
 - o Button “Select File” dùng để chọn file dữ liệu để huấn luyện.

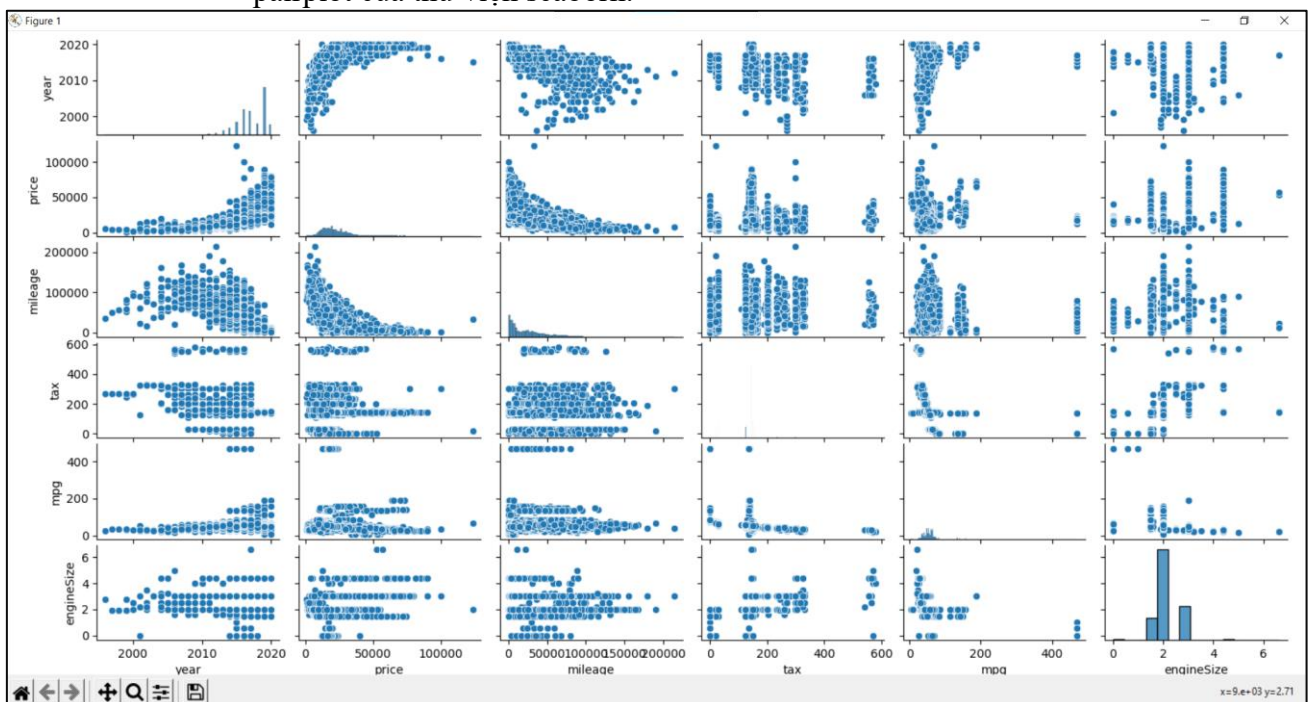


Hình 3.4 Giao diện sự kiện chọn file dữ liệu



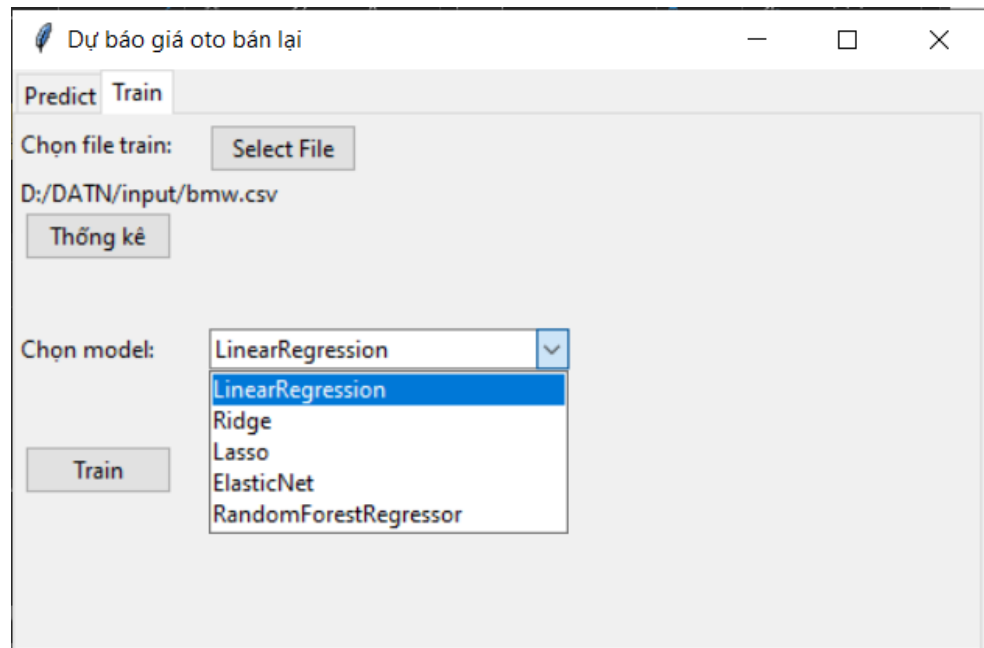
Hình 3.5 Popup sự kiện đọc file không thành công

- Button “Thống kê” chỉ hoạt động khi đã chọn file dữ liệu, dùng để cung cấp cho người sử dụng góc nhìn sơ bộ về dữ liệu đã chọn thông qua hàm pairplot của thư viện seaborn.



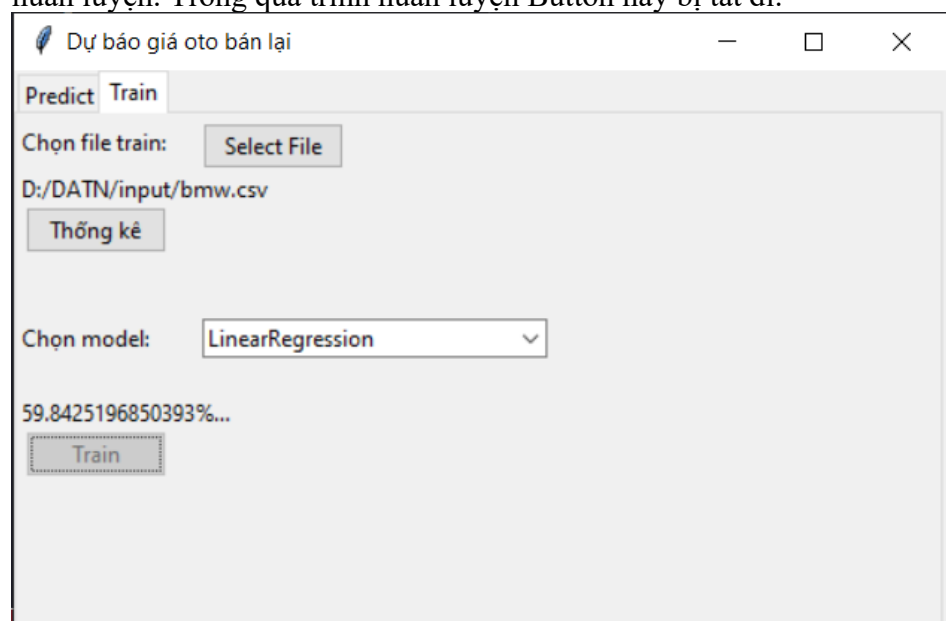
Hình 3.6 Giao diện sự kiện thống kê dữ liệu

- Combobox dạng drop down lựa chọn thuật toán sẽ dùng để huấn luyện mô hình, thành phần này cũng chỉ hoạt động sau khi đã chọn file dữ liệu huấn luyện.

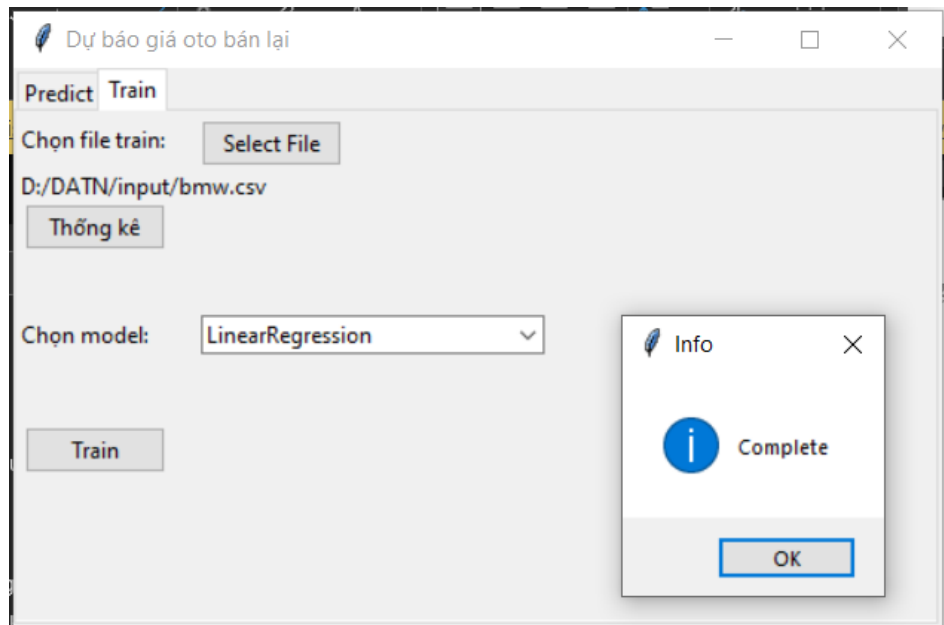


Hình 3.7 Giao diện sự kiện chọn mô hình huấn luyện

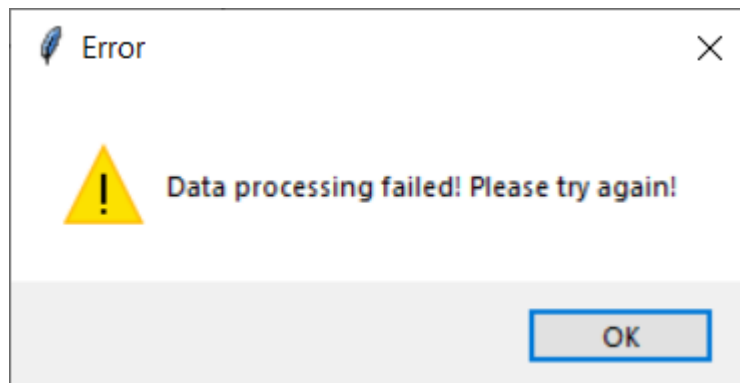
- Button Train thực hiện huấn luyện mô hình theo mô hình đã chọn ở combobox. Dữ liệu sẽ được xử lý và tiền xử lý theo quy trình đã đề ra ở phần thiết kế mô hình chương 2. Kết quả của việc huấn luyện sẽ được lưu dưới dạng file pickle (.pkl) bằng hàm dump trong thư viện pickle. Quá trình huấn luyện sẽ thực hiện 2^x lần cho từng bộ tổ hợp để đảm bảo ứng dụng có khả năng đưa ra dự đoán khi thiếu dữ liệu bất kì. Button này được hoạt động khi đã chọn dữ liệu và sau khi hoàn thành huấn luyện. Trong quá trình huấn luyện Button này bị tắt đi.



Hình 3.8 Giao diện khung Train khi đang trong quá trình huấn luyện

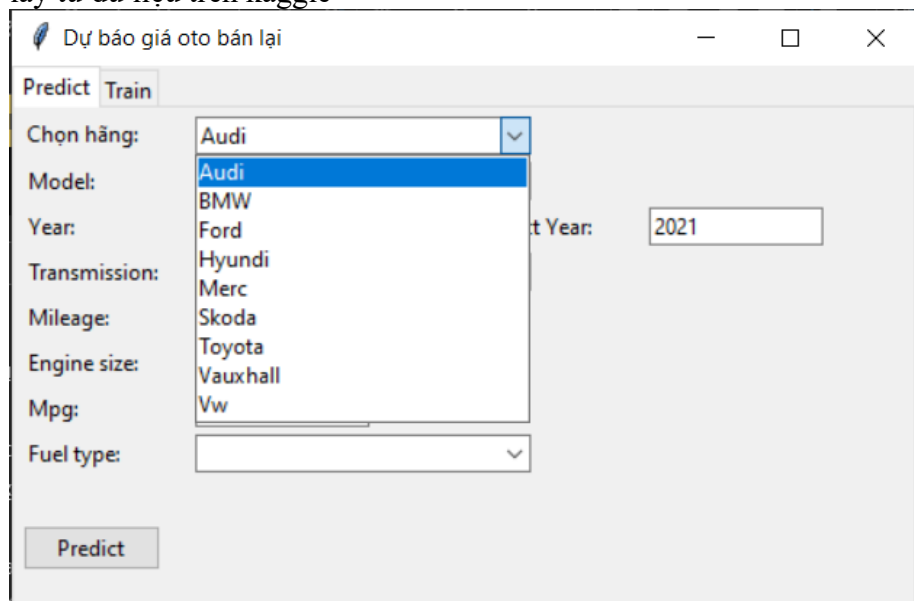


Hình 3.9 Giao diện và popup khi đã hoàn thành huấn luyện



Hình 3.10 Popup sự kiện xử lý dữ liệu thất bại

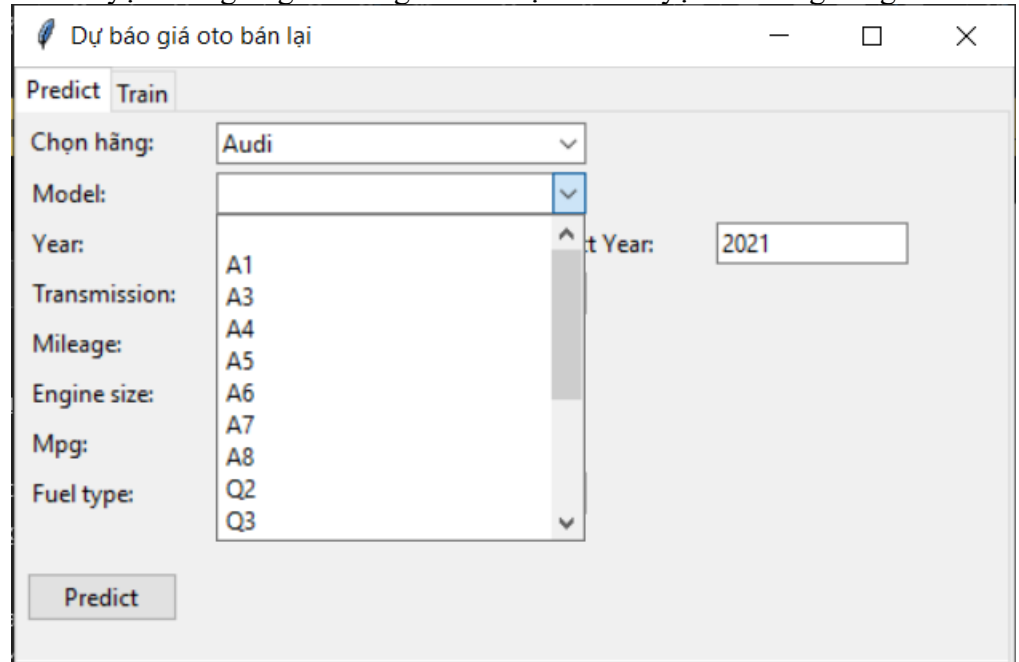
- Khung Predict bao gồm các thành phần:
 - Combobox dạng dropdown chọn hãng xe tương ứng với số lượng hãng xe lấy từ dữ liệu trên kaggle



Hình 3.11 Giao diện combobox chọn hãng xe

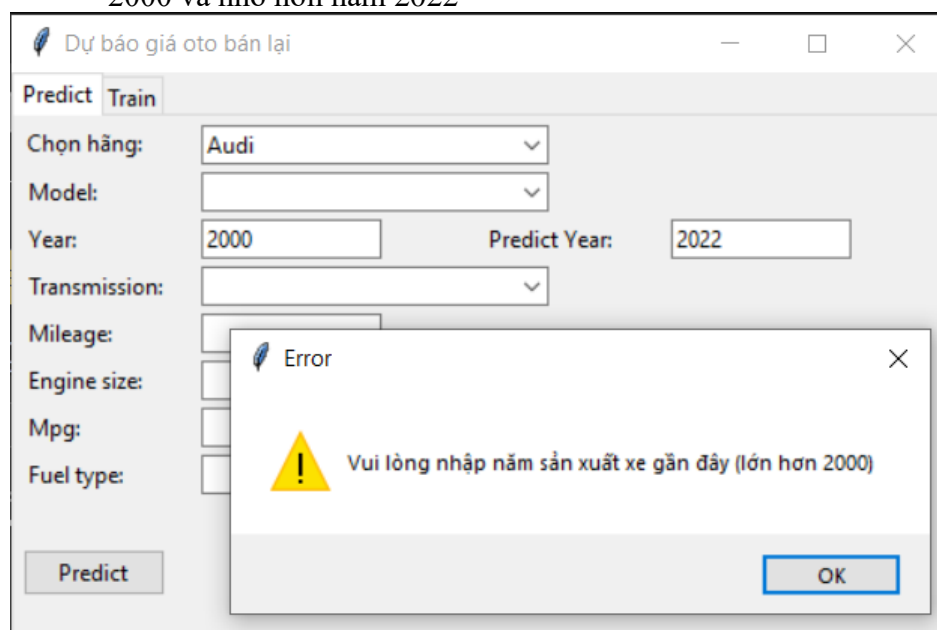
Ứng dụng sẽ load lại dữ liệu đã huấn luyện nếu thay đổi hãng xe đã chọn bằng hàm load của thư viện pickle, nếu không tìm thấy file dữ liệu huấn luyện, button Predict sẽ bị tắt.

- Combobox chọn mẫu xe, combobox này sẽ load các mẫu xe đã dùng trong huấn luyện tương ứng với từng file dữ liệu huấn luyện của từng hãng xe.

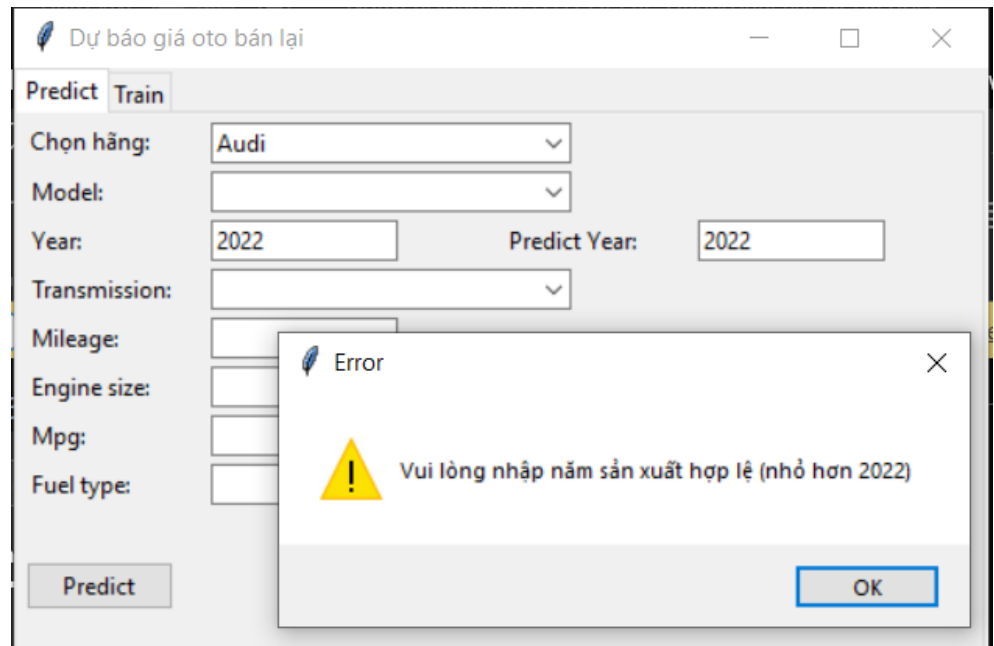


Hình 3.12 Giao diện combobox chọn mẫu xe

- Ô input nhập năm sản xuất của xe, nội dung ô này phải là số năm lớn hơn 2000 và nhỏ hơn năm 2022

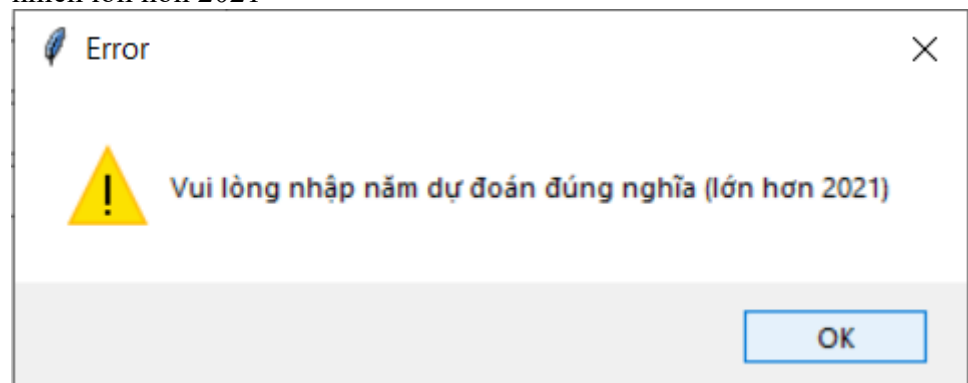


Hình 3.13 Giao diện popup sự kiện năm sản xuất cách quá xa



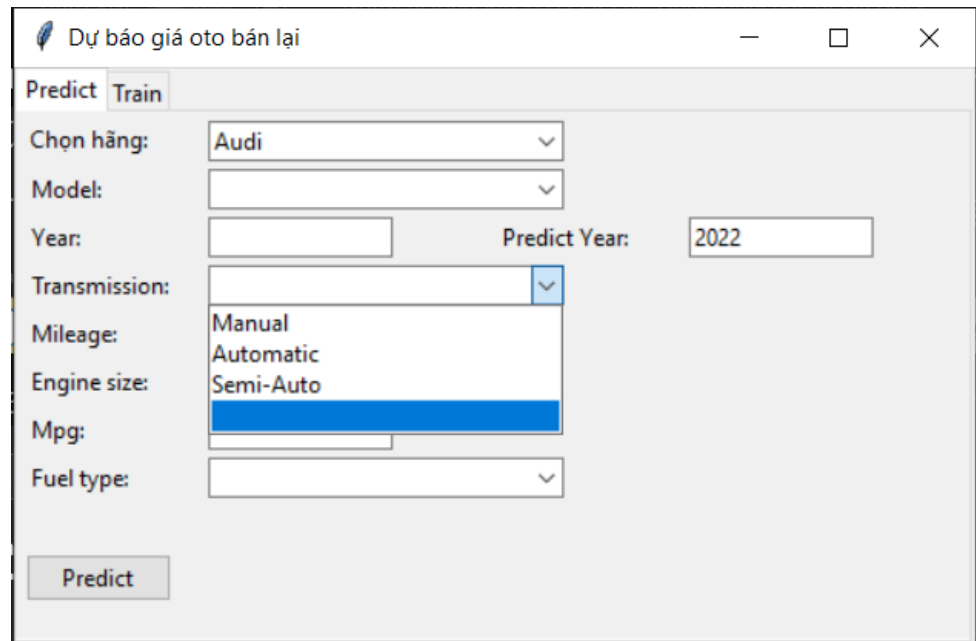
Hình 3.14 Giao diện popup sự kiện năm sản xuất không nhỏ hơn 2022

- Ô input năm dự đoán, mặc định giá trị 2022, nội dung này phải là số tự nhiên lớn hơn 2021



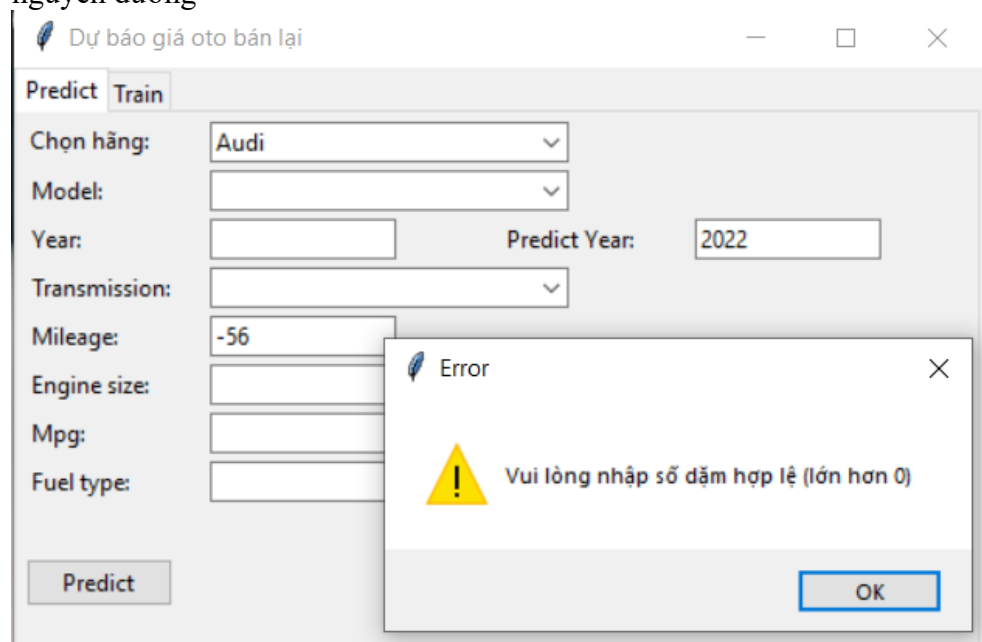
Hình 3.15 Giao diện popup sự kiện năm dự đoán ở quá khứ

- Combobox chọn loại hộp số của xe, nội dung combobox cũng tương ứng với dữ liệu đã huấn luyện



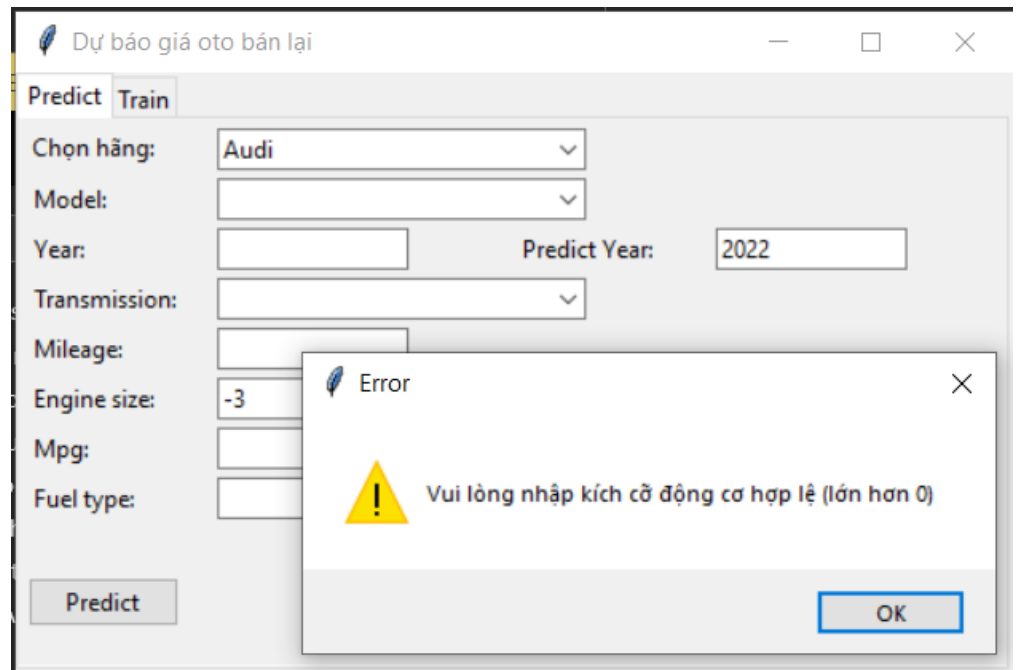
Hình 3.16 Giao diện combobox chọn loại hộp số

- Ô input nhập dữ liệu số dặm đã chạy của xe, nội dung này phải là số nguyên dương

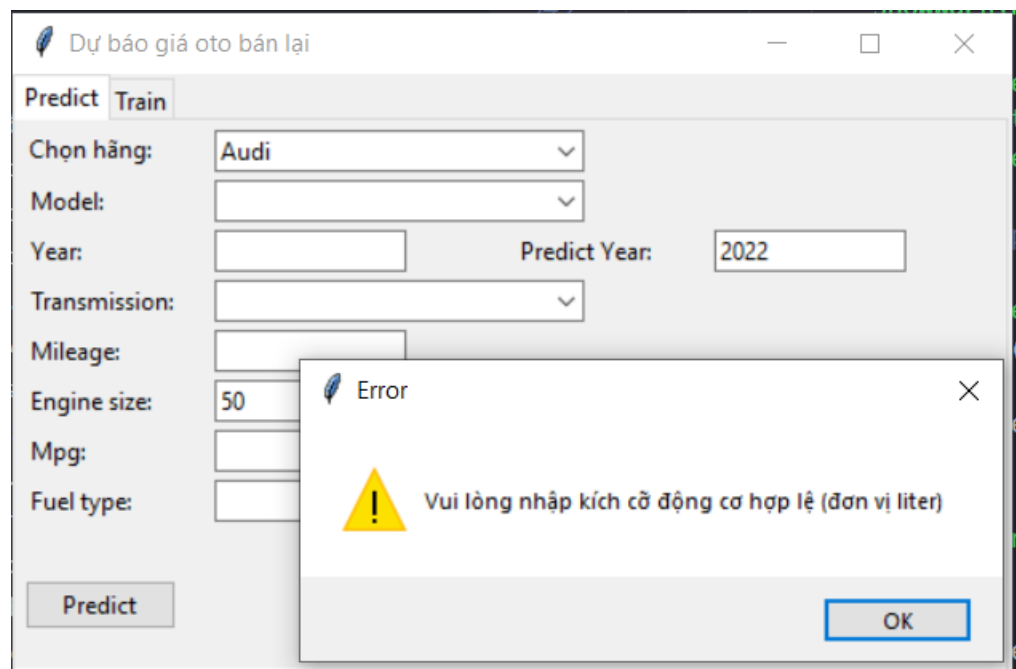


Hình 3.17 Giao diện popup sự kiện số dặm không hợp lệ

- Ô input nhập kích cỡ động cơ, nội dung này phải là số thập phân lớn hơn 0, và phải nhỏ hơn 50 lít (ngưỡng được lấy một cách chủ quan để nhắc nhở người dùng nếu sai đơn vị)

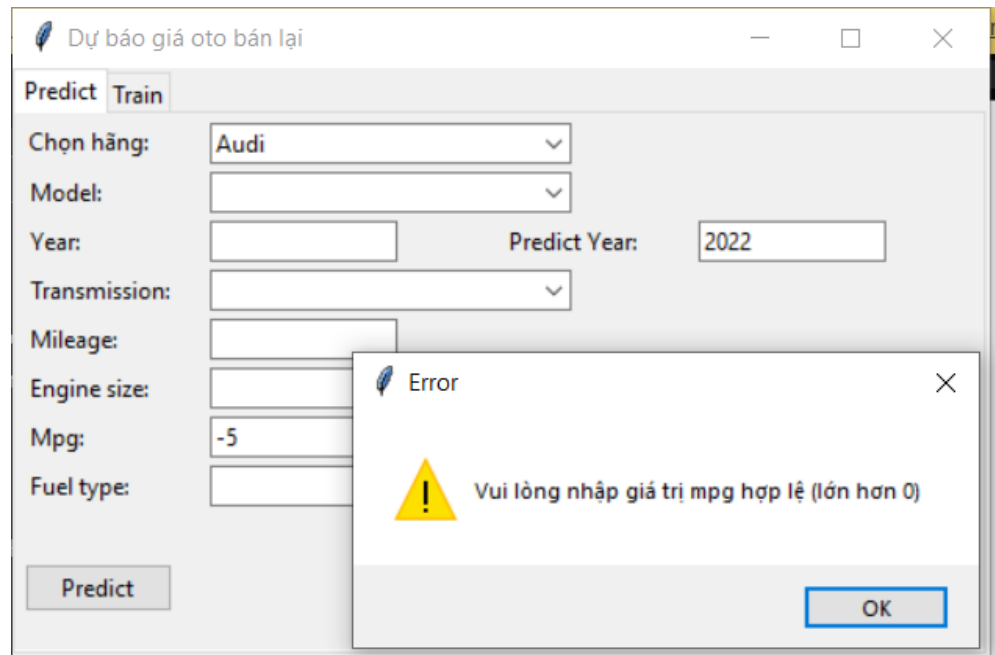


Hình 3.18 Giao diện popup sự kiện kích cỡ động cơ không hợp lệ



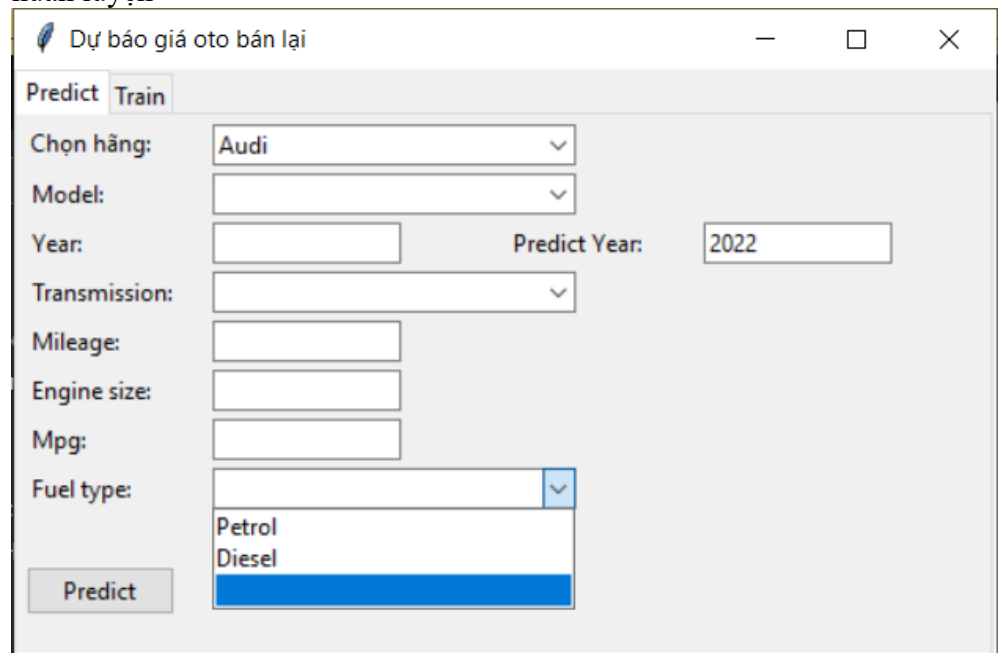
Hình 3.19 Giao diện popup sự kiện kích cỡ động cơ quá lớn

- Ô input nhập mức độ tiết kiệm nhiên liệu (mpg - miles per gallon), nội dung này phải là số thập phân lớn hơn 0



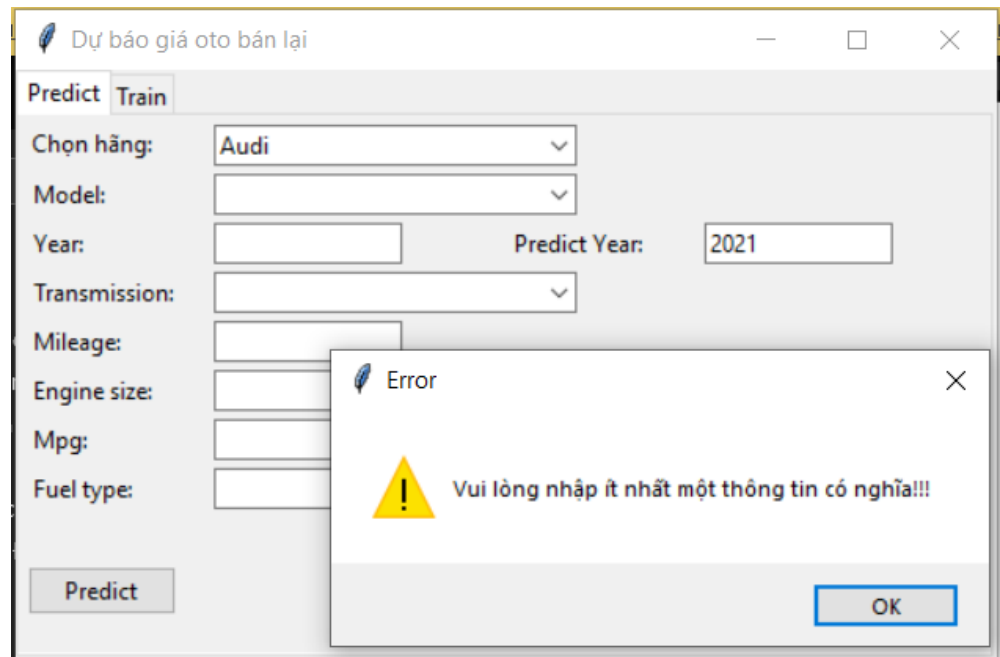
Hình 3.20 Giao diện popup sự kiện dữ liệu mpg không hợp lệ

- Combobox chọn loại nhiên liệu, nội dung này tương ứng với dữ liệu đã huấn luyện



Hình 3.21 Giao diện combobox loại nhiên liệu

- Button Predict thực hiện dự đoán giá, phải nhập ít nhất một dữ liệu có nghĩa, và các dữ liệu phải đúng định dạng

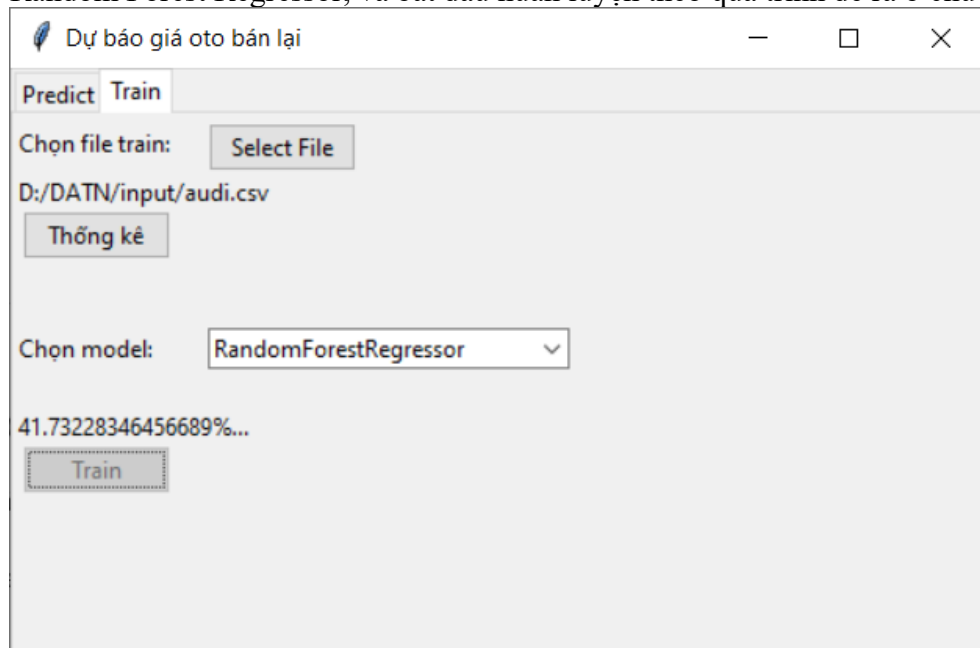


Hình 3.22 Giao diện popup sự kiện dự đoán thất bại do thiếu dữ kiện

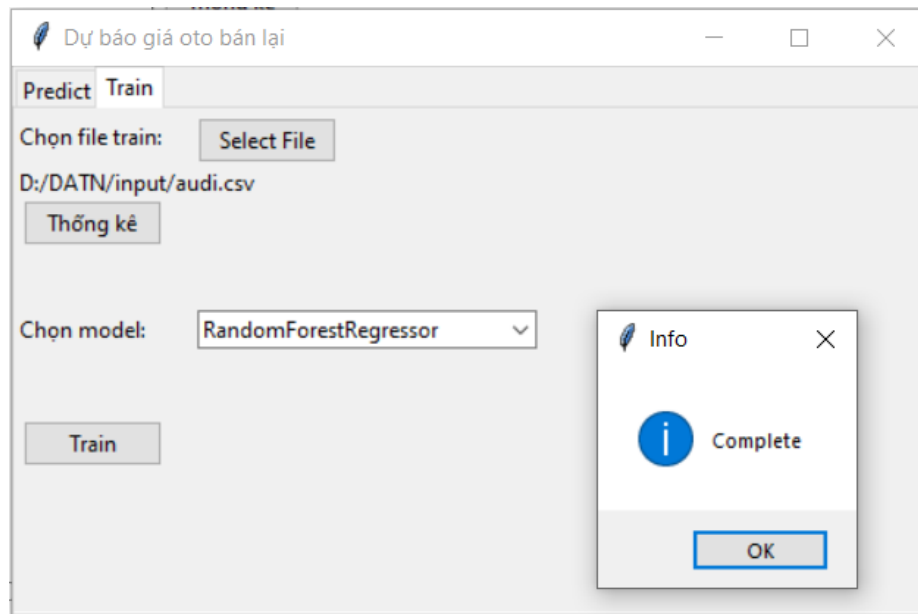
3.3 Chạy thử ứng dụng

- Vì đã phân tích và đánh giá mô hình dự báo ở chương 2, ta đã kết luận mô hình này có thể sử dụng trong việc dự báo, tiếp đó, ta đánh giá lần nữa tính ứng dụng thực tế của mô hình.
- Lấy ngẫu nhiên 5 giá trị từ tập dữ liệu, sau đó xóa các dữ liệu này ra khỏi tập dataset và cho chương trình huấn luyện lại.
- Chạy ứng dụng và kiểm tra kết quả so với dữ liệu thử nghiệm.

Bước 1: Chọn lại file dữ liệu đã lấy ra 5 giá trị ngẫu nhiên, chọn thuật toán Random Forest Regressor, và bắt đầu huấn luyện theo quá trình đề ra ở chương 2

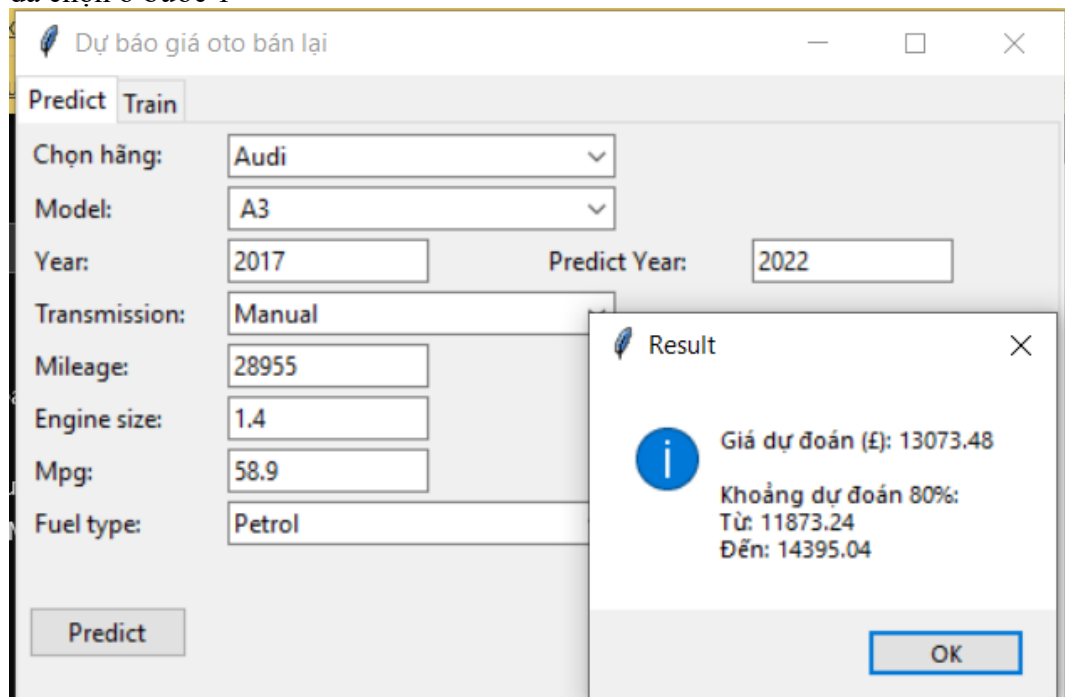


Hình 3.23 Giao diện chạy thử phần huấn luyện



Hình 3.24 Giao diện đã hoàn thành huấn luyện

Bước 2: Sau khi đã hoàn thành, chuyển qua khung Predict và nhập lần lượt 5 xe đã chọn ở bước 1



Hình 3.25 Ví dụ 1 chạy thử ứng dụng

Lần lượt dự đoán các xe còn lại ta được kết quả là bảng sau:

Mẫu xe	Năm sản xuất	Loại hộp số	Số dặm	Loại nhiên liệu	Mpg	Kích cỡ động cơ	Giá trị thực tế năm 2020	Giá trị dự đoán vào năm 2022
A3	2017	Manual	28955	Petrol	58.9	1.4	16100	13073.48
A5	2017	Automatic	37100	Diesel	67.3	2	17300	15192.66
A6	2016	Automatic	34030	Diesel	58.9	2	19400	15084.99
Q3	2017	Automatic	19319	Petrol	40.4	2	21800	19990.05
Q5	2016	Semi-Auto	49649	Diesel	42.2	3	30000	21527.58

Bảng 3.1 Bảng kết quả chạy thử ứng dụng với các mẫu thử ngẫu nhiên

❖ Nhận xét

- Sau 2 năm tính từ thời điểm thu thập dữ liệu, tất cả giá xe đều giảm, điều này đúng với thực tế.
- Xe càng cũ (sản xuất năm 2016) thì chênh lệch giá giữa năm 2020 so với năm 2022 càng lớn nếu so với xe năm 2017, điều này cũng đúng với thực tế (đồ thị hàm log)

Kết luận: Chênh lệch giá dự đoán không lớn lắm, có thể gần đúng với giá xe trong tương lai. Ta có thể kết luận ứng dụng này làm việc tốt trong thực tế.

CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

4.1 Kết quả đạt được

❖ Lý thuyết

- Hiểu được ý nghĩa của mô hình dự báo trong thị trường kinh doanh
- Hiểu được khái niệm mô hình hồi quy
- Hiểu được các mô hình hồi quy thông dụng
- Hiểu được khái niệm hồi quy kernel
- Hiểu được một số tiêu chí đánh giá mô hình hồi quy
- Nắm được phương pháp xây dựng mô hình dự báo sử dụng hồi quy
- Hiểu được vai trò của dữ liệu trong hồi quy

❖ Thực hành

- Xây dựng được ứng dụng dự báo giá ô tô phù hợp

4.2 Hạn chế

- Giao diện còn nhiều thô sơ, khó ứng dụng vào thực tế

4.3 Hướng khắc phục

- Xây dựng ứng dụng trên điện thoại, web app để dễ dàng tiếp cận

4.4 Hướng mở rộng

- Dự báo giá cho thị trường xe Việt Nam

DANH MỤC TÀI LIỆU THAM KHẢO

Danh mục các Website tham khảo:

1. Ý nghĩa của dự báo: <http://www.dankinhhte.vn/y-nghia-va-vai-tro-cua-phan-tich-va-du-bao-trong-qua-trinh-ra-quyet-dinh-kinh-doanh/>
2. Khái niệm mô hình hồi quy: http://www.ctump.edu.vn/DesktopModules/NEWS/DinhKem/6424_LT-Buoi-5---hoi-quy-tuong-quan---Trung.pdf
3. Một số kỹ thuật hồi quy: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/?utm_source=blog&utm_medium=RideandLassoRegressionarticle
4. Phương pháp xây dựng mô hình dự báo: <https://www.miai.vn/2020/07/27/da-ds-trien-khai-mo-hinh-du-bao-don-gian-chuong-1-2/>
5. <https://dinhnghia.vn/r-square-la-gi-cong-thuc-r-square.html>
6. <https://corporatefinanceinstitute.com/resources/knowledge/other/variance-inflation-factor-vif/>
7. <https://svensmark.jp/blog/car-prices/>
8. <https://www.kaggle.com/ajayabbaraju/car-price-regression>
9. <https://couhpcode.wordpress.com/2018/01/24/random-forest-the-nao-la-mot-rung-ngau-nhien/>