BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KINH TẾ - TÀI CHÍNH THÀNH PHỐ HỒ CHÍ MINH

ĐỒ ÁN HỌC PHẦN

"PHÂN TÍCH DỮ LIỆU VỚI R"

TÊN ĐỀ TÀI:

PHÂN TÍCH THỊ TRƯỜNG XE HƠI ĐÃ QUA SỬ DỤNG TẠI VƯƠNG QUỐC ANH

Giảng viên hướng dẫn: ThS. Tống Thanh Văn

Sinh viên thực hiện :

Ngô Thị Thuỳ Lam 215210030 21D1DA01

TP. Hồ Chí Minh, thứ hai, ngày 08 tháng 4, năm 2024

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KINH TẾ - TÀI CHÍNH THÀNH PHỐ HỒ CHÍ MINH

ĐỒ ÁN HỌC PHẦN

"PHÂN TÍCH DỮ LIỆU VỚI R"

TÊN ĐỀ TÀI:

PHÂN TÍCH THỊ TRƯỜNG XE HƠI ĐÃ QUA SỬ DỤNG TẠI VƯƠNG QUỐC ANH

Giảng viên hướng dẫn: ThS. Tống Thanh Văn

Sinh viên thực hiện :

Ngô Thị Thuỳ Lam 215210030 21D1DA01

TP. Hồ Chí Minh, thứ hai, ngày 08 tháng 4, năm 2024

MỤC LỤC

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

CHƯƠNG 1: TỔNG QUAN

1.1. Tóm tắt nghiên cứu

Trong thế giới ngày càng phát triển và cạnh tranh, việc hiểu rõ hành vi của khách hàng trở nên vô cùng quan trọng, kể cả trong ngành kinh doanh xe hơi đã qua sử dụng. Bài nghiên cứu này nhằm mục đích khám phá và phân tích hành vi của khách hàng khi mua xe hơi đã qua sử dụng, từ đó đề xuất những chiến lược kinh doanh hiệu quả cho công ty.

Tôi sử dụng phương pháp phân tích dữ liệu khoa học để nắm bắt xu hướng, thói quen và yêu cầu của khách hàng. Dựa trên những phân tích này, tôi sẽ đề xuất các chiến lược kinh doanh cụ thể, nhằm tối ưu hóa lợi nhuận và tăng cường sự hài lòng của khách hàng.

Bài nghiên cứu này không chỉ giúp công ty hiểu rõ hơn về thị trường mà mình đang hoạt động, mà còn cung cấp những kiến thức quý giá về cách thức tạo ra giá trị thực sự cho khách hàng, qua đó tạo ra sự khác biệt và cạnh tranh hiệu quả trên thị trường.

1.2. Đồ án

1.2.1. Nhiệm vụ đồ án

1. Tính cấp thiết và lý do hình thành đề tài

Tính cấp thiết: Phân tích hành vi khách hàng và đề ra chiến lược kinh doanh phù hợp cho công ty kinh doanh xe hơi cũ là một đề tài quan trọng và cấp thiết trong lĩnh vực kinh doanh và tiếp thị. Các nhà kinh doanh, các nhà quản lý, và các tổ chức kinh doanh sẽ rất quan tâm đến việc phân tích trước những xu hướng mua sắm của khách hàng để có thể giúp họ đưa ra quyết định kinh doanh thông minh và hiệu quả nhất. Mặc dù chắc chắn sẽ có những sai số tương đối nhưng nó cũng sẽ giúp được cho các nhà kinh doanh giảm thiểu được các rủi ro không mong muốn và tối ưu hóa lợi nhuận.

Lý do hình thành đề tài: Việc có thể phân tích chính xác hành vi mua sắm của khách hàng và đề ra chiến lược kinh doanh phù hợp có tác dụng lớn trong việc giúp các nhà kinh doanh, nhà quản lý và các tổ chức kinh doanh đưa ra những quyết định chiến lược và phòng

ngừa các rủi ro lớn liên quan đến kinh doanh và tiếp thị. Đồng thời, việc này cũng giúp tối ưu hóa lợi nhuận và nâng cao sự hài lòng của khách hàng.

1.2.2. Cấu trúc đồ án

- ✓ CHƯƠNG 1: TỔNG QUAN

Bao gồm tóm tắt nghiên cứu là sự trình bày ngắn gọn về những nghiên cứu đã có và trình bày nên những nhiệm vụ của đồ án và cấu trúc của đồ án một cách ngắn gọn.

- ✓ CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

Chương tập trung chủ yếu vào việc giải thích các cụm từ, nếu những khái niệm khoa học về những từ ngữ mà chúng ta sẽ nghiên cứu trong đề tài này cũng như giới thiệu sơ lược về bộ dữ liệu được sử dụng trong đề tài.

- ✓ CHƯƠNG 3: PHÂN TÍCH THIẾT KẾ

Đây là chương sẽ giới thiệu về ngôn ngữ lập trình đã sử dụng, mô hình phân tích chuỗi thời gian sử dụng trong đề tài này và các lý thuyết cơ bản về mô hình đó.

- ✓ CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

Chương này sẽ cho thấy từng bước thực hiện mô hình đã lý giải trong chương 3 và khởi chạy chi tiết từng bước rõ ràng, cho thấy kết quả, giải thích và nhận xét chung về những kết quả đó.

- ✓ CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ

Chương cuối cùng trong bài báo cáo sẽ rút ra các kết luận cuối cùng sau khi chạy mô hình và dự đoán ở chương 4. Sau đó sẽ đưa ra những kiến nghị về kết quả dự báo cũng như sẽ kiến nghị về các chiến lược đầu tư khả thi nhất dựa theo kết quả dự đoán.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Giới thiệu về dữ liệu

Dữ liệu về các xe hơi cũ mà bài nghiên cứu này đã sử dụng được cập nhật từ trang web "deepvisualmarketing.github.io". Bộ dữ liệu này được công khai với mục đích hỗ trợ nghiên cứu và ứng dụng liên quan đến ngành công nghiệp ô tô như thiết kế hình thức xe, phân tích người tiêu dùng và mô hình hóa bán hàng.

Bộ dữ liệu này bao gồm:

- Dữ liệu bán hàng và thông số kỹ thuật: Bao gồm các thuộc tính cơ bản của xe như tên mô hình, ID mô hình và tên thương hiệu. Dữ liệu bán hàng trong vòng mười năm ở Vương quốc Anh. Bảng giá cho mức giá mới nhất của từng mô hình xe qua các năm.
- Dữ liệu hình ảnh: Gồm 1,451,784 hình ảnh từ 899 mô hình xe phổ biến ở thị trường Vương quốc Anh. Các mô hình xe này bao gồm các mô hình trong hai thập kỷ qua. Tất cả các hình ảnh đều được chỉnh sửa về kích thước 300x300 với nền được loại bỏ.

Trong bài nghiên cứu này, tôi sử dụng 3 bảng dữ liệu gồm:

Ad_table(extra)	Chứa thông tin của 0.27 triệu quảng cáo xe hơi đã qua sử dụng, bao gồm hãng xe, dòng xe, năm đăng kí, số dặm đã đi, thông số động cơ,…
Price	Chứa giá bán của xe mới.
Sales	Nó chứa dữ liệu bán xe của thị trường vương quốc Anh.

2.2. Ngôn ngữ lập trình

Trong đồ án lần này, tôi sử dụng "ngôn ngữ R" làm ngôn ngữ lập trình chính trong việc thực hiện phân tích và xây dựng mô hình dự báo giá xe.

Ngôn ngữ R là một ngôn ngữ lập trình và môi trường có năng lực tính toán mạnh mẽ, rất phổ biến và cực kì linh hoạt được sử dụng rộng rãi trong các lĩnh vực thống kê và phân tích dữ liệu. R cung cấp một loạt các gói thư viện, phần mềm phong phú và có chức năng thống kê mạnh mẽ, khiến nó trở thành một trong những công cụ yêu thích của nhiều nhà khoa học dữ liệu và những nhà nghiên cứu.

R có một cú pháp dễ hiểu và tương đối linh hoạt, cho phép người dùng có thể tạo, kiểm tra và triển khai các mô hình phân tích mang khuynh hướng phức tạp. Với cú pháp dễ đọc, R giúp người dùng dễ dàng thực hiện các phân tích thống kê phức tạp và tạo ra những đồ thị để trực quan hóa kết quả.

2.3. Mô hình hồi quy tuyến tính (Linear Regression)

Mô hình hồi quy tuyến tính (Linear Regression) là một thuật toán học có giám sát (supervised learning) trong máy học. Dưới đây là một số khái niệm cơ bản về mô hình này:

Khái niệm cơ bản: Hồi quy tuyến tính là một phương pháp thống kê dùng để ước lượng mối quan hệ giữa các biến độc lập (input features) và biến phụ thuộc (output target). Nói cách khác, nó mô tả mối quan hệ giữa đầu vào và đầu ra bằng một hàm tuyến tính.

Phương trình hồi quy tuyến tính: Phương trình hồi quy tuyến tính đơn biến có dạng như phương trình đường thẳng

$$y = ax + b$$

với x là biến độc lập và y là biến phụ thuộc vào x. Đối với Hồi quy tuyến tính đa biến, bạn có thể hiểu một cách đơn giản là sẽ có nhiều biến độc lập x_1, x_2,…, x_n và nhiều hệ số a_1, a_2, …, a_n thay vì chỉ một biến x duy nhất.

CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM

Lần lượt đọc và gán 3 bảng Ad_table(extra), Price, Sales vào 3 biến infor, price, sale.

3.1. Tiền xử lí dữ liệu

3.1.1. Đơn giản hoá dữ liệu

Để thuận tiện cho việc kết nối các bảng lại với nhau, trước tiên cần tiến hành đơn giản hoá tên nhà sản xuất và tên dòng xe bằng cách chuyển đổi kí tự thành chữ thường để giảm thiểu rủi ro bất đồng bộ trong quá trình liên kết.

```
convert_to_lower <- function(df, col1, col2) {
  df[[col1]] <- tolower(df[[col1]])
  df[[col2]] <- tolower(df[[col2]])
  return(df)
}

infor <- convert_to_lower(infor, 'Maker', 'Genmodel')
sale <- convert_to_lower(sale, 'Maker', 'Genmodel')
price <- convert_to_lower(price, 'Maker', 'Genmodel')
```

3.1.2. Kết nối bảng

Mức độ sụt giảm của giá bán lại so với giá mua là một trong những phương thức đánh giá quan trọng trong chiến lược kinh doanh, vì vậy tôi sẽ tiến hành kết nối 2 bảng “infor” và “price” thông qua tên nhà sản xuất và dòng xe.

```
infor <- merge(infor, price, by = c('Maker', 'Genmodel', 'Genmodel_ID'),
all = TRUE)
```

3.1.3. Định dạng dữ liệu

```
Maker:         'character' Genmodel:         'character' Genmodel_ID:         'character' Adv_ID:         'character' Adv_year:
       'integer' Adv_month:         'integer' Color:         'character' Reg_year:         'integer' Bodytype:         'character' Runned_Miles:
       'character' Engin_size:         'character' Gearbox:         'character' Fuel_type:         'character' Price:         'integer'
Engine_power:         'numeric' Annual_Tax:         'character' Wheelbase:         'integer' Height:         'integer' Width:         'integer'
Length:         'integer' Average_mpg:         'character' Top_speed:         'character' Seat_num:         'integer' Door_num:
       'integer' Entry_price:         'integer' classify:         'character'
```

Hình 1: Kiểu dữ liệu ban đầu của bảng infor

Để thuận tiện cho việc tính toán, thống kê, tôi cần giữ các cột Runned_miles, Engin_size, Average_mpg, Top_speed, Reg_year ở dạng số thực, vì vậy tôi đã tiến hành loại bỏ các kí tự không cần thiết như đơn vị đo (mile, mpg, L, …) để chuyển thành dạng *numeric/ integer*.

```
# Xây dựng hàm runned_miles để loại bỏ kí tự "mile" (nếu có)
runned_miles <- function(x) {
  x <- as.character(x)
  if (grepl("mile", x)) {
    x <- gsub(" mile", "", x)
    return(x)
  } else {
    return(x)
  }
}

# Áp dụng hàm runned_miles cho cột 'Runned_Miles' của dataframe 'infor'
infor$Runned_Miles <- sapply(infor$Runned_Miles, runned_miles)

# Kiểm tra xem cột 'Runned_Miles' có chứa chuỗi 'mile' không
infor[infor$Runned_Miles %in% grep("mile", infor$Runned_Miles, value =
TRUE), 'Runned_Miles']

# Chuyển đổi cột 'Runned_Miles' thành kiểu dữ liệu float
infor$Runned_Miles <- as.numeric(infor$Runned_Miles)

# Loại bỏ chuỗi 'L' khỏi cột 'Engin_size'
infor$Engin_size <- gsub("L", "", infor$Engin_size)

# Chuyển đổi cột 'Engin_size' thành kiểu dữ liệu float
infor$Engin_size <- as.numeric(infor$Engin_size)

# Loại bỏ chuỗi ' mpg' khỏi cột 'Average_mpg'
infor$Average_mpg <- gsub(" mpg", "", infor$Average_mpg)

# Chuyển đổi cột 'Average_mpg' thành kiểu dữ liệu float
infor$Average_mpg <- as.numeric(infor$Average_mpg)

# Loại bỏ chuỗi ' mph' khỏi cột 'Top_speed'
infor$Top_speed <- gsub(" mph", "", infor$Top_speed)

# Chuyển đổi cột 'Top_speed' thành kiểu dữ liệu float
infor$Top_speed <- as.numeric(infor$Top_speed)
```

```
# Chuyển đổi cột 'Reg_year' thành kiểu dữ liệu integer
infor2$Reg_year <- as.integer(infor2$Reg_year)
```

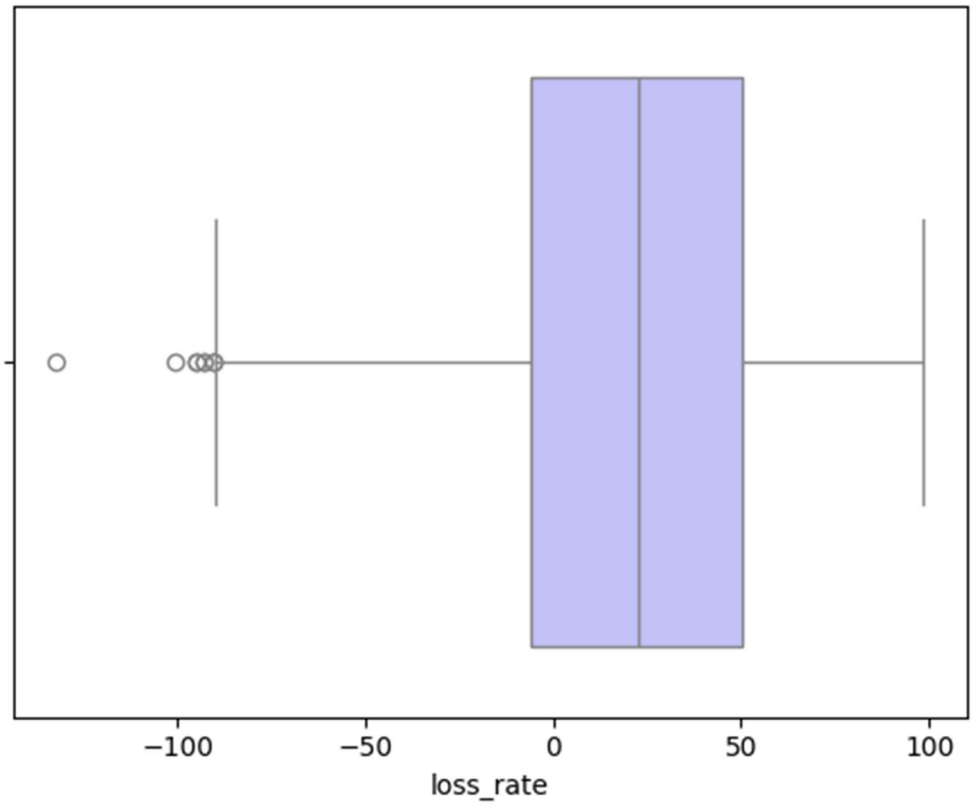
3.2. Phân tích dữ liệu

3.2.1. Tỉ lệ hao hụt về giá trị của xe đã qua sử dụng

Tỉ lệ chênh lệch giữa giá mua và giá bán là một thước đo quan trọng trong việc kinh doanh hàng hoá đã qua sử dụng, kể cả trong ngành công nghiệp xe hơi. Vì vậy tôi đã tính tỉ lệ này với mục đích làm thước đo cho mọi phân tích.

Tỉ lệ hao hụt = (Giá gốc – Giá bán) / Giá bán * 100

```
# Tính 'loss_rate' và thêm vào dataframe 'infor'
infor$loss_rate <- (infor$Entry_price - info2$Price) / infor$Entry_price *
100
```



Hình 2: Sự phân bố của tỉ lệ hao hụt giá khi bán lại xe hơi tại thị trường nước Anh

Từ biểu đồ hộp trên có thể thấy rằng một nửa số xe hơi bán lại có tỉ lệ sụt giảm từ 0 – 50% giá so với giá mua ban đầu, tuy nhiên vẫn có đến gần 25% trường hợp tỉ lệ hao hụt âm, chứng tỏ người bán vẫn có thể kiếm lời từ việc bán lại xe hơi đã qua sử dụng, thậm chí có thể sinh lời lên hơn 100%.

3.2.2. Phân loại tốc độ tăng trưởng doanh thu của các hãng xe

Trong bài phân tích này, tôi lựa chọn phương pháp phân tích hành vi khách hàng theo phân khúc xe. Dựa vào dữ liệu thuộc bảng sales, tôi có sự phân bố dữ liệu về tốc độ tăng trưởng trung bình của các hãng xe như sau:

```
summary(average_growth_speed)
```

```
   maker           average_growth_speed
Length:73          Min.    : -42.857
Class :character   1st Qu.:   1.236
Mode  :character   Median :  10.984
                   Mean    : 149.459
                   3rd Qu.:  19.568
                   Max.    :5200.399
```

25% hãng xe có tốc độ tăng trưởng từ -42% đến 1%

50% hãng xe có tốc độ tăng trưởng từ 1% đến 19%.

25% còn lại có tốc độ tăng trưởng trên 19% và có thể lên đến tối đa là 5200%.

Từ sự phân bố trên, tôi sử dụng phương pháp IQR tìm ra cận trên và cận dưới để phân khác các hãng xe:

```
# Tính IQR
IQR <- IQR(average_growth_speed$average_growth_speed)

# Tính cận trên và cận dưới
upper <- quantile(average_growth_speed$average_growth_speed, 0.75) + 1.5 *
IQR
lower <- quantile(average_growth_speed$average_growth_speed, 0.25) - 1.5 *
IQR
```

Theo phương pháp trên, tôi có được cận trên bằng 47% và cận dưới bằng -26%. Tôi sử dụng kết quả này chia dữ liệu thành 3 mẫu:

- 5 hãng xe có tốc độ phát triển vượt bậc (tốc độ phát triển doanh thu cao nhất)
- 5 hãng xe có tốc độ phát triển cao (dưới 47%)
- 5 hãng xe có tốc độ phát triển thấp (trên -26%)

Kết quả phân chia như sau:

Tốc độ phát triển vượt bậc	"dacia", "smart", "mclaren", "ssangyong", "abarth"
Tốc độ phát triển cao	"mg", "bentley", "lamborghini", "cadillac", "corvette"
Tốc độ phát triển thấp	“ds”, “daewoo”, “daihatsu”, “dodge”, “maybach”

3.2.3. Sự tương quan giữa các biến liên tục

Để xác định mối liên hệ giữa các biến liên tục trong bộ dữ liệu, tôi sử dụng hệ số tương quan (Correlation Coefficient).

Hệ số tương quan (Correlation Coefficient) là chỉ số thống kê đo lường mức độ mạnh yếu của mối quan hệ giữa hai biến số. Công thức tính hệ số tương quan Pearson (một loại phổ biến của hệ số tương quan) được biểu diễn như sau:

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

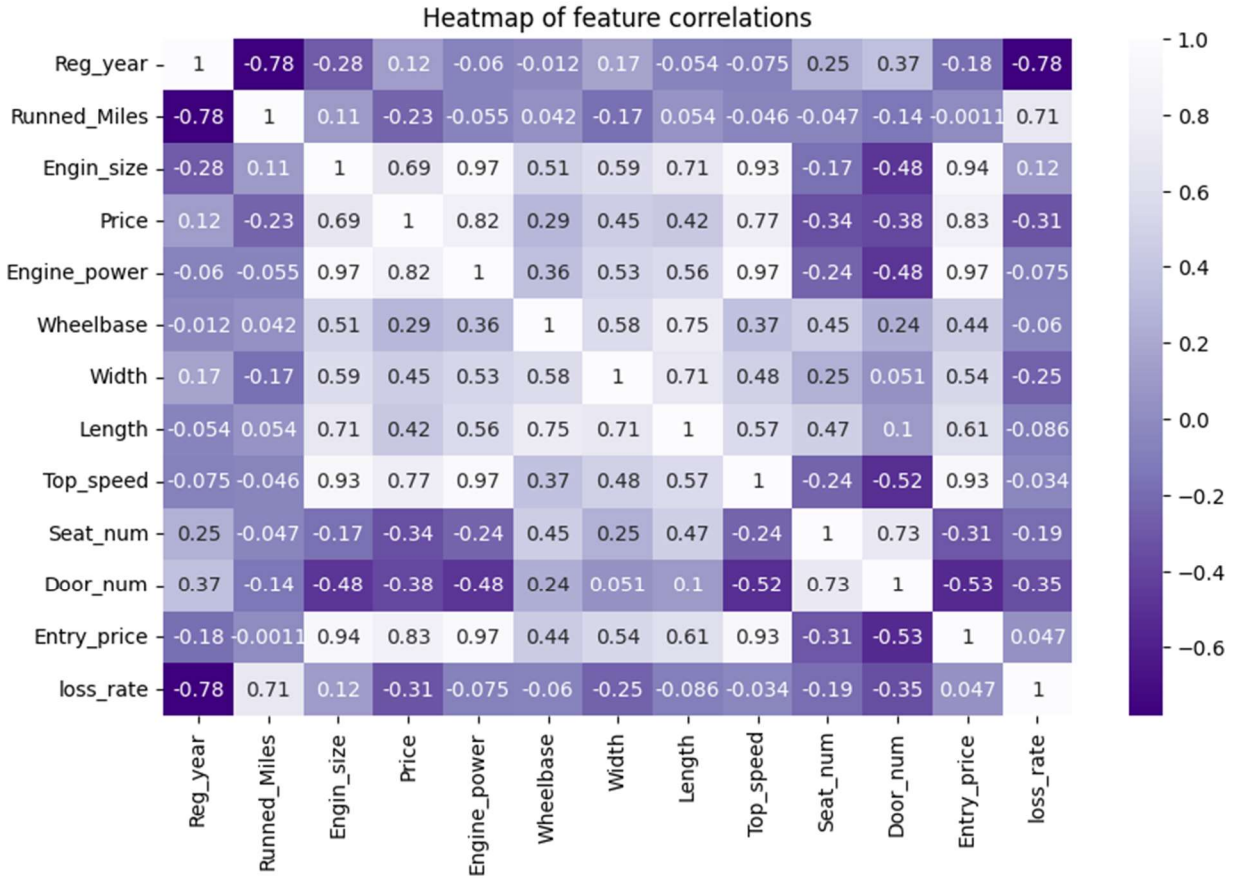
Trong đó:

ρxy: Hệ số tương quan Pearson.

Cov(x,y): Hiệp phương sai của biến x và y.

σx: Độ lệch chuẩn của x.

σy: Độ lệch chuẩn của y



Hình 3: Hệ số tương quan giữa các biến liên tục

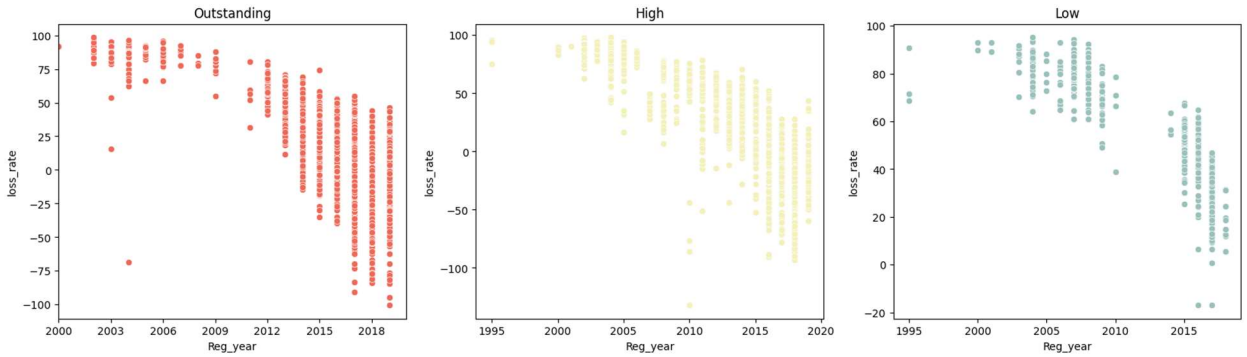
Hai yếu tố có tương quan mạnh mẽ nhất đến sự hao hụt về giá là năm đăng kí và số dặm đã qua sử dụng (Reg_year, Runned_Miles) với hệ số tương quan lần lượt là -0.78 và 0.71.

Các yếu tố ảnh hưởng đến giá bán lại của 1 chiếc xe gồm có kích thước động cơ, công suất động cơ, tốc độ tối đa và giá bán khởi điểm với hệ số tương quan lần lượt là 0.69, 0.82, 0.77 và 0.83.

Để làm rõ hơn về những mối tương quan này, tôi tiến hành phân tích cụ thể trên từng biến.

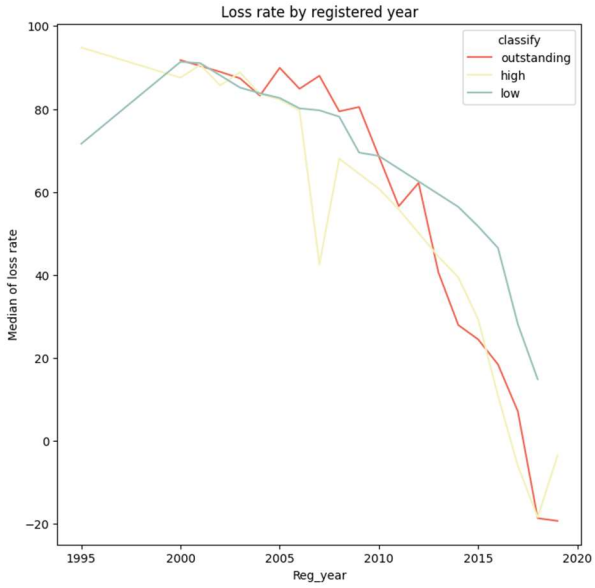
3.2.3.1 Năm đăng kí

Năm đăng kí của xe tỉ lệ nghịch với tỉ lệ hao hụt về giá (hệ số tương quan bằng -0.78), đồng nghĩa với việc những chiếc xe có năm đăng kí càng gần thì sẽ có khả năng bán lại với giá cao hơn.



Hình 4: Sự tương quan giữa năm đăng kí và tỉ lệ hao hụt về giá

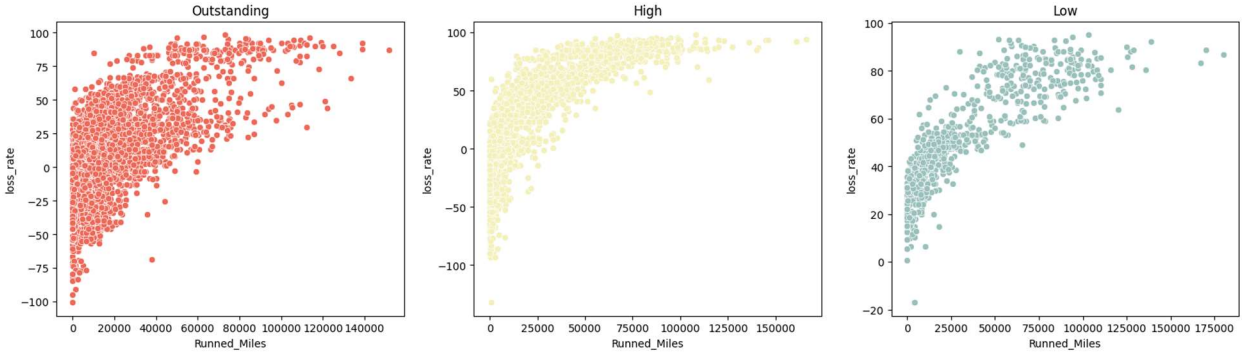
Những chiếc xe thuộc những hãng xe bán chạy với doanh thu vượt bậc có khả năng mang được đăng kí những năm 2017, 2018, 2019 có khả năng mang lại lợi nhuận cao. Đó cũng là xu hướng chung đối với hai phân khúc còn lại, tuy nhiên khả năng sinh lời của những hãng xe phát triển chậm chỉ nằm ở mức tối đa 20% với một tỉ lệ rất thấp.



Hình 5: Tỉ lệ mất giá trung bình theo năm đăng kí

Cụ thể, đối với những dòng xe được đăng kí từ năm 2010 trở về trước, nhóm phát triển vượt bậc có tỉ lệ mất giá cao nhất và hầu như không có dòng xe nào có thể mang lại lợi nhuận. Ngược lại, từ năm 2010 trở lại đây tỉ lệ mất giá của xe giảm dần, đặc biệt tỉ lệ này của xe thuộc những phân khúc phát triển và phát triển vượt bậc giảm đáng kể và chạm mức sinh lời ở những năm 2018, 2019.

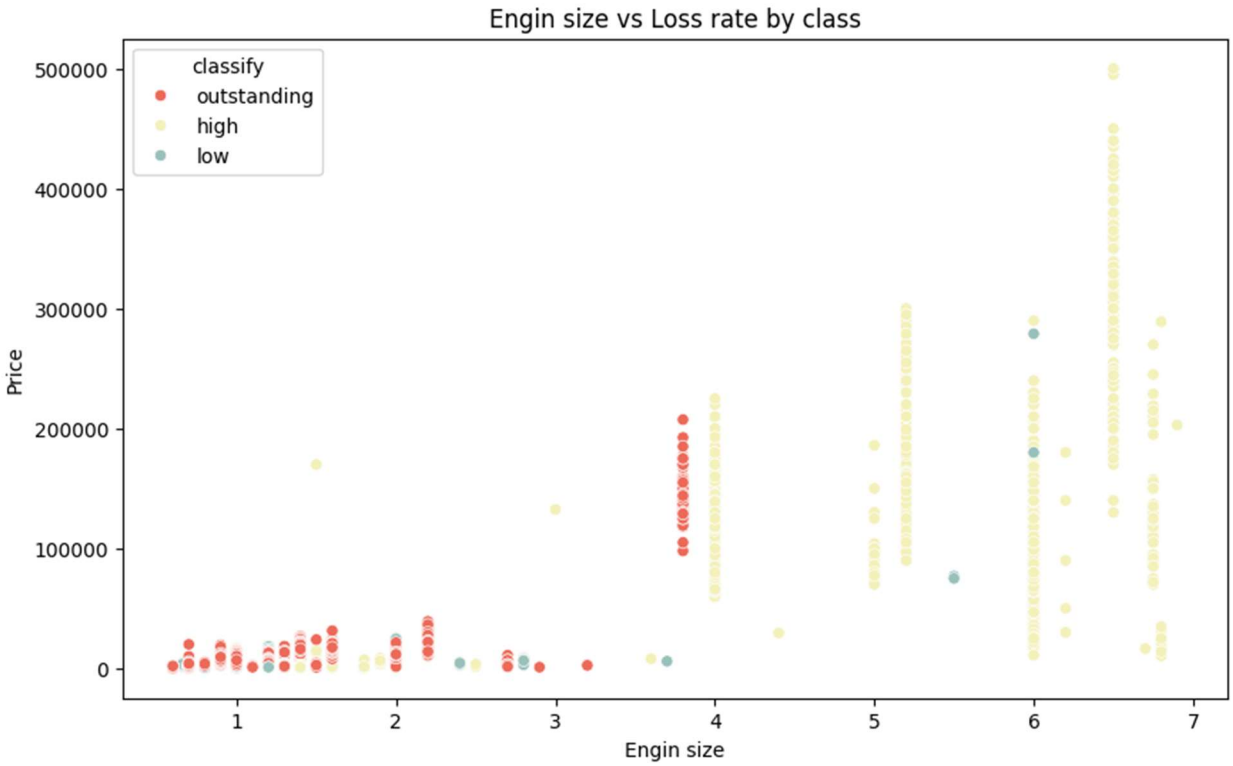
Ngược lại với năm đăng kí, quãng đường đã đi tỉ lệ thuận với tỉ lệ hao hụt (hệ số tương quan bằng 0.71), điều đó có nghĩa rằng một chiếc xe đã được sử dụng càng nhiều thì giá trị của nó sẽ càng giảm.



Hình 6: Sự tương quan giữa quãng đường đã đi và tỉ lệ mất giá

Đối với các hãng xe có tốc độ phát triển vượt bậc, những xe đã chạy dưới 5000 dặm đã có khả năng sinh lời. Những dòng xe thuộc phân khúc tốc độ phát triển cao có cơ hội sinh lời thấp hơn, những xe đã chạy khoảng dưới 3500 dặm mới có khả năng mang lại lợi nhuận. Ngược lại, hầu như không có xe nào thuộc phân khúc phát triển thấp có thể bán lại với giá cao hơn giá mua, thậm chí khi số dặm đã đi dưới 1000.

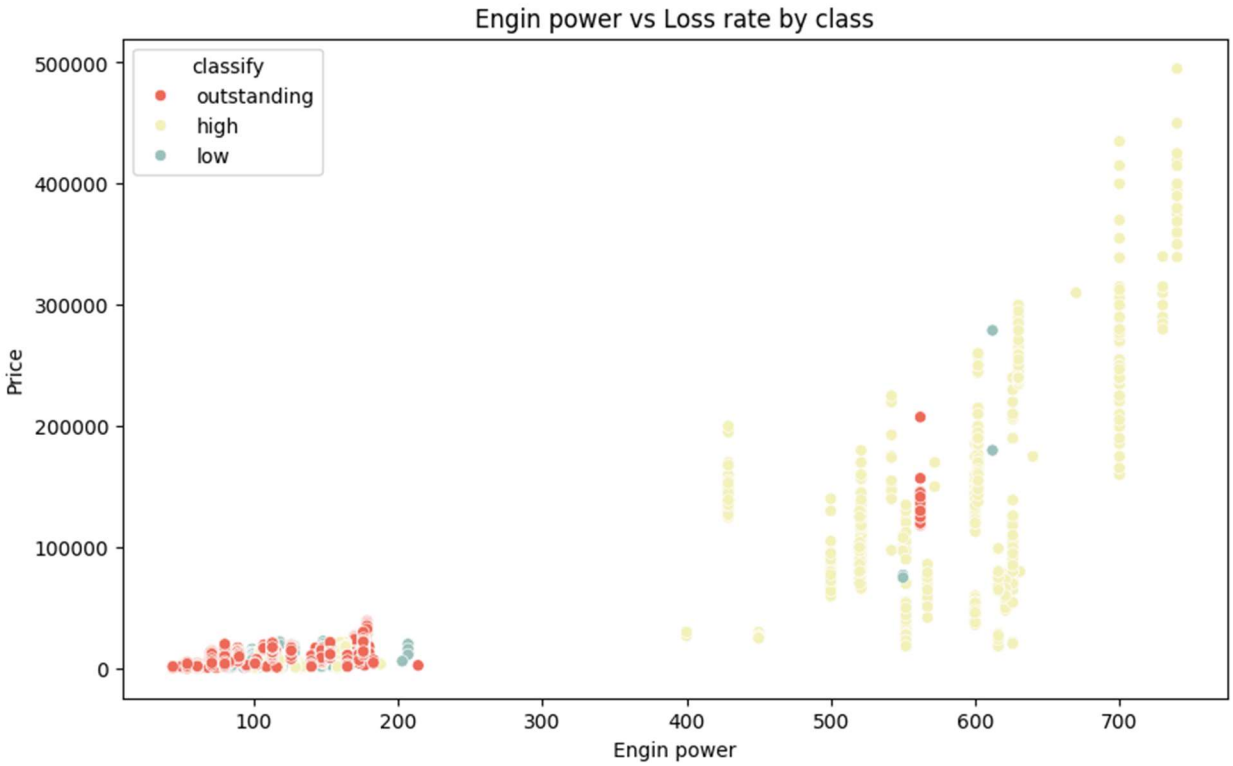
3.2.3.2 Kích thước động cơ



Hình 7: Sự tương quan giữa kích thước động cơ và giá

Kích thước động cơ của 2 phân khúc phát triển vượt bậc và phát triển cao cũng được phân thành 2 nhóm tương đối rõ rệt. Các dòng xe có tốc độ phát triển vượt bậc thường có kích thước động cơ dưới 4 lít trong khi hầu hết các xe có kích thước từ trên 4 lít đến 7 lít đều thuộc phân khúc tốc độ tăng trưởng cao. Trong các nhóm này, sự tương quan giữa giá và kích thước động cơ được thể hiện rõ ràng nhất ở phân khúc tăng trưởng vượt bậc khi xe có kích thước 4 lít có giá lớn hơn từ 2 đến 5 lần so với xe có kích thước động cơ dưới 3 lít.

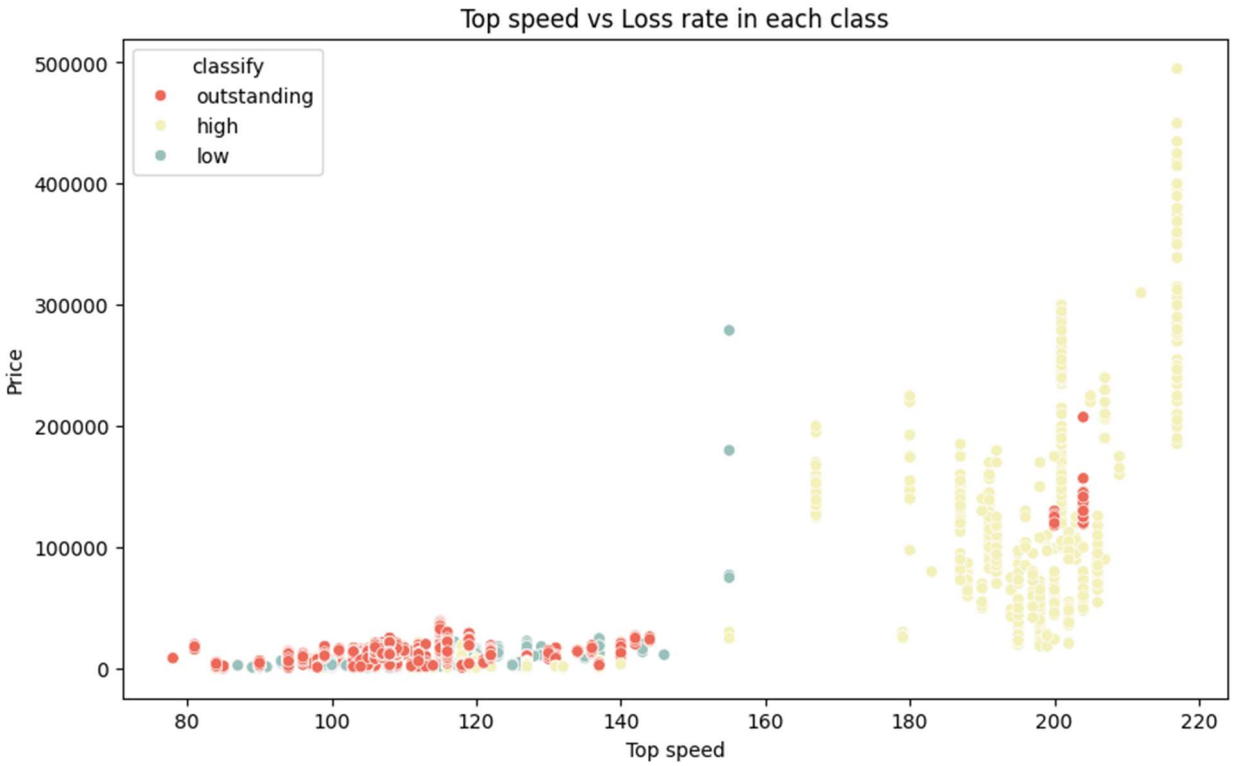
3.2.3.3 Công suất động cơ



Hình 8: Sự tương quan giữa công suất động cơ và giá

Tương tự như kích thước động cơ, công suất động cơ của 2 phân khúc phát triển vượt bậc và phát triển cao cũng được phân thành 2 nhóm tương đối rõ rệt. Tuy nhiên mối tương quan giữa công suất và giá được thể hiện rõ nhất ở phân khúc tốc độ phát triển cao, khi thông số công suất càng lớn thì giá xe bán lại càng tăng. Mối tương quan này của 2 phân khúc còn lại không được thể hiện rõ ràng.

3.2.3.4 Tốc độ tối đa



Hình 9: Sự tương quan giữa tốc độ tối đa và giá

Tốc độ tối đa cũng được phân cụm theo phân khúc các hãng xe với tốc độ từ 80 đến 150km/h thường thuộc phân khúc tốc độ phát triển vượt bậc, còn các loại xe có tốc độ tối đa cao – từ 180 đến 220km/h thường thuộc phân khúc tốc độ phát nhiển cao. Tuy nhiên mức độ tương quan không được thể hiện rõ nét ở cả 3 phân khúc, dựa vào biểu đồ phân tán chỉ có thể nhận định rằng các xe có tốc độ tối đa từ 80 đến 140km có giá bán lại rất thấp – dưới 50,000, còn các dòng xe có tốc độ tối đa trên 160km/h có giá bán cao hơn nhưng khoảng giá tương đối lớn – từ 40,000 đến 300,000. Riêng những dòng xe có tốc độ đạt đến 220km/h có giá bán lại rất cao, khởi điểm từ khoảng 200,000 và có thể lên đến 500,000.

3.2.4. Sự tương quan giữa các biến phân loại

Để xác định mối liên hệ giữa các biến phân loại trong bộ dữ liệu, tôi sử dụng phương pháp phân tích phương sai (kiểm định ANOVA).

Kiểm định ANOVA (Analysis of Variance) là một kỹ thuật thống kê tham số được sử dụng để phân tích sự khác nhau giữa giá trị trung bình của các biến phụ thuộc. Kiểm định ANOVA giúp xác định ảnh hưởng của các biến độc lập đối với biến phụ thuộc trong nghiên cứu hồi quy [(1)].

Cụ thể, kiểm định ANOVA có chức năng đánh giá sự khác biệt tiềm năng trong một biến phụ thuộc mức quy mô bằng một biến mức danh nghĩa có từ 2 loại trở lên. Các nhà phân tích sử dụng thử nghiệm ANOVA để xác định ảnh hưởng của các biến độc lập đối với biến phụ thuộc trong nghiên cứu hồi quy [(1)].

Kiểm định ANOVA bao gồm 3 phương pháp chính:

- ANOVA một chiều (One-way ANOVA): Đánh giá tác động của một biến độc lập duy nhất lên một biến phản hồi duy nhất.
- ANOVA hai chiều (Two-way ANOVA): Quan sát sự tương tác giữa hai yếu tố và kiểm tra sự ảnh hưởng của 2 yếu tố đó lên biến phụ thuộc cùng một lúc.
- ANOVA đa biến (MANOVA).

Trong bài phân tích này, tôi sử dụng ANOVA một chiều để lần lượt đánh giá tác động của các biến thành phần lên tỉ lệ mất giá của xe hơi.

```
categorical_vars <- c('Color', 'Bodytype', 'Gearbox', 'Fuel_type')

for (categorical_var in categorical_vars) {
  # Loại bỏ giá trị NA
  df_no_na <- na.omit(infor2[, c(categorical_var, 'loss_rate')])

  # Thực hiện kiểm định ANOVA một chiều
  result <- aov(as.formula(paste('loss_rate ~', categorical_var)), data =
df_no_na)

  # In kết quả
  print(paste('For variable', categorical_var, ':'))
  print(summary(result))
}
```

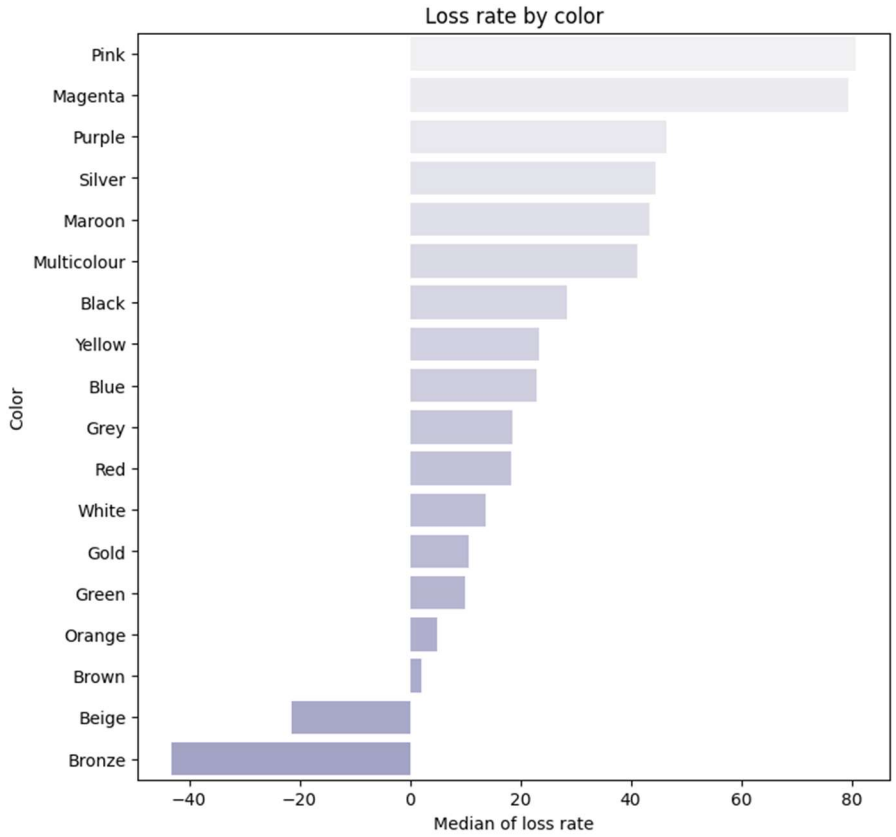
```
[1] "For variable Color :"
             Df  Sum Sq Mean Sq F value Pr(>F)
Color         18  571561   31753    22.3 <2e-16 ***
Residuals   6203 8830895    1424
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "For variable Bodytype :"
             Df  Sum Sq Mean Sq F value Pr(>F)
Bodytype      10 2097949  209795   178.4 <2e-16 ***
Residuals   6211 7304507    1176
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "For variable Gearbox :"
             Df  Sum Sq Mean Sq F value Pr(>F)
Gearbox       3  128855   42952    28.8 <2e-16 ***
Residuals   6218 9273601    1491
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "For variable Fuel_type :"
             Df  Sum Sq Mean Sq F value Pr(>F)
Fuel_type     6  399585   66597   45.98 <2e-16 ***
Residuals   6215 9002872    1449
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hình: Kết quả phân tích phương sai của các biến phân loại

Kết quả của kiểm định ANOVA trên các biến màu sắc, kiểu dáng xe, hộp số, loại nhiên liệu đều cho kết quả p-value rất bé và có ý nghĩa thống kê. Điều đó chứng tỏ các biến phân loại đều có ảnh hưởng đến tỉ lệ mất giá của xe hơi đã qua sử dụng.

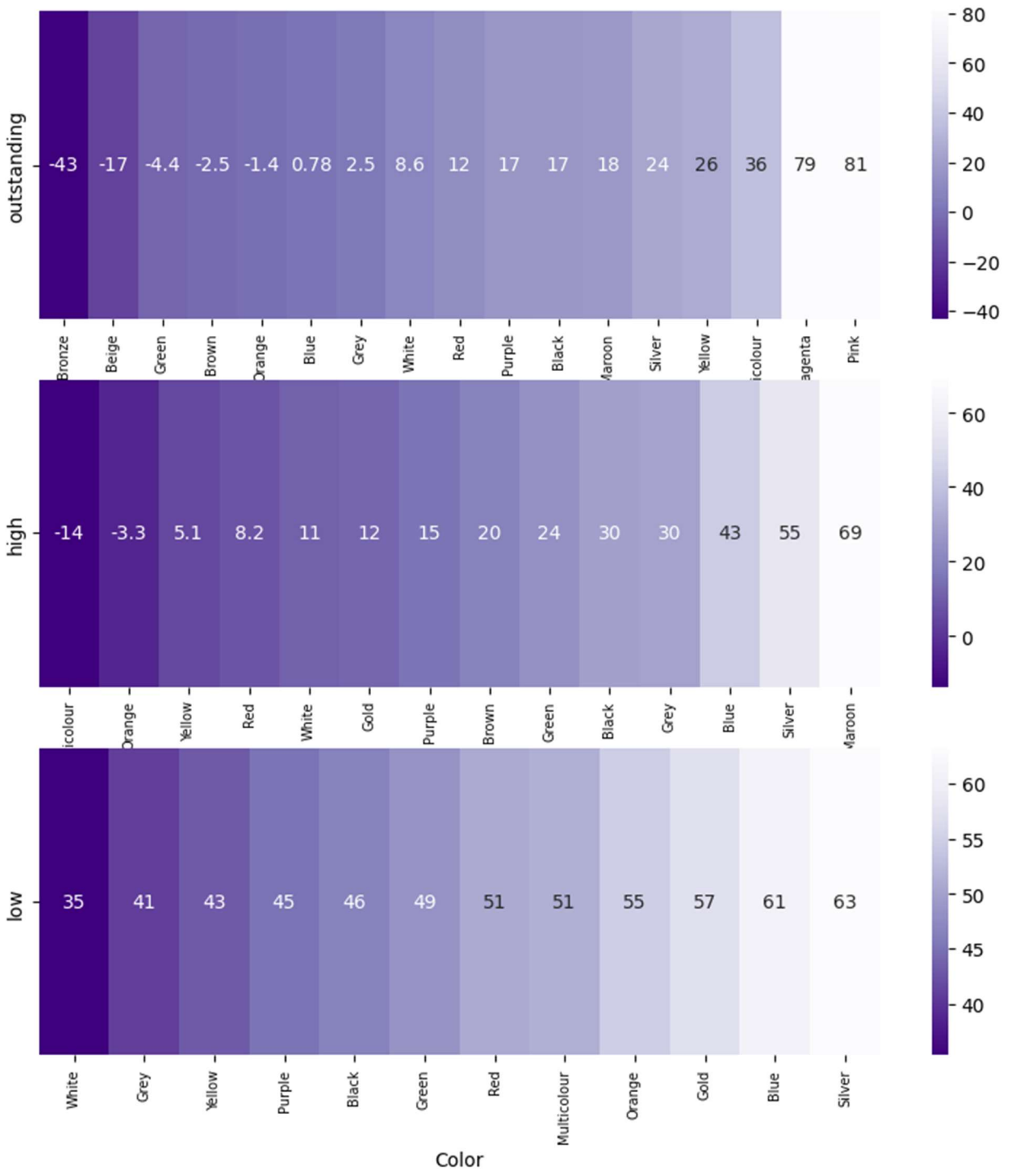
Tuy nhiên, phương pháp ANOVA chỉ có thể kiểm định giả thiết liệu các biến có tương quan với nhau hay không mà không thể kiểm định được mức độ tương quan giữa hai biến, vì vậy tôi sẽ tiến hành phân tích từng biến để làm rõ mối tương quan này.

3.2.4.1 Màu sắc



Hình 10: Trung vị tỉ lệ mất giá theo màu sắc xe

Dựa vào trung vị của tỉ lệ mất giá, có thể thấy rằng màu đồng và màu be là hai màu sắc xe được ưa chuộng nhất và chỉ hai màu này là mang lại giá bán cao hơn giá trị ban đầu của xe. Trong khi đó, những màu sắc mang tính nữ tính và tươi sáng như tím, hồng lại có tỉ lệ mất giá rất cao. Các mày sắc phổ biến như vàng đồng, đen, trắng có tỉ lệ mất giá khoảng 20%.

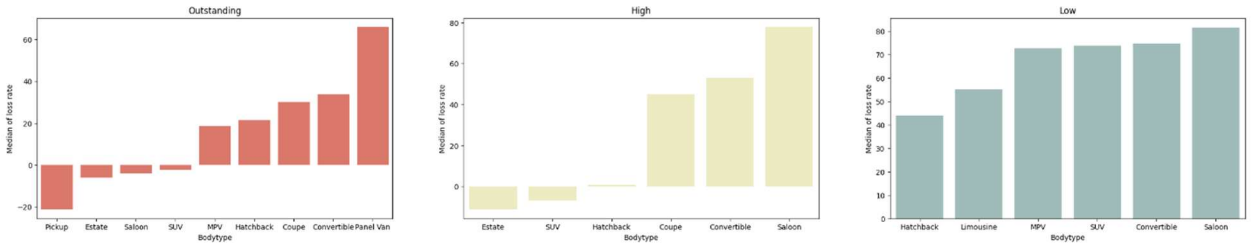


Hình 11: Trung vị tỉ lệ mất giá của màu sắc theo phân khúc xe

Cụ thể hơn về mối tương quan, ta thấy rằng các dòng xe ở phân khúc tốc độ phát triển vượt bậc có tương quan tương đối rõ ràng theo màu sắc khi những màu sắc thuộc gam nóng và

ánh cam thường có xu hướng có thể mang lại lợi nhuận, các màu phổ biến như trắng, đỏ, đen có tỉ lệ mất giá tương đối thấp (dưới 20%) và những màu nổi bật mang tính cá nhân hoá như hồng, vàng hay phối hợp nhiều màu lại có tỉ lệ mất giá cao. Đối với phân khúc tốc độ phát triển cao, mối tương quan này thể hiện qua gam màu nóng và gam màu lạnh với các gam màu nóng mang lại tỉ lệ mất giá thấp hơn (dưới 20%) và ngược lại đối với các gam màu lạnh. Đối với phân khúc tốc độ phát triển chậm thì hầu như sự biến đổi về màu sắc không có tác động đáng kể đến tỉ lệ mất giá.

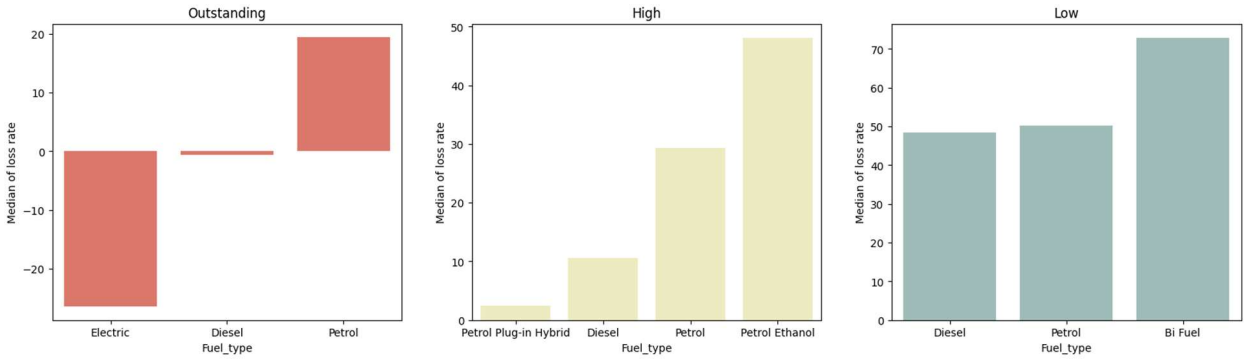
3.2.4.2 Kiểu dáng xe



Hình 12: Trung vị tỉ lệ mất giá của kiểu dáng xe theo phân khúc xe

Nhìn chung, có nhiều điểm tương đồng giữa phân khúc tốc độ phát triển vượt bậc và cao. “Estate” và “SUV” là những dáng xe được ưa chuộng ở 2 phân khúc này, “Hatback” lại là sự lựa chọn tương đối an toàn ở cả 3 phân khúc. Riêng phân khúc tốc độ phát triển vượt bậc có dáng xe “Pick-up” là dòng xe bán tải lại là dòng xe mang lại lợi nhuận lớn nhất, ngoài ra, khác với phân khúc tốc độ phát triển cao, “Saloon” cũng là một dáng xe được ưa chuộng của phân khúc này tuy lợi nhuận mang lại không quá lớn. Từ đó ta có thể thấy rằng những dòng xe giá rẻ và có khoang hành lí tách biệt thường được bán lại với giá trị cao hơn so với giá mua.

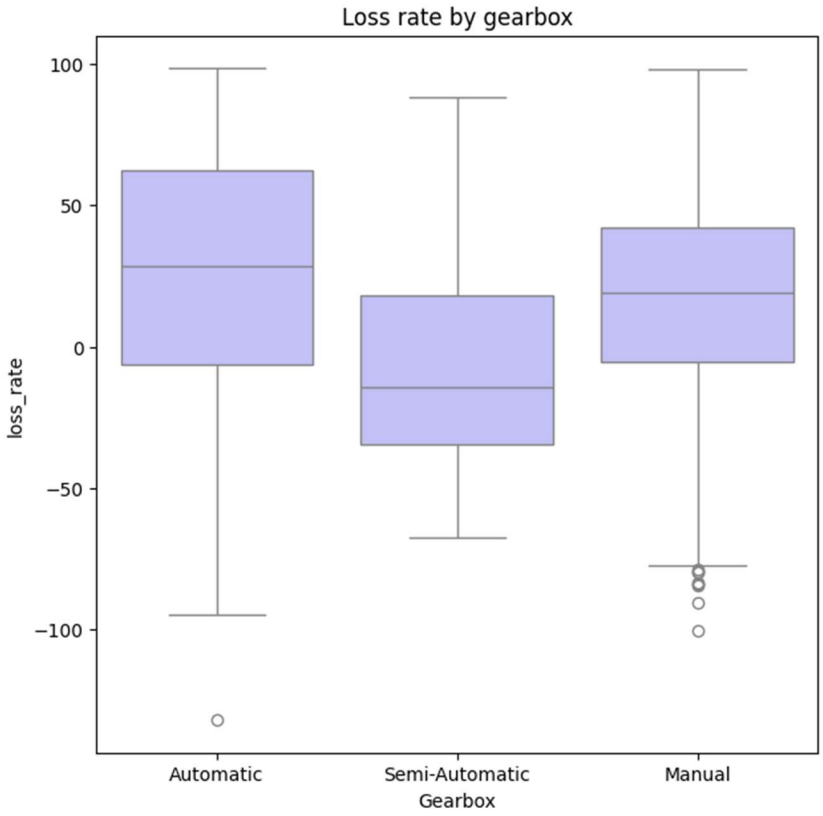
3.2.4.3 Loại nhiên liệu



Hình 13: Trung vị tỉ lệ mất giá của loại nhiên liệu theo phân khúc xe

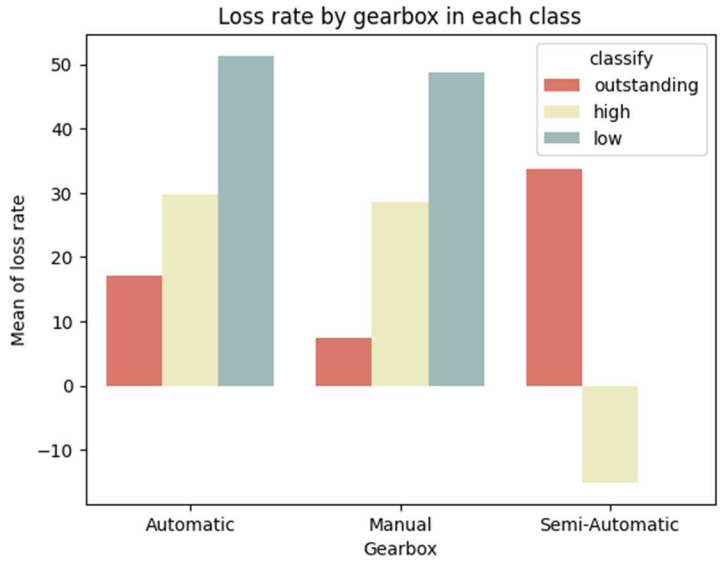
Thông qua phân tích loại nhiên liệu, có thể thấy rằng xe điện dần đang trở thành xu hướng khi nó là thể loại duy nhất mang lại lợi nhuận cao (trên 20%), kể cả các loại xe điện lai (Petrol Plig-in Hybrid) cũng có tỉ lệ mất giá thấp hơn đáng kể so với các dòng xe còn lại. Riêng các dòng xe thuộc phân khúc tốc độ phát triển chậm chạy bằng nhiên liệu tự nhiên có tỉ lệ trượt giá rất lớn (hơn 70%), điều này có thể bởi vì sự đắt đỏ của nhiên liệu tự nhiên khiến nó không trở thành lựa chọn ưu tiên của khách hàng.

3.2.4.4 Hộp số



Hình 14: Sự phân bố của tỉ lệ mất giá theo các loại hộp số

Từ biểu đồ hộp, ta thấy rằng không có nhiều sự khác biệt giữa sự phân bố dữ liệu khi phân chia theo hộp số.



Hình 15: Trung vị tỉ lệ mất giá của các loại hộp số theo phân khúc xe

Cụ thể hơn, đối với phân khúc tốc độ phát triển vượt bậc, xe số sàn là loại xe có tỉ lệ mất giá thấp nhất trong khi xe bán tự động lại có tỉ lệ mất giá tương đối cao (khoảng 35%). Đối với phân khúc tốc độ phát triển cao, xe bán tự động là dòng xe mang lại giá trị cao nhất (lợi nhuận trên 10%) trong khi hai thể loại còn lại có sự chênh lệch không đáng kể (tỉ lệ mất giá khoảng 30%). Phân khúc tốc độ phát triển chậm cũng không có sự khác biệt đáng kể, điều đó cho thấy hộp số có thể không phải là mối quan tâm lớn của khách hàng trong quyết định mua xe.

3.3. Xây dựng mô hình dự báo giá xe hơi

3.3.1. Mô hình thử nghiệm

Để dự báo giá xe hơi dựa trên hãng xe và các thông số của xe, tôi lựa chọn mô hình Linear Regression.

3.3.1.1 Lựa chọn tính năng

Dựa trên kết quả hệ số tương quan và kiểm định ANOVA, ban đầu tôi sử dụng các biến nhà sản xuất, dòng xe, màu sắc, kiểu dáng, loại nhiên liệu, kích thước động cơ, công

suất động cơ, tốc độ tối đa và giá mua ban đầu với vai trò là đặc trưng đầu vào cho mô hình.

```
features <- c('Maker', 'Genmodel', 'Genmodel_ID', 'Color', 'Bodytype',
'Fuel_type', 'Engin_size', 'Engine_power', 'Top_speed', 'Entry_price')
X <- infor[features]
y <- infor['Price']
```

3.3.1.2 Mã hoá các biến phân loại

Trước khi đưa các tính năng vào mô hình máy học, ta cần mã hoá các biến phân loại thành dạng số. Cụ thể, trong mô hình này cần mã hoá các biến nhà sản xuất, dòng xe, màu sắc, kiểu dáng, loại nhiên liệu.

```
# Chọn các categorical features
categorical_features <- c('Maker', 'Genmodel', 'Color', 'Bodytype',
'Fuel_type')

# Chuyển đổi các categorical features thành dạng số
X_cat <- X[categorical_features] %>%
  mutate(across(everything(), as_factor)) %>%
  mutate(across(everything(), as.numeric))
```

3.3.1.3 Chuẩn hoá dữ liệu

Để tránh được sự chi phối của các đặc trưng có giá trị lớn hơn so với các đặc trưng hoặc biến khác và cân bằng tỷ lệ của các đặc trưng, tôi tiến hành chuẩn hoá dữ liệu để biến đổi khoảng giá trị thành phạm vi từ 0 đến 1 bằng phương pháp Min-Max-Scaler.

Công thức biến đổi của Min-Max-Scaler là:

$$Xstd=Xmax−XminX−Xmin$$

$$Xscaled=Xstd×(max−min)+min$$

Trong đó min, max là phạm vi mong muốn của dữ liệu sau khi biến đổi.

Min-Max-Scaler không giảm bớt tác động của các giá trị ngoại lai, nhưng nó co giãn chúng theo cách tuyến tính xuống một phạm vi cố định, nơi điểm dữ liệu lớn nhất tương ứng với giá trị tối đa và điểm dữ liệu nhỏ nhất tương ứng với giá trị tối thiểu.

```
# Tạo một bản sao của dữ liệu để không làm thay đổi dữ liệu gốc
df_normalized <- X

# Danh sách các cột cần chuẩn hóa
columns_to_normalize <- colnames(X)

# Thực hiện chuẩn hóa Min-Max cho mỗi cột
df_normalized <- df_normalized %>%
  mutate(across(all_of(columns_to_normalize), ~rescale(.x, to = c(0, 1))))
```

3.3.1.4 Xây dựng và kiểm định mô hình

Để xây dựng và kiểm định mô hình, tôi tiến hành phân chia bộ dữ liệu thành 2 tập – tập huấn luyện và tập kiểm tra với tỉ lệ 8:2. Tập huấn luyện được sử dụng để xây dựng mô hình học máy và tập kiểm tra được dùng để dự báo và so sánh kết quả dự báo với kết quả thực tế để đánh giá mô hình.

```
# Chia tập dữ liệu thành train và test
set.seed(41)
split <- sample.split(df_normalized, SplitRatio = 0.8)
X_train <- subset(df_normalized, split == TRUE)
X_test <- subset(df_normalized, split == FALSE)
y_train <- subset(y, split == TRUE)
y_test <- subset(y, split == FALSE)

# Chuyển 'y_train' và 'y_test' thành dataframe
y_train <- data.frame(Price = y_train)
y_test <- data.frame(Price = y_test)

# Xây dựng mô hình logistic regression
lr <- lm(Price ~ ., data = cbind(X_train, y_train))

# Dự đoán giá trên tập test
y_pred <- predict(lr, newdata = cbind(X_test, y_test))
```

```
Call:
lm(formula = Price ~ ., data = cbind(X_train, y_train))

Residuals:
    Min      1Q  Median      3Q     Max
-122394   -3604       4    3550  903628

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)    24701.5      312.1   79.153  < 2e-16 ***
Maker           5104.1     1301.4    3.922 8.79e-05 ***
Genmodel        7371.2     3825.3    1.927  0.05399 .
Genmodel_ID   -13541.9     4242.3   -3.192  0.00141 **
Color           2810.3      179.8   15.633  < 2e-16 ***
Bodytype        2510.6      229.8   10.926  < 2e-16 ***
Fuel_type      14416.0      306.0   47.112  < 2e-16 ***
Engin_size   -125878.8      574.2 -219.224  < 2e-16 ***
Engine_power  188362.1      851.5  221.214  < 2e-16 ***
Top_speed     -49976.0      678.4  -73.667  < 2e-16 ***
Entry_price   182219.4      878.6  207.403  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10570 on 151509 degrees of freedom
Multiple R-squared:  0.692,     Adjusted R-squared:  0.692
F-statistic: 3.404e+04 on 10 and 151509 DF,  p-value: < 2.2e-16
```

Qua kết quả kiểm định mô hình, R bình phương hiệu chỉnh (Ajusted R-squared) bằng 69%, cho thấy các biến độc lập giải thích được 69% sự biến thiên của biến phụ thuộc (giá xe). Kết quả này cho thấy mô hình thật sự đáng tin cậy.

Để đánh giá kết quả dự báo, tôi sử dụng 2 thước đo là trung bình trị tuyệt đối sai số (MAE) và lỗi bình phương trung bình gốc (RMSE).

```
[1] "Mean Absolute Error: 5747.11119129467"
[1] "Mean Squared Error: 100795896.761965"
[1] "Root Mean Squared Error: 10039.7159701839"
```

Kết quả cho thấy mô hình có MAE = 5747 và RMSE = 10039, so với giá trị trung bình của biến phụ thuộc là 12500 thì kết quả trên cho thấy kết quả dự báo từ mô hình đang thiếu tính ổn định và chính xác, điều này cho thấy điểm bất lợi của mô hình này trong việc dự báo giá xe. Điều này có thể xảy ra bởi sự ảnh hưởng của các đặc trưng đầu vào, vì vậy tôi tiến hành quay lại bước lựa chọn đặc trưng và thực hiện xây dựng lại mô hình hồi quy tuyến tính.

3.3.2. Mô hình lựa chọn

3.3.2.1 Lựa chọn tính năng

Để xác định lại các đặc trưng phù hợp với mô hình, tôi sử dụng công cụ lựa chọn đặc trưng tự động thuộc gói Caret R để xác định được các đặc trưng quan trọng với mô hình. Qua đó, công cụ đã lựa chọn các biến độc lập sau: 'Maker', 'Genmodel', 'Runned_Miles', 'Engin_size', 'Engine_power', 'Height','Length', 'Top_speed', 'Entry_price', 'loss_rate'.

3.3.2.2 Mã hoá các biến phân loại

Tương tự như mô hình ban đầu, ta cần mã hoá các biến phân loại thành dạng số. Cụ thể, trong mô hình này cần mã hoá các biến nhà sản xuất và dòng xe.

3.3.2.3 Chuẩn hoá dữ liệu

Cũng tương tự như mô hình thử nghiệm, để tránh sự chi phối của các đặc trưng có giá trị lớn hơn so với các đặc trưng hoặc biến khác và cân bằng tỷ lệ của các đặc trưng, tôi tiến hành chuẩn hoá dữ liệu để biến đổi khoảng giá trị thành phạm vi từ 0 đến 1 bằng phương pháp Min-Max-Scaler.

3.3.2.4 Xây dựng và kiểm định mô hình

Mô hình này cũng phân chia bộ dữ liệu thành 2 tập – tập huấn luyện và tập kiểm tra với tỉ lệ 8:2. Tập huấn luyện được sử dụng để xây dựng mô hình học máy và tập kiểm tra được dùng để dự báo và so sánh kết quả dự báo với kết quả thực tế để đánh giá mô hình.

```
# Chia tập dữ liệu thành train và test
set.seed(41)
split <- sample.split(df_normalized, SplitRatio = 0.8)
X_train <- subset(df_normalized, split == TRUE)
X_test <- subset(df_normalized, split == FALSE)
y_train <- subset(y, split == TRUE)
y_test <- subset(y, split == FALSE)

# Chuyển 'y_train' và 'y_test' thành dataframe
y_train <- data.frame(Price = y_train)
y_test <- data.frame(Price = y_test)

# Xây dựng mô hình logistic regression
lr <- lm(Price ~ ., data = cbind(X_train, y_train))

# Dự đoán giá trên tập test
y_pred <- predict(lr, newdata = cbind(X_test, y_test))
```

```
Call:
lm(formula = Price ~ ., data = cbind(X_train, y_train))

Residuals:
    Min      1Q  Median      3Q     Max
-124928   -2351     641    2943  648796

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   309422.9     1010.3  306.26   <2e-16 ***
Maker         -10552.2      840.7  -12.55   <2e-16 ***
Genmodel        9186.9      861.1   10.67   <2e-16 ***
Runned_Miles  376743.8     5108.4   73.75   <2e-16 ***
Engin_size    -46662.4      474.3  -98.39   <2e-16 ***
Engine_power   71202.2      699.1  101.84   <2e-16 ***
Height        -32426.4      537.1  -60.37   <2e-16 ***
Length         -8084.3      278.2  -29.06   <2e-16 ***
Top_speed     -36102.1      628.9  -57.41   <2e-16 ***
Entry_price   220285.5      643.9  342.10   <2e-16 ***
loss_rate    -286315.4      943.5 -303.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7610 on 151509 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8405
F-statistic: 7.983e+04 on 10 and 151509 DF,  p-value: < 2.2e-16
```

Qua kết quả kiểm định mô hình, R bình phương hiệu chỉnh (Ajusted R-squared) bằng 84%, tốt hơn rất nhiều so với mô hình thử nghiệm (64%). Điều đó chứng tỏ các đặc trưng được lựa chọn cho mô hình này mang lại hiệu quả tích cực.

```
[1] "Mean Absolute Error: 3979.35862369157"
[1] "Mean Squared Error: 56861560.957378"
[1] "Root Mean Squared Error: 7540.66051200941"
```

Kết quả cho thấy mô hình có MAE và RMSE thấp hơn so với mô hình thử nghiệm. Mức MAE khoảng 3900 cho thấy kết quả sai số của dự báo nằm trong phạm vi chấp nhận được. RMSE thấp hơn đáng kể cso với mô hình thử nghiệm chứng tỏ mô hình sau khi được lựa chọn lại đặc trưng có khả năng dự báo ổn định hơn. Điều đó chứng tỏ đây là mô hình phù hợp và ổn định cho bộ dữ liệu này.

CHƯƠNG 4: KẾT LUẬN VÀ KIẾN NGHỊ

4.1. Kết luận

Thông qua phân tích dữ liệu về thị trường xe hơi đã qua sử dụng tại Vương Quốc Anh, có thể thấy rằng những hãng xe bình dân như "dacia", "smart", "ssangyong", "abarth" lại là những hãng xe được ưa chuộng nhất tại thị trường này. Giá cả của những chiếc xe này thấp và thường là những chiếc xe không có thông số ấn tượng (dung tích động cơ và công suất động cơ thấp). Tuy nhiên, những chiếc xe thuộc phân khúc này có thể bán lại với giá trị cao hơn giá trị mua lúc ban đầu, đặc biệt là các dáng xe có khoang chưa hành lí rộng đến rất rộng (như xe bán tải) và các dòng xe điện, điều đó mở ra cơ hội sinh lời cho người bán nói riêng và các đơn vị kinh doanh xe hơi đã qua sử dụng nói chung.

Ngoài ra, các hãng xe có mức độ tăng trưởng ổn định và vừa phải thường là những hãng xe xa xỉ với giá thành cao hơn rất nhiều so với phân khúc các dòng xe có tốc độ tăng trưởng vượt bậc. Các dòng xe này thường có công suất động cơ và kích thước động cơ lớn, giá cả của xe bán ra cũng bị ảnh hưởng rất lớn bởi hai thông số này. Dòng xe bán tự động và các xe có màu sắc thuộc gam màu nóng của phân khúc này rất được ưa chuộng.

Đối với các xe có tốc độ tăng trưởng chậm (tỉ lệ tăng trưởng âm), không có quá nhiều đặc điểm riêng biệt của nhóm xe này, điều này có thể bởi vì các hãng xe này không còn phổ biến trên thị trường nên dẫn đến việc chênh lệch về mẫu dữ liệu. Thông qua các phân tích về thông số và kiểu dáng xe có thể thấy, tỉ mất giá của các dòng xe này luôn nằm ở mức cao đến rất cao (có thể lên đến 100%) và tỉ lệ mất giá không phụ thuộc nhiều vào các thông số đã phân tích, điều này có thể được giải thích bởi yếu tố thương hiệu quyết định đến giá xe nhiều hơn trong trường hợp này.

4.2. Kiến nghị

Thông qua mô hình hồi quy tuyến tính dự báo giá xe, doanh nghiệp kinh doanh xe hơi đã qua sử dụng có thể ước tính được giá mua vào để tối ưu hoá lợi nhuận. Việc ước lượng giá xe bán ra phụ thuộc vào hãng xe, số quãng đường đã đi, các thông số về động cơ, kích thước, tốc độ.

Tuy nhiên, vẫn có những sai số từ mô hình học máy có thể ảnh hưởng đến kết quả dự báo giá xe, vì vậy cần cân nhắc đến các yếu tố khác có thể ảnh hưởng đến giá xe đã được đề cập trong bài phân tích.

TÀI LIỆU THAM KHẢO

(1) Kiểm Định Anova Trong Spss: Khái Niệm, Phân Loại, Cách Chạy (trithuccongdong.net)

(2) Kiểm Định Anova Trong Spss: Khái Niệm, Phân Loại, Cách Chạy (trithuccongdong.net)