ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC KINH TẾ
KHOA THƯƠNG MẠI ĐIỆN TỬ
*****

BÁO CÁO
ĐỀ ÁN THỰC HÀNH 2
Đề tài:

ỨNG DỤNG MÔ HÌNH XGB REGRESSOR NHẰM ĐỊNH GIÁ BẢO HIỂM XE Ô TÔ

Giảng viên hướng dẫn : *ThS. Nguyễn Văn Chức*
Lớp học phần : *ELC3011_46K29.1/46K29.2*
Nhóm : *05*
Thành viên nhóm:

1. *Nguyễn Thị Thu Huyền*
2. *Nguyễn Văn Linh*
3. *Lê Thị Hồng Ny*

Đà Nẵng, ngày 28 tháng 05 năm 2023

PHẦN TRĂM ĐÓNG GÓP

Bảng 1: Phần trăm đóng góp

STT	Họ và tên	Lớp	Nhiệm vụ	Phần trăm đóng góp
1	Nguyễn Thị Thu Huyền	46K29.1	Lý thuyết, tìm dữ liệu, phân tích dữ liệu, xây dựng mô hình	35%
2	Nguyễn Văn Linh	46K29.1	Lý thuyết, tìm dữ liệu, phân tích dữ liệu, xây dựng ứng dụng	35%
3	Lê Thị Hồng Ny	46K29.2	Lý thuyết, phân tích dữ liệu	30%

MỤC LỤC

MỤC LỤC BẢNG

MỤC LỤC HÌNH ẢNH

LỜI CẢM ƠN

"Đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến Khoa Thương Mại Điện Tử đã đưa môn học Đề án thực hành 2 vào chương trình giảng dạy. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến giảng viên hướng dẫn - Thầy Nguyễn Văn Chức đã truyền đạt những kiến thức quý báu cho em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia nghiên cứu, chúng em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để chúng em có thể vững bước sau này.

Đề án thực hành 2 là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Tuy nhiên, do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ. Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài báo cáo khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong thầy/cô xem xét và góp ý để bài báo cáo của chúng em được hoàn thiện hơn.

Xin chân thành cảm ơn!"

LỜI MỞ ĐẦU

Hiện nay, trong bối cảnh nền kinh tế thế giới và trong nước còn nhiều diễn biến khó lường, bất ổn do đại dịch Covid-19, thị trường bảo hiểm Việt Nam vẫn duy trì đà tăng trưởng và ngày càng thể hiện được vai trò, vị trí quan trọng đối với nền kinh tế – xã hội của đất nước. Bảo hiểm đang ngày càng khẳng định vai trò quan trọng. Những năm gần đây thì quy mô thị trường bảo hiểm ngày càng lớn, góp phần ổn định kinh tế vĩ mô và đã góp phần bổ trợ cho các chính sách an sinh xã hội.

Định giá bảo hiểm xe ô tô là một vấn đề quan trọng đối với các chủ xe ô tô, bởi nó ảnh hưởng đến chi phí bảo hiểm hàng năm. Tuy nhiên, việc định giá này không chỉ đơn giản là tính toán giá trị của chiếc xe, mà còn phải xem xét nhiều yếu tố khác nhau.

Một trong những yếu tố quan trọng nhất là mức độ rủi ro của chiếc xe. Những chiếc xe có nguy cơ gây tai nạn hoặc bị mất cắp cao hơn sẽ có mức phí bảo hiểm cao hơn. Các yếu tố khác cũng có thể ảnh hưởng đến giá bảo hiểm, chẳng hạn như tuổi của chủ xe, kinh nghiệm lái xe, vị trí đỗ xe, loại xe và mục đích sử dụng của xe.

Để định giá bảo hiểm xe ô tô một cách chính xác, các công ty bảo hiểm thường sử dụng các mô hình phân tích rủi ro và thu thập thông tin chi tiết về các yếu tố ảnh hưởng đến giá bảo hiểm. Từ đó, họ có thể tính toán được mức phí bảo hiểm phù hợp với mức độ rủi ro của chiếc xe và đảm bảo sự bảo vệ tài sản của chủ xe.

PHẦN 1. GIỚI THIỆU TỔNG QUAN

1.1. Lý do chọn đề tài

Nền kinh tế Việt Nam đang trong công cuộc đổi mới, hội nhập vì vậy muốn phát triển đất nước cần có sự đóng góp của tất cả các ngành, các lĩnh vực. Góp phần bảo đảm an toàn, ổn định tài chính cho các cá nhân, gia đình và mọi tổ chức doanh nghiệp giúp khôi phục đời sống và hoạt động sản xuất kinh doanh, đồng thời đóng góp vai trò trong việc huy động các nguồn lực tài chính đáp ứng nhu cầu vốn đầu tư dài hạn của nền kinh tế, bảo hiểm ngày càng chứng tỏ được vị trí quan trọng trong nền kinh tế quốc dân. Thực tế cho thấy, nền kinh tế càng phát triển, đời sống càng cao thì nhu cầu của con người càng phong phú và đang dạng trong đó có nhu cầu bảo đảm an toàn - an tâm - ổn định cuộc sống. Vì vậy ngày nay bảo hiểm đã đi vào cuộc sống của từng cá nhân, từng hộ gia đình và doanh nghiệp qua đó cũng cho thấy sự phát triển lớn mạnh.

Bảo hiểm xe ô tô là một vấn đề quan trọng đối với các chủ xe ô tô, là một hợp đồng giữa người sở hữu xe ô tô và công ty bảo hiểm cam kết chi trả một khoản tiền đền bù cho người sở hữu xe trong trường hợp xảy ra mất mát, hư hỏng hoặc thương tích liên quan đến xe ô tô. Vì vậy, việc định giá bảo hiểm ô tô là rất quan trọng giúp hiểu rõ về cách các công ty bảo hiểm xác định giá trị bảo hiểm và đóng phí, từ đó tăng khả năng hiểu và đưa ra quyết định thông minh khi mua bảo hiểm. Bên cạnh đó còn giúp xác định mức chi phí mà khách hàng phải trả hàng tháng tương ứng với gói bảo hiểm ô tô phù hợp dựa vào các yếu tố quan trọng như độ tuổi của người lái, loại và tuổi của xe, vị trí địa lý, lịch sử lái xe và các yếu tố khác.

Để việc định giá bảo hiểm xe ô tô một cách chính xác, các công ty bảo hiểm phải thu thập thông tin chi tiết về khách hàng, ô tô sở hữu và sử dụng các mô hình phân tích rủi ro và các yếu tố ảnh hưởng đến giá bảo hiểm. Từ đó, họ có thể tính toán được mức phí bảo hiểm phù hợp với mức độ rủi ro của chiếc xe và đảm bảo sự bảo vệ tài sản và sức khỏe của chủ xe.

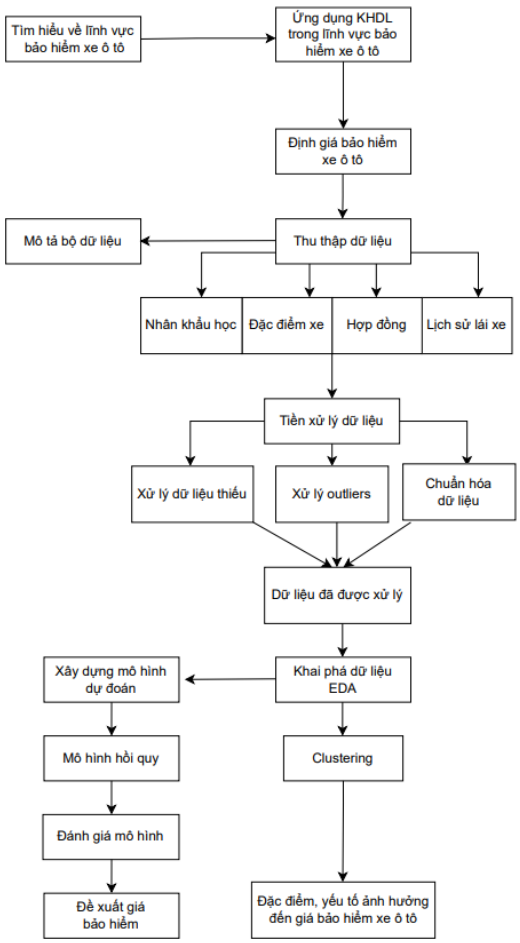
1.2. Mục tiêu nghiên cứu

Mục tiêu chính của nghiên cứu là xác định các yếu tố ảnh hưởng đến việc định giá bảo hiểm ô tô. Các yếu tố này có thể bao gồm tuổi và kinh nghiệm lái xe của chủ sở hữu, loại xe, lịch sử lái xe của người sử dụng, cũng như các yếu tố địa lý như địa điểm. Bằng cách nghiên cứu và phân tích những yếu tố này, chúng tôi hy vọng có thể xác minh được cách thức ảnh hưởng của các yếu tố trên đến việc định giá bảo hiểm ô tô.

Nghiên cứu cũng sẽ đưa ra một cái nhìn tổng quan về quy trình định giá bảo hiểm ô tô. Điều này bao gồm việc tìm hiểu các phương pháp định giá được sử dụng bởi các công ty bảo hiểm và các quy tắc và quy định liên quan đến việc định giá. Chúng tôi sẽ tìm hiểu về các mô hình và thuật toán tính toán giá cả bảo hiểm, cũng như cách định giá căn cứ vào dữ liệu thống kê và lịch sử. Điều này sẽ giúp hiểu rõ hơn về quy trình định giá và những yếu tố quan trọng trong việc xác định giá cả bảo hiểm ô tô.

Ngoài ra, nghiên cứu cũng cung cấp thông tin hữu ích và đóng góp phần vào việc cải thiện hiệu quả và công bằng trong quy định giá bảo hiểm ô tô. Bằng cách nắm bắt được những yếu tố quan trọng và áp dụng các phương pháp tối ưu, chúng tôi hy vọng rằng nghiên cứu này có thể hỗ trợ các công ty bảo hiểm trong việc định giá bảo hiểm ô tô để xác định một cách chính xác hơn, từ đó mang lại lợi ích cho cả công ty và khách hàng. Đồng thời, nghiên cứu cũng có thể đóng góp vào sự phát triển của lĩnh vực bảo hiểm ô tô, mở ra những cơ hội mới trong việc tăng cường tính minh bạch, đa dạng hóa sản phẩm và tăng cường niềm tin của khách hàng trong lĩnh vực bảo hiểm ô tô.

1.3. Mô tả bài toán



Hình 1.1: Mô tả bài toán

Quy trình nghiên cứu định giá bảo hiểm xe ô tô gồm 3 giai đoạn chính như sau:

- **Tìm hiểu lĩnh vực bảo hiểm ô tô:** Lĩnh vực bảo hiểm ô tô là một phần quan trọng trong lĩnh vực bảo hiểm, đảm bảo bồi thường cho các rủi ro liên quan đến ô tô và các hoạt động lái xe. Đây là một lĩnh vực phức tạp và đa dạng, liên quan đến nhiều khía cạnh khác nhau, bao gồm cả kỹ thuật, tài chính và pháp lý. Các khía cạnh cơ bản của lĩnh vực bảo hiểm xe ô tô bao gồm: loại hình bảo hiểm ô tô, quy định giá bảo hiểm xe ô tô, bồi thường và xử lý yêu cầu bồi thường, đổi mới trong lĩnh vực bảo hiểm xe ô tô, tầm quan trọng trong lĩnh vực bảo hiểm xe ô tô, ...
- **Tiền xử lý dữ liệu:** Làm sạch, chuẩn hóa dữ liệu, loại bỏ các giá trị gây nhiễu. Sau đó phân tích tổng quan và trực quan hóa các thuộc tính có trong dữ liệu.
- **Xây dựng mô hình:** Áp dụng các thuật toán học máy phù hợp để có thể xây dựng mô hình dự đoán giá bảo hiểm.

PHẦN 2. CƠ SỞ LÝ THUYẾT

2.1. Sơ lược về Big data

Trong thế giới ngày nay, dữ liệu có tầm quan trọng rất lớn. Để xử lý và sử dụng dữ liệu này, khoa học dữ liệu đang giúp thế giới bằng các ứng dụng hoa học dữ liệu khác nhau. Trong thế giới cạnh tranh ngày nay, có nhiều ngành đang sử dụng Khoa học dữ liệu để xử lý dữ liệu lớn hiệu quả và đạt được sự tăng trưởng trong ngành của họ. Tầm quan trọng của Big Data hiện nay không thể phủ nhận và đang ngày càng gia tăng. Những lợi ích và tầm quan trọng của Big Data trong thế giới hiện đại:

- **Hiểu khách hàng và cá nhân hóa:** Big Data cho phép các doanh nghiệp thu thập và phân tích thông tin về khách hàng từ nhiều nguồn dữ liệu khác nhau. Điều này giúp các doanh nghiệp hiểu rõ hơn về nhu cầu, sở thích và hành vi của khách hàng, từ đó tạo ra trải nghiệm cá nhân hóa, tăng cường tương tác và nâng cao sự hài lòng của khách hàng.
- **Ra quyết định thông minh:** Big Data cung cấp cho doanh nghiệp một nguồn thông tin phong phú và chi tiết để đưa ra quyết định kinh doanh thông minh hơn. Phân tích dữ liệu lớn giúp phát hiện xu hướng, tìm ra mối quan hệ ẩn giữa các yếu tố và dự đoán kết quả. Điều này hỗ trợ các nhà quản lý và nhà lãnh đạo đưa ra quyết định dựa trên sự hiểu biết chính xác và phân tích đáng tin cậy.
- **Tối ưu hóa hoạt động kinh doanh:** Big Data cho phép doanh nghiệp phân tích quá trình kinh doanh hiện tại và tìm ra cách cải thiện hiệu suất và hiệu quả. Bằng cách áp dụng phân tích dữ liệu, doanh nghiệp có thể tối ưu hóa chuỗi cung ứng, quản lý tồn kho, dự đoán nhu cầu và tối đa hóa hiệu quả sản xuất.
- **Phát hiện và phòng ngừa gian lận:** Big Data giúp phát hiện các hoạt động gian lận và rủi ro trong nhiều lĩnh vực. Bằng cách phân tích dữ liệu, các mô hình và thuật toán có thể xác định các mô hình bất thường và cảnh báo về các hoạt động gian lận, bảo vệ doanh nghiệp và khách hàng khỏi các rủi ro tài chính và danh tiếng.
- **Nghiên cứu và phát triển:** Big Data đóng vai trò quan trọng trong việc nghiên cứu và phát triển. Dữ liệu lớn cung cấp nguồn tài nguyên quý giá cho các nhà nghiên cứu trong nhiều lĩnh vực, từ y học, sinh học, vật lý, tài chính đến khoa học xã hội. Big Data cung cấp khả năng xử lý và phân tích dữ liệu lớn, giúp nhà nghiên

cứu khám phá mối quan hệ, xu hướng và mô hình mới trong dữ liệu. Điều này mang lại những khám phá và hiểu biết sâu sắc về các lĩnh vực này, đồng thời tạo ra cơ hội cho sự phát triển và cải tiến.

- **Dự đoán và dự báo:** Big Data cho phép dự đoán và dự báo trong nhiều lĩnh vực. Dữ liệu lớn được sử dụng để xây dựng các mô hình dự đoán, từ việc dự báo thị trường tài chính, thời tiết, xu hướng tiêu dùng đến việc dự đoán sự thành công của một sản phẩm hoặc dự án. Những dự đoán và dự báo này giúp các tổ chức và doanh nghiệp đưa ra quyết định chiến lược và lập kế hoạch tốt hơn.

2.2. Ứng dụng Big data trong lĩnh vực bảo hiểm xe ô tô

Khoa học dữ liệu đã có ứng dụng rộng rãi trong lĩnh vực định giá bảo hiểm xe ô tô, đóng vai trò quan trọng trong việc cung cấp thông tin quan trọng và chi tiết để đưa ra quyết định định giá chính xác và công bằng.

- **Định giá bảo hiểm:** Big Data cho phép các công ty bảo hiểm thu thập và phân tích dữ liệu lớn về các yếu tố liên quan đến xe ô tô, như tuổi của tài xế, khu vực lưu trú, lịch sử lái xe và các yếu tố an toàn của xe. Bằng cách phân tích dữ liệu này, các công ty có thể xây dựng mô hình định giá bảo hiểm chính xác hơn, từ đó đưa ra các mức phí bảo hiểm phù hợp với từng khách hàng.
- **Đánh giá rủi ro:** Big Data giúp các công ty bảo hiểm ô tô đánh giá rủi ro liên quan đến việc bảo hiểm xe hơi. Dữ liệu từ các cảm biến trên xe, hệ thống định vị và các nguồn dữ liệu khác cho phép theo dõi và phân tích hành vi lái xe, tình trạng xe và môi trường lái xe. Điều này giúp đánh giá rủi ro cá nhân của từng khách hàng và xác định mức độ bảo hiểm thích hợp.
- **Phát hiện gian lận:** Big Data được sử dụng để phát hiện các hoạt động gian lận trong lĩnh vực bảo hiểm ô tô. Bằng cách phân tích dữ liệu lớn, các mô hình và thuật toán có thể xác định các mô hình bất thường và cảnh báo về các hoạt động gian lận, chẳng hạn như gửi thông tin sai lệch về sự cố xe hay kỹ thuật lái xe.
- **Dự đoán tai nạn và thiệt hại:** Big Data cung cấp cho các công ty bảo hiểm khả năng dự đoán và phân tích các mô hình tai nạn và thiệt hại trong lĩnh vực ô tô. Dữ liệu từ các nguồn như hệ thống báo cáo tai nạn, điều kiện thời tiết và địa lý được sử dụng để phát hiện các xu hướng và mô hình, từ đó giúp công ty bảo hiểm ước tính mức độ rủi ro và thiệt hại, đồng thời quản lý nguồn lực một cách hiệu quả.

Hình 2.1: Ứng dụng của Bigdata trong lĩnh vực bảo hiểm ô tô

2.3. Giới thiệu về định giá bảo hiểm xe ô tô

Định giá bảo hiểm xe ô tô là quá trình quan trọng để xác định mức phí bảo hiểm hợp lý cho chủ sở hữu xe. Các công ty bảo hiểm đánh giá nhiều yếu tố để định giá, bao gồm thông tin về xe, lịch sử lái xe, địa điểm và môi trường sống, mục đích sử dụng xe và phạm vi bảo hiểm.

Định giá bảo hiểm xe ô tô mang lại nhiều lợi ích quan trọng cho chủ sở hữu xe. Trước hết, việc có một chính sách bảo hiểm xe ô tô giúp bảo vệ tài sản của bạn khỏi những rủi ro không mong muốn. Tai nạn, hư hỏng, mất mát do cướp, cháy nổ hay thiên tai có thể gây thiệt hại nặng nề cho xe của bạn. Tuy nhiên, với bảo hiểm, bạn sẽ được bồi thường hoặc sửa chữa xe theo điều khoản hợp đồng, giúp giảm thiểu rủi ro tài chính đáng kể, không chỉ đảm bảo bảo vệ tài sản và giảm thiểu rủi ro tài chính mà còn mang lại sự an tâm, sự hỗ trợ pháp lý và sự linh hoạt trong việc tùy chỉnh chính sách. Việc có một chính sách bảo hiểm xe ô tô đáng tin cậy sẽ mang lại sự yên tâm và bình an trong suốt hành trình lái xe.

Hình 2.2: Định giá bảo hiểm xe ô tô

2.4. Các kỹ thuật phân tích trong nghiên cứu

2.4.1. Phân cụm

2.4.1.1. Giới thiệu phân cụm

Phân cụm dữ liệu là quy trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm càng giống nhau càng tốt và các đối tượng khác cụm thì càng khác nhau nhau càng tốt.

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Có rất nhiều kỹ thuật phân cụm như phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ... Tuy nhiên, không có tiêu chí nào được xem là tốt nhất để đánh giá hiệu quả của phân tích phân cụm, điều này phụ thuộc vào mục đích của bài toán phân cụm.

Phân cụm dữ liệu mang lại nhiều lợi ích quan trọng trong việc xử lý và hiểu dữ liệu. Nó giúp khám phá cấu trúc ẩn trong dữ liệu mà không cần thông tin nhãn trước đó. Điều này cho phép chúng ta phát hiện ra các mẫu, quy luật hoặc mối quan hệ giữa các mẫu dữ liệu. Phân cụm dữ liệu cung cấp cơ sở để thực hiện phân tích và khai thác dữ liệu sâu hơn. Bằng cách xem xét các cụm riêng biệt và sự khác biệt giữa chúng, chúng ta có thể nhận ra các đặc trưng, xu hướng và mô hình ẩn trong dữ liệu. Điều này hỗ trợ trong việc đưa ra quyết định thông minh và tối ưu hóa.

2.4.1.2. Giới thiệu K-Means

K-Means là một thuật toán phân cụm dữ liệu (clustering) phổ biến trong Machine Learning và Data Mining. K-Means được sử dụng để phân chia một tập dữ liệu cho trước thành K cụm (clusters) khác nhau, với mỗi cụm chứa các điểm dữ liệu có tính chất tương tự nhau.

Thuật toán K-Means hoạt động bằng cách chọn K điểm dữ liệu ngẫu nhiên làm tâm của K cụm ban đầu, sau đó lặp lại các bước sau cho đến khi đạt được điều kiện dừng:

- Gán từng điểm dữ liệu vào cụm gần nhất với tâm cụm
- Cập nhật lại tâm cho mỗi cụm bằng cách tính trung bình của tất cả các điểm dữ liệu trong cụm
- Các bước trên được lặp lại cho đến khi không còn có sự thay đổi về việc gán điểm dữ liệu vào các cụm. Kết quả cuối cùng của thuật toán K-Means là K cụm chứa các điểm dữ liệu có tính chất tương tự nhau, và tâm của mỗi cụm là một điểm dữ

liệu trung bình của các điểm trong cụm đó.

K-Means là một trong những thuật toán phân cụm dữ liệu đơn giản nhất và phổ biến nhất, nhưng cũng có một số hạn chế. Ví dụ, thuật toán K-Means có thể dẫn đến kết quả không tối ưu nếu chọn sai số lượng cụm K ban đầu, hoặc nếu dữ liệu có kích thước lớn hoặc có tính chất phức tạp. Ngoài ra, K-Means không thể xử lý được dữ liệu có tính chất phi tuyến tính hoặc không phân cách được bằng các đường thẳng.

2.4.2. Phân lớp

2.4.2.1. Giới thiệu phân lớp

Phân lớp dữ liệu là kĩ thuật dựa trên tập huấn luyện và những giá trị hay hay là nhãn của lớp trong một thuộc tính phân lớp và sử dụng nó trong việc phân lớp dữ liệu mới.

Mục tiêu của phương pháp phân lớp dữ liệu là dự đoán nhãn lớp cho các mẫu dữ liệu. Không giống như phân cụm dữ liệu, phân lớp dữ liệu là học bằng ví dụ, trong khi phân cụm dữ liệu có thể coi là một cách học bằng quan sát.

Dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu định danh như khu vực sinh sống, hạng xe, tình trạng việc làm của người lái xe,… Thuật toán phổ biến dùng để xây dựng cây phân lớp là ID3, J48, C4.5, C5.0.

2.4.2.2. Các mô hình phân lớp

Bảng 2-1: So sánh các mô hình phân lớp

	XGBClassifier	Navie Bayes Classifier	MLPClassifier
Đặc điểm	Sử dụng phương pháp tối ưu hóa Gradient Boosting để xây dựng một loạt các cây quyết định yếu và kết hợp chúng lại để tạo ra một cây quyết định mạnh hơn.	Mô hình phân loại dựa trên giả thiết độc lập giữa các đặc trưng, tính toán xác suất của từng lớp và dựa trên xác suất đó để phân loại các mẫu dữ liệu mới.	Một thuật toán học có giám sát dựa trên mạng nơ-ron nhân tạo. Nó có thể xử lý các bộ dữ liệu phi tuyến tính và có khả năng học các mô hình phức tạp.
Ưu điểm	– Có khả năng xử lý các tập dữ liệu lớn. – Độ chính xác cao. – Có khả năng giảm thiểu overfitting.	– Có thể được huấn luyện nhanh chóng và hoạt động hiệu quả trên các tập dữ liệu lớn.	– Có khả năng học các quan hệ phi tuyến giữa các đặc trưng. – Có thể xử lý các tập dữ liệu lớn.

			– Có khả năng học các đặc trưng phức tạp và được sử dụng rộng rãi trong các bài toán phân loại ảnh và ngôn ngữ tự nhiên.
Nhược điểm	Cần nhiều thời gian để huấn luyện.	Không hoạt động tốt trên các bộ dữ liệu có tính tương quan cao giữa các tính năng.	Có thể bị overfitting nếu số lượng lớp ẩn và số lượng nơ-ron không được chọn đúng cách.

2.4.3. Hồi quy

2.4.3.1. Giới thiệu hồi quy

Phân tích hồi quy là kỹ thuật thống kê dùng để ước lượng phương trình phù hợp nhất với các tập hợp kết quả quan sát của biến phụ thuộc và biến độc lập. Nó cho phép đạt được kết quả ước lượng tốt nhất về mối quan hệ chân thực giữa các biến số.

Trong phân tích hồi quy, những yếu tố đó được gọi là biến. Biến phụ thuộc – yếu tố chính mà bạn đang cố gắng hiểu hoặc dự đoán. Phân tích hồi quy bao gồm một số biến thể, chẳng hạn như tuyến tính, nhiều tuyến tính và phi tuyến tính. Các mô hình phổ biến nhất là tuyến tính đơn giản và nhiều tuyến tính. Phân tích hồi quy phi tuyến thường được sử dụng cho các tập dữ liệu phức tạp hơn trong đó các biến phụ thuộc và độc lập thể hiện mối quan hệ phi tuyến.

Phân tích hồi quy cung cấp nhiều ứng dụng trong các lĩnh vực khác nhau, bao gồm cả bảo hiểm. Dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu định lượng như số lượng hợp đồng bảo hiểm mà một chủ xe đã mua, tổng tiền yêu cầu bồi thường, giá trị trọn đời…

2.4.3.2. Giới thiệu XGB Regressor

XGBRegressor là một mô hình học máy sử dụng thuật toán Gradient Boosting và được xây dựng trên nền tảng thư viện XGBoost. Nó được sử dụng để dự đoán giá trị liên tục cho các bài toán regression.

XGBRegressor sử dụng một loạt các cây quyết định để học từ dữ liệu và tối ưu hóa một hàm mất mát bằng cách sử dụng kỹ thuật Gradient Boosting. Nó có thể xử lý tập dữ liệu lớn, có khả năng xử lý các tính năng phức tạp và giảm thiểu các vấn đề về

overfitting.

XGBRegressor cũng có thể được sử dụng để đánh giá độ quan trọng của các tính năng trong dữ liệu, giúp cho việc lựa chọn tính năng trở nên dễ dàng hơn.

Tổng quan, XGBRegression là một công cụ mạnh mẽ để xây dựng các mô hình regression và được sử dụng rộng rãi trong các bài toán dự đoán giá trị liên tục trong nhiều lĩnh vực, như tài chính, y tế, kinh doanh và nghiên cứu khoa học.

PHẦN 3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Mô tả dữ liệu

Link dữ liệu: https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data

Mô tả cột: Có 24 cột x 9134 dòng

Bảng 3-1: Các nhóm đặc điểm của dữ liệu

Class	Name
1	Nhân khẩu học
2	Đặc điểm xe
3	Hợp đồng
4	Lịch sử lái xe

Bảng 3-2: Mô tả dữ liệu

#	Column	Info	Value	Class
1	Customer	Mã khách hàng	ID	0
2	State	Bang đang sinh sống	Washington, Arizona, Nevada, California, Oregon	1
3	Customer Lifetime Value	Giá trị trọn đời của khách hàng	Tổng giá trị mà khách hàng dự kiến sẽ mang lại cho công ty bảo hiểm trong suốt thời gian khách hàng ở lại với công ty.	3
4	Response	phản hồi của công ty bảo hiểm sau khi nhận được yêu cầu bồi thường từ chủ xe	Yes: Công ty bảo hiểm đồng ý chi trả bồi thường cho chủ xe No: Công ty bảo hiểm từ chối chi trả bồi thường cho chủ xe	3
5	Coverage	Loại bảo hiểm được mua để bảo vệ chiếc xe	Basic: mức bảo hiểm tối thiểu được yêu cầu để lái xe trên đường, bao gồm phần bảo trì và sửa chữa xe, phí y tế và thiệt hại cho người khác hoặc xe hơi khác trong trường hợp gây ra tai nạn. Extended: Basic + cung cấp bảo hiểm cho các sự cố khác như mất cắp, thiệt hại do thời tiết hoặc động đất, hoặc thiệt hại do đâm vào động vật hoang dã Premium: Extended + cùng với một số	3

#	Column	Info	Value	Class
			tiện ích bổ sung như bảo hiểm cho người lái hoặc phụ xe trong trường hợp tai nạn, bảo hiểm chỗ ở tạm thời trong trường hợp xe của bạn bị hư hỏng và bạn cần phải sửa chữa trong một thời gian dài	
6	Education	Trình độ học vấn	Bachelor: Cử nhân College: Cao đẳng Master: Giáo sư Doctor: Bác sĩ High School or Below: từ cấp 2 trở xuống	1
7	Effective To Date	ngày bắt đầu hiệu lực của hợp đồng bảo hiểm	Ngày bắt đầu hiệu lực của chính sách bảo hiểm.	3
8	Employment Status	Tình trạng việc làm của người lái xe	Employed: Có việc làm Unemployed: Không có việc làm Retired: Nghỉ hưu Disabled: Tàn tật Medical Leave: Đang phục hồi	1
9	Gender	Giới tính	F: Nữ M: Nam	1
10	Income	Thu nhập	Money	1
11	Location Code	Khu vực sinh sống	Suburban: Ngoại ô Rural: Nông thôn Urban: Thành thị	1
12	Marital Status	Tình trạng hôn nhân	Married: Kết hôn Single: Độc thân Divorced: Ly hôn	1
13	Monthly Premium Auto	Khoản tiền bảo hiểm hàng tháng mà khách hàng phải trả	Số tiền bảo hiểm hàng tháng mà khách hàng đang trả cho công ty bảo hiểm.	3
14	Months Since Last Claim	Số tháng kể từ lần gửi yêu cầu bồi thường cuối cùng		4

#	Column	Info	Value	Class
15	Months Since Policy Inception	thời gian kể từ khi chính sách bảo hiểm được mua	Số tháng kể từ khi khách hàng đã mua chính sách bảo hiểm.	3
16	Number of Open Complaints	số lượng khiếu nại đang mở của khách hàng tại thời điểm hiện tại	1,2,3,4,5	4
17	Number of Policies	số lượng hợp đồng bảo hiểm mà một chủ xe đã mua	Bao gồm nhiều loại bảo hiểm khác nhau như bảo hiểm trách nhiệm dân sự, bảo hiểm tai nạn cá nhân, bảo hiểm tai nạn xe hơi và nhiều loại bảo hiểm khác. Số lượng hợp đồng bảo hiểm càng nhiều, chủ xe sẽ có mức độ bảo vệ cao hơn đối với chiếc xe của mình. Tuy nhiên, điều này cũng có thể dẫn đến chi phí bảo hiểm cao hơn.	3
18	Policy Type	Loại chính sách	Corporate Auto: Ô tô công ty Personal Auto: Ô tô cá nhân Special Auto: Ô tô đặc biệt	3
19	Policy	Chính sách bảo hiểm	Corporate L1, Corporate L2, Corporate L3, Personal L1, Personal L2, Personal L3, Special L1, Special L2, Special L3	3
20	Renew Offer Type	Gia hạn loại ưu đãi	Offer1, Offer2, Offer3, Offer4	3
21	Sales Channel	Kênh bán hàng	Agent: Đại lý Call Center: Trung tâm cuộc gọi Web: Qua mạng Branch: Chi nhánh	3
22	Total Claim Amount	Tổng tiền yêu cầu bồi thường	Money	3
23	Vehicle Class	Hạng xe	Two-Door Car: Xe 2 cửa Four-Door Car: Xe 4 cửa Sports Car: Xe thể thao Luxury: Xe sang	2

#	Column	Info	Value	Class
24	Vehicle Size	Kích thước xe	Large: Cỡ lớn Medsize: Cỡ trung small: Nhỏ	2

3.2. Tiền xử lý dữ liệu

– Chọn các cột đặc trưng trong bộ dữ liệu:

Qua tìm hiểu thì thấy được các cột Customer, Vehicle Class, Coverage, Marital Status, Location Code, Months Since Driving, Age, EmploymentStatus, Monthly Premium Auto, Months Since Last Claim, State là các cột có ảnh hưởng đến chi phí mà khách hàng sẽ trả cho công ty bảo hiểm ô tô.

```
df_model=df[['Customer','Vehicle Class','Coverage','Marital Status','Location Code','Months Since Driving',
             'Age','EmploymentStatus','Monthly Premium Auto','Months Since Last Claim','State']]
df_model
```

✓ 0.0s Python

	Customer	Vehicle Class	Coverage	Marital Status	Location Code	Months Since Driving	Age	EmploymentStatus	Monthly Premium Auto	Months Since Last Claim	State
0	BU79786	Two-Door Car	Basic	Married	Suburban	236	41	Employed	69	32	Washington
1	QZ44356	Four-Door Car	Extended	Single	Suburban	5	39	Unemployed	94	13	Arizona
2	AI49188	Two-Door Car	Premium	Married	Suburban	23	54	Employed	108	18	Nevada
3	WW63253	SUV	Basic	Married	Suburban	12	23	Unemployed	106	18	California
4	HB64268	Four-Door Car	Basic	Single	Rural	643	70	Employed	73	12	Washington
...	...	...	...	...	...	...	...	...	...	...	...
9129	LA72316	Four-Door Car	Basic	Married	Urban	263	57	Employed	73	18	California
9130	PK87824	Four-Door Car	Extended	Divorced	Suburban	262	49	Employed	79	14	California
9131	TD14365	Four-Door Car	Extended	Single	Suburban	162	46	Unemployed	85	9	California
9132	UP19263	Four-Door Car	Extended	Married	Suburban	253	65	Employed	96	34	California
9133	Y167826	Two-Door Car	Extended	Single	Suburban	374	49	Unemployed	77	3	California

9134 rows × 11 columns

Hình 3.1: Các cột đặc trưng trong dữ liệu

– Kiểm tra giá trị null → Không có giá trị null

```
df.isnull().sum()
```

✓ 0.0s

```
Customer                       0
State                          0
Customer Lifetime Value        0
Response                       0
Coverage                       0
Education                      0
Effective To Date              0
EmploymentStatus               0
Gender                         0
Income                         0
Location Code                  0
Marital Status                 0
Monthly Premium Auto           0
Months Since Last Claim        0
Months Since Policy Inception  0
Number of Open Complaints      0
Number of Policies             0
Policy Type                    0
Policy                         0
Renew Offer Type               0
Sales Channel                  0
Total Claim Amount             0
Vehicle Class                  0
Vehicle Size                   0
Age                            0
Months Since Driving           0
dtype: int64
```

Hình 3.2: Kiểm tra giá trị null

– Kiểm tra giá trị trùng lặp → Không có giá trị trùng lặp

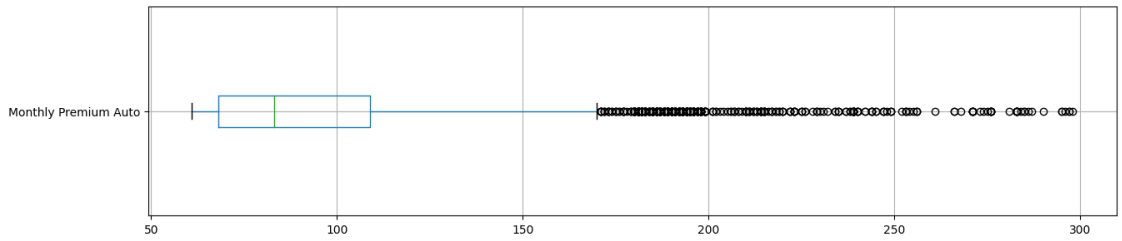
```
df.duplicated().sum()
✓ 0.1s
0
```

Hình 3.3: Kiểm tra giá trị trùng lặp

– Kiểm tra overfitting

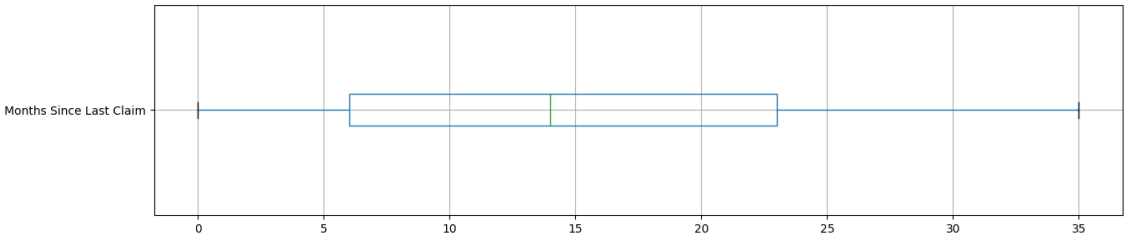
```
fig, axes = plt.subplots(nrows=4, ncols=1, figsize=(15, 15))
df[["Monthly Premium Auto"]].boxplot(ax=axes[0], vert=False)
df[["Months Since Last Claim"]].boxplot(ax=axes[1], vert=False)
df[["Age"]].boxplot(ax=axes[2], vert=False)
df[["Months Since Driving"]].boxplot(ax=axes[3], vert=False)
✓ 0.7s
```

Hình 3.4: Kiểm tra overfitting



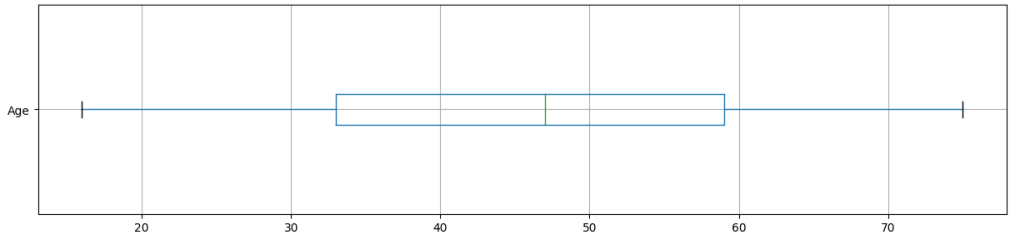
Hình 3.5: Biểu đồ Boxplot của Monthly Premium Auto

+ Cột *Monthly Premium Auto* có giá trị ngoại lai tuy nhiên đây cũng thể hiện giá bảo hiểm có sự chênh lệch và đặc trưng của giá bảo hiểm đối với từng nhóm khách hàng → Nó chứa thông tin quan trọng nên không xử lý.



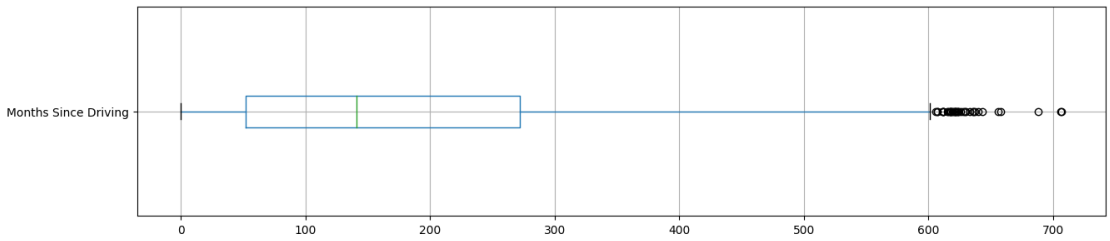
Hình 3.6: Biểu đồ Boxplot của Months Since Last Claim

+ Cột Months Since Last Claim không có giá trị ngoại lại



Hình 3.7: Biểu đồ Boxplot của Age

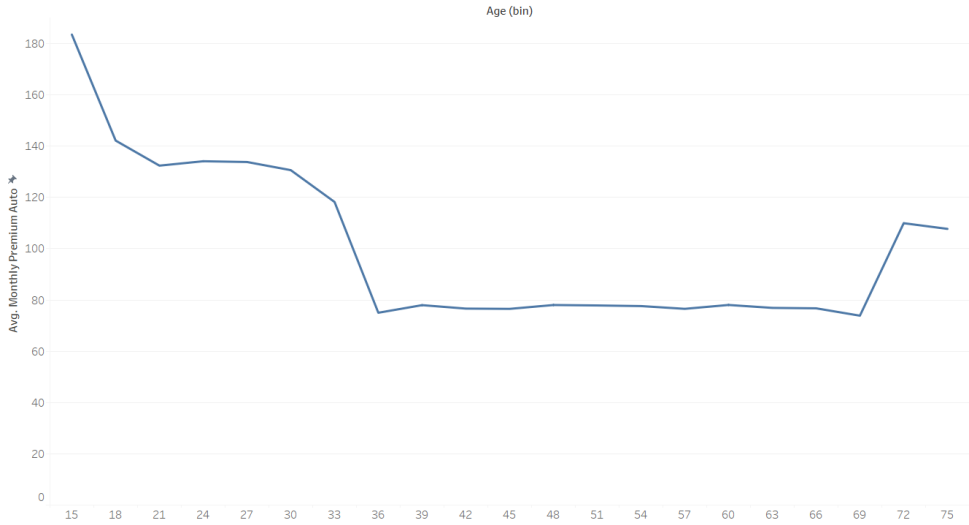
+ Cột Age không có giá trị ngoại lại



Hình 3.8: Biểu đồ Boxplot của Months Since Driving

+ Cột Months Since Driving có giá trị ngoại lai tuy nhiên nếu loại bỏ các giá trị ngoại lai này có thể sẽ làm mất mát một phần thông tin quan trọng và làm giảm độ chính xác của mô hình → Không xử lý.

3.3. Trực quan hóa dữ liệu



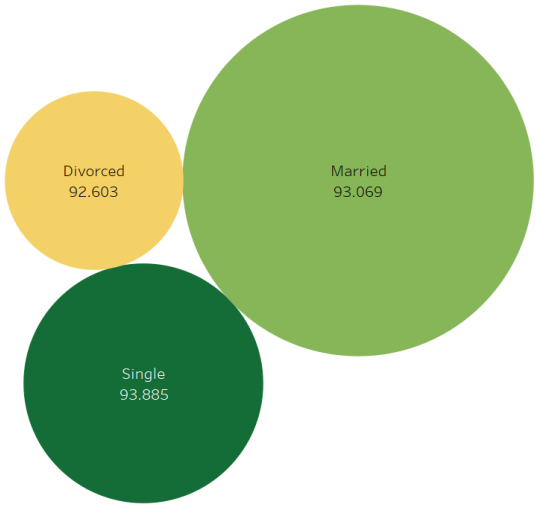
Hình 3.9: Biểu đồ đường giữa tuổi và trung bình giá bảo hiểm hàng tháng

Chi phí bảo hiểm xe ô tô có thể thay đổi dựa trên độ tuổi của tài xế. Thông thường, các công ty bảo hiểm xem xét độ tuổi như một yếu tố quan trọng để đánh giá rủi ro và tính phí bảo hiểm. Dưới đây là một số thông tin tổng quát về cách độ tuổi có thể ảnh hưởng đến chi phí bảo hiểm xe ô tô:

– Tài xế trẻ: Các tài xế trẻ, thường là độ tuổi 25 trở lại, có thể đối mặt với mức phí bảo hiểm cao nhất. Bởi vì tài xế trẻ thường được xem là nhóm rủi ro cao, có khả năng gây tai nạn hoặc có thể gặp nguy hiểm khi lái xe cao hơn những người trung niên.

– Tài xế trung niên: Các tài xế ở độ tuổi trung niên là nhóm có rủi ro thấp nhất. Vì thường họ là những người đã có kinh nghiệm sử dụng ô tô lâu năm và đã đủ để có thể kiểm soát hành vi của mình tốt nên sẽ là nhóm ít gây tai nạn nhất trong ba nhóm tuổi.

– Tài xế là người lớn tuổi: Trong một số trường hợp, các công ty bảo hiểm có thể tính phí cao hơn cho người có độ tuổi cao vì khả năng sức khỏe có thể không đảm bảo và khả năng phản ứng chậm hơn khi lái xe.



Hình 3.10: Biểu đồ giữa trình độ hôn nhân và trung bình giá bảo hiểm hàng tháng

Số người đã kết hôn chiếm đa số nên có thể nói rằng người đã kết hôn thường sẽ mua bảo hiểm xe ô tô nhiều hơn. Một số lý do gây nên hiện tượng đó:

– **Trách nhiệm gia đình:** Người kết hôn thường có trách nhiệm hơn đối với gia đình và người thân, bao gồm việc cung cấp cho con cái, trang trải chi phí sinh hoạt và chăm sóc cho vợ/chồng. Vì vậy, họ có xu hướng mua bảo hiểm để đảm bảo rằng gia đình được bảo vệ tốt nhất trong trường hợp xảy ra sự cố.

– **Lợi ích kinh tế:** Đôi khi, việc kết hôn có thể mang lại lợi ích tài chính, bao gồm giảm chi phí sống và tăng thu nhập. Những lợi ích này có thể dẫn đến khả năng mua bảo hiểm cao hơn cho cả hai người trong đôi vợ chồng.

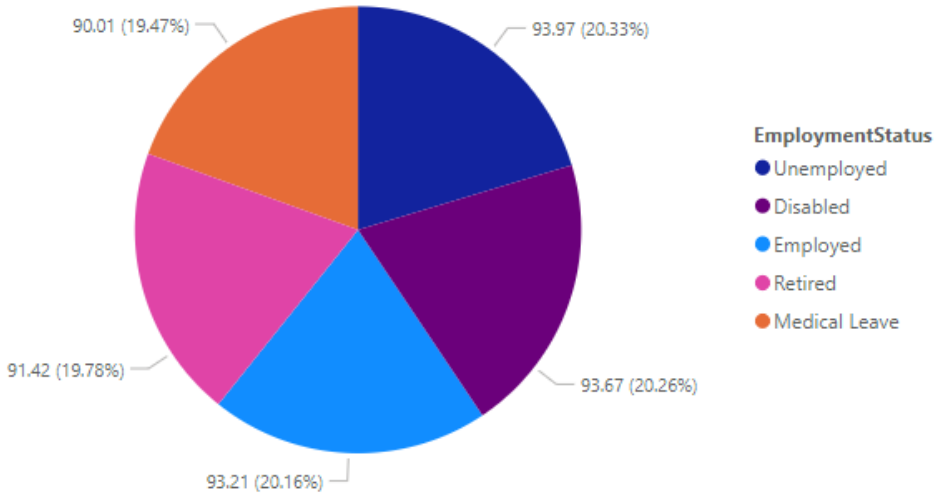
– **Khả năng tài chính ổn định hơn:** Người kết hôn có thể có khả năng tài chính ổn định hơn so với những người độc thân. Họ có thể có thu nhập ổn định và tài sản chung, giúp tăng khả năng mua bảo hiểm cao hơn để bảo vệ tài sản và gia đình.

– **Tăng cường ý thức bảo hiểm:** Khi kết hôn, người ta có xu hướng thảo luận và suy nghĩ về tương lai, bao gồm cả việc bảo vệ gia đình trong mọi trường hợp. Do đó, việc mua bảo hiểm được coi là một phần của kế hoạch tài chính của đôi vợ chồng.

Tuy nhiên trung bình mức phí đóng bảo hiểm của khách hàng độc thân vẫn cao

nhất với mức giá là 93,885$, nhưng vẫn không có chênh lệch nhiều với 2 nhóm còn lại.

Nguyên nhân gây nên hiện tượng đó có thể là do người đó không có vợ hoặc chồng, do đó họ không có ai để chia sẻ rủi ro trong trường hợp xấu nhất. Điều này có nghĩa là người đó sẽ phải chịu một khoản bảo hiểm cao hơn so với những người kết hôn, vì những người kết hôn sẽ chia sẻ rủi ro với đối tác của họ. Ngoài ra, giá bảo hiểm cũng phụ thuộc vào nhiều yếu tố khác như tình trạng sức khỏe, độ tuổi, nghề nghiệp, lịch sử lái xe, v.v. Nếu như các yếu tố này giống nhau giữa các nhóm Married, Single và Divorced, thì giá bảo hiểm của các nhóm này sẽ tương đương nhau. Hoặc các công ty bảo hiểm cũng có thể áp dụng các chiến lược khác nhau cho từng nhóm khách hàng để tối ưu hóa lợi nhuận của họ. Chẳng hạn, công ty bảo hiểm có thể áp dụng giảm giá cho khách hàng kết hôn hoặc có con cái để thu hút khách hàng và giữ chân họ. Do đó, giá bảo hiểm của các nhóm Married, Single, Divorced có thể tương đương nhau mặc dù giá bảo hiểm của Single cao hơn so với các nhóm khác.



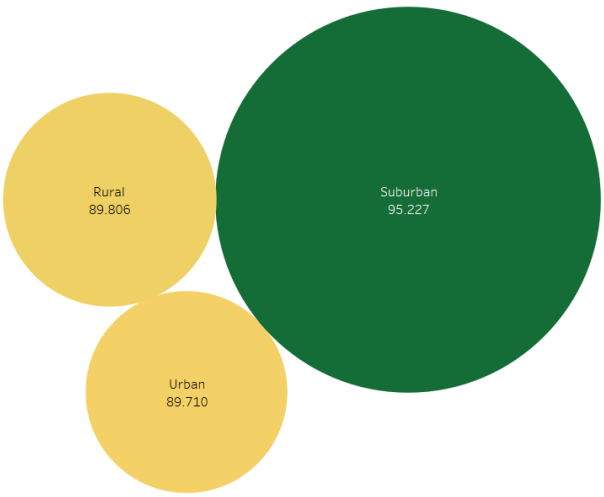
Hình 3.11: Biểu đồ giữa tình trạng việc làm và trung bình giá bảo hiểm hàng tháng

Giá bảo hiểm của một khoản bảo hiểm được tính dựa trên rủi ro xảy ra trong tương lai. Vì vậy, các khoản bảo hiểm khác nhau sẽ có mức giá khác nhau tùy vào mức độ rủi ro mà chúng đại diện.

- Trong trường hợp bảo hiểm người thất nghiệp (Unemployed), rủi ro là khả năng người đó không có nguồn thu nhập ổn định để trang trải chi phí cuộc sống và đóng góp cho khoản tiết kiệm hưu trí. Do đó, giá bảo hiểm cho người thất nghiệp sẽ cao hơn so với những người có thu nhập ổn định.
- Người khuyết tật (Disable) cũng có giá bảo hiểm cao hơn vì họ có khả năng mất

đi khả năng lao động và thu nhập trong tương lai. Rủi ro này sẽ dẫn đến chi phí để hỗ trợ cuộc sống của người khuyết tật trong thời gian dài, bao gồm chi phí y tế, phương tiện trợ giúp và các chi phí khác. Do đó, giá bảo hiểm cho người khuyết tật cũng sẽ cao hơn so với những người có khả năng lao động và thu nhập bình thường.

- Với những người có việc làm (Employed), rủi ro là khả năng mất việc và thu nhập trong tương lai. Tuy nhiên, do họ đang có nguồn thu nhập ổn định trong thời điểm hiện tại, rủi ro này thấp hơn so với những người thất nghiệp hoặc khuyết tật. Do đó, giá bảo hiểm cho người có việc làm sẽ thấp hơn so với những khoản bảo hiểm khác.
- Người về hưu (Retired) cũng có giá bảo hiểm thấp hơn do họ đã nghỉ hưu và không còn thu nhập từ việc làm nữa. Tuy nhiên, rủi ro của họ là sức khỏe có thể suy giảm và chi phí y tế có thể tăng cao hơn trong tương lai. Do đó, giá bảo hiểm cho người về hưu sẽ cao hơn so với những người có khả năng lao động và thu nhập bình thường, nhưng thấp hơn so với những khoản bảo hiểm cho người thất nghiệp hoặc khuyết tật.
- Cuối cùng, người nghỉ phép y tế (Medical Leave) có giá bảo hiểm thấp nhất vì họ đang được trả lương trong thời gian nghỉ phép y tế và rủi ro của họ thường là tạm thời và không kéo dài lâu dài. Tuy nhiên, nếu rủi ro của họ kéo dài hoặc trở nên nghiêm trọng hơn, giá bảo hiểm có thể tăng lên để phản ánh rủi ro đó.

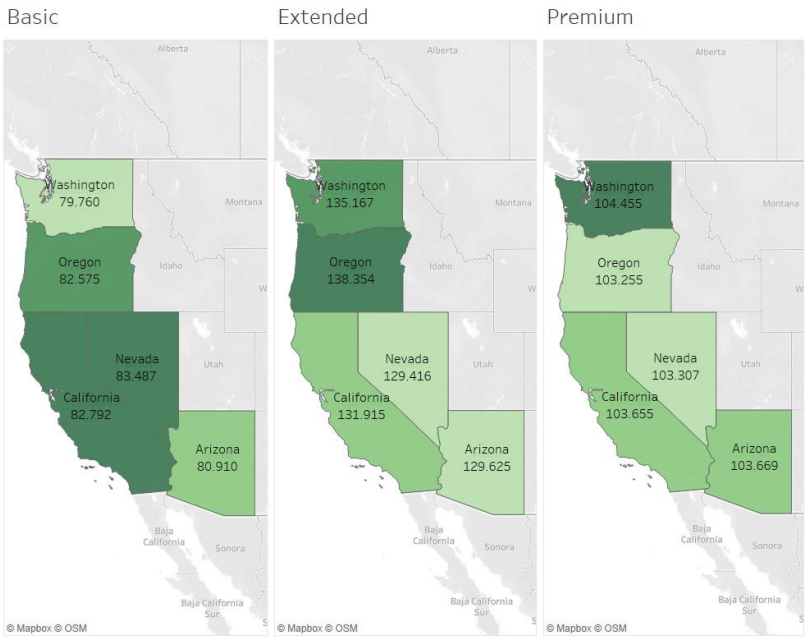


Hình 3. 12: Biểu đồ giữa khu vực sống và trung bình giá bảo hiểm hàng tháng

Vì các bang đang xét đều tập trung ở ngoại ô của Mỹ nên các khách ở ngoại ô

chiếm đa số và mức giá bảo hiểm trung bình ở vùng này cũng cao nhất với 95,227$. Hai vùng nông thôn và thành thị có số lượng và mức giá bảo hiểm trung bình tương đương nhau với mức giá là 89,806$ và 89,710$. Vị trí địa lý của một khu vực cũng là một yếu tố quan trọng trong việc xác định giá bảo hiểm xe hơi.

- Một trong những lý do chính là tần suất tai nạn ở vùng ngoại ô thường cao hơn so với các vùng khác. Điều này có thể do các tài xế chạy xe lâu hơn để đi đến nơi làm việc hoặc các địa điểm tham quan giải trí, dẫn đến mức độ mệt mỏi và cảnh giác giảm; cũng như tần suất gặp phải các tình huống khó khăn hơn, chẳng hạn như xe tải lớn hay xe buýt đi qua. Ngoài ra, vùng ngoại ô thường có tốc độ xe cao hơn so với khu vực đô thị và nông thôn, dẫn đến nguy cơ tai nạn nghiêm trọng hơn. Bên cạnh đó, các vùng ngoại ô thường có giá trị tài sản cao hơn so với các vùng khác, vì có nhiều nhà và xe hơi, cũng như các mặt hàng đắt tiền khác. Việc bảo vệ tài sản này đòi hỏi chi phí bảo hiểm cao hơn.
- Trong khi đó, Urban và Rural có giá bảo hiểm xe hơi tương đương nhau vì các yếu tố an toàn và rủi ro tương đối ổn định. Ở khu vực đô thị, mật độ xe cộ cao và tốc độ di chuyển chậm hơn có thể giảm thiểu nguy cơ tai nạn. Trong khi đó, ở khu vực nông thôn, tội phạm và tình trạng đường xá kém có thể làm tăng nguy cơ tai nạn, nhưng vì số lượng xe cộ và tốc độ di chuyển thường thấp hơn, nên đây không phải là một vấn đề lớn.

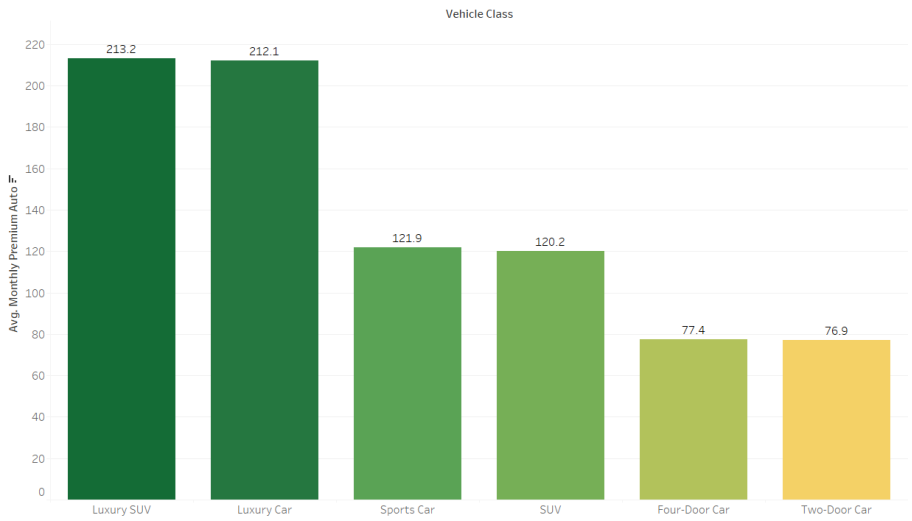


Hình 3.13: Biểu đồ giữa bang đang sinh sống và trung bình giá bảo hiểm hàng tháng

Ta thấy được với từng loại hợp đồng bảo hiểm, loại Basic thì sẽ trả phí nhiều nhất là từ 82-83 ở 2 bang California và Nevada, với loại hợp đồng Extended thì có chi phí cao nhất là hơn 138$ ở bang Oregon, cuối cùng là loại hợp đồng Premium thì chi phí bảo hiểm cho ô tô nhiều nhất là ở bang Washington với hơn 104$.

Với sự khác nhau ở từng loại hợp đồng cho thấy được chi phí bảo hiểm xe ô tô có thể khác nhau, tùy theo từng bang hoặc tiểu bang ở mỗi quốc gia. Một số yếu tố quan trọng ảnh hưởng đến chi phí bảo hiểm xe ô tô bao gồm:

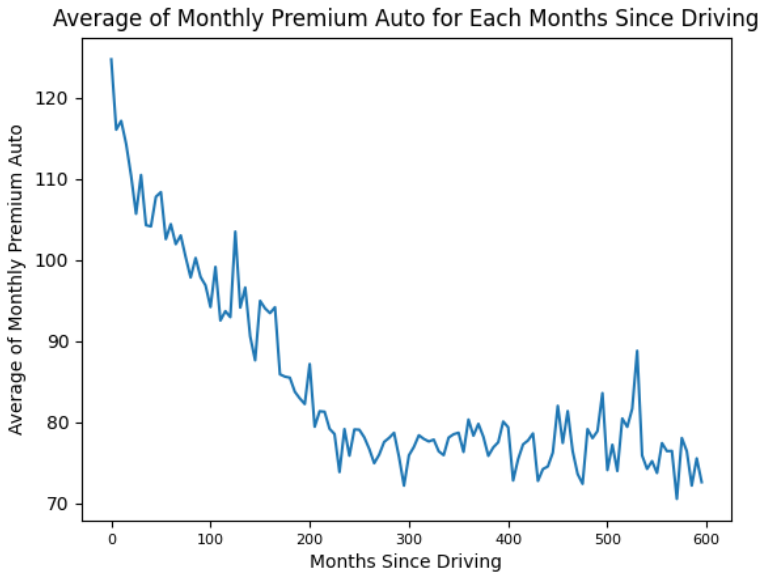
- Luật pháp địa phương: Luật bảo hiểm và quy định pháp lý liên quan đến bảo hiểm có thể khác nhau giữa các bang hoặc tiểu bang. Các yêu cầu bắt buộc, như mức tối thiểu của bảo hiểm trách nhiệm dân sự và các loại bảo hiểm bổ sung có thể khác nhau.
- Thống kê tai nạn: Tỉ lệ tai nạn và mức độ tổn thất trong từng bang cũng có thể ảnh hưởng đến chi phí bảo hiểm xe ô tô. Các bang với tỉ lệ tai nạn cao hơn hoặc mức độ tổn thất lớn hơn có thể có mức phí bảo hiểm cao hơn.
- Điều kiện địa phương: Các yếu tố như mức độ trộm cắp xe, tình trạng giao thông, điều kiện thời tiết, và cơ sở hạ tầng đường xá trong từng bang cũng có thể ảnh hưởng đến chi phí bảo hiểm. Các bang có môi trường giao thông nguy hiểm hoặc mức độ trộm cắp xe cao có thể yêu cầu mức phí bảo hiểm cao hơn.
- Thông tin cá nhân: Các yếu tố cá nhân như độ tuổi, lịch sử lái xe, loại xe và mục đích sử dụng cũng có thể ảnh hưởng đến chi phí bảo hiểm. Những người trẻ tuổi, tài xế mới, hoặc tài xế có lịch sử lái xe không tốt có thể phải trả mức phí cao hơn.



Hình 3.14: Biểu đồ giữa loại xe và trung bình giá bảo hiểm hàng tháng

Chi phí bảo hiểm xe ô tô có thể thay đổi tùy theo loại xe. Loại xe cụ thể sẽ ảnh hưởng đến mức độ rủi ro và chi phí bảo hiểm liên quan, được chia thành ba phân khúc khác nhau rõ. Dưới đây là một số thông tin tổng quát về các loại xe có thể ảnh hưởng đến chi phí bảo hiểm:

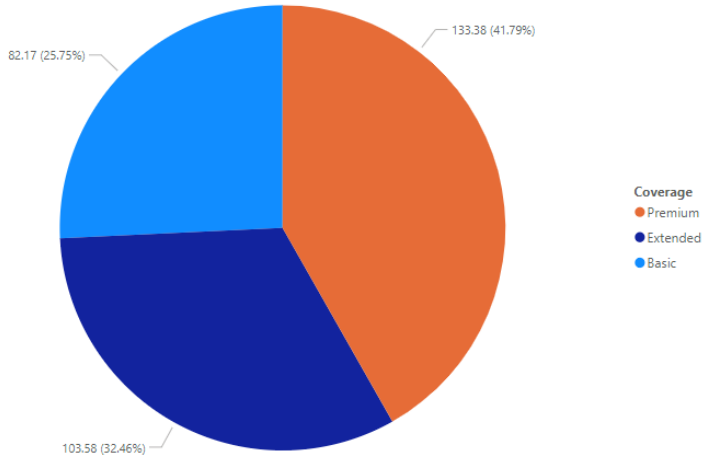
- Xe hạng sang và cao cấp: Các loại xe hạng sang và cao cấp thường có giá trị cao hơn và chi phí sửa chữa đắt đỏ hơn. Do đó, chi phí bảo hiểm cho các loại xe này thường cao hơn so với các loại xe khác, giao động từ 212-213. Các yếu tố khác như tính năng an toàn, hiệu suất động cơ và khả năng tăng tốc cũng có thể ảnh hưởng đến chi phí bảo hiểm.
- Xe thể thao: Xe thể thao có khả năng tăng tốc nhanh và thường được sử dụng với tốc độ cao. Vì vậy, chi phí bảo hiểm cho xe thể thao có thể cao hơn so với các loại xe thông thường, nhưng vẫn thấp hơn loại xe cao cấp chỉ giao động từ 120-122. Tốc độ lái xe cao và khả năng lái xe không an toàn có thể tạo ra mức rủi ro cao hơn và cần sử dụng đến bảo hiểm.
- Xe SUV cỡ lớn: Xe gia đình và xe du lịch thông thường được sử dụng cho mục đích đi lại hàng ngày và có khả năng an toàn cao hơn. Do đó, chi phí bảo hiểm cho các loại xe này thường ở mức trung bình, khoảng từ 76-78.



Hình 3.15: Biểu đồ giữa số tháng đã lái xe và trung bình giá bảo hiểm hàng tháng

Số tháng đã lái xe là một trong yếu tố chính ảnh hưởng mạnh đến giá của bảo hiểm. Như trên biểu đồ ta có thể thấy, giá của chi phí bảo hiểm sẽ càng cao khi thời gian lái

xe càng ít, giá sẽ giảm dần khi kinh nghiệm của người lái xe tăng lên nhưng tới gần cuối thì có sự tăng trở lại vì với số tháng lái xe đó chứng tỏ người lái xe có thể đã ở độ tuổi khá cao, khả năng điều khiển phương tiện sẽ xuống thấp dẫn tới việc dễ gây tai nạn.



Hình 3.16: Biểu đồ giữa gói bảo hiểm và trung bình giá bảo hiểm hàng tháng

Các gói bảo hiểm xe ô tô như Premium, Basic và Extended đều cung cấp các mức bảo vệ khác nhau và có ảnh hưởng đến mức phí phải đóng.

- Gói Basic là gói bảo hiểm cơ bản và có mức phí trung bình thấp hơn so với gói Premium. Gói này thường bao gồm bảo hiểm tai nạn cơ bản và ít dịch vụ hỗ trợ hơn. Gói Cơ bản có thể phù hợp với những người có ngân sách hạn chế hoặc chỉ quan tâm đến việc đảm bảo rủi ro cơ bản.
- Gói Extended nằm giữa gói Premium và gói Basic về phạm vi bảo vệ và mức phí. Gói này thường cung cấp một phạm vi bảo hiểm rộng hơn so với gói Cơ bản và có thể bao gồm một số dịch vụ hỗ trợ nâng cao. Phí trung bình cho gói Extended thường cao hơn so với gói Basic nhưng thấp hơn so với gói Premium.
- Gói Premium được xem là gói bảo hiểm rủi ro cao cấp, với mức phí trung bình cao hơn do cung cấp phạm vi bảo vệ toàn diện và các dịch vụ hỗ trợ khác. Gói này thích hợp cho những người muốn có độ bảo hiểm rộng rãi và sự an toàn tối đa trong trường hợp xảy ra tai nạn hoặc tổn thất toàn bộ xe.

Lựa chọn gói bảo hiểm xe ô tô phù hợp cần xem xét nhiều yếu tố như ngân sách cá nhân, mức độ bảo vệ mong muốn và nhu cầu sử dụng xe. Trước khi quyết định, nên tham khảo ý kiến và tư vấn trực tiếp với các công ty bảo hiểm để hiểu rõ hơn về mức phí công cụ có thể và phạm vi bảo hiểm của từng gói.

3.4. Phân cụm dữ liệu

3.4.1. Xây dựng mô hình K-means

Mục đích: Giúp công ty xem khách hàng của mình có bao nhiêu phân khúc và đặc trưng của mỗi nhóm như thế nào.

Xây dựng mô hình:

– Bộ dữ liệu chia làm 3 bộ nhỏ với 3 loại gói bảo hiểm khác nhau:

```
basic=df_model[df_model['Coverage']=='Basic']
extended=df_model[df_model['Coverage']=='Extended']
premium=df_model[df_model['Coverage']=='Premium']
df_model
```

✓ 0.0s Python

	Customer	Vehicle Class	Coverage	Marital Status	Location Code	Months Since Driving	Age	EmploymentStatus	Monthly Premium Auto	Months Since Last Claim	State
0	BU79786	Two-Door Car	Basic	Married	Suburban	236	41	Employed	69	32	Washington
1	QZ44356	Four-Door Car	Extended	Single	Suburban	5	39	Unemployed	94	13	Arizona
2	AI49188	Two-Door Car	Premium	Married	Suburban	23	54	Employed	108	18	Nevada
3	WW63253	SUV	Basic	Married	Suburban	12	23	Unemployed	106	18	California
4	HB64268	Four-Door Car	Basic	Single	Rural	643	70	Employed	73	12	Washington
...	...	...	...	...	...	...	...	...	...	...	...
9129	LA72316	Four-Door Car	Basic	Married	Urban	263	57	Employed	73	18	California
9130	PK87824	Four-Door Car	Extended	Divorced	Suburban	262	49	Employed	79	14	California
9131	TD14365	Four-Door Car	Extended	Single	Suburban	162	46	Unemployed	85	9	California
9132	UP19263	Four-Door Car	Extended	Married	Suburban	253	65	Employed	96	34	California
9133	Y167826	Two-Door Car	Extended	Single	Suburban	374	49	Unemployed	77	3	California

9134 rows × 11 columns

Hình 3.17: Tách dữ liệu

– Xây dựng hàm cluster() để phân cụm các bộ dữ liệu, trong đó:

+ Sử dụng các cột Customer, Vehicle Class, Coverage, Marital Status, Location Code, Months Since Driving, Age, EmploymentStatus, Monthly Premium Auto, Months Since Last Claim, State làm đầu vào của mô hình.

+ Tạo biến giả cho các biến định tính.

+ Sử dụng phương pháp standardization để chuẩn hóa dữ liệu đầu vào

+ Fit dữ liệu đã xử lý vào mô hình K-Means

```
def cluster(df_model,n):
    do_dummy_cols = ['Vehicle Class', 'Marital Status','Location Code','EmploymentStatus','State']
    clus_model= pd.get_dummies(df_model, columns=do_dummy_cols)
    clus_model=clus_model.drop(columns=['Customer','Coverage'])
    clus_model = clus_model.dropna()
    scaled_df = StandardScaler().fit_transform(clus_model)
    kmeans = KMeans(init="random", n_clusters=n, n_init=10, random_state=1)
    kmeans.fit(scaled_df)
    df_model['Cluster']=kmeans.labels_
    return df_model
```

✓ 0.0s

Hình 3.18: Xây dựng hàm cluster()

– Sử dụng hàm cluster() đã xây dựng trên với 3 bộ dữ liệu nhỏ (chọn K=2 theo từng bộ) sau đó gán từng cụm dữ liệu tạo ra bằng mô hình vào các biến lần lượt là B1, B2, E1, E2, P1, P2

```
basic = cluster(basic,2)
B1=basic[basic['Cluster']==0]
B2=basic[basic['Cluster']==1]
extended = cluster(extended,2)
E1=extended[extended['Cluster']==0]
E2=extended[extended['Cluster']==1]
premium = cluster(premium,2)
P1=premium[premium['Cluster']==0]
P2=premium[premium['Cluster']==1]
```

Hình 3.19: Sử dụng hàm cluster()

3.4.2. Đặc điểm của 6 nhóm

	Cluster	Count	State	Vehicle Class	Coverage	Marital Status	Location Code	EmploymentStatus	Months Since Driving	Age	Months Since Last Claim	Monthly Premium Auto
0	B1	1542	[California, Arizona, Oregon, Washington, Nevada]	[SUV, Sports Car, Luxury Car, Luxury SUV]	[Basic]	[Married, Divorced, Single]	[Suburban, Rural, Urban]	[Unemployed, Employed, Medical Leave, Disabled...	0.0 - 225.0	16.0 - 73.0	0.0 - 35.0	100.0 - 199.0
1	B2	4026	[Washington, Oregon, California, Nevada, Arizona]	[Two-Door Car, Four-Door Car, Sports Car, SUV]	[Basic]	[Married, Single, Divorced]	[Suburban, Rural, Urban]	[Employed, Medical Leave, Unemployed, Disabled...	0.0 - 706.0	36.0 - 75.0	0.0 - 35.0	61.0 - 119.0
2	E1	769	[Washington, Nevada, Oregon, California, Arizona]	[SUV, Luxury SUV, Sports Car, Luxury Car]	[Extended]	[Married, Divorced, Single]	[Urban, Suburban, Rural]	[Disabled, Employed, Unemployed, Medical Leave...	0.0 - 215.0	16.0 - 66.0	0.0 - 35.0	121.0 - 249.0
3	E2	1973	[Arizona, Oregon, Washington, California, Nevada]	[Four-Door Car, Two-Door Car, SUV, Sports Car]	[Extended]	[Single, Married, Divorced]	[Suburban, Urban, Rural]	[Unemployed, Employed, Disabled, Retired, Medi...	0.0 - 688.0	37.0 - 75.0	0.0 - 35.0	76.0 - 139.0
4	P1	579	[Nevada, Arizona, California, Oregon, Washington]	[Two-Door Car, Four-Door Car]	[Premium]	[Married, Single, Divorced]	[Suburban, Urban, Rural]	[Employed, Unemployed, Disabled, Medical Leave...	0.0 - 707.0	30.0 - 75.0	0.0 - 35.0	101.0 - 119.0
5	P2	245	[Oregon, Arizona, Washington, California, Nevada]	[SUV, Luxury SUV, Sports Car, Luxury Car]	[Premium]	[Married, Single, Divorced]	[Rural, Urban, Suburban]	[Disabled, Employed, Unemployed, Retired, Medi...	0.0 - 533.0	16.0 - 74.0	0.0 - 35.0	140.0 - 298.0

Hình 3.20: Đặc điểm của các cụm

Bảng 3-3: Đặc điểm của các cụm

Cụm	**B1**	**B2**	**E1**	**E2**	**P1**	**P2**
Số lượng	1542	4026	769	1973	579	245
Vehicle Class	SUV, Sports Car, Luxury SUV, Luxury Car	Four-Door Car, Two-Door Car	SUV,Sports Car, Luxury SUV, Luxury Car	Four-Door Car, Two-Door Car	Four-Door Car, Two-Door Car	SUV, Sports Car, Luxury SUV, Luxury Car
Coverage	Basic	Basic	Extended	Extended	Premium	Premium
Months Since Driving	Chủ yếu <=24 tháng	Chủ yếu >=36 tháng	Chủ yếu <=24 tháng	Chủ yếu >=36 tháng	Chủ yếu >=36 tháng	Chủ yếu <=24 tháng
Age	Chủ yếu <=35 tuổi	>35 tuổi	Chủ yếu <=35 tuổi	>35 tuổi	>35 tuổi	Chủ yếu <=35 tuổi
Monthly Premium Auto	**100$ - 199$**	**61$ - 119$**	**121$ - 249$**	**76$ - 139$**	**101$ - 119$**	**140$ - 298$**

3.5. Xây dựng mô hình

3.5.1. Xây dựng mô hình

3.5.1.1. Chuẩn bị dữ liệu

– Ghép 3 bảng nhỏ được chia theo 3 gói lại với nhau (đã có giá trị label cụm)

– Xóa cột *Customer* và chuyển các giá trị định tính về dạng định lượng bằng phương pháp get_dummies()

```
basic['Cluster']=basic['Cluster'].replace([0,1],[0,1])
extended['Cluster']=extended['Cluster'].replace([0,1],[2,3])
premium['Cluster']=premium['Cluster'].replace([0,1],[4,5])
df_model=pd.concat([basic,extended,premium],ignore_index=True)
df_model = df_model.drop(columns=['Customer'])
do_dummy_cols = ['Vehicle Class','Coverage', 'Marital Status','Location Code','EmploymentStatus','State']
df_model= pd.get_dummies(df_model, columns=do_dummy_cols)
df_model = df_model.replace([True,False],[1,0])
df_model.info()
```
✓ 0.1s

Hình 3.21: Chuẩn bị dữ liệu để xây dựng mô hình

– Kết quả thu được:

```
RangeIndex: 9134 entries, 0 to 9133
Data columns (total 30 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Months Since Driving           9134 non-null   int64
 1   Age                            9134 non-null   int64
 2   Monthly Premium Auto           9134 non-null   int64
 3   Months Since Last Claim        9134 non-null   int64
 4   Cluster                        9134 non-null   int32
 5   Vehicle Class_Four-Door Car    9134 non-null   int64
 6   Vehicle Class_Luxury Car       9134 non-null   int64
 7   Vehicle Class_Luxury SUV       9134 non-null   int64
 8   Vehicle Class_SUV              9134 non-null   int64
 9   Vehicle Class_Sports Car       9134 non-null   int64
 10  Vehicle Class_Two-Door Car     9134 non-null   int64
 11  Coverage_Basic                 9134 non-null   int64
 12  Coverage_Extended              9134 non-null   int64
 13  Coverage_Premium               9134 non-null   int64
 14  Marital Status_Divorced        9134 non-null   int64
 15  Marital Status_Married         9134 non-null   int64
 16  Marital Status_Single          9134 non-null   int64
 17  Location Code_Rural            9134 non-null   int64
 18  Location Code_Suburban         9134 non-null   int64
 19  Location Code_Urban            9134 non-null   int64
...
 28  State_Oregon                   9134 non-null   int64
 29  State_Washington               9134 non-null   int64
dtypes: int32(1), int64(29)
memory usage: 2.1 MB
```

Hình 3.22: Kết quả thu được sau khi chuẩn bị dữ liệu

3.5.1.2. Xây dựng mô hình phân lớp

– Mục đích: Phân loại khách hàng, để xem khách hàng nên được phân vào nhóm nào.

– Chuẩn bị dữ liệu với các cột đầu vào là *Vehicle Class, Coverage, Marital Status, Location Code, Months Since Driving, Age, EmploymentStatus, Months Since Last Claim, State* gán cho X

```
y = df_model['Cluster']
X = df_model.drop(columns=['Monthly Premium Auto','Cluster'])
X
```

0.0s Python

	Months Since Driving	Age	Months Since Last Claim	Vehicle Class_Four-Door Car	Vehicle Class_Luxury Car	Vehicle Class_Luxury SUV	Vehicle Class_SUV	Vehicle Class_Sports Car	Vehicle Class_Two-Door Car	Coverage_Basic	...	EmploymentStatus_Disabled	EmploymentStatus_Employed	Employmen
0	236	41	32	0	0	0	0	0	1	1	...	0	1	
1	12	23	18	0	0	0	1	0	0	1	...	0	0	
2	643	70	12	1	0	0	0	0	0	1	...	0	1	
3	42	43	14	0	0	0	0	0	1	1	...	0	1	
4	514	60	0	1	0	0	0	0	0	1	...	0	1	

```
print('Shape X: ',X.shape)
print('Shape Y: ',y.shape)
```

0.0s Python Python

```
Shape X:  (9134, 28)
Shape Y:  (9134,)
```

Hình 3.23: Tách dữ liệu X, Y cho mô hình phân lớp

– Sử dụng phương pháp standardization để chuẩn hóa dữ liệu đầu vào sau đó chia dữ liệu thành 2 bộ train và test với tỷ lệ là 7:3

```
sc = StandardScaler()
X= sc.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

0.0s

Hình 3.24: Chia dữ liệu Train, Test

– Xây dựng hàm multiclass_roc_auc_score() để tính chỉ số ROC_AUC

– Xây dựng hàm model_eval() để xuất ra các chỉ số đánh giá

```
def multiclass_roc_auc_score(y_test, y_pred, average="macro"):
    lb = LabelBinarizer()
    lb.fit(y_test)
    y_test = lb.transform(y_test)
    y_pred = lb.transform(y_pred)
    return roc_auc_score(y_test, y_pred, average=average)
def model_eval(clf, y_test, y_pred):
    print(clf)
    print('ROC_AUC_Score:', multiclass_roc_auc_score(y_test, y_pred))
    print('accuracy: ',accuracy_score(y_test, y_pred))
    print('precision: ',precision_score(y_test, y_pred, average = 'macro'))
    print(classification_report(y_test, y_pred))
    cm=confusion_matrix(y_test, y_pred, labels=y_test.unique())
    disp = ConfusionMatrixDisplay(cm, display_labels=y_test.unique())
    disp.plot(cmap='cividis')
```

0.1s

Hình 3.25: Xây dựng hàm model_eval()

– Xây dựng mô hình với XGBClassifier

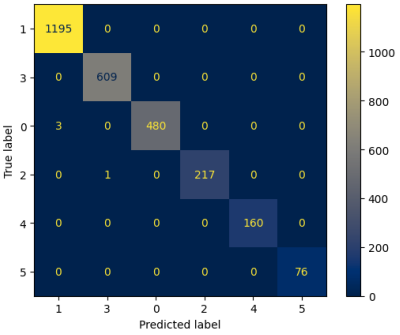
```
xgb_clf = XGBClassifier(n_estimators=500, max_depth=1, max_leaves=2, random_state=0)
xgb_clf.fit(X_train, y_train)
y_pred = xgb_clf.predict(X_test)
clf = 'XGBoost'
model_eval(clf, y_test, y_pred)
```

✓ 2.3s Python

```
XGBoost
ROC_AUC_Score: 0.9988993440974417
accuracy:  0.9985406785844583
precision:  0.9993094136858334
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       483
           1       1.00      1.00      1.00      1195
           2       1.00      1.00      1.00       218
           3       1.00      1.00      1.00       609
           4       1.00      1.00      1.00       160
           5       1.00      1.00      1.00        76

    accuracy                           1.00      2741
   macro avg       1.00      1.00      1.00      2741
weighted avg       1.00      1.00      1.00      2741
```



Hình 3.26: Kết quả của mô hình XGB Classifier

– Xây dựng mô hình với GaussianNB

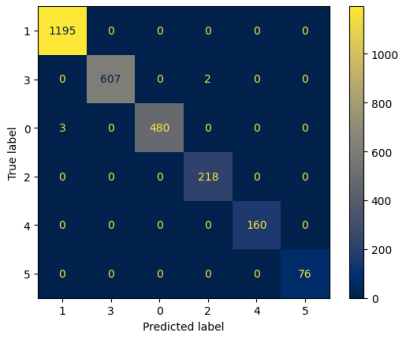
```
from sklearn.naive_bayes import GaussianNB
nv = GaussianNB()
nv=nv.fit(X_train, y_train)
y_pred = nv.predict(X_test)
clf = 'naive bayes classifier'
model_eval(clf, y_test, y_pred)
```

✓ 0.5s Python

```
naive bayes classifier
ROC_AUC_Score: 0.998980962411688
accuracy:  0.9981758482305728
precision:  0.9980674862143978
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       483
           1       1.00      1.00      1.00      1195
           2       0.99      1.00      1.00       218
           3       1.00      1.00      1.00       609
           4       1.00      1.00      1.00       160
           5       1.00      1.00      1.00        76

    accuracy                           1.00      2741
   macro avg       1.00      1.00      1.00      2741
weighted avg       1.00      1.00      1.00      2741
```



Hình 3.27: Kết quả của mô hình Navie Bayes

– Xây dựng mô hình với Neural Network

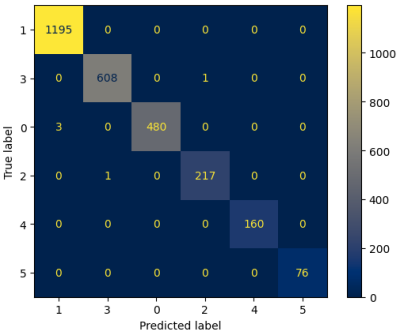
```
from sklearn.neural_network import MLPClassifier
mlp= MLPClassifier(hidden_layer_sizes=(100,50), max_iter=1000)
mlp=mlp.fit(X_train, y_train)
y_pred = mlp.predict(X_test)
clf = 'neural network classifier'
model_eval(clf, y_test, y_pred)
```

✓ 3.2s Python

```
neural network classifier
ROC_AUC_Score: 0.9987294782914286
accuracy:  0.9981758482305728
precision:  0.9985444390481998
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       483
           1       1.00      1.00      1.00      1195
           2       1.00      1.00      1.00       218
           3       1.00      1.00      1.00       609
           4       1.00      1.00      1.00       160
           5       1.00      1.00      1.00        76

    accuracy                           1.00      2741
   macro avg       1.00      1.00      1.00      2741
weighted avg       1.00      1.00      1.00      2741
```



Hình 3.28: Kết quả của mô hình Neural Network

– Qua các chỉ số đánh giá, mô hình XGB Classifier cho kết quả đánh giá tốt nhất trong 3 mô hình.

3.5.1.3. Xây dựng mô hình hồi quy dự đoán

– Mục đích: Dự đoán chính xác chi phí bảo hiểm mà người mua bảo hiểm phải bỏ ra.

– Xây dựng:

+ Xây dựng mô hình trên bộ dữ liệu tổng hợp:

- Thiết lập giá trị đầu vào và đầu ra của mô hình sau đó chuẩn hóa dữ liệu bằng phương pháp standardization
- Chia dữ liệu thành 2 bộ train test với tỷ lệ 85:15
- Fit dữ liệu train với mô hình XGBRegressor()
- Kết quả thấy được R Squared = 0,96 gần với 1 nên thấy được mô hình này phù hợp tốt với dữ liệu và giải thích được 96% phương sai của biến phụ thuộc

```
from pdpbox import pdp
import matplotlib.pyplot as plt
y = df_model['Monthly Premium Auto']
X = df_model.drop(columns=['Monthly Premium Auto','Cluster'])
sc = StandardScaler()
X= sc.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, random_state=0)
xgb_r = XGBRegressor()
xgb_r.fit(X_train,y_train)
y_pred = xgb_r.predict(X_test)
print(f'R Squared Score of XGBRegressor: {r2_score(y_pred, y_test)}')
```

✓ 0.8s

```
R Squared Score of XGBRegressor: 0.9621639160680483
```

Hình 3.29: Xây dựng mô hình hồi quy trên bộ dữ liệu tổng hợp

+ Tuy nhiên giá dự đoán chưa thực sự đúng với từng nhóm đã được phân cụm nên cần phải chia ra thành 6 mô hình với 6 bộ dữ liệu.

+ Xây dựng mô hình trên 6 bộ dữ liệu theo nhóm với hàm xgb():

- Thiết lập giá trị đầu vào và đầu ra của mô hình sau đó chuẩn hóa dữ liệu bằng phương pháp standardization
- Chia dữ liệu thành 2 bộ train test với tỷ lệ 85:15
- Fit dữ liệu train với mô hình XGBRegressor()
- Xuất ra các chỉ số MAE, MSE và R^2 để đánh giá

```python
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
def xgb(a):
    print(a)
    y = df_model[df_model['Cluster']==a]['Monthly Premium Auto']
    X = df_model[df_model['Cluster']==a].drop(columns=['Monthly Premium Auto','Cluster'])
    # print(len(X.columns))
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, random_state=0)
    xgb_r = XGBRegressor(n_estimators=100, max_depth=5, learning_rate=0.05)
    xgb_r.fit(X_train,y_train)
    y_pred = xgb_r.predict(X_test)
    mae = mean_absolute_error(y_test, y_pred) # Mean Absolute Error
    mse = mean_squared_error(y_test, y_pred) # Mean Squared Error
    rmse = mean_squared_error(y_test, y_pred, squared=False) # Root Mean Squared Error
    r2 = r2_score(y_test, y_pred) # Coefficient of Determination
    print("Mean Absolute Error:", mae)
    print("Mean Squared Error:", mse)
    # print("Root Mean Squared Error:", rmse)
    print('R Squared Score:',r2)
```

✓ 0.0s

Hình 3.30: Xây dựng mô hình hồi quy

- Sử dụng hàm xgb(), thu được kết quả:

```
0
Mean Absolute Error: 4.7465874902133285
Mean Squared Error: 31.270351280818556
R Squared Score: 0.9547140210618511
1
Mean Absolute Error: 3.4489571021882113
Mean Squared Error: 16.188214047728085
R Squared Score: 0.4990399792857726
2
Mean Absolute Error: 6.008815107674434
Mean Squared Error: 65.72198283879472
R Squared Score: 0.9392987869930517
3
Mean Absolute Error: 5.764957685728331
Mean Squared Error: 46.209037513465844
R Squared Score: 0.23672672691939545
4
Mean Absolute Error: 4.887103201329023
Mean Squared Error: 31.5665125858879
R Squared Score: 0.010826548746861953
5
Mean Absolute Error: 15.287636421822214
Mean Squared Error: 390.7687038286539
R Squared Score: 0.8116726282489014
```

Hình 3.31: Kết quả của mô hình hồi quy

- Chỉ số R^2 ở mỗi cụm đều lớn hơn 0 tuy nhiên có một số cụm không phù hợp (cụm 3,4), giải thích được số phần trăm phương sai của biến phụ thuộc dưới 50%.
- Chỉ số MAE ở mỗi cụm khá cao nhưng ở mức chấp nhận được.

3.5.2. Kiểm tra mô hình

– Đầu tiên, gán biến z với dataframe biểu thị các giá trị trung bình, độ lệch chuẩn, min, max, … của các cột đầu vào

```
X = df_model.drop(columns=['Monthly Premium Auto','Cluster'])
z=X.describe()
z
```

✓ 0.1s Python

	Months Since Driving	Age	Months Since Last Claim	Vehicle Class_Four-Door Car	Vehicle Class_Luxury Car	Vehicle Class_Luxury SUV	Vehicle Class_SUV	Vehicle Class_Sports Car	Vehicle Class_Two-Door Car	Coverage_Basic	...	EmploymentStatus_Disabled	EmploymentStatus_Em
count	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	...	9134.00000	9134.
mean	177.558791	45.717320	15.097000	0.505912	0.017845	0.020145	0.196628	0.052989	0.206481	0.609591	...	0.04434	0.
std	149.118392	15.674646	10.073257	0.499992	0.132397	0.140502	0.397470	0.224023	0.404802	0.487869	...	0.20586	0.
min	0.000000	16.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00000	0.
25%	52.000000	33.000000	6.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00000	0.
50%	141.000000	47.000000	14.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	...	0.00000	1.
75%	272.000000	59.000000	23.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	...	0.00000	1.
max	707.000000	75.000000	35.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.00000	1.

8 rows × 28 columns

Hình 3.32: Mô tả dữ liệu xây dựng mô hình

– Xây dựng hàm TEST() với mục đích là chuẩn hóa giá trị nhập vào.

```
def TEST(test,z):
    listtest=[]
    for k in X.columns:
        l=k.split('_')
        if len(l)==1:
            x=(test[l[0]]-z[k][1])/z[k][2]
        else:
            if test[l[0]]==l[1]:
                x=(1-z[k][1])/z[k][2]
            else:
                x=(0-z[k][1])/z[k][2]
        listtest.append(x)
    import numpy as np
    user_input=np.array([listtest])
    return user_input
```

✓ 0.0s

Hình 3.33: Xây dựng hàm TEST()

– Xây dựng hàm ketqua() để in ra kết quả dự đoán.

```
def ketqua(test):
    dl=TEST(test,z)
    gr=xgb_clf.predict(dl)[0]
    if gr==0:
        gr='Basic B1'
        y_pred = xgb_0.predict(TEST(test,z))[0]
    elif gr==1:
        gr='Basic B2'
        y_pred = xgb_1.predict(TEST(test,z))[0]
    elif gr==2:
        gr='Extended E1'
        y_pred = xgb_2.predict(TEST(test,z))[0]
    elif gr==3:
        gr='Extended E2'
        y_pred = xgb_3.predict(TEST(test,z))[0]
    elif gr==4:
        gr='Premium P1'
        y_pred = xgb_4.predict(TEST(test,z))[0]
    elif gr==5:
        gr='Premium P2'
        y_pred = xgb_5.predict(TEST(test,z))[0]
    y_p=round(y_pred)
    print(f'Gói bảo hiểm phù hợp: {gr}')
    print(f'Số tiền cần phải trả cho công ty là: {y_p}$')
✓ 0.0s
```

Hình 3.34: Xây dựng hàm ketqua()

– Với dữ liệu test như dưới đây, thì thu được kết quả:

```
test={'Months Since Driving':100,'Vehicle Class':'Two-Door Car','Age':24,'Coverage':'Extended','Marital Status':'Married',
      'Location Code':'Suburban','EmploymentStatus':'Employed','Months Since Last Claim':48,'State':'Washington'}
ketqua(test)
✓ 0.0s
```

```
Gói bảo hiểm phù hợp: Extended E1
Số tiền cần phải trả cho công ty là: 143$
```

Hình 3.35: Kết quả test dữ liệu mới

Gói bảo hiểm phù hợp: **Extended E1**

Số tiền cần phải trả cho công ty là: **143$**

- Giải thích kết quả:

  + Gói bảo hiểm:

    Vì Coverage là Extended nên sẽ phù hợp với nhóm E1 và E2

    Vì Vehicle là Two-Door Car nên sẽ phù hợp với E2 nhưng Age là 24 và State là Washington (phổ biến hơn ở E1) nên sẽ phù hợp với E1

    ⇒ E1 là gói bảo hiểm phù hợp nhất

  + Số tiền cần phải trả: Với mức giá dự đoán 143$ phù hợp với gói E1 vì giá của gói này nằm trong khoảng từ 121$ - 249$.

3.6. Xây dựng ứng dụng

– Xây dựng ứng dụng:

+ Cài đặt thư viện flask và import các thư viện cần dùng.

+ Xây dựng hàm TEST() để chuẩn hóa dữ liệu đầu vào.

+ Tiếp theo sẽ sử dụng @app.route() để xử lý các yêu cầu HTTP đến endpoint '/' của web.

  - Nếu yêu cầu HTTP là GET, hàm main() sẽ trả về trang web "index.html" bằng cách sử dụng hàm render_template() của Flask để render trang HTML từ các mẫu được định nghĩa trước đó.

  - Nếu yêu cầu HTTP là POST, hàm main() sẽ lấy giá trị của các trường input trong "index.html" từ form được gửi kèm theo yêu cầu HTTP bằng cách sử dụng đối tượng request của Flask. Sau đó, hàm sử dụng giá trị này để gọi hàm TEST() để dự đoán chi phí bảo hiểm mà khách hàng bỏ ra, và trả về kết quả bằng cách sử dụng hàm render_template() để render trang HTML "results.html" với giá trị chi phí đã dự đoán, sau đó được truyền vào dưới dạng các biến trong jinja2 template engine của Flask.

```
from flask import Flask, render_template, request
import pandas as pd
import pickle
z = pd.read_excel('z.xlsx')
X = pd.read_excel('X.xlsx')
def TEST(test,z):
    listtest=[]
    for k in X.columns:
        l=k.split('_')
        if len(l)==1:
            x=(test[l[0]]-z[k][1])/z[k][2]
        else:
            if test[l[0]]==l[1]:
                x=(1-z[k][1])/z[k][2]
            else:
                x=(0-z[k][1])/z[k][2]
        listtest.append(x)
    import numpy as np
    user_input=np.array([listtest])
    return user_input
app = Flask(__name__)
@app.route('/', methods=['GET', 'POST'])
def main():
    if request.method == 'GET':
        return render_template('index.html')
    if request.method == 'POST':
        age= int(request.form["Age"])
        mar= request.form["Marital"]
        emp= request.form["Employed"]
        loc= request.form["Location"]
```

```
        sta= request.form["State"]
        veh= request.form["Vehicle"]
        dri= int(request.form["Driving"])
        cla= int(request.form["Claim"])
        cov= request.form["Coverage"]
        test={'Months Since Driving':dri,'Vehicle Class':veh,'Age':age,'Coverage':cov,'Marital Status':mar,
       'Location Code':loc,'EmploymentStatus':emp,'Months Since Last Claim':cla,'State':sta}
        dl=TEST(test,z)
        class_model = pickle.load(open('classifier.pkl', 'rb'))
        gr=class_model.predict(dl)[0]
        if gr==0:
            reg_model = pickle.load(open('xgb0.pkl', 'rb'))
        elif gr==1:
            reg_model = pickle.load(open('xgb1.pkl', 'rb'))
        elif gr==2:
            reg_model = pickle.load(open('xgb2.pkl', 'rb'))
        elif gr==3:
            reg_model = pickle.load(open('xgb3.pkl', 'rb'))
        elif gr==4:
            reg_model = pickle.load(open('xgb4.pkl', 'rb'))
        elif gr==5:
            reg_model = pickle.load(open('xgb5.pkl', 'rb'))
        price = reg_model.predict(dl)[0]
        return render_template('results.html', pricing=price)

if __name__ == '__main__':
#     # app.run(host="127.0.0.1:5000", port=8080, debug=True)
    app.run(debug=True)
```

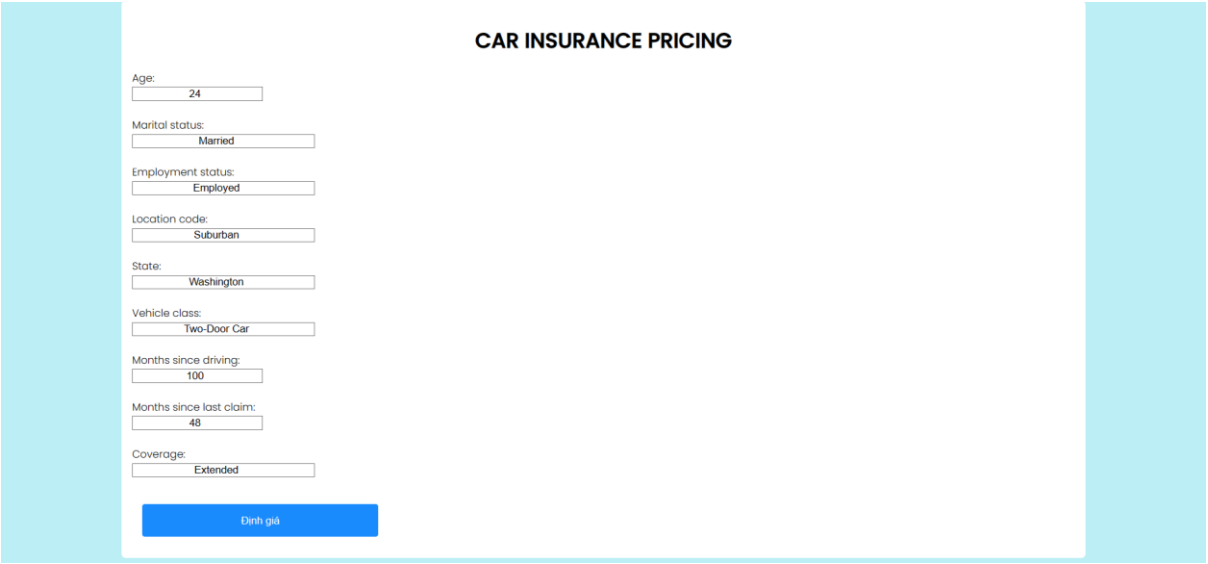
Hình 3.36: Xây dựng ứng dụng html

+ Khi chạy đoạn code trên thì trên màn hình Terminal sẽ xuất hiện như sau:

```
 * Serving Flask app 'app'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 555-652-132
```

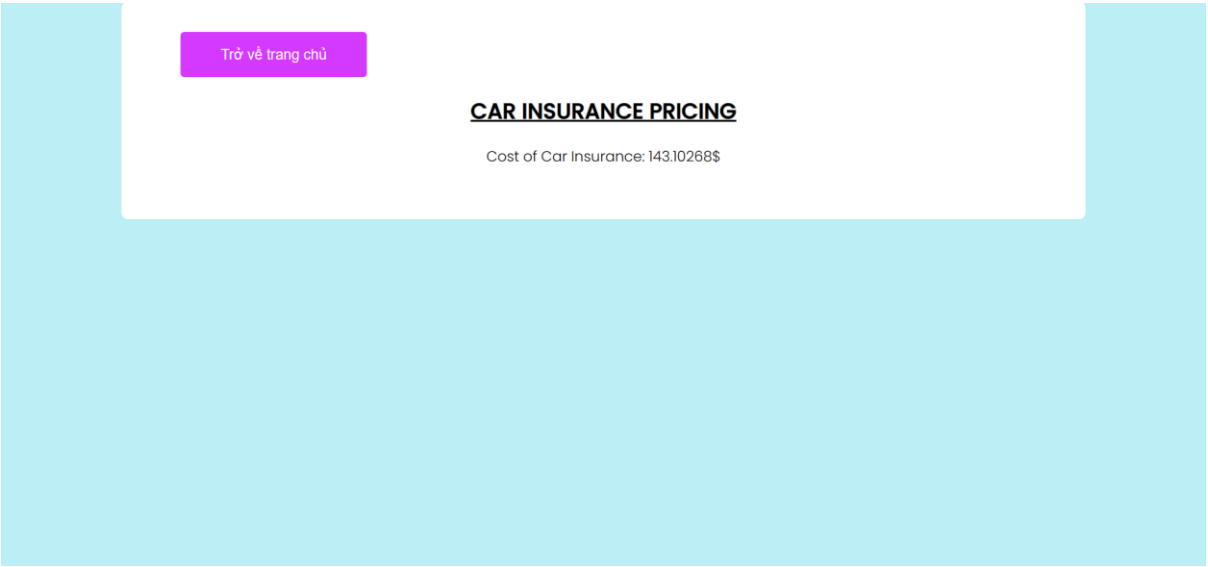
Hình 3.37: Kết quả sau khi chạy code ứng dụng

+ Khi đó chỉ cần đi theo đường link http://127.0.0.1:5000 để đến với app đề xuất sách. Đây là giao diện khi vào trang



Hình 3.38: Giao diện của ứng dụng

+ Ở các ô thì nhập các đặc điểm của khách hàng như tuổi, tình trạng hôn nhân, tình trạng việc làm,... Sau đó click "*Định giá"* thì trang sẽ tự động chuyển sang trang như hình dưới:



Hình 3.39: Giao diện kết quả của ứng dụng

PHẦN 4. KẾT LUẬN

Định giá bảo hiểm xe ô tô là một phần quan trọng trong ngành bảo hiểm, đóng vai trò quyết định trong việc xác định mức độ rủi ro và tính phí bảo hiểm hợp lý cho khách hàng. Các yếu tố quan trọng được sử dụng để định giá bảo hiểm xe ô tô, bao gồm thông tin về người sử dụng ô tô, lịch sử lái xe, khu vực giao thông, điều kiện đường và nhiều yếu tố khác. Đồng thời, chúng tôi đã thảo luận về vai trò của dữ liệu lớn và phân tích dữ liệu trong quá trình định giá bảo hiểm, đưa ra những ứng dụng cụ thể của Big Data trong lĩnh vực này. Hơn nữa, sự phát triển của big data đã cung cấp cơ hội cho việc tạo ra các dịch vụ bảo hiểm cá nhân hóa và cung cấp trải nghiệm khách hàng tốt hơn. Các công ty bảo hiểm có thể hiểu rõ hơn về nhu cầu và sở thích của khách hàng thông qua phân tích dữ liệu, từ đó đề xuất các sản phẩm và dịch vụ tùy chỉnh. Điều này mang lại sự hài lòng và trung thực trong quan hệ giữa công ty bảo hiểm và khách hàng.

Tóm lại, việc áp dụng big data trong định giá bảo hiểm xe ô tô đã đem lại nhiều lợi ích đáng kể. Công nghệ này không chỉ cải thiện quy trình định giá bảo hiểm mà còn tăng cường khả năng dự đoán rủi ro, ngăn chặn gian lận và cung cấp trải nghiệm tốt hơn cho khách hàng. Với tiềm năng không ngừng phát triển, big data tiếp tục đóng vai trò quan trọng trong lĩnh vực bảo hiểm xe ô tô và đóng góp vào sự phát triển và tăng trưởng của ngành này. Với khả năng tận dụng triệt để nguồn tài nguyên này, ngành bảo hiểm xe ô tô sẽ tiếp tục đi đầu trong việc ứng dụng công nghệ và sáng tạo để mang lại những dịch vụ bảo hiểm tốt nhất cho khách hàng và đáp ứng nhu cầu ngày càng cao của thị trường bảo hiểm trong tương lai.

TÀI LIỆU THAM KHẢO

"Average Cost of Car Insurance: 2023 Rate and Price Factors," [Online].

[1] Url: https://www.marketwatch.com/guides/insurance-services/average-cost-of-car-insurance/.

"How to Become a Data Scientist in Insurance," [Online].

[2] Url: https://365datascience.com/career-advice/how-to-become-a-data-scientist-in-insurance/.