ĐẠI HỌC ĐÀ NẴNG

TRƯỜNG ĐẠI HỌC KINH TẾ ĐÀ NẴNG

KHOA THƯƠNG MẠI ĐIỆN TỬ

CAPSTONE PROJECT 2

BÀI BÁO CÁO

DỰ ĐOÁN CHẤP NHẬN CÁC ĐỀ NGHỊ CHO VAY TIÊU DÙNG CÁ NHÂN CỦA CÔNG TY TÀI CHÍNH

Giảng viên hướng dẫn : Th.s Trần Văn Lộc

Nhóm: 1

Sinh viên thực hiện:

Lê Thị Oanh Mã SV: 201124029237

Nguyễn Thị Mỹ Thúy Mã SV: 201124029250

Phan Thị Thảo Ngân Mã SV: 201124029225

Đà Nẵng, ngày 13 tháng 5 năm 2023

LỜI CAM ĐOAN

Chúng em xin cam đoan đề án dưới đây là công trình nghiên cứu của chúng em dưới sự hướng dẫn của Thầy Trần Văn Lộc. Những nhận định được nêu ra trong đồ án cũng là kết quả từ sự nghiên cứu trực tiếp, nghiêm túc, độc lập của chúng em dựa và các cơ sở tìm kiếm, hiểu biết và nghiên cứu tài liệu khoa học hay bản dịch khác đã được công bố. Đề án vẫn sẽ giúp đảm bảo được tính khách quan, trung thực và khoa học.

LỜI CẢM ƠN

Trước hết, chúng em xin bày tỏ tình cảm và lòng biết ơn chân thành đến GVHD – Thầy Trần Văn Lộc đã chỉ bảo, góp ý, bổ sung và hướng dẫn tận tình để nhóm có thể rút kinh nghiệm từ những sai sót và phát huy thế mạnh của bản thân và việc hoàn thành đề án.

Để hoàn thành bài đề án thực hành 1 này, chúng em đã nhận được rất nhiều sự giúp đỡ, đóng góp ý kiến từ giáo viên hướng dẫn. Chúng em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc tới các quý thầy cô trong Khoa Thương mại điện tử đã trang bị cho chúng em những kiến thức nền tảng cơ bản đến chuyên ngành làm tiền đề để chúng em thực hiện và hoàn thành đúng thời hạn đề án.

Một lần nữa, chúng em xin chân thành cảm ơn!

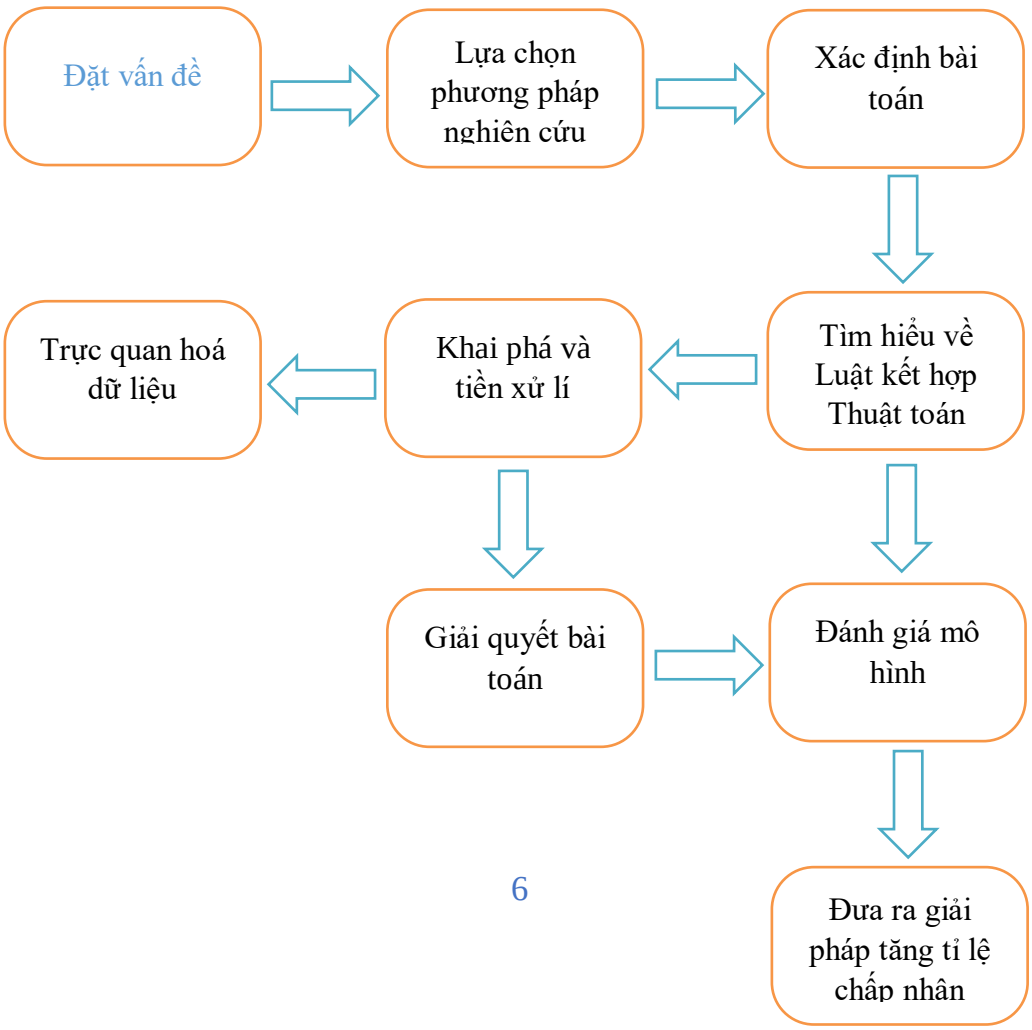
MỤC LỤC

CHƯƠNG I: GIỚI THIỆU ĐỀ TÀI

1.1 Câu chuyện cho vay tiêu dùng cá nhân

Trong ngành công ty tài chính và công ty tài chính, việc cung cấp các sản phẩm vay tiêu dùng cá nhân đã trở thành một lĩnh vực quan trọng, mang lại lợi ích cho cả khách hàng và tổ chức tài chính. Hiểu rõ nhu cầu ngày càng tăng về tiền mặt và nhu cầu tiêu dùng cá nhân, công ty tài chính và công ty tài chính đang đưa ra những nỗ lực để phát triển và tăng cường doanh số của sản phẩm này, việc đưa ra các quyết định về chấp nhận hay từ chối các đề nghị vay tiêu dùng cá nhân đóng vai trò quan trọng. Để nâng cao hiệu quả và đạt được mục tiêu kinh doanh, công ty tài chính cần phải dự đoán một cách chính xác khả năng chấp nhận hay từ chối từng đề nghị vay tiêu dùng cá nhân. Trong bài báo cáo này, chúng ta sẽ tìm hiểu về phương pháp dự đoán chấp nhận các đề nghị cho vay tiêu dùng cá nhân của một công ty tài chính, nhằm tối ưu hóa quy trình và đạt được kết quả tốt nhất trong việc xử lý đề nghị vay.

1.2 Ý tưởng thực hiện



1.3 Mục tiêu - tầm quan trọng của dự đoán

Trường hợp công ty tài chính có một dữ liệu khách hàng với các đặc điểm khác nhau của khách hàng. Ban quản lý đã xây dựng một sản phẩm mới - Khoản vay cá nhân và thực hiện một chiến dịch nhỏ nhằm bán Sản phẩm mới cho khách hàng của họ. Sau một thời gian, 9% khách hàng vay tiêu dùng cá nhân từ công ty tài chính.

Mục tiêu xây dựng đề tài: Bán thêm các sản phẩm Vay tiêu dùng cá nhân cho khách hàng của công ty tài chính. Để đưa ra các chiến dịch nhằm tiếp thị mục tiêu tốt hơn để tăng tỷ lệ thành công với ngân sách tối thiểu. Để xác định những khách hàng tiềm năng có xác suất mua khoản vay cao hơn

Các chiến dịch cho khoản vay cá nhân không chỉ là công cụ thông báo cho những khách hàng tiềm năng về lợi ích của khoản vay cá nhân bên cạnh đó cũng xây dựng niềm tin với khách hàng bằng cách nêu bật tính minh bạch và bảo mật các quy trình cho vay.

Chiến dịch có thể nhắm mục tiêu nhân khẩu học hoặc phân khúc cụ thể, chẳng hạn như thế hệ thiên niên kỷ hoặc người vay lần đầu và sử dụng các kênh tiếp thị khác nhau như mạng xã hội, tiếp thị qua email và quảng cáo được nhắm mục tiêu để tiếp cận họ. Mục tiêu cuối cùng là tăng số lượng đơn xin vay cá nhân và phê duyệt, điều này sẽ giúp người vay đạt được các mục tiêu tài chính của họ đồng thời thúc đẩy tăng trưởng cho tổ chức cho vay.

Dự đoán vấn đề này là một công việc phụ đối với các ngân hàng, nhưng nếu họ có thể dự đoán khách hàng nào sẽ chấp nhận đề nghị cho vay cá nhân, thì họ có thể kiếm được lợi nhuận tốt hơn.

1.4 Công ty tài chính

Công ty tài chính là một loại hình tổ chức tín dụng phi ngân hàng chịu sự điều chỉnh của Luật các tổ chức tín dụng.

Tổ chức tín dụng phi ngân hàng là những tổ chức tín dụng được thực hiện một số các hoạt động ngân hàng như là nội dung kinh doanh thường xuyên nhưng không được nhận tiền gửi không kỳ hạn, không được làm dịch vụ thanh toán.

1.4.1 Tài chính công ty tài chính là gì?

Tài chính - công ty tài chính là một lĩnh vực trong tài chính, tập trung vào các hoạt động tài chính của các tổ chức tín dụng và các hoạt động công ty tài chính. Tài chính công ty tài chính bao gồm các hoạt động như quản lý tài sản và khoản nợ của công ty tài chính, định giá và quản lý rủi ro, cung cấp dịch vụ tài chính cho khách hàng…

1.4.2 Tài chính tiêu dùng là gì?

Tài chính tiêu dùng là các hoạt động tài chính nhằm mua sắm, sử dụng các mặt hàng và dịch vụ để đáp ứng nhu cầu cá nhân trong cuộc sống hàng ngày. Bao gồm các khoản vay để mua sắm những món đồ lớn như ô tô hoặc nhà, hoặc các khoản vay nhỏ như thẻ tín dụng, vay tiêu dùng các mặt hàng gia đình như tivi, tủ lạnh, điện thoại di động, đồ gia dụng và nhu yếu phẩm hàng ngày. Tài chính tiêu dùng rất quan trọng để hỗ trợ các hoạt động mua sắm và tiêu dùng của mỗi người trong xã hội.

1.4.3 Đặc điểm tài chính tiêu dùng

- Các khoản vay có quy mô thấp
- Lãi suất cao hơn các khoản vay thường

1.4.4 Các sản phẩm tài chính tiêu dùng

Đa dạng, đáp ứng mọi nhu cầu tiêu dùng của người dân. Từ các sản phẩm có giá trị nhỏ như xe máy, xe đạp điện, đồ gia dụng, nội thất, điện máy,... đến các sản phẩm có giá trị lớn như vay mua nhà, mua ô tô,…

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Phân tích EDA

Phân tích EDA là quá trình biến dữ liệu thô thành các biểu diễn trực quan. Thông thường, những hình dung đó ở dạng biểu đồ và đồ thị. Mục đích của phân tích EDA là làm cho dữ liệu dễ hiểu và nhanh hơn, ngay cả đối với những người không được đào tạo về phân tích hoặc thường giỏi về các con số.

2.1.1 Lợi ích của phân tích EDA

- Cung cấp cho người đọc phương tiện để nhanh chóng tiếp thu thông tin, cải thiện hiểu biết sâu sắc và đưa ra quyết định nhanh hơn.
- Cung cấp một phương tiện dễ dàng để phân phối thông tin, mang đến cho người dùng nhiều cơ hội hơn để chia sẻ những hiểu biết của họ với mọi người tham gia vào dự án.
- Tăng cường hiểu biết về các bước mà một tổ chức phải thực hiện để cải thiện chính mình.
- Loại bỏ nhu cầu phụ thuộc quá mức vào các nhà khoa học dữ liệu vì nó dễ tiếp cận và dễ hiểu hơn

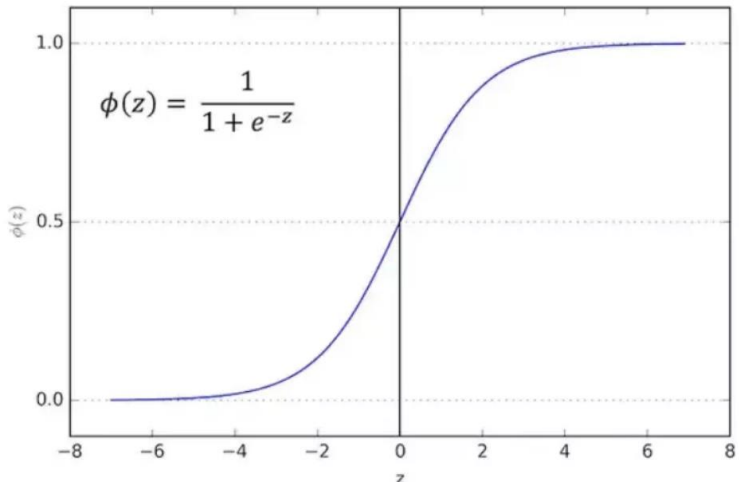
2.1.2 Các bước phân tích EDA

- Bước 1: **Xác định yêu cầu**
- Bước 2: **Thu thập dữ liệu**
- Bước 3: **Phân tích dữ liệu**
- Bước 4: **Phân tích EDA**

2.2 Mô hình hồi quy Log

2.2.1 Hồi quy Log là gì?

- Log là một phương pháp phân loại trong Machine Learning, được sử dụng để dự đoán kết quả của một biến phụ thuộc nhị phân dựa trên các biến độc lập.
- Mục tiêu của Log là tìm ra một hàm số logistic (Sigmoid function) để ước lượng xác suất của biến phụ thuộc đạt giá trị 1 hay 0. Hàm số logistic có dạng S-shaped curve và giá trị đầu ra nằm trong khoảng [0,1].



Hình 2.1: Hồi quy Log

- Hàm liên tục và luôn đưa ra giá trị trong khoảng (0, 1)
- Có đạo hàm tại mọi điểm nên có thể dùng gradient descent
- Log là một trong những phương pháp phân loại đơn giản và hiệu quả, thường được sử dụng trong các bài toán phân loại nhị phân, ví dụ như phân loại email spam, phân loại tin tức, phát hiện bệnh, v.v.

2.2.2 Lợi ích của việc sử dụng hồi quy Log

- Tính đơn giản:

Các mô hình hồi quy logistic ít phức tạp về mặt toán học hơn các phương pháp ML khác. Do đó, bạn có thể triển khai chúng ngay cả khi đội ngũ của bạn không ai có chuyên môn sâu về ML.

- Tốc độ:

Các mô hình hồi quy logistic có thể xử lý khối lượng lớn dữ liệu ở tốc độ cao bởi chúng cần ít khả năng điện toán hơn, chẳng hạn như bộ nhớ và sức mạnh xử lý. Điều này khiến các mô hình hồi quy logistic trở nên lý tưởng đối với những tổ chức đang bắt đầu với các dự án ML để đạt được một số thành tựu nhanh chóng.

- Sự linh hoạt:

Bạn có thể sử dụng hồi quy logistic để tìm đáp án cho các câu hỏi có hai hoặc nhiều kết quả hữu hạn. Bạn cũng có thể sử dụng phương pháp này để xử lý trước dữ liệu.

- Khả năng hiển thị

Phân tích hồi quy logistic cung cấp cho nhà phát triển khả năng nhìn nhận các quy trình phần mềm nội bộ rõ hơn so với các kỹ thuật phân tích dữ liệu khác. Khắc phục sự cố và sửa lỗi cũng trở nên dễ dàng hơn do các phép toán ít phức tạp hơn.

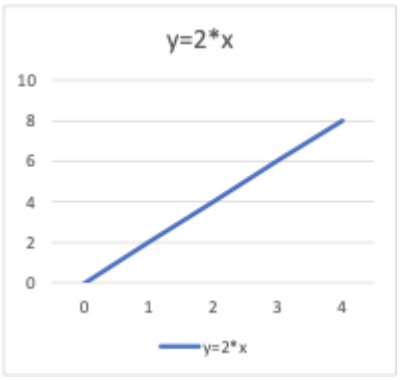
2.2.3 Phân tích hồi quy Log hoạt động như thế nào?

Để hiểu rõ về mô hình hồi quy logistic, trước tiên chúng ta phải hiểu các phương trình và biến.

- Phương trình:

Trong toán học, phương trình cho ta mối quan hệ giữa hai biến: x và y. Bạn có thể sử dụng các phương trình hoặc hàm này để vẽ đồ thị theo trục x và trục y bằng cách

nhập các giá trị khác nhau của x và y. ***Ví dụ:*** nếu bạn vẽ đồ thị cho hàm y = 2*x, bạn sẽ có một đường thẳng như hình dưới đây. Do đó hàm này còn được gọi là hàm tuyến tính.



Hình 2.2: Hàm tuyến tính

– Biến:

Trong thống kê, biến là các yếu tố dữ liệu hoặc thuộc tính có giá trị khác nhau. Bất kỳ phân tích nào cũng có một số biến nhất định là biến độc lập hoặc biến giải thích. Những thuộc tính này là nguyên nhân của một kết quả. Các biến khác là biến phụ thuộc hoặc biến đáp ứng; giá trị của chúng phụ thuộc vào các biến độc lập. Nhìn chung, hồi quy logistic khám phá cách các biến độc lập ảnh hưởng đến một biến phụ thuộc bằng cách xem xét các giá trị dữ liệu lịch sử của cả hai biến.

Trong ví dụ ở trên của chúng tôi, x được gọi là biến độc lập, biến dự đoán hoặc biến giải thích vì nó có một giá trị đã xác định. Y được gọi là biến phụ thuộc, biến kết quả hoặc biến đáp ứng vì giá trị của nó không xác định.

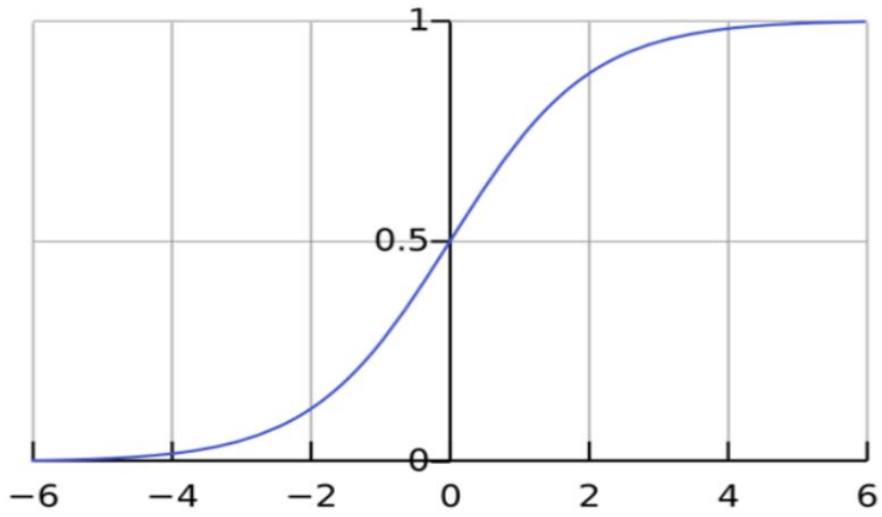
– Hàm hồi quy logistic

Hồi quy logistic là một mô hình thống kê sử dụng hàm logistic, hay hàm logit trong toán học làm phương trình giữa x và y. Hàm logit ánh xạ y làm hàm sigmoid của x.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Hình 2.3: Hàm logit ánh xạ y làm hàm sigmoid của x

Nếu vẽ phương trình hồi quy logistic này, bạn sẽ có một đường cong hình chữ S như hình dưới đây.



Hình 2.4: Phương trình hàm logit ánh xạ y làm hàm sigmoid của x

Như bạn có thể thấy, hàm logit chỉ trả về các giá trị giữa 0 và 1 cho biến phụ thuộc, dù giá trị của biến độc lập là gì. Đây là cách hồi quy logistic ước tính giá trị của biến phụ thuộc. Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc.

– Phân tích hồi quy logistic với nhiều biến độc lập

Trong nhiều trường hợp, nhiều biến giải thích ảnh hưởng đến giá trị của biến phụ thuộc. Để lập mô hình các tập dữ liệu đầu vào như vậy, công thức hồi quy logistic phải giả định mối quan hệ tuyến tính giữa các biến độc lập khác nhau. Bạn có thể sửa đổi hàm sigmoid và tính toán biến đầu ra cuối cùng như sau:

$$y = f(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n)$$

Ký hiệu β đại diện cho hệ số hồi quy. Mô hình logit có thể đảo ngược tính toán các giá trị hệ số này khi bạn cho nó một tập dữ liệu thực nghiệm đủ lớn có các giá trị đã xác định của cả hai biến phụ thuộc và biến độc lập.

– Log của tỷ số odds

Mô hình logit cũng có thể xác định tỷ số thành công trên thất bại hay log của tỷ số odds. Ví dụ: nếu bạn đang chơi poker với bạn bè và thắng bốn ván trên mười ván, tỷ số chiến thắng của bạn là bốn phần sáu, hoặc 4/6, và đó là tỷ số thành công trên thất bại của bạn. Mặt khác, xác suất thắng là 4/10.

Về mặt toán học, tỉ số odds về mặt xác suất của bạn là p/(1 - p) và log của tỷ số odds là log (p/(1 - p)). Bạn có thể biểu diễn hàm logistic bằng log của tỷ số odds như hình dưới đây:

$$\text{Logit Function} = \log\left(\frac{p}{1-p}\right)$$

Hình 2.5: Hàm logit bằng log tỷ số odds

2.2.5 Ưu điểm - nhược điểm của hồi quy Log

– Ưu điểm:

Hồi quy logistic dễ thực hiện hơn nhiều so với các phương pháp khác, đặc biệt là trong Machine Learning: Mô hình Machine Learning có thể được mô tả như một mô tả toán học của một quá trình trong thế giới thực. Quá trình thiết lập mô hình học máy yêu cầu đào tạo và thử nghiệm mô hình. Huấn luyện là quá trình tìm kiếm các mẫu trong dữ liệu đầu vào, để mô hình có thể ánh xạ một đầu vào cụ thể (ví dụ, một hình ảnh) tới một loại đầu ra nào đó, chẳng hạn như một nhãn. Hồi quy logistic dễ đào tạo và triển khai hơn so với các phương pháp khác.

Hồi quy logistic hoạt động tốt đối với các trường hợp tập dữ liệu có thể phân tách tuyến tính: Tập dữ liệu được cho là có thể phân tách tuyến tính nếu có thể vẽ một đường thẳng có thể tách hai lớp dữ liệu khỏi nhau. Hồi quy logistic được sử dụng khi biến Y của bạn chỉ có thể nhận hai giá trị và nếu dữ liệu có thể phân tách tuyến tính, thì việc phân loại nó thành hai lớp riêng biệt sẽ hiệu quả hơn.

Hồi quy logistic cung cấp những hiểu biết hữu ích: Hồi quy logistic không chỉ cho phép đo lường mức độ liên quan của một biến độc lập (tức là (kích thước hệ số), mà còn cho chúng ta biết về hướng của mối quan hệ (tích cực hoặc tiêu cực). Hai biến được cho là có một liên kết tích cực khi sự gia tăng giá trị của một biến số cũng làm tăng giá trị của biến số khác. Ví dụ: bạn càng dành nhiều giờ tập luyện, bạn càng trở nên giỏi hơn trong một môn thể thao cụ thể. Tuy nhiên: Điều quan trọng là phải biết mối tương quan đó Nói cách khác, hồi quy logistic có thể cho bạn thấy rằng có mối tương quan thuận giữa nhiệt độ ngoài trời và doanh số bán hàng, nhưng điều này không nhất thiết có nghĩa là doanh số bán hàng tăng do nhiệt độ.

– Nhược điểm:

Hồi quy logistic không dự đoán được kết quả liên tục. Hãy xem xét một ví dụ để hiểu rõ hơn về hạn chế này. Trong các ứng dụng y tế, hồi quy logistic không thể được sử dụng để dự đoán nhiệt độ của bệnh nhân viêm phổi sẽ tăng cao như thế nào. Điều này là do quy mô đo lường là liên tục (hồi quy logistic chỉ hoạt động khi biến phụ thuộc hoặc biến kết quả là lưỡng phân).

Hồi quy logistic giả định tính tuyến tính giữa biến dự đoán (phụ thuộc) và biến dự báo (độc lập). Tại sao đây là một hạn chế? Trong thế giới thực, rất khó có khả năng các quan sát được phân tách tuyến tính. Hãy tưởng tượng bạn muốn phân loại cây diên vĩ thành một trong hai họ: sentosa hoặc versicolor. Để phân biệt giữa hai loại, bạn sẽ phân biệt kích thước cánh hoa và kích thước đài hoa. Bạn muốn tạo ra một thuật toán để phân loại cây diên vĩ, nhưng thực sự không có sự phân biệt rõ ràng — một cánh hoa kích thước 2cm có thể đủ tiêu chuẩn cho cây trồng cho cả hai loại màu xanh lá và màu sắc. Vì vậy, trong khi dữ liệu có thể phân tách tuyến tính là giả định cho hồi quy logistic, trên thực tế, nó không phải lúc nào cũng thực sự khả thi.

Hồi quy logistic có thể không chính xác nếu kích thước mẫu quá nhỏ. Nếu kích thước mẫu ở mức nhỏ, thì mô hình được tạo ra bằng hồi quy logistic dựa trên số lượng quan sát thực tế nhỏ hơn. Điều này có thể dẫn đến trang bị quá nhiều. Trong thống kê, overfitting là một lỗi mô hình hóa xảy ra khi mô hình quá khớp với một bộ dữ liệu hạn chế vì thiếu dữ liệu đào tạo. Hay nói cách khác, không có đủ dữ liệu đầu vào để mô hình tìm ra các mẫu trong đó. Trong trường hợp này, mô hình không thể dự đoán chính xác kết quả của một tập dữ liệu mới hoặc trong tương lai.

2.3 Random Forest

2.3.1 Random Forest là gì?

Random Forests là thuật toán học có giám sát (supervised learning). Random Forest sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định. Người ta nói rằng càng có nhiều cây thì rừng càng mạnh. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu.

2.3.2 Ý tưởng thuật toán

Mô hình rừng cây được huấn luyện dựa trên sự phối hợp giữa luật kết hợp (ensembling) và quá trình lấy mẫu tái lặp (bootstrapping). Cụ thể thuật toán này tạo ra nhiều cây quyết định mà mỗi cây quyết định được huấn luyện dựa trên nhiều mẫu con khác nhau và kết quả dự báo là bầu cử (voting) từ toàn bộ những cây quyết định. Như vậy một kết quả dự báo được tổng hợp từ nhiều mô hình nên kết quả của chúng sẽ không bị chệch. Đồng thời kết hợp kết quả dự báo từ nhiều mô hình sẽ có phương sai nhỏ hơn so với chỉ một mô hình.

2.3.3 Thuật toán Random Forest hoạt động như thế nào?

- Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
- Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
- Hãy bỏ phiếu cho mỗi kết quả dự đoán.
- Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.

2.4 Naive Bayes

2.4.1 Naive Bayes là gì?

Naive Bayes là một trong nhóm các thuật toán áp dụng định lý Bayes với một giả định khá ngây thơ rằng mọi features đầu vào đều độc lập với nhau. Naive Bayes là bộ phân loại theo xác suất (probability classifier) nên chúng ta sẽ đi tính toán xác suất bằng cách sử dụng định lý Bayes.

2.4.2 Một số kiểu mô hình Naive Bayes

- Multinomial Naive Bayes

Mô hình này chủ yếu được sử dụng trong phân loại văn bản. Đặc trưng đầu vào ở đây chính là tần suất xuất hiện của từ trong văn bản đó.

- Bernoulli Naive Bayes

Mô hình này được sử dụng khi các đặc trưng đầu vào chỉ nhận giá trị nhị phân 0 hoặc 1 (phân bố Bernoulli).

- Gaussian Naive Bayes

Mô hình này được sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục.

2.4.3 Hoạt động của Naive Bayes

Naive Bayes Classifiers (NBC) thường được sử dụng trong các bài toán Text Classification.

NBC có thời gian training và test rất nhanh. Điều này có được là do giả sử về tính độc lập giữa các thành phần, nếu biết class.

Nếu giả sử về tính độc lập được thoả mãn (dựa vào bản chất của dữ liệu), NBC được cho là cho kết quả tốt hơn so với SVM và Log khi có ít dữ liệu training.

NBC có thể hoạt động với các feature vector mà một phần là liên tục (sử dụng Gaussian Naive Bayes), phần còn lại ở dạng rời rạc (sử dụng Multinomial hoặc Bernoulli).

Khi sử dụng Multinomial Naive Bayes, Laplace smoothing thường được sử dụng để tránh trường hợp 1 thành phần trong test data chưa xuất hiện ở training data.

2.4.4 Ưu điểm - nhược điểm của Naive Bayes

– Ưu điểm:

Dễ sử dụng và nhanh khi cần đoán nhãn của dữ liệu test. Thực hiện khá tốt trong multi class prediction (test later).

Khi giả định rằng các feature của dữ liệu là độc lập với nhau thì Naive Bayes chạy tốt hơn so với các thuật toán khác như Log và cũng cần ít dữ liệu hơn.

– Nhược điểm:

Độ chính xác của Naive Bayes nếu so với các thuật toán khác thì không được cao.

Trong thế giới thực, hầu như bất khả thi khi các feature của dữ liệu test là độc lập với nhau.

CHƯƠNG 3: BỘ DỮ LIỆU “THERE BANK”

3.1 Mô tả dữ liệu

Bộ dữ liệu có sẵn công khai được lấy từ Kaggle , bộ dữ liệu thuộc về ‘There Bank’.

Tệp chứa dữ liệu nhân khẩu học của 5000 khách hàng như cột “Tuổi” và “Thu nhập” cũng như mối quan hệ với dữ liệu công ty tài chính như cột “Thế chấp” và “Tài

khoản chứng khoán". Và phản hồi của khách hàng về chiến dịch vừa qua, như chuyên mục Vay tiêu dùng cá nhân.

Dữ liệu "There Bank" gồm:

- Tổng số hàng: 5000
- Tổng số cột: 15
- –Gồm các cột: Index (['ID', 'Age', 'Sex', 'Experience', 'Income', 'ZIP Code', 'Family', 'CCAvg', 'Education', 'Mortgage', 'Securities Account', 'CD Account', 'Online', 'CreditCard', 'Personal Loan'], dtype='object')

	ID	Age	Sex	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities Account	CD Account	Online	CreditCard	Personal Loan
0	1	25.0	A	1	49	91107	4.0	1.6	1.0	0	1	0	0	0	0
1	2	45.0	B	19	34	90089	3.0	1.5	1.0	0	1	0	0	0	0
2	3	39.0	A	15	11	94720	1.0	1.0	NaN	0	0	0	0	0	0
3	4	35.0	A	9	100	94112	1.0	2.7	2.0	0	0	0	0	0	0
4	5	35.0	B	8	45	91330	4.0	1.0	2.0	0	0	0	0	1	0
5	6	37.0	A	13	29	92121	4.0	0.4	2.0	155	0	0	1	0	0
6	7	53.0	B	27	72	91711	2.0	1.5	2.0	0	0	0	1	0	0
7	8	NaN	A	24	22	93943	1.0	0.3	3.0	0	0	0	0	1	0
8	9	35.0	A	10	81	90089	3.0	0.6	2.0	104	0	0	1	0	0
9	10	34.0	B	9	180	93023	1.0	8.9	3.0	0	0	0	0	0	1
10	11	65.0	A	39	105	94710	4.0	2.4	3.0	0	0	0	0	0	0
11	12	29.0	B	5	45	90277	3.0	0.1	2.0	0	0	0	1	0	0
12	13	48.0	B	23	114	93106	2.0	3.8	3.0	0	1	0	0	0	0
13	14	59.0	B	32	40	94920	4.0	2.5	2.0	0	0	0	1	0	0
14	15	67.0	B	41	112	91741	NaN	2.0	1.0	0	1	0	0	0	0
15	16	60.0	B	30	22	95054	1.0	1.5	3.0	0	0	0	1	1	0

Hình 3.1: Dữ liệu "There Bank"

- Các biến danh nghĩa:
  - ID - Mã khách hàng
  - ZIP Code - Địa chỉ nhà Mã ZIP của khách hàng
  - Các biến phân loại thông thường:
  - Family - Số thành viên trong gia đình của khách hàng
  - Education - Trình độ học vấn của khách hàng (khoảng từ 1 đến 3 tương ứng là Dưới Đại học, Cao học và Sau đại học
- Biến khoảng:
  - Tuổi - Tuổi của khách hàng

- Kinh nghiệm - Số năm kinh nghiệm của khách hàng đã có
- Thu nhập - Annula Thu nhập của khách hàng tính bằng đô la
- CCAvg - Trung bình chi tiêu trên thẻ tín dụng mỗi tháng bằng đô la
- Thế Chấp - Trị Giá Thế Chấp Căn Nhà
- Biến phân loại nhị phân:
- Tài khoản CD - Khách hàng có Tài khoản CD với công ty tài chính hay không
- Tài Khoản Bảo Mật - Khách hàng có Tài Khoản Bảo Mật với công ty tài chính hay không
- Online - Khách hàng có tiện ích Online banking với công ty tài chính hay không
- Thẻ tín dụng - Khách hàng có thẻ tín dụng do Universal Bank phát hành hay không
- Khoản vay cá nhân - Đây là biến mục tiêu của chúng tôi mà chúng tôi phải dự đoán => Điều này cho biết khách hàng có chấp nhận vay hay không

Kiểu dữ liệu:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  5000 non-null   int64
 1   Age                 4992 non-null   float64
 2   Sex                 5000 non-null   object
 3   Experience          5000 non-null   int64
 4   Income              5000 non-null   int64
 5   ZIP Code            5000 non-null   int64
 6   Family              4996 non-null   float64
 7   CCAvg               5000 non-null   float64
 8   Education           4993 non-null   float64
 9   Mortgage            5000 non-null   int64
 10  Securities Account  5000 non-null   int64
 11  CD Account          5000 non-null   int64
 12  Online              5000 non-null   int64
 13  CreditCard          5000 non-null   int64
 14  Personal Loan       5000 non-null   int64
dtypes: float64(4), int64(10), object(1)
memory usage: 586.1+ KB
```

Hình 3.2: Các kiểu dữ liệu

Bảng 3.1: Mô tả bộ dữ liệu "There Bank"

STT	TÊN CỘT	MÔ TẢ

1	ID	Mã khách hàng
2	Tuổi	Tuổi tính theo năm đầy đủ của khách hàng
3	Kinh nghiệm	Số năm kinh nghiệm chuyên môn
4	Thu nhập	Thu nhập của khách hàng hằng năm
5	Family	Quy mô gia đình của khách hàng
6	CCavg	Trung bình chi tiêu bằng thẻ tín dụng mỗi tháng của khách hàng
7	Học vấn	Trình độ học vấn của khách hàng
8	Mortgage	Giá trị thuế chấp của căn nhà nếu có
9	Khoản vay cá nhân	Khách hàng này có chấp nhận khoản vay cá nhân được cung cấp trong chiến dịch trước không?
10	Tài khoản chứng khoán	Khách hàng có tài khoản chứng khoán tại công ty tài chính không?
11	Tài khoản CD	Khách hàng có tài khoản chứng nhận tiền gửi (CD) với công ty tài chính không?
12	Trực tuyến	Khách hàng có đang sử dụng tiện ích công ty tài chính trực tuyến không?
13	Thẻ tín dụng	Khách hàng có sử dụng thẻ tín dụng do công ty tài chính phát hành không?
14	Zip code	Mã bưu chính nơi ở khách hàng

3.2 Tiền xử lý dữ liệu

Kiểm tra các giá trị null: có 19 chỗ trống.

```
ID                    0
Age                   8
Sex                   0
Experience            0
Income                0
ZIP Code              0
Family                4
CCAvg                 0
Education             7
Mortgage              0
Securities Account    0
CD Account            0
Online                0
CreditCard            0
Personal Loan         0
dtype: int64
```

	ID	Age	Sex	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities Account	CD Account	Online	CreditCard	Personal Loan
2	3	39.0	A	15	11	94720	1.0	1.0	NaN	0	0	0	0	0	0
7	8	NaN	A	24	22	93943	1.0	0.3	3.0	0	0	0	0	1	0
14	15	67.0	B	41	112	91741	NaN	2.0	1.0	0	1	0	0	0	0
17	18	42.0	A	18	81	94305	4.0	2.4	NaN	0	0	0	0	0	0
23	24	44.0	A	18	43	91320	2.0	0.7	NaN	163	1	0	0	0	0
27	28	46.0	B	20	158	90064	1.0	2.4	NaN	0	0	0	1	1	0
54	55	NaN	B	5	44	95819	1.0	0.2	3.0	0	0	0	1	0	0
109	110	43.0	B	17	49	94542	1.0	2.8	NaN	0	0	0	1	0	0
227	228	NaN	B	23	148	94551	2.0	7.5	1.0	0	0	1	1	1	0
247	248	NaN	A	29	120	92626	4.0	2.7	2.0	111	1	1	1	0	1
334	335	NaN	B	23	45	95053	1.0	1.3	2.0	0	0	0	1	0	0
383	384	NaN	B	18	53	94608	1.0	0.2	1.0	0	0	0	1	0	0
522	523	36.0	A	11	72	91007	NaN	2.8	1.0	224	0	0	0	0	0
596	597	48.0	A	22	152	94022	NaN	3.5	3.0	0	0	0	1	0	1
651	652	28.0	B	4	58	92121	NaN	1.5	1.0	131	0	0	0	0	0
681	682	NaN	B	9	164	94720	1.0	6.0	3.0	0	0	0	1	0	1
845	846	NaN	A	17	29	94706	3.0	1.0	2.0	0	0	0	1	1	0

Hình 3.4: Các giá trị null của dữ liệu

Nhận xét:

- Cột 'Age' có 8 trường hợp có giá trị null: Điền giá trị trung bình
- Cột 'Family' có 4 giá trị null: Điền giá trị chiếm số đông
- Cột 'Education' có 7 giá trị null: Điền giá trị chiếm số đông
- Dữ liệu trùng lặp

```
data.duplicated().any()
```

```
False
```

Hình 3.5: Các giá trị trùng lặp

Nhận xét:

– Không có dữ liệu bị trùng

Dữ liệu:

	count	mean	std	min	25%	50%	75%	max
ID	5000.0	2500.500000	1443.520003	1.0	1250.75	2500.5	3750.25	5000.0
Age	5000.0	45.341146	11.458895	23.0	35.00	45.0	55.00	67.0
Experience	5000.0	20.104600	11.467954	-3.0	10.00	20.0	30.00	43.0
Income	5000.0	73.774200	46.033729	8.0	39.00	64.0	98.00	224.0
ZIP Code	5000.0	93152.503000	2121.852197	9307.0	91911.00	93437.0	94608.00	96651.0
Family	5000.0	2.396000	1.147801	1.0	1.00	2.0	3.00	4.0
CCAvg	5000.0	1.937938	1.747659	0.0	0.70	1.5	2.50	10.0
Education	5000.0	1.881000	0.839869	1.0	1.00	2.0	3.00	3.0
Mortgage	5000.0	56.498800	101.713802	0.0	0.00	0.0	101.00	635.0
Securities Account	5000.0	0.104400	0.305809	0.0	0.00	0.0	0.00	1.0
CD Account	5000.0	0.060400	0.238250	0.0	0.00	0.0	0.00	1.0
Online	5000.0	0.596800	0.490589	0.0	0.00	1.0	1.00	1.0
CreditCard	5000.0	0.294000	0.455637	0.0	0.00	0.0	1.00	1.0
Personal Loan	5000.0	0.096000	0.294621	0.0	0.00	0.0	0.00	1.0

Hình 3.6: Nhận xét dữ liệu

Nhận xét:

– Thấy ở cột Experience giá trị min = -3

– Kiểm tra cột Experience:

– Có 52 hàng có giá trị âm. Với giá trị -1 có 33 hàng; với giá trị -2 có 15 hàng; giá trị -3 có 4 hàng.

```
print(data.groupby([data['Experience'] < 0])['Experience'].count().sort_values(ascending=False))
data[data['Experience'] < 0]['Experience'].value_counts()
```

```
Experience
False    4948
True       52
Name: Experience, dtype: int64

-1    33
-2    15
-3     4
Name: Experience, dtype: int64
```

	ID	Age	Sex	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities Account	CD Account	Online	CreditCard	Personal Loan
89	90	25	A	-1	113	94303	4.0	2.30	3.0	0	0	0	0	1	0
226	227	24	A	-1	39	94085	2.0	1.70	2.0	0	0	0	0	0	0
315	316	24	A	-2	51	90630	3.0	0.30	3.0	0	0	0	1	0	0
451	452	28	B	-2	48	94132	2.0	1.75	3.0	89	0	0	1	0	0
524	525	24	B	-1	75	93014	4.0	0.20	1.0	0	0	0	1	0	0
536	537	25	B	-1	43	92173	3.0	2.40	2.0	176	0	0	1	0	0
540	541	25	A	-1	109	94010	4.0	2.30	3.0	314	0	0	1	0	0
576	577	25	A	-1	48	92870	3.0	0.30	3.0	0	0	0	0	1	0
583	584	24	B	-1	38	95045	2.0	1.70	2.0	0	0	0	1	0	0
597	598	24	B	-2	125	92835	2.0	7.20	1.0	0	1	0	0	1	0
649	650	25	A	-1	82	92677	4.0	2.10	3.0	0	0	0	1	0	0
670	671	23	B	-1	61	92374	4.0	2.60	1.0	239	0	0	1	0	0
686	687	24	A	-1	38	92612	4.0	0.60	2.0	0	0	0	1	0	0
793	794	24	A	-2	150	94720	2.0	2.00	1.0	0	0	0	1	0	0
889	890	24	A	-2	82	91103	2.0	1.60	3.0	0	0	0	1	1	0
909	910	23	A	-1	149	91709	1.0	6.33	1.0	305	0	0	0	1	0
1173	1174	24	B	-1	35	94305	2.0	1.70	2.0	0	0	0	0	0	0

Hình 3.7: Dữ liệu Experience bị nhầm lẫn

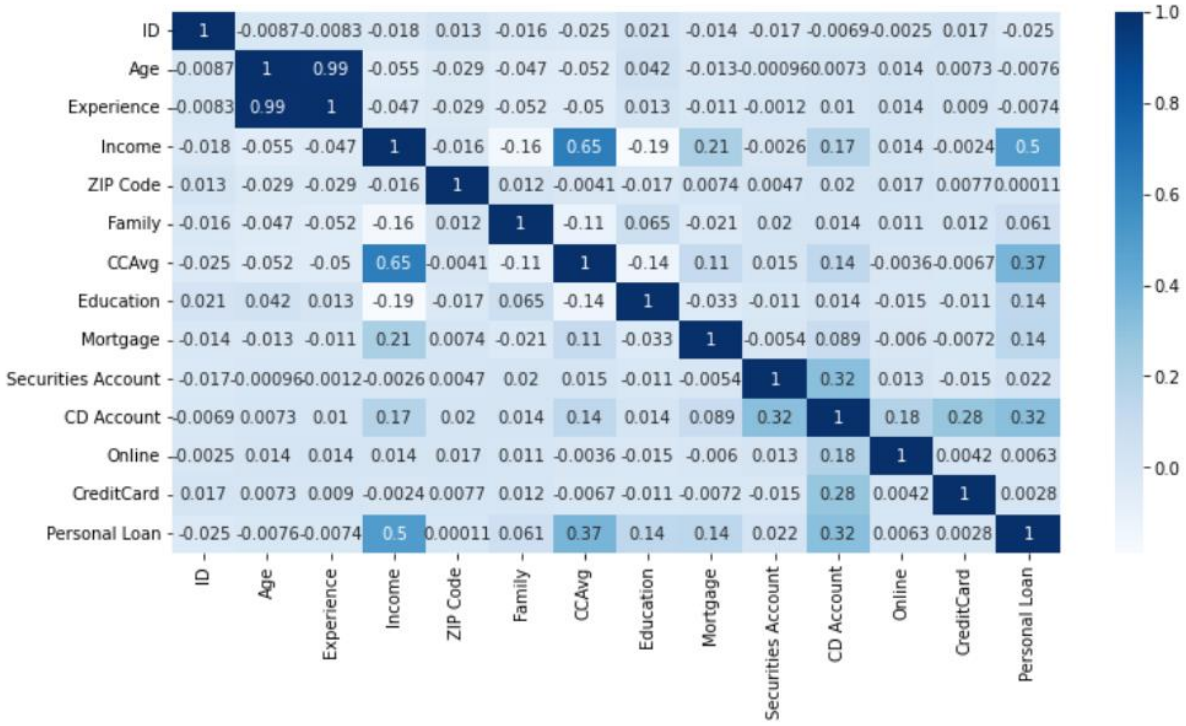
Nhận xét:

– Đa số những khách hàng có Experience < 0 là những người có tuổi còn khá trẻ. Có vẻ lúc điền thông tin đã có sai sót.

```
data.Experience = data.Experience.replace(-1, 1)
data.Experience = data.Experience.replace(-2, 2)
data.Experience = data.Experience.replace(-3, 1)
```

– Lựa chọn cột:

- Biến ID không thêm bất kỳ thông tin thú vị nào. Không có mối liên hệ nào giữa ID khách hàng của một người và khoản vay, nó cũng không đưa ra bất kỳ kết luận chung nào cho các khách hàng tiềm năng cho khoản vay trong tương lai. Chúng ta có thể bỏ qua thông tin này để dự đoán mô hình của mình.
- Vì biến Age và Experience có mức độ 'tương quan' cao vì thế mà vai trò của 2 biến là như nhau nên có thể xóa 1 trong 2.



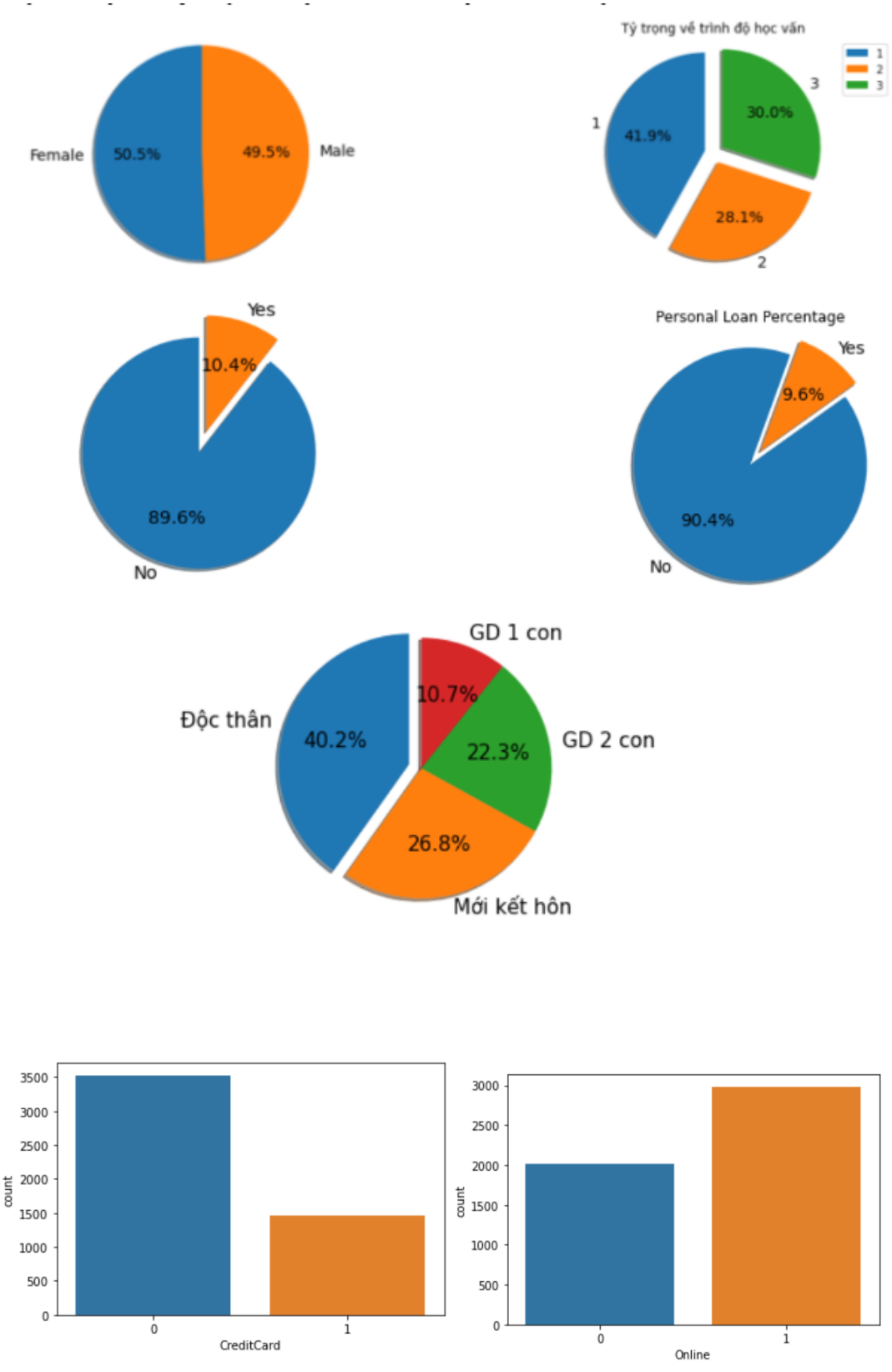
Hình 3.8: Tương quan hai biến Age và Experience

– Biến Zip code: mã nơi ở của khách hàng cũng không có mối quan hệ với biến cần dự đoán nên có thể xóa.

CHƯƠNG 4: PHÂN TÍCH DỮ LIỆU - XÂY DỰNG MÔ HÌNH

4.1 Phân tích EDA

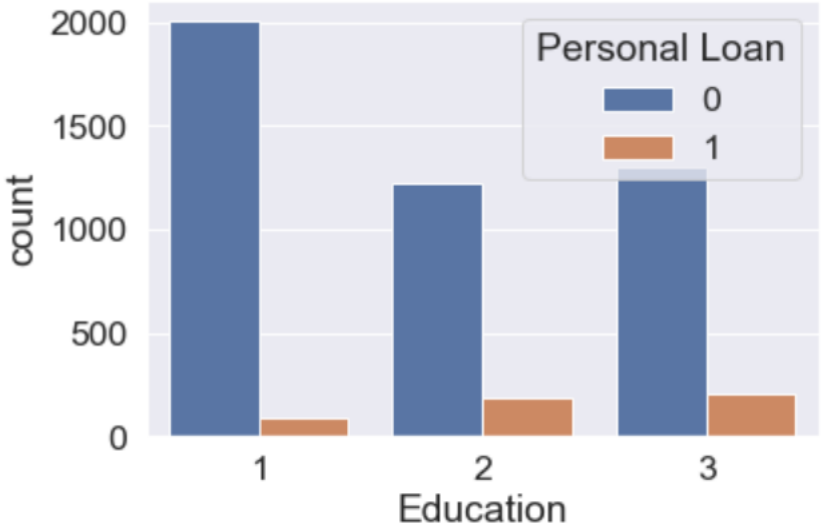
Từ bộ dữ liệu, nhóm đã phân tích EDA để có thể hiểu rõ hơn về dữ liệu. Nhận thấy một số đặc điểm nổi bật của dữ liệu như sau:



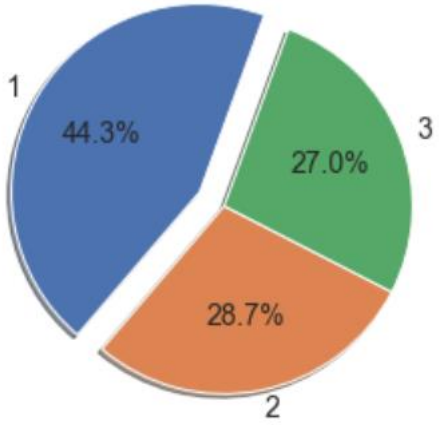
Hình 4.1: Tổng quan phân tích EDA

Từ dữ liệu đã phân tích EDA, nhìn chung nhận thấy tỉ lệ khách hàng có giới tính nữ nhỉnh hơn so với giới tính nam, đồng thời tỉ lệ tình trạng hôn nhân là ‘’độc thân’’ khá lớn khi chiếm hơn 40% trên tổng số khách hàng. Số lượng khách hàng có tài khoản trực tuyến lớn hơn nhóm khách hàng còn lại. Tương tự như thế, nhóm khách hàng có thẻ tín dụng gấp 2 lần số lượng khách hàng không có thẻ tín dụng. Đặc biệt, dữ liệu thấy rõ có sự chênh lệch rất lớn đối với nhóm khách hàng không chấp nhận khoản vay cá nhân. Vậy nên, đề xuất có thể đưa ra ý kiến mô hình sẽ có xu hướng hoạt động tốt hơn trong việc dự đoán những khách hàng nào không chấp nhận khoản vay cá nhân.

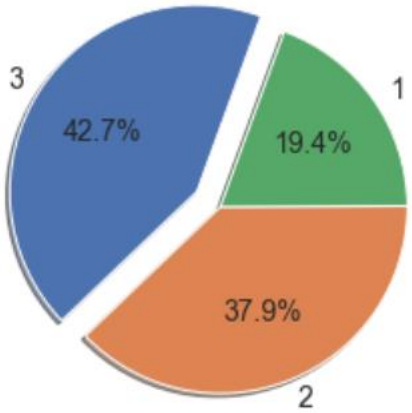
4.1.1 Trình độ học vấn của khách hàng



Hình 4.2: Thống kê khách hàng có khoản vay theo trình độ học vấn



Hình 4.3: Tỷ lệ khách hàng không đồng ý vay theo trình độ học vấn

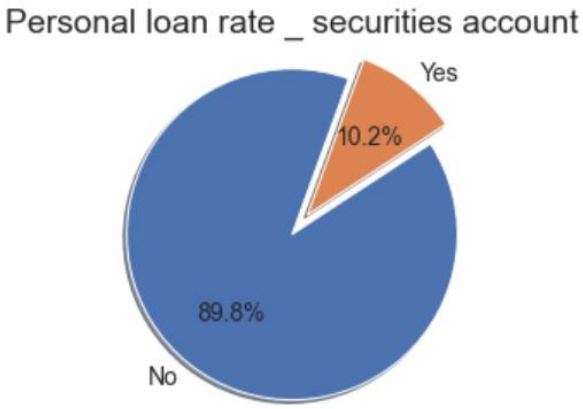


Hình 4.4: Tỷ lệ khách hàng đồng ý vay theo trình độ học vấn

Từ biểu đồ hộp trên ta có thể thấy trong số những khách hàng không đồng ý vay tiền, có một sự phân chia rõ rệt về học vấn. Học vấn 1 chiếm 44.3%, học vấn 2 chiếm 28.7%, và học vấn 3 chiếm 27%. Điều này cho thấy khách hàng có học vấn cao (học vấn 3) có xu hướng ít đồng ý vay tiền hơn so với khách hàng có học vấn thấp hơn (học vấn 1 và 2). Trong số những khách hàng đồng ý vay tiền, cũng có sự phân chia về học vấn. Tuy nhiên, phân bổ này có sự khác biệt so với tỷ lệ không đồng ý. Học vấn 1 chiếm 19.4%, học vấn 2 chiếm 37.9%, và học vấn 3 chiếm 42.7%. Nhìn chung, khách hàng có học vấn cao (học vấn 3) có xu hướng đồng ý vay tiền hơn so với khách hàng có học vấn thấp hơn (học vấn 1 và 2). Những khách hàng có trình độ học vấn là 1 và không vay có thế chấp cao hơn so với những khách hàng có cùng trình độ học vấn. Những khách hàng có trình độ học vấn là 2 và 3 và không vay có thế chấp ít hơn so với những khách hàng có cùng trình độ học vấn.

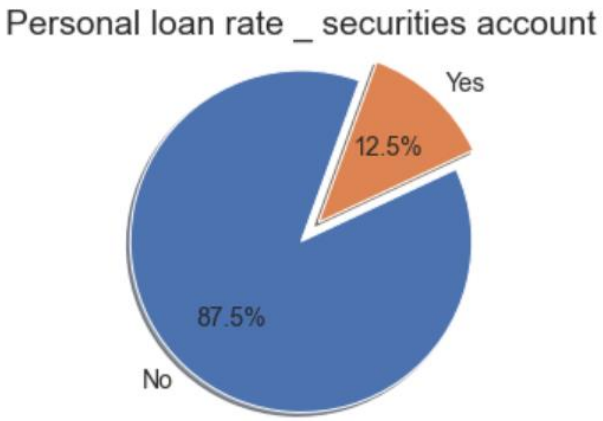
4.1.2 Khách hàng với chỉ tiêu có sử dụng tài khoản chứng khoán

Hình 4.5: Thống kê khách hàng có khoản vay theo chỉ tiêu có sử dụng tài khoản chứng khoán



Hình 4.6: Tỷ lệ khách hàng có tài khoản chứng khoán trên tổng số khách hàng không đồng ý khoản vay

Trong số khách hàng không đồng ý vay tiền, 10.2% có tài khoản chứng khoán. Điều này cho thấy một phần nhỏ khách hàng không đồng ý vay tiền vẫn sở hữu tài khoản chứng khoán. Có thể giải thích rằng những khách hàng này có lý do khác nhau để không đồng ý vay tiền, có thể liên quan đến sự ưu tiên đầu tư vào thị trường chứng khoán hoặc không có nhu cầu vay tiền.



Hình 4.7: Tỷ lệ khách hàng có tài khoản chứng khoán trên tổng số khách hàng đồng ý khoản vay

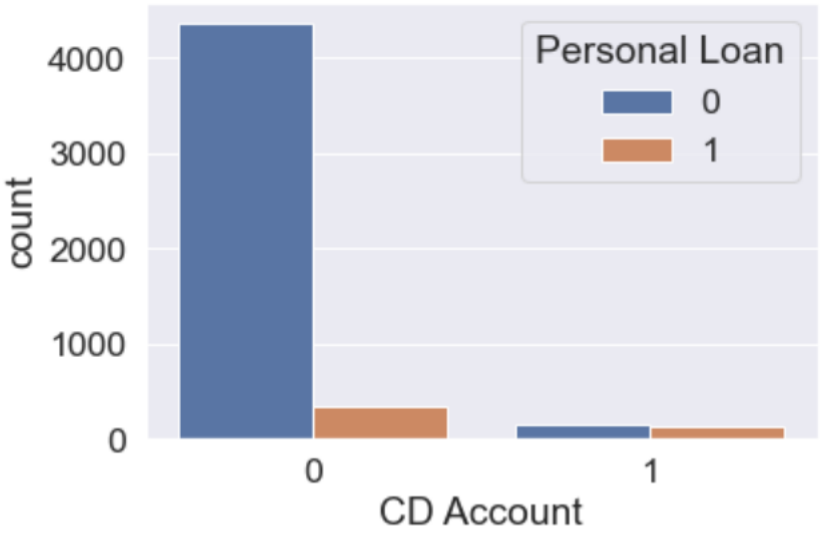
Trong số khách hàng đồng ý vay tiền, 12.5% có tài khoản chứng khoán. Điều này cho thấy một số khách hàng có quyết định vay tiền cũng đồng thời sở hữu tài khoản chứng khoán. Có thể giải thích rằng việc sở hữu tài khoản chứng khoán không ảnh hưởng quyết định vay tiền, hoặc có thể liên quan đến mục đích sử dụng vay tiền.

Khách hàng có tài khoản chứng khoán thường dễ vay hơn. Đa số khách hàng chưa vay vốn đều không có tài khoản chứng khoán.

4.1.3 Family

Hình 4.8: Thống kê khách hàng có khoản vay theo chỉ tiêu quy mô gia đình

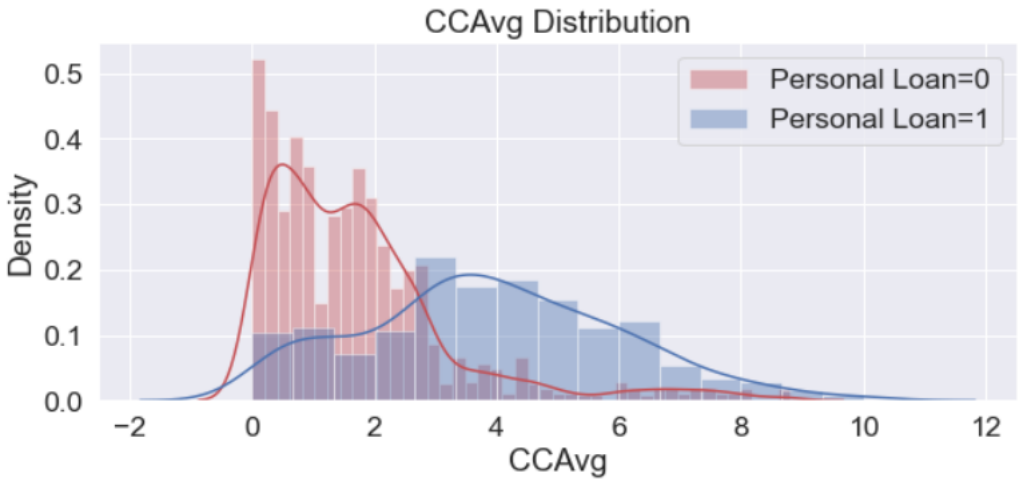
Quy mô gia đình không có bất kỳ tác động nào trong khoản vay cá nhân. Nhưng có vẻ như các gia đình có quy mô 3 và 4 người có nhiều khả năng vay hơn.



Hình 4.9: Thống kê khách hàng có khoản vay theo chỉ tiêu tài khoản tiền gửi

Khách hàng không có tài khoản chứng chỉ gửi tiền thì cũng không có khoản vay. Điều này dường như là đa số. Nhưng hầu như tất cả các khách hàng có tài khoản chứng chỉ gửi tiền đều có khoản vay. Điều này cho thấy rằng khách hàng sở hữu tài khoản chứng chỉ gửi tiền có khả năng và sẵn lòng vay tiền. Có thể giải thích rằng họ có thu nhập ổn định hoặc lịch sử tín dụng tốt, điều này làm cho họ trở thành ứng viên phù hợp để vay tiền từ ngân hàng hoặc tổ chức tài chính.

4.1.4 Chi tiêu bằng thẻ tín dụng của 2 nhóm khách hàng

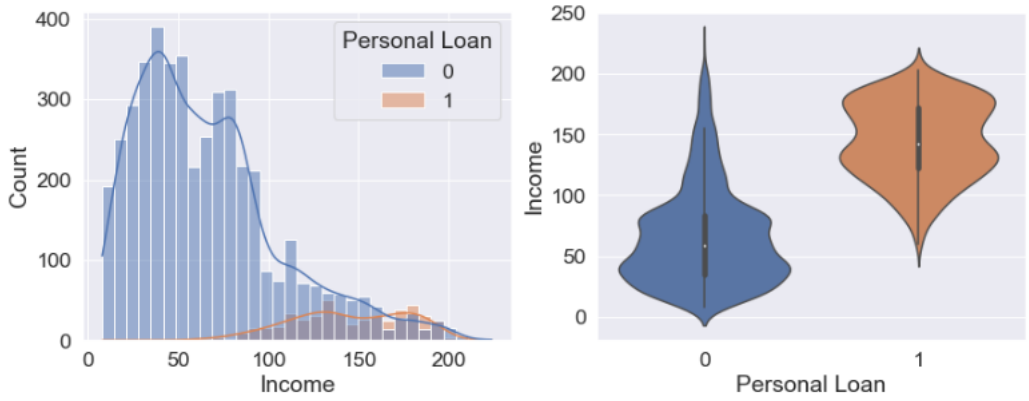


Hình 4.10: Thống kê khách hàng có khoản vay theo chỉ tiêu thẻ tín dụng

Những khách hàng đã vay cá nhân có mức trung bình thẻ tín dụng cao hơn so với những khách hàng không vay. Vì vậy, mức trung bình thẻ tín dụng cao dường như là yếu tố dự đoán tốt về việc khách hàng có vay cá nhân hay không. Biểu đồ cho thấy

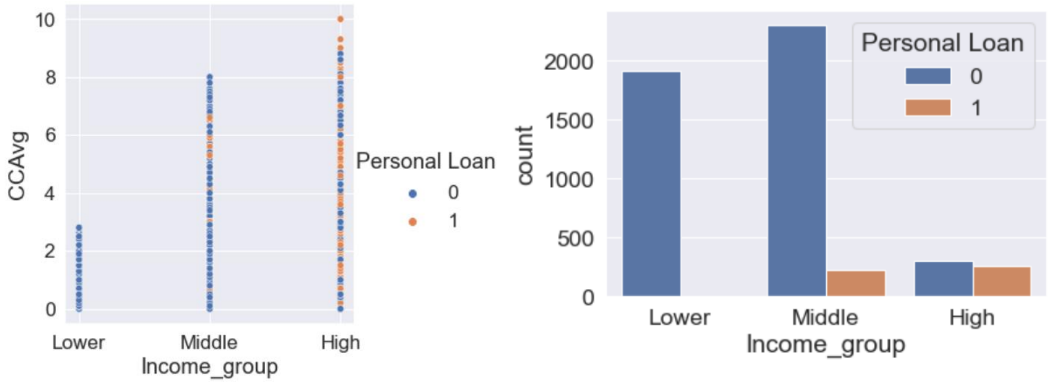
những người có khoản vay cá nhân có mức trung bình thẻ tín dụng cao hơn. Chi tiêu thẻ tín dụng trung bình với mức trung bình là 3800 đô la cho thấy xác suất vay cá nhân cao hơn. Chi tiêu bằng thẻ tín dụng thấp hơn với mức trung bình là 1400 đô la thì ít có khả năng vay hơn.

4.1.5 Thu nhập của 2 nhóm khách hàng



Hình 4.11: Thống kê khách hàng có khoản vay theo chỉ tiêu thu nhập

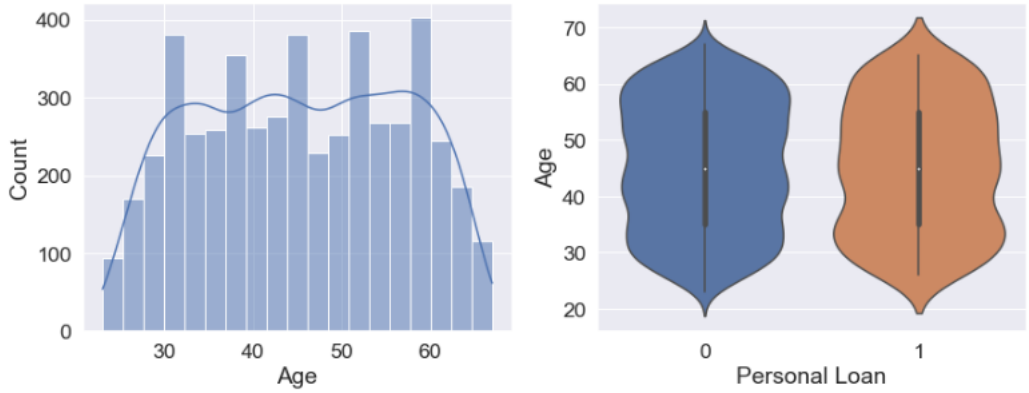
Những khách hàng đã vay cá nhân có thu nhập cao hơn những người không vay. Điều này cho thấy một xu hướng rõ rệt rằng khách hàng có thu nhập cao có khả năng và sẵn lòng vay tiền cá nhân hơn. Lý do có thể là do khách hàng có thu nhập cao có khả năng trả nợ tốt hơn và có sự ổn định tài chính, làm cho họ trở thành ứng viên phù hợp để vay tiền.Vì vậy, thu nhập cao dường như là yếu tố dự đoán tốt về việc khách hàng có vay tiền cá nhân hay không.



Hình 4.12: Thống kê khách hàng chấp nhận khoản vay theo chỉ tiêu thu nhập

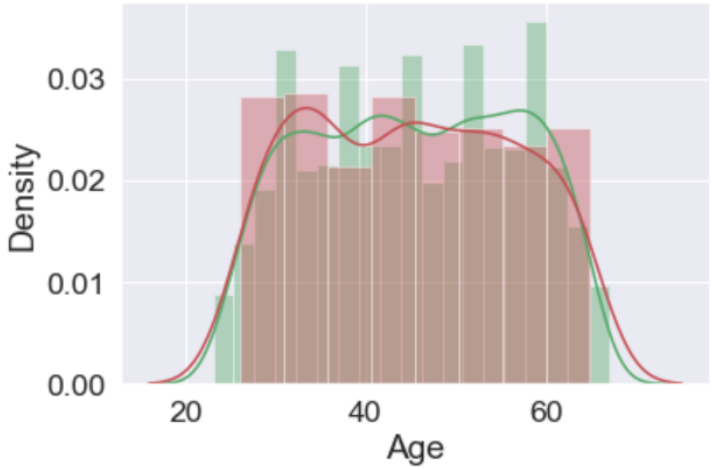
Trong biểu đồ này, chúng ta có thể thấy rằng chúng ta có sự phân biệt rõ ràng về thu nhập giữa những người có chấp nhận khoản vay cá nhân hay không, những người đã chấp nhận khoản vay có thu nhập cao hơn những người không chấp nhận.

4.1.6 Phân phối về tuổi



Hình 4.13: Thống kê khách hàng theo chỉ tiêu độ tuổi

Dữ liệu tuổi có độ phân phối qua nhiều lứa tuổi. Độ tuổi của khách hàng nằm trong khoảng từ 23 - 67, với giá trị trung bình và trung vị là ~45.



Hình 4.14: Tỷ trọng khách hàng đã từng chọn khoản vay cá nhân trước đây

Phân tích EDA, thấy được những người có thu nhập cao hơn đã từng chọn khoản vay cá nhân trước đây. Những người có thế chấp cao đã chọn vay. Khách hàng có mức sử dụng tín dụng trung bình hàng tháng cao hơn đã chọn vay. Khách hàng có thu nhập cao hơn có mức sử dụng thẻ tín dụng và thế chấp trung bình cao hơn. Cao học và Cao cấp/Chuyên nghiệp có mức sử dụng thẻ tín dụng hàng tháng cao hơn và đã vay công ty tài chính.

CHƯƠNG 5: XÂY DỰNG MÔ HÌNH

5.1 Hồi quy Logistic:

– Chia dữ liệu train, test: theo tỷ lệ 70 % - 30%

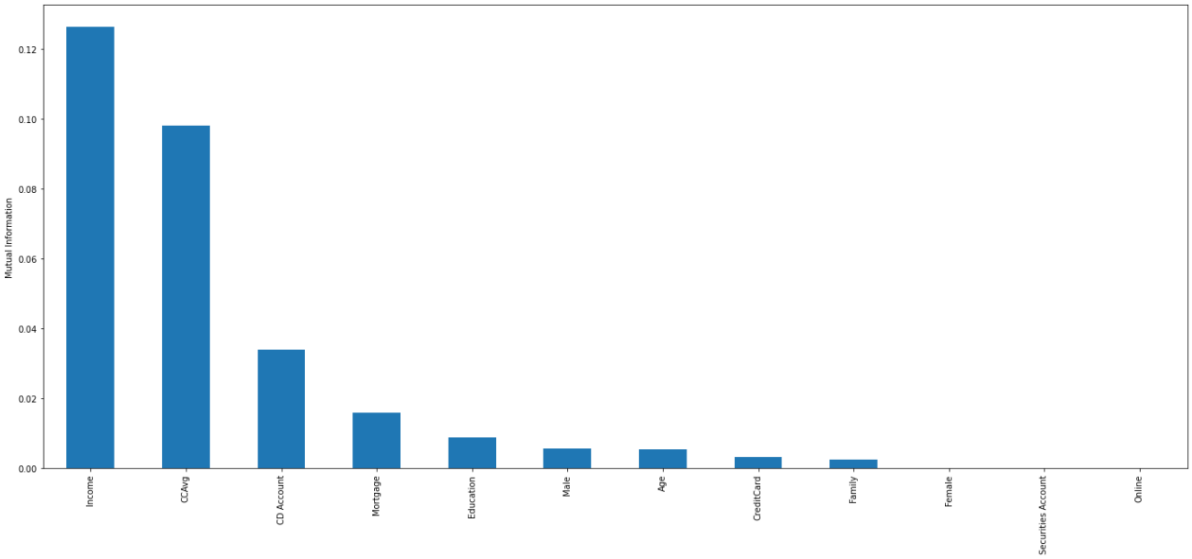
```
from sklearn.model_selection import train_test_split
```

```
x_train, x_test, y_train, y_test = train_test_split(
                                    data_X,data_Y, test_size= 0.3)
```

```
print('Số dòng dùng để đào tạo mô hình: ', len(x_train))
print('Số dòng dùng để thử nghiệm mô hình: ', len(x_test))
```

```
Số dòng dùng để đào tạo mô hình:  3500
Số dòng dùng để thử nghiệm mô hình:  1500
```

– Lựa chọn các biến mang lại nhiều thông tin:



Nhận xét:

- Có một số tính năng có giá trị thông tin lẫn nhau cao. Ngoài ra cũng có các tính năng có giá trị thông tin lẫn nhau thấp gần như bằng 0.
- Dựa vào đó ta chọn các tính năng: chọn ***k*** tính năng có chỉ số giá trị thông tin lẫn nhau cao.
- Nhưng với bài toán mà nhóm đang thực hiện số cột và số dòng dữ liệu còn khá ít cùng với sự tham khảo ý kiến của GVHD nên nhóm quyết định giữ lại các cột.

Ở đây chúng ta có thể thấy dữ liệu bị mất cân bằng. Nhãn '0' nhiều hơn nhãn '1'. Điều này có thể làm giảm hiệu suất và độ chính xác của mô hình trên các lớp thiểu số. Vì vậy cần CÂN BẰNG DỮ LIỆU để mô hình được đào tạo tốt hơn.

```
print("Trước khi cân bằng dữ liệu, số dòng thuộc nhãn '1':", len(y_train_1))
print("Trước khi cân bằng dữ liệu, số dòng thuộc nhãn '0':", len(y_train_0))
```

```
Trước khi cân bằng dữ liệu, số dòng thuộc nhãn '1': 331
Trước khi cân bằng dữ liệu, số dòng thuộc nhãn '0': 3169
```

```
print("Sau khi cân bằng dữ liệu, số dòng thuộc nhãn '1':", len(y_train_res_1 == 1))
print("Sau khi cân bằng dữ liệu, số dòng thuộc nhãn '0':", len(y_train_res_0 == 0))
```

```
Sau khi cân bằng dữ liệu, số dòng thuộc nhãn '1': 3169
Sau khi cân bằng dữ liệu, số dòng thuộc nhãn '0': 3169
```

Đào tạo mô hình:

```
from sklearn.linear_model import LogisticRegression
```

```
classifier = LogisticRegression(random_state = 0)
classifier.fit(x_train_res, y_train_res)
y_pred = classifier.predict(x_test)
```

Dự đoán ngẫu nhiên 20 người:

```
#dự đoán 20 người bất kì
DD = pd.DataFrame({'True values ':y_test ,'y_pred':predictions})
DD.sample(10)
```

	True values	y_pred
1415	0	0
1402	0	0
142	0	0
609	0	1
1434	0	0
909	1	0
693	0	1
358	0	0
1341	0	1
1268	0	1

Các chỉ số đánh giá mô hình:

a. Độ chính xác của mô hình:

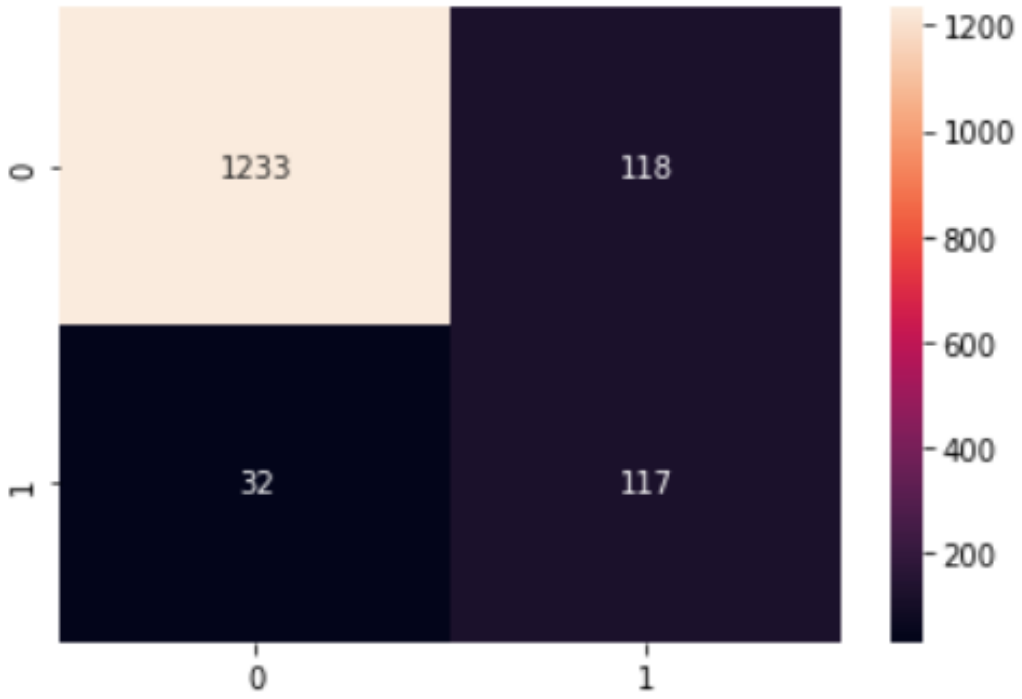
```
print("Độ chính xác của mô hình: {:.2f} %"
      .format(metrics.accuracy_score(y_test, y_pred)*100))
```

```
Độ chính xác của mô hình: 90.00 %
```

Nhận xét:

– Với độ chính xác của mô hình dự đoán chấp nhận các đề nghị cho vay tiêu dùng cá nhân của công ty tài chính là 90%, có thể kết luận rằng mô hình đạt được mức độ chính xác khá tốt trong việc dự đoán xem một đề nghị cho vay tiêu dùng cá nhân nên được chấp nhận hay không.

b. Ma trận nhầm lẫn:



Nhận xét:

– Trong số 1233 dự đoán khách hàng chấp nhận đề nghị khoản vay tiêu dùng cá nhân thì có thực sự 1351 khách hàng chấp nhận.

– Có 1233 dự đoán đúng những khách hàng chấp nhận đề nghị khoản vay tiêu dùng cá nhân trong số 1266 khách hàng thực sự chấp nhận.

- Có 1350 dự đoán đúng những khách hàng chấp nhận và không chấp nhận đề nghị khoản vay tiêu dùng cá nhân trong tổng số 1500 người được dự đoán
- Có 117 dự đoán đúng những khách hàng không chấp nhận đề nghị khoản vay tiêu dùng cá nhân trong số 235 khách hàng thực sự không chấp nhận.

c. Các chỉ số đánh giá khác:

	precision	recall	f1-score	support
0	0.97	0.91	0.94	1351
1	0.50	0.79	0.61	149
accuracy			0.90	1500
macro avg	0.74	0.85	0.78	1500
weighted avg	0.93	0.90	0.91	1500

Nhận xét:

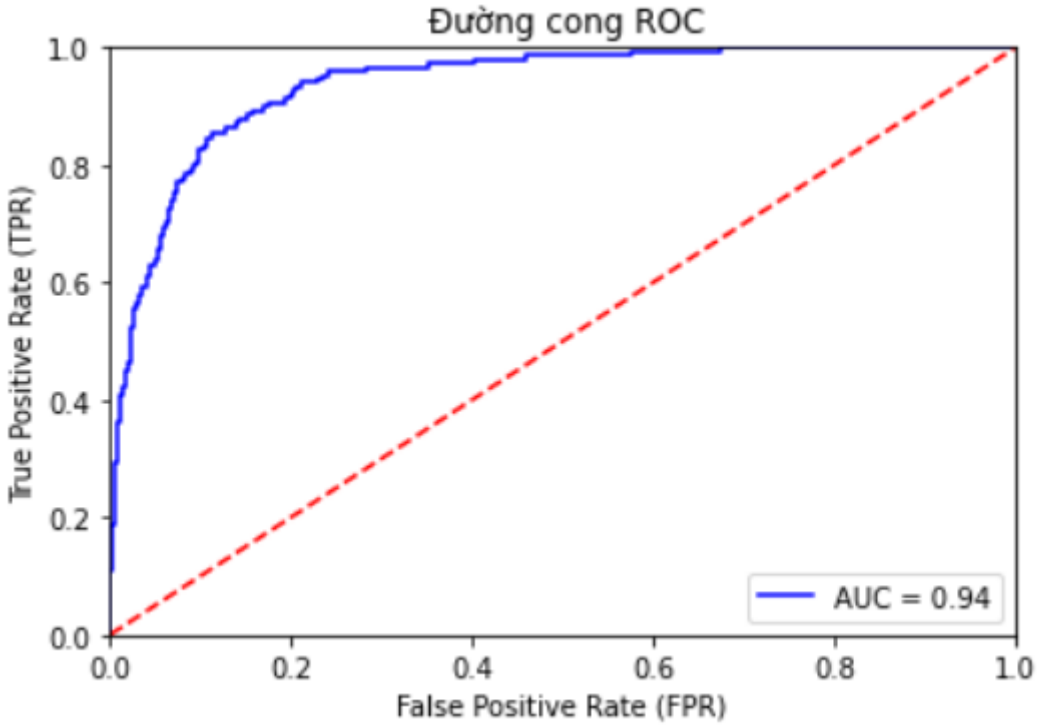
- Precision: Mô hình đạt được độ chính xác phân loại lớp 0 là 0,97 và lớp 1 là 0,5. Điều này cho thấy mô hình có khả năng dự đoán chính xác các trường hợp không chấp nhận đề nghị vay (lớp 0) cao hơn so với trường hợp chấp nhận đề nghị vay (lớp 1).
- Recall: Mô hình đạt được tỷ lệ phát hiện (recall) của lớp 0 là 0,91 và lớp 1 là 0,79. Điều này cho thấy mô hình có khả năng tìm ra nhiều trường hợp chấp nhận đề nghị vay (lớp 1) hơn so với trường hợp không chấp nhận đề nghị vay (lớp 0).
- F1-score: F1-score là một phép đo kết hợp giữa precision và recall. Mô hình đạt được F1-score của lớp 0 là 0,94 và lớp 1 là 0,61. Điều này cho thấy mô hình có hiệu suất tốt trong việc phân loại trường hợp không chấp nhận đề nghị vay (lớp 0), nhưng hiệu suất trong việc phân loại trường hợp chấp nhận đề nghị vay (lớp 1) còn có thể cải thiện.
- Tổng thể, mô hình có khả năng dự đoán chính xác trường hợp không chấp nhận đề nghị vay (lớp 0) tốt hơn, trong khi hiệu suất phân loại trường hợp chấp nhận đề

nghị vay (lớp 1) có thể cải thiện. Cần xem xét thêm các biện pháp để cải thiện hiệu suất phân loại trên lớp thiểu số.

d. Đường cong ROC:

```
print("Roc Auc Score: ", roc_auc_score(y_test,preds))
```

```
Roc Auc Score:  0.9374214476972065
```



Nhận xét:

– Mô hình dự đoán chấp nhận các đề nghị cho vay tiêu dùng cá nhân của công ty tài chính đạt được mức độ chính xác cao. Giá trị Roc Auc Score là 0.94, gần với giá trị tối đa là 1. Điều này cho thấy mô hình có khả năng phân loại tốt giữa các trường hợp chấp nhận và không chấp nhận đề nghị vay.

e. Chỉ số hồi quy:

```
print('Mean Absolute Error: ', metrics.mean_absolute_error(y_test,y_pred))
print('Root Mean Squared Error: ', np.sqrt(metrics.mean_squared_error(y_test,y_pred)))
```

```
Mean Absolute Error:  0.1
Root Mean Squared Error:  0.31622776601683794
```

Nhận xét:

Dựa trên hai độ đo Mean Absolute Error (MAE) và Root Mean Squared Error (RMSE) của mô hình dự đoán chấp nhận các đề nghị cho vay tiêu dùng cá nhân của công ty tài chính, có thể đưa ra nhận xét như sau:

- Mean Absolute Error (MAE): MAE đo lường độ lệch trung bình giữa giá trị dự đoán và giá trị thực tế. Giá trị MAE là 0.1, đây là một giá trị thấp, cho thấy mô hình có xu hướng dự đoán chính xác và gần với giá trị thực tế. Điều này cho thấy mô hình có độ chính xác tương đối cao trong việc dự đoán khả năng chấp nhận đề nghị cho vay tiêu dùng cá nhân.
- Root Mean Squared Error (RMSE): RMSE cũng đo lường sự sai khác giữa giá trị dự đoán và giá trị thực tế, nhưng trọng số lớn hơn cho các sai khác lớn hơn. Giá trị RMSE là 0.3162, cũng cho thấy sự sai khác giữa giá trị dự đoán và thực tế là tương đối nhỏ. Một giá trị RMSE thấp hơn cho thấy mô hình có khả năng dự đoán chính xác hơn và có hiệu suất tốt.

5.2 Một vài mô hình khác:

5.2.1 SVM:

```
print("Độ chính xác của mô hình: {:.2f} %"
      .format(metrics.accuracy_score(y_test, y_pred_2)*100))
```

```
Độ chính xác của mô hình: 90.00 %
```

5.2.2 Decision Tree:

```
y_pred_3 = decision_tree.predict(x_test)
print("Độ chính xác của mô hình: {:.2f} %"
      .format(metrics.accuracy_score(y_test, y_pred_3)*100))
```

```
Độ chính xác của mô hình: 99.07 %
```

5.2.3 Random Forest:

```
y_pred_4 = RFClassifier.predict(x_test)
print("Độ chính xác của mô hình: {:.2f} %"
      .format(metrics.accuracy_score(y_test, y_pred_4)*100))
```

```
Độ chính xác của mô hình: 98.33 %
```

5.2.4 Naive Bayes:

```
y_pred_5 = GaussianNB_model.predict(x_test)
print("Độ chính xác của mô hình: {:.2f} %"
      .format(metrics.accuracy_score(y_test, y_pred_5)*100))
```

```
Độ chính xác của mô hình: 86.13 %
```

CHƯƠNG 6: ÁP DỤNG MÔ HÌNH ĐỂ ĐƯA RA DỰ ĐOÁN

Mục đích của đề tài là dự đoán chính xác liệu khách hàng đó sau khi tiếp xúc với chiến lược marketing của công ty tài chính sẽ đưa ra lựa chọn như thế nào. Có năm thuật toán phân loại đã được sử dụng trong dự án này: Hồi quy Log, SVM, Decision Tree, Random Forest và GausianNB. Từ quá trình triển khai, có vẻ như Trình phân loại rừng ngẫu nhiên có độ chính xác cao nhất, nhưng không thể phủ nhận độ chính xác của các mô hình khác là khá cao và không chênh lệch gì mấy. Cùng với việc biến mục tiêu Personal Loan là biến nhị phân đưa về 2 giá trị dự đoán 0 và 1. Nhóm quyết định sẽ chọn mô hình hồi quy Logistic.

Nhập thông tin để dự đoán:

```
def main():
    age                 = int(input("Độ tuổi của bạn : "))
    income              = float(input("Thu nhập của bạn: "))
    sex_female          = int(input("Nhập 1 nếu bạn là nữ, 0 nếu là nam: "))
    sex_male            = int(input("Nhập 1 nếu bạn là nam, 0 nếu là nữ: "))
    family              = int(input("Số người trong gia đình bạn: "))
    education           = int(input("Trình độ học vấn của bạn : "))
    ccavg               = float(input("Trung bình chi tiêu bằng thẻ tín dụng mỗi tháng: "))
    mortgage            = int(input("Giá trị thuế chấp của căn nhà nếu có: "))
    securities_account  = int(input("Bạn có sử dụng tài khoản tại ngân hàng:"))
    CD_account          = int(input("Bạn có tài khoản tiền gửi ngân hàng: "))
    credit_card         = int(input("Bạn có sử dụng thẻ tín dụng do ngân hàng phát hành: "))
    online              = int(input("Bạn có sử dụng các tiện ích trực tuyến của ngân hàng: "))
    x_input = [[age, sex_female, sex_male, income, family, ccavg, mortgage,
                education,securities_account, CD_account, credit_card, online]]

    result = predict(x_input)
    print('Kết quả nhận được là:',result[0])
```

```
Bạn đang nhập các chỉ số!
Độ tuổi của bạn : 40
Thu nhập của bạn: 60
Nhập 1 nếu bạn là nữ, 0 nếu là nam: 0
Nhập 1 nếu bạn là nam, 0 nếu là nữ: 1
Số người trong gia đình bạn: 4
Trình độ học vấn của bạn : 3
Trung bình chi tiêu bằng thẻ tín dụng mỗi tháng: 4.0
Giá trị thuế chấp của căn nhà nếu có: 200
Bạn có sử dụng tài khoản tại ngân hàng:1
Bạn có tài khoản tiền gửi ngân hàng: 1
Bạn có sử dụng thẻ tín dụng do ngân hàng phát hành: 1
Bạn có sử dụng các tiện ích trực tuyến của ngân hàng: 1
Kết quả nhận được là: 0
```

CHƯƠNG 7: KẾT LUẬN

Thuật toán hồi quy logistic được sử dụng rộng rãi trong các bài toán phân loại nhị phân. Với sự kết hợp giữa hàm sigmoid và phương pháp tối ưu gradient descent, thuật toán này có khả năng tìm ra giá trị tối ưu của các tham số để tạo ra một mô hình phân loại hiệu quả.Một trong những ưu điểm quan trọng của thuật toán hồi quy logistic là tính linh hoạt trong việc xử lý dữ liệu đầu vào. Nó không yêu cầu các điều kiện đặc biệt về phân phối của dữ liệu và có thể làm việc với các biến đầu vào liên tục, rời rạc hoặc thậm chí các biến hạng mục. Điều này giúp cho thuật toán trở thành một công cụ mạnh mẽ và linh hoạt để áp dụng vào nhiều bài toán thực tế. Thuật toán hồi quy logistic là một công cụ quan trọng và phổ biến trong lĩnh vực học máy và phân loại dữ liệu. Với tính linh hoạt và khả năng tìm ra giá trị tối ưu, nó cho phép chúng ta xây dựng các mô hình phân loại hiệu quả cho nhiều bài toán thực tế.