TRƯỜNG ĐẠI HỌC KINH TẾ LUẬT
KHOA TÀI CHÍNH - NGÂN HÀNG

BÁO CÁO CUỐI KÌ

Mô hình rủi ro tín dụng với R/ Python

THẨM ĐỊNH RỦI RO TÍN DỤNG BẰNG MÔ HÌNH HỒI QUY LOGISTIC VÀ CÂY QUYẾT ĐỊNH.

GVHD	***Cô Phạm Thị Thanh Xuân***
Sinh viên	***Huỳnh Thị Hà Thanh***
	Vũ Thành Đạt
	Nguyễn Cao Long
	Lê Nguyễn Phú Lộc
	Giáp Hoàng Long
Lớp	***K19414C***

TP Hồ Chí Minh, ngày tháng năm 2019

MỤC LỤC

THẨM ĐỊNH RỦI RO TÍN DỤNG BẰNG MÔ HÌNH HỒI QUY LOGISTIC VÀ CÂY QUYẾT ĐỊNH.

1. Giới thiệu

Hoạt động tín dụng của các ngân hàng đã và đang là một trong những hoạt động mang lại nhiều lợi nhuận và góp phần không nhỏ vào tổng lợi nhuận của các ngân hàng. Nhưng hình thức hoạt động này cũng mang lại không ít rủi ro cho các ngân hàng. Rủi ro từ việc khách hàng không đủ khả năng hoặc không thực hiện nghĩa vụ trả nợ cho ngân hàng. Từ đó có thể dẫn đến các rủi ro vỡ nợ dây chuyền cũng như nợ xấu ảnh hưởng trực tiếp đến ngân hàng đang cho vay. Bên cạnh đó sẽ ảnh hưởng tới các bên liên quan. Những hậu quả trên sẽ được khắc phục khi các ngân hàng chủ động nâng cao công tác thẩm định, đánh giá hồ sơ của khách hàng một cách chủ động, khoa học hơn. Phần lớn công việc đánh giá hồ sơ khách hàng sẽ phụ thuộc vào kinh nghiệm, năng lực và yếu tố cảm tính của cán bộ tín dụng. Nhằm hỗ trợ ngân hàng hoạt động tốt và chính xác hơn trong việc đánh giá hồ sơ khách hàng, nhóm tác giả tiến hành nghiên cứu khả năng trả nợ đúng hạn của khách hàng cá nhân bằng mô hình LOGISTIC và DECISION TREE.

2. Khung lý thuyết

2.1 Khái niệm khả năng trả nợ đúng hạn

Ngân hàng Nhà Nước. (2013) Tại Việt Nam, theo khoản 8 điều 3 chương I của Thông tư số 02/2013/TT-NHNN có quy định nợ xấu là nợ thuộc các nhóm 3, 4 và 5, trong đó điều 11 mục 1 chương II có quy định rõ:
khả năng trả nợ đúng hạn của khách hàng là việc khách hàng có khả năng trả nợ đầy đủ và đúng hạn với bên cho vay hay không. Hiện nay vẫn chưa có định nghĩa thống nhất về "khả năng trả nợ đúng hạn", những gì đúc kết ra được vẫn chỉ là những dấu hiệu của "không có khả năng trả nợ đúng hạn". Từ đó ta có thể rút ra rằng những khách hàng "không có khả năng trả nợ đúng hạn" là một tập hợp không thuộc những khách hàng "có khả năng trả nợ đúng hạn". Dờn, N. Đ. (2016) Theo Hiệp ước Basel II có 2 tình trạng sau có thể dùng làm căn cứ để đánh giá khả năng không trả được nợ đúng hạn của khách hàng:

- Khách hàng không có khả năng thực hiện nghĩa vụ thanh toán đầy đủ khi đến hạn mà chưa tính đến việc ngân hàng bán tài sản (nếu có) để hoàn trả
- Khách hàng có các khoản nợ xấu có thời gian quá hạn trên 90 ngày. Trong đó, những khoản thấu chi được xem là quá hạn khi khách hàng vượt hạn mức hoặc được thông báo một hạn mức nhỏ hơn dư nợ hiện tại.

Để phù hợp với Hiệp ước Basel II và khung pháp lý của Việt Nam, nghiên cứu này tiến hàng xác định khả năng trả nợ đúng hạn của khách hàng và phân làm hai nhóm:

- Nhóm 0: Khách hàng trả nợ không đúng hạn.
- Nhóm 1: Khách hàng trả nợ đúng hạn.

2.2 Lược khảo các mô hình thẩm định rủi ro tín dụng

Mai, N. C. (2022) đã thực hiện một bài nghiên cứu đánh giá khả năng trả nợ và phân loại khách hàng vay tín chấp tại NHTMCP Việt Nam Thịnh Vượng. Theo kiểm định t-test kết quả cho thấy có sự khác biệt giữa độ tuổi, thu nhập hàng tháng, thời gian và lãi suất vay của hai nhóm khách hàng có và không có nợ quá hạn, trong khi đó không có sự khác biệt giữa thời gian còn lại khoản vay, dư nợ gốc của hai nhóm khách hàng này. Theo kiểm định Pearson's Chi-squared kết quả cho thấy với mức ý nghĩa trên 99% các biến rời rạc như giới tính, tình trạng hôn nhân, trình độ học vấn và chức vụ công tác của khách hàng đều có mối quan hệ với biến nợ quá hạn. Kết quả mô hình hồi quy Logistic dựa trên dữ liệu huấn luyện cho thấy với mức ý nghĩa trên 99% tất cả 11 yếu tố đã đề cập ở trên đều ảnh hưởng đến khả năng có nợ quá hạn của khách hàng vay tín chấp, bao gồm: độ tuổi, giới tính, tình trạng hôn nhân, trình độ học vấn, chức vụ công tác, thu nhập hàng tháng của khách hàng, số sản phẩm khách hàng nắm giữ, thời hạn cho vay, thời gian còn lại khoản vay, lãi suất cho vay và dư nợ gốc còn lại của khoản vay.

Phong, T. T., Bình, N. T., Trang, L. X., & Phượng, Đ. T. (2020) nghiên cứu về khả năng trả nợ của khách hàng cá nhân được thực hiện tại Agribank Tân Hưng, Tỉnh Long An. Kết quả kiểm định độ phù hợp của mô hình ta có Sig < 0.05, cho thấy mối tương quan giữa biến phụ thuộc và các biến độc lập trong mô hình có ý nghĩa thống kê tổng quát ở mức 99%. Kết quả nghiên cứu cho thấy có 6 thành phần tác động có ý nghĩa thống kê tới khả năng trả nợ của khách hàng cá nhân gồm: (1) Nhà ở, (2) Mục đích khoản vay, (3) Số người phụ thuộc, (4) Thu nhập, (5) Tuổi, (6) Quy mô khoản vay.

Viễn, N. T., Xuân, P. T. T., Huyền, L. T. T.(2020) đã chỉ ra những thuộc tính trọng tâm trong việc đánh giá khả năng trả nợ của khách hàng ở các nghiên cứu trước đó. Những thuộc tính bao gồm: thu nhập, kinh nghiệm làm việc, trình độ học vấn, lãi suất vay, tuổi tác, số tiền vay. Ngoài ra trong mô hình nghiên cứu, nhóm tác giả đã sử dụng thêm nhiều thuộc tính khác cũng có ý nghĩa thống kê đối với biến phụ thuộc. Đối với nhóm khách hàng có thu nhập thấp thì chưa thể kết luận khả năng trả nợ đúng hạn/ quá hạn của khách hàng chỉ với biến thu nhập, nên tác giả đã xét thêm thuộc tính số nguồn thu nhập.

3. Phương pháp và dữ liệu

3.1 Dữ liệu

Dữ liệu được mô phỏng gồm 100 khách hàng có quan hệ tín dụng. Trong đó có 54 khách hàng trả nợ đúng hạn và 46 khách hàng trả nợ không đúng hạn. Dữ liệu cân đối gần bằng 1:1 giữa 2 nhóm: Nhóm khách hàng trả nợ đúng hạn và nhóm khách hàng trả nợ không đúng hạn.

Mô tả các biến sử dụng

Biến dự báo là khả năng trả nợ đúng hạn của khách hàng. Đối với khách hàng trả nợ đúng hạn là 1 và khách hàng không trả nợ đúng hạn là 0.

Tên biến	Thang đo	Kí hiệu	Lí do lựa chọn các biến
Biến dự báo			
Khả năng trả nợ vay khách hàng cá nhân	1: Trả nợ đúng hạn	KNTN	Từ các bài nghiên cứu trước đó như nghiên cứu đánh giá khả năng trả nợ đúng hạn và phân loại khách hàng vay tín chấp tại NHTMCP Việt Nam Thịnh Vượng, nghiên cứu về khả năng trả nợ đúng hạn của khách hàng cá nhân được thực hiện tại Agribank Tân Hưng, Tỉnh Long An đã chỉ ra sự tương quan của các biến thu nhập, số tiền vay, số năm công tác, tài sản đảm bảo, số nguồn thu nhập, độ tuổi có sự ảnh hưởng đến khả năng trả nợ đúng hạn vay của khách hàng. Chính vì thế, ở bài nghiên cứu này nhóm tác giả đã lựa chọn các biến chứa thông tin dự báo đó. Bên cạnh đó, biến thu nhập còn lại được nhóm phát triển dựa trên biến thu nhập và chi phí dự phòng của khách hàng cá nhân.
	0: Trả nợ không đúng hạn		
Các biến chứa thông tin dự báo			
Thu nhập	0: Thấp (<= 20.000.000)	TN	
	1: Trung bình (>20.000.000 và <= 35.000.000)		
	2: Cao (>35.000.000)		
Số tiền vay	1: Nhỏ hơn KNTN****	STV	
	0: Lớn hơn KNTN		
Thu nhập còn lại	1: Cao (>20.000.000)	TNCL	
	0: Thấp (<= 20.000.000)		
Số năm công tác	0: 0 năm	SNCT	
	1: 1 năm		
	2: 2 năm		
	3: 3 năm		
	4: 4 năm		
	5: 5 năm		

	6: Trên 5 năm		
Số nguồn thu nhập	0: Ít nguồn thu nhập (X < 2)	SNTN	
	1: Nhiều nguồn thu nhập (X ≥ 2)		
Tài sản đảm bảo	0: Không thuộc sở hữu người vay (bảo lãnh)	TSDB	
	1: Thuộc sở hữu người vay		
Độ tuổi	0: X ≤ 30	DT	
	1: 30 < X ≤ 50		
	2: X > 50		
Trình độ	1: Cao (Sau đại học, đại học, cao đẳng)	TD	
	0: Thấp (Trung cấp, phổ thông, khác)		

Bảng 1: Mô tả các biến

*Thu nhập còn lại bằng thu nhập trừ đi các chi phí. Số năm công tác được tính bằng số năm khách hàng đang làm công việc hiện tại.

3.2 Các phương pháp sàng lọc biến

3.2.1 Correlation

Hệ số tương quan là một thước đo thống kê về độ mạnh yếu của mối quan hệ giữa các chuyển động tương đối của hai biến. Các giá trị nằm trong khoảng từ -1,0 đến 1,0. Một số được tính toán lớn hơn 1,0 hoặc nhỏ hơn -1,0 có nghĩa là đã xảy ra lỗi trong phép đo tương quan. Tương quan -1,0 cho thấy mối tương quan âm tuyệt đối, trong khi mức tương quan 1,0 cho thấy mối tương quan dương tuyệt đối. Tương quan 0,0 cho thấy không có mối quan hệ tuyến tính giữa chuyển động của hai biến. Phổ biến nhất là hệ số tương quan Pearson (R). Hệ số này chỉ ra độ mạnh và hướng của quan hệ tuyến tính giữa hai biến. Giá trị chính xác bằng 1,0 có nghĩa là có một mối quan hệ dương tuyệt đối giữa hai biến. Đối với một biến số tăng dương, thì biến số thứ hai cũng tăng dương. Giá trị -1,0 có nghĩa là có một mối quan hệ âm tuyệt đối giữa hai biến. Điều này cho thấy rằng các biến chuyển động ngược chiều nhau - đối với một biến số tăng dương thì

biến số thứ hai sẽ giảm xuống. Nếu mối tương quan giữa hai biến là 0 thì kết luận không có mối quan hệ tuyến tính giữa chúng. Độ mạnh của mối quan hệ thay đổi theo mức độ dựa trên giá trị của hệ số tương quan

3.2.2 Feature Importance

Feature Importance dùng để đánh giá tầm quan trọng của mỗi biến cho mỗi quyết định mà cây lựa chọn. Mỗi biến sẽ có giá trị Feature Importance trong khoảng từ 0 đến 1. Nếu biến mang giá trị 0, hoàn toàn không được sử dụng và giá trị 1 mang ý nghĩa hoàn hảo trong dùng để dự đoán mục tiêu. Tổng của feature importances luôn bằng 1. Feature Importance luôn mang giá trị dương, tuy nhiên nó không có giá trị nói lên mối liên hệ giữa biến và biến mục tiêu. Biến có feature importance thấp không có nghĩa là nó không chứa bất kì thông tin gì, nó chỉ thể hiện mô cây quyết định không chọn biến đó để phân loại, bởi vì có thể các biến khác đã bao hàm cả thông tin của biến không được chọn.

3.3 Cây quyết định

3.3.1 Giới thiệu cây quyết định

Thuật toán cây quyết định được xây dựng thành nhiều mô hình sử dụng rộng rãi cho cả việc phân loại cũng như hồi quy. Về bản chất, chúng học từ thứ bậc của các câu hỏi if/else và đưa ra quyết định. Mỗi nút trong cây đại diện cho một câu hỏi hoặc nút đầu cuối (còn được gọi là lá) có chứa câu trả lời. Các cạnh có thể kết nối với một câu trả lời hoặc câu hỏi tiếp theo. Một lá của cây khi chỉ chứa toàn dữ liệu mục tiêu thì được gọi là pure (thuần khiết)

3.3.2 Các kiểu cây quyết định

Các kiểu cây quyết định bao gồm:

- ID3: thuật toán nổi tiếng để xây dựng Decision Tree, áp dụng cho bài toán Phân loại (Classification) mà tất các các thuộc tính để ở dạng category
- C4.5: Kế thừa của ID3
- CART: Cây phân loại và cây hồi quy
- CHAID: Nhận diện tự động tương tác Chi bình phương…
- MARS: Mở rộng nhiều cây quyết định để giải quyết các biến dữ liệu số tốt hơn.
- Conditional Inference Trees: Cây suy luận có điều kiện

3.3.3 Ưu điểm và hạn chế

- Ưu điểm

Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bởi những lợi ích của nó:
Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả.
Có thể làm việc với cả dữ liệu số và dữ liệu phân loại.
Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê.

Có khả năng là việc với dữ liệu lớn

- Nhược điểm

Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
Cây quyết định hay gặp vấn đề overfitting.

3.3.4 Những ứng dụng của cây quyết định

1. Đánh giá các cơ hội tăng trưởng trong tương lai

Một trong những ứng dụng của cây quyết định liên quan đến việc đánh giá các cơ hội tăng trưởng tiềm năng cho các doanh nghiệp dựa trên dữ liệu lịch sử. Dữ liệu lịch sử về doanh số có thể được sử dụng trong cây quyết định có thể dẫn đến việc thực hiện những thay đổi căn bản trong chiến lược của một doanh nghiệp để giúp mở rộng và tăng trưởng.

2. Sử dụng dữ liệu nhân khẩu học để tìm khách hàng tiềm năng

Một ứng dụng khác của cây quyết định là sử dụng dữ liệu nhân khẩu học để tìm khách hàng tiềm năng. Chúng có thể giúp sắp xếp hợp lý ngân sách tiếp thị và đưa ra quyết định sáng suốt về thị trường mục tiêu mà doanh nghiệp đang tập trung vào. Trong trường hợp không có cây quyết định, doanh nghiệp có thể chi tiêu thị trường tiếp thị của mình mà không tính đến nhân khẩu học cụ thể, điều này sẽ ảnh hưởng đến doanh thu tổng thể của doanh nghiệp.

3. Phục vụ như một công cụ hỗ trợ trong một số lĩnh vực

Người cho vay cũng sử dụng cây quyết định để dự đoán xác suất khách hàng không trả được nợ bằng cách áp dụng tạo mô hình dự đoán sử dụng dữ liệu trong quá khứ của khách hàng. Việc sử dụng công cụ hỗ trợ cây quyết định có thể giúp người cho vay đánh giá mức độ tín nhiệm của khách hàng để ngăn ngừa tổn thất.
Cây quyết định cũng có thể được sử dụng trong nghiên cứu hoạt động trong việc lập kế hoạch hậu cần và quản lý chiến lược. Họ có thể giúp xác định các chiến lược phù hợp sẽ giúp công ty đạt được các mục tiêu đã định. Các lĩnh vực khác mà cây quyết định có thể được áp dụng bao gồm kỹ thuật, giáo dục, luật, kinh doanh, chăm sóc sức khỏe và tài chính.

3.4 Mô hình Logistic

3.4.1. Khái niệm về mô hình Logistic

Mô hình Logistic (Maddala[1], 1984) là mô hình định lượng trong đó biến phụ thuộc là biến giả, chỉ nhận 2 giá trị là 0 hoặc 1. Mô hình này được ứng dụng rộng rãi trong phân tích kinh tế nói chung và rủi ro tín dụng nói riêng. Cụ thể hơn, mô hình này có thể giúp Ngân hàng xác định khả năng khách hàng sẽ có rủi ro tín dụng (biến phụ thuộc) trên cơ sở sử dụng các nhân tố có ảnh hưởng đến khách hàng (biến độc lập).

3.4.2 Ưu điểm và hạn chế

- Ưu điểm

Mô hình hồi quy Logistic huấn luyện nhanh và đưa ra kết quả dự báo rất nhanh.
Có thể áp dụng dễ dàng với tập dữ liệu lớn và thưa thớt.
Dễ dàng trong việc hiểu cách đưa ra dự đoán của mô hình.

Dễ dàng giải thích tầm quan trọng của các biến đối với mô hình và sự ảnh hưởng của các biến đó.

- Nhược điểm

Thường khó trong việc hiểu được hết cách đưa ra các hệ số trong mô hình, điều này thường đặc biệt đúng nếu trong tập dữ liệu có các biến tương quan cao, và trong những trường hợp đó khó để giải thích.

3.4.3 Các loại mô hình Logistic

Có ba loại mô hình hồi quy logistic:

- Mô hình hồi quy logistic nhị phân là một loại hồi quy logistic trong đó biến phản hồi chỉ có thể thuộc về hai loại.
- Mô hình hồi quy logistic đa thức là một loại hồi quy logistic trong đó biến phản hồi có thể thuộc một trong ba hoặc nhiều loại và không có thứ tự tự nhiên giữa các loại.
- Mô hình hồi quy logistic thông thường là một loại hồi quy logistic trong đó biến phản hồi có thể thuộc một trong ba hoặc nhiều loại và có thứ tự tự nhiên giữa các loại.

3.4.4 Những ứng dụng của mô hình Logistic

Hồi quy logistic thường được sử dụng cho các bài toán dự đoán và phân loại. Dưới đây là một trong số những ứng dụng của mô hình Logistic:

- Phát hiện gian lận: Mô hình hồi quy logistic có thể giúp các nhóm xác định các bất thường về dữ liệu, dự đoán về gian lận. Một số hành vi hoặc đặc điểm nhất định có thể có mối liên hệ cao hơn với các hoạt động gian lận, điều này đặc biệt hữu ích cho ngân hàng và các tổ chức tài chính khác trong việc bảo vệ khách hàng của họ. Các công ty dựa trên SaaS cũng đã bắt đầu áp dụng các phương pháp này để loại bỏ các tài khoản người dùng giả mạo khỏi bộ dữ liệu của họ khi tiến hành phân tích dữ liệu xung quanh hiệu suất kinh doanh.
- Dự đoán bệnh tật: Trong y học, phương pháp phân tích này có thể được sử dụng để dự đoán khả năng mắc bệnh hoặc ốm đau cho một nhóm dân số nhất định. Các tổ chức chăm sóc sức khỏe có thể thiết lập dịch vụ chăm sóc phòng ngừa cho những cá nhân có xu hướng mắc các bệnh cụ thể cao hơn.
- Dự đoán sự xáo trộn: Các hành vi cụ thể có thể là dấu hiệu cho thấy sự xáo trộn trong các phòng ban khác nhau của tổ chức. Ví dụ, đội ngũ nhân sự và quản lý muốn biết liệu có những người có thành tích cao trong công ty có nguy cơ rời bỏ tổ chức hay không. Loại thông tin chi tiết này có thể thúc đẩy các cuộc trò chuyện giữa các nhân viên để hiểu thêm về các vấn đề trong công ty, chẳng hạn như văn hóa hoặc lương thưởng. Ngoài ra, tổ chức kinh doanh có thể muốn tìm hiểu xem khách hàng nào của họ có rủi ro khi kinh doanh ở nơi khác. Điều này có thể thúc đẩy các nhóm thiết lập chiến lược duy trì để tránh bị mất doanh thu.

4. Kết quả

4.1 Kết quả thống kê mô tả

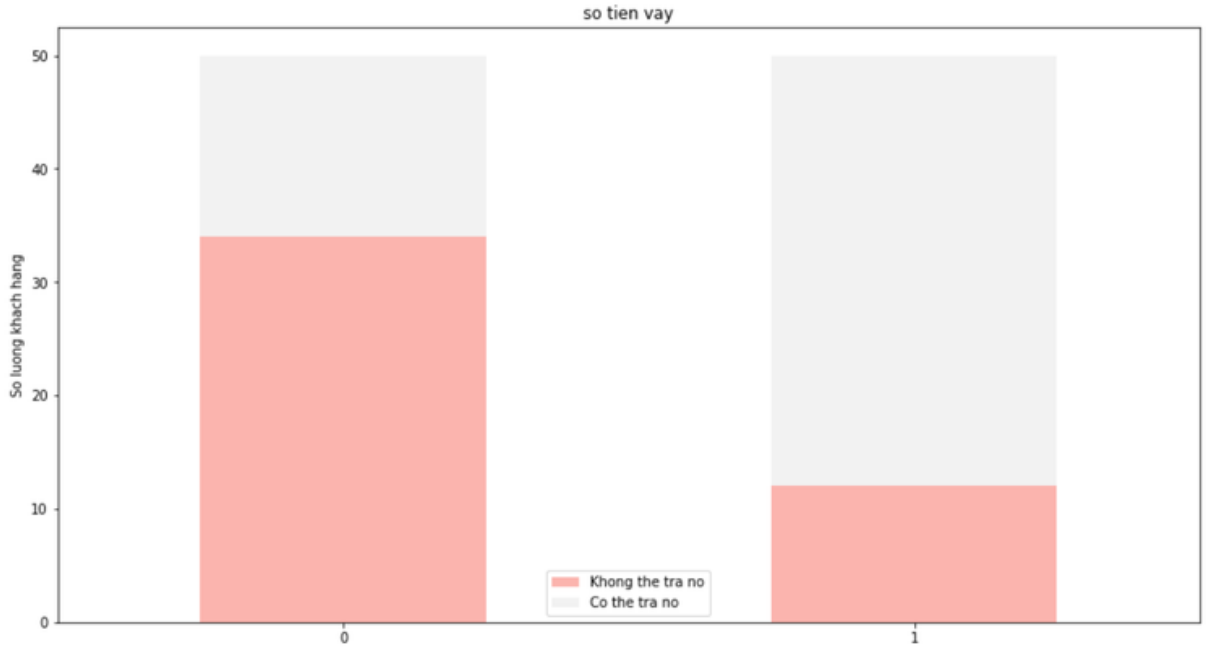
4.1.1 Thống kê mô tả các biến chứa thông tin dự báo

	Khả năng trả nợ vay khách hàng cá nhân	Thu nhập	Số tiền vay	Thu nhập còn lại	Số năm công tác**	Số nguồn thu nhập *	Tài sản đảm bảo	Độ tuổi *	Trình độ *
count	100	100	100	100	100	100	100	100	100
unique	2	3	2	2	7	2	2	3	2
top	1: Trả nợ đúng hạn	0: Thấp	1: Nhỏ hơn KNTN****	1: Cao	6: Trên 5 năm	0: Ít nguồn thu nhập (X < 2)	1: Thuộc sở hữu người vay	1: 30 < X ≤ 50	1: Cao (Sau đại học, đại học, cao đẳng)
freq	54	37	50	51	23	61	56	46	61

Bảng 2: Thống kê mô tả các biến chứa thông tin dự báo

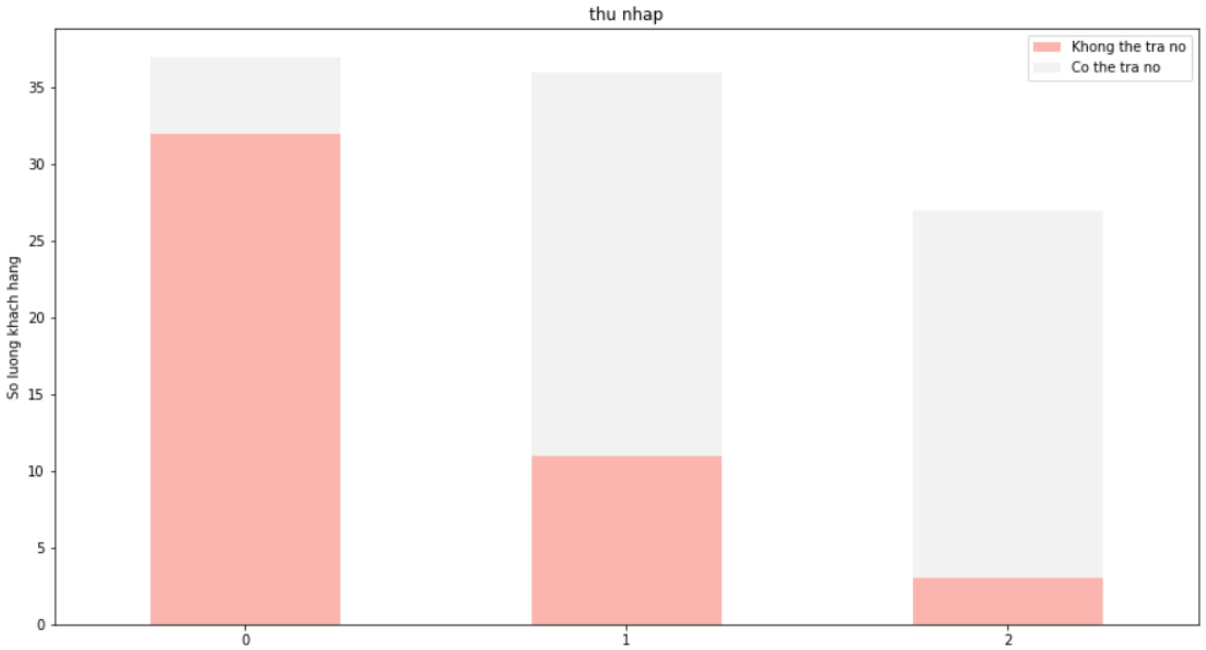
Bộ dữ liệu gồm 100 quan sát 1 biến phụ thuộc và 8 biến độc lập. Biến mục tiêu *khả năng trả nợ đúng hạn vay khách hàng cá nhân* bao gồm 2 nhóm trong đó *1: Trả nợ đúng hạn* xuất hiện nhiều nhất với 54 quan sát. Biến *Thu nhập* có 3 nhóm giá trị trong đó *0: Thấp* xuất hiện nhiều nhất với 37 quan sát. Biến *Số tiền vay* có 2 nhóm trong đó *1: Nhỏ hơn KNTN**** có số quan sát bằng với nhóm 0 là 50 quan sát. Biến *Thu nhập còn lại* có 2 nhóm trong đó *1: Cao* có số quan sát là 51 quan sát. Biến *Số năm công tác*** có 7 nhóm trong đó *6: Trên 5 năm* xuất hiện nhiều nhất có số quan sát là 23 quan sát. Biến *Số nguồn thu nhập* có 2 nhóm trong đó *0: Ít nguồn thu nhập (X < 2)* xuất hiện nhiều nhất có số quan sát là 61 quan sát. Biến *Tài sản đảm bảo* có 2 nhóm trong đó *1: Thuộc sở hữu người vay* xuất hiện nhiều nhất có số quan sát là 56 quan sát.
Biến *Độ tuổi ** có 3 nhóm trong đó *1: 30 < X ≤ 50* xuất hiện nhiều nhất có số quan sát là 46 quan sát. Biến *Trình độ* có 2 nhóm trong đó *1: Cao (Sau đại học, đại học, cao đẳng)* xuất hiện nhiều nhất có số quan sát là 61 quan sát.

4.1.2 Sự ảnh hưởng của các biến chứa thông tin dự báo đến biến dự báo



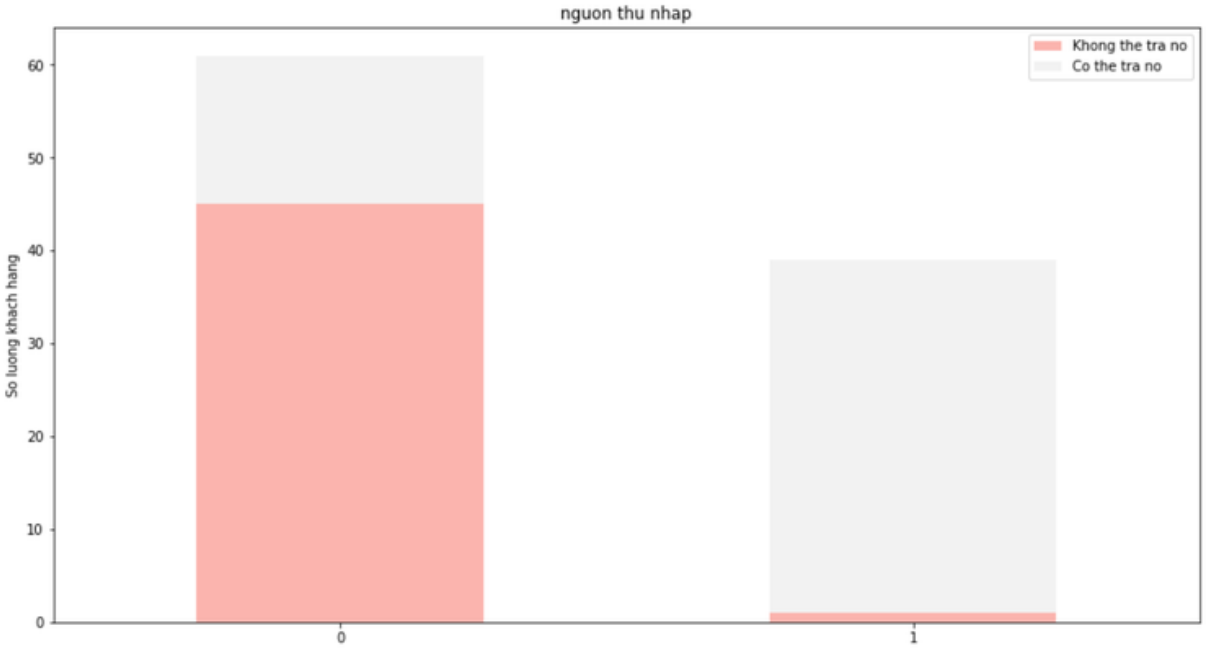
Hình 1: Sự ảnh hưởng của số tiền vay đến khả năng trả nợ đúng hạn

Số tiền vay của khách hàng lớn hơn khả năng trả nợ đúng hạn thì khách hàng đó có khả năng trả nợ không đúng hạn cao hơn so với những khách hàng có số tiền vay nhỏ hơn khả năng trả nợ đúng hạn của họ.



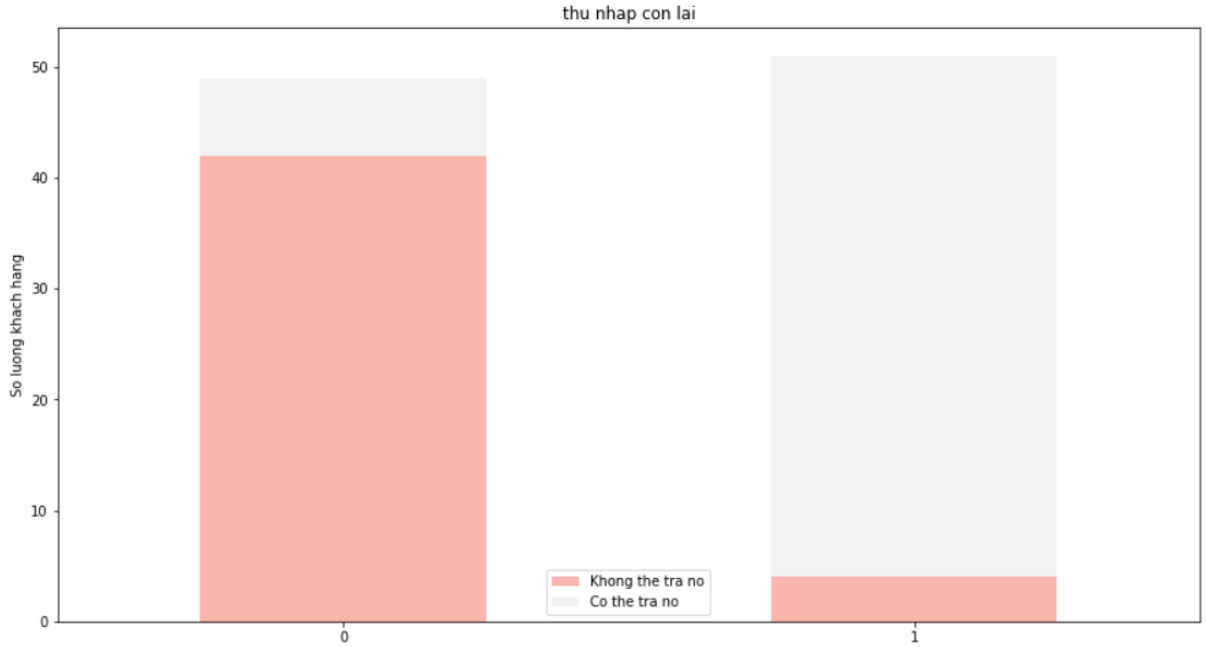
Hình 2: Sự ảnh hưởng của thu nhập đến khả năng trả nợ đúng hạn

Biến thu nhập có ảnh hưởng đến khả năng trả nợ đúng hạn của khách hàng cá nhân. Thu nhập càng thấp (<20.000.000) thì khả năng trả nợ đúng hạn càng thấp và ngược lại. Có thể thay khi mức thu nhập ở mức trung bình và cao thì số trường hợp trả nợ không đúng hạn đã giảm đi.



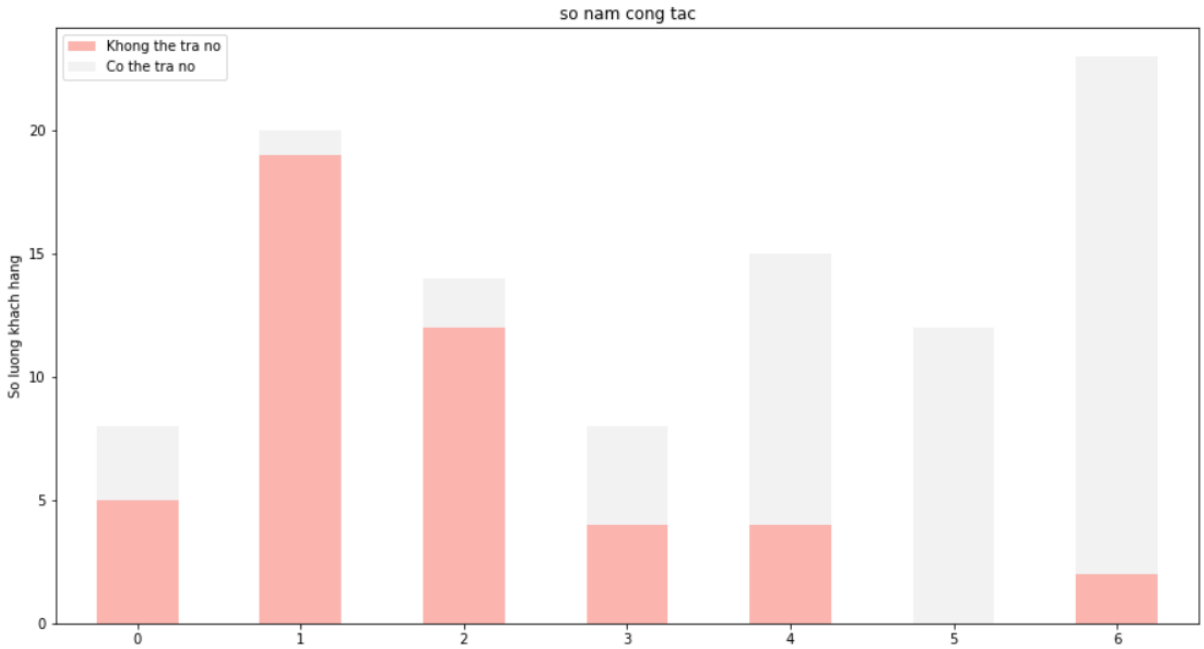
Hình 3: Sự ảnh hưởng của số nguồn thu nhập đến khả năng trả nợ đúng hạn

Số nguồn thu nhập cũng có ảnh hưởng đến khả năng trả nợ đúng hạn của khách hàng. Có thể thấy rằng khách hàng có ít nguồn thu nhập (dưới 2) thì có khả năng trả nợ đúng hạn kém hơn so với nhóm có nhiều nguồn thu nhập.



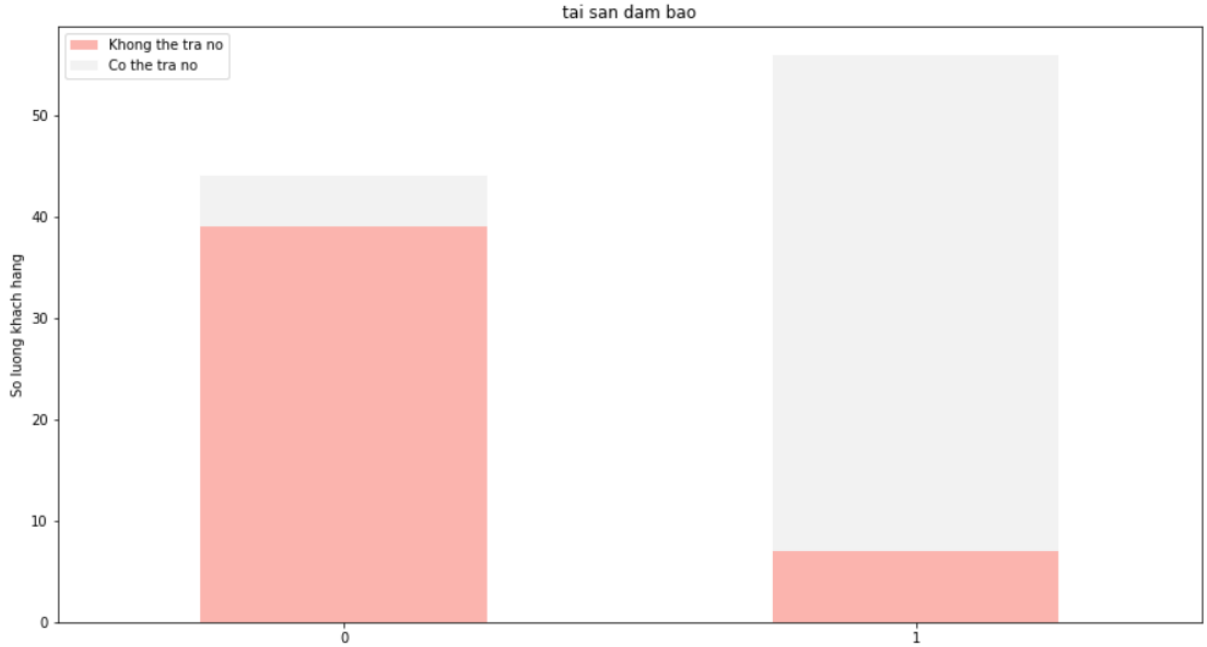
Hình 4: Sự ảnh hưởng của thu nhập còn lại đến khả năng trả nợ đúng hạn

Khi thu nhập còn lại của khách hàng càng cao (>20.000.000) thì khả năng trả nợ đúng hạn được tăng lên và ngược lại.



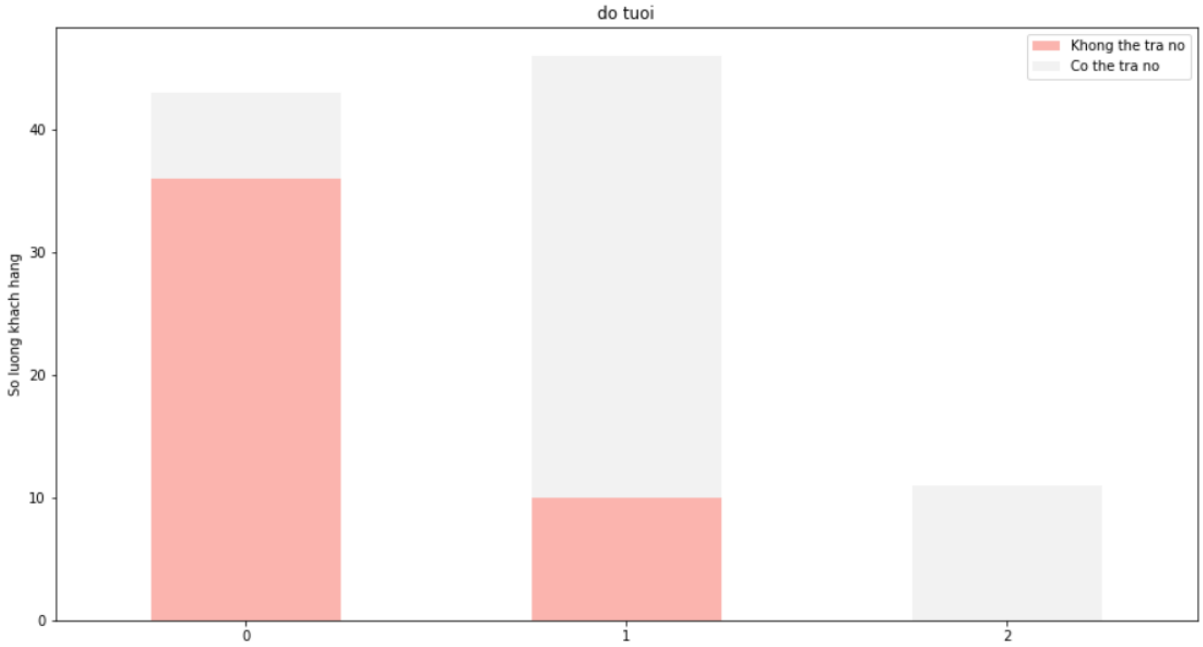
Hình 5: Sự ảnh hưởng của số năm công tác đến khả năng trả nợ đúng hạn

Số năm công tác cũng ảnh hưởng đến khả năng trả nợ đúng hạn của khách hàng. Số khách hàng trả nợ đúng hạn tăng lên khi họ có số năm công tác càng cao. Nhóm khách hàng không trả nợ được đúng hạn chủ yếu tập trung ở khách hàng có số năm công tác từ 0 năm đến 2 năm.



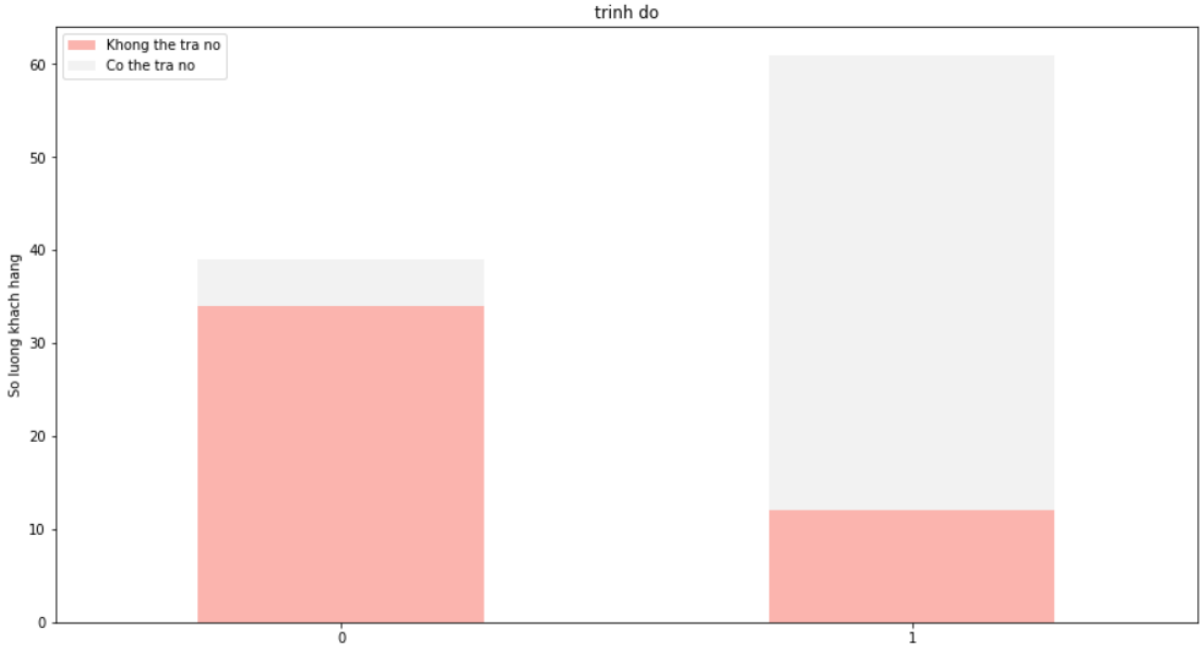
Hình 6: Sự ảnh hưởng của tài sản đảm bảo vay đến khả năng trả nợ đúng hạn

Số lượng khách hàng không thể trả nợ đúng hạn được tập trung chủ yếu ở những khách hàng không có tài sản đảm bảo và ngược lại.



Hình 7: Sự ảnh hưởng của độ tuổi đến khả năng trả nợ đúng hạn

Thuộc tính độ tuổi cũng có sự ảnh hưởng đến khả năng trả nợ đúng hạn. Số lượng khách hàng không trả nợ được đúng hạn tập trung chủ yếu ở nhóm khách hàng có độ tuổi dưới 30 tuổi và được giảm dần khi tăng độ tuổi lên.



Hình 8: Sự ảnh hưởng của trình độ đến khả năng trả nợ đúng hạn

Số lượng khách hàng không trả nợ được đúng hạn tập trung chủ yếu ở nhóm trình độ thấp (trung cấp, phổ thông, khác) và ngược lại.

4.2 Kết quả sàng lọc biến

4.2.1 Correlation

	kha nang tra no cua khach hang
thu nhap	0.616799
so tien vay	0.441415
thu nhap con lai	0.781059
so nam cong tac	0.668034
so nguon thu nhap	0.696852
tai san dam bao	0.758292
do tuoi	0.645442
trinh do	0.660652

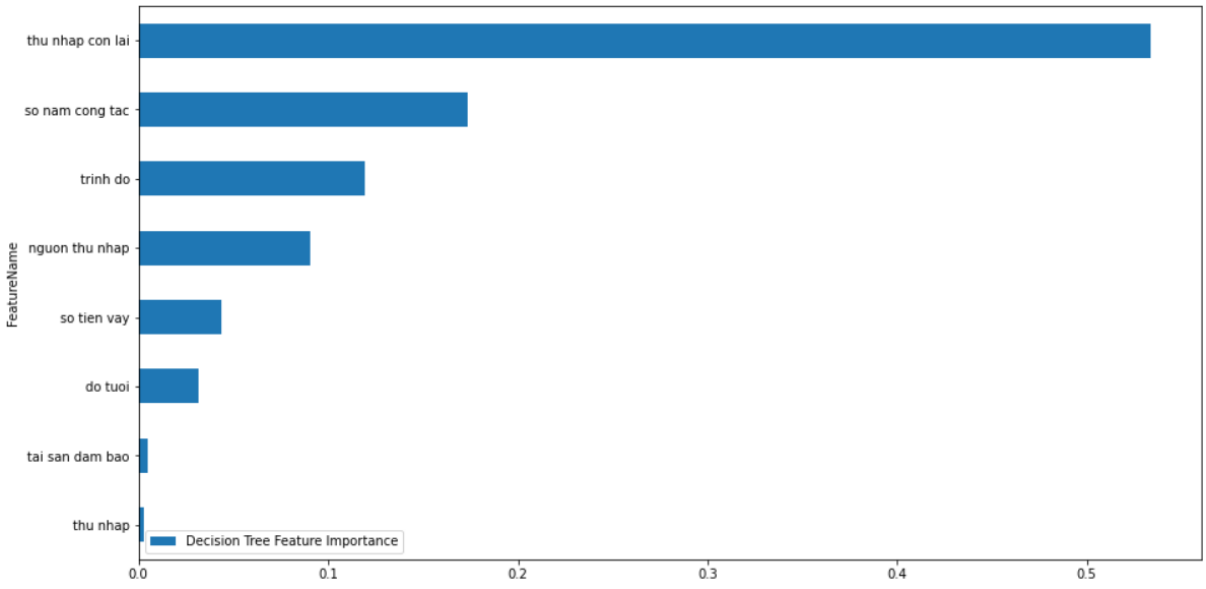
Bảng 3: Kết quả Correlation của các biến chứa thông tin dự báo và biến dự báo

Kết quả: Tất cả các biến chứa thông tin dự báo trong bài nghiên cứu này đều có mức độ tương quan hay sự ảnh hưởng đối với biến khả năng trả nợ đúng hạn.
Với biến thu nhập còn lại có sự ảnh hưởng lớn nhất với hệ số tương quan là 0.78, khi tăng thu nhập còn lại lên thì khả năng trả nợ đúng hạn của khách hàng cũng tăng lên. Tương tự, biến tài sản đảm bảo có sự ảnh hưởng lớn đến khả năng trả nợ đúng hạn với hệ số tương quan là khoảng 0.76.
Các biến như thu nhập, số năm công tác, số nguồn thu nhập, độ tuổi, trình độ đều có sự ảnh hưởng đối với khả năng trả nợ đúng hạn và có hệ số tương quan giao động từ 0.61 đến 0.697. Tất cả các biến trên đều có sự tương quan dương với khả năng trả nợ đúng hạn, điều đó có nghĩa là khi tăng giá trị của các biến thì khả năng trả nợ đúng hạn cũng được tăng lên.
Riêng biến số tiền vay cũng có sự ảnh hưởng đến khả năng trả nợ đúng hạn với hệ số tương quan là 0.44, nhưng không có sự ảnh hưởng lớn như những biến còn lại.

4.2.2 Important Feature

	FeatureName	Decision Tree Feature Importance
2	chi phi du phong	0.533553
3	so nam cong tac	0.168597
7	trinh do	0.11959
4	so nguon thu nhap	0.090236
1	so tien vay	0.043758
6	do tuoi	0.031478
5	tai san dam bao	0.009837
0	thu nhap	0.002951

Bảng 4: Kết quả Important Feature



Hình 9: Kết quả Feature Importance

Như hình trên, biến được sử dụng để phân tách dữ liệu đầu tiên là *thu nhap con lai*, nó cũng mang ý nghĩa là biến quan trọng nhất. Nếu nhìn vào sơ đồ cây quyết định thì biến này thuộc tầng đầu tiên để bắt đầu phân loại trong sơ đồ là hoàn toàn hợp lý. Thuộc tính *so nam cong tac* xếp thứ hai thể hiện mức độ quan trọng của thuộc tính này trong việc phân loại dữ liệu, đối chiếu với sơ đồ cây thuộc tính *so nam cong tac* xuất hiện nhiều

nhất với 5 lần. Ở tầng tiếp theo của sơ đồ cây quyết định là hai biến *trinh do* và *so nguon thu nhap*, hai biến này ở trong bảng xếp hạng feature importances xếp lần lượt ở thứ 3 và thứ 4. Thuộc tính thu nhập có kết quả feature importances thấp nhất cho thấy khả năng phân loại của biến này trong mô hình là chưa rõ rệt.

4.3 Kết quả mô hình Logistic

LogisticRegression(random_state=44)				
	Precision	Recall	F1_score	Support
0	1	0.85	0.92	13
1	0.89	1	0.94	17
Accuracy			0.93	30
Macro average	0.95	0.92	0.93	30
Weighted average	0.94	0.93	0.93	30

Bảng 5: Kết quả mô hình hồi quy Logistic

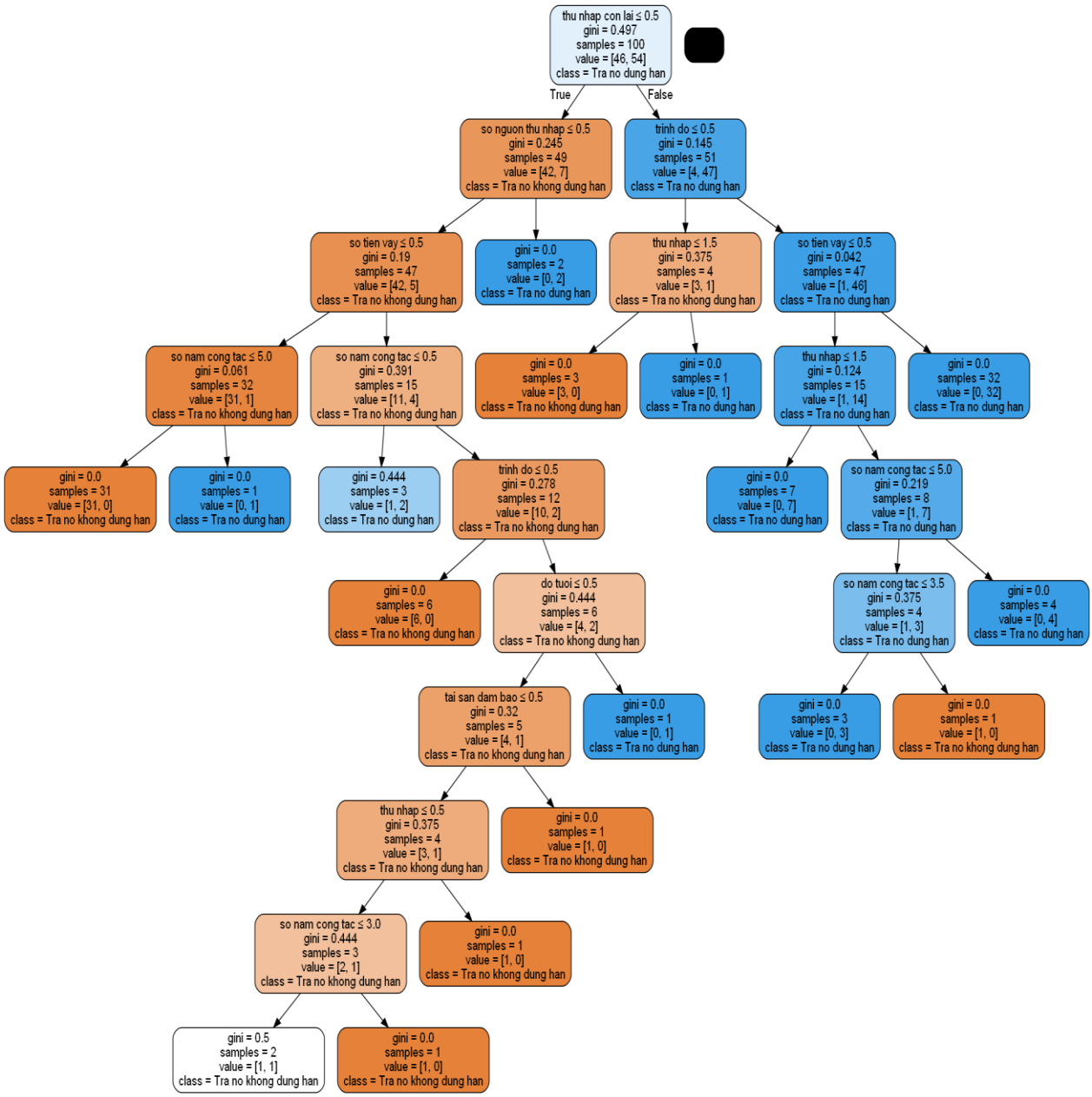
Khả năng dự báo chính xác của mô hình là 93% cho cả 2 lớp *0: tỉ lệ trả nợ không đúng hạn và 1: Trả nợ đúng hạn.* Nhìn vào bảng báo cáo phân loại ta có thể thấy ở loại *0: tỉ lệ trả nợ không đúng hạn* có số quan sát là 13 và tỉ lệ dự đoán đúng trên trên tổng số dự đoán là 100% tuy nhiên khả năng dự báo đúng so với số quan sát thực tế chỉ là 85% đạt mức *F1_score* là 92%. Ở lớp mục tiêu là *1: Trả nợ đúng hạn* có số quan sát là 17 và tỉ lệ dự đoán đúng trên trên tổng số dự đoán là 89% tuy nhiên khả năng dự báo đúng so với số quan sát thực tế là 100% đạt mức *F1_score* là 94% cho thấy được khả năng không bỏ sót bất kì biến mục tiêu *1: Trả nợ đúng hạn.* Mô hình có khả năng dự báo tốt nhóm trả được nợ đúng hạn tuy nhiên về rủi ro không trả được nợ đúng hạn của khách hàng mô hình dự báo chưa thực sự tốt.

4.4 Kết quả cây quyết định

DecisionTreeClassifier(random_state=44)				
	Precision	Recall	F1_score	Support
0	0.75	0.92	0.83	13
1	0.93	0.76	0.84	17
Accuracy			0.83	30
Macro average	0.84	0.84	0.83	30
Weighted average	0.85	0.83	0.83	30

Bảng 6: Kết quả mô hình cây quyết định

Khả năng dự báo chính xác của mô hình là 83% cho cả 2 lớp *0: trả nợ không đúng hạn và 1: Trả nợ đúng hạn.* Nhìn vào bảng báo cáo phân loại ta có thể thấy ở loại *0: tỉ lệ trả nợ không đúng hạn* có số quan sát là 13 và tỷ lệ dự đoán đúng trên trên tổng số dự đoán là 75% tuy nhiên khả năng dự báo đúng so với số quan sát thực tế là rất tốt 92% đạt mức *F1_score* là 83%.
Ở lớp mục tiêu là *1: Trả nợ đúng hạn* có số quan sát là 17 và tỷ lệ dự đoán đúng trên trên tổng số dự đoán là 93% tuy nhiên khả năng dự báo đúng so với số quan sát thực tế chỉ là 76% đạt mức *F1_score* là 84%. Mô hình có khả năng phát hiện nhóm *0: trả nợ không đúng hạn* có thể được dùng để giúp các ngân hàng tránh phải các rủi ro không trả được nợ đúng hạn. Tuy nhiên sẽ bỏ sót nhiều đối tượng cho vay tiềm năng vì khả năng dự báo nhóm trả nợ được là chưa thực sự tốt.



Hình 10: Kết quả cây quyết định

Yếu tố kỹ thuật

Bộ dữ liệu trước khi được đưa vào huấn luyện cây quyết định bao gồm 100 khách hàng (46 khách hàng không thể trả nợ và 54 khách hàng không thể trả nợ). Hệ số gini là 0.497 (giá trị lớn nhất của gini là 0.5 và nhỏ nhất là 0, mục tiêu của thuật toán là sử dụng các điều kiện để chia ra các nhóm sao cho gini bé nhất và giá trị đó sẽ giảm dần sau mỗi lần phân tách), điều đó chứng tỏ rằng dữ liệu hiện tại của 2 nhóm khách hàng đang đồng đều và chưa thể phân loại rõ ràng. Sau khi đưa bộ dữ liệu vào huấn luyện, cây quyết định đã dự đoán bằng cách xếp hạng các thuộc tính quan trọng nhất theo thứ tự giảm dần cùng với sử dụng mệnh đề điều kiện “nếu”-”thì” để rẽ nhánh từ “node gốc”.

Đầu tiên, cây quyết định xác định thuộc tính thu nhập còn lại là thuộc tính quan trọng nhất.

Trong trường hợp thu nhập còn lại thấp (<= 0.5) có 49 khách hàng được phân vào nhóm (Nhóm 1) gồm: 42 khách hàng không thể trả nợ đúng hạn và 7 khách hàng có thể trảnợ đúng hạn. Hệ số gini là 0.245 và bé hơn hệ số gini ban đầu trước khi phân tách (0.497). Trong trường hợp thu nhập còn lại cao, có 51 khách hàng được phân vào nhóm (Nhóm 2) gồm có 47 khách hàng có thể trả nợ được đúng hạn và 4 khách hàng không thể trả được nợ đúng hạn. Hệ số gini là 0.145 và bé hơn hệ số gini ban đầu trước khi phân tách (0.497). Dựa vào kết quả (gini) của lần phân tách đầu tiên dựa vào thuộc tính thu nhập còn lại, chúng ta đã có thể phân loại được khá rõ ràng 2 nhóm khách hàng với giả thiết rằng những khách hàng có thu nhập còn lại thấp nhiều khả năng sẽ không trả được nợ đúng hạn và ngược lại và độ chính xác trong dự đoán khi sử dụng giả thiết này là 42/49 (85.7%) đối với nhóm khách hàng không trả được nợ đúng hạn và 47/51 (92.2%) đối với nhóm khách hàng trả được nợ đúng hạn

Đối với Nhóm 1, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số nguồn thu nhập. Trong trường hợp số nguồn thu nhập bé hơn 2 (<= 0.5), có 47 khách hàng được phân vào nhóm (Nhóm 3) gồm 42 khách hàng không thể trả nợ đúng hạn và 5 khách hàng có thể trả nợ đúng hạn. Hệ số gini là 0.19 và bé hơn hệ số gini ban đầu trước khi phân tách (0.245). Trong trường hợp điều kiện nguồn thu nhập lớn hơn 2, có 2 khách hàng được phân vào nhóm (Nhóm 4) gồm 2 khách hàng có thể trả nợ đúng hạn và 0 có khách hàng nào không thể trả được nợ đúng hạn. Hệ số gini là 0, đã bé hơn hệ số gini ban đầu trước khi phân tách (0.245) do đó thuật toán sẽ không phân tách lớp ở nhóm này. Dựa vào kết quả (gini) của lần phân tách thứ hai, dựa vào hai thuộc tính là thu nhập còn lại và số nguồn thu nhập đã có thể nâng cao độ chính xác khi phân loại 2 nhóm khách hàng với giả thiết rằng những khách hàng có thu nhập còn lại thấp và số nguồn thu nhập bé hơn 2 nhiều khả năng sẽ không trả được nợ đúng hạn và ngược lại những khách hàng có thu nhập còn lại thấp nhưng số nguồn thu nhập lớn hơn 2 nhiều khả năng sẽ trả được nợ đúng hạn với độ chính xác trong dự đoán khi sử dụng giả thiết này là 42/47 (89.4%) đối với nhóm khách hàng không trả được nợ đúng hạn và 2/2 (100%) đối với nhóm khách hàng trả được nợ đúng hạn.

Tiếp tục với Nhóm 2, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là trình độ. Trong trường hợp trình độ thấp (<= 0.5), có 4 khách hàng được phân vào nhóm (Nhóm 5) gồm 3 khách hàng không thể trả nợ đúng hạn và 1 khách hàng có thể trả nợ đúng hạn. Hệ số gini là 0.375 và lớn hơn hệ số gini ban đầu trước khi phân tách (0.145), điều này xảy ra khi tỷ lệ chênh lệch giữa 2 nhóm khách hàng không lớn bằng Nhóm 2 (3:1 so với 47:4). Trong trường hợp trình độ cao, có 47 khách hàng được phân vào nhóm (Nhóm 6) gồm 47 khách hàng có thể trả nợ đúng hạn và chỉ có 1 khách hàng không thể trả được nợ đúng hạn. Hệ số gini là 0.042 và bé hơn hệ số gini ban đầu trước khi phân tách (0.145). Từ kết quả (gini) của lần phân tách thứ ba dựa vào hai thuộc tính là thu nhập khả dụng còn lại và trình độ, chúng ta có thể nâng cao độ chính xác khi phân loại 2 nhóm khách hàng với giả thiết rằng những khách hàng có thu nhập khả dụng còn lại cao nhưng trình độ thấp nhiều khả năng không trả được nợ đúng hạn. Ngược lại, những khách hàng có thu nhập khả dụng còn lại cao và trình độ cũng cao thì nhiều khả năng sẽ trả được nợ đúng hạn với độ chính xác trong dự đoán khi sử dụng giả thiết này là 3/4 (75%) đối với nhóm khách hàng không trả được nợ đúng hạn và 46/47 (97.9%) đối với nhóm khách hàng trả được nợ đúng hạn.

Đối với với Nhóm 3, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số tiền vay. Nếu số tiền vay lớn hơn khả năng trả nợ có 32 khách hàng được phân vào nhóm (Nhóm 7) gồm 31 khách hàng không thể trả nợ đúng hạn và 1 khách hàng có thể trả nợ đúng hạn. Hệ số gini là 0.061 và bé hơn hệ số gini ban đầu trước khi phân tách (0.19). Trong trường hợp số tiền vay bé hơn khả năng trả nợ đúng hạn, có 15 khách hàng được phân vào nhóm (Nhóm 8) gồm 11 khách hàng không thể trả nợ đúng hạn và 4 khách hàng có thể trả nợ đúng hạn. Hệ số gini là 0.391 và lớn hơn hệ số gini ban đầu trước khi phân tách (0.19), điều này là do tỷ lệ chênh lệch giữa 2 nhóm khách hàng không lớn bằng Nhóm 3 (11:4 so với 42:5). Dựa vào kết quả (gini) của lần phân tách thứ tư chúng ta có thể thấy rằng, thuật toán đã thêm 1 thuộc tính vào 2 thuộc tính sẵn có ở lần phân tách trước để có thể phân loại chính xác hơn khi cùng với một giả thiết là những khách hàng có thu nhập còn lại thấp và số nguồn thu nhập bé hơn 2. Nếu số tiền vay lớn hơn khả năng trả nợ thì sẽ có khả năng vỡ nợ cao hơn so với khi số tiền vay bé hơn khả năng trả nợ đúng hạn với xác suất khách hàng không trả được nợ đúng hạn lần lượt là 31/32 (96.9%) và 11/15 (73.3%).

Tiếp tục với Nhóm 5, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số nguồn thu nhập. Trong trường hợp nguồn thu nhập bé hơn 2, có 3 khách hàng được phân vào nhóm (Nhóm 9) và cả 3 khách hàng đều không có khả năng trả nợ đúng hạn. Hệ số gini là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.375). Trong trường hợp số nguồn thu nhập lớn hơn 2 thì 1 khách hàng được phân vào nhóm (Nhóm 10) và khách hàng đó có khả năng trả nợ đúng hạn. Hệ số gini là 0, bé hơn hệ số gini ban đầu trước khi phân tách (0.375). Dựa vào kết quả (gini) của lần phân tách thứ năm thuật toán đã thêm 1 thuộc tính vào 2 thuộc tính sẵn có ở lần phân tách trước để có thể phân loại chính xác hơn khi cùng với một giả thiết là những khách hàng có thu nhập còn lại cao nhưng trình độ thấp, nếu số nguồn thu nhập bé hơn 2 thì chắc chắn sẽ không trả được nợ đúng hạn còn nếu nguồn thu nhập cao vẫn sẽ trả được nợ đúng hạn với xác suất dự đoán là 100% (3/3 đối với khách hàng không trả nợ được đúng hạn và 1/1 đối với khách hàng trả nợ đúng hạn). Đồng thời hệ số gini của cả Nhóm 9 và Nhóm 10 đều bằng 0 nên thuật toán sẽ không tiếp tục phân tách ở hai nhóm này.

Tiếp tục với Nhóm 6, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số tiền vay. Trong trường hợp số tiền cho vay lớn hơn khả năng trả nợ, có 15 khách hàng được phân vào nhóm (Nhóm 11) gồm 14 khách hàng có thể trả nợ và 1 khách hàng không có khả năng trả nợ đúng hạn. Hệ số gini là 0.124 và lớn hơn hệ số gini ban đầu trước khi phân tách (0.042), điều này do tỉ lệ chênh lệch giữa 2 nhóm khách hàng không lớn bằng Nhóm 6 (1:14 so với 1:46). Trong trường hợp số tiền cho vay bé hơn khả năng trả nợ, có 32 khách hàng được phân vào nhóm (Nhóm 12) và cả 32 khách hàng này đều có khả năng trả nợ đúng hạn. Hệ số gini bằng 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.042). Điều đó cũng đồng nghĩa thuật toán sẽ không tiếp tục phân tách ở nhóm này. Dựa vào kết quả (gini) của lần phân tách thứ sáu chúng ta có thể thấy rằng, thuật toán đã thêm 1 thuộc tính vào 2 thuộc tính sẵn có ở lần phân tách trước để có thể phân loại chính xác hơn khi cùng với một giả thiết là những khách hàng có thu nhập còn lại cao và trình độ cao, nếu số tiền vay bé hơn khả năng trả nợ thì sẽ có khả năng trả nợ

đúng hạn cao hơn so với khi số tiền vay lớn hơn khả năng trả với xác suất khách hàng trả đúng hạn lần lượt là 14/15 (93.3%) và 32/32 (100%).

Tiếp tục với Nhóm 7, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số năm công tác. Trong trường hợp số năm công tác ít hơn hoặc bằng 5 năm, có 31 khách hàng được phân vào nhóm (Nhóm 13) gồm 31 khách hàng đều không có khả năng trả nợ đúng hạn. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.061). Trong trường hợp số năm công tác lớn hơn 5 năm, có 1 khách hàng được phân vào nhóm (Nhóm 14) và khách hàng đó có khả năng trả nợ đúng hạn. Hệ số gini là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.061). Dựa vào kết quả (gini) của lần phân tách thứ bảy thuật toán đã thêm 1 thuộc tính vào 3 thuộc tính sẵn có ở lần phân tách trước để có thể phân loại chính xác hơn khi cùng với một giả thiết là những khách hàng có thu nhập còn lại thấp, số nguồn thu nhập bé hơn 2 có số tiền vay nhỏ hơn khả năng trả nợ đúng hạn, nếu có số năm công tác dưới 5 năm thì sẽ không có khả năng trả nợ đúng hạn và nếu có số năm công tác trên 5 năm thì sẽ có khả năng trả nợ đúng hạn với xác suất dự đoán là 100% (31/31 đối với khách hàng không trả nợ được và 1/1 đối với khách hàng trả nợ được). Đồng thời hệ số gini của cả Nhóm 13 và Nhóm 14 đều bằng 0 nên thuật toán sẽ không tiếp tục phân tách ở hai nhóm này.

Tiếp tục với Nhóm 8, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số năm công tác ít hơn hoặc bằng 0.5 năm. Trong trường hợp điều kiện này đúng, có 3 khách hàng được phân vào nhóm (Nhóm 15) gồm có 1 khách hàng không trả được nợ đúng hạn và 2 khách hàng có thể trả được nợ đúng hạn. Hệ số gini là 0.444 và cao hơn hệ số gini ban đầu trước khi phân tách (0.391). Trong trường hợp điều kiện này sai, có 12 khách hàng được phân vào nhóm (Nhóm 16) gồm có 10 khách hàng không thể trả nợ và 2 khách hàng có thể trả nợ. Hệ số gini lúc này là 0.278 và bé hơn hệ số gini ban đầu trước khi phân tách (0.391). Dựa vào kết quả (gini) của lần phân tách thứ tám chúng ta có thể thấy rằng, thuật toán đã thêm 1 thuộc tính vào 3 thuộc tính sẵn có ở lần phân tách trước để có thể phân loại chính xác hơn khi cùng với một giả thiết là những khách hàng có thu nhập còn lại thấp, số nguồn thu nhập bé hơn 2 và số tiền vay lớn hơn khả năng trả nợ đúng hạn. Nếu số năm công tác ít hơn hoặc bằng 0.5 năm thì có khả năng trả nợ đúng hạn và ngược lại có số năm công tác nhiều hơn 0.5 năm thì sẽ không có khả năng trả được nợ đúng hạn với xác suất dự đoán là 2/3 (66.7%) đối với nhóm khách hàng có khả năng trả nợ đúng hạn và 10/12 (83.3%) đối với nhóm khách hàng không có khả năng trả nợ đúng hạn.

Để đơn giản hóa, đối với những lần phân tách sau của thuật toán, nhóm nghiên cứu sẽ chỉ đưa ra kết quả của các thông số kỹ thuật. Những ý nghĩa chuyên môn rút ra cũng như các khuyến nghị sẽ được nhóm tổng hợp ở phần sau.

Tiếp tục với Nhóm 11, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là thu nhập với điều kiện là thu nhập trung bình thấp (<= 1.5). Trong trường hợp điều kiện này đúng, có 7 khách hàng được phân vào nhóm (Nhóm 17), trong đó cả 7 khách hàng đều trả được nợ đúng hạn. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.124). Trong trường hợp điều kiện này sai, có 8 khách hàng được phân vào nhóm (Nhóm 18), trong đó có 7 khách hàng trả được nợ đúng hạn và 1 khách hàng

không trả được nợ đúng hạn. Hệ số gini lúc này là 0.219 và lớn hơn hệ số gini ban đầu trước khi phân tách (0.124).

Tiếp tục với Nhóm 16, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là trình độ với điều kiện là trình độ thấp (<= 0.5). Trong trường hợp điều kiện này đúng, có 6 khách hàng được phân vào nhóm (Nhóm 19), trong đó cả 6 khách hàng đều không thể trả nợ đúng hạn. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.278). Trong trường hợp điều kiện này sai, có 6 khách hàng được phân vào nhóm (Nhóm 20), trong đó có 4 khách hàng không thể trả nợ đúng hạn và 2 khách hàng có thể trả nợ đúng hạn, hệ số gini lúc này là 0.444, cao hơn hệ số gini ban đầu trước khi phân tách (0.278).

Tiếp tục với Nhóm 18, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số năm công tác với điều kiện là số năm công tác ít hơn hoặc bằng 5 năm. Trong trường hợp điều kiện này đúng, có 4 khách hàng được phân vào nhóm (Nhóm 21), trong đó có 1 khách hàng không thể trả nợ đúng hạn và 3 khách hàng có thể trả nợ đúng hạn, hệ số gini lúc này là 0.375, lớn hơn hệ số gini ban đầu trước khi phân tách (0.219). Trong trường hợp điều kiện này sai, có 4 khách hàng được phân vào nhóm (Nhóm 22), trong đó cả 4 khách hàng đều có thể trả nợ. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.219).

Tiếp tục với Nhóm 20, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là độ tuổi với điều kiện là độ tuổi dưới 30 (<= 0.5). Trong trường hợp điều kiện này đúng, có 5 khách hàng được phân vào nhóm (Nhóm 23) gồm có 4 khách hàng đều không thể trả nợ đúng hạn và 1 khách hàng có thể trả nợ đúng hạn. Hệ số gini lúc này là 0.32 và bé hơn hệ số gini ban đầu trước khi phân tách (0.444). Trong trường hợp điều kiện này sai, có 1 khách hàng được phân vào nhóm (Nhóm 24) và khách hàng đó có thể trả nợ. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.444).
Tiếp tục với Nhóm 21, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số năm công tác với điều kiện là số năm công tác ít hơn hoặc bằng 3.5 năm. Trong trường hợp điều kiện này đúng, có 3 khách hàng được phân vào nhóm (Nhóm 25), trong đó cả 3 khách hàng đều có thể trả nợ. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.375). Trong trường hợp điều kiện này sai, có 1 khách hàng được phân vào nhóm (Nhóm 26), và khách hàng đó không thể trả nợ đúng hạn. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.375).
Tiếp tục với Nhóm 23, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là thu nhập với điều kiện là thu nhập thấp (<= 0.5). Trong trường hợp điều kiện này đúng, có 4 khách hàng được phân vào nhóm (Nhóm 27), trong đó có 3 khách hàng không thể trả nợ đúng hạn và 1 khách hàng có thể trả nợ đúng hạn. Hệ số gini lúc này là 0.375 và lớn hơn hệ số gini ban đầu trước khi phân tách (0.32). Trong trường hợp điều kiện này sai, có 1 khách hàng được phân vào nhóm (Nhóm 28), và khách hàng đó không thể trả nợ đúng hạn, hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.32).
Tiếp tục với Nhóm 27, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là tài sản đảm bảo với điều kiện là tài sản đảm bảo không thuộc sở hữu người vay (<= 0.5). Trong trường hợp điều kiện này đúng, có 3 khách hàng được phân vào nhóm (Nhóm 29), trong đó có 2 khách hàng không thể trả nợ đúng hạn và 1 khách hàng có thể trả nợ đúng hạn.

Hệ số gini lúc này là 0.444 và lớn hơn hệ số gini ban đầu trước khi phân tách (0.375). Trong trường hợp điều kiện này sai, có 1 khách hàng được phân vào nhóm (Nhóm 30) và khách hàng đó không thể trả nợ đúng hạn. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.375).
Tiếp tục với Nhóm 29, thuật toán đã lựa chọn thuộc tính để phân tách nhóm là số năm công tác với điều kiện là số năm công tác ít hơn hoặc bằng 3 năm. Trong trường hợp điều kiện này đúng, có 2 khách hàng được phân vào nhóm (Nhóm 31), trong đó có 1 khách hàng không thể trả nợ và 1 khách hàng có thể trả nợ đúng hạn. Hệ số gini lúc này là 0.5 và lớn hơn hệ số gini ban đầu trước khi phân tách (0.44). Trong trường hợp điều kiện này sai, có 1 khách hàng được phân vào nhóm (Nhóm 32) và khách hàng đó không thể trả nợ. Hệ số gini lúc này là 0 và bé hơn hệ số gini ban đầu trước khi phân tách (0.44).
Thuật toán kết thúc phân tách ở đây. Ta có thể thấy ở tất cả những nút mà ở đó thuật toán không tiếp tục phân nhánh đều có hệ số gini bằng 0 (hay xác suất dự đoán cho nhóm khách hàng là 100%) ngoại trừ Nhóm 31 khi hệ số gini ở đó là tối đa (0.5) và thuật toán không thể xác định thêm thuộc tính nào để có thể phân loại được những khách hàng thuộc nhóm đó thì thuật toán cũng sẽ kết thúc việc phân nhánh ở nhóm đó.

4.5 Quy tắc cho vay

Khi ngân hàng xem xét liệu có nên cho một khách hàng vay hay không thì yếu tố đầu tiên cần xem xét là thu nhập còn lại để đưa ra quyết định.
Nếu khách hàng có thu nhập còn lại cao hơn 20.000.000, trình độ cao (sau đại học, cao đẳng) và số tiền vay bé hơn khả năng trả nợ thì ngân hàng có thể đồng ý cho vay. Điều này giúp cho các ngân hàng rút gọn được quy trình thẩm định và tiết kiệm được thời gian lẫn chi phí.
Tuy nhiên, nếu có thu nhập còn lại cao hơn 20.000.000 đồng, trình độ sau bậc đại học, cao đẳng nhưng số tiền cho vay lớn hơn khả năng trả nợ thì ngân hàng phải xem xét tiếp yếu tố thu nhập của khách hàng. Nếu khách hàng có thu nhập thấp/trung bình bé hơn 20.000.000/35.000.000 thì thuộc nhóm trả nợ đúng hạn và ngân hàng có thể xem xét để cho vay. Tương tự, nếu khách hàng có thu nhập cao hơn 35.000.0000 và số năm công tác bé hơn 3.5 năm hoặc lớn hơn 5 năm thì khách hàng đó cũng thuộc nhóm trả nợ đúng hạn. Do đó, ngân hàng có thể xem xét cho vay nếu một khách hàng đáp ứng đủ các điều kiện trên. Nếu khách hàng có thu nhập còn lại cao hơn 20.000.000, trình độ sau đại học, cao đẳng, số tiền cho vay lớn hơn khả năng trả nợ và thu nhập cao hơn 35.000.000 nhưng số năm công tác lớn hơn 3.5 năm và bé hơn 5 năm thì thuộc nhóm trả nợ không đúng hạn. Vì vậy, ngân hàng không nên cho vay để tránh rủi ro.
Ở trường hợp nếu một khách hàng có thu nhập còn lại thấp hơn 20.000.000 thì ngân hàng xem xét tiếp yếu tố số nguồn thu nhập. Nếu khách hàng đó có từ 2 nguồn thu nhập trở lên thì khách hàng thuộc nhóm có khả năng trả nợ đúng hạn. Do đó, ngân hàng có thể quyết định cho vay mà không cần thẩm định tiếp để rút ngắn quá trình thẩm định. Tuy nhiên, nếu khách hàng đó có ít nguồn thu nhập (bé hơn 2 nguồn thu nhập) thì ngân hàng phải xem xét tiếp yếu tố số tiền vay liệu có lớn hơn khả năng trả nợ hay không. Nếu khách hàng có số tiền vay lớn hơn khả năng trả nợ và số năm công tác bé hơn 5 năm thì thuộc nhóm trả nợ không đúng hạn và ngân hàng quyết định không cho vay để tránh rủi ro. Ngược lại, nếu khách hàng có số năm công tác lớn hơn 5 năm thì lại thuộc nhóm khách hàng trả nợ đúng hạn. Do đó, ngân hàng có thể đồng ý cho vay.

Nếu một khách hàng có nguồn thu nhập còn lại thấp hơn 20.000.000, số tiền vay bé hơn khả năng trả nợ và số năm công tác dưới 1 năm thì khách hàng đó thuộc nhóm trả nợ đúng hạn. Tuy nhiên, ngân hàng nên cân nhắc trường hợp này khi cho vay vì những trường hợp này này cũng có nguy cơ rủi ro. Ngược lại, nếu khách hàng có đầy đủ các điều kiện trên nhưng số năm công tác lớn hơn 1 năm và có trình độ thấp (dưới đại học, cao đẳng) thì khách hàng đó thuộc nhóm không trả nợ đúng hạn. Do đó, ngân hàng có thể quyết định ngay là không cho vay.
Nếu một khách hàng có nguồn thu nhập còn lại thấp hơn 20.000.000, số tiền vay bé hơn khả năng trả nợ, số năm công tác lớn hơn 1 năm, có trình độ sau đại học, cao đẳng và độ tuổi lớn hơn 30 thì xếp vào nhóm có khả năng trả nợ đúng hạn. Vì thế, ngân hàng nên cho vay đối với những khách hàng thỏa mãn những điều kiện này. Tuy nhiên, nếu khách hàng đáp ứng những điều kiện trên nhưng có độ tuổi bé hơn 30 tuổi thì đều rơi vào nhóm khách hàng không trả nợ đúng hạn hoặc chưa xác định được bất kể họ có thu nhập bao nhiêu, có tài sản đảm bảo hay không và số năm công tác là bao nhiêu. Do đó, để tránh rủi ro và rút ngắn quy trình thẩm định thì ngân hàng không nên hoặc xem xét cho vay để tránh rủi ro mặc dù có thể bỏ lỡ những khách hàng tiềm năng ở nhóm khách hàng chưa xác định được.

Vậy từ những kết quả trên, nhóm tác giả nhận thấy rằng những yếu tố quan trọng nhất để đưa ra quyết định cho vay là: thu nhập còn lại cao hơn 20.000.000, trình độ sau đại học cao đẳng và số tiền vay bé hơn khả năng trả nợ, số năm công tác hơn 5 năm, có hơn 2 nguồn thu nhập trở lên. Những yếu tố quyết định không cho vay và cân nhắc là: Trình độ thấp (trước đại học, cao đẳng), số tiền vay lớn hơn khả năng trả nợ, độ tuổi bé hơn 30 tuổi, có ít nguồn thu nhập (bé hơn 2).

***Thảo luận kết quả**

Thông qua bài nghiên cứu có thể thấy rằng khả năng trả nợ đúng hạn của khách hàng bị ảnh hưởng bởi rất nhiều thuộc tính khác nhau như: thu nhập còn lại, số năm công tác, trình độ, số nguồn thu nhập, số tiền vay, độ tuổi, thu nhập.
Cụ thể như sau: Thu nhập còn lại là thuộc tính quyết định và ảnh hưởng nhiều nhất đến khả năng trả nợ đúng hạn. Nếu thu nhập còn còn lại cao hơn 20.000.000 thì sẽ có khả năng trả nợ đúng hạn cao và ngược lại.
Số năm công tác cũng có mức độ ảnh hưởng đến khả năng trợ nợ đúng hạn của khách hàng vì nó thể hiện chuyên môn cũng như mức độ chuyên cần, ổn định của một cá nhân từ đó nó cũng tương quan dương với khả năng trả nợ đúng hạn. Yếu tố này tương tự như bài nghiên cứu của Viễn, N. T., Xuân, P. T. T., Huyền, L. T. T.(2020).
Số tiền vay khi lớn hơn khả năng trả nợ của khách hàng làm tăng nguy cơ trả nợ không đúng hạn. Hai bài nghiên cứu của Phong, T. T., Bình, N. T., Trang, L. X., & Phượng, Đ. T. (2020) và Viễn, N. T., Xuân, P. T. T., Huyền, L. T. T.(2020) cũng đã cho ra kết quả tương tự.
Bên cạnh đó, trình độ càng cao (sau đại học, cao đẳng) thì khả năng trả nợ đúng hạn càng cao tương tự như độ tuổi (lớn hơn 50 tuổi), thu nhập cao hơn 35.000.000, số nguồn thu nhập (lớn hơn 2) càng cao thì khả năng trả nợ đúng hạn càng cao. Kết quả cũng tương đồng với Viễn, N. T., Xuân, P. T. T., Huyền, L. T. T.(2020) và Mai, N. C. (2022) cũng đã cho thấy sự ảnh hưởng của trình độ học vấn đối với khả năng trả nợ.

Kết quả của nghiên cứu này tương đối khá cao dựa vào các thuộc tính trên. Cây quyết định sẽ giúp cho các cơ quan tín dụng rút ngắn được khả năng thẩm định của tổ chức và đồng thời giảm thiểu được chi phí thẩm định cũng như giúp thủ tục được hoàn thiện một cách nhanh chóng hơn.

5. Kết luận

5.1 Kết luận

Bài nghiên cứu này không thể áp dụng vào thực tế, vì dữ liệu được sử dụng trong bài được nhóm tác giả mô phỏng và cân bằng 2 nhóm giữa trả nợ đúng hạn và không đúng hạn ngay từ đầu. Trong khi đó, dữ liệu thực tế của quá trình thẩm định có xu hướng mất cân bằng và nghiêng về nhóm có khả năng trả nợ đúng hạn. Ngoài ra, để đưa ra được kết luận liệu một khách hàng có khả năng trả nợ được đúng hạn để ngân hàng có nên cho vay hay không thì cần rất nhiều thuộc tính quan trọng khác như giới tính, số lần quan hệ tín dụng,...

5.2 Điểm khác biệt so với tài liệu tham khảo

Trong những thuộc tính đã được nhóm tác giả thu thập, chỉ có một số thuộc tính thể hiện sự ảnh hưởng mạnh mẽ đến khả năng trả nợ đúng hạn của khách hàng. Những thuộc tính đó có thể kể đến như: Số năm công tác, trình độ, số nguồn thu nhập. Nhóm tác giả đã phát triển và xây dựng thêm một thuộc tính mới là thu nhập còn lại và có sự ảnh hưởng nhất đến khả năng trả nợ đúng hạn của khách hàng.

TÀI LIỆU THAM KHẢO

Cây Quyết Định (Decision Tree). (2019, June 6). Trí Tuệ Nhân Tạo. https://trituenhantao.io/kien-thuc/decision-tree/

What are Decision Trees, their types and why are they important? (n.d.). Www.linkedin.com. https://www.linkedin.com/pulse/what-decision-trees-types-why-important-srinivas-kilambi/

Müller, A. C., & Guido, S. (2017). Introduction to machine learning with Python : a guide for data scientists. O'reilly.

MÔ HÌNH LOGISTIC TRONG XẾP HẠNG RỦI RO TÍN DỤNG. (2013, June 8). MFEDE. https://tuanvanle.wordpress.com/2013/06/08/mo-hinh-logistic-trong-xep-hang-rui-ro-tin-dung/

Doan, N. (2018, October 28). Cây quyết định (Decision Tree) là gì? Tìm hiểu Thuật toán ID3. 1UP Note. https://1upnote.me/post/2018/10/ds-ml-decision-tree-id3/#:~:text=Iterative%20Dichotomiser%203%20%28ID3%29%20l%C3%A0%20thu%E1%BA%ADt%20to%C3%A1n%20n%E1%BB%95i

Ngân hàng Nhà nước Việt Nam (2013). Thông tư số 02/2013/TT-NHNN: Quy định về phân loại tài sản nợ, phương pháp trích lập dự phòng rủi ro và việc sử dụng dự phòng để xử lý rủi ro trong hoạt động của tổ chức tín dụng, chi nhánh ngân hàng nước ngoài.

Hệ số tương quan (Correlation Coefficient) là gì và ứng dụng. (n.d.). CareerLink.vn. https://www.careerlink.vn/en/careertools/economic-knowledge/he-so-tuong-quan-correlation-coefficient-la-gi-va-ung-dung

Thanh Phong, T., Thanh Bình, N., Xuân Trang, L. and Thị Phượng, Đ., (2021). Đánh giá khả năng trả nợ của khách hàng cá nhân – Nghiên cứu trường hợp Ngân hàng Nông nghiệp và Phát triển Nông thôn Việt Nam – Chi nhánh huyện Tân Hưng, tỉnh Long An. *Tạp chí Nghiên cứu Tài chính - Marketing*, (57).

Trường, N. H. (2017). Đánh giá khả năng trả nợ và phân loại khách hàng vay tín chấp tại Ngân hàng Thương mại Cổ phần Việt Nam Thịnh vượng.

Viễn, N. T., Xuân, P. T. T., Huyền, L. T. T.(2020). Nghiên cứu khả năng trả nợ của khách hàng cá nhân bằng mô hình cây quyết định. *Tạp chí Kinh tế và Ngân hàng Châu Á, (168).*