ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

BÁO CÁO CUỐI KỲ

Môn: MÔ HÌNH RỦI RO TÍN DỤNG TRONG R/PYTHON

Giảng viên: TS. Phạm Thị Thanh Xuân

Nhóm thực hiện: Nhóm 1

1. Bùi Nguyễn Thùy Như (Nhóm trưởng) K194141737
2. Lê Trung Chính K194141716
3. Phạm Văn Vĩnh Lộc K194141731
4. Nguyễn Thị Huệ Minh K194141733
5. Hồ Ngọc Quỳnh Như K194141738

Thành phố Hồ Chí Minh, ngày 31 tháng 5 năm 2022

MỤC LỤC

1. GIỚI THIỆU ĐỀ TÀI

Đánh giá rủi ro tín dụng là một vấn đề đầy thách thức đối với các ngân hàng thương mại và các nhà nghiên cứu kinh tế đã thúc đẩy nhiều nghiên cứu về lĩnh vực này trong vài thập kỷ qua. Quản lý tốt yếu tố rủi ro này có vai trò vô cùng quan trọng để tăng khả năng sinh lời của ngân hàng. Theo Ruziqa (2013), rủi ro tín dụng được đo lường bằng tỷ lệ nợ xấu. Tuy nhiên, việc các chủ nợ không có khả năng đánh giá chính xác rủi ro tín dụng tiềm ẩn của những người đi vay sẽ có những tác động tiêu cực không nhỏ đến hệ thống tài chính toàn cầu và hoạt động kinh tế nói chung. Một số nghiên cứu cho rằng tăng trưởng tín dụng quá mức,kém hiệu quả và quản lý rủi ro tín dụng không đầy đủ là những nguyên nhân chính dẫn đến tình trạng khủng hoảng tài chính toàn cầu gần đây. Với tiến bộ của khoa học công nghệ, các ngân hàng đã cố gắng giảm chi phí, phát triển các hệ thống và mô hình mạnh mẽ và tinh vi để dự đoán và quản lý rủi ro tín dụng. Mục tiêu của mô hình rủi ro tín dụng là đánh giá danh mục rủi ro của người đi vay và sau đó xác định xác suất vỡ nợ của người đó để từ đó ngân hàng có những cơ sở khách quan để đánh giá có nên cho đối tượng cần vay cho vay tín dụng hay không. Logistic Regression và Decision Tree là hai trong số những thuật toán dự đoán phân loại cho mô hình đang được sử dụng phổ biến nhất hiện nay, thông qua việc chia bộ dữ liệu và lựa chọn các biến cần thiết, nhóm nghiên cứu đưa ra những kết quả mô hình sau khi đã chạy thuật toán và đánh giá việc phân loại nhóm khách hàng theo dự đoán và trên thực tế.

2. TỔNG QUAN LÝ LUẬN

2.1. Lý thuyết Khả năng trả nợ

Tại Việt Nam, theo khoản 8 điều 3 chương I của Thông tư số 02/2013/TT-NHNN có quy định nợ xấu (NPL) 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 của khách hàng là việc khách hàng có khả năng trả nợ đầy đủ và đúng hạn với bên cho vay hay không (Ngân hàng Nhà Nước, 2013)". 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 của khách hàng (Nguyễn Đăng Dờn, 2016):

- 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.

Căn cứ theo định nghĩa của Quỹ tiền tệ quốc tế (IMF) thì: "Nợ xấu là khoản nợ khi quá hạn trả lãi và/hoặc gốc trên 90 ngày; hoặc các khoản lãi chưa trả từ 90 ngày trở lên đã được nhập gốc, tái cấp vốn hoặc đồng ý chậm theo thỏa thuận, hoặc các khoản phải thanh toán đã quá hạn 90 ngày nhưng có lý do để chắc chắn để nghi ngờ về khả năng khoản vay sẽ không được thanh toán đầy đủ" (Nguyễn Đăng Dờn, 2016).

Để thống nhất cách hiểu trong toàn bộ bài nghiên cứu, nhóm thống nhất việc đánh giá "khả năng trả nợ" của khách hàng theo Thông tư số 02/2013/TT-NHNN, đánh giá của Basel, IMF nghiên cứu này xác định khả năng trả nợ đúng/quá hạn của khách hàng và phân thành hai nhóm như sau:

- Nhóm 1: Khách hàng trả nợ đúng hạn – Trả nợ quá hạn dưới 10 ngày.

- Nhóm 2: Khách hàng trả nợ quá hạn từ 90 ngày trở lên.

2.2. Lý thuyết Decision Tree

Decision Tree là một thuật toán thuộc nhóm Supervised Learning được sử dụng cho cả classification và regression. Đây là thuật toán theo mô hình cây để xác định kết quả của hành động. Mỗi nhánh cây đại diện cho một quyết định, sự xuất hiện hay phản ứng có thể xảy ra. Decision Tree thường được sử dụng để dự đoán và phân loại trong nhiều lĩnh vực khác nhau như: Business Management, Fault Diagnosis và Customer Relationship Management. Tựu trung, mô hình này thường được sử dụng trong nghiên cứu hoạt động, đặc biệt trong phân tích quyết định, giúp xác định chiến lược có khả năng đạt được mục tiêu cao nhất. Về ý tưởng, Decision Tree sẽ tìm các tính năng (feature) mô tả có chứa "thông tin" nhất về target feature và sau đó chia tập dữ liệu dọc theo các giá trị target feature cho các tập con (sub dataset) càng "thuần" càng tốt, từ đó có thể thấy các tính năng mô tả dẫn đến target feature "thuần khiết" nhất được xem là các tính năng có giá trị "thông tin" nhất. Quá trình tìm kiếm các tính năng có giá trị "thông tin" nhất được thực hiện cho đến khi chúng ta hoàn thành điều kiện dừng khi cuối cùng kết thúc ở nút lá (leaf node). Các nút lá chứa các thông tin dự đoán mà chúng ta sẽ thực hiện cho các thực thể mới được trình bày cho mô hình đã được huấn luyện.

Điều này là khả thi vì mô hình đã học được cấu trúc cơ bản của dữ liệu huấn luyện (training data), và do đó có thể đưa ra một số giả định, đưa ra các dự đoán về giá trị target feature (class) của các thực thể chưa biết. Điều này được thực hiện thông qua một quá trình xây dựng Decision Tree được khái quát như sau:

- Bắt đầu với tất cả các sample tại một node.

- Các sample phân vùng dựa trên input để tạo tập con thuần khiết nhất.

- Quá trình phân vùng dữ liệu được lặp lại để cho vào các tập con "thuần khiết" hơn.

2.3. Lược khảo nghiên cứu

Kohansal & Mansoori (2009) đã tiến hành một nghiên cứu thực nghiệm tại Khorasan-Razavi, Iran trên dữ liệu khách hàng cá nhân và phân tích bằng mô hình logit. Kết quả nghiên cứu cho thấy, có bảy thuộc tính tác động đến khả năng trả nợ gồm: (1) Kinh nghiệm làm việc; (2) Thu nhập; (3) Số tiền vay; (4) Giá trị tài sản đảm bảo; (5) Lãi suất vay; (6) Tổng chi phí ứng dụng – Chi phí cho việc sử dụng máy móc,

thiết bị nông nghiệp; và (7) Số lần trả nợ. Trong số đó, lãi suất vay là thuộc tính tác động quan trọng nhất, tiếp đến là kinh nghiệm làm việc và tổng chi phí ứng dụng.

Gupta & Goyal (2019) đã nghiên cứu mạng nơron nhân tạo và các mô hình hồi quy tuyến tính để dự đoán tình trạng vỡ nợ tín dụng. Cả hệ thống đã được đào tạo về dữ liệu cho vay do kaggle.com cung cấp. Kết quả của cả hai hệ thống cho thấy tác động ngang nhau trên tập dữ liệu và do đó rất hiệu quả với độ chính xác là 97,67575% của mạng nơ-ron nhân tạo và 97,69609%. Hệ thống phân loại biến đầu ra một cách chính xác với sai số rất thấp. Vì vậy, cả hai quy trình này có thể được sử dụng để xác định tình trạng vỡ nợ tín dụng với độ chính xác như nhau. Ngoài ra, mạng nơron biểu diễn một phương pháp hộp đen nên rất khó giải thích kết quả so với mô hình hồi quy tuyến tính. Do đó, việc sử dụng mô hình nào phụ thuộc vào ứng dụng mà người ta phải sử dụng. Hơn nữa, trong khi điều chỉnh một mô hình sử dụng quy trình mạng thần kinh, người dùng cần quan tâm nhiều hơn đến các thuộc tính và chuẩn hóa dữ liệu để cải thiện hiệu suất. Để kết luận, mạng nơ-ron cung cấp bằng chứng mạnh mẽ để dự đoán hiệu quả tình trạng vỡ nợ tín dụng đối với một đơn xin vay.

Sihem Khemakhem & Younes Boujelbene (2017) đã so sánh độ chính xác dự đoán của cả hai kỹ thuật ANN và Decision Tree trên một nhóm các công ty Tunisia, có tính đến các yếu tố dự đoán thông thường, chẳng hạn như tỷ số tài chính, các biến số tài chính khác, chẳng hạn như thời gian nghiên cứu của báo cáo tín dụng, đảm bảo, quy mô của số công ty và khoản vay, và các biến số phi tài chính như cơ cấu sở hữu, thời hạn quan hệ ngân hàng doanh nghiệp và hình thức pháp lý. Công việc này kết hợp các kỹ thuật trí tuệ nhân tạo. Các kỹ thuật này đã được áp dụng trong lĩnh vực dự đoán mức độ tín nhiệm của các công ty như được minh họa bằng việc sử dụng ANN và Decision Tree. Kết quả đầu tiên cho thấy tỷ suất sinh lời, khả năng trả nợ và khả năng thanh toán là một trong những tỷ số có ý nghĩa dự báo khả năng mất khả năng thanh toán của công ty. Kết quả cũng cho thấy tầm quan trọng của thời gian nghiên cứu của một báo cáo tín dụng, các khoản đảm bảo, quy mô của công ty RAF 17,3 334 và số khoản vay trong đánh giá tín dụng. Ngoài ra, nhóm nghiên cứu đã tìm thấy mối quan hệ đáng kể giữa cấu trúc sở hữu, thời hạn quan hệ ngân hàng doanh nghiệp và rủi ro không thanh toán các khoản vay do ngân hàng cấp trong nghiên cứu. Thứ hai, để đánh giá các tiêu chí hoạt động, nhóm đã sử dụng tỷ lệ chính xác và các chi phí phân loại sai khác nhau. Vấn đề với tham số này là hành vi sai lệch của tỷ lệ chính xác tốt với dữ liệu không cân bằng có thể dẫn đến việc lựa chọn mô hình dự đoán không tốt. Nghiên cứu cũng đã bao gồm các lỗi loại I và loại II, cho phép nhóm đánh giá hiệu suất của từng lớp. Dường như để chọn thước đo đánh giá hiệu suất phù hợp nhất cần tính đến tính đặc thù của dữ liệu mà mô hình dự đoán sẽ được áp dụng. Hơn nữa, kết quả thu được cho thấy Decision Tree hiệu quả hơn ANN về dự đoán rủi ro tín dụng sử dụng dữ liệu cân bằng.

Tại Việt Nam, khả năng trả nợ của người vay gần đây cũng đã được quan tâm nghiên cứu. Trương Đông Lộc & Nguyễn Thanh Bình (2011) đã nghiên cứu về các

thuộc tính ảnh hưởng đến khả năng trả nợ đúng hạn của nông hộ ở tỉnh Hậu Giang. Nghiên cứu sử dụng mô hình hồi quy Probit kiểm định trên tập dữ liệu gồm 436 mẫu. Kết quả của nghiên cứu chỉ ra những thuộc tính tác động đến khả năng trả nợ đúng hạn của các nông hộ là: (1) Thu nhập; (2) Trình độ học vấn; (3) Số thành viên gia đình có thu nhập; và (4) Lãi suất vay. Nghiên cứu cũng chỉ ra rằng những khoản vay được sử dụng đúng mục đích cũng sẽ cho xác suất trả nợ đúng hạn cao hơn. Ngoài ra, kết quả phân tích định lượng cũng cho thấy khả năng trả nợ của những hộ đi vay vốn phục vụ cho sản xuất nông nghiệp cao hơn những hộ đi vay vốn sử dụng cho mục đích phi nông nghiệp. Trần Thế Sao (2017) trong nghiên cứu của mình, các thuộc tính ảnh hưởng khả năng trả nợ ngân hàng của các nông hộ trên địa bàn huyện Bến Lức tỉnh Long An. Ông sử dụng mô hình hồi quy Binary Logistic kiểm định trên tập dữ liệu gồm 250 mẫu nhằm xác định mức độ tác động của các thuộc tính. Kết quả nghiên cứu cho thấy các thuộc tính: (1) Trình độ học vấn; (2) Diện tích đất canh tác; (3) Thu nhập phi nông nghiệp; và (4) Thời hạn trả nợ có tác động tích cực đến khả năng trả nợ đúng hạn của nông hộ. Các thuộc tính (5) Số người phụ thuộc; và (6) Số tiền vay có tác động tiêu cực. Những thuộc tính còn lại là tuổi tác, tình trạng hôn nhân, kinh nghiệm làm việc và số lần đến thăm của cán bộ tín dụng sau khi giải ngân đều không có ý nghĩa thống kê. Tóm lại, mặc dù mỗi nghiên cứu được thực hiện trên phạm vi, đối tượng và mô hình khác nhau nhưng tựu chung lại, có thể thấy: thu nhập, kinh nghiệm làm việc, trình độ học vấn, lãi suất vay, tuổi tác, số tiền vay là những thuộc tính trọng tâm trong việc đánh giá khả năng trả nợ đúng hạn/quá hạn của khách hàng cá nhân ở các nghiên cứu trước đó. Đây chính là cơ sở khoa học được kế thừa trong việc xây dựng mô hình nghiên cứu cho trường hợp khách hàng cá nhân.

Võ Văn Tài & Nguyễn Thị Hồng Dân & Nghiêm Quang Thường (2017) dựa trên các số liệu thực tế thu được và lý thuyết đã trình bày, thực hiện việc đánh giá khả năng trả nợ vay của khách hàng trên địa bàn thành phố Cần Thơ. Đối tượng khách hàng được khảo sát là các doanh nghiệp hoạt động trên các lĩnh vực quan trọng: nông nghiệp, công nghiệp và thương mại. Gồm 214 doanh nghiệp, trong đó 143 doanh nghiệp trả nợ được đúng hạn (TN) và 71 không trả nợ được đúng hạn (KTN). Số liệu nghiên cứu được cung cấp bởi cơ quan có trách nhiệm quản lý trên địa bàn thành phố Cần Thơ năm 2013, trong một đề tài nghiên cứu về doanh nghiệp trên địa bàn. Bài nghiên cứu đã trình bày các phương pháp phân loại và vấn đề tính toán của chúng, trong đó đã đề nghị thuật toán xác định xác suất tiên nghiệm trong phân loại bằng phương pháp Bayes. Thuật toán này đã chứng minh ưu điểm, khi làm giảm được xác suất sai lầm phân loại trong tất cả các trường hợp với bộ số liệu thực tế được khảo sát. Bài báo đã xem xét vấn đề tính toán trong áp dụng thực tế của các phương pháp, trong đó đã thiết lập các chương trình để giải quyết vấn đề tính toán của phương pháp Bayes với thuật toán tìm xác suất tiên nghiệm đề nghị.

3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Dữ liệu

3.1.1. Lý do chọn biến

Về lý do chọn biến, dựa trên các gợi ý của giảng viên và các lược khảo nghiên cứu, nhóm nhận thấy có những biến độc lập quan trọng cần phải có đối với bài thực hành đó là thu nhập, tài sản đảm bảo và số nguồn thu nhập. Ngoài ra các biến tự chọn liên quan đến cá nhân đối tượng cho vay như hôn nhân, tuổi được nhóm thử nghiệm và thay đổi nhiều lần để chọn ra một mô hình có kết quả tối ưu nhất.

Bảng 3.1: Mô tả biến sử dụng trong nghiên cứu

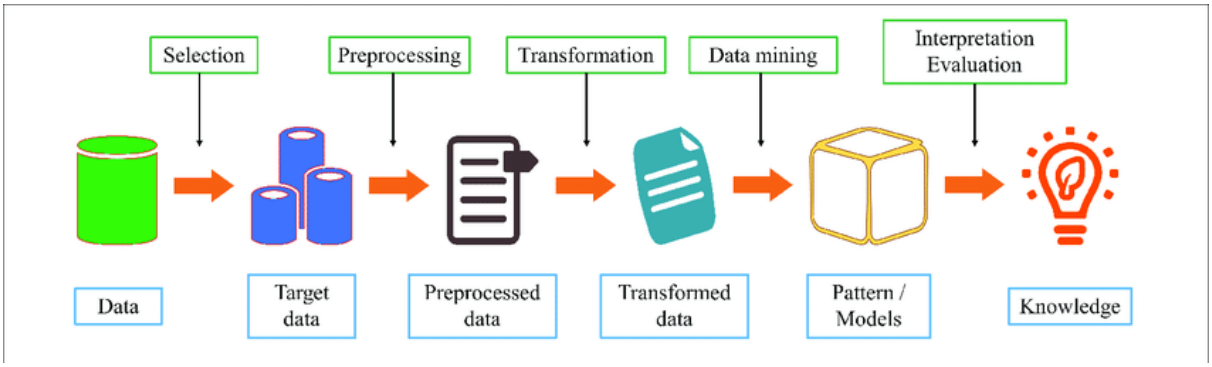
Tên biến	**Thang đo**	**Thể hiện trên bảng dữ liệu**
Biến phụ thuộc		
Khả năng trả nợ vay khách hàng cá nhân	1: Trả nợ đúng hạn	Khả năng trả nợ
	2: Trả nợ không đúng hạn	
Biến độc lập		
Thu nhập	1: Thấp (≤ 39.425.000)	Thu nhập trung bình tháng của một khách hàng
	2: Trung bình (39.425.000 < X ≤ 61.500.000)	
	3: Cao (61.500.000 < X ≤ 98.479.500)	
	4: Rất cao (X ≥ 98.479.500)	
Tài sản đảm bảo	0: Không thuộc sở hữu người vay (bảo lãnh)	Tài sản đảm bảo có hay không thuộc sở hữu người đi vay
	1: Thuộc sở hữu người vay	
Số nguồn thu nhập	1: Ít nguồn thu nhập (X < 2)	Số nguồn thu nhập của khách hàng
	2: Nhiều nguồn thu nhập (X ≥ 2)	

Số năm công tác	1: X ≤ 2 (năm)	Số năm công tác/làm việc của khách hàng
	2: 2 < X ≤ 5 (năm)	
	3: X > 5 (năm)	
Số lần quan hệ tín dụng	1: QHTD lần đầu (X = 1)	Số lần quan hệ tín dụng (đã có đi vay từ trước hay không)
	2: QHTD nhiều lần (X > 1)	
Hôn nhân	1: Độc thân	Tình trạng hôn nhân hiện tại của khách hàng
	2: Đã kết hôn	
	3: Ly dị	
Tuổi	1. Thanh niên (X <= 35 tuổi)	Số tuổi hiện tại của khách hàng
	2. Trung niên (35 < X <= 65 tuổi)	
	3. Cao tuổi (X > 65 tuổi)	

3.1.2. Xử lý dữ liệu

Sau khi thu thập dữ liệu, nhóm nhận thấy bộ dữ liệu thu thập được có nhiều lỗi nghiêm trọng như người nhập nhập sai đơn vị đo, nhập thiếu, trùng lắp,.. Chính vì thế để có một bộ dữ liệu phù hợp để đưa vào mô hình, nhóm đã tiến hành nhiều bước làm để chỉnh sửa và xóa bỏ những phần dữ liệu lỗi. Có thể kể đến như xóa bỏ những dữ liệu trống, trùng lặp, chỉnh sửa giá trị bị nhập sai,.. Việc thực hiện những bước làm này tuy sẽ làm mất một phần lớn dữ liệu ban đầu nhưng đây là điều cần phải làm vì nếu dữ liệu có nhiều lỗi khi đưa vào mô hình sẽ hoạt động không hiệu quả và không có ý nghĩa.

3.2. Mô hình nghiên cứu



Hình 3.1: Quy trình thực hiện nghiên cứu

Nguồn: Harri Niska, 2012

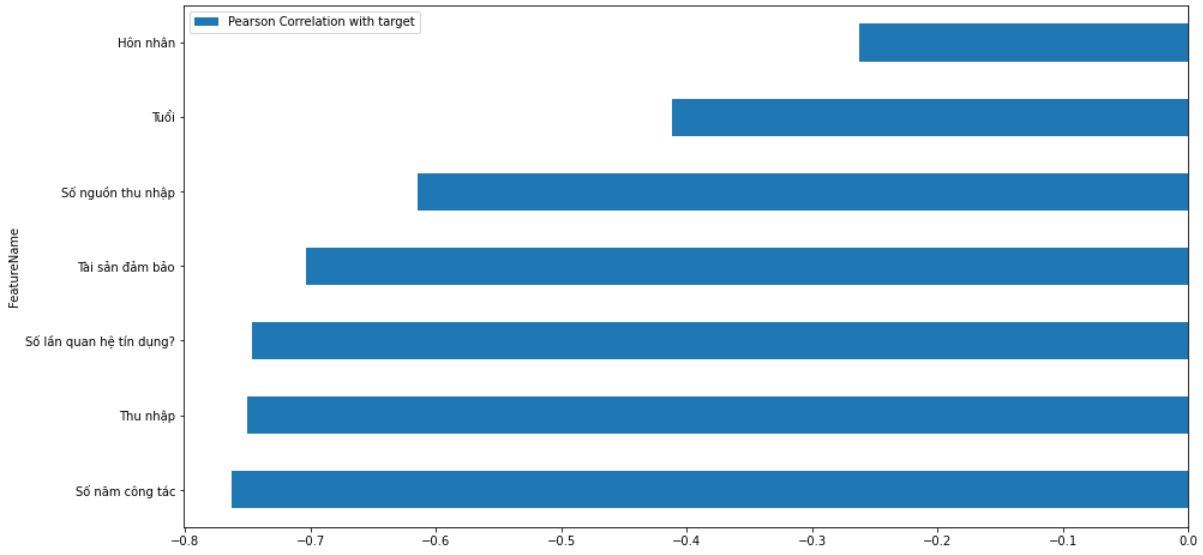
- Bước 1: Tiến hành import dữ liệu gốc, chọn lọc trong data gốc để lựa ra những biến mục tiêu, loại bỏ những biến không cần thiết.

- Bước 2: Sau khi chọn lọc, tiến hành xử lý bộ data mới, thống kê mô tả để xem xét mức độ bất thường và tính ổn định của dữ liệu.

- Bước 3: Sau tiền xử lý data, tiến hành biến đổi bộ dữ liệu sao cho phù hợp với mục đích sử dụng.

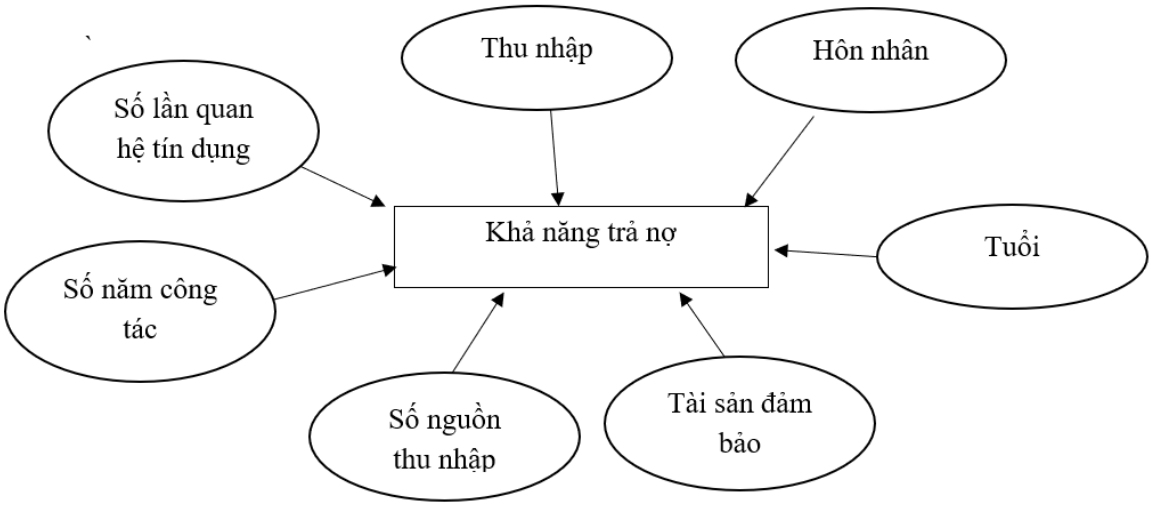
- Bước 4: Xây dựng các mô hình, thiết lập các bảng giá trị (Classification Report), sử dụng biểu đồ để trực quan hóa, nhận xét, đánh giá và so sánh giữa các mô hình.

- Bước 5: Kết luận, đánh giá sự ổn định và phù hợp giữa các mô hình, xem xét ưu nhược và đưa ra kết luận.



Hình 3.2: Tương quan Pearson giữa các biến độc lập và biến phụ thuộc

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab



Hình 3.3: Mô hình nghiên cứu

Nguồn: Tác giả

4. KẾT QUẢ NGHIÊN CỨU

4.1. Thống kê mô tả

	Khả năng trả nợ	Thu nhập	Tài sản đảm bảo	Số nguồn thu nhập	Số năm công tác	Số lần quan hệ tín dụng?	Hôn nhân	Tuổi
count	150	150	150	150	150	150	150	150
unique	2	4	2	2	3	2	3	3
top	2: Trả nợ không đúng hạn	1: Thấp (X<= 39.425.000)	1: Thuộc sở hữu người vay	2: Nhiều nguồn thu nhập (X>=2)	1: X<=2 năm	2: QHTD nhiều lần (X > 1)	2. Đã kết hôn	2. Trung niên (35<X <=65 tuổi)
freq	75	46	88	91	55	76	75	76

Hình 4.1: Thống kê mô tả của dữ liệu

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

	Khả năng trả nợ	Thu nhập	Tài sản đảm bảo	Số nguồn thu nhập	Số năm công tác	Số lần quan hệ tín dụng?	Hôn nhân	Tuổi
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000
mean	0.500000	2.366667	0.586667	1.606667	2.000000	1.506667	1.833333	1.826667
std	0.501675	1.131707	0.494081	0.490126	0.859218	0.501630	0.689486	0.682934
min	0.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	0.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
50%	0.500000	2.000000	1.000000	2.000000	2.000000	2.000000	2.000000	2.000000
75%	1.000000	3.000000	1.000000	2.000000	3.000000	2.000000	2.000000	2.000000
max	1.000000	4.000000	1.000000	2.000000	3.000000	2.000000	3.000000	3.000000

Hình 4.2: Thống kê mô tả chi tiết dữ liệu

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Hình 4.1 và 4.2 cho biết các trường dữ liệu đều có 150 biến quan sát, đại diện cho 150 trường hợp khách hàng đi vay. Cụ thể:

- "Khả năng trả nợ" có 2 nhóm, trong đó, cả hai đều có số lượng quan sát bằng nhau là 75/150, do đó giá trị trung bình là 0.5.

- "Thu nhập" có 4 nhóm, trong đó nhóm khách hàng có thu nhập thấp (dưới 39,425,000 đồng) chiếm tỷ trọng lớn nhất với 46/150 quan sát.

- “Tài sản đảm bảo” có 2 nhóm, trong đó nhóm khách hàng có tài sản thuộc sở hữu chiếm tỷ trọng lớn nhất với 88/150 quan sát.

- “Số nguồn thu nhập” có 2 nhóm, trong đó nhóm khách hàng có từ 2 nguồn thu nhập trở lên chiếm tỷ trọng lớn nhất với 91/150 quan sát.

- “Số năm công tác” có 3 nhóm, trong đó nhóm khách hàng có số năm công tác từ 2 năm trở xuống chiếm đa số với 55/150 quan sát.

- “Số lần quan hệ tín dụng” có 2 nhóm, trong đó nhóm khách hàng có số lần quan hệ tín dụng lớn hơn 1 chiếm đa số với 76/150 quan sát.

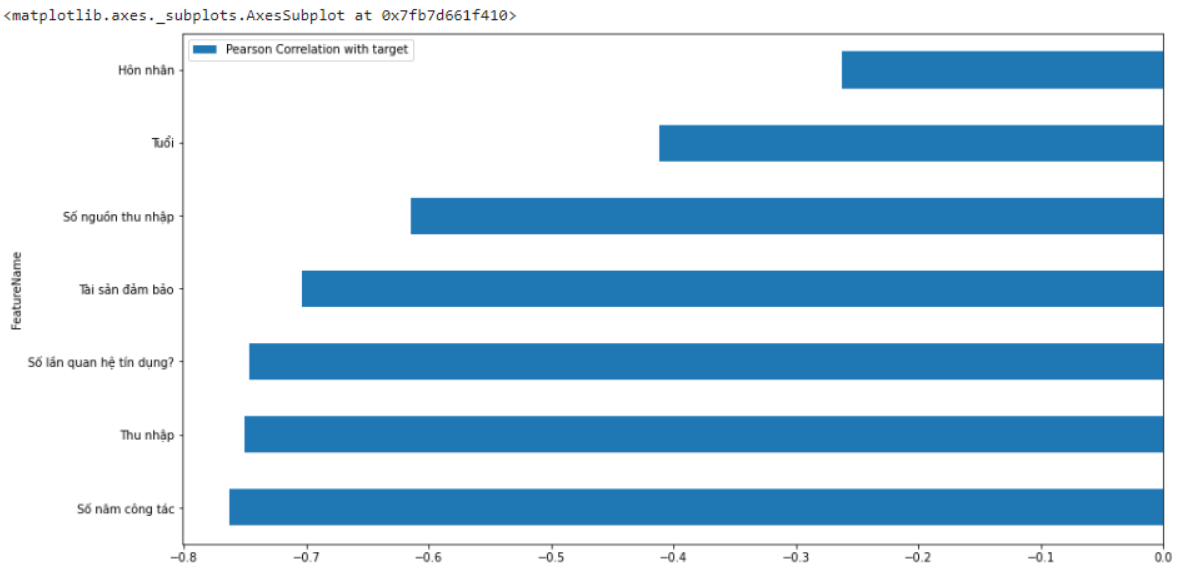
- “Hôn nhân” có 3 nhóm, trong đó nhóm khách hàng đã kết hôn chiếm đa số với 75/150 quan sát.

- “Tuổi” có 3 nhóm, trong đó nhóm khách hàng từ 35 đến 65 tuổi chiếm đa số với 76/150 quan sát.

	FeatureName	Pearson Correlation with target
3	Số năm công tác	-0.762929
0	Thu nhập	-0.750639
4	Số lần quan hệ tín dụng?	-0.746733
1	Tài sản đảm bảo	-0.703989
2	Số nguồn thu nhập	-0.614138
6	Tuổi	-0.411370
5	Hôn nhân	-0.261938

Hình 4.3: Tương quan giữa các biến với “Khả năng trả nợ” sắp xếp theo thứ tự giảm dần

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab



Hình 4.4: Biểu đồ thể hiện tương quan giữa các biến với biến "Khả năng trả nợ"

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Hình 4.3 và 4.4 cho thấy các biến độc lập đều có tác động nhất định đối với biến phụ thuộc, biến có tác động mạnh nhất là "Số năm công tác" và biến có tác động yếu nhất là "Hôn nhân", trong đó, tất cả các biến độc lập đều tương quan âm với "Khả năng trả nợ".

4.2. Kết quả của Logistic Regression

4.2.1. Confusion matrix

Quy ước: Trả nợ đúng hạn là nhóm 0, Trả nợ không đúng hạn là nhóm 1.

```
[[10  1]
 [ 0 12]]
```

Hình 4.5: Confusion matrix của Logistic regression

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

- True Positive = 10: Mô hình dự đoán đúng 10 người trả nợ đúng hạn.

- False Positive = 0: Không có trường hợp trả nợ không đúng hạn bị mô hình dự báo sai.

- False negative = 1: Có 1 trường hợp trả nợ đúng hạn nhưng bị mô hình phát hiện sai.

- True negative = 12: Có 12 trường hợp trả nợ không đúng hạn được mô hình phát hiện đúng.

Trong trường hợp này, mô hình đã cho ra kết quả dự báo khá tốt khi số lượng False Positive và False Negative đều thấp. Mô hình có ưu điểm hơn khi không bỏ nhầm một trường hợp trả nợ không đúng hạn nào, trong thực tế thì việc bỏ qua trường hợp trả nợ không đúng hạn gây ra hậu quả nghiêm trọng hơn, nên yêu cầu ngân hàng cần phải cẩn trọng và lưu ý nhiều. Mô hình đã dự báo một False Negative, tuy nhiên điều này không quá nghiêm trọng và ảnh hưởng, mô hình chỉ đang đưa ra báo động giả.

4.2.2. Classification report

```
[[10  1]
 [ 0 12]]
              precision    recall  f1-score   support

           0       1.00      0.91      0.95        11
           1       0.92      1.00      0.96        12

    accuracy                           0.96        23
   macro avg       0.96      0.95      0.96        23
weighted avg       0.96      0.96      0.96        23

Logistic Regression accuracy:  0.9565217391304348
```

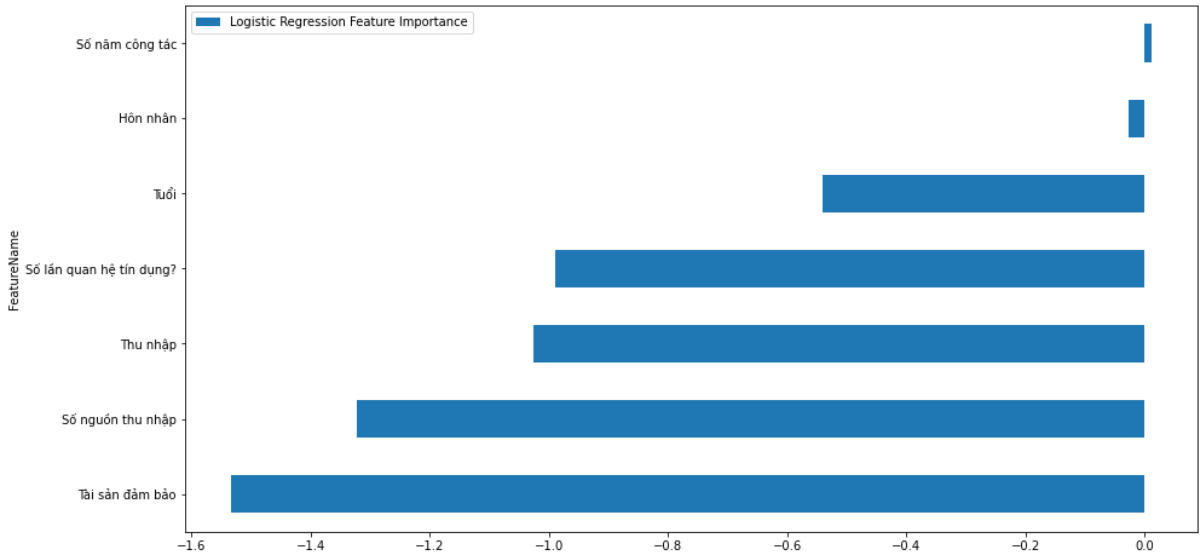
Hình 4.6: Classification report của Logistic regression

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Support[0] = 11, support[1] = 12: dữ liệu của mô hình bao gồm 11 trường hợp trả nợ đúng hạn và 12 trường hợp trả nợ không đúng hạn với tổng quan sát là 23. Bộ dữ liệu được cân bằng. Precision[0] = 1: Có 100% trường hợp trả nợ đúng hạn được mô hình phân loại đúng. Precision[1] = 0.92: Trong số các trường hợp được phân loại là trả nợ không đúng hạn, có 92% trường hợp thực sự là trả nợ không đúng hạn. Recall[0] = 0.91: Mô hình chỉ phát hiện được 91% trường hợp trả nợ đúng hạn trong số các trường hợp thực sự đúng hạn. Recall[1] = 1: Mô hình phát hiện được tất cả các trường hợp trả nợ không đúng hạn F1-score[0] = 0.95: Cả precision và recall đều cao khiến cho F1-score (trung bình điều hòa lên tới 95%. Điều này cho thấy mô hình ở lớp trả nợ đúng hạn có hiệu suất tốt. F1-score[1] = 0.96: Precision và recall ở lớp 1 đều cao giúp chỉ số trung bình điều hòa lên tới 96%. Chỉ số này cho thấy mô hình dự báo trả nợ không đúng hạn có hiệu suất tốt. Vì bộ dữ liệu không bị mất cân bằng nên các chỉ số macro average và weighted average đều giống nhau. Macro/weighted average của precision = 0.96, trung bình mô hình dự báo đúng 96% trường hợp. Với macro average của recall = 0.95, mô hình phát hiện đúng được 95% trường hợp. Weighted average của recall =0.96, mô hình phát hiện đúng được 96% trường hợp. F1-score ở macro average và weighted average đều là 96% bằng với accuracy. Mô hình logistic regression có accuracy là 95.65%, cho thấy mô hình dự đoán đúng 95.65%.

Việc mô hình dự báo đúng 100% trường hợp trả nợ đúng hạn và dự báo đúng 100% trường hợp trả nợ không đúng hạn trên toàn bộ tập dêdữ liệu. Điều này cho thấy, mô hình không bỏ sót bất kỳ trường hợp trả nợ không đúng hạn nào. Việc dự báo về các trường hợp đúng hạn cũng đúng hoàn toàn. Vì vậy có thể tin tưởng vào mô hình dự báo để thực hiện phê duyệt tín dụng.

4.2.3. Feature importance



Hình 4.7: Feature importance của Logistic regression

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Thực hiện đánh giá mức độ quan trọng của từng thuộc tính trong mô hình, các thuộc tính được sắp xếp theo thứ tự tăng dần như sau: Số năm công tác, Hôn nhân, Tuổi, Số lần quan hệ tín dụng, Thu nhập, Số nguồn thu nhập, Tài sản đảm bảo. Kết quả đánh giá cho thấy tài sản đảm bảo là yếu tố quyết định quan trọng bậc nhất trong mô hình. Ngoài yếu tố số năm công tác, các yếu tố khác đều có ảnh hưởng âm đối với khả năng trả nợ.

4.3. Kết quả của Decision Tree

4.3.1. Confusion matrix

Quy ước: Trả nợ đúng hạn là nhóm 0, Trả nợ không đúng hạn là nhóm 1.

```
[[10  1]
 [ 1 11]]
```

Hình 4.8: Confusion matrix của Decision Tree

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Kết quả của confusion matrix cho thấy:

- True Positive = 10: Mô hình dự đoán đúng 10 người trả nợ đúng hạn.

- False Positive = 1: Có 01 trường hợp trả nợ không đúng hạn bị mô hình dự báo sai.

- False negative = 1: Có 01 trường hợp trả nợ đúng hạn nhưng bị mô hình phát hiện sai.

- True negative = 11: Có 11 trường hợp trả nợ không đúng hạn được mô hình phát hiện đúng.

Kết quả dự báo của mô hình decision tree khá tốt khi số lượng False Positive và False Negative đều thấp. Trong thực tế thì việc bỏ qua trường hợp trả nợ không đúng hạn gây ra hậu quả nghiêm trọng hơn, ngân hàng cần phải cẩn trọng và lưu ý nhiều. Mô hình dự báo sai 01 trường hợp trả nợ không đúng hạn thành trả nợ đúng hạn vì vậy mô hình cần phải khắc phục để giảm việc dự báo sai ảnh hưởng đến ngân hàng. Mô hình đã dự báo một False Negative, tuy nhiên điều này không quá nghiêm trọng và ảnh hưởng, mô hình chỉ đang đưa ra báo động giả.

4.3.2. Classification report

```
[[10  1]
 [ 1 11]]
              precision    recall  f1-score   support

           0       0.91      0.91      0.91        11
           1       0.92      0.92      0.92        12

    accuracy                           0.91        23
   macro avg       0.91      0.91      0.91        23
weighted avg       0.91      0.91      0.91        23

Decision Tree accuracy:  0.9130434782608695
```

Hình 4.9: Classification report của Decision Tree

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

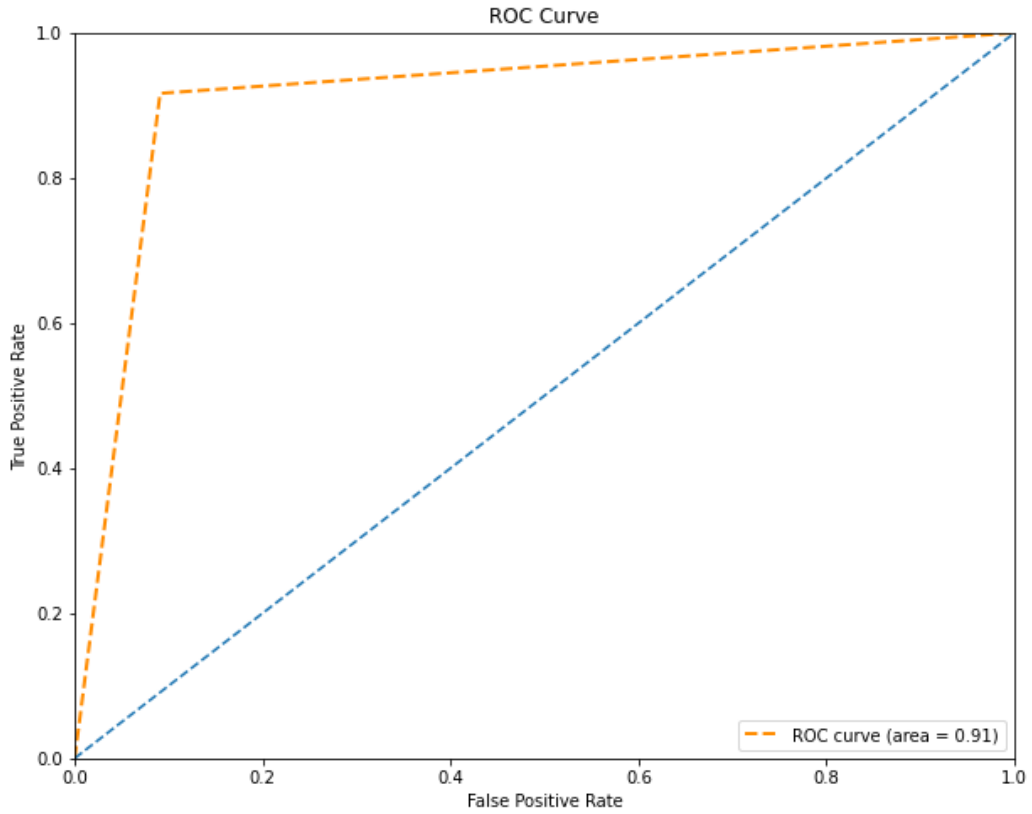
Các chỉ số của classification report ở mô hình decision tree đều thấp hơn ở mô hình logistic regression.

Precision[0] = 0.91: Có 91% trường hợp trả nợ đúng hạn được mô hình phân loại đúng. Precision[1] = 0.92: Trong số các trường hợp được phân loại là trả nợ không đúng hạn, có 92% trường hợp thực sự là trả nợ không đúng hạn. Recall[0] = 0.91: Mô hình chỉ phát hiện được 91% trường hợp trả nợ đúng hạn trong số các trường hợp thực sự đúng hạn. Recall[1] = 0.92: Mô hình dự báo được 92% trường hợp trên tổng số các trường hợp trả nợ không đúng hạn. F1-score[0]= 0.95: Cả precision và recall đều cao khiến cho F1-score (trung bình điều hòa lên tới 95%. Điều này cho thấy mô hình ở lớp trả nợ đúng hạn có hiệu suất tốt. F1-score[1] = 0.96: Precision và recall ở lớp 1 đều cao giúp chỉ số trung bình điều hòa lên tới 96%. Chỉ số này cho thấy mô hình dự báo trả nợ không đúng hạn có hiệu suất tốt. Vì bộ dữ liệu không bị mất cân bằng nên các chỉ số Macro Average và Weighted Average đều giống nhau. Macro/weighted average của Precision = 0.96, trung bình mô hình dự báo đúng 96% trường hợp. Với Macro Average của Recall = 0.95, mô hình phát hiện đúng được 95% trường hợp. Weighted Average của Recall = 0.96, mô hình phát hiện đúng được 96% trường hợp. F1-score ở macro average và weighted average đều là 96% bằng với accuracy.

Với tính chất của mô hình cần nghiên cứu, chỉ số recall của nhóm có rủi ro (nhóm 1) là điều quan trọng hơn. Trong quy trình phê duyệt tín dụng, việc đánh giá

một khách hàng không có rủi ro trở thành có rủi ro ít quan trọng hơn. Ngược lại, nếu đánh giá một khách hàng có rủi ro thành không có rủi ro và chấp thuận khoản vay, khoản vay này có thể trở thành khoản nợ trả chậm hoặc nợ xấu. Điều này gây ảnh hưởng lớn đến tài chính của ngân hàng. Vì vậy việc không bỏ sót các trường hợp có rủi ro trở nên quan trọng hơn. Mô hình chỉ dự báo đúng 92% trường hợp có rủi ro trên toàn bộ tập dữ liệu, chỉ số này vẫn cần gia tăng để giúp giảm thiểu rủi ro trong việc phê duyệt tín dụng của ngân hàng.

4.3.3. Kết quả của đường ROC

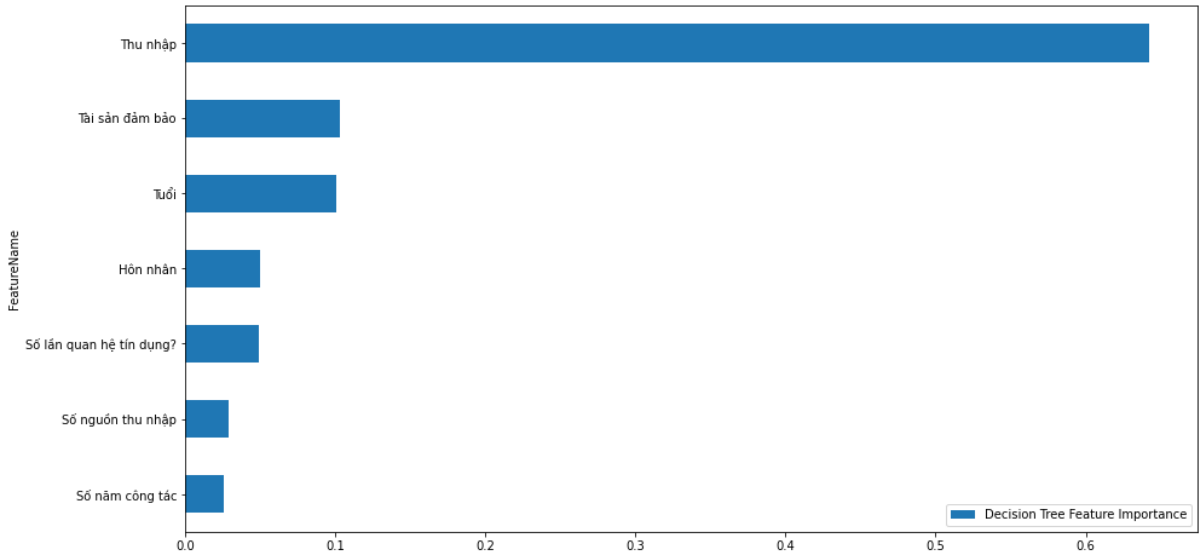


Hình 4.10: Đường ROC

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Độ chính xác (accuracy) được đo lường bằng diện tích dưới đường cong ROC. Nếu diện tích bằng 1 là test rất tốt và nếu bằng 0,5 thì test không có giá trị. Trong hình trên kết quả diện tích ROC xấp xỉ 0.91 thuộc mức cao giúp việc nhận diện và phân loại của thuật toán về khả năng trả nợ của khách hàng tốt hơn.

4.3.4. Feature importance



Hình 4.11: Feature importance của Decision Tree

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Thực hiện đánh giá mức độ quan trọng của từng thuộc tính trong mô hình, các thuộc tính được sắp xếp theo thứ tự giảm dần như sau: Thu nhập, Số tài sản đảm bảo, Tuổi, Hôn nhân, Số lần quan hệ tín dụng, Số nguồn thu nhập, Số năm công tác. Kết quả đánh giá cho thấy thu nhập là yếu tố quyết định quan trọng bậc nhất trong mô hình, theo đó, tất cả các khách hàng trong tập dữ liệu sẽ được phân loại dựa trên thu nhập của họ thành từng nhánh cây khác nhau và mỗi nhánh cây sẽ tiếp tục xem xét các thuộc tính tiếp theo. Mặt khác, số nguồn thu nhập và số năm công tác được mô hình lựa chọn là yếu tố ít quan trọng nhất đối với khả năng trả nợ của khách hàng.

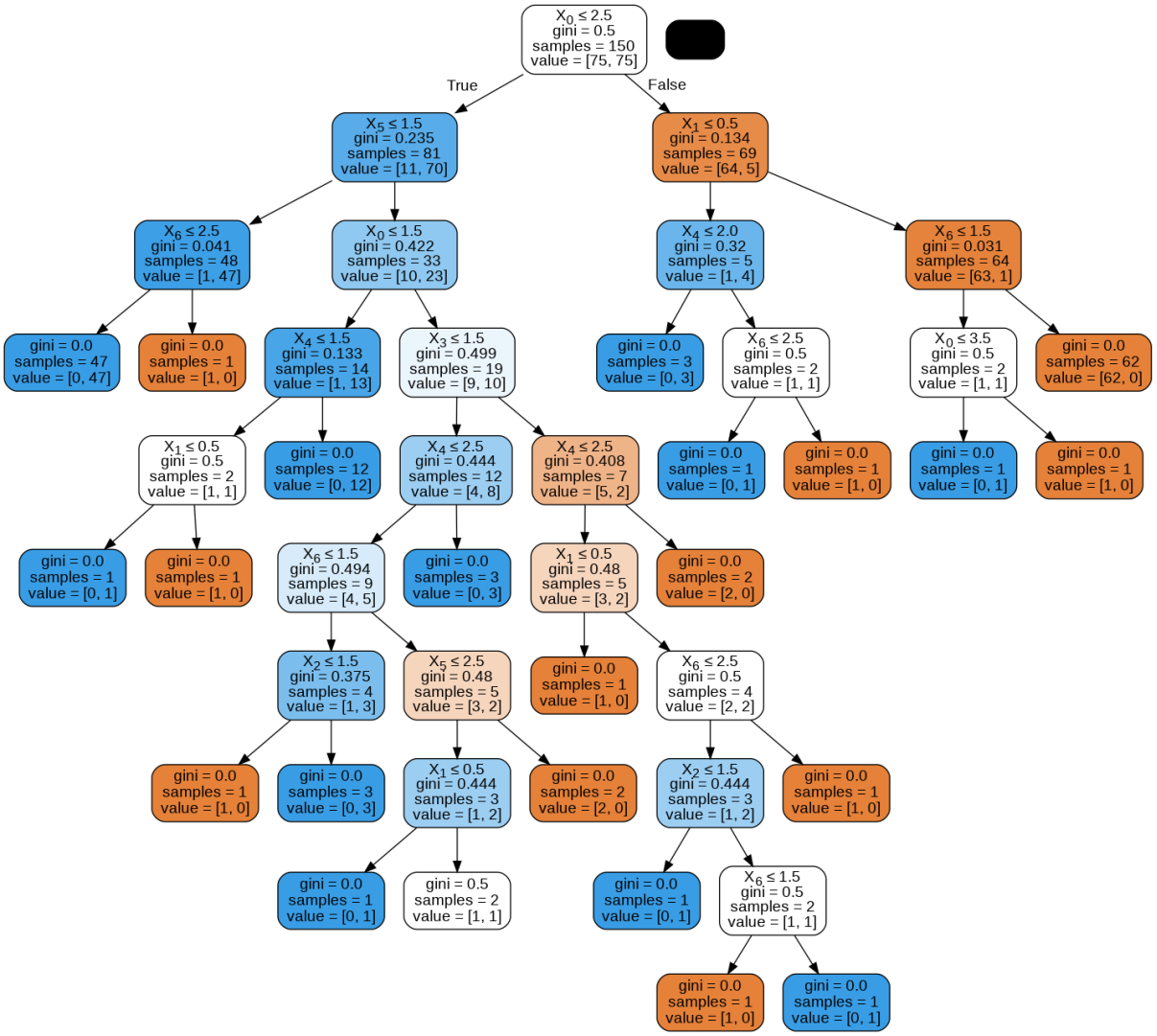
4.3.5. Kết quả của biểu đồ Decision Tree

```
X_0: Thu nhập
X_1: Tài sản đảm bảo
X_2: Số nguồn thu nhập
X_3: Số lần quan hệ tín dụng?
X_4: Hôn nhân
X_5: Tuổi
X_6: Số năm công tác
```

Hình 4.12: Biểu diễn các thuộc tính của bài nghiên cứu theo tên biến

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Kết quả phân tích cho biết "Thu nhập" là yếu tố quyết định đến khả năng trả nợ của khách hàng, tiếp theo sau đó lần lượt là "Tài sản đảm bảo", "Số nguồn thu nhập", "Số lần quan hệ tín dụng", "Hôn nhân", "Tuổi" và cuối cùng là "Số năm công tác".



Hình 4.13: Kết quả của biểu đồ Decision Tree

Nguồn: Trích xuất từ kết quả lập trình trên Google Colab

Theo kết quả của mô hình Decision Tree, thuộc tính quyết định quan trọng nhất ảnh hưởng đến đến khả năng trả nợ của khách hàng cá nhân đối với ngân hàng là "Thu nhập". Cây quyết định gợi ý phân loại khách hàng theo bốn nhóm thu nhập: (1) Khách hàng có thu nhập thấp (từ 39.425.000 đồng/tháng trở xuống); (2) Khách hàng có thu nhập trung bình (trên 39.425.000 đến 61.500.000 đồng/tháng); (3) Khách hàng có thu nhập cao (trên 61.500.000 đến 98.479.500 đồng/tháng); và (4) Khách hàng có thu nhập rất cao (trên 98.479.500 đồng/tháng). Cây quyết định phân loại khách hàng theo "Thu nhập" bao gồm 2 nhánh chính: Nhóm khách hàng có thu nhập từ dưới 39.425.000 đến 61.500.000 đồng/tháng trở xuống và từ 61.500.000 đồng/tháng đến 98.479.500 trở lên đồng/tháng.

Đối với nhóm khách hàng có thu nhập từ 39.425.000 đến 61.500.000 đồng/tháng trở xuống thì chưa đủ để kết luận khách hàng bị nợ quá hạn, Decision Tree xét thêm thuộc tính "Tuổi". Nếu khách hàng có số tuổi nhỏ hơn hoặc bằng 35 tuổi và sau đó đồng thời thỏa thuộc tính "Số năm công tác" dưới hoặc bằng 2 năm thì mô hình

Decision Tree đưa ra đánh giá tiềm ẩn rủi ro không trả được nợ ở mức rất cao, khi có 47/48 trường hợp thuộc trường hợp quá hạn có những thuộc tính này. Trong khi đó, chỉ có một trường hợp tuy có thu nhập trung bình trên 39.425.000 đến 61.500.000 đồng/tháng, độ tuổi của khách hàng nhỏ hoặc bằng 35 tuổi nhưng số năm công tác từ 2 đến 5 năm trở lên lại trả nợ đúng hạn. Điều này là một mang ý nghĩa rất lớn cho ngân hàng, các tổ chức tài chính nên lưu tâm đến mối quan hệ tương quan giữa số tuổi và số năm công tác của một khách hàng. Nếu khách hàng có số tuổi lớn và số năm công tác đồng thời lâu năm, điều đó chứng tỏ rằng khách hàng có sự ổn định trong tài chính, là người đi làm lâu năm, có nguồn thu nhập ổn định, vì thế hạn chế được rủi ro trả nợ quá hạn, ngân hàng nên cho phép vay. Ngược lại, nếu khách hàng có số tuổi lớn nhưng số năm công tác lại nhỏ, điều đó chứng tỏ rằng vị khách hàng này dù lớn tuổi nhưng không có sự ổn định trong công việc, hoặc là thường xuyên chuyển việc, nếu cho vay sẽ gây ra rủi ro quá hạn rất lớn.

Xem xét Decision Tree ở mức độ phức tạp hơn, đối với nhóm khách hàng có thu nhập dưới hoặc nằm trong khoảng 39.425.000 đến 61.500.000 đồng/tháng, xét thuộc tính "Tuổi", khách hàng có độ tuổi trung niên (từ 35 tuổi tới 65 tuổi) hay cao tuổi (trên 65 tuổi). Nếu nhóm khách hàng có mức thu nhập thấp từ dưới 39.425.000 đồng/tháng và thuộc diện đã kết hôn hoặc đã ly hôn, Decision Tree dự báo có 12/14 trường hợp trả nợ không đúng hạn. Ngược lại, nếu nhóm khách hàng thuộc nhóm thu nhập cao trong khoảng trên 61.500.000 đồng/tháng đến 98.479.500 đồng/tháng trở lên, Decision Tree sẽ xem xét thêm yếu tố số lần quan hệ tín dụng và hôn nhân. Đối với nhóm khách hàng thực hiện quan hệ tín dụng lần đầu và đã ly hôn, khách hàng có thể không được cấp tín dụng vì Decision Tree xác định có 3/12 trường hợp trả nợ không đúng hạn. Ngược lại đối với nhóm khách hàng đã li hôn nhưng quan hệ tín dụng trên 2 lần, có 2/7 trường hợp trả nợ đúng hạn nên có thể ra quyết định cấp tín dụng. Có thể nhận thấy, đối với nhóm khách hàng đã li dị, quan hệ tín dụng trở nên rất quan trọng vì tuy rằng người ly hôn sẽ bớt đi gánh nặng tài chính và nghĩa vụ vợ chồng, nhưng đồng thời cũng thiếu đi ràng buộc cũng như sự giúp đỡ tài chính của người vợ/chồng. Lúc này quyết định cho vay phụ thuộc vào mức độ uy tín, quan hệ tín dụng mà họ đã xây dựng trước đó.

Đối với nhóm khách hàng có thu nhập cao trên 61.500.000 đồng/tháng đến 98.479.500 trở lên, Decision Tree sẽ xem xét thêm thuộc tính "Tài sản đảm bảo". Nếu tài sản đảm bảo thuộc sở hữu người vay thì Decision Tree sẽ xem xét tiếp thuộc tính "Số năm công tác". Nếu số năm công tác của khách hàng từ 2 đến 5 năm trở lên thì ngân hàng đã đủ cơ sở để cấp tín dụng khi có 62/64 trường hợp khách hàng trả nợ đúng hạn. Phức tạp hơn, nếu số năm công tác của khách hàng từ bé hơn 2 năm thì phải xem xét tới thuộc tính "Thu nhập", tuy nhiên số mẫu cho thuộc tính này chỉ có 2/64 quan sát. Vì vậy với nhánh cây này, không đủ thông tin cũng như dữ liệu để ngân hàng có thể dựa vào đưa ra quyết định, nên khuyến khích không nên xem xét tiếp.

Vẫn xét ở nhóm khách hàng có thu nhập cao trên 61.500.000 đồng/tháng đến 98.479.500 trở lên, nếu khách hàng có thuộc tính "Tài sản đảm bảo" là không thuộc sở hữu người vay thì Decision Tree sẽ xem xét đến thuộc tính "Hôn nhân". Nếu khách hàng còn độc thân hoặc đã kết hôn, Decision Tree dự báo có 3/5 trường hợp trả nợ không đúng hạn. Điều này là hợp lý với thực tiễn, vì nếu khách hàng vừa độc thân hay đã kết hôn mà tài sản đảm bảo không thuộc sở hữu của bản thân, có thể khách hàng vẫn chưa độc lập tài chính, vẫn còn phụ thuộc vào người khác, vì thế rủi ro trả nợ không đúng hạn của nhóm khách hàng này cao. Có thể thấy thuộc tính "Tài sản đảm bảo" ở nhánh Decision Tree là quan trọng, nếu tài sản không phải của người đi vay, người đi vay không cần sợ mất đi tài sản đảm bảo vì thế trách nhiệm trả nợ cũng giảm đi đáng kể, ngân hàng phải đối diện với rủi ro nếu cấp tín dụng cho nhóm khách hàng này.

4.3. Thảo luận

Nghiên cứu này chính là sử dụng bài toán phân loại (classification) của Machine Learning để dự báo nhóm khách hàng có khả năng trả nợ đúng hạn/không đúng hạn. Vì vậy, sử dụng thuật toán Logistic Regression và Decision Tree là phù hợp với yêu cầu nghiên cứu. Tuy nhiên, mặc dù chỉ số accuracy của thuật toán Logistic Regression là 95,7% cao hơn so với thuật toán Decision Tree là 91,3% nhưng ở bài nghiên cứu này khuyến nghị sử dụng mô hình của thuật toán Decision Tree. Bởi vì, đối với Logistic Regression có một hạn chế là yêu cầu các thuộc tính của dữ liệu được tạo ra một cách độc lập với nhau. Trên thực tế, các thuộc tính trong bộ dữ liệu này bị ảnh hưởng bởi nhau. Điển hình như thuộc tính "Tuổi" và "Số năm công tác" có tương quan dương ở một số điểm dữ liệu. Trong khi đó, kết quả của Decision Tree cho thấy rõ các thuộc tính có tác động mạnh/yếu trực quan, cũng như sinh ra các quy tắc dễ hiểu cho người đọc. Vì vậy kết quả của Logistic Regression sẽ thành một kênh tham khảo cho nghiên cứu.

Kết quả nghiên cứu thông qua việc áp dụng Decision Tree đã đưa ra những đánh giá quan trọng về khả năng trả nợ của khách hàng cá nhân tại ngân hàng. Khả năng trả nợ đúng hay quá hạn chịu tác động đồng thời bởi nhiều thuộc tính. Cụ thể, "Thu nhập" là thuộc tính quan trọng nhất quyết định khả năng trả nợ của khách hàng cá nhân. Trong thực tế, thu nhập càng cao thì khả năng trả nợ càng được đảm bảo, kết quả này hoàn toàn trùng khớp với lược khảo nghiên cứu trước đó của đề tài của Kohansal & ctg (2009) và Trương Đông Lộc & ctg (2011). Như lược khảo, hai nghiên cứu này đều cho thấy tác động cùng chiều của thu nhập đến khả năng trả nợ của khách hàng. Tuy nhiên, có thể thấy thu nhập là yếu tố quan trọng nhưng chưa đủ để đánh giá khả năng trả nợ của khách hàng mà còn phải xem xét thêm ít nhất 2 yếu tố khác.

Trong đó yếu tố "Tuổi" và "Tài sản đảm bảo" cũng là những yếu tố quan trọng để đưa ra quyết định cấp tín dụng. Theo kết quả của mô hình, thuộc tính "Tuổi" có mối quan hệ tương quan thuận với thuộc tính "Số năm công tác" - thể hiện số năm kinh nghiệm trong lĩnh vực chuyên môn khách hàng đang công tác. Nếu số tuổi và số năm

công tác đồng thời cùng cao, điều đó chứng tỏ rằng khách hàng có sự ổn định trong tài chính và có cơ sở tín nhiệm để được cấp khoản vay cao hơn. Kết quả này cũng tương đồng với kết quả nghiên cứu của Trương Đông Lộc & ctg (2011). Thêm vào đó, thuộc tính "Tài sản đảm bảo" được chứng minh thông qua mô hình có mối tương quan tích cực lên khả năng trả nợ đúng hạn của khách hàng.

5. KẾT LUẬN VÀ KHUYẾN NGHỊ

5.1. Kết luận

Để hỗ trợ cho quá trình thẩm định tín dụng của khách hàng tại các ngân hàng, nghiên cứu đã chỉ ra được các thuộc tính tín dụng cho phép dự báo khả năng trả nợ đúng hạn/không đúng hạn của khách hàng cá nhân bao gồm: Thu nhập, Số tài sản đảm bảo, Tuổi, Hôn nhân, Số lần quan hệ tín dụng, Số nguồn thu nhập và Số năm công tác thông qua hai mô hình phân loại là Logistic regression và Decision tree, trong đó, Decision tree đã thể hiện nhiều ưu thế hơn nhờ khả năng biểu diễn mối quan hệ giữa các thuộc tính thành quy luật dưới dạng biểu đồ nhánh cây nhằm xác định trường hợp nào khách hàng trả nợ đúng hạn hoặc không đúng hạn. Kết quả của nghiên cứu này có độ chính xác tương đối cao trên cơ sở các điều kiện tiên quyết cho mô hình đều được thỏa mãn. Đóng góp của nghiên cứu là chỉ ra mối quan hệ giữa các thuộc tính tín dụng và biểu diễn chúng thành cấu trúc dưới dạng Decision tree, đồng thời cho thấy việc ứng dụng thuật toán Decision tree đối với các bài toán phân loại liên quan đến hành vi của khách hàng là tối ưu so với thuật toán Logistic regression.

5.2. Hạn chế

Để xây dựng mô hình phân loại đạt độ chính xác cao, bộ dữ liệu đầu vào đòi hỏi phải có tỷ lệ cân bằng giữa các quan sát, đặc biệt là các quan sát của thuộc tính quyết định "Thu nhập", đây là hạn chế thường gặp của thuật toán Decision tree trong quá trình xây dựng mô hình. Bên cạnh, dữ liệu sử dụng trong bài nghiên cứu là dữ liệu khảo sát với hình thức thu thập câu trả lời thông qua bảng câu hỏi với quy mô nhỏ, vì vậy dữ liệu chưa đủ khách quan trong việc đánh giá hay đại diện cho nhóm khách hàng cụ thể. Biểu đồ cấu trúc Decision tree của nghiên cứu chỉ mang tính chất tham khảo. Đối với các bài nghiên cứu sau này, vấn đề dữ liệu cần phải được lựa chọn và xử lý một cách cẩn trọng, khoa học và khách quan.

5.3. Khuyến nghị

Dựa trên kết quả phân tích mô hình Decision Tree, mô hình xây dựng một số gợi ý chính cho ngân hàng trong quản trị rủi ro tín dụng, bao gồm:

Thứ nhất, ngân hàng cần cẩn trọng đưa thu nhập thành yếu tố hàng đầu sẽ xét duyệt hồ sơ cho vay. Yếu tố thu nhập cao/thấp của khách hàng và tính ổn định tài chính chính là điều kiện thứ yếu giúp ngân hàng tránh được rủi ro không đáng có. Vì thu nhập chính là nguồn đảm bảo khả năng trả nợ, thông qua thu nhập, ngân hàng đánh giá đây là thước đo chính xác để đánh giá năng lực tài chính của khách hàng.

Thứ hai, khi xem xét hồ sơ vay vốn, ngân hàng cần ưu tiên cho khách hàng lớn tuổi. Khách hàng càng lớn tuổi thì càng chín chắn và sử dụng vốn hợp lí hơn, hiệu quả hơn khách hàng nhỏ tuổi. Đồng thời nhóm khách hàng này có nguồn tài chính ổn định.

Thứ ba, tài sản đảm bảo là nguồn thu thứ cấp để thu hồi vốn khi có rủi ro xảy ra. Vì vậy cần phải có sự xem xét kỹ càng tài sản thuộc quyền sở hữu của người đi vay hay không, thẩm định và định giá kỹ càng tài sản để hạn chế tình trạng phát sinh tranh chấp không đáng có cũng như tạo cho khách hàng động lực trả nợ đúng hạn.

Thứ tư, về tình trạng hôn nhân, khi xem xét hồ sơ vay vốn, ngân hàng cần ưu tiên những khách hàng đã có gia đình ổn định vì những khách hàng này luôn xem trọng sự tồn tại của gia đình mình và uy tín của gia đình đối với xã hội. Cũng như có cả hai nguồn tài chính từ vợ và chồng đảm bảo cho khả năng trả nợ tốt nhất

Cuối cùng, số lần quan hệ tín dụng của khách hàng là một yếu tố uy tín để ngân hàng dựa vào để đưa ra quyết định cho vay. Khi khách hàng có nhiều lần quan hệ tín dụng nghĩa khách hàng đã có lịch sử tín dụng tốt để ngân hàng cho vay nhiều lần. Điều này góp phần đưa ra quyết định đúng đắn cho ngân hàng.

Trên đây là những kiến nghị dựa trên kết quả phân tích với mục đích nâng cao khả năng trả nợ của khách hàng.

THAM KHẢO

Kohansal, M.R. & Mansoori, H. (2009), Factors Affecting Loan Repayment Performance of Farmers in Khorsan-Razavi Province of Iran, Conference on International Research on Food Security, Natural Resource Management and Rural Development, University of Hamburg.

Kumar Gupta, D., & Goyal, S. (2018). Credit Risk Prediction Using Artificial Neural Network Algorithm. International Journal Of Modern Education And Computer Science.P

Nguyễn Thị Viễn, Phạm Thị Thanh Xuân, Lê Thị Thanh Huyền (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, số 168.

Sihem Khemakhem & Younes Boujelbene (2018). "Predicting credit risk on the basis of financial and non-financial variables and data mining," Review of Accounting and Finance, Emerald Group Publishing, vol. 17(3), pages 316-340, August.

Trương Đông Lộc, Nguyễn Thanh Bình (2011). Các nhân tố ảnh hưởng đến khả năng trả nợ vay đúng hạn của nông hộ tỉnh Hậu Giang. Tạp chí Công nghệ ngân hàng, số 64.

Võ Văn Tài, Nguyễn Thị Hồng Dân và Nghiêm Quang Thường (2017). Đánh giá khả năng trả nợ vay của khách hàng bằng các phương pháp phân loại. Tạp chí Khoa học Trường Đại học Cần Thơ. 49a: 110-117.