TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

KHOA TÀI CHÍNH – NGÂN HÀNG

❖ ❖ ❖

BÁO CÁO CUỐI KỲ

MÔ HÌNH RỦI RO TÍN DỤNG TRONG R/PYTHON

Đề tài:

Đánh giá rủi ro tín dụng của khách hàng

GVHD: Phạm Thị Thanh Xuân

Nhóm sinh viên	**MSSV**
Lâm Nhựt Thịnh	K194141751
Thái Tuấn Kha	K194141725
Nguyễn Tuấn Hưng	K194141723
Trần Thanh Trúc	K194141755
Đoàn Thị Ngọc Diệu	K194141717

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

1. Giới thiệu:

- Tìm hiểu tín dụng của các ngân hàng hiện nay:
  + Ngân hàng BIDV: vừa công bố triển khai Chương trình hỗ trợ dành cho cán bộ y tế công tác tại các Bệnh viện, Cơ sở y tế trên toàn quốc (2022) với tên gọi "Đồng hành cùng ngành Y, chung tay vượt đại dịch". Bao gồm gói Tín dụng hỗ trợ nhu cầu đời sống (lãi suất vay cố định 1%/năm, số tiền vay lên đến 50 triệu, phương thức vay theo món/ Thấu chi/ Thẻ tín dụng); Gói Tín dụng hỗ trợ nhu cầu nhà ở (lãi suất 5.5%/năm trong 24 tháng).
  + Ngân hàng VCB: chương trình trả góp cho thẻ tín dụng VCB. Chủ thẻ Vietcombank có thể tận hưởng ưu đãi mua sắm ngay và trả dần trong vòng tối đa 12 tháng tại các Đối tác liên kết với Vietcombank. Một số lợi ích nổi bật của thẻ tín dụng: lãi suất 0% áp dụng trong toàn thời gian trả góp, Mua trước trả dần mỗi tháng trong vòng 3,6,9 hoặc 12 tháng).
  + Ngân hàng Shinhanbank: các chương trình hot nhất với thẻ tín dụng cá nhân hàng tháng (giai đoạn 10/05/2022 đến 30/06/2022), khách hàng sẽ được chính sách giá ưu đãi cho từng hạng thẻ theo từng gói tập, hoàn tiền cho chủ thẻ đạt chi tiêu bằng ngoại tệ tối thiểu trên thẻ tín dụng Shinhan sớm nhất mỗi tuần, quà tặng đặc biệt cho chủ thẻ đạt tổng chi tiêu ngoại tệ cao nhất mỗi tuần và không thấp hơn mức chi tiêu tối thiểu, Trả góp 0% lãi suất và không phí chuyển đổi dành cho chủ thẻ tín dụng cá nhân Shinhan (Áp dụng tại cửa hàng Digibox Estella), Giảm 110.000 đồng cho đơn hàng từ 1.100.000 đồng khi thanh toán bằng thẻ tín dụng Shinhan tại Tiki vào mỗi Ngày ưu đãi.
  + Ngân hàng TPBank: triển khai các chương trình khuyến mãi siêu khủng cho thẻ tín dụng khách hàng cá nhân: Đón hè sang săn ưu đãi siêu khủng bằng thẻ TPBank (Giảm 500k khi mua sắm tại Điện máy xanh - cho giao dịch thanh toán từ 8 triệu VNĐ trở lên, X10 lần tích dặm Bông Sen Vàng khi mua vé tại Vietnam Airline, Hoàn tiền 100k khi mua sắm tại UNIQLO, ZARA với chi tiêu từ 800k, áp dụng từ thứ 6 hàng tuần).
  + ……

- Ưu nhược điểm các nghiên cứu hiện nay về khả năng trả nợ của khách hàng: Ưu điểm: Đa phần các bài nckh sử dụng nhiều mô hình khác nhau chẳng hạn : Logistic regression, Random Forests, Naive Bayes, LightGBM, Neural networks,... sau đó dùng accuracy để lựa chọn mô hình tốt nhất.
    + Hồi quy logistic: mô hình này là mô hình cơ bản nhất để chạy thử các vấn đề phân loại. Bởi vì các giả định của hồi quy logistic hầu như không thực thi trong dữ liệu cuộc sống.
    + Random Forests: Random Forests thường thực hiện trên bộ dữ liệu mất cân bằng. Một lợi ích khác của việc chạy mô hình phân loại Random Forests trên bộ dữ liệu là nó cung cấp cho chúng tôi một cách nhìn trực quan các tính năng bằng cách liệt kê tầm quan trọng của từng Feature, trong đó cho chúng ta trực giác vào các yếu tố ảnh hưởng đến một người Khả năng trả nợ cho vay.
    + Naive Bayes: Naive Bayes sử dụng giả định ngây thơ về các Feature, có nghĩa là các Feature là độc lập có điều kiện với nhau. Naive Bayes Là một phương pháp tổng thể, khác với hai thuật toán trước đó. Sau đó, chúng ta có thể so sánh Hiệu suất giữa các phương pháp phân biệt đối xử cơ bản và phương pháp tổng quát này.

2. Khung lý thuyết

Tín dụng cá nhân là hình thức tín dụng mà theo đó ngân hàng chuyển quyền sử dụng vốn cho cá nhân, hộ gia đình có đăng ký kinh doanh cá thể trong một khoảng thời gian nhất định với một khoản chi phí nhất định nhằm mục đích phục vụ đời sống hoặc phục vụ sản xuất kinh doanh dưới hình thức hộ gia đình kinh doanh cá thể.

Khả năng trả nợ của khách hàng là việc khách hàng có khả năng trả nợ đầy đủ với bên cho vay hay không. Hiện tại vẫn chưa có định nghĩa thống nhất về khái niệm "khả năng trả nợ" mà chỉ có những dấu hiệu về việc khách hàng "không có khả năng trả nợ", đó là:

- 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.

Thông qua định nghĩa của IMF và các dấu hiệu mà Hiệp ước Basel II mô tả có thể thấy, thông thường việc khách hàng phát sinh nợ xấu đồng nghĩa với việc khách hàng không có khả năng trả nợ. Để thống nhất cách hiểu trong toàn bộ luận văn, nghiên cứu thống nhất việc đánh giá "khả năng trả nợ" của khách hàng sẽ được đánh giá thông qua nhóm nợ cao nhất tại các Tổ chức tín dụng khách hàng có quan hệ tín dụng. Những khách hàng hiện đang có nợ nhóm 3, 4, 5 được hiểu là nhóm khách hàng không có khả năng trả nợ, những trường hợp còn lại được hiểu là khách hàng có khả năng trả nợ.

Đã có nhiều nghiên cứu trong và ngoài nước về vấn đề này, sau đây là tóm tắt một vài nghiên cứu trước:

Bảng 1: Lược khảo nghiên cứu

STT	**Nghiên cứu**	**Nội dung nghiên cứu**	**Các biến độc lập trong mô hình nghiên cứu**	**Kết quả nghiên cứu**
1	Jonathan Crook (1995)	Giải pháp nhằm nâng cao khả năng trả nợ của khách hàng	10 biến: Độ tuổi, thu nhập, thu nhập ròng, sở hữu nhà riêng, giới tính, trình độ học vấn, nhu cầu vay, dư nợ, ngành kinh doanh, lãi suất.	Khả năng trả nợ chịu ảnh hưởng tích cực từ yếu tố độ tuổi của chủ hộ, yếu tố thu nhập, thu nhập ròng và sở hữu nhà riêng.
2	Roslan &	Các yếu tố ảnh hưởng đến khả năng trả nợ	10 biến: Giới tính, độ tuổi, thu nhập, tỷ	Kết quả nghiên cứu cho thấy

	Karim (2009)	của các đối tượng tín dụng vi mô tại Agribank	lệ nợ quá hạn, lĩnh vực sản xuất, quy mô khoản vay, thời hạn cho vay, thâm niên, người phụ thuộc, lãi suất.	những người vay hoạt động trong lĩnh vực dịch vụ ít rủi ro hơn đối với những người hoạt động trong lĩnh vực sản xuất, (iv) quy mô khoản vay càng lớn thì tỷ lệ nợ quá hạn càng thấp, (v) thời hạn cho vay tác động ngược chiều và có ý nghĩa thống kê, và (vi) thời gian cho vay càng dài thì tỷ lệ nợ quá hạn càng cao.
3	Bekhet & Eletter (2014)	Nghiên cứu về nhu cầu vay nợ của các hộ gia đình.	13 biến: Độ tuổi, giới tính, thu nhập, thâm niên, kinh nghiệm, người phụ thuộc, loại hình công ty, nguồn trả nợ dự phòng, tỷ số nợ/thu nhập, thu nhập, kỳ hạn vay, quy mô khoản vay,	7 biến có ý nghĩa thống kê với quyết định cấp tín dụng từ ngân hàng. Đó là độ tuổi, giới tính, tổng thu nhập, loại hình công ty khách hàng làm việc,

			và lãi suất.	nguồn trả nợ, tỷ số nợ/thu nhập, và tổng thu nhập.
4	Đặng Thị Cẩm Nhung (2015)	Phân tích yếu tố tác động đến khả năng trả nợ của khách hàng cá nhân tại Agribank Long An	16 biến: Độ tuổi, giới tính, nghề nghiệp, tình trạng hôn nhân, sở hữu nhà, người phụ thuộc, thời gian cư trú, thời gian làm công việc hiện tại, lịch sử tín dụng, thời hạn vay, thu nhập, chi tiêu, quy mô khoản vay, tài sản đảm bảo, lãi suất.	Tài sản thế chấp là động sản thì khả năng trả nợ vay tốt hơn các tài sản thế chấp khác. Thời hạn vay càng dài thì khả năng trả nợ vay tốt hơn những hộ vay thời gian ngắn. Thu nhập bình quân của hộ càng cao thì càng đảm bảo khả năng trả nợ tốt hơn.

Tín dụng là nguồn tài chính quan trọng cho các hoạt động tiêu dùng, đầu tư kinh doanh, sản xuất của các cá nhân và hộ gia đình. Giúp cải thiện dòng tiền của các hộ gia đình, tăng cơ hội sản xuất kinh doanh, cải thiện nguồn vốn, thực hiện hóa các ý tưởng kinh doanh. Hoạt động tín dụng là nguồn thu đáng kể của các ngân hàng tín dụng, tuy nhiên hoạt động này chứa nhiều rủi ro. Nguyên nhân chính là từ những hạn chế trong khả năng nhận diện và đo lường khả năng trả nợ của khách hàng. Vì vậy có

thể nói rằng khả năng trả nợ đúng hạn và các thuộc tính tác động hiện đang là vấn đề được các tổ chức tín dụng và giới nghiên cứu rất quan tâm.

Hiện nay, ngành Ngân hàng việt Nam đang trong quá trình hòa nhập với ngành ngân hàng trong khu vực và trên thế giới nên tính an toàn trong hoạt động tín dụng vẫn còn chưa cao. Hiểu rõ những yếu tố tác động đến khả năng trả nợ của khách hàng sẽ giúp ngân hàng có cơ sở đưa ra quyết định nên cho vay hay không đối với các khách hàng mới, và đối với các khách hàng hiện hữu thì việc đánh giá này sẽ giúp ngân hàng có thể nhận diện được những khách hàng có khả năng trả nợ quá hạn, từ đó đưa ra những biện pháp xử lý để hạn chế rủi ro từ những khoản vay của những khách hàng này.

Đã có nhiều công trình nghiên cứu liên quan đến các yếu tố ảnh hưởng đến khả năng trả nợ của khách hàng cá nhân. Nguyễn Hải Trường (2017) nghiên cứu phân loại khách hàng (“tốt” – không có nợ quá hạn, “xấu” – có nợ quá hạn) trên tập dữ liệu kiểm tra dựa vào giá trị ngưỡng xác suất 0.5, nghĩa là nếu xác suất khách hàng có nợ quá hạn lớn hơn 0.5 thì khách hàng đó được gọi là khách hàng “xấu". Kết quả mô hình hồi quy Logistic dựa trên dữ liệu huấn luyện cho thấy với mức ý nghĩa trên 99% các yếu tố 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.

T.T.Phong và Cộng sự (2019) nghiên cứu về khả năng trả nợ của khách hàng cá nhân được thực hiện tại Agribank Tân Hưng, Tỉnh Long An. Với kích thước n = 300 và chọn 300 khách hàng gần nhất. Kết quả nghiên cứu cho thấy có biến thành phần tác động có ý nghĩa thống kê tới khả năng trả nợ của khách hàng cá nhân: Tài sản (Asset) cho thấy khi khách hàng có tài sản thì ý thức trả nợ càng cao vì họ biết rằng đó là điều kiện đảm bảo cho khoản vay với ngân hàng. Ngoài ra chỉ số EXP = 6,519 cho thấy người có nhà thì khả năng trả nợ cao hơn người không có nhà 6,519 lần (nếu các yếu tố khác không đổi). Theo nghiên cứu này thì yếu tố Tài sản là yếu tố quan trọng nhất cần được xem xét vì nhà ở là tài sản đảm bảo thu hồi vốn cho ngân hàng.

Với biến thu nhập, nghiên cứu Kohansal & Mansoori (2009), Acquah & Addo (2011), Trương Đông Lộc & Nguyễn Thanh Bình (2011) chỉ ra rằng thu nhập có mối tương quan thuận đến khả năng trả nợ của khách hàng. Khi khách hàng có thu nhập càng cao thì khả năng trả nợ càng cao và ngược lại.

Với biến công việc, nghiên cứu của Valentina Michelangeli and Enrico Sette (2016), I Baklouti (2013) chỉ ra rằng những người vay có công việc tốt có khả năng trả lại các khoản nợ vay tốt hơn.

3. Phương pháp dữ liệu:

3.1. Dữ liệu

Bộ dữ liệu được sử dụng để chạy mô hình dự báo bao gồm 10000 quan sát được nhóm tác giả tự mô phỏng sau khi đã lược khảo những bài nghiên cứu đi trước, những dòng đầu tiên của dữ liệu được thể hiện qua bảng dưới đây:

Bảng 2: Dữ liệu

GENDER	COLLATERAL	OWN_JOB	MARRIED	INCOME	TARGET
1	1	0	0	1	0
0	0	0	0	2	1
1	1	0	0	0	0
0	1	0	0	1	1
1	1	1	1	1	0
…	…	…	…	…	…

Quan sát bảng trên có thể nhận thấy được có tất cả 6 biến được nhóm tác giả lựa chọn. Các biến này được phân chia và giải thích như sau:

Biến độc lập:

- *GENDER*: biến này thể hiện giới tính của khách hàng trong quan sát đó (0: Nữ, 1: Nam).
- *COLLATERAL*: biến này xác định xem tài sản đảm bảo cho khoản vay của khách hàng có thuộc quyền sở hữu của chính họ hay không hay là thuộc quyền sở hữu của người khác (bảo lãnh) (0: Không, 1: Có).
- *OWN_JOB*: biến này xác định xem khách hàng vay đang có việc làm hay không (0: Không, 1: Có).
- *MARRIED*: biến này xác định tình trạng hôn nhân của của khách hàng (0: độc thân, 1: đã có gia đình).
- *INCOME*: biến này xác định mức thu nhập của khách hàng vay tiền (0: Thấp (thu nhập ≤ 39.425.000), 1: Trung bình (39.425.000 < thu nhập ≤ 61.500.000), 2: Cao (61.500.000 < thu nhập).

Biến phụ thuộc:

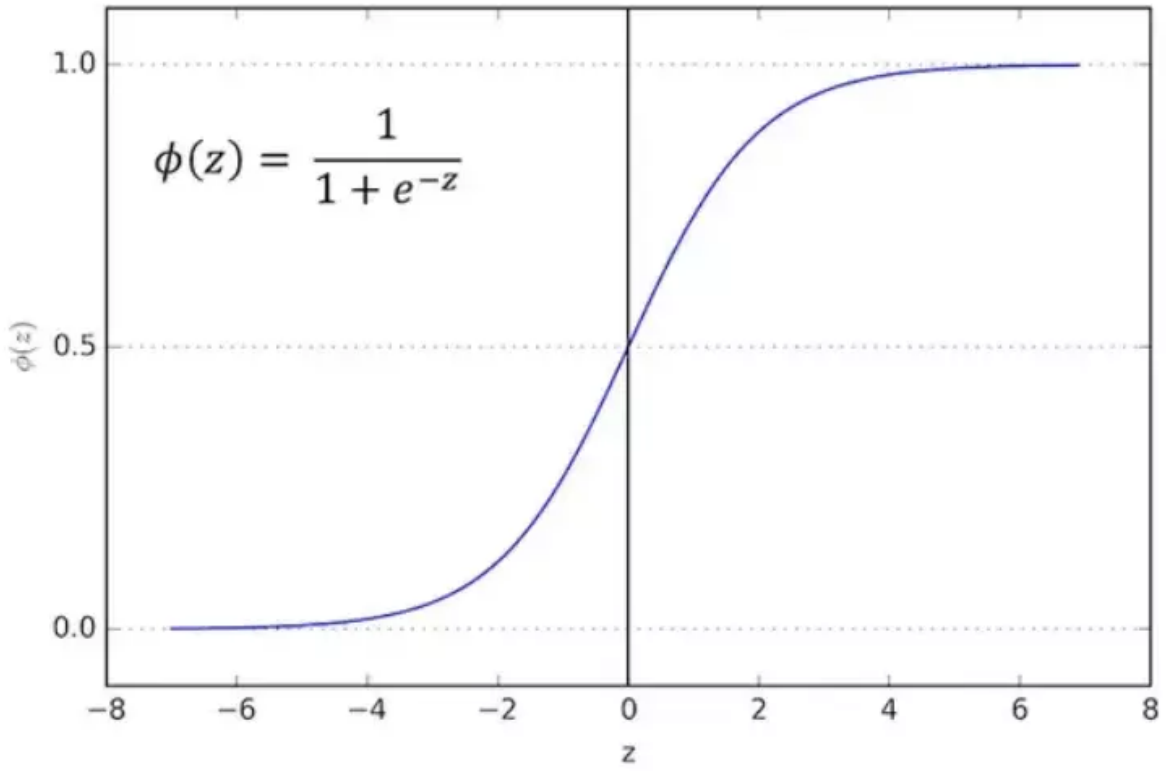
TARGET: đây là biến phụ thuộc của mô hình, biến này sẽ xác định khả năng trả nợ của khách hàng. (0: Trả nợ đúng hạn, 1: Trả nợ không đúng hạn)

3.2. Dự báo bằng mô hình Logistic

Logistic Regression là 1 thuật toán được dùng để phân loại các đối tượng cho 1 lớp nào đó (class). Các lớp thường được mô phỏng thông qua một tập giá trị rời rạc (như 0, 1, 2, ...). Một ví dụ rất phổ biến là bài toán phân loại email có phải là một email rác (spam) hay không. Thuật toán trên dùng hàm sigmoid logistic để đưa ra đánh giá và phân loại theo xác suất. Ví dụ: Email này 60% là spam, giao dịch này 90% là gian lận, …

Hàm Sigmoid (hay còn gọi là hàm số Logistic) là một hàm số toán học có đường cong dạng hình chữ “S” với công thức và đồ thị như sau:

Hình 1: Minh họa phương trình hàm Sigmoid



Nguồn: https://viblo.asia/

Mọi giá trị khi đi qua hàm Sigmoid sẽ nằm trong miền giá trị số thực chạy từ 0 đến 1. Hàm sigmoid sẽ trả về giá trị xác suất tương ứng với mỗi giá trị được đưa vào. Tùy thuộc vào ngưỡng xác suất (threshold) được đặt ra (giá trị này tùy thuộc vào tính chất của từng bài toán khác nhau), các quan sát sẽ được phân vào các lớp khác nhau.

3.3. Dự báo bằng mô hình Decision Tree

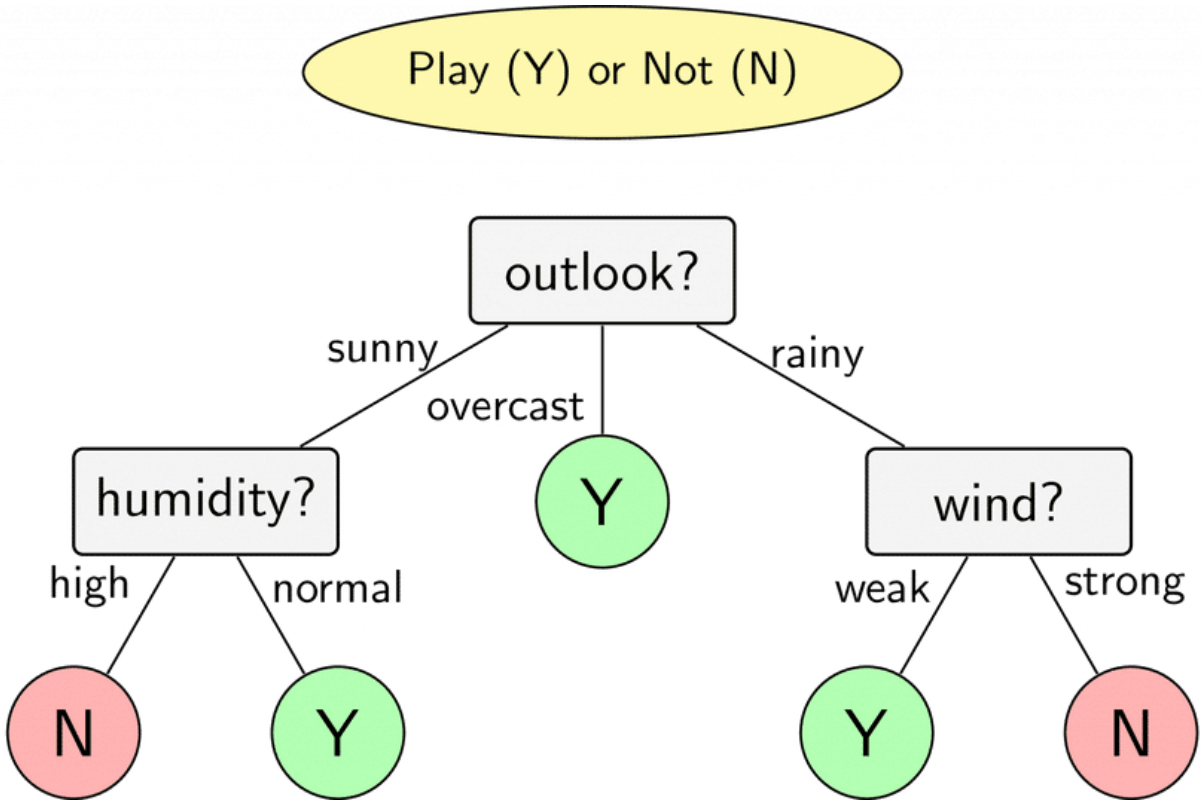
Thuật toán Decision Tree là thuật toán học có giám sát trong Machine learning. Mô hình được sử dụng rộng rãi với vai trò là mô hình giải quyết các bài toán phân loại và hồi quy. Thuật toán hoạt động dựa trên cách thức phân cụm thứ bậc thông qua lệnh

If/else để dẫn đến các quyết định, vì thế nên mô hình mô phỏng khá tương đồng với suy nghĩ của con người.

Xây dựng mô hình Decision Tree hay còn có nghĩa là xây dựng một chuỗi các câu hỏi If/else để trả về câu trả lời chính xác một cách nhanh chóng nhất. Trong Machine Learning, các câu hỏi này được gọi là tests. Decision tree còn là một mô hình liên tiếp hợp nhất loạt test cơ bản một cách hiệu quả và gắn kết trong đó giá trị số được so sánh với giá trị ngưỡng (Threshold) trong mỗi test. Với bài toán được đặt ra trong bài nghiên cứu này, mô hình phân loại cây quyết định (Decision Tree Classifier Model) sẽ được sử dụng để dự báo.

Decision Tree Classifier là mô hình phân loại hoạt động dưới dạng một cây phân cấp có cấu trúc được dùng để phân loại các đối tượng dựa vào dãy các luật (rules). Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau.

Hình 2: Minh họa mô hình Decision Tree Classifier

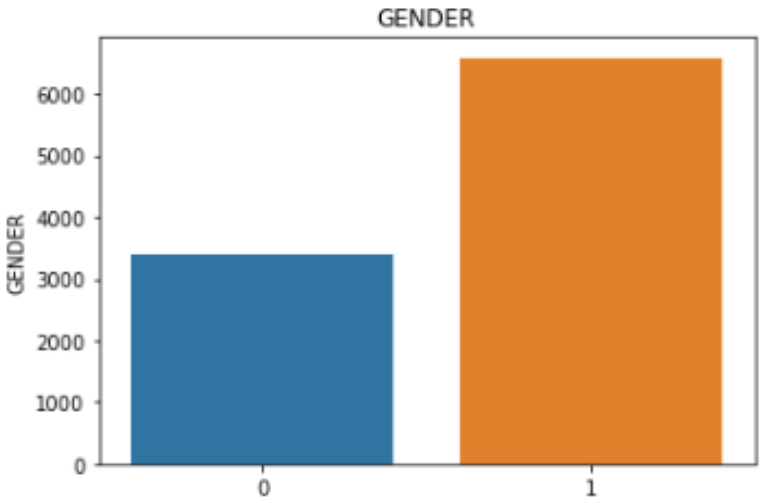


Nguồn: https://trituenhantao.io/

4. Kết quả

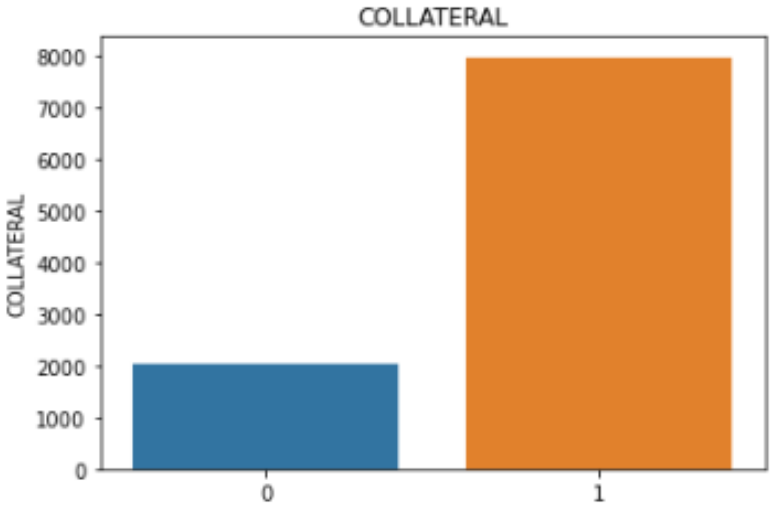
4.1. Thống kê mô tả

Hình 3: Thống kê mô tả biến GENDER



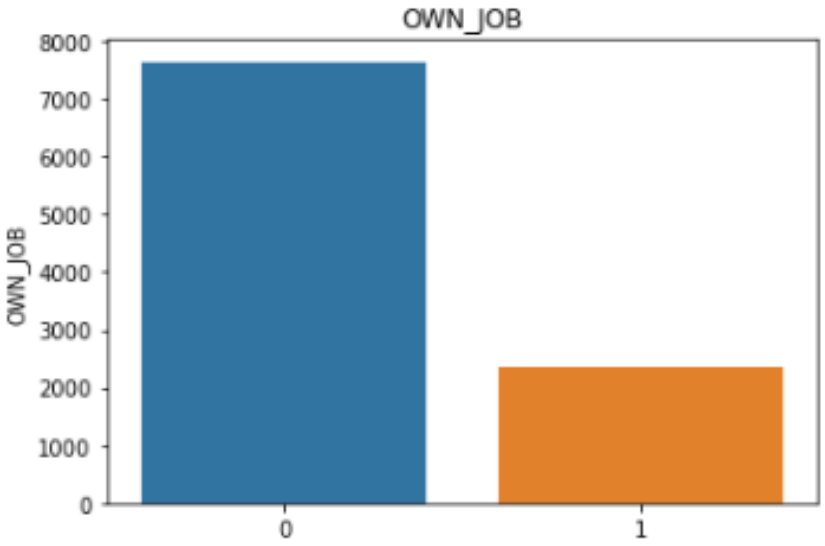
Trong bộ dữ liệu, đa số khách hàng đi vay là nam với hơn 6000 người (chiếm hơn 60%), chiếm phần thiểu số với hơn 3000 người là các khách hàng nữ.

Hình 4: Thống kê mô tả biến COLLATERAL



Số lượng khách hàng đi vay sở hữu tài sản đảm bảo khá áp đảo với gần 80% (gần 8000 người), còn lại các khách hàng được người khác bảo lãnh khoản vay chiếm số lượng khá khiêm tốn với khoảng 2000 người.

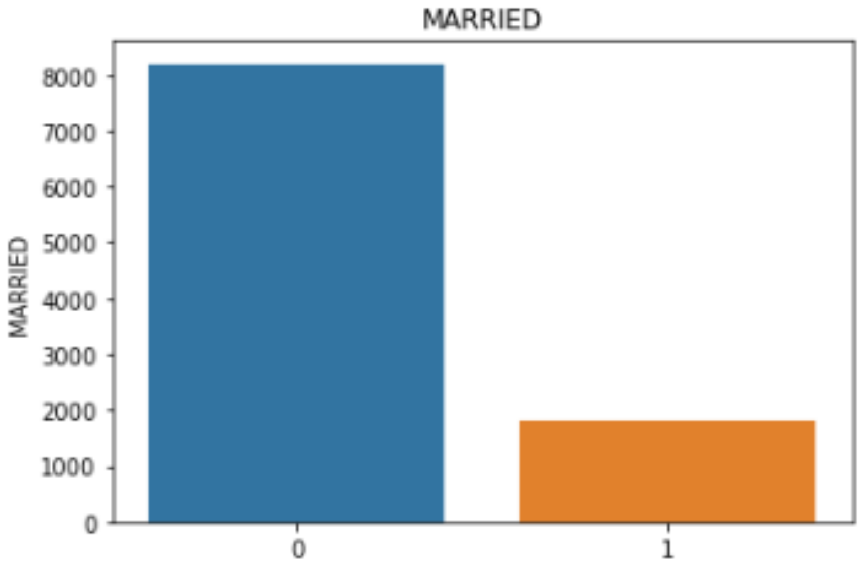
Hình 5: Thống kê mô tả biến OWN_JOB



Phần lớn khách hàng trong bộ dữ liệu đi vay trong tình trạng không có việc làm với gần 8000 người. Điều này có vẻ khá mâu thuẫn với thống kê của biến COLLATERAL khi đại đa số các khách hàng có sở hữu tài sản đảm bảo cho khoản vay. Tuy nhiên, từ

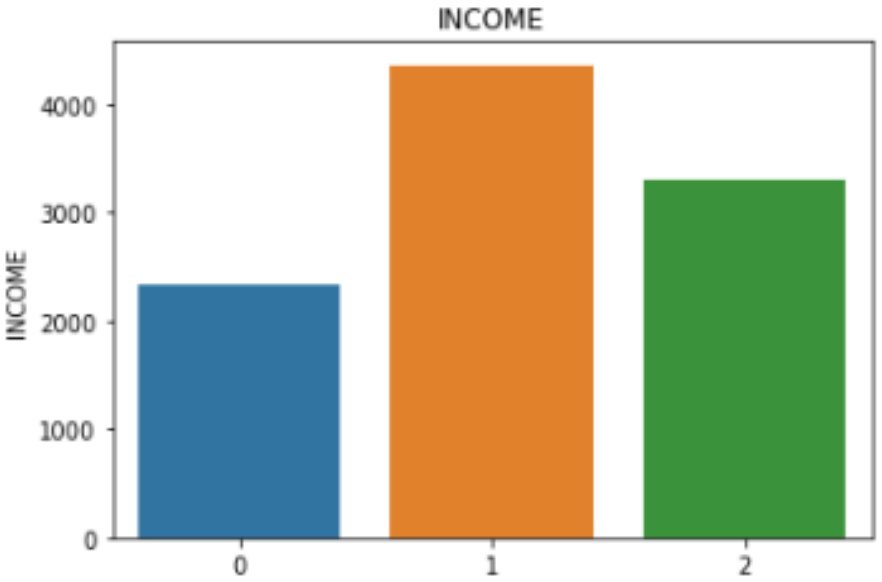
các thống kê có vẻ trái ngược đó có thể nhận xét được rằng trong bộ dữ liệu của bài nghiên cứu này có một số lượng không nhỏ khách hàng thất nghiệp nhưng lại sở hữu tài sản đảm bảo cho khoản vay. Đối chiếu với thực tế thì đây có thể là những người lớn tuổi đã về hưu, đã sở hữu tài sản trong tay tuy nhiên lại thiếu tiền để phục vụ cho mục đích nhất định nào đó nên đã đi vay.

Hình 6: Thống kê mô tả biến MARRIED



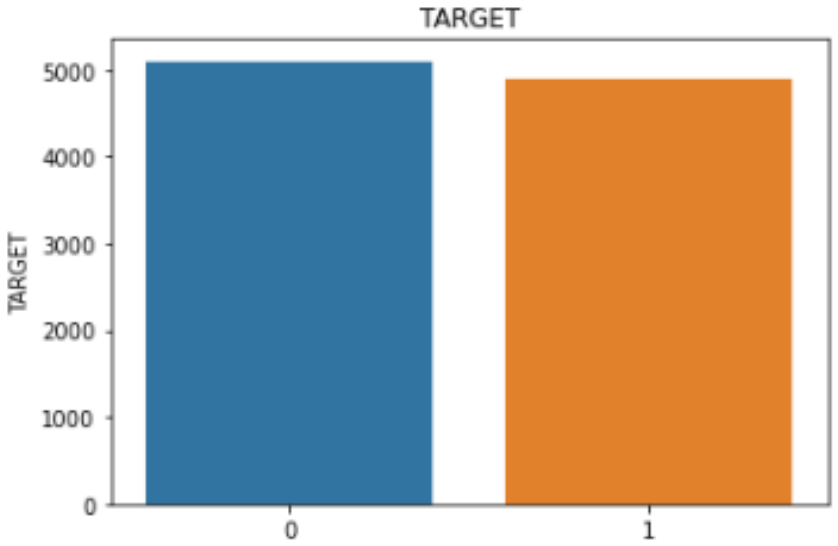
Có hơn 80% khách hàng độc thân đi vay trong bộ dữ liệu.

Hình 7: Thống kê mô tả biến INCOME



Thu nhập của phần lớn khách hàng đi vay nằm ở mức trung bình, tiếp theo sau đó là những người có thu nhập ở mức cao và chiếm số lượng ít nhất là các khách hàng có thu nhập thấp.

Hình 8: Thống kê mô tả biến TARGET

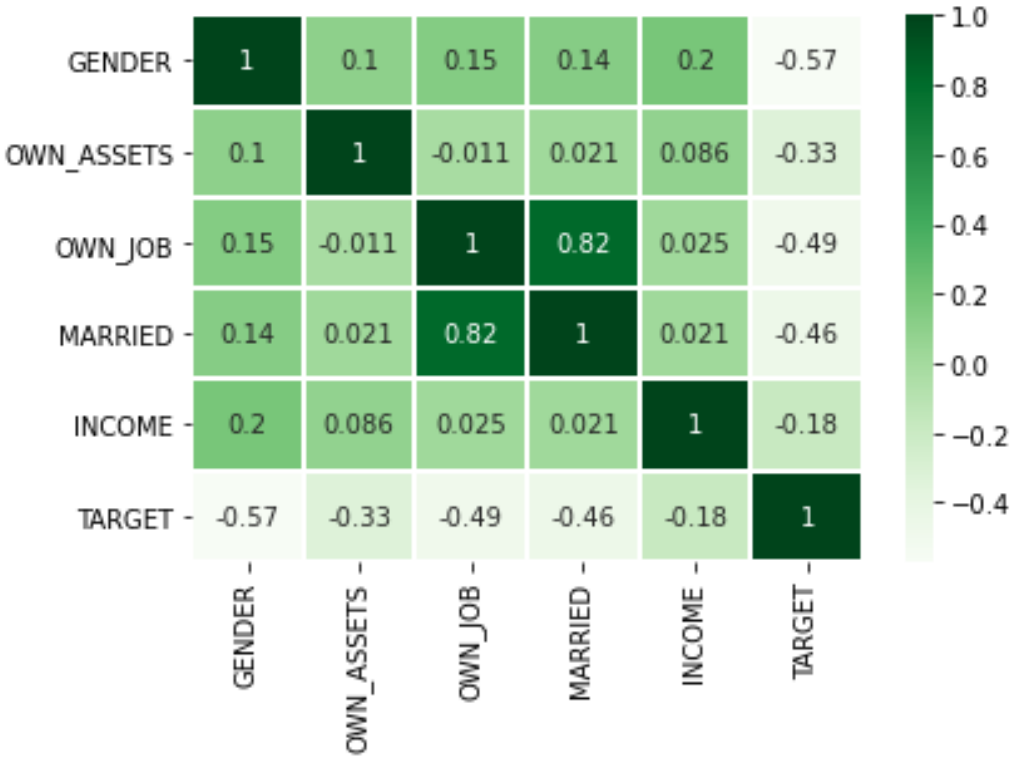


Nhóm người đi vay trả nợ đúng hạn (class = 0) và nhóm người đi vay trả nợ trễ hạn (class = 1) có số lượng khá ngang bằng nhau. Nhóm trả nợ đúng hạn chỉ nhỉnh hơn một chút.

4.2. Feature Selection

4.2.1. Correlation

Hình 9: Ma trận tương quan



Ngoài 2 biến OWN_JOB và MARRIED có tương quan dương lớn với nhau (0.82), các cặp biến còn lại không có tương quan quá lớn với nhau. Biến mục tiêu (TARGET) có tương quan âm với tất cả các biến độc lập trong mô hình, từ đó có thể suy ra một số nhận xét ban đầu như sau:

- Khách hàng nam có xu hướng trả nợ đúng hạn.
- Khách hàng sở hữu tài sản đảm bảo có xu hướng trả nợ đúng hạn.
- Khách hàng đang có việc làm có xu hướng trả nợ đúng hạn.
- Khách hàng đã lập gia đình có xu hướng trả nợ đúng hạn.
- Khách hàng có thu nhập càng cao thì khả năng trả nợ đúng hạn càng cao.

4.2.2. Feature Importance

Feature Importance là một thuật toán xác định mức độ "quan trọng" của mỗi biến độc lập trong việc dự báo giá trị của biến phụ thuộc, mức độ "quan trọng" của các biến được sắp xếp dựa trên chỉ số "điểm quan trọng" (Importance Score).

Bảng 3: Logistic Regression Importance Score

FEATURE NAME	**LOGISTIC REGRESSION IMPORTANCE SCORE**
COLLATERAL	-4.365954
GENDER	-3.546528
OWN_JOB	-3.436177
MARRIED	-3.432777
INCOME	-0.364668

Ở mô hình Logistic Regression, importance score được tính dựa trên coefficient của từng biến mà mô hình đã tính toán ra được. Biến nào có độ lớn (trị tuyệt đối) của coefficient càng lớn thì mức độ "quan trọng" của biến đó càng lớn. Biến COLLATERAL (tài sản đảm bảo có do người đi vay nắm giữ hay không?) là biến quan trọng nhất trong việc dự báo biến TARGET với importance score đạt mức khoảng -4.37, theo sau là biến GENDER (giới tính) với importance score đạt mức khoảng -3.55. Biến INCOME (thu nhập) được xác định là "ít quan trọng nhất" trong mô hình này với importance score chỉ khoảng -0.36.

Bảng 4: Decision Tree Importance Score

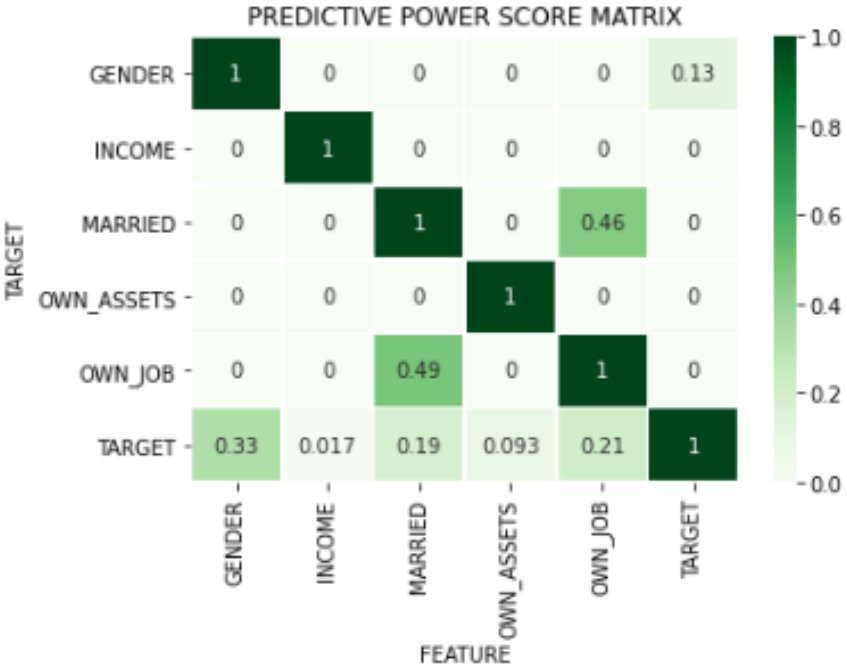
FEATURE NAME	**DECISION TREE IMPORTANCE SCORE**
GENDER	0.368977
OWN_JOB	0.327338
COLLATERAL	0.248702
MARRIED	0.047406
INCOME	0.007577

Ở mô hình Decision Tree, importance score của biến nào càng lớn thì mức độ "quan trọng" của biến đó càng lớn. Biến quan trọng nhất trong việc dự báo TARGET là GENDER (giới tính) với importance score cao nhất (khoảng 0.37). Theo sau là biến OWN_JOB (người vay có sở hữu công việc hay không?) với importance score đạt mức khoảng 0.33. Biến INCOME (thu nhập) được xác định là "ít quan trọng nhất" trong mô hình này với importance score chỉ khoảng 0.008.

4.2.3. Predictive Power Score

Predictive power score là điểm số thể hiện mối quan hệ tuyến tính hoặc phi tuyến tính giữa hai cột. Điểm số dao động từ 0 (không có sức mạnh dự đoán) đến 1 (sức mạnh dự đoán hoàn hảo).

Hình 10: Ma trận Predictive Power Score



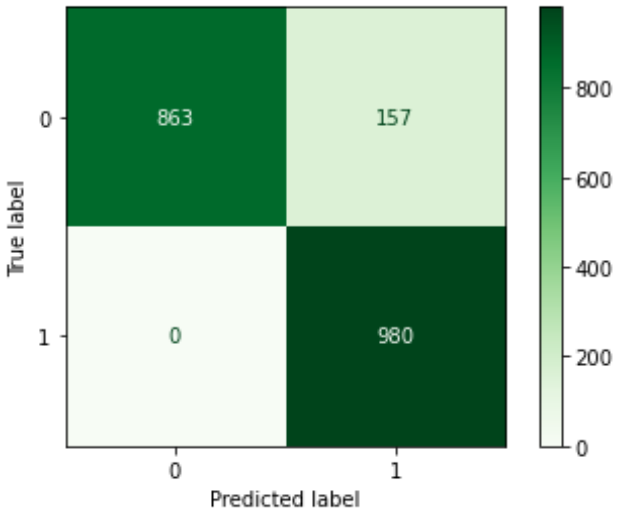
Hầu hết các biến đều không có quan hệ tuyến tính hay phi tuyến với nhau khi predictive power score giữa các biến này đều đạt giá trị là 0 (không có sức mạnh dự đoán) ngoại trừ cặp biến MARRIED và OWN_JOB có độ tương quan (correlation) lớn được đề cập ở mục 4.2.1 có predictive power score là 0.49 và 0.46. Điều này có nghĩa là mỗi biến trong cặp biến này đều có "sức mạnh dự báo" trong bài toán dự báo biến còn lại. Xét đến predictive power score giữa biến TARGET và các biến độc lập thì nhận thấy biến GENDER (giới tính) có "sức mạnh dự báo" cao nhất với predictive power score đạt mức 0.33, biến INCOME (thu nhập) có "sức mạnh dự báo" thấp nhất với predictive power score chỉ đạt mức 0.017.

4.3. Kết quả Logistic Regression

Hình 11: Kết quả Logistic Regression

	precision	recall	f1-score	support
0	1.00	0.85	0.92	1020
1	0.86	1.00	0.93	980
accuracy			0.92	2000
macro avg	0.93	0.92	0.92	2000
weighted avg	0.93	0.92	0.92	2000

Logistic Regression Accuracy: 92.15%



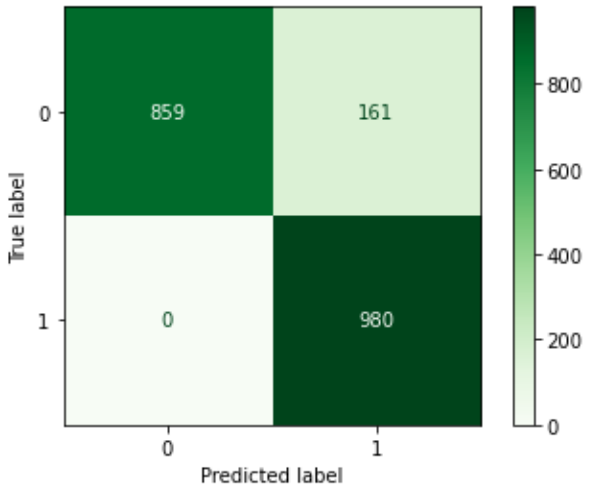
Độ chính xác trong dự báo của mô hình (Accuracy) Logistic Regression là 92.15%, một con số khá tốt. Chi tiết kết quả dự báo thể hiện đầy đủ trên confusion matrix. Có thể nhận thấy rằng các quan sát thuộc class 1 (nhóm những người chậm trả nợ) đều được dự đoán chính xác, giá trị recall(class=1) = 1 cũng thể hiện điều này. Đây là một tín hiệu tốt đối với bài toán phân loại đã đặt ra vì thông thường các ngân hàng hay các tổ chức tài chính rất chú trọng vào nhóm các đối tượng trả nợ trễ vì nhóm này có khả năng gây ra rủi ro tín dụng không mong muốn cho họ. Vì vậy việc dự báo chính xác tất cả những khách hàng có khả năng trả nợ trễ giúp các định chế tài chính có những biện pháp tiếp cận và giải quyết sớm hơn để có thể giảm thiểu rủi ro tín dụng mà họ có thể mắc phải, hoặc thậm chí họ có thể không cho vay ngay từ đầu.

4.4. Kết quả Decision Tree

Hình 12: Kết quả Decision Tree

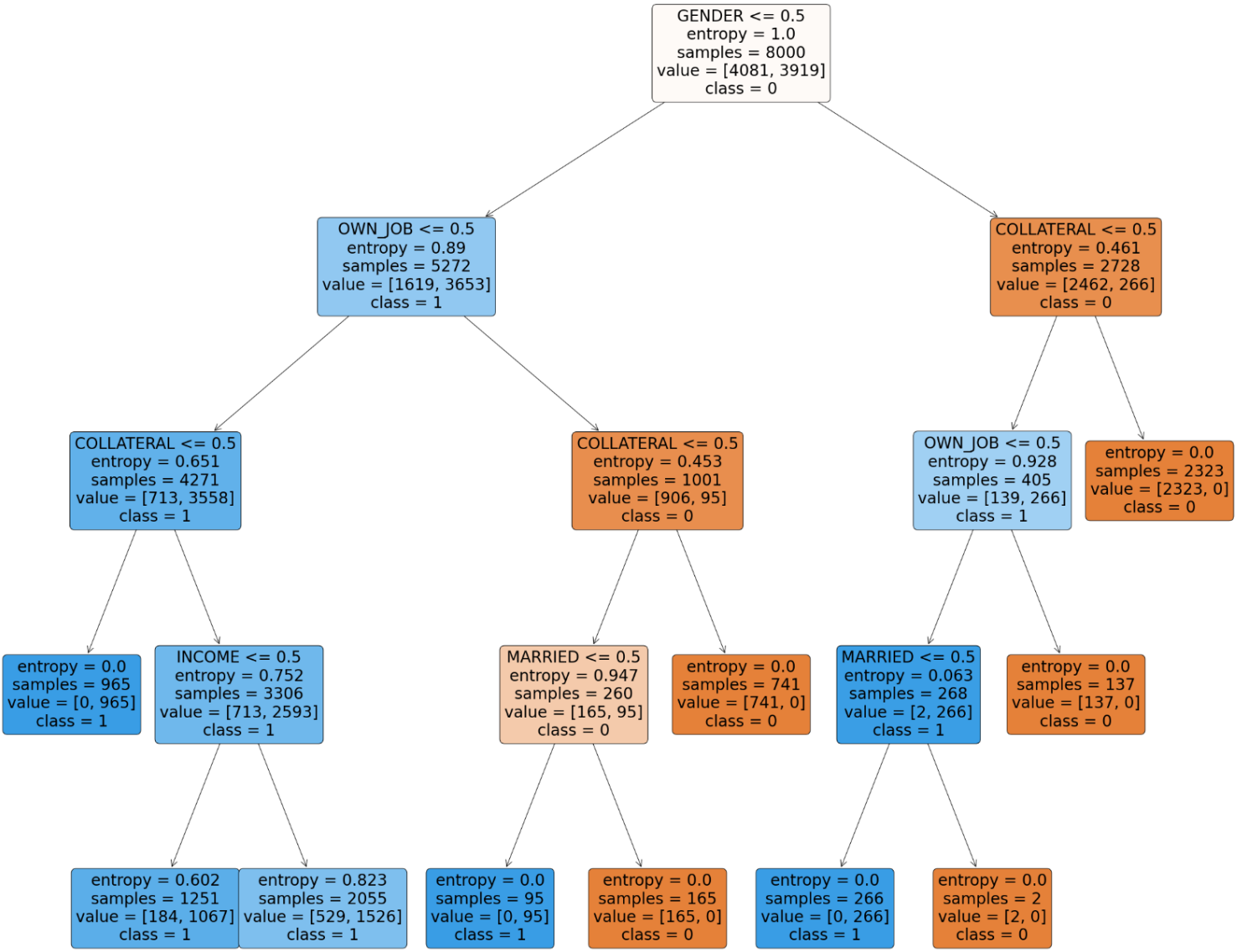
	precision	recall	f1-score	support
0	1.00	0.84	0.91	1020
1	0.86	1.00	0.92	980
accuracy			0.92	2000
macro avg	0.93	0.92	0.92	2000
weighted avg	0.93	0.92	0.92	2000

Decision Tree Accuracy: 92.0%



Độ chính xác trong dự báo của mô hình Decision Tree đạt mức 92%, thấp hơn một chút so với mô hình Logistic. Theo như phân tích ở mục 4.3, nhóm các khách hàng trả nợ trễ hạn (class = 1) là nhóm được quan tâm nhất và cần phải được dự báo với tỉ lệ chính xác cao nhất (thể hiện qua chỉ số recall(class=1)). Và giống với mô hình Logistic, mô hình Decision Tree đã dự báo một cách chính xác, không để bỏ sót nhóm các khách hàng chậm trả nợ (giá trị recall(class=1) = 1). Vì vậy đây cũng là một kết quả tốt. Quá trình phân loại của cây quyết định được mô tả qua hình ảnh dưới đây:

Hình 13: Mô phỏng quá trình phân loại mô hình Decision Tree



Từ cây quyết định, mỗi một đường dẫn từ gốc đến nút lá trong cây tạo thành một luật, luật này có vế trái là một bộ giá trị của các thuộc tính được chọn để phân lớp, vế phải là một trong các giá trị của thuộc tính kết quả.

Trong thuật toán Cây quyết định, Entropy là phương pháp lựa chọn cách phân nhánh tối ưu dựa trên cơ sở tối đa hóa lượng thông tin nhận vào "Information maximization", tức giảm thiểu tối đa độ hỗn độn và nhiễu loạn trong từng node. Các node phân nhánh được lựa chọn theo phương pháp này phải thể hiện tối đa thông tin cần thiết để cây quyết định có thể phân loại chính xác đối tượng dữ liệu vào các tập con có chứa nhãn,

là giá trị/ thuộc tính của biến đầu vào. Nếu tất cả các đối tượng trong một node, một tập con đều mang giá trị là nhãn của phân nhánh, hay chính tập con đó thì Entropy có giá trị bằng 0, ngược lại nếu các đối tượng chứa nhiều giá trị khác nhau thì Entropy sẽ cao nhất và bằng 1.

Theo kết quả mô hình cây quyết định, thuộc tính quyết định quan trọng bậc nhất để phân tách các nhóm khả năng trả nợ của khách hàng cá nhân là "GENDER", tức "Giới tính". Cây quyết định gợi ý phân loại khách hàng theo 2 nhóm giới tính: (1) Khách hàng có giới tính là nam; (0) Khách hàng có giới tính là nữ. Ta có thể thấy ở node này, giá trị entropy = 1 - giá trị cao nhất, nghĩa là node này có mức độ hỗn độn dữ liệu thuộc mỗi nhóm của biến phụ thuộc cao nhất (trong số 8000 dữ liệu đầu vào thì có 4081 khách hàng trả nợ đúng hạn và 3919 khách hàng trả nợ không đúng hạn). Trong trường hợp thỏa mãn điều kiện 0, có 5272 khách hàng được phân vào nhóm giới tính nữ, trong đó có 3653 khách hàng trả nợ không đúng hạn. Trong trường hợp thỏa mãn điều kiện 1, có 2728 khách hàng được phân vào nhóm giới tính nam, trong đó có chỉ có 266 khách hàng trả nợ không đúng hạn.

- **Cây quyết định gợi ý đối với nhóm khách hàng là nữ (GENDER = 0):** Sau khi đã tách được 5272 khách hàng có giới tính nữ sau lần phân tách đầu tiên thì thuật toán sẽ xét đến thuộc tính tiếp theo để tách nhóm, đó là khách hàng vay đang có việc làm hay không - "OWN_JOB" (0: Không, 1: Có).

    + Trong trường hợp khách hàng thỏa mãn điều kiện đang không có việc làm: có 4271 khách hàng được phân vào nhóm không có việc làm, trong đó có 3558 khách hàng không trả nợ đúng hạn, giá trị entropy là 0,651. Với nhóm khách hàng thỏa đang có việc làm, cây quyết định sẽ thêm thuộc tính tiếp theo là tài sản đảm bảo cho khoản vay của khách hàng có thuộc quyền sở hữu của chính họ hay không - "COLLATERAL" (0: Không, 1: Có). Nếu khách hàng thỏa mãn điều kiện 0 ở thuộc tính "COLLATERAL" thì tiềm ẩn rủi ro không trả được nợ khi cây quyết định đã chỉ ra 965/965 trường hợp nợ quá hạn. Ngược lại, trong trường

hợp khách hàng thỏa mãn điều kiện 1 ở thuộc tính "COLLATERAL", cây quyết định sẽ thêm một thuộc tính khác để phân chia, đó là mức thu nhập của khách hàng vay tiền - "INCOME" (0: Thấp, 1: Trung bình, 2: Cao), giá trị entropy lúc này là 0,752. Đối với nhóm khách hàng có thu nhập từ trung bình đến cao, có 2055 khách hàng được phân vào nhóm này, trong đó có 1526/2055 trường hợp khách hàng trả nợ không đúng hạn, giá trị entropy là 0,823. Đối với nhóm khách hàng có thu nhập thấp, cây quyết định cho thấy tiềm ẩn rủi không trả được nợ cao khi có đến 1067/1251 trường hợp khách hàng trả nợ không đúng hạn, giá trị entropy ở nhóm này là 0,602. Như vậy có thể thấy tiêu chí mức thu nhập của khách hàng vay tiền không tác động tới việc có tiềm ẩn rủi ro hay không vì ở mức thu nhập nào thì thuật toán đều phân vào nhóm trả nợ không đúng hạn, chỉ khác nhau ở xác suất xảy ra rủi ro trả nợ không đúng hạn.

+ Trong trường hợp khách hàng thỏa mãn điều kiện đang có việc làm: có đến 1001 khách hàng được phân vào nhóm có việc làm, nhưng chỉ có 95 khách hàng không trả nợ đúng hạn. Lúc này, trong trường hợp thỏa mãn điều kiện 1 ở thuộc tính "COLLATERAL", rủi ro xảy khả năng trả không trả được nợ đúng hạn là rất thấp khi có 741/741 khách hàng trả được nợ đúng hạn. Trong trường hợp thỏa mãn điều kiện 0 ở thuộc tính "COLLATERAL", thuật toán đã thêm một thuộc tính khác vào các thuộc tính có ở các lần phân tách trước để có thể phân loại chính xác hơn khi cùng với một giả thiết là những khách hàng là nữ, đang có việc làm, tài sản đảm bảo cho khoản vay của khách hàng không thuộc quyền sở hữu của chính họ, đó là thuộc tính tình trạng hôn nhân của của khách hàng (0: độc thân, 1: đã có gia đình) - "MARRIED". Trong trường hợp điều kiện 0 ở thuộc tính "MARRIED" được thỏa mãn, rủi ro trả nợ không đúng hạn là rất cao khi có 95/95 khách hàng trả nợ không đúng hạn. Ngược lại, nếu trường hợp điều kiện 1 ở thuộc tính "MARRIED" được thỏa mãn thì có đến 165/165 khách hàng trả nợ đúng hạn. Cả 2

trường hợp này đều có giá trị entropy là 0 nên thuật toán sẽ không tiếp tục phân tách ở 2 nhóm này nữa.

- **Cây quyết định gợi ý đối với nhóm khách hàng là nam (GENDER = 1):** Sau khi đã tách được 2728 khách hàng có giới tính nam sau lần phân tách đầu tiên thì thuật toán sẽ xét đến thuộc tính tài sản đảm bảo cho khoản vay của khách hàng có thuộc quyền sở hữu của chính họ hay không - "COLLATERAL (0: Không, 1: Có) để tách nhóm.

    + Trong trường hợp khách hàng thỏa mãn điều kiện tài sản đảm bảo cho khoản vay của khách hàng không thuộc quyền sở hữu của chính họ: có 405 khách hàng được phân vào nhóm có tài sản đảm bảo cho khoản vay thuộc quyền sở hữu của chính họ, trong đó có 266 khách hàng không trả nợ đúng hạn, giá trị entropy là 0,928. Với nhóm khách hàng thỏa mãn điều kiện có tài sản đảm bảo cho khoản vay thuộc quyền sở hữu của chính họ, cây quyết định sẽ đến thuộc tính tiếp theo, đó là "OWN_JOB" (0: Không, 1: Có). Nếu khách hàng thỏa mãn điều kiện 1 ở thuộc tính "OWN_JOB" thì rủi ro không trả được nợ là rất thấp khi cây quyết định đã chỉ ra 137/137 trường hợp trả nợ đúng hạn. Ngược lại, trong trường hợp khách hàng thỏa mãn điều kiện 0 ở thuộc tính "OWN_JOB" , cây quyết định sẽ thêm một thân chia, đó là tình trạng hôn nhân của của khách hàng (0: độc thân, 1: đã có gia đình) - "MARRIED" với giá trị entropy là 0,063. Đối với nhóm khách hàng độc thân, rủi ro không trả được nợ đúng hạn là 100% khi cây quyết định chỉ ra rằng 266/266 trường hợp trả nợ không đúng hạn. Ngược lại, đối với nhóm khách hàng đã có gia đình, có 2 khách hàng được phân vào nhóm này và cả 2 người này đều trả nợ đúng hạn, giá trị entropy ở hai nhóm này là đều là 0 nên thuật toán không phân tách nữa.

    + Trong trường hợp khách hàng thỏa mãn điều kiện tài sản đảm bảo cho khoản vay của khách hàng có thuộc quyền sở hữu của chính họ: có 2323

khách hàng được phân vào nhóm có tài sản đảm bảo cho khoản vay không thuộc quyền sở hữu của chính họ, và cả 2323 người này đều trả nợ đúng hạn, giá trị entropy là 0 nên thuật toán cũng dừng phân tách ở nhóm này.

Ta có thể rút ra một số luật từ cây quyết định vừa xây dựng:

- Nếu khách hàng là nữ, đang không có việc làm và tài sản đảm bảo cho khoản vay của khách hàng không thuộc quyền sở hữu của chính họ thì ngân hàng có thể cân nhắc không 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. Tương tự, nếu khách hàng đáp ứng những điều kiện trên nhưng tài sản đảm bảo cho khoản vay của khách hàng thuộc quyền sở hữu của chính họ thì họ đều rơi vào nhóm khách hàng có khả năng trả nợ không đúng hạn hoặc chưa xác định được, bất kể họ có thu nhập 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.

- Ở trường hợp nếu một khách hàng là nữ nhưng đang có việc làm và tài sản đảm bảo cho khoản vay của khách hàng không thuộc quyền sở hữu của chính họ thì ngân hàng phải xem xét tiếp tình trạng hôn nhân của khách hàng. Nếu khách hàng đã có gia đình thì được phân vào nhóm trả nợ đúng hạn và ngân hàng có thể xem xét để cho vay. Tương tự, nếu khách hàng là nữ, đang có việc làm và tài sản đảm bảo cho khoản vay của khách hàng thuộc quyền sở hữu của chính họ 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 là nữ, đang có việc làm, tài sản đảm bảo cho khoản vay của khách hàng không thuộc quyền sở hữu của chính họ nhưng vẫn đang còn độc thân thì thuộc nhóm trả nợ không đúng hạn. Vì vậy, ngân hàng không nên cho vay để tránh tiềm ẩn rủi ro.

- Ở trường hợp nếu một khách hàng là nam, tài sản đảm bảo cho khoản vay của khách hàng không thuộc quyền sở hữu của chính họ, thì ngân hàng xem xét tiếp yếu tố khách hàng đang có việc làm hay không. Nếu khách hàng đó đang có việc làm 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 đang không có việc làm có thì ngân hàng phải xem xét tiếp yếu tố tình trạng hôn nhân của khách hàng. Nếu khách hàng còn độc thân 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 đã có gia đình thì lại thuộc nhóm khách hàng trả nợ đúng hạn. Tuy nhiên, ngân hàng cần cân nhắc và thẩm định kĩ càng hơn để đưa ra quyết định có cho vay hay không đối với những hồ sơ này vì những trường hợp này này vẫn có nguy cơ rủi ro khi số khách hàng trả nợ đúng hạn là cực kì thấp.

- Nếu một khách hàng là nam, tài sản đảm bảo cho khoản vay của khách hàng thuộc quyền sở hữu của chính họ 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.

5. Kết luận và khuyến nghị cho ngân hàng

5.1. Kết luận

Để hạn chế rủi ro tín dụng, các ngân hàng cần nâng cao khả năng đánh giá tình hình tài chính và thẩm định khả năng thanh toán nợ vay của khách hàng. Với báo cáo này, kết quả nghiên cứu phân tích các nhân tố ảnh hưởng đến khả năng trả nợ của khách hàng cá nhân thông qua thuật toán Logistic Regression và Decision Tree cho thấy các yếu tố đều có tác động đến khả năng trả nợ, bao gồm: GENDER, MARRIED, OWN_JOB, COLLATERAL, INCOME. Trong đó thuộc tính quyết định quan trọng nhất đến khả năng trả nợ của khách hàng cá nhân trong bộ dữ liệu là giới tính (GENDER). Điều này phù hợp với kết quả nghiên cứu của Nguyễn Hải Trường (2017). Bên cạnh đó yếu tố tài sản (COLLATERAL) cũng được đánh giá là có tác

động không kém so với biến GENDER, tài sản thế chấp nên là động sản thì khả năng trả nợ vay tốt hơn các tài sản thế chấp khác (T.T.Phong và Cộng sự (2019)). Mô hình cây quyết định đem lại hiệu quả dự báo rất tốt, bên cạnh đó sơ đồ cây quyết định đã cung cấp một góc nhìn trực quan, dễ hiểu về quá trình cũng như các tiêu chí phân loại của cây quyết định. Ngoài ra, mô hình Logistic Regression cũng đem lại những kết quả rất ấn tượng.

5.2. Khuyến nghị

Dựa vào kết quả của mô hình, nhóm đưa ra một số khuyến nghị như sau:

Thứ nhất, ngân hàng cần đặc biệt chú ý đến giới tính của người đi vay, đây là điều kiện hàng đầu xét duyệt khả năng cho vay.

Thứ hai, công việc là một yếu tố thứ yếu liên quan đến khả năng trả nợ của khách hàng. Một cá nhân có công việc sẽ giúp họ có được nguồn thu nhập từ đó làm giảm thiểu khả năng trả nợ không đúng hạn. ở bậc cao hơn, ngân hàng cần chú ý xem liệu khách hàng đang có công việc ổn định hay là công việc không ổn định. Ngân hàng cũng cần liên kết với các doanh nghiệp nơi khách hàng làm việc để đánh giá về hiệu suất làm việc của khách hàng.

Thứ ba, với tài sản đảm bảo, ngân hàng cần có các phương pháp xác định về nguồn gốc tài sản đảm bảo, phương pháp định giá tài sản đảm bảo. Ngân hàng cần xem xét liệu tài sản đảm bảo có đang bị vấn đề về pháp lý, khả năng thanh khoản kém. Có phương pháp dự phòng rủi ro tài sản đảm bảo. Đặc biệt, cần hạn chế các khoản vay có tài sản đảm bảo từ bên thứ ba.

Thứ tư, với thu nhập của khách hàng. Thu nhập của khách hàng thể hiện khả năng dòng tiền của khách hàng. Ngân hàng cần đảm bảo cho việc thẩm định chính xác tình hình tài chính của khách hàng, có các mô hình thích hợp phát hiện gian lận các nguồn thu nhập của khách hàng.

Trên đây là một số khuyến nghị dựa vào mô hình cây quyết định. Từ nghiên cứu trên nhóm cũng đề xuất các ngân hàng cần có mô hình rủi ro tín dụng phù hợp với xu thế 4.0, thẩm định dựa trên công nghệ blockchain, liên kết việc thẩm định, đánh giá giữa ngân hàng với doanh nghiệp nơi khách hàng làm việc, giữa ngân hàng với hệ thống dữ liệu chính phủ, từ đó giúp cho quá trình diễn ra nhanh chóng, chính xác hơn.

REFERENCES

[1] Baklouti, I. (2013). Determinants of microcredit repayment: The case of Tunisian Microfinance Bank. African Development Review, 25(3), 370-382.

[2] Bekhet, H. A., & Eletter, S. F. K. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. Review of Development Finance, 4(1), 20-28.

[3] Crook, J. (1995). Time series explanations of merger activity: some econometric results. *International Review Of Applied Economics*, *9*(1), 59-85.

[4] Dang Thi Cam, N. (2022). Phân Tích Yếu Tố Tác Động Đến Khả Năng Trả Nợ Của Khách Hàng Cá Nhân Tại Agribank Long An.

[5] Kohansal, M. R., & Mansoori, H. (2009, October). Factors affecting on loan repayment performance of farmers in Khorasan-Razavi province of Iran. In *Conference on International Research on Food Security, Natural Resource Management and Rural Development, University of Hamburg* (Vol. 26, pp. 359-366).

[6] Michelangeli, V., & Sette, E. (2016). How does bank capital affect the supply of mortgages? Evidence from a randomized experiment. Evidence from a Randomized Experiment (February 25, 2016). Bank of Italy Temi di Discussione (Working Paper) No, 1051.

[7] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), 275-285.

[8] Nguyen Hai, T. (2022). 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.

[9] Roslan, A.H. and Karim, M.Z.A. (2009) Determinants of Microcredit Repayment in Malaysia The Case of Agrobank. Humanity & Social Science Journal, 4, 45-52.

[10] Thanh Phong, T., Thanh Bình, N., Xuân Trang, L., & 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).

[11] Truong Dong, L., & Nguyen Thanh, B. (2022). Các Nhân Tố Ảnh Hưởng Đến Khả Năng Trả Nợ Vay Đúng Hạn Của Nông Hộ Ở Tỉnh Hậu Giang.