BỘ TÀI CHÍNH BỘ GIÁO DỤC VÀ ĐÀO TẠO

HỌC VIỆN CHÍNH SÁCH VÀ PHÁT TRIỂN

HOÀNG THU HƯƠNG

KHÓA LUẬN TỐT NGHIỆP

CHUYÊN NGÀNH: PHÂN TÍCH DỮ LIỆU LỚN TRONG KINH TẾ VÀ KINH DOANH

TÊN ĐỀ TÀI: XÂY DỰNG MÔ HÌNH HỌC MÁY TRONG DỰ BÁO VÀ TỐI ƯU HẠN MỨC TÍN DỤNG TẠI NGÂN HÀNG THƯƠNG MẠI VIỆT NAM

Hà Nội, năm 2025

BỘ TÀI CHÍNH BỘ GIÁO DỤC VÀ ĐÀO TẠO

HỌC VIỆN CHÍNH SÁCH VÀ PHÁT TRIỂN

KHÓA LUẬN TỐT NGHIỆP

Giáo viên hướng dẫn: **TS. Giang Thành Trung**

Sinh viên thực hiện: **Hoàng Thu Hương**

Mã sinh viên: **7123112094**

Lớp: **DLL12**

Hà Nội, năm 2025

LỜI CẢM ƠN

Trước tiên, em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến giảng viên trực tiếp hướng dẫn thầy TS. Giang Thành Trung – người đã tận tình hướng dẫn, định hướng và hỗ trợ em trong suốt quá trình thực hiện khóa luận tốt nghiệp. Những góp ý chuyên môn quý báu và sự đồng hành của thầy không chỉ giúp em hoàn thiện đề tài mà còn là nguồn động lực lớn để em nỗ lực không ngừng trong quá trình nghiên cứu.

Tiếp theo, em xin chân thành cảm ơn toàn thể các thầy cô cán bộ, giảng viên trong trường nói chung và các thầy cô Khoa Kinh tế số nói riêng đã tận tình giảng dạy, truyền đạt kiến thức và tạo điều kiện thuận lợi cho em trong quá trình học tập.

Dù đã rất cố gắng nghiên cứu và tìm hiểu nhưng do còn hạn chế về kiến thức và kinh nghiệm, khóa luận của em khó tránh khỏi những thiếu sót. Em hy vọng sẽ nhận được những ý kiến đóng góp quý báu từ các thầy cô để hoàn thiện khóa luận tốt nghiệp của mình hơn nữa.

Em xin chân thành cảm ơn!

Sinh viên thực hiện

Hương

Hoàng Thu Hương

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

LỜI MỞ ĐẦU

1. Lý do chọn đề tài

Trong bối cảnh chuyển đổi số đang diễn ra mạnh mẽ như hiện nay, việc ứng dụng dữ liệu lớn (Big Data) và học máy (Machine Learning) đã trở thành xu hướng tất yếu trong ngành tài chính. Những công nghệ này không chỉ giúp các tổ chức tài chính nâng cao khả năng phân tích mà còn tạo ra những đột phá trong việc tối ưu hóa các quy trình tín dụng, đặc biệt là trong việc dự báo hạn mức tín dụng cho khách hàng. Hạn mức tín dụng đóng vai trò quan trọng trong việc duy trì sự ổn định của hệ thống tài chính và đảm bảo quản lý rủi ro tín dụng hiệu quả. Việc xác định và điều chỉnh chính xác hạn mức tín dụng giúp các tổ chức tài chính hạn chế các rủi ro tiềm ẩn trong khi vẫn đảm bảo khách hàng có thể tiếp cận các sản phẩm tín dụng phù hợp với khả năng tài chính của mình. Báo cáo của ngành tài chính cho thấy, học máy mang lại tiềm năng to lớn trong việc tự động hóa quy trình, giảm thiểu sai sót đồng thời cải thiện hiệu suất hoạt động của các doanh nghiệp. Các ứng dụng như dự báo tín dụng, phát hiện gian lận và tối ưu hóa danh mục đầu tư đã được triển khai rộng rãi và mang lại hiệu quả rõ rệt.

Theo báo cáo của Ngân hàng Nhà nước Việt Nam vào năm 2022 một số ngân hàng lớn như Ngân hàng VPBank (VPB), Ngân hàng HDBank (HDB), Ngân hàng Quân Đội (MBB) và Ngân hàng Vietcombank (VCB) đã được điều chỉnh tăng hạn mức tín dụng, qua đó giúp tăng trưởng tín dụng toàn ngành lên 19,2% so với đầu năm đạt 530,1 nghìn tỷ đồng. Con số này phản ánh sự quan trọng trong việc điều chỉnh hạn mức tín dụng nhằm đáp ứng nhu cầu vay vốn ngày càng cao của nền kinh tế, đồng thời duy trì sự ổn định của các tổ chức tài chính. Mặc dù vậy quá trình xác định hạn mức tín dụng không hề đơn giản, đòi hỏi sự phân tích kỹ lưỡng và chính xác về khả năng chi trả của khách hàng đồng thời phải tính toán các yếu tố rủi ro tác động đến quyết định tín dụng. Theo báo cáo từ Ngân hàng Nhà nước, tỷ lệ tín dụng/GDP của Việt Nam đã đạt mức 14% trong năm 2021, cho thấy mức độ tín dụng trong nền kinh tế đang tăng trưởng

nhanh chóng tạo ra áp lực trong việc kiểm soát và điều chỉnh hạn mức tín dụng một cách hợp lý.

Với những lý do trên, tôi lựa chọn đề tài "Xây dựng mô hình học máy trong dự báo và tối ưu hạn mức tín dụng tại ngân hàng thương mại Việt Nam" để hỗ trợ việc dự báo hạn mức tín dụng chính xác. Đây không chỉ là cơ hội để sinh viên áp dụng những kiến thức đã học vào thực tế mà còn phản ánh xu hướng công nghệ hiện đại trong ngành tài chính và tín dụng, nhằm nâng cao hiệu quả hoạt động, quản lý rủi ro và đáp ứng sự thay đổi mạnh mẽ của thị trường.

2. Mục tiêu nghiên cứu

2.1. Mục tiêu chung

Mục tiêu chung của khóa luận này là xây dựng mô hình học máy hỗ trợ quá trình dự đoán hạn mức tín dụng. Mô hình nhằm hỗ trợ các doanh nghiệp, tổ chức tài chính trong việc xác định được hạn mức tín dụng phù hợp với từng đối tượng khách hàng, từ đó quản lý rủi ro hiệu quả và tối ưu hóa quy trình cấp tín dụng.

2.2. Mục tiêu cụ thể

Nắm vững nghiệp vụ và phân tích dữ liệu hiệu quả: Hiểu rõ nghiệp vụ, xác định các yếu tố ảnh hưởng đến việc điều chỉnh hạn mức tín dụng, tiến hành phân tích dữ liệu và chọn lọc các biến quan trọng phục vụ cho việc xây dựng mô hình.

Xây dựng và tối ưu hóa mô hình: Phát triển các mô hình học máy có độ chính xác cao, tối ưu hóa tham số và so sánh hiệu suất giữa các mô hình để chọn ra giải pháp tối ưu nhất.

Ứng dụng thực tiễn: Mô hình hoạt động hiệu quả đủ để hỗ trợ các doanh nghiệp, tổ chức tài chính ra quyết định điều chỉnh hạn mức tín dụng hợp lý, đồng thời đưa ra các khuyến nghị cải thiện quy trình nghiệp vụ.

3. Đối tượng và phạm vi nghiên cứu

3.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là khách hàng sử dụng thẻ tín dụng với trọng tâm phân tích các yếu tố tác động đến hạn mức tín dụng. Nghiên cứu xem xét lịch sử thanh toán, hạn mức hiện tại, thói quen chi tiêu, đặc điểm nhân khẩu học và rủi ro tín dụng để đề xuất chiến lược quản lý hạn mức phù hợp.

3.2. Phạm vi nghiên cứu

Phạm vi nội dung: Tập trung vào việc ứng dụng các phương pháp học máy, phân tích các yếu tố ảnh hưởng đến hạn mức tín dụng dựa trên lịch sử thanh toán của khách hàng. Cụ thể, nghiên cứu xem xét các biến số như hạn mức tín dụng hiện tại, tình trạng thanh toán trong quá khứ, số dư tín dụng, số tiền thanh toán hàng tháng để xây dựng mô hình dự đoán hạn mức tín dụng phù hợp với từng đối tượng khách hàng, giúp tối ưu hóa lợi nhuận và quản lý rủi ro.

Phạm vi không gian: Dữ liệu được thu thập từ các khách hàng sử dụng thẻ tín dụng của một ngân hàng tại Việt Nam, phản ánh hành vi tài chính và khả năng thanh toán trong môi trường tín dụng.

4. Phương pháp nghiên cứu

4.1. Nghiên cứu lý thuyết

Phương pháp phân tích và tổng hợp lý thuyết: Phân tích, tổng hợp các tài liệu và nghiên cứu liên quan đến lĩnh vực tín dụng, dữ liệu lớn và học máy để xây dựng cơ sở lý luận cho nghiên cứu.

Phương pháp phân loại và hệ thống hóa lý thuyết: Tiến hành hệ thống hóa các khái niệm chính như về tín dụng, hạn mức tín dụng, đặc điểm khách hàng và các yếu tố tài chính liên quan. Phân loại các phương pháp học máy dựa trên đặc điểm dữ liệu và bài toán cần giải quyết, từ đó đề xuất hướng tiếp cận phù hợp

Phương pháp nghiên cứu tài liệu: Nghiên cứu các bài báo khoa học, báo cáo ngành và nghiên cứu thực tiễn liên quan đến việc ứng dụng học máy

trong dự đoán và tối ưu hạn mức tín dụng. Đặc biệt tập trung vào các mô hình cũng như các kỹ thuật tối ưu hóa mô hình.

4.2. Nghiên cứu thực nghiệm

Phương pháp thu thập thông tin: Dữ liệu được thu thập từ hệ thống quản lý khách hàng của một ngân hàng, bao gồm thông tin về hạn mức tín dụng, lịch sử thanh toán, số dư thẻ, thói quen chi tiêu và tình trạng tín dụng của khách hàng. Dữ liệu được tổng hợp và xử lý để đảm bảo tính chính xác, phục vụ cho việc phân tích và đề xuất chiến lược điều chỉnh hạn mức tín dụng.

Phương pháp định tính: Phân tích và diễn giải các yếu tố ảnh hưởng đến việc ra quyết định điều chỉnh hạn mức tín dụng.

Phương pháp định lượng: Sử dụng thuật toán Random Forest để xây dựng mô hình, đánh giá hiệu quả bằng các chỉ số như Root Mean Squared Error (RMSE), Mean Absolute Error (MSE), R-squared Score (R2 Score).

5. Kết cấu luận văn

Ngoài phần mở đầu, kết luận, tài liệu tham khảo và phụ lục bài khóa luận tốt nghiệp kết cấu gồm 3 chương:

- Chương 1: Cơ sở lý luận
- Chương 2: Thực nghiệm bài toán
- Chương 3: Đề xuất ứng dụng và cải tiến mô hình

CHƯƠNG 1: CƠ SỞ LÝ LUẬN

1.1. Tổng quan về tín dụng và quản lý hạn mức tín dụng

1.1.1. Tổng quan về tín dụng

a) Định nghĩa

Tín dụng là khái niệm thể hiện mối quan hệ giữa người cho vay và người vay. Tín dụng ra đời, tồn tại qua nhiều hình thái kinh tế - xã hội. Quan hệ tín dụng được phát sinh từ thời kỳ chế độ công xã nguyên thủy bắt đầu tan rã. Khi chế độ tư hữu về tư liệu sản xuất xuất hiện, đồng thời quan hệ trao đổi hàng hóa cũng xuất hiện. Thời kỳ này, tín dụng được thực hiện dưới hình thức vay mượn bằng hiện vật – hàng hóa. Xuất hiện sở hữu tư nhân tư liệu sản xuất làm cho xã hội có sự phân hóa: giàu, nghèo, người nắm quyền lực, người không có gì. Khi người nghèo gặp phải những khó khăn không thể tránh thì buộc họ phải đi vay mà những người giàu thì cấu kết với nhau để ấn định lãi suất cao, chính vì thế tín dụng nặng lãi cao ra đời. Trong giai đoạn tín dụng nặng lãi tín dụng lãi suất cao nhất là 40 – 50%, do việc sử dụng tín dụng nặng lãi không phục vụ việc sản xuất mà chỉ phục vụ cho mục đích tín dụng nên nền kinh tế bị kìm hãm động lực phát triển. Về sau, tín dụng đã chuyển sang hình thức vay mượn bằng tiền tệ.

Tín dụng là việc một bên (bên cho vay) cung cấp nguồn tài chính cho đối tượng khác (bên đi vay) trong đó bên đi vay sẽ hoàn trả tài chính cho bên vay trong một thời hạn thỏa thuận và thường kèm theo lãi suất. Do đó, tín dụng phản ánh mối quan hệ giữa hai bên, một bên là người cho vay và một bên là người đi vay. Quan hệ giữa hai bên ràng buộc bởi cơ chế tín dụng, thỏa thuận thời gian cho vay, lãi suất phải trả. Thực chất, tín dụng là biểu hiện mối quan hệ kinh tế gắn liền với quá trình tạo lập và sử dụng quỹ tín dụng nhằm mục đích thỏa mãn nhu cầu vốn tạm thời cho quá trình sản xuất và đời sống theo nguyên tắc hoàn trả.

b) Đặc điểm

Dựa trên sự tin tưởng: Người cho vay chỉ cấp tín dụng khi có lòng tin vào việc người vay sử dụng vốn vay đúng mục đích, hiệu quả và có khả năng hoàn trả đúng hạn.

Có tính tạm thời: Việc cho vay chỉ là nhường quyền sử dụng tạm thời một lượng vốn trong một thời hạn nhất định.

Có tính hoàn trả cả gốc lẫn lãi: Đến thời hạn người vay có nghĩa vụ và trách nhiệm phải hoàn trả cả vốn gốc và lãi vô điều kiện.

c) Vai trò

Thúc đẩy phân bổ nguồn vốn: Góp phần tăng lượng vốn đầu tư và hiệu quả đầu tư thông qua việc luân chuyển nguồn vốn tạm thời nhàn rỗi giữa các cá nhân, hộ gia đình, doanh nghiệp và Chính phủ đến những người đang cần vốn, đồng thời đòi hỏi người đi vay phải nỗ lực sử dụng vốn hiệu quả.

Công cụ điều tiết kinh tế vĩ mô: Nhà nước sử dụng các chính sách tín dụng để kiểm soát lạm phát, ổn định thị trường và thúc đẩy các mục tiêu kinh tế - xã hội.

Nâng cao đời sống cá nhân và cộng đồng: Thông qua các hình thức tín dụng tiêu dùng cá nhân và hộ gia đình có thể tiếp cận nguồn vốn để đáp ứng nhu cầu tiêu dùng, cải thiện chất lượng cuộc sống và thúc đẩy sự phát triển của cộng đồng.

d) Các hình thức tín dụng phổ biến

Tín dụng thương mại: Là quan hệ tín dụng giữa các doanh nghiệp dưới hình thức mua bán chịu hàng hóa. Đây là quan hệ tín dụng giữa các nhà sản xuất – kinh doanh được thể hiện dưới hình thức mua bán, bán chịu hàng hóa. Hành vi mua bán chịu hàng hóa được xem là hình thức tín dụng người bán chuyển giao cho người mua quyền sử dụng vốn tạm thời trong một thời gian nhất định và khi đến thời hạn đã thỏa thuận, người mua phải hoàn lại vốn cho người bán dưới hình thức tiền tệ và cả phần lãi cho người bán chịu.

Đặc điểm của tín dụng thương mại:

- Tín dụng thương mại vốn cho vay dưới dạng hàng hóa hay một bộ phận của vốn sản xuất chuẩn bị chuyển hóa thành tiền, chưa phải là tiền nhàn rỗi.
- Người cho vay và người đi vay đều là những doanh nghiệp trực tiếp tham gia vào quá trình sản xuất và lưu thông hàng hóa.
- Khối lượng tín dụng lớn hay nhỏ phụ thuộc vào tổng giá trị của khối lượng hàng hóa được đưa ra mua bán chịu.

Tín dụng ngân hàng: Là giao dịch tài sản giữa ngân hàng với bên đi vay là các tổ chức kinh tế, cá nhân trong nền kinh tế. Trong đó ngân hàng chuyển giao tài sản cho bên đi vay sử dụng trong một thời gian nhất định theo thỏa thuận và bên đi vay có trách nhiệm hoàn trả vô điều kiện cả gốc và lãi cho ngân hàng khi đến hạn thanh toán.

Đặc điểm của tín dụng ngân hàng:

- Phù hợp với nhiều đối tượng, đáp ứng nhu cầu của hầu hết đối tượng khách hàng.
- Cho vay bằng tiền: Vì ngân hàng chủ yếu huy động vốn từ nhiều nguồn trên thị trường nên thường cung cấp hình thức vay tiền phổ biến.
- Linh hoạt thời gian cho vay: Hình thức tín dụng ngân hàng cho phép khách hàng vay ngắn hạn, trung hạn và dài hạn. Do đó, ngân hàng linh hoạt điều chỉnh nguồn vốn đáp ứng nhu cầu thời hạn vay của từng khách hàng.
- Đáp ứng nhu cầu về vốn một cách tối đa: Nguồn vốn ngân hàng huy động từ nhiều nguồn có thể đáp ứng tối đa nhu cầu tài chính của khách hàng.

Tín dụng Nhà nước: Là quan hệ tín dụng giữa Nhà nước với doanh nghiệp, các tổ chức kinh tế - xã hội và các cá nhân. Tín dụng Nhà nước xuất hiện nhằm thỏa mãn những nhu cầu chi tiêu của ngân sách Nhà nước trong điều kiện nguồn thu không đủ để đáp ứng, nó còn là công cụ để Nhà nước hỗ trợ cho

các ngành kinh tế yếu kém, ngành mũi nhọn, khu vực kinh tế kém phát triển và là công cụ quan trọng để Nhà nước quản lý, điều hành vĩ mô.

Đặc điểm của tín dụng Nhà nước:

- Chủ thể là Nhà nước, các pháp nhân và thể nhân.
- Hình thức đa dạng, phong phú.
- Tín dụng Nhà nước chủ yếu là loại hình trực tiếp không thông qua tổ chức trung gian.

Tín dụng tiêu dùng: Là quan hệ tín dụng giữa dân cư với doanh nghiệp, ngân hàng và các công ty cho thuê tài chính.

Đặc điểm của tín dụng tiêu dùng:

- Đáp ứng nhu cầu tiêu dùng cho dân cư.
- Hình thức là hàng hóa hoặc tiền tệ.
- Dân cư là người vay; ngân hàng, công ty cho thuê tài chính và doanh nghiệp là người cho vay.

1.1.2. Tổng quan về quản lý hạn mức tín dụng

a) Định nghĩa

Hạn mức tín dụng là số tiền tối đa mà ngân hàng hoặc tổ chức tài chính cho phép khách hàng vay hoặc chi tiêu trong một khoảng thời gian nhất định, nhằm hạn chế mức dư nợ tín dụng tối đa đến với nền kinh tế. Hạn mức tín dụng có thể được điều chỉnh theo yêu cầu và thỏa thuận giữa người dùng và nhà cung cấp thẻ.

Ví dụ một doanh nghiệp được ngân hàng cấp cho hạn mức tín dụng là 400 triệu đồng/tháng, nghĩa là trong một tháng số tiền tối đa mà doanh nghiệp được vay là 400 triệu đồng. Trường hợp doanh nghiệp vay và thanh toán 200 triệu đồng ngay trong tháng thì cũng chỉ được vay thêm 200 triệu đồng do số dư khoản vay cuối tháng của doanh nghiệp không được vượt quá hạn mức 400 triệu đồng.

Theo Điều 1 Quyết định 43/QĐ-NH14 hạn mức tín dụng là công cụ quan trọng trong việc thực hiện chính sách tiền tệ nhằm kiểm soát tổng mức dư nợ

tín dụng trong nền kinh tế đối với các tổ chức tín dụng. Việc quản lý hạn mức tín dụng đóng vai trò then chốt trong công tác quản lý rủi ro tín dụng, góp phần bảo vệ sự ổn định và phát triển lâu dài của nền kinh tế, được thể hiện qua các khía cạnh sau:

Quản lý hạn mức tín dụng giúp ngân hàng kiểm soát rủi ro tín dụng bằng cách hạn chế sự tập trung tín dụng vào một nhóm khách hàng hoặc ngành nghề cụ thể, từ đó nâng cao khả năng dự báo và giảm thiểu nguy cơ vỡ nợ. Tổ chức Hợp tác và Phát triển Kinh tế (OECD, 2018) cho rằng việc quản lý hạn mức tín dụng hiệu quả giúp giảm thiểu rủi ro tín dụng và nâng cao hiệu quả sử dụng vốn. Khi hạn mức tín dụng được thiết lập phù hợp với khả năng tài chính của khách hàng ngân hàng có thể giảm thiểu rủi ro vỡ nợ và các thiệt hại tài chính liên quan.

Nâng cao hiệu quả hoạt động của ngân hàng, tối ưu hóa dòng tiền, cân bằng giữa tăng trưởng tín dụng và duy trì thanh khoản giúp bảo đảm dòng tiền ổn định cho hoạt động và quản lý hiệu quả tài sản nợ, tài sản có từ đó tăng cường tỷ suất lợi suất lợi nhuận trên tài sản và vốn chủ sở hữu.

Tăng cường khả năng cạnh tranh, đáp ứng các yêu cầu pháp lý và tiêu chuẩn quốc tế. Giúp ngân hàng xây dựng uy tín, phát triển quan hệ khách hàng bền vững và tăng khả năng cạnh tranh trên thị trường

b) Thực trạng quản lý hạn mức tín dụng tại các Ngân hàng thương mại

- Quản lý hạn mức tín dụng dựa trên mức độ tăng trưởng tín dụng

Trong giai đoạn 2015 – 2024, tỷ lệ tăng trưởng tín dụng của ngành ngân hàng Việt Nam đã có sự biến động rõ rệt phản ánh những thay đổi trong nền kinh tế và nhu cầu tín dụng của các doanh nghiệp cũng như cá nhân. Trong ba năm đầu (2015 – 2017), tỷ lệ tăng trưởng tín dụng đạt mức khá cao dao động từ 17 – 18%. Đây là giai đoạn nền kinh tế Việt Nam phục hồi mạnh mẽ sau khủng hoảng kinh tế toàn cầu năm 2008, kéo theo nhu cầu tín dụng từ các doanh nghiệp và cá nhân tăng cao đã tạo điều kiện thuận lợi cho các ngân hàng thương mại trong việc cấp tín dụng. Tuy nhiên, sự tăng trưởng tín dụng quá mức cũng

tiềm ẩn nhiều nguy cơ rủi ro tín dụng, đặc biệt là khi các khoản vay chưa được kiểm soát chặt chẽ.

Từ năm 2018 – 2020, tỷ lệ tăng trưởng tín dụng bắt đầu giảm nhẹ xuống dưới mức 14%, chủ yếu do tác động của các yếu tố kinh tế vĩ mô đặc biệt là ảnh hưởng của đại dịch Covid-19 khiến các ngân hàng trở nên thận trọng hơn trong việc cấp tín dụng. Trong giai đoạn này các ngân hàng đã chú trọng hơn vào việc ứng dụng công nghệ để kiểm soát tín dụng và hạn mức tín dụng nhằm giảm thiểu rủi ro. Một số công cụ phân tích tín dụng tự động, phần mềm đánh giá tín dụng và các hệ thống dự báo rủi ro đã được triển khai để duy trì sự ổn định trong mức tăng trưởng tín dụng.

Từ năm 2021 – 2024, tỷ lệ tăng trưởng tín dụng có sự phục hồi nhẹ đạt 15.08% vào năm 2024. Tuy nhiên, mức tăng này vẫn thấp hơn so với giai đoạn 2015 – 2017, chủ yếu do các ngân hàng đã thực hiện các biện pháp kiểm soát tín dụng chặt chẽ hơn để giảm thiểu rủi ro. Trong giai đoạn này các ngân hàng đã ứng dụng phần mềm phân tích dữ liệu lớn và trí tuệ nhân tạo (AI) để nâng cao quy trình phê duyệt tín dụng, xác định mức tín dụng phù hợp cho từng khách hàng và phân tích các yếu tố rủi ro.

- Quản lý hạn mức tín dụng dựa trên mức độ phân bổ dư nợ tín dụng

Trong bối cảnh hội nhập và chuyển đổi số việc phân bổ dư nợ tín dụng theo nhóm khách hàng giúp các ngân hàng tối ưu hóa chiến lược tín dụng và giảm thiểu rủi ro. Tuy vào đối tượng khách hàng và theo loại hình doanh nghiệp có thể phân loại khách hàng thành các nhóm như: Khách hàng doanh nghiệp lớn, khách hàng doanh nghiệp nhỏ và vừa, khách hàng cá nhân và các nhóm khác.

Doanh nghiệp lớn chiếm 13,52% tại ngân hàng thương mại Nhà nước và 0,97% tại ngân hàng thương mại cổ phần lớn, điều này phản ánh chiến lược tập trung hỗ trợ các dự án quốc gia. Tuy nhiên nhóm này tiềm ẩn rủi ro tín dụng lớn và đòi hỏi ngân hàng phải áp dụng các công nghệ như AI để đánh giá và theo dõi từ đó kiểm soát hạn mức tín dụng hiệu quả hơn.

Doanh nghiệp nhỏ và vừa chiếm 41,37% tại ngân hàng thương mại Nhà nước và 36,57% tại ngân hàng thương mại cổ phần lớn. Những doanh nghiệp này có năng lực tài chính hạn chế và dễ bị tác động bởi những biến động kinh tế, do đó các ngân hàng thường triển khai chương trình tín dụng ưu đãi và chính sách hỗ trợ riêng biệt nhằm đảm bảo hạn mức tín dụng hợp lý và giảm thiểu rủi ro nợ xấu.

Khách hàng cá nhân chiếm tỷ trọng lớn nhất với 36,57% tại ngân hàng thương mại Nhà nước và 42,53% tại ngân hàng thương mại cổ phần lớn. Các khoản vay cá nhân thường được sử dụng cho mục đích tiêu dùng, mua nhà hoặc kinh doanh nhỏ. Để quản lý hạn mức tín dụng hiệu quả các ngân hàng áp dụng công cụ hiện đại như phân tích hành vi tài chính cá nhân và hệ thống chấm điểm tín dụng tự động, giúp xác định hạn mức phù hợp với nhu cầu và khả năng trả nợ của từng khách hàng.

Nhóm khách hàng khách còn lại bao gồm các hợp tác xã, tổ chức kinh tế khác chiếm tỷ trọng nhỏ nhất với 11,6% tại nhóm ngân hàng thương mại Nhà nước và 0,21% tại nhóm ngân hàng thương mại cổ phần. Mặc dù chiếm tỷ trọng thấp nhưng đây là nhóm khách hàng đặc thù đòi hỏi các ngân hàng phải xây dựng quy trình quản lý tín dụng hợp lý, từ việc xác định hạn mức đến kiểm soát dòng tiền và hiệu quả sử dụng vốn.

- Quản lý hạn mức tín dụng thông quan chính sách quản lý hạn mức tín dụng

Chính sách hạn mức tín dụng của các ngân hàng thương mại Việt Nam không chỉ phụ thuộc vào quy định nội bộ của từng ngân hàng mà còn chịu sự điều chỉnh từ Ngân hàng Nhà nước. Ngân hàng Nhà nước đã ban hành nhiều chính sách liên quan đến hạn mức tín dụng, đặc biệt là các quy định về cho vay và mức tín dụng đối với các ngành nghề khác nhau. Bên cạnh việc phát hành các thông tư, Ngân hàng Nhà nước còn thực hiện các biện pháp giám sát chặt chẽ hoạt động cấp tín dụng của các ngân hàng và điều chỉnh dư nợ tín dụng để đảm bảo không vượt quá mức an toàn. Dựa trên tình hình kinh tế vĩ mô và thực

tế, Ngân hàng Nhà nước cũng đưa ra các khuyến cáo về việc kiểm soát tỷ lệ tín dụng của các ngân hàng nhằm hạn chế rủi ro tín dụng và bảo hệ thống tài chính khỏi những tác động tiêu cực do việc cấp tín dụng quá mức hoặc cho vay vào các ngành nghề tiềm ẩn nhiều rủi ro.

c) Các yếu tố ảnh hưởng đến hạn mức tín dụng

- Yếu tố chủ quan:

Mức thu nhập hàng tháng: Đây là một yếu tố chính và quan trọng nhất để ngân hàng xem xét hạn mức tín dụng của khách hàng. Một khách hàng có thu nhập cao và ổn định sẽ có nhiều cơ hội được cấp hạn mức tín dụng cao hơn so với những người có thu nhập không ổn định.

Lịch sử tín dụng: Lịch sử tín dụng đóng vai trò quan trọng trong việc xác định hạn mức tín dụng của khách hàng. Nếu một khách hàng có lịch sử vay và trả nợ đúng hạn trong nhiều năm ngân hàng có thể coi là khách hàng đáng tin cậy. Ngược lại nếu khách hàng có lịch sử thanh toán trễ hạn hoặc từng có khoản vay bị quá hạn nghiêm trọng, ngân hàng sẽ áp đặt mức tín dụng thấp hoặc thậm chí từ chối cấp tín dụng. Các ngân hàng thường xem xét ít nhất 12 – 24 tháng dữ liệu tín dụng gần nhất để đưa ra quyết định về hạn mức tín dụng.

Dựa trên hạn mức các thẻ tín dụng cùng một chủ sở hữu: Ngân hàng mở thẻ tín dụng sẽ căn cứ vào lịch sử tín dụng của các thẻ đã mở. Nếu uy tín và không có khoản nợ quá hạn nào thì hồ sơ của khách hàng sẽ được phê duyệt một cách dễ dàng.

Mục đích vay: Ngân hàng và các tổ chức tín dụng có chính sách ưu đãi dành cho các khoản vay của một số ngành nghề. Ngoài ra, các khoản vay tiêu dùng thường có hạn mức thấp hơn các khoản vay kinh doanh, vay mua nhà, mua đất.

Ngành nghề và công việc của khách hàng: Những người làm trong các ngành có thu nhập ổn định thường được cấp hạn mức tín dụng cao hơn so với những người làm nghề tự do hoặc trong các ngành nghề không có thu nhập ổn

định, nguyên nhân là do ngân hàng sẽ đánh giá mức độ rủi ro về khả năng trả nợ của từng ngành nghề.

Điểm tín dụng: Đây là một thước đo tổng hợp về mức độ tin cậy tài chính của khách hàng, thường được tính toán dựa trên lịch sử tín dụng, tần suất thanh toán đúng hạn, tỷ lệ nợ và thời gian sử dụng tín dụng. Tại Việt Nam, Trung tâm Thông tin Tín dụng Quốc gia (CIC) cung cấp điểm tín dụng cho khách hàng cá nhân và doanh nghiệp dựa trên dữ liệu từ các tổ chức tài chính. Điểm tín dụng càng cao khách hàng càng có cơ hội được cấp hạn mức tín dụng cao hơn. Ví dụ, một khách hàng có điểm tín dụng từ 750 trở lên có thể nhận được hạn mức tín dụng cao hơn so với khách hàng có điểm dưới 600.

- Yếu tố khách quan:

Chính sách tín dụng của ngân hàng: Mỗi ngân hàng có chính sách tín dụng riêng dựa vào khẩu vị rủi ro và chiến lược kinh doanh. Một số ngân hàng chấp nhận rủi ro cao hơn để mở rộng thị phần, trong khi những ngân hàng khác thường sẽ có xu hướng hạn chế rủi ro hơn. Ví dụ, Vietcombank có thể áp dụng chính sách kiểm soát rủi ro chặt chẽ, hạn chế cấp tín dụng cho khách hàng có điểm tín dụng thấp. Trong khi đó các ngân hàng bán lẻ như TPBank, VPBank có thể linh hoạt hơn trong việc cấp hạn mức tín dụng nhằm thu hút thêm nhiều khách hàng.

Lãi suất thị trường: Nếu lãi suất thị trường tăng cao ngân hàng thường sẽ có xu hướng siết chặt hạn mức tín dụng để giảm thiểu rủi ro. Điều này là do chi phí vốn tăng lên khiến ngân hàng thận trọng hơn trong việc cấp tín dụng. Ngược lại, khi lãi suất thấp ngân hàng có thể mở rộng hạn mức tín dụng để khuyến khích khách hàng vay nhiều hơn.

Tỷ lệ nợ xấu: Nếu tỷ lệ nợ xấu trong hệ thống ngân hàng tăng cao thì ngân hàng sẽ có xu hướng hạn chế cấp hạn mức tín dụng cao. Ngược lại nếu tỷ lệ nợ xấu ở mức thấp ngân hàng thường sẽ linh hoạt hơn trong việc mở rộng hạn mức tín dụng.

Các quy định pháp lý: Ngân hàng Nhà nước Việt Nam có thể áp đặt giới hạn tín dụng theo từng thời kỳ để kiểm soát lạm phát và tăng trưởng kinh tế. Khi ngân hàng bị giới hạn về hạn mức tín dụng họ sẽ phải điều chỉnh cách phân bổ hạn mức tín dụng cho từng nhóm khách hàng.

1.2. Học máy và ứng dụng trong quản lý hạn mức tín dụng

1.2.1. Giới thiệu về học máy

Học máy (Machine Learning) là một nhánh của trí tuệ nhân tạo (AI), là quá trình máy tính sử dụng các thuật toán để học từ dữ liệu mà không cần phải lập trình chi tiết từng bước. Mục tiêu chính của học máy là xây dựng các mô hình có khả năng nhận biết mẫu trong dữ liệu và sử dụng chúng để dự đoán các kết quả mới. Học máy được chia thành ba nhóm chính: học có giám sát (supervised learning), học không giám sát (unsupervised learning) và học tăng cường (reinforcement learning).

Học có giám sát: Đây là phương pháp phổ biến nhất, trong đó dữ liệu huấn luyện bao gồm cả đầu vào và đầu ra mong muốn. Các mô hình học có giám sát thường được sử dụng để giải quyết các bài toán như phân loại (classification) và hồi quy (regression). Ví dụ, dự đoán giá nhà dựa trên thông tin về vị trí và diện tích là một bài toán hồi quy.

Học không giám sát: Phương pháp này không yêu cầu đầu ra mong muốn mà chỉ phân tích cấu trúc của dữ liệu để tìm ra các mẫu hoặc các cụm. Phân khúc khách hàng dựa trên hành vi tiêu dùng là một ứng dụng điển hình của bài toán học không giám sát.

Học tăng cường: Đây là phương pháp trong đó mô hình học qua tương tác với môi trường và nhận phần thưởng hoặc hình phạt. Học tăng cường thường được áp dụng trong robot tự động hoặc các trò chơi.

1.2.2. Ứng dụng của học máy trong quản lý hạn mức tín dụng

a) Vai trò của học máy trong quản lý hạn mức tín dụng

Học máy đang ngày càng trở nên phổ biến và được ứng dụng nhiều trong quản lý hạn mức tín dụng giúp các ngân hàng và tổ chức tín dụng tối ưu hóa

việc cấp tín dụng và điều chỉnh hạn mức một cách chính xác hơn. Các mô hình học máy như Random Forest, XGBoost hay Neural Netwworks được sử dụng để phân tích các yếu tố như thu nhập, lịch sử tín dụng và hành vi chi tiêu của khách hàng từ đó ra quyết định về hạn mức tín dụng phù hợp. Các ngân hàng lớn như JPMorgan Chase, Bank of America đã áp dụng các thuật toán học máy để phân tích dữ liệu tín dụng giúp xác định và điều chỉnh hạn mức tín dụng phù hợp cho từng khách hàng.

Tại Việt Nam, trong giai đoạn trước đây nhiều ngân hàng thương mại chủ yếu áp dụng các quy trình quản lý tín dụng thủ công như tập trung vào việc thẩm định hồ sơ, phân tích tài chính và phê duyệt tín dụng qua nhiều cấp độ quản lý. Quy trình này không chỉ kéo dài thời gian xử lý mà còn phụ thuộc nhiều vào kinh nghiệm của cán bộ tín dụng. Hiện nay, ngân hàng thương mại đã áp dụng các công cụ hiện đại để phân tích dữ liệu hay dự đoán rủi ro. Chẳng hạn, ngân hàng thương mại cổ phần Kỹ thương Việt Nam (Techcombank) đã ứng dụng hệ thống quản lý tài sản bảo đảm và hạn mức tín dụng (CLIMS) cho phép quản lý toàn diện các quan hệ tín dụng, kiểm soát giới hạn hạn mức tín dụng theo quy định và hỗ trợ đánh giá rủi ro hiệu quả. Tại ngân hàng thương mại cổ phần Ngoại thương Việt Nam (Vietcombank), hệ thống CLIMS được triển khai để quản lý hiệu quả các khoản vay, theo dõi chi tiết hạn mức tín dụng, kiểm soát rủi ro và đảm bảo tính minh bạch trong quá trình xử lý.

Việc ứng dụng các công nghệ tiên tiến như dữ liệu lớn, AI và đặc biệt là học máy đã hỗ trợ ngân hàng trong việc tự động hóa quy trình xét duyệt tín dụng, phân tích hành vi tài chính, lịch sử tín dụng và dòng tiền khách hàng. Hiện nay, ngoài hệ thống chấm điểm tín dụng của Trung tâm Thông tin tín dụng Quốc gia Việt Nam (CIC) các ngân hàng như Vietcombank đã sử dụng các thuật toán để phân tích hành vi tài chính, lịch sử tín dụng và dòng tiền của khách hàng từ đó đưa ra quyết định cấp hạn mức tín dụng phù hợp. BIDV triển khai hệ thống phân tích tín dụng dựa trên dữ liệu lớn giúp tăng khả năng phát hiện rủi ro tín dụng lên đến 20% so với phương pháp truyền thống. Việc áp dụng

học máy không chỉ rút ngắn thời gian xử lý từ vài ngày xuống vài giờ mà còn cải thiện độ chính xác trong dự báo rủi ro và ra quyết định điều chỉnh hạn mức tín dụng.

b) Thách thức khi ứng dụng học máy trong quản lý hạn mức tín dụng

Ngân hàng có các phân khúc khách hàng đa dạng từ cá nhân, doanh nghiệp, tổ chức đến các định chế tài chính với những đặc điểm, nhu cầu và hành vi tín dụng khác nhau. Do đó, việc áp dụng công nghệ nói chung và học máy nói riêng trong dự báo hạn mức tín dụng đòi hỏi phải bao quát được tất cả phân khúc khách hàng, hệ thống thông tin của các ngân hàng phải có khả năng xử lý khối lượng dữ liệu lớn và phức tạp. Điều này gây áp lực lên năng lực công nghệ của ngân hàng, đòi hỏi phải có sự đầu tư vào hệ thống máy chủ, cơ sở dữ liệu và các công cụ phân tích.

Ứng dụng học máy trong dự báo và quản lý hạn mức tín dụng đòi hỏi những khoản đầu tư lớn cho việc triển khai hạ tầng công nghệ như máy chủ, phần mềm phân tích hay các công nghệ hiện đại. Ngoài ra, chi phí bảo trì, nâng cấp hệ thống cũng như đào tạo nhân sự cũng là một gánh nặng đáng kể cho các ngân hàng, đặc biệt là các ngân hàng có quy mô vừa và nhỏ khi năng lực tài chính còn hạn chế.

Hạ tầng công nghệ của các ngân hàng chưa đủ mạnh để xử lý khối lượng dữ liệu ngày càng tăng cao dẫn đến tốc độ xử lý giao dịch chậm, chưa hoàn toàn bảo đảm an toàn bảo mật. Hiện nay các ngân hàng chủ yếu sử dụng nhiều hệ thống để quản lý hạn mức tín dụng, tuy thuộc vào các nhóm sản phẩm và nhóm khách hàng. Việc sử dụng các hệ thống học máy dẫn đến khó khăn trong việc tích hợp hệ thống mới vào hệ thống cũ, các hệ thống được tích hợp hoạt động thiếu liền mạch, giảm hiệu quả của các công cụ quản lý hạn mức tín dụng hoặc xảy ra xung đột về tín dụng.

Việc triển khai các thuật toán học máy để dự báo và quản lý hạn mức tín dụng đòi hỏi đội ngũ nhân sự phải có kiến thức chuyên môn sâu về công nghệ thông tin, phân tích dữ liệu và quản trị rủi ro. Tuy nhiên, không phải tất cả các

ngân hàng đều có đội ngũ nhân sự đáp ứng các yêu cầu trên. Hiện nay, một số ngân hàng thương phải tuyển bổ sung các nhân sự có kinh nghiệm và chuyên môn phù hợp, đặc biệt là trong lĩnh vực công nghệ thông tin và phân tích dữ liệu để đảm bảo chất lượng triển khai hiệu quả.

1.3. Mô hình học máy sử dụng trong bài toán

1.3.1. Mô hình Random Forest

Random Forest là một thuật toán học máy thuộc nhóm ensemble learning, trong dựa trên việc kết hợp nhiều cây quyết định (decision tree) để tạo ra một mô hình mạnh mẽ hơn. Thuật toán này sử dụng phương pháp bagging để huấn luyện một tập hợp các cây quyết định trên các phân mẫu ngẫu nhiên của dữ liệu, sau đó kết hợp các kết quả dự đoán của các cây này để đưa ra quyết định cuối cùng.

Quy trình hoạt động:

1. Bagging (Bootstrap Aggregating)
   - Dữ liệu gốc được lấy mẫu ngẫu nhiên để tạo ra các tập con (bootstrap samples).
   - Mỗi tập con được sử dụng để huấn luyện một cây quyết định riêng biệt, giúp giảm thiểu độ phức tạp và cải thiện độ chính xác của mô hình.
2. Cấu trúc cây quyết định
   - Mỗi cây quyết định trong Random Forest được huấn luyện bằng cách chia nhỏ dữ liệu dựa trên các đặc trưng tối ưu.
   - Chỉ sử dụng một tập hợp ngẫu nhiên các đặc trưng tại mỗi nút để giảm sự tương quan giữa các cây.
3. Tổng hợp dự đoán
   - Đối với bài toán hồi quy, kết quả là giá trị trung bình của các dự đoán từ tất cả các cây.
   - Đối với bài toán phân loại, kết quả được lấy dựa trên số phiếu (majority voting).

Lý do sử dụng XGBoost trong bài toán dự đoán vỡ nợ:

- Khả năng chống overfitting tốt: Random Forest sử dụng phương pháp bagging và kết hợp nhiều cây quyết định để giảm thiểu overfititng và tăng tính ổn định của mô hình.
- Khả năng xử lý dữ liệu phức tạp: Với khả năng học các mối quan hệ phi tuyến tính trong dữ liệu, mô hình có thể áp dụng cho các bài toán phức tạp mà các mô hình học máy khác gặp khó khăn.
- Hiệu suất ổn định: Mặc dù Random Forets đòi hỏi thời gian huấn luyện lâu nhưng mô hình này vẫn cho kết quả ổn định và chính xác trong hầu hết các tình huống.

Các chỉ số đánh giá hiệu quả mô hình: Hiệu suất của mô hình được đo lường bằng các chỉ số sau:

- RMSE (Root Mean Square Error): Sai số căn bậc hai trung bình, đo lường sai số trung bình giữa giá trị thực tế và giá trị dự đoán.
- MAE (Mean Absolute Error): Sai số tuyệt đối trung bình, phản ánh mức độ sai lệch thực tế của mô hình.
- R^2 (Conefficient of Determination): Đánh giá tỷ lệ biến thiên của dữ liệu đầu ra được giải thích bởi mô hình.

1.3.2. Mô hình XGBoost

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy dựa trên phương pháp boosting, được tối ưu hóa để dự đoán các giá trị liên tục (regression tasks).

Quy trình hoạt động:

1. Gradient Boosting Framework

- Bắt đầu với mô hình đơn giản như dự đoán giá trị trung bình.
- Các mô hình tiếp theo được xây dựng dựa trên gradient của sai số (loss) từ mô hình trước, nhằm cải thiện kết quả dự đoán và tối ưu hóa mô hình.

2. Regularization (Điều chỉnh)

- Tích hợp L1 (Lasso) và L2 (Ridge) để giảm thiểu overfitting và giúp mô hình tổng quát tốt hơn khi làm việc với dự liệu lớn.

3. Tối ưu hóa song song: XGBoost sử dụng cấu trúc cây nhị phân tối ưu và hỗ trợ GPU để tăng tốc độ quá trình huấn luyện, giúp tiết kiệm thời gian và tài nguyên.

Lý do sử dụng XGBoost:

- Hiệu năng cao: XGBoost rất mạnh trong việc xử lý các bài toán hồi quy với khả năng tối ưu hóa siêu tham số và tăng tốc nhờ hỗ trợ GPU.
- Khả năng xử lý phi tuyến tính: Mô hình học tốt các mối quan hệ phi tuyến tính giữa các đặc trưng.
- Hạn chế overfitting: XGBoost tíhc hợp các kỹ thuật regularization (L1, L2) giúp giảm nguy cơ overfititng khi làm việc với dữ liệu lớn.

Các chỉ số đánh giá hiệu quả mô hình: Hiệu suất của mô hình được đo lường bằng các chỉ số sau:

- RMSE (Root Mean Square Error): Sai số căn bậc hai trung bình, đo lường sai số trung bình giữa giá trị thực tế và giá trị dự đoán.
- MAE (Mean Absolute Error): Sai số tuyệt đối trung bình, phản ánh mức độ sai lệch thực tế của mô hình.
- R^2 (Conefficient of Determination): Đánh giá tỷ lệ biến thiên của dữ liệu đầu ra được giải thích bởi mô hình.

1.3.3. Mô hình SVR (Support Vector Regression)

Support Vector Regression là thuật toán học máy dùng để giải quyết các bài toán hồi quy. SVR là biến thể của SVM, được thiết kế để dự đoán giá trị liên tục thay vì phân loại. SVR tìm mặt phảng hồi quy sao cho sai số giữa giá trị dự đoán và thực tế nằm trong phạm vi cho phép (epsilon tube).

Quy trình hoạt động:

1. Tạo mặt phăng hồi quy:

- SVR cố gắng tìm ra một mặt phẳng hồi quy sao cho sai số giữa giá trị dự đoán và thực tế được giữ trong phạm vi epsilon. Mô hình sẽ cố gắng không làm tăng sai số quá mức.

2. Xử lý dữ liệu phi tuyến tính
   - SVR sử dụng kernel trick để xử lý các mối quan hệ phi tuyến tính giữa các đặc trưng đầu vào và giá trị dự đoán.
   - Việc sử dụng kernel giúp mô hình có thể học các mối quan hệ phức tạp mà không cần phải chuyển dữ liệu sang dạng phi tuyến tính.
3. Tối ưu hóa và phạt
   - SVR sử dụng phương pháp tối ưu hóa để tìm ra một hàm hồi quy tối ưu với sai số nhỏ nhất. Các điểm nằm ngoài epsilon tube sẽ bị phạt và được đưa vào quá trình tối ưu hóa để giảm thiểu sai số tổng thể

Lý do sử dụng mô hình SVR:

- Khả năng xử lý dữ liệu phi tuyến tính: SVR rất mạnh trong việc mô hình hóa các mối quan hệ phi tuyến tính giữa các đặc trưng và gia strij mục tiêu, đặc biệt khi dữ liệu không tuân theo các mối quan hệ tuyên stisnh đơn giản.
- Hiệu quả với dữ liệu nhiễu: SVR hoạt động tốt với ngay cả khi số lượng mẫu không lớn và có khả năng xử lý dữ liệu nhiễu tốt hơn một số thuật toán hồi quy khác.
- Chống overfitting: SVR có khả năng điều chỉnh sai số cho phép giúp giảm thiểu overfitting và chỉ phạt các điểm dữ liệu quan trọng vượt ra ngoài phạm vi epsilon.

Các chỉ số đánh giá hiệu quả mô hình: Hiệu suất của mô hình được đo lường bằng các chỉ số sau:

- RMSE (Root Mean Square Error): Sai số căn bậc hai trung bình, đo lường sai số trung bình giữa giá trị thực tế và giá trị dự đoán.

- MAE (Mean Absolute Error): Sai số tuyệt đối trung bình, phản ánh mức độ sai lệch thực tế của mô hình.
- R^2 (Conefficient of Determination): Đánh giá tỷ lệ biến thiên của dữ liệu đầu ra được giải thích bởi mô hình.

1.4. Giới thiệu chung về thư viện streamlit

Streamlit là một framework mã nguồn mở được thiết kế đặc biệt cho Python giúp các nhà khoa học dữ liệu, kỹ sư học máy và lập trình viên tạo ra các ứng dụng web tương tác với người dùng. Mục tiêu của streamlit là đơn giản hóa quy trình xây dựng giao diện người dùng cho các mô hình và dữ liệu phân tích.

1.4.1. Tính năng nổi bật của streamlit

Đơn giản và dễ sử dụng: Streamlit cho phép xây dựng web, tạo widget đơn giản như thanh trượt, biểu đồ, nút nhấn và các biểu mẫu tương tác một cách dễ dàng

Cập nhật trực tiếp: Khi mã Python thay đổi streamlit có thể tự động cập nhật mà không cần phải làm lại từ đầu

Tích hợp với các thư viện: Streamlit hỗ trợ tích hợp với nhiều thư viện phổ biến như Matplotlib, pandas, numpy, scikit learn giúp hiển thị dữ liệu và đồ họa trực quan hơn.

Cấu hình dễ dàng: Tạo các ứng dụng web với cấu hình thấp, chỉ cần nhập st.write() hoặc st.plot_chart() để hiển thị dữ liệu.

1.4.2. Ứng dụng của streamlit

Ứng dụng phân tích dữ liệu: Tạo các bảng điều khiển phân tích dữ liệu nhanh chóng, trực quan chẳng hạn như các ứng dụng phân tích tài chính, phân tích dữ liệu và các báo cáo dữ liệu.

Ứng dụng học máy: Streamlit phù hợp để triển khai mô hình học máy và cung cấp giao diện người dùng để kiểm tra và đánh giá các mô hình AI.

Ứng dụng chia sẻ: Streamlit cho phép người dùng dễ dàng chia sẻ kết quả các mô hình học máy, phân tích dữ liệu bằng một ứng dụng streamlit mà không cần thiết lập máy chủ phức tạp.

TIỂU KẾT CHƯƠNG 1

Chương 1 của khóa luận tập trung vào việc xây dựng các cơ sở lý luận về tín dụng và quản lý hạn mức tín dụng, làm rõ các khái niệm, vai trò và hình thức tín dụng phổ biến trong hệ thống tài chính. Phân tích thực trạng quản lý hạn mức tín dụng tại các ngân hàng thương mại Việt Nam, chỉ ra các yếu tố tác động đến hạn mức tín dụng và các thách thức trong việc quản lý hiệu quả quy trình này.

Bên cạnh đó, nội dung chương 1 cũng tập trung giới thiệu về học máy, các mô hình học máy và ứng dụng của các mô hình học máy trong việc tối ưu hóa hạn mức tín dụng. Ngoài ra, chương 1 còn giới thiệu về Streamlit – một công cụ mạnh mẽ hỗ trợ xây dựng giao diện người dùng, giúp trực quan hóa và triển khai kết quả từ các mô hình học máy. Những kiến thức này sẽ là cơ sở cho việc ứng dụng công nghệ vào công tác dự báo, tối ưu hạn mức tín dụng nhằm nâng cao hiệu quả và tối ưu hóa quy trình ra quyết định tại các tổ chức tài chính.

CHƯƠNG 2: THỰC NGHIỆM BÀI TOÁN

2.1. Mô tả bài toán và các bước thực hiện bài toán

2.1.1. Mô tả bài toán

Trong lĩnh vực tài chính, việc xác định hạn mức tín dụng chính xác là một yếu tố quan trọng giúp tối ưu hóa chiến lược tài chính và quản lý rủi ro hiệu quả. Mục tiêu của bài toán này là dựa vào các yếu tố nhân khẩu học, hành vi và lịch sử tín dụng của khách hàng để xác định hạn mức tín dụng phù hợp. Bằng cách áp dụng các mô hình học máy các tổ chức tài chính có thể đưa ra quyết định nhanh chóng, chính xác, giảm thiểu rủi ro và tối ưu hóa lợi nhuận và duy trì sự ổn định.

- Dự đoán hạn mức tín dụng: Mô hình học máy được áp dụng để dự đoán hạn mức tín dụng của khách hàng. Mô hình được huấn luyện dựa trên các yếu tố tài chính như lịch sử thanh toán, tỷ lệ sử dụng tín dụng, số lần thanh toán trễ, thu nhập ước tính và các yếu tố nhân khẩu học của khách hàng.
- So sánh kết quả: Đánh giá và so sánh kết quả dự đoán từ các mô hình Random Forest, XGBoost và Support Vector Regression (SVR) dựa trên độ chính xác và khả năng dự đoán của mô hình.
- Xây dựng web dự đoán hạn mức tín dụng: Sử dụng thư viện streamlit để xây dựng giao diện người dùng, thu thập thông tin người dùng và sử dụng mô hình học máy đã huấn luyện để dự đoán hạn mức tín dụng.

2.1.2. Các bước thực hiện bài toán

- Bước 1: Tiền xử lý dữ liệu.
- Bước 2: Khám phá dữ liệu
- Bước 3: Xây dựng mô hình Random Forest, XGBoost, SVR dự đoán hạn mức tín dụng của khách hàng.
- Bước 4: Đánh giá và so sánh kết quả giữa ba mô hình.

2.2. Dữ liệu thực nghiệm

Bộ dữ liệu chứa thông tin về 30000 khách hàng sử dụng thẻ tín dụng, bao gồm 25 cột tập trung vào các yếu tố liên quan đến hành vi sử dụng thẻ tín dụng, tình trạng tài chính và lịch sử thanh toán của khách hàng.

Nhóm đặc điểm nhân khẩu học: Nhóm này tập trung vào các đặc điểm cơ bản của khách hàng, có ảnh hưởng mạnh đến hành vi tài chính và nhu cầu sử dụng thẻ tín dụng của khách hàng.

- AGE: Độ tuổi ảnh hưởng mạnh đến hành vi sử dụng thẻ tín dụng của khách hàng. Các khách hàng ở độ tuổi trẻ (dưới 30 tuổi) có thể sử dụng thẻ tín dụng chủ yếu cho chi tiêu cá nhân và ít duy trì số dư lớn. Khách hàng trung niên thường sẽ có xu hướng duy trì thẻ tín dụng với hạn mức cao hơn và sử dụng để quản lý tài chính gia đình. Khách hàng lớn tuổi có thể sử dụng ít hơn do nhu cầu tài chính thay đổi.
- SEX: Giới tính có thể ảnh hưởng đến cách thức sử dụng thẻ tín dụng. Nam giới thường có xu hướng sử dụng thẻ tín dụng với hạn mức cao hơn để đầu tư vào sản phẩm tài chính lâu dài. Phụ nữ có thể tập trung vào các sản phẩm bảo hiểm sức khỏe hoặc các khoản vay ngắn hạn hơn.
- EDUCATION: Những khách hàng có trình độ học vấn cao thường có thu nhập ổn định và nhận thức tài chính tốt hơn, giúp họ dễ dàng quản lý thẻ tín dụng và yêu cầu hạn mức tín dụng cao hơn. Họ có thể sử dụng thẻ tín dụng cho các mục đích dài hạn như đầu tư hoặc các khoản chi tiêu lớn.
- MARRIAGE: Những khách hàng đã kết hôn thường có nhu cầu tài chính lớn hơn do phải chi tiêu cho gia đình, nuôi dưỡng con cái và các khoản tiền chi tiêu khác nên nhóm khách hàng này thường có xu hướng yêu cầu hạn mức tín dụng cao so với những khách hàng chưa kết hôn.

Nhóm đặc điểm tài chính: Các đặc trưng này phản ánh khả năng tài chính của khách hàng và ảnh hưởng đến quyết định điều chỉnh hạn mức tín dụng.

- LIMIT_BAL: Hạn mức tín dụng của khách hàng được xác định chủ yếu bởi khả năng tài chính của khách hàng. Những khách hàng có thu nhập cao hơn và số dư tài khoản ổn định sẽ được cấp hạn mức tín dụng cao hơn vì họ có khả năng trả nợ tốt hơn. Ngược lại, hạn mức tín dụng thấp cho thấy khách hàng có khả năng tài chính hạn hẹp.
- PAY_STATUS_Sept – PAY_STATUS_Apr (Tình trạng thanh toán trong 6 tháng gần nhất): Khách hàng thanh toán đúng hạn được xem là có khả năng tài chính tốt hơn và thường được cấp hạn mức tín dụng cao hơn. Trong khi đó, những khách hàng thường xuyên thanh toán trễ có thể bị giới hạn hạn mức tín dụng hoặc gặp khó khăn khi yêu cầu tăng hạn mức tín dụng.
- BILL_BALANCE_Sept – BILL_BALANCE_Apr (Số dư hóa đơn trong 6 tháng gần nhất): Số tiền khách hàng còn nợ trên thẻ tín dụng vào cuối kỳ thanh toán, nếu số dư hóa đơn khách hàng có thể đang gặp khó khăn về tài chính hoặc chi tiêu vượt mức ảnh hưởng đến việc cấp hạn mức tín dụng vì ngân hàng thấy rủi ro cao.
- PAY_AMOUNT_Sept – PAY_AMOUNT_Apr (Số tiền thanh toán trong 6 tháng gần nhất): Phản ánh mức độ tin cậy của khách hàng, nếu khách hàng thanh toán đầy đủ tức là khách hàng có khả năng tài chính tốt.
- Default payment next month: Khách hàng có nguy cơ vỡ nợ cao (default payment = 1) sẽ bị giảm hạn mức tín dụng hoặc không yêu cầu tăng hạn mức tín dụng trong tương lai. Những khách hàng không có nguy cơ vỡ nợ (default payment = 0) sẽ có cơ hội tăng hạn mức tín dụng.

2.3. Tiền xử lý dữ liệu

2.3.1. Import các thư viện cần thiết

Để bắt đầu quy trình phân tích và dự báo hạn mức tín dụng, các thư viện sau được nhập vào. Mỗi thư viện đóng vai trò quan trọng trong các bước xử lý, phân tích, huấn luyện và đánh giá mô hình. Cụ thể:

- Xử lý dữ liệu:
    - Pandas: Đọc, xử lý, cung cấp cấu trúc dữ liệu và phân tích dữ liệu.
    - Numpy: Xử lý các phép toán số học, các phép toán mảng và ma trận.
- Chia dữ liệu và xây dựng mô hình học máy:
    - Các công cụ từ sklearn.model_selection như tran_test_split được sử dụng để chia dữ liệu thành các tập huấn luyện và kiểm tra.
    - XGBRegressor, RandomForestRegressor, SVR: Xây dựng mô hình dự đoán hạn mức tín dụng của khách hàng.
- Đánh giá mô hình: Các chỉ số từ sklearn.metrics như r2_score, mean_absolute_error (MAE), mean_squared_error (RMSE).
- Chuẩn bị dữ liệu đầu vào:
    - StandardScaler: Chuẩn hóa dữ liệu, đảm bảo rằng tất cả các tính năng có cùng một phạm vi giá trị.
    - LabelEncoder: Mã hóa các biến phân loại thành số nguyên.
- Tắt cảnh báo: warnings.filterwarnings

2.3.2. Kiểm tra cấu trúc và chất lượng dữ liệu

Quá trình tiền xử lý dữ liệu bắt đầu bằng việc đọc dữ liệu gốc, kiểm tra các thông tin như kiểu dữ liệu, các cột, số lượng quan sát và số lượng giá trị thiếu. Điều này giúp hiểu rõ hơn về cấu trúc và chất lượng dữ liệu trước khi tiến hành các bước tiền xử lý dữ liệu tiếp theo.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          30000 non-null  int64
 1   LIMIT_BAL                   30000 non-null  int64
 2   SEX                         30000 non-null  object
 3   EDUCATION                   30000 non-null  int64
 4   MARRIAGE                    30000 non-null  object
 5   AGE                         30000 non-null  int64
 6   PAY_STATUS_Sept             30000 non-null  int64
 7   PAY_STATUS_Aug              30000 non-null  int64
 8   PAY_STATUS_Jul              30000 non-null  int64
 9   PAY_STATUS_Jun              30000 non-null  int64
 10  PAY_STATUS_May              30000 non-null  int64
 11  PAY_STATUS_Apr              30000 non-null  int64
 12  BILL_BALANCE_Sept           30000 non-null  int64
 13  BILL_BALANCE_Aug            30000 non-null  int64
 14  BILL_BALANCE_Jul            30000 non-null  int64
 15  BILL_BALANCE_Jun            30000 non-null  int64
 16  BILL_BALANCE_May            30000 non-null  int64
 17  BILL_BALANCE_Apr            30000 non-null  int64
 18  PAY_AMOUNT_Sept             30000 non-null  int64
 19  PAY_AMOUNT_Aug              30000 non-null  int64
 20  PAY_AMOUNT_Jul              30000 non-null  int64
 21  PAY_AMOUNT_Jun              30000 non-null  int64
 22  PAY_AMOUNT_May              30000 non-null  int64
 23  PAY_AMOUNT_Apr              30000 non-null  int64
 24  default payment next month  30000 non-null  int64
dtypes: int64(23), object(2)
memory usage: 5.7+ MB
None
```

Hình 2.1: Cấu trúc và chất lượng dữ liệu gốc

Đoạn code kiểm tra cấu trúc và chất lượng dữ liệu được trình bày chi tiết tại phụ lục 1.1. Qua kiểm tra cho thấy bộ dữ liệu bao gồm 30000 quan sát và 25 cột, chủ yếu chứa các thông tin tài chính và nhân khẩu học như hạn mức tín dụng, tình trạng thanh toán, số dư hóa đơn. Dữ liệu có đầy đủ thông tin và không có dữ liệu thiếu. Dữ liệu chủ yếu chưa các giá trị số (int64), nhưng vẫn có vài cột có kiểu dữ liệu chưa phù hợp cần được xử lý.

2.3.3. Mã hóa các biến phân loại

Đối với biến "SEX" biểu thị giới tính của khách hàng, ban đầu có giá trị ở dạng chuỗi bao gồm "Nam và "Nữ". Tuy nhiên trong các thuật toán học máy các giá trị chuỗi không thể trực tiếp sử dụng được. Do đó cần phải chuyển đổi các giá trị này thành dạng số để các thuật toán có thể xử lý được. Cụ thể, trong trường hợp biến "SEX" phương thức map() trong Python được sử dụng, mỗi giá trị duy nhất trong biến danh mục được gán một mã số . Ví dụ , giá trị "Nam" sẽ được gán là 1 và giá trị "Nữ" được gán là 0.

Biến "MARRIAGE" thể hiện tình trạng hôn nhân của khách hàng ban đầu chứa các giá trị "Đã kết hôn", "Độc thân", "Khác" và "0". Phương thức

map() trong Python được áp dụng để mã hóa các giá trị "Đã kết hôn" được gán là 1, "Độc thân" được gán là 2, giá trị "Khác" được gán là 3 và giá trị "0" cũng được gán là 3. Cách thức mã hóa này được trình bày chi tiết trong phụ lục 1.2.1.

Trong bộ dữ liệu, cột "EDUCATION" biểu thị trình độ học vấn của khách hàng với các giá trị bao gồm 1 là "Cao học", 2 là "Đại học", 3 tức là "Trung học phổ thông" và giá trị 4 là "Khác". Tuy nhiên, trong bộ dữ liệu vẫn tồn tại những giá trị không có ý nghĩa chẳng hạn như các giá trị 0, 5, 6. Những giá trị này không thuộc vào ba nhóm chính, điều này có thể gây khó khăn trong việc phân tích và xây dựng mô hình học máy vì các giá trị này không cung cấp thêm thông tin hữu ích. Vì vậy, để đảm bảo chất lượng dữ liệu phục vụ cho quá trình xây dựng mô hình học máy, các giá trị 0, 5, và 6 được gộp thành một nhóm "Khác" với mã giá trị là 4. Việc gộp nhóm này sẽ giúp mô hình học máy có thể học thông tin một cách có hiệu quả hơn. Mã code thực hiện quy trình này được trình bày ở phụ lục 1.2.1.

2.3.4. Tạo các đặc trưng mới

Để mô hình học máy hoạt động hiệu quả và chính xác hơn việc tạo thêm các đặc trưng mới từ dữ liệu gốc là một bước quan trọng trong quá trình tiền xử lý dữ liệu. Mặc dù dữ liệu thô có thể chứa các thông tin hữu ích nhưng việc tạo thêm các đặc trưng sẽ giúp mô hình có thêm nhiều thông tin để học. Đoạn mã thực hiện quy trình tạo các đặc trưng mới được trình bày tại phụ lục 1.2.3. Các đặc trưng mới có thể cung cấp các thông tin bổ sung, giúp mô hình nhận diện các mẫu hành vi, tăng khả năng phân biệt giữa các lớp và nâng cao hiệu quả hoạt động.

Tỷ lệ sử dụng tín dụng: Được tạo ra từ các cột BILL_BALANCE và LIMIT_BAL, đại diện cho mức độ sử dụng hạn mức tín dụng của khách hàng. Cụ thể, tỷ lệ sử dụng tín dụng được tính bằng cách lấy trung bình của các giá trị dư nợ hàng tháng và chia cho hạn mức tín dụng của khách hàng. Đặc trưng này đo lường mức độ sử dụng tín dụng của khách hàng so với khả năng chi trả của họ. Tỷ lệ sử dụng tín dụng là một chỉ số quan trọng để đánh giá khả năng

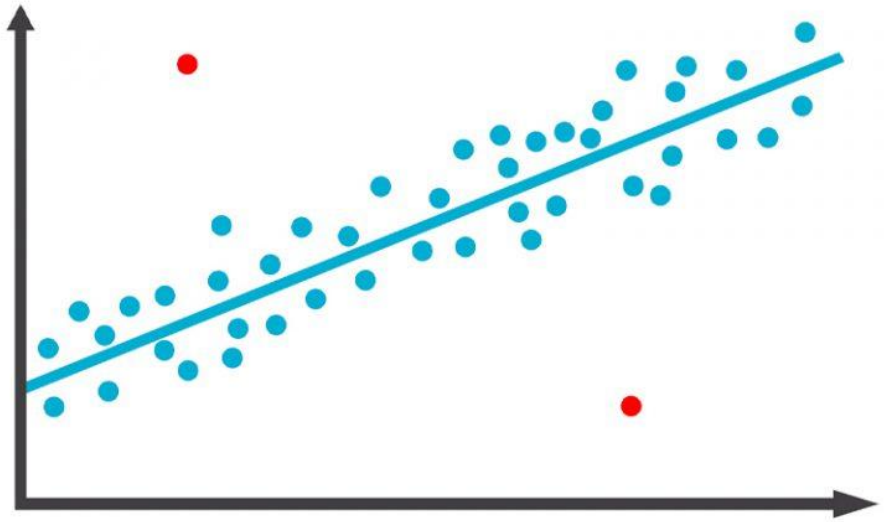
tài chính của khách hàng và giúp mô hình học máy nhận diện các khách hàng có khả năng trả nợ thấp hơn, từ đó hỗ trợ cho quy trình dự đoán hạn mức tín dụng của khách hàng chính xác hơn.

Số dư hóa đơn trung bình: Được tính bằng cách lấy trung bình của số tiền khách hàng đã chi tiêu trên thẻ tín dụng trong 6 tháng. Chỉ số này phản ánh mức độ chi tiêu của khách hàng, nếu khách hàng chi tiêu nhiều tức là số dư hóa đơn cao, ngược lại khách hàng có số dư hóa đơn trung bình thấp được xem là có khả năng tài chính ổn định và ít rủi ro hơn.

Thu nhập: Thu nhập là yếu tố quan trọng trong việc xác định hạn mức tín dụng. Thông thường các ngân hàng sẽ cấp hạn mức tín dụng cho khách hàng cao hơn lương từ 2 – 3 lần nên trong khóa luận này thu nhập được tính bằng 1/3 hạn mức tín dụng của khách hàng.

2.3.5. Xử lý ngoại lai

Dữ liệu ngoại lai là những giá trị được ghi nhận có sự khác biệt bất thường so với những giá trị dữ liệu khác, không theo một quy tắc chung nào và có thể gây ra sự sai lệch trong kết quả phân tích và việc xây dựng các thuật toán dự đoán.



Hình 2.2: Dữ liệu ngoại lai

Để xử lý ngoại lai trong bộ dữ liệu này, phương pháp IQR (Interquartile Range) được áp dụng, sử dụng công thức: Q1 – 1.5 * IQR và Q3 + 1.5 * IQR nhằm xác định các giá trị nằm ngoài phạm vi hợp lý và tiến hành loại bỏ. Việc xử lý ngoại lai giúp cải thiện chất lượng dữ liệu, từ đó giúp mô hình phân tích và dự đoán chính xác hơn. Quy trình xử lý ngoại lai được trình bày tại phụ lục 1.2.4.

2.3.6. Chuẩn hóa dữ liệu

Dữ liệu được chuẩn hóa bằng StandardScaler, đây là bước quan trọng để đưa các biến số về cùng thang đo giúp giảm ảnh hưởng của các biến có giá trị lớn hơn chi phối mô hình. StandardScale hoạt động bằng cách tính trung bình và độ lệch chuẩn mỗi cột, sau đó chuẩn hóa từng giá trị theo công thức. Kết quả là tất cả các cột chuẩn hóa sẽ có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Nếu mã hóa các biến danh mục giúp mô hình hiểu được các biến phi số thì quá trình chuẩn hóa giúp mô hình có thể học nhanh hơn và giảm thiểu lỗi trong quá trình tối ưu.

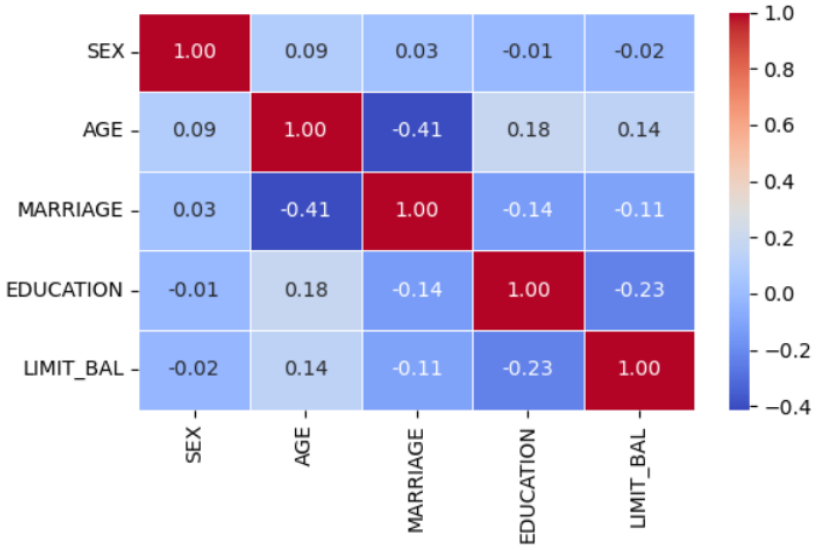
2.4. Khám phá dữ liệu

2.4.1. Mối quan hệ giữa các yếu tố nhân khẩu học và hạn mức tín dụng

Biểu đồ ma trận tương quan cho thấy mối quan hệ giữa các yếu tố nhân khẩu học và hạn mức tín dụng. Các giá trị trong ma trận tương quan phản ánh mức độ liên kết giữa từng cặp biến. Từ biểu đồ có thể thấy biến “SEX” có tương quan rất yếu với tất cả các yếu tố khác, đặc biệt với “LIMIT_BAL” cho thấy trong bài toán này giới tính không phải yếu tố quan trọng trong việc xác định hạn mức tín dụng.

AGE có mối tương quan nhẹ với LIMT_BAL (0.14), tuy nhiên mối quan hệ giữa hai biến không quá rõ ràng. Giữa AGE và MARRIAGE có mối tương quan âm (-0.41) cho thấy xu hướng giảm tỷ lệ kết hôn khi độ tuổi tăng lên. Giữa MARRIAGE và LIMIT_BAL cũng là sự tương quan âm (-0.11), điều này có nghĩa là tình trạng hôn nhân không có ảnh hưởng đến hạn mức tín dụng. EDUCATION cũng có sự tương quan âm (-0.23) với LIMIT_BAL cho thấy

trong bài toán này khách hàng có trình độ học vấn cao có thể ít yêu cầu hạn mức tín dụng cao, điều này có thể do những khách hàng này có thu nhập ổn định và ít cần đến hạn mức tín dụng lớn.



Hình 2.3: Ma trận tương quan giữa các yếu tố nhân khẩu học và hạn mức tín dụng

Hạn mức tín dụng có mối tương quan rất yếu với các yếu tố khác, điều này cho thấy rằng các yếu tố nhân khẩu học như giới tính, độ tuổi, tình trạng hôn nhân và trình độ học vấn có ảnh hưởng không đáng kể đến việc xác định hạn mức tín dụng.

2.4.2. Mối quan hệ giữa hạn mức tín dụng và độ tuổi

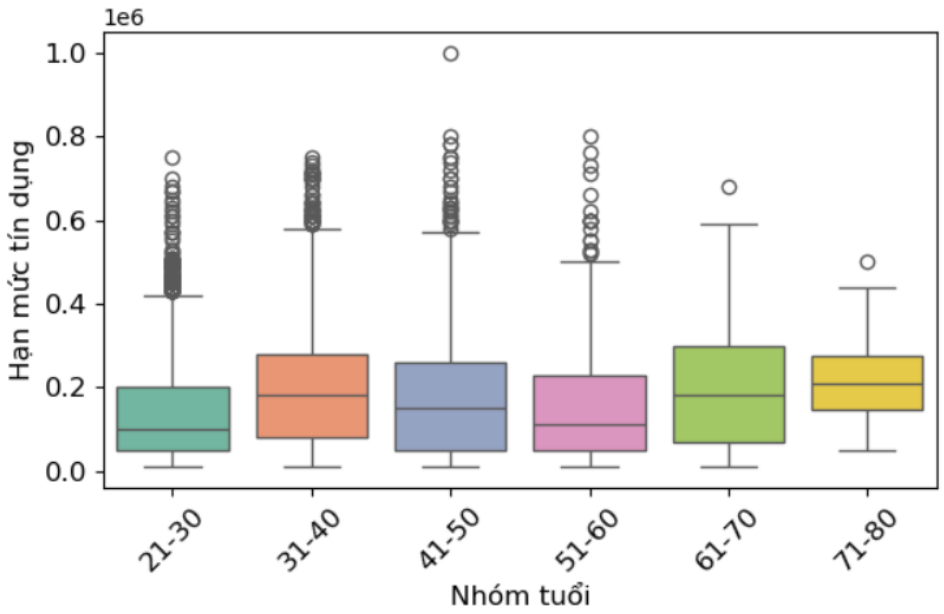
Biểu đồ boxplot thể hiện sự phân phối của từng hạn mức tín dụng cho từng nhóm tuổi. Trong đó, trục hoành thể hiển các nhóm tuổi và trục tung hiển thị hạn mức tín dụng. Mỗi nhóm tuổi được đại diện bằng một boxplot cùng với các điểm ngoại lai.

Đối với nhóm tuổi 21 – 30, hạn mức tín dụng của nhóm tuổi này có sự phân bố khá rộng với mức trung vị xung quanh 0.2 triệu. Tuy nhiên, nhóm tuổi này lại có số lượng các điểm ngoại lai lớn ở phía trên cho thấy một số ít khách hàng có hạn mức tín dụng rất cao. Điều này có nghĩa là mặc dù nhóm tuổi trẻ có hạn mức tín dụng thấp nhất nhưng một số cá nhân vẫn có thể có hạn mức tín dụng cao, có thể do các yếu tố tài chính riêng biệt hoặc thu nhập cao.

Hạn mức tín dụng của nhóm tuổi từ 31 – 40 cao hơn so với nhóm tuổi 21 – 30, với mức trung vị khoảng 0.4 triệu. Xu hướng này có thể chỉ ra rằng khách hàng ở độ tuổi trưởng thành từ 31 – 40 có thể có sự ổn định tài chính hơn, thu nhập ổn định và khả năng quản lý tài chính tốt hơn nên có hạn mức tín dụng cao hơn so với nhóm tuổi 21 – 30.

Những khách hàng ở nhóm tuổi 41 – 50 có sự phân bố hạn mức tín dụng khá ổn định, nhóm tuổi này có hạn mức tín dụng cao nhất trong tất cả các nhóm tuổi. Đây là nhóm khách hàng có tài chính ổn định, có thể hoàn thành các nghĩa vụ tài chính trước đó, thu nhập cao và có xu hướng yêu cầu hạn mức tín dụng cao hơn.

Với nhóm tuổi 51 – 60, hạn mức tín dụng ở nhóm khách hàng này có sự giảm nhẹ so với nhóm khách hàng từ 41 – 50 tuổi, nhưng vẫn ở mức khá ổn định với mức trung vị khoảng 0.5 triệu. Nhóm khách hàng ở độ tuổi này có thể đang ở giai đoạn nghỉ hưu hoặc thay đổi trong tài chính cá nhân nên có xu hướng yêu cầu hạn mức tín giảm một chút so với độ tuổi trước.



Hình 2.4: Mối quan hệ giữa hạn mức tín dụng và độ tuổi

Nhóm khách hàng ở độ tuổi từ 61 – 70 có hạn mức tín dụng giảm đáng kể, phân phối này cho thấy hầu hết khách hàng trong nhóm tuổi này có hạn mức

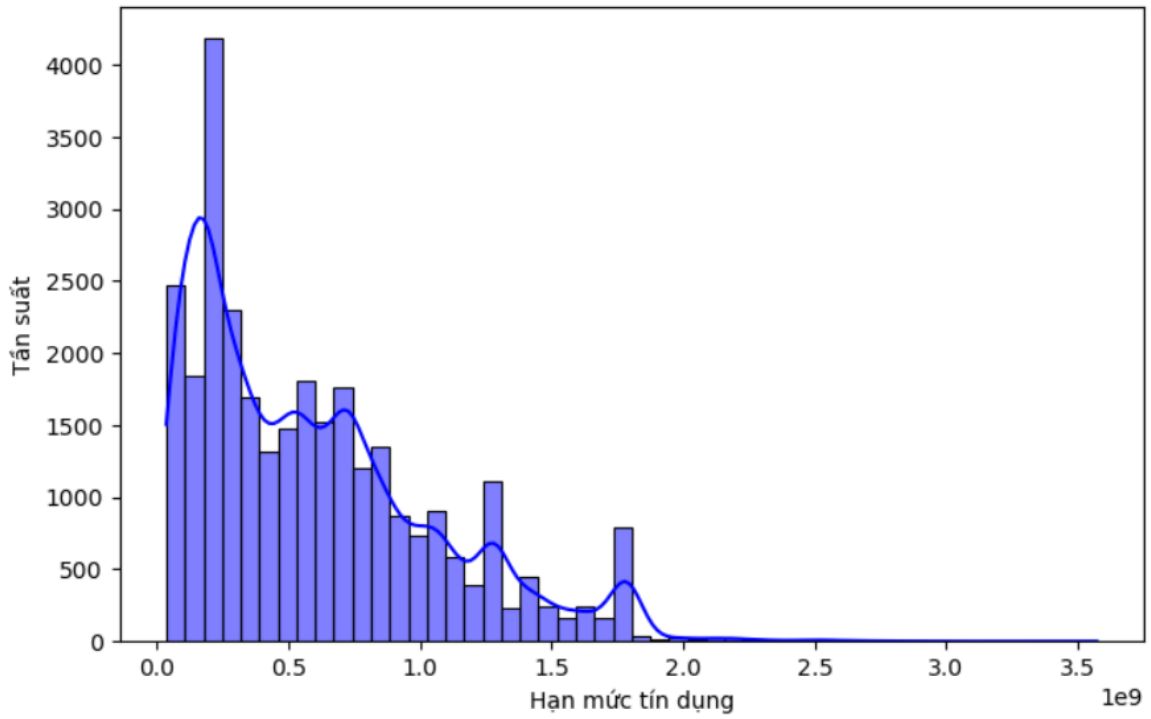
tín dụng thấp hơn. Những khách hàng trong độ tuổi này có hạn mức tín dụng thấp có thể do thu nhập giảm.

Nhóm tuổi từ 71 – 80 là nhóm có hạn mức tín dụng thấp nhất, điều này phản ánh sự giảm dần trong nhu cầu tín dụng khi độ tuổi tăng lên, đặc biệt trong nhóm khách hàng cao tuổi khi họ có ít nhu cầu sử dụng tín dụng và có thu nhập giảm.

Hạn mức tín dụng có xu hướng tăng lên từ nhóm tuổi 21 – 30 đến nhóm tuổi 41 – 50 rồi giảm dần ở các nhóm tuổi cao hơn. Nhóm tuổi 21 – 30 có hạn mức tín dụng thấp có thể do họ thường có ít kinh nghiệm quản lý tài chính và thu nhập không ổn định. Những khách hàng ở độ tuổi 31 – 50 có xu hướng hạn mức tín dụng cao vì thu nhập ổn định và khả năng quản lý tài chính tốt. Nhóm khách hàng từ 61 – 80 có hạn mức tín dụng thấp do nhu cầu tín dụng giảm khi không còn làm việc và thu nhập hạn chế.

2.4.3. Phân phối hạn mức tín dụng

Biểu đồ phân phối hạn mức tín dụng với trục hoành biểu thị hạn mức tín dụng, trục tung thể hiện tần suất của các giá trị hạn mức tín dụng. Các giá trị này cho thấy số lượng khách hàng có hạn mức tín dụng rơi vào từng khoảng giá trị trên trục hoành. Biểu đồ histogram được vẽ dưới dạng các cột, mỗi cột biểu thị số lượng khách hàng có hạn mức tín dụng trong một khoảng nhất định. Đường cong KDE (Kernel Density Estimate) thể hiện mật độ xác suất của phân phối hạn mức tín dụng. Đỉnh của biểu đồ nằm ở khoảng 0 đến 0,1 cho thấy phần lớn khách hàng có hạn mức tín dụng thấp. Biểu đồ có sự phân bố lệch phải với một đỉnh cao phía gần giá trị 0 và giảm dần về phía bên phải. Điều này cho thấy đa số khách hàng có hạn mức tín dụng thấp nhưng một số khách hàng lại có hạn mức tín dụng cao dẫn đến phân phối lệch.



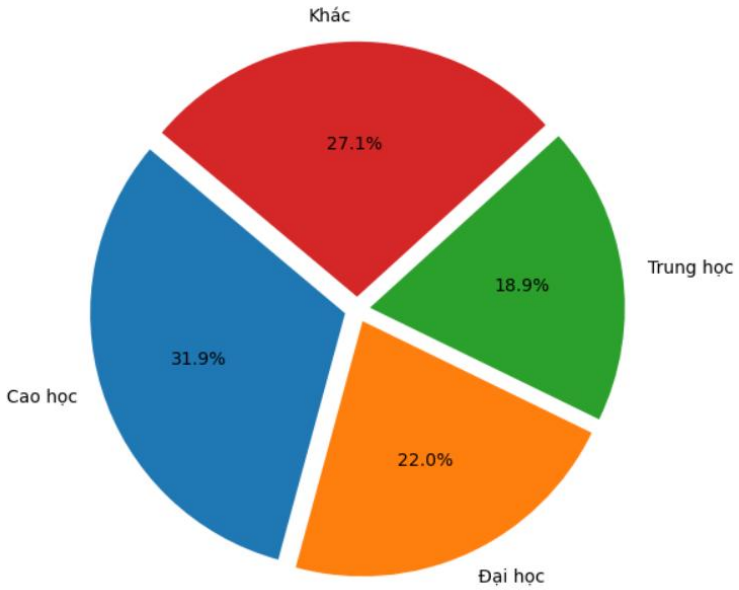
Hình 2.5: Phân phối hạn mức tín dụng

Từ biểu đồ, có thể thấy rằng hạn mức tín dụng thấp dưới 0,2 là phổ biến nhất. Đây là một đặc điểm thường thấy trong các khoản vay tiêu dùng, đa số là các khách hàng cá nhân. Phần dữ liệu bên phải của biểu đồ rất ít với tần suất giảm dần khi hạn mức tăng lên, số lượng khách hàng với hạn mức tín dụng cao chỉ chiếm một phần nhỏ nhưng những khách hàng này có thể ảnh hưởng đến tổng số tiền cho vay hoặc các chỉ số tài chính khác. Biểu đồ cung cấp cái nhìn sâu sắc về cách hạn mức tín dụng được phân bố trong tập dữ liệu khách hàng. Phần lớn khách hàng có hạn mức tín dụng thấp nhưng vẫn có số ít khách hàng có hạn mức cao. Việc phân tích sự phân phối này giúp các tổ chức tín dụng hiểu rõ hơn về khách hàng của mình để xây dựng các chiến lược tín dụng phù hợp để giảm thiểu rủi ro và tối ưu hóa việc cấp tín dụng.

2.4.4. Phân bổ hạn mức tín dụng trung bình dựa trên trình độ học vấn

Biểu đồ này thể hiện phân bổ hạn mức tín dụng trung bình của khách hàng theo trình độ học vấn, cụ thể: Nhóm khách hàng có trình độ học vấn cao học chiếm tỷ lệ lớn nhất trong biểu đồ với 31,9%. Điều này cho thấy những người có trình độ học vấn cao thường có công việc ổn định, thu nhập tốt hơn

nên có thể được cấp hạn mức tín dụng cao hơn. Nhóm khách hàng "Khác" có tỷ lệ lớn thứ hai chiếm 27,1% đây là những khách hàng không thuộc các nhóm học vấn chính thức nhưng vẫn khả năng tài chính tốt. Nhóm khách hàng có trình độ học vấn đại học chiếm tỷ lệ 22%, những người có trình độ học vấn đại học thường có công việc ổn định và thu nhập tốt nhưng vẫn kém hơn so với nhóm khách hàng có trình độ học vấn cao học. Tuy nhiên đây vẫn là nhóm khách hàng tiềm năng được cấp hạn mức tín dụng ở mức trung bình. Nhóm khách hàng có trình độ học vấn trung học chiếm tỷ lệ thấp nhất, đây là nhóm khách hàng có trình độ học vấn thấp hơn nên thu nhập và khả năng tài chính ít ổn định hơn vì vậy họ có hạn mức tín dụng thấp hơn các nhóm khác.



Hình 2.6: Hạn mức tín dụng trung bình dựa trên trình độ học vấn

Dựa trên biểu đồ có thể thấy mối liên hệ giữa trình độ học vấn và hạn mức tín dụng, những khách hàng có trình độ học vấn cao hơn thường có thu nhập và khả năng tài chính tốt hơn do đó họ được cấp hạn mức tín dụng cao hơn so với những nhóm khách hàng khác. Ngược lại, những khách hàng có trình độ học vấn thấp hơn có thể có thu nhập không ổn định dẫn đến hạn mức tín dụng thấp. Nhóm "Khác" có tỷ lệ khá cao với 27,1%, đây là nhóm khách hàng không có trình độ học vấn rõ ràng nhưng vẫn có khả năng tài chính tốt.

Điều này cho thấy các tổ chức tài chính ngoài đánh giá trình độ học vấn thì còn xem xét nhiều yếu tố khác khi cấp hạn mức tín dụng.

2.5. Xây dựng mô hình dự đoán hạn mức tín dụng

2.5.1. Chia tập dữ liệu

Quy trình xây dựng mô hình bắt đầu với việc phân chia tập dữ liệu gốc thành dữ liệu huấn luyện và dữ liệu kiểm tra. Đầu tiên, các cột không liên quan như "ID" được loại bỏ khỏi tập dữ liệu, cột "LIMIT_BAL" được tách ra thành nhãn y và là biến mục tiêu. Phương pháp train_test_split từ thư viện scikit-learn được sử dụng để thực hiện phân chia tập dữ liệu thành hai phần với 80% dữ liệu được dùng để huấn luyện và 20% dữ liệu để kiểm tra mô hình. Đoạn mã thực hiện phân chia tập dữ liệu được trình bày tại phụ lục 1.4.1.

Cụ thể, tập huấn luyện bao gồm 24000 quan sát, tập kiểm tra bao gồm 6000 quan sát. Việc phân chia dữ liệu được thực hiện theo tỷ lệ 80:20 nhằm đảm bảo mô hình có đủ dữ liệu để học và cũng có khối lượng dữ liệu đủ lớn để đánh giá hiệu quả của mô hình.

2.5.2. Xây dựng mô hình Random Forest

Khởi tạo mô hình: Mô hình Random Forest Regressor được khởi tạo với các tham số để tối ưu quá trình học. Cụ thể, mô hình sử dụng n_estimators để xác định số lượng cây quyết định, min_samples_split và min_samples_leaf giúp điều chỉnh số lượng mẫu tối thiểu để một nút cây có thể chia và có ít nhất bao nhiêu mẫu tại mỗi lá cây. Sử dụng tham số max_feature xác định số lượng đặc trưng được sử dụng trong mỗi nút cây, max_depth để xác định độ sâu tối đa của mỗi cây giúp tránh hiện tượng overfitting. Các tham số khác như bootstrap và random_state giúp điều chỉnh cách dữ liệu được lấy mẫu và tạo tính ngẫu nhiên trong quá trình huấn luyện.

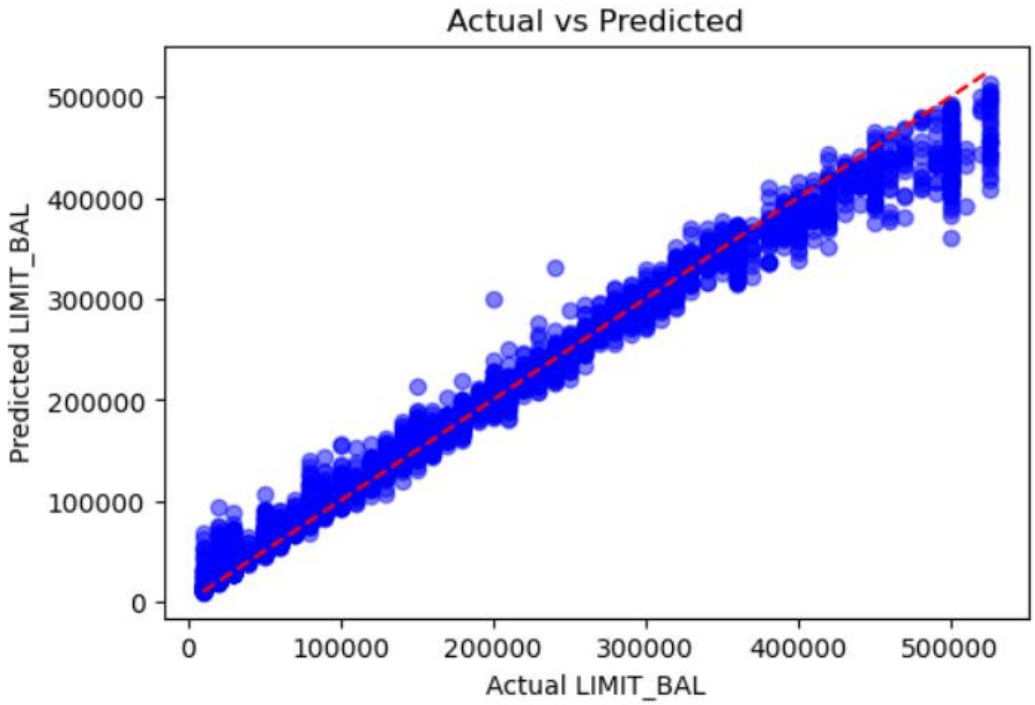
Huấn luyện mô hình: Mô hình được huấn luyện bằng cách sử dụng tập dữ liệu huấn luyện X_train_scaled và nhãn kết quả y_train, quá trình huấn luyện giúp mô hình học được các mối quan hệ giữa đặc trưng đầu vào và kết quả đầu ra. Sau khi huấn luyện, mô hình sẽ được sử dụng để dự đoán trên tập kiểm tra,

mô hình sử dụng các thông tin đã học từ tập huấn luyện để dự đoán giá trị cho các mẫu trong tập kiểm tra.

Đánh giá mô hình: Sử dụng các chỉ số đánh giá như RMSE (đo lường sai lệch bình phương trung bình), MAE (đo lường sai lệch tuyệt đối trung bình) và R^2 (đo lường mức độ giải thích của mô hình) để tính toán và đánh giá hiệu suất. Tỷ lệ của các chỉ số RMSE, MAE và R^2 so với giá trị trung bình của biến mục tiêu được tính để chuyển đổi sai số thành tỷ lệ phần trăm, giúp dễ dàng so sánh mức độ sai số trong mối tương quan với giá trị thực tế của biến mục tiêu. Quy trình xây dựng mô hình Random Forest được trình bày chi tiết tại phụ lục 1.4.2.

Kết quả đánh giá của mô hình Random Forest trên tập kiểm tra cho thấy hiệu suất của mô hình rất tốt. Chỉ số RMSE (Root Mean Square Error) đạt 14137,02 tương đương với 8,42% giá trị trung bình của biến mục tiêu. Điều này cho thấy sai số bình phương trung bình tương đối thấp so với giá trị thực tế. Chỉ số MAE (Mean Absolute Error) đạt 7651,65 tương đương 4,55% giá trị trung bình, phản ánh rằng sai số trung bình tuyệt đối giữa giá trị dự đoán và giá trị thực tế cũng khá thấp. Đặc biệt, chỉ số R^2 đạt 0,987804 gần như hoàn hảo cho thấy mô hình có khả năng giải thích tới 98,74% sự biến thiên của biến mục tiêu.

Để đánh giá chi tiết hơn về mức độ chính xác của mô hình có thể dựa vào biểu đồ dưới đây, biểu đồ này thể hiện sự phân bố của các điểm dữ liệu dựa trên mối quan hệ giữa giá trị thực tế và giá trị dự đoán từ mô hình. Trong đó, trục hoành đại diện cho giá trị thực tế, trục tung đại diện cho giá trị dự đoán, các điểm dữ liệu trong biểu đồ thể hiện sự phân tán giữa giá trị thực tế và giá trị dự đoán.



Hình 2.7: Giá trị thực tế và dự đoán mô hình Random Forest

Qua biểu đồ có thể thấy mối quan hệ tuyến tính rõ rệt giữa giá trị thực tế và giá trị dự đoán với phần lớn các điểm dữ liệu gần sát với đường chéo đỏ. Đường chéo này thể hiện sự khớp giữa giá trị thực tế và giá trị dự đoán. Mặc dù mô hình Random Forest có thể giải thích phần phương sai trong dữ liệu với R2 = 0,987804 nhưng vẫn tồn tại sự lệch nhỏ đối với các giá trị cao (phía trên biểu đồ). Điều này có nghĩa là mặc dù mô hình đã học được phần lớn các mối quan hệ giữa các đặc trưng nhưng mô hình vẫn chưa tối ưu hoàn toàn.

Mặc dù có một số sai lệch ở các giá trị cao nhưng nhìn chung biểu đồ cho thấy mô hình Random Forest có khả năng dự đoán chính xác và gần như hoàn hảo. Sự phân tán hạn chế giữa các điểm dữ liệu và đường chéo lý tưởng chứng tỏ mô hình có khả năng học các mối quan hệ trong dữ liệu rất tốt và đạt hiệu suất ổn định. Tuy nhiên, mô hình vẫn cần cải tiến thêm các tham số hoặc sử dụng các phương pháp như outlier detection, resampling techniques để cải thiện khả năng dự đoán.

2.5.3. Xây dựng mô hình XGBoost

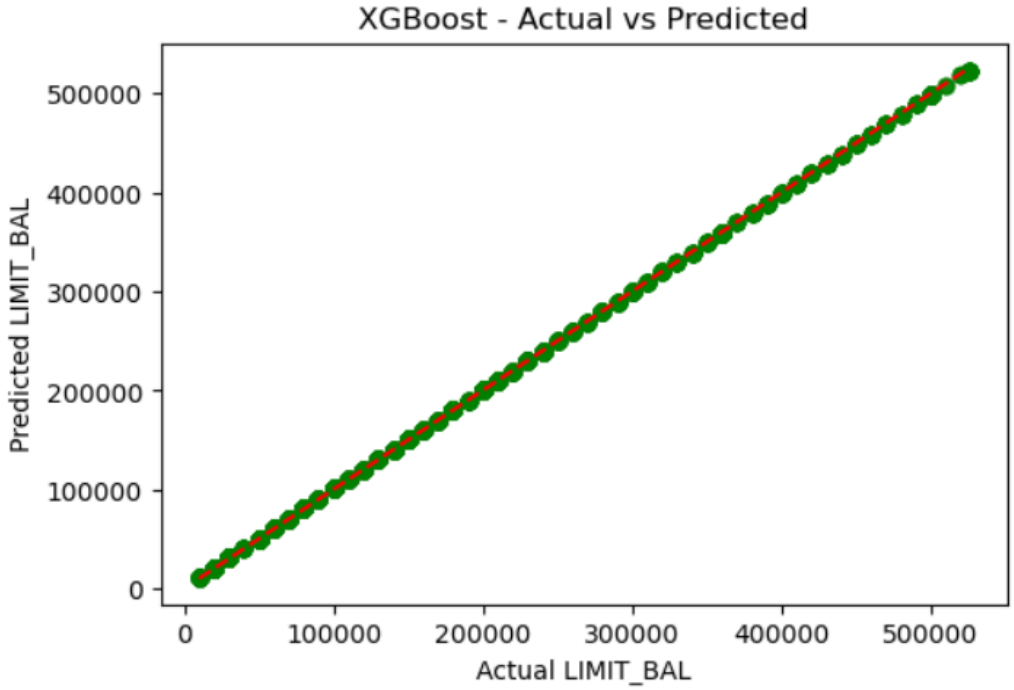
Khởi tạo mô hình: Mô hình XGBoost Regressor được thiết lập với các tham số, bao gồm: n_estimators để xác định số lượng cây trong mô hình, max_depth để điều chỉnh độ sâu tối đa của mỗi cây giúp mô hình học được các mối quan hệ phức tạp trong dữ liệu. Tham số learning_rate kiểm soát tốc độ học của mô hình, giúp mô hình điều chỉnh nhẹ qua mỗi vòng lặp. Để tránh hiện hiện tượng overfititng, tham số subsample giới hạn tỷ lệ mẫu được chọn trong mỗi cây, colsample_bytree quyết định tỷ lệ đặc trưng được sử dụng trong mỗi cây. Cuối cùng, tham số random_state đảm bảo tính tái lập của mô hình giúp kết quả huấn luyện trở nên ổn định và có thể tái tạo.

Huấn luyện mô hình: Mô hình được huấn luyện bằng cách sử dụng tập dữ liệu huấn luyện X_train_scaled và nhãn kết quả y_train, quá trình huấn luyện giúp mô hình học được các mối quan hệ giữa đặc trưng đầu vào và kết quả đầu ra. Sau khi huấn luyện, mô hình sẽ được sử dụng để dự đoán trên tập kiểm tra, mô hình sử dụng các thông tin đã học từ tập huấn luyện để dự đoán giá trị cho các mẫu trong tập kiểm tra.

Đánh giá mô hình: Sử dụng các chỉ số đánh giá như RMSE (đo lường sai lệch bình phương trung bình), MAE (đo lường sai lệch tuyệt đối trung bình) và R^2 (đo lường mức độ giải thích của mô hình) để tính toán và đánh giá hiệu suất. Tỷ lệ của các chỉ số RMSE, MAE và R^2 so với giá trị trung bình của biến mục tiêu được tính để chuyển đổi sai số thành tỷ lệ phần trăm, giúp dễ dàng so sánh mức độ sai số trong mối tương quan với giá trị thực tế của biến mục tiêu. Quy trình xây dựng mô hình XGBoost được trình bày tại phụ lục 1.4.3.

Kết quả từ mô hình XGBoost trên tập kiểm tra cho thấy mô hình đạt hiệu suất rất cao. RMSE của mô hình đạt 671,67 tương đương với 0,4% giá trị trung bình của biến mục tiêu cho thấy sai số bình phương trung bình rất nhỏ. Chỉ số MAE của mô hình đạt 543,46 tương đương với 0,32%, điều này phản ánh rằng sai lệch tuyệt đối trung bình giá trị thực tế và giá trị dự đoán rất thấp. Quan trọng hơn, chỉ số R^2 của mô hình đạt 0,999972 gần như hoàn hảo, cho thấy mô

hình giải thích gần như toàn bộ sự biến thiên của dữ liệu với 99,99% phương sai của biến mục tiêu được giải thích. Mô hình XGBoost cho thấy sự ưu việt trong việc dự đoán hạn mức tín dụng của khách hàng.



Hình 2.8: Giá trị thực tế và giá trị dự đoán mô hình XGBoost

Biểu đồ trên thể hiện mối quan hệ giữa giá trị dự đoán và giá trị thực tế của mô hình XGBoost, trong đó trục hoành biểu thị giá trị thực tế, trục tung biểu thị giá trị mà mô hình XGBoost dự đoán, mỗi điểm dữ liệu trên biểu đồ là sự kết hợp giữa giá trị dự đoán và giá trị thực tế. Qua biểu đồ có thể thấy mối quan hệ rất chặt chẽ giữa giá trị thực tế và giá trị dự đoán. Các điểm trên biểu đồ hầu như nằm trên một đường chéo, tức là mô hình dự đoán gần như chính xác tuyệt đối. Không có bất kỳ sự phân tán đáng kể nào quanh đường chéo, điều này cho thấy mô hình không có sai lệch giữa dự đoán và giá trị thực tế. Sự khớp hoàn hảo giữa các điểm và đường chéo cho thấy mô hình XGBoost đã được huấn luyện tốt và việc dự đoán hạn mức tín dụng có độ chính xác cao.

2.5.4. Xây dựng mô hình SVR (Support Vector Regression)

Khởi tạo mô hình: Mô hình Support Vector Regression (SVR) được khởi tạo với tham số kernel = ‘rbf’ cho phép hạt nhân Gaussian (Radial Basis Function) để chuyển dữ liệu vào không gian có độ phân giải cao, giúp mô hình

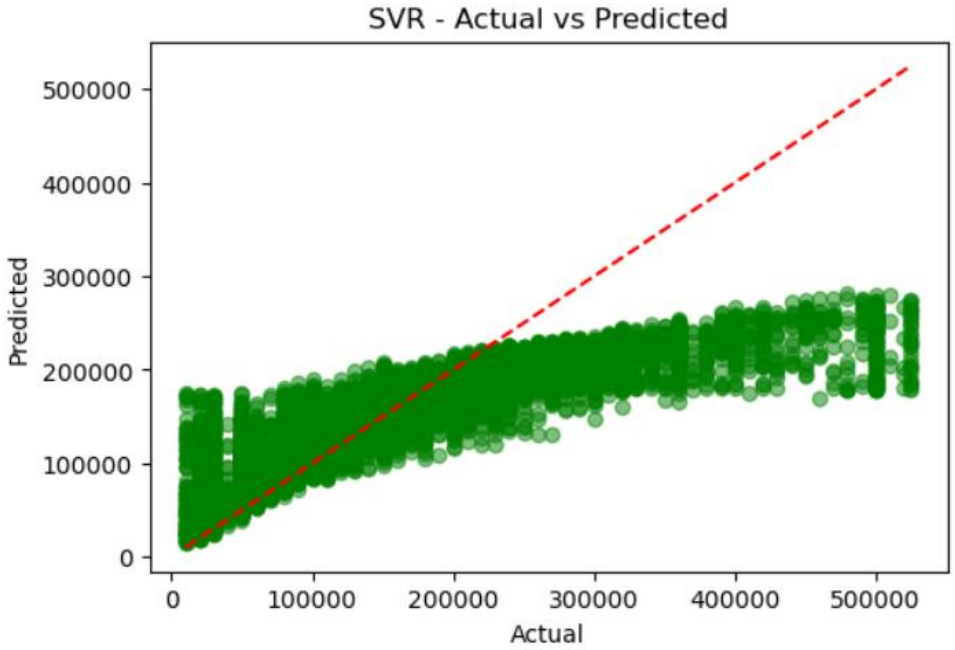
học được các mối quan hệ phi tuyến tính giữa các đặc trưng đầu vào và kết quả đầu ra. Tham số C = 100 điều chỉnh mức độ phạt cho các sai số trong quá trình huấn luyện, với giá trị càng cao sẽ làm mô hình cố gắng tối ưu hóa các sai số này. Tham số epsilon được sử dụng để điều chỉnh biên độ trong đó không có sự phạt cho các sai số nhỏ, giúp mô hình có thể bỏ qua các sai số nhỏ mà không ảnh hưởng đến quá trình học.

Huấn luyện mô hình: Sau khi khởi tạo mô hình SVR được huấn luyện dựa trên tập dữ liệu huấn luyện và nhãn kết quả đầu ra. Quá trình này giúp mô hình học được mối quan hệ giữa các đặc trưng đầu vào và nhãn đầu ra. Sau khi huấn luyện, mô hình được sử dụng để dự đoán cho tập kiểm tra.

Đánh giá mô hình: Sử dụng các chỉ số đánh giá như RMSE (đo lường sai lệch bình phương trung bình), MAE (đo lường sai lệch tuyệt đối trung bình) và R^2 (đo lường mức độ giải thích của mô hình) để tính toán và đánh giá hiệu suất. Tỷ lệ của các chỉ số RMSE, MAE và R^2 so với giá trị trung bình của biến mục tiêu được tính để chuyển đổi sai số thành tỷ lệ phần trăm, giúp dễ dàng so sánh mức độ sai số trong mối tương quan với giá trị thực tế của biến mục tiêu. Quy trình xây dựng mô hình SVR được trình bày tại phụ lục 1.4.4.

Trong quá trình đánh giá mô hình SVR trên tập kiểm tra các chỉ số đánh giá như RMSE, MAE, R^2 đã được sử dụng để đo lường mức độ chính xác của mô hình trong việc dự đoán giá trị hạn mức tín dụng. Kết quả cho thấy RMSE đạt 83864,69 tương ứng với 49,92% giá trị trung bình của biến mục tiêu. Đây là chỉ số cho thấy phương sai trung bình khá lớn, mức độ sai lệch đáng kể giữa giá trị dự đoán và giá trị thực tế. Chỉ số MAE đạt 54720,36 tương đương với 32,57% giá trị trung bình của biến LIMIT_BAL đo lường sai số tuyệt đối trung bình, kết quả này cho thấy sai số của mô hình SVR khá lớn. R^2 đạt 0,570809 có nghĩa là mô hình chỉ giải thích được 57,08% sự biến thiên trong dữ liệu, đây là một kết quả khá thấp. Điều này cho thấy mô hình SVR chưa thể giải thích hết sự biến động của biến mục tiêu và vẫn còn nhiều yếu tố chưa được mô hình xem xét hoặc dự đoán đúng.

Để minh họa rõ hơn về kết quả của mô hình SVR, biểu đồ so sánh giữa giá trị dự đoán và giá trị thực tế được sử dụng, đồng thời biểu đồ cũng phản ánh mức độ chính xác của mô hình trong việc dự đoán giá trị mục tiêu. Biểu đồ bao gồm trục hoành đại diện cho giá trị hạn mức tín dụng thực tế, trục tung đại diện cho giá trị hạn mức tín dụng mô hình SVR dự đoán, mỗi điểm màu xanh lá đại diện cho một giá trị thực tế và giá trị dự đoán hạn mức tín dụng. Đường chéo đỏ thể hiện mối quan hệ lý tưởng giữa giá trị thực tế và giá trị dự đoán, nếu tất cả các điểm dữ liệu đều nằm trên đường này tức là mô hình có thể dự đoán chính xác cao.



Hình 2.9: Giá trị thực tế và dự đoán mô hình SVR

Qua biểu đồ trên có thể thấy các điểm dữ liệu có sự phân tán khá lớn khỏi đường chéo, cho thấy có sự sai lệch lớn giữa giá trị dự đoán và thực tế. Mối quan hệ giữa giá trị thực tế và giá trị dự đoán có xu hướng tăng dần, thể hiện qua việc các điểm dữ liệu tập trung theo đường chéo đi lên. Đường chấm đỏ trong biểu là đường lý tưởng, nếu các giá trị dự đoán hoàn toàn trùng với đường này tức là mô hình có độ chính xác cao. Tuy nhiên, tại biểu đồ này có sự phân tán giữa các điểm cho thấy mô hình có độ chính xác không cao. Mô hình SVR chỉ dự đoán tương đối tốt ở các điểm có giá trị dưới 200000. Khi giá trị thực tế

tăng cao, sự phân tán của các điểm dữ liệu cũng tăng lên cho thấy mô hình ít chính xác ở các giá trị cao. Nhìn chung mô hình có thể dự đoán tốt ở các điểm có giá trị thấp nhưng mất đi độ chính xác khi các mức giá trị càng tăng cao. Điều này cho thấy mô hình SVR cần có những điều chỉnh hoặc tối ưu hóa để cải thiện độ chính xác.

2.6. Đánh giá và so sánh kết quả ba mô hình học máy

Kết quả đầu ra trên tập kiểm tra được hiển thị bao gồm các chỉ số RMSE, MAE và R^2. RMSE đo lường mức độ sai lệch lớn giữa giá trị thực tế và dự đoán, MAE tập trung vào sai số trung bình tuyệt đối, chỉ số R^2 đánh giá khả năng giải thích của mô hình đối với biến mục tiêu. Nếu RMSE và MAE thấp, đồng thời R^2 gần bằng 1 mô hình được coi là hoạt động tốt. Sau khi xây dựng và huấn luyện mô hình Random Forest, XGBoost, SVR kết quả thu được như sau:

Bảng 2.1: Kết quả đánh giá ba mô hình học máy

Mô hình	RMSE (%)	MAE (%)	R^2
Random Forest	8,42	4,55	0,987804
XGBoost	**0,4**	**0,32**	**0,999972**
Support Vector Regression	49,92	32,57	0,570809

(Nguồn: Sinh viên thực hiện)

Đối với chỉ số RMSE (Root Mean Squared Error), mô hình XGBoost có độ sai lệch thấp nhất 0,4%, điều này cho thấy mô hình XGBoost có mức độ dự đoán chính xác rất cao. Trong khi đó, mô hình Random Forest có giá trị RMSE là 8,42% mức độ sai lệch tuy không quá cao nhưng vẫn kém hơn XGBoost. SVR là mô hình có giá trị RMSE cao nhất 49,92%, có nghĩa là mô hình này có sự sai lệch lớn giữa các giá trị thực tế và giá trị dự đoán, đồng thời thể hiển khả năng dự đoán kém hơn hẳn so với hai mô hình Random Forest và XGBoost.

Về chỉ số MAE (Mean Absolute Error), mô hình XGBoost tiếp tục thể hiện sự vượt trội với giá trị MAE chỉ đạt 0,32% cho thấy sự sai lệch tuyệt đối

giữa giá trị thực tế và giá trị dự đoán rất thấp. Random Forest có MAE ở mức 4,55% cao hơn XGBoost nhưng đây vẫn là kết quả tương đối tốt, có thể được sử dụng trong nhiều ứng dụng. Tuy nhiên, mô hình SVR có mức sai số trung bình tuyệt đối rất cao 32,57%, điều này cho thấy mô hình này hoạt động không hiệu quả và cần cải thiện nhiều.

Chỉ số R^2 (Confficient of Determination) đánh giá khả năng giải thích sự biến động của dữ liệu từng mô hình. Mô hình XGBoost đạt giá trị R^2 lên tới 0,999972 gần như hoàn hảo, chứng tỏ mô hình có khả năng giải thích rất tốt sự biến động của dữ liệu. Mô hình Random Forest cũng đạt giá trị R^2 rất cao 0,987804 cho thấy mô hình mặc dù không đạt kết quả cao như XGBoost nhưng cũng có khả năng giải thích sự biến động của dữ liệu rất tốt. Ngược lại mô hình SVR có giá trị R^2 chỉ đạt 0,570809 cho thấy mô hình này chỉ giải thích được một phần nhỏ sự biến động của dữ liệu và có thể không phù hợp với bài toán dự đoán hạn mức tín dụng.

Nhìn chung, XGBoost là mô hình hiệu quả nhất trong ba mô hình Random Forest, XGBoost và SVR với các chỉ số đánh giá đều đạt kết quả tối ưu. Mô hình XGBoost không chỉ có mức độ chính xác cao mà còn giải thích tốt sự biến động của dữ liệu. Random Forest cũng là một lựa chọn tốt với độ chính xác tương đối cao. Tuy nhiên mô hình SVR với các chỉ số đánh giá kém, đặc biệt là RMSE và MAE cao cùng với chỉ số R^2 thấp không phải lựa chọn tối ưu trong bài toán này và cần được cải thiện hoặc thay thế để đạt kết quả dự đoán chính xác cao hơn.

Kết quả huân luyện của mô hình XGBoost sẽ được lưu lại và sử dụng trong quá trình xây dựng giao diện người dùng, tạo ra một ứng dụng tương tác giúp người dùng có thể dễ dàng nhập thông tin và nhận về kết quả hạn mức tín dụng tối ưu dựa trên các thông tin mà người dùng cung cấp.

TIỂU KẾT CHƯƠNG 2

Chương 2 tập trung vào trình bày chi tiết quy trình thực hiện bài toán dự đoán hạn mức tín dụng, bao gồm các bước như mô tả bài toán, tiền xử lý dữ liệu, xây dựng mô hình và đánh giá kết quả. Các bước được triển khai một cách hệ thống đảm bảo tính logic và chặt chẽ trong quá trình giải quyết bài toán. Đầu tiên bài toán được mô tả rõ ràng tạo tiền đề cho việc xác định các mục tiêu và yêu cầu cần đạt được. Tiếp theo, dữ liệu được xử lý thông qua các bước tiền xử lý dữ liệu nhằm đảm bảo chất lượng dữ liệu đầu vào, góp phần tối ưu hóa khả năng học của mô hình.

Quá trình xây dựng mô hình được thực hiện với ba phương pháp học máy bao gồm RandomForest, XGBoost, SVR. Mỗi mô hình được triển khai và tinh chỉnh sao cho phù hợp với đặc điểm bài toán và đạt được độ chính xác cao nhất. Sau khi mô hình được xây dựng và tối ưu, kết quả của ba mô hình được đánh giá và so sánh nhằm lựa chọn ra phương pháp tối ưu nhất cho bài toán dự đoán hạn mức tín dụng.

CHƯƠNG 3: ĐỀ XUẤT ỨNG DỤNG VÀ CẢI TIẾN MÔ HÌNH

3.1. Một số khuyến nghị cải tiến và kết hợp mô hình

3.1.1. Cải thiện chất lượng dữ liệu đầu vào

Sử dụng các kỹ thuật chuẩn hóa khác như OneHot Encoding để chuẩn hóa dữ liệu về cùng một phạm vi, tránh cho mô hình bị nhầm lẫn giữa các giá trị phân loại khi dữ liệu không có mối quan hệ thứ tự.

Bổ sung thêm các nguồn dữ liệu khác như lịch sử tín dụng, hành vi chi tiêu của khách hàng, giúp cải thiện khả năng dự đoán của mô hình bằng cách cung cấp thêm thông tin cho mô hình.

Áp dụng các kỹ thuật như synthetic data generation để tạo thêm dữ liệu từ các khách hàng hiện tại, giúp mô hình có nhiều mẫu dữ liệu hơn từ đó cải thiện độ chính xác của mô hình khi dự đoán.

Sử dụng các phương pháp như SHAP (Shapley Additive Explanations), RFE (Recursive Feature Elimination) để đánh giá tầm quan trọng của từng đặc trưng trong mô hình, từ đó loại bỏ những đặc trưng ít quan trọng giúp giảm độ phức tạp và tránh hiện tượng overfitting.

3.1.2. Tối ưu hóa mô hình

Sử dụng thêm các kỹ thuật như k-fold cross validation để đánh giá mô hình trên nhiều tập con dữ liệu, giúp hiểu rõ hơn về độ ổn định và khả năng tổng quát của mô hình khi dự đoán các tập dữ liệu khác nhau.

Sử dụng các phương pháp như Bayesian Optimization, Grid Search để tìm kiếm các tham số tối ưu cho từng mô hình

Áp dụng stacking kết hợp các mô hình XGBoost, Random Forest và SVR trong một mô hình để tận dụng điểm mạnh của mỗi mô hình. Trong mô hình stacking kết quả dự đoán của các mô hình sẽ là cơ sở được sử dụng để làm đặc trưng đầu vào cho một mô hình thứ hai.

Áp dụng Early Stopping trong quá trình huấn luyện mô hình để ngừng huấn luyện khi mô hình không còn cải thiện độ chính xác giúp giảm thiểu overfititing.

Thử nghiệm mô hình với các giá trị random seed khác nhau giúp kiểm tra độ ổn định và chính xác của mô hình khi dự đoán nhiều tập dữ liệu khác nhau, đảm bảo mô hình không phụ thuộc vào bất kỳ khởi tạo ngẫu nhiên nào và có sự tổng quát tốt hơn.

3.1.3. Kết hợp mô hình dự đoán hạn mức tín dụng với mô hình phân khúc khách hàng

Mục tiêu: Phân khúc khách hàng theo nhóm hành vi và cấp hạn mức tín dụng phù hợp theo nhóm khách hàng.

Quy trình triển khai:

- Sử dụng các kỹ thuật phân khúc khách hàng như K-means để phân loại khách hàng thành các nhóm theo hành vi chi tiêu hoặc lịch sử tín dụng.
- Sau khi phân nhóm mô hình dự đoán hạn mức tín dụng có thể được điều chỉnh theo từng nhóm khách hàng. Ví dụ: nhóm khách hàng thường xuyên thanh toán đúng hạn và có mức chi tiêu cao có thể được cấp hạn mức tín dụng cao.
- Kết quả của mô hình phân khúc khách hàng có thể được sử dụng như một yếu tố đầu vào bổ sung cho mô hình dự đoán hạn mức tín dụng.

3.2. Đề xuất ứng dụng

Để giúp người dùng có thể dễ dàng tương tác và sử dụng các ứng dụng của học máy mà không cần biết quá nhiều kiến thức chuyên sâu, một giao diện người dùng được xây dựng kết nối trực tiếp với mô hình học máy đã được huấn luyện. Để xây dựng giao diện này, thư viện streamlit được sử dụng cho phép tạo ra các ứng dụng web tương tác mà không cần lập trình phức tạp. Streamlit giúp tạo giao diện đơn giản nhưng hiệu quả, người dùng có thể dễ dàng thao tác về nhận được kết quả mong muốn.

3.2.1. Quy trình triển khai

- Tải mô hình học máy: Sử dụng thư viện joblib để tải mô hình học máy XGBoost từ file đã lưu sau quá trình huấn luyện, mô hình XGBoost sẽ

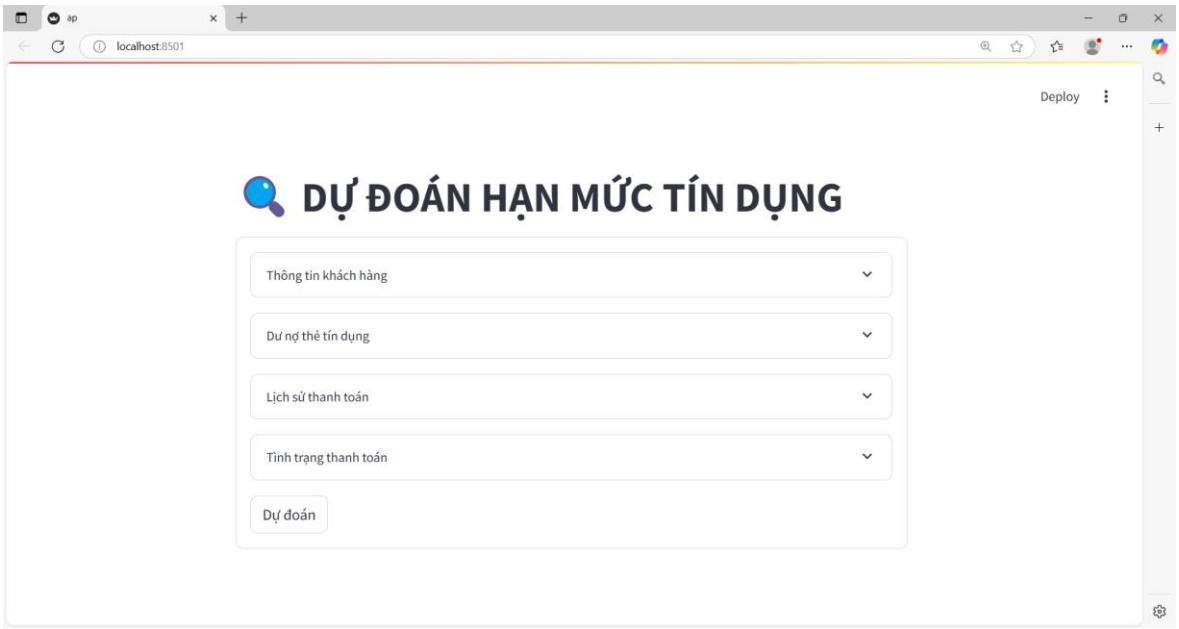
được sử dụng để thực hiện dự đoán hạn mức tín dụng dựa trên các thông tin người dùng nhập vào.

- Tạo các trường nhập thông tin: Sử dụng các hàm như st.number_input(), st.selectbox() để tạo các trường nhập thông tin người dùng cần thiết cho quá trình dự đoán hạn mức tín dụng như thu nhập, giới tính, tuổi.
- Chuẩn hóa dữ liệu đầu vào: Sau khi người dùng nhập đầy đủ các thông tin, các thông tin này sẽ được chuẩn hóa về cùng một phạm vi trước khi đưa vào mô hình dự đoán.
- Dự đoán và hiển thị kết quả: Sau khi chuẩn hóa, dữ liệu được đưa vào mô hình XGBoost để thực hiện dự đoán, mô hình dự đoán hạn mức tín dụng tối ưu cho người dùng dựa trên các thông tin đã nhập. Sử dụng hàm st.write() để trả về kết quả dự đoán về hạn mức tín dụng được tính toán và hiển thị trên giao diện người dùng.
- Quy trình xây dựng giao diện người dùng được trình bày chi tiết tại phụ lục 1.5.

3.2.2. Mô tả giao diện

Ứng dụng dự đoán hạn mức tín dụng được thiết kế với các trường nhập liệu rõ ràng giúp người dùng có thể dễ dàng nhập các thông tin cần thiết cho mô hình dự đoán. Các thành phần chính trong ứng dụng, bao gồm:

- Các trường nhập thông tin cá nhân như tên, tuổi và thu nhập.
- Các thẻ lựa chọn cho các tùy chọn như giới tính, trình độ học vấn và tình trạng hôn nhân.
- Các trường nhập liệu các chỉ số tài chính và lịch sử thanh toán như: dư nợ hóa đơn, tình trạng thanh toán.
- Nút “Dự đoán”: Sau khi nhập đầy đủ thông tin người dùng có thể nhấn nút này để nhận kết quả dự đoán hạn mức tín dụng tối ưu từ mô hình XGBoost.

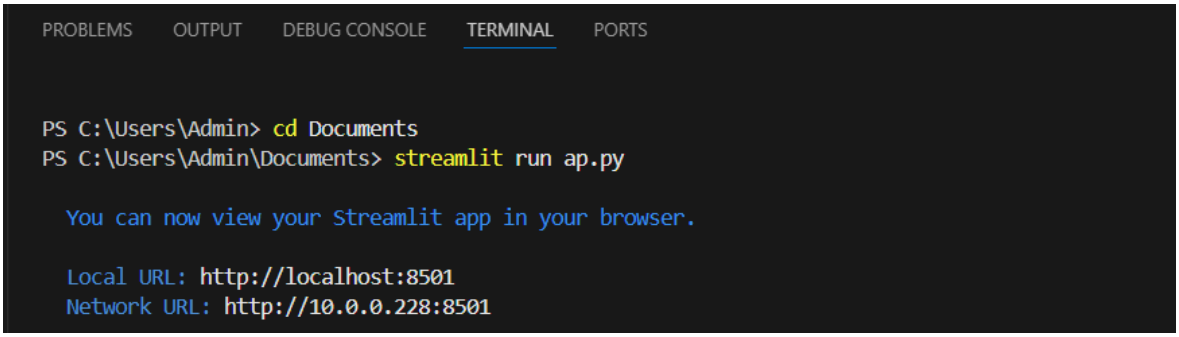


Hình 3.1: Giao diện web dự đoán hạn mức tín dụng

3.2.3. Hướng dẫn thực hiện

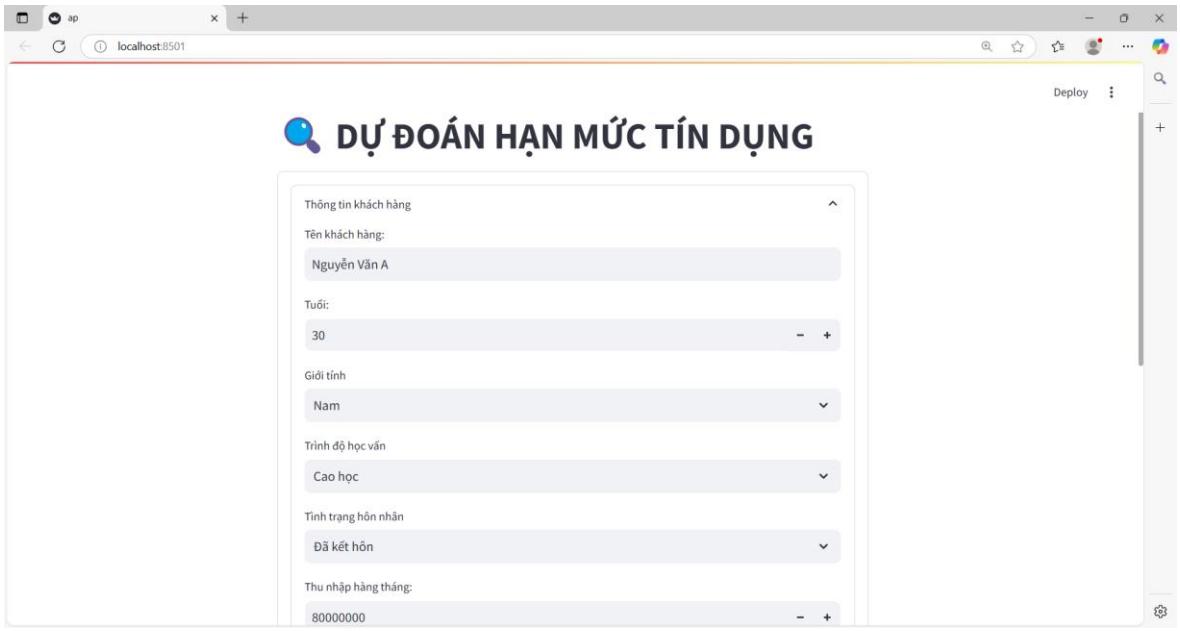
Bước 1: Tại giao diện dòng lệnh (Terminal) nhập lệnh "cd" để di chuyển đến thư mục lưu trữ file chứa mã nguồn ứng dụng streamlit. Ví dụ, nếu file ap.py nằm trong thư mục "Documents" thì nhập lệnh "cd Documents".

Bước 2: Để khởi động ứng dụng tiến hành nhập lệnh "streamlit run ap.py", lệnh này sẽ bắt đầu chạy ứng dụng và tự động mở ra một trang web trong trình duyệt.



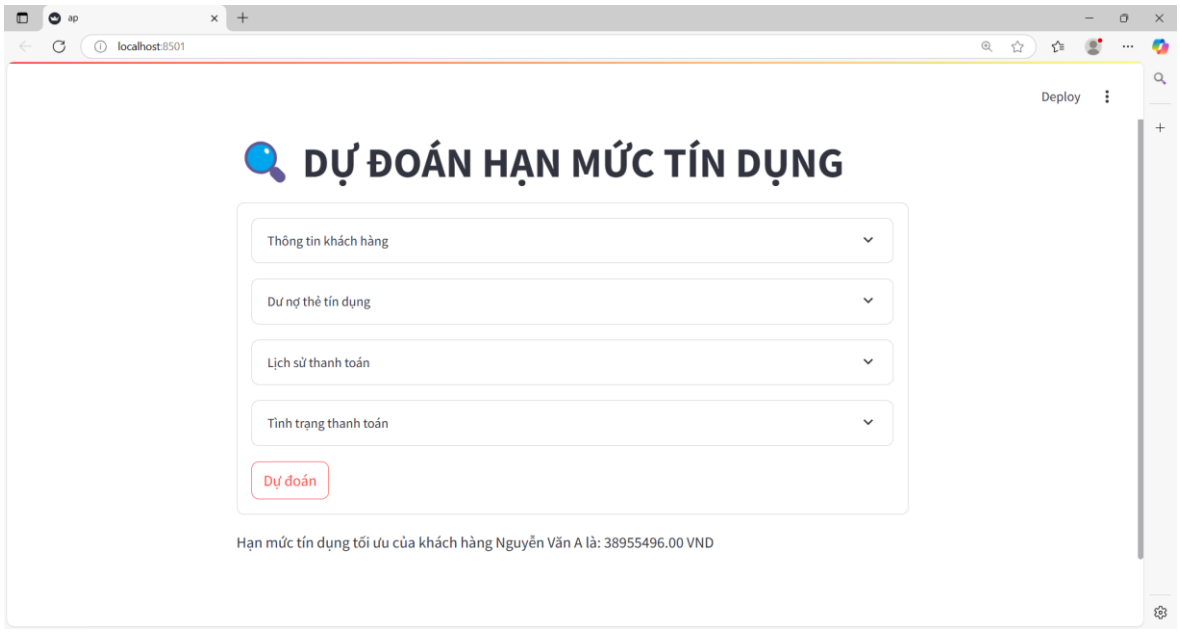
Hình 3.2: Khởi động ứng dụng dự đoán hạn mức tín dụng

Bước 3: Nhập đầy đủ các thông tin như thông tin khách hàng, tình trạng thanh toán, dư nợ thẻ tín dụng, lịch sử thanh toán.



Hình 3.3: Giao diện nhập thông tin khách hàng

Bước 4: Sau khi nhập đầy đủ các thông tin thì nhấn "Dự đoán" để nhận kết quả dự đoán hạn mức tín dụng của khách hàng.



Hình 3.4: Kết quả dự đoán hạn mức tín dụng

3.2.4. Một số tính năng nổi bật

Dự đoán nhanh chóng: Người dùng có thể nhận kết quả dự đoán hạn mức tín dụng ngay lập tức sau khi nhập đầy đủ các thông tin, giúp tiết kiệm thời gian và hỗ trợ quá trình ra quyết định nhanh chóng.

Giao diện dễ sử dụng: Giao diện trực quan, dễ dàng nhập thông tin và có hướng dẫn rõ ràng cho từng trường thông tin, giúp người dễ dàng sử dụng và tương tác.

Dự đoán chính xác: Sử dụng mô hình học máy đã được huấn luyện và có độ chính xác cao để tính toán giúp nâng cao mức độ chính xác khi đưa ra kết quả.

Ứng dụng cho phép người dùng có thể dễ dàng nhận được kết quả dự đoán nhanh chóng và chính xác mà không cần biết các kiến thức chuyên sâu về học máy hay lập trình python.

TIỂU KẾT CHƯƠNG 3

Chương 3 của khóa luận đã tập trung trình bày các khuyến nghị cải tiến và kết hợp mô hình nhằm nâng cao độ chính xác của mô hình học máy. Các phương pháp cải tiến được xây dựng dựa trên cơ sở lý thuyết đã phân tích ở các chương trước, đồng thời có sự điều chỉnh để phù hợp với yêu cầu bài toán. Cụ thể các cải tiến được áp dụng như áp dụng các kỹ thuật lựa chọn tính năng, tối ưu hóa tham số và mở rộng kết hợp với các bài toán khác.

Ngoài việc nâng cao độ chính xác của mô hình học máy dự đoán hạn mức tín dụng, chương 3 còn đề xuất xây dựng ứng dụng web để triển khai mô hình học máy vào thực tế, giúp người dùng có thể dễ dàng sử dụng và tương tác linh hoạt hơn. Giao diện web được thiết kế để hỗ trợ người dùng dễ dàng nhập thông tin và nhận kết quả dự đoán mà không cần quá nhiều kiến thức về lập trình. Điều này không chỉ giúp cải thiện trải nghiệm người dùng mà còn mở ra cơ hội ứng dụng mô hình vào các lĩnh vực khác trong thực tiễn.

KẾT LUẬN

Khóa luận này tập trung vào việc xây dựng các mô hình học máy để dự đoán hạn mức tín dụng của khách hàng tại các ngân hàng thương mại Việt Nam và triển khai ứng dụng vào thực tế thông qua một giao diện người dùng. Kết quả của đề tài không chỉ chứng minh tính hiệu quả của việc áp dụng học máy vào các quy trình quản lý tín dụng mà còn mở ra nhiều ứng dụng thực tiến hữu ích trong thực tế.

Ưu điểm nổi bật:

- Hiệu suất mô hình đạt kết quả cao: Mô hình XGBoost và Random Forest cho thấy khả năng vượt trội trong việc xử lý dữ liệu phức tạp và đạt được các chỉ số đánh giá cao vượt trội, chứng minh sự phù hợp của mô hình đối với bài toán dự đoán hạn mức tín dụng.
- Ứng dụng thực tế: Kết quả huấn luyện của mô hình được sử dụng để xây dựng một giao diện người dùng thông qua thư viện streamlit, giúp người dùng có thể dễ dàng tương tác, nhập dữ liệu và nhận được kết quả dự đoán một cách nhanh chóng và hiệu quả mà không đòi hỏi các kiến thức về công nghệ và học máy.
- Giá trị chuyên ngành: Quá trình thực hiện khóa luận mang lại những bài học hữu ích về việc áp dụng phân tích dữ liệu và học máy vào trong thực tế.

Hạn chế:

- Dữ liệu giả định: Một số phần dữ liệu mang tính giả định và chưa đủ phong phú có thể làm giảm độ chính xác của mô hình khi áp dụng vào các tình huống thực tế.
- Dữ liệu chưa đa dạng: Dữ liệu thu thập được vẫn còn thiếu một số yếu tố quan trọng để phù hợp với thực tế hơn, việc bổ sung dữ liệu sẽ giúp cải thiện độ chính xác và phù hợp với các tình huống thực tế hơn.
- Hạn chế trong việc sử dụng các kỹ thuật học máy: Các phương pháp học máy được sử dụng trong khoa luận có thể thiếu linh hoạt trong việc xử

lý dữ liệu phức tạp và phi tuyến tính, đặc biệt khi dữ liệu lớn và có sự thay đổi nhanh chóng.

Quá trình thực hiện khóa luận không chỉ cung cấp cho sinh viên cơ hội được áp dụng kiến thức vào các bài toán trong thực tế mà còn tạo nền tảng để sinh viên tiếp tục tìm hiểu, mở rộng các ứng dụng phân tích dữ liệu và học máy trong lĩnh vực kinh tế và kinh doanh. Đây là cơ hội để sinh viên nâng cao kỹ năng và kiến thức, đóng góp vào các giải pháp sáng tạo và góp phần thúc đẩy ứng dụng học máy trong các lĩnh vực đời sống.

TÀI LIỆU THAM KHẢO

[1] Đào Mỹ Hằng & Nguyễn Thị Hòa (13/03/2025). *Hoạt động quản lý hạn mức tín dụng tại các ngân hàng thương mại Việt Nam trong bối cảnh chuyển đổi số*. Truy xuất ngày 15/03/2025 từ https://tapchinganhang.gov.vn/hoat-dong-quan-ly-han-muc-tin-dung-tai-cac-ngan-hang-thuong-mai-viet-nam-trong-boi-canh-chuyen-doi-so-15597.html

[2] Đỗ Minh Hiếu (06/02/2024). *Hạn mức tín dụng là gì? Hạn mức tối đa của thẻ tín dụng là bao nhiêu?* Truy xuất ngày 15/03/2025 từ https://thuvienphapluat.vn/banan/tin-tuc/han-muc-tin-dung-la-gi-han-muc-toi-da-cua-the-tin-dung-la-bao-nhieu-9320

[3] Wikipedia (27/02/2024). *Hạn mức tín dụng*. Truy xuất ngày 15/03/2025 từ https://vi.wikipedia.org/wiki/H%E1%BA%A1n_m%E1%BB%A9c_t%C3%ADn_d%E1%BB%A5ng

[4] Wikipedia (03/10/2024). *Tín dụng.* Truy xuất ngày 15/03/2025 từ https://vi.wikipedia.org/wiki/T%C3%ADn_d%E1%BB%A5ng

[5] VPBank (22/09/2023). *Hạn mức tín dụng là gì? Hướng dẫn cách nâng hạn mức tín dụng*. Truy xuất ngày 15/03/2025 từ https://www.vpbank.com.vn/bi-kip-va-chia-se/retail-story-and-tips/loans-category/han-muc-tin-dung

[6] Khoa Kinh tế Quốc tế, Đại học Văn Hiến. (n.d). *Ứng dụng trí tuệ nhân tạo và phân tích dữ liệu lớn trong chuyển đổi số ngành ngân hàng*. Truy xuất ngày 15/03/2025 từ https://ktqt.vhu.edu.vn

[7] MCI Vietnam. (2023). Ứng dụng Machine Learning trong ngành tài chính: Dự đoán rủi ro tín dụng, phát hiện gian lận và tối ưu hóa. Truy xuất ngày 17/03/2025 từ https://www.mcivietnam.com

[8] Ng, A.(2018). Machine learning yearning: Technical strategy for AI engineers in the era of deep learning. Truy xuất ngày 17/03/2025 từ https://www.deeplearning.ai

[9] Zhou, Z., & Yang, Y. (2019). Credit scoring using machine learning: A survey. *Processdings of the 2019 IEEE 3rd International Conference on*

Computing, Data Communication and Intelligent System (CDIS), 123-128. Retrieved March 17, 2025 from https://ieeexplore.ieee.org/document/8981329

[10] Ledhem, A. (2022). Predicting bank performance using machine learning algorithms. *Journal of Economic Development*, 24(1), 15-29. Retrieved March 17, 2025 from https://www.jedjournal.com

PHỤ LỤC

ĐOẠN MÃ XÂY DỰNG MÔ HÌNH HỌC MÁY TRONG DỰ BÁO VÀ TỐI ƯU HẠN MỨC TÍN DỤNG TẠI NGÂN HÀNG THƯƠNG MẠI VIỆT NAM

1.1. Đọc dữ liệu

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
import warnings
warnings.filterwarnings('ignore')

# Đọc dữ liệu từ file CSV
file_path = "D:/New folder/default of credit card clients.csv"
df = pd.read_csv(file_path)
print(df.info())
```

1.2. Tiền xử lý dữ liệu

1.2.1. Mã hóa dữ liệu

```
# Mã hóa SEX
df["SEX"] = df["SEX"].map({"Nam": 1, "Nữ": 0})

# Mã hóa MARRIAGE
marriage_mapping = {"Đã kết hôn": 1, "Độc thân": 2, "Khác": 3, "0": 3}
df["MARRIAGE"] = df["MARRIAGE"].map(marriage_mapping)

# Gộp nhóm EDUCATION (0, 4, 5, 6 thành "Khác" với giá trị là 4)
df["EDUCATION"] = df["EDUCATION"].replace({0: 4, 5: 4, 6: 4})
```

1.2.2. Xóa cột

```
# Xóa cột 'default payment next month'
df = df.drop(columns=['default payment next month'])
```

1.2.3. Tạo các đặc trưng mới

```
# Tỷ lệ sử dụng tín dụng
bill_balance_cols = ["BILL_BALANCE_Apr", "BILL_BALANCE_May", "BILL_BALANCE_Jun",
                     "BILL_BALANCE_Jul", "BILL_BALANCE_Aug", "BILL_BALANCE_Sept"]
df["CUR"] = df[bill_balance_cols].mean(axis=1) / (df["LIMIT_BAL"] + 1)

# Số dư hóa đơn
df["bill_balance"] = df[bill_balance_cols].mean(axis=1)

# Ước lượng thu nhập dựa trên hạn mức tín dụng
df["estimated_income"] = df["LIMIT_BAL"] * 1/3
```

1.2.4. Xử lý ngoại lai

```
# Xử lý ngoại lai
def remove_outliers(df, columns):
    for col in columns:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        # Loại bỏ các dòng có giá trị ngoài phạm vi IQR
        df = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)]
    return df

# Các cột cần xử lý ngoại lai
columns_to_check = ['LIMIT_BAL', 'CUR', 'estimated_income', 'BILL_BALANCE_Apr', 'BILL_BALANCE_May',
                    'BILL_BALANCE_Jun', 'BILL_BALANCE_Jul', 'BILL_BALANCE_Aug', 'BILL_BALANCE_Sept',
                    'PAY_AMOUNT_Apr', 'PAY_AMOUNT_May', 'PAY_AMOUNT_Jun', 'PAY_AMOUNT_Jul', 'PAY_AMOUNT_Aug', 'PAY_AMOUNT_Sept']

# Áp dụng xử lý ngoại lai
df_cleaned = remove_outliers(df, columns_to_check)

# Kiểm tra số dòng sau khi xử lý
print(f"Số dòng sau khi xử lý ngoại lai: {df_cleaned.shape[0]}")
```

1.3. Khám phá dữ liệu

1.3.1. Mối quan hệ giữa các yếu tố nhân khẩu học và hạn mức tín dụng

```
demo = df[['SEX', 'AGE', 'MARRIAGE', 'EDUCATION', 'LIMIT_BAL']]
correlation_matrix_demo = demo.corr()
plt.figure(figsize=(6, 4))
sns.heatmap(correlation_matrix_demo, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.tight_layout()
plt.show()
```

1.3.2. Mối quan hệ giữa hạn mức tín dụng và độ tuổi

```
bins = [20, 30, 40, 50, 60, 70, 80]
names = ['21-30', '31-40', '41-50', '51-60', '61-70', '71-80']
df['AGE_GROUP'] = pd.cut(df['AGE'], bins=bins, labels=names, right=True)
plt.figure(figsize=(6, 4))
sns.boxplot(x='AGE_GROUP', y='LIMIT_BAL', data=df, palette='Set2')
plt.xlabel('Nhóm tuổi', fontsize=12)
plt.ylabel('Hạn mức tín dụng', fontsize=12)
plt.xticks(rotation=45, fontsize=12)
plt.yticks(fontsize=12)
plt.tight_layout()
plt.show()
```

1.3.3. Phân phối hạn mức tín dụng

```
# Phân tích phân phối của hạn mức tín dụng
plt.figure(figsize=(8, 5))
sns.histplot(df["LIMIT_BAL"], bins=50, kde=True, color="blue")
plt.title("Phân phối hạn mức tín dụng")
plt.xlabel("Hạn mức tín dụng")
plt.ylabel("Tần suất")
plt.show()
```

1.3.4. Phân bổ hạn mức tín dụng trung bình dựa trên trình độ học vấn

```
edu_avg_limit = df.groupby('EDUCATION')['LIMIT_BAL'].mean()
labels = ['Cao học', 'Đại học', 'Trung học', 'Khác']
plt.figure(figsize=(6,6))
plt.pie(edu_avg_limit, labels=labels, autopct='%1.1f%%', startangle=140, explode=[0.05]*4)
plt.axis('equal')
plt.show()
```

1.4. Xây dựng mô hình học máy

1.4.1. Chia tập dữ liệu

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
X = df.drop(columns=["LIMIT_BAL", "ID"])
y = df["LIMIT_BAL"]

# Chia dữ liệu thành tập huấn luyện và kiểm tra
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

1.4.2. Xây dựng mô hình Random Forest

```
# Khởi tạo các tham số
rf_model = RandomForestRegressor(
    n_estimators=100,
    min_samples_split=2,
    min_samples_leaf=1,
    max_features='log2',
    max_depth=30,
    bootstrap=False,
    random_state=42
)
rf_model.fit(X_train_scaled, y_train)

# Dự đoán trên tập kiểm tra
y_pred_rf = rf_model.predict(X_test_scaled)

# Đánh giá mô hình
rmse = mean_squared_error(y_test, y_pred_rf, squared=False)
mae = mean_absolute_error(y_test, y_pred_rf)
r2 = r2_score(y_test, y_pred_rf)
```

```
plt.figure(figsize=(6,4))
plt.scatter(y_test, y_pred_rf, color='blue', alpha=0.5)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', linestyle='--')
plt.title('Actual vs Predicted')
plt.xlabel('Actual LIMIT_BAL')
plt.ylabel('Predicted LIMIT_BAL')
plt.show()
```

1.4.3. Xây dựng mô hình XGBoost

```
# Khởi tạo và huấn luyện mô hình
from xgboost import XGBRegressor
xgb_model = XGBRegressor(
    n_estimators=50,
    max_depth=5,
    learning_rate=0.1,
    subsample=0.8,
    colsample_bytree=1.0,
    random_state=42
)

# Huấn luyện mô hình
xgb_model.fit(X_train_scaled, y_train)

# Dự đoán trên tập kiểm tra
y_pred_xgb = xgb_model.predict(X_test_scaled)

# Đánh giá mô hình
rmse = mean_squared_error(y_test, y_pred_xgb, squared=False)
mae = mean_absolute_error(y_test, y_pred_xgb)
r2 = r2_score(y_test, y_pred_xgb)
```

```
plt.figure(figsize=(6,4))
plt.scatter(y_test, y_pred_xgb, color='green', alpha=0.5)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', linestyle='--')
plt.title('XGBoost - Actual vs Predicted')
plt.xlabel('Actual LIMIT_BAL')
plt.ylabel('Predicted LIMIT_BAL')
plt.show()
```

1.4.4. Xây dựng mô hình SVR

```
from sklearn.svm import SVR
# Khởi tạo mô hình Support Vector Regression (SVR)
svr_model = SVR(
    kernel='rbf',
    C=100,
    epsilon=0.1
)

# Huấn luyện mô hình
svr_model.fit(X_train_scaled, y_train)

# Dự đoán trên tập kiểm tra
y_pred_svr = svr_model.predict(X_test_scaled)
# Đánh giá mô hình
rmse = mean_squared_error(y_test, y_pred_svr, squared=False)
mae = mean_absolute_error(y_test, y_pred_svr)
r2 = r2_score(y_test, y_pred_svr)
```

```
plt.figure(figsize=(6,4))
plt.scatter(y_test, y_pred_svrsvr, color='green', alpha=0.5)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', linestyle='--')
plt.title('SVRSVR - Actual vs Predicted')
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.show()
```

1.5. Xây dựng giao diện người dùng

```
import streamlit as st
import pandas as pd
import numpy as np
import joblib
from sklearn.preprocessing import StandardScaler

# Tải mô hình đã huấn luyện từ file
xgb_model = joblib.load('xgb_model.pkl')
scaler = joblib.load('scaler.pkl')

# Tạo giao diện web
st.title('🔍 DỰ ĐOÁN HẠN MỨC TÍN DỤNG')
AGE = st.number_input("Tuổi: ", min_value=18, format="%d")
SEX = st.selectbox("Giới tính", ["Nam", "Nữ"])
EDUCATION = st.selectbox("Trình độ học vấn", ["Cao học", "Đại học", "Trung học phổ thông", "Khác"])
MARRIAGE = st.selectbox("Tình trạng hôn nhân", ["Đã kết hôn", "Độc thân", "Khác"])
estimated_income = st.number_input("Thu nhập hàng tháng: ", min_value=0, format="%d")
bill_balance = st.number_input("Số dư hóa đơn trung bình trong 6 tháng: ", min_value=0.0, format="%f")
CUR = st.number_input("Tỷ lệ sử dụng tín dụng:", min_value=0.0)
```

```python
# Khi nhấn nút dự đoán
if st.button("Dự đoán"):
    # Chuyển đổi các giá trị nhập vào thành một vector cho mô hình
    SEX = 1 if SEX == "Nam" else 0
    EDUCATION = 1 if EDUCATION == "Cao học" else (2 if EDUCATION == "Đại học" else (3 if EDUCATION == "Trung học phổ thông" else 4))
    MARRIAGE = 1 if MARRIAGE == "Đã kết hôn" else (2 if MARRIAGE == "Độc thân" else 3)

    # Tạo feature input cho mô hình
    features = np.array([AGE, SEX, EDUCATION, estimated_income, bill_balance,
                         BILL_BALANCE_Apr, BILL_BALANCE_May, BILL_BALANCE_Jun,
                         BILL_BALANCE_Jul, BILL_BALANCE_Aug, BILL_BALANCE_Sept,
                         PAY_STATUS_Apr, PAY_STATUS_May, PAY_STATUS_Jun,
                         PAY_STATUS_Jul, PAY_STATUS_Aug, PAY_STATUS_Sept,
                         PAY_AMOUNT_Apr, PAY_AMOUNT_May, PAY_AMOUNT_Jun,
                         PAY_AMOUNT_Jul, PAY_AMOUNT_Aug, PAY_AMOUNT_Sept,
                         MARRIAGE, CUR]).reshape(1, -1)

    # Chuẩn hóa dữ liệu với scaler đã huấn luyện
    features_scaled = scaler.transform(features)

    # Dự đoán từ mô hình
    prediction = xgb_model.predict(features_scaled)

    # Hiển thị kết quả dự đoán
    st.write(f"Hạn mức tín dụng tối ưu: {prediction[0]:.2f} VND")
```